UNIVERSITY OF TECHNOLOGY, SYDNEY
FACULTY OF INFORMATION TECHNOLOGY

A Strategic Analytics Methodology

By Marcel van Rooyen

A dissertation submitted in fulfillment of the requirements for the degree of Master in Computing Sciences

2005

.

Acknowledgements

Associate Professor Simeon Simoff for his vision of what we had to achieve together in this novel area of research, his insightful guidance about how to achieve it, and his encouragement to perseverance at times of frustration. Professor Ken Miller for his direction about marketing subject matter.

Family and friends for their faith in me during a mid-life career change, and for the neglect they endured for it. Yong Mei Tang (Grace) for her unquestionable support during unquestionable neglect, and for supporting with graphic design.

The University of Technology for an RTS Scholarship and the fantastic research infrastructure.

SAS Institute Australia Pty. Ltd. and key individuals there, for access to their software, training support, professional networking opportunities, access to proprietary research literature, and the introduction to the industrial research partner.

An Australian mobile phone company for the industrial research partnership. In particular their Retention and Business Intelligence Managers for their support and interest in the project, and two analysts who endured me while transferring invaluable knowledge to me.

This work is a celebration of the pleasant surprises life presents to the fortuitous and adventurous

Table of contents

Certificate of Authorship and Originality

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that this thesis has been written by me. Any help that I received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that al information sources and literature used are indicated in the thesis.

Signature of Candidate

_____

List of figures

List of tables

Abstract

Commercial organisations are dependent on generating profit from competitive advantage. Central to this approach, is the Strategic Planning Cycle (SPC). SPC converts new information and new subject matter expertise into competitive knowledge, and then converts that knowledge into executable solutions best suited to the organisation's internal and external circumstances and resources. SPC also maintains the relevance and efficiency of the executed solutions over time.

In order to optimise competitiveness, organisations seek to improve SPC in a number of ways. First, they improve the quality of the informational inputs to SPC. Second, they improve the quality of the knowledge which they develop from that information. Third, they optimise the executibility of the solutions, which were based on the knowledge, for the organisation's particular circumstances and resources. Four, they improve the solutions over time, maintaining competitiveness. All four ways of improving SPC are supported by data analytics. It is therefore a necessity ever to improve the integration of data analytics with SPC.

Data mining is an advanced analytics approach, which has been shown to support SPC. Recognising the complexity of integrating data analytics with the business at the turn of the 21st century, the analytics community developed data mining project methodologies to facilitate the integration. The most widely published methodology is CRISP-DM. SAS Institute's SAS Data Mining Projects Methodology (SDMPM) is a second, albeit proprietary, methodology which is also widely used.

Despite the availability of packaged data mining software and project methodologies for more than a decade now, organisations are still finding the integration of data mining with the SPC process complex and daunting. The current situation is that business leaders and data analysts often express the need for better integration of data analytics with SPC and business goals.

The researcher hypothesized that the data mining project methodologies may be a major contributor to the above situation. The researcher therefore formulated the research objective of evaluating data mining methodology for its support of the SPC process. The CRISP-DM methodology was chosen for evaluation because it is in the public domain

and therefore available to other researchers. (The researcher has evaluated SDMPM in a separate paper.)

The research method chosen was Participatory Action Research, specifically that of action science or *expert reflection-in-action*. The research was industry-based, using data from a real-life Telco customer retention management problem. The researcher and the Telco formulated a data analytics project using CRISP-DM. The project was in support of the Telco's strategic initiative drastically to reduce customer churn in their consumer business.

The data mining project would support the initiative in three ways. First, it would predict customer churn behaviour within an upcoming time window. Second, it would segment the most at-risk customers in strategic marketing dimensions. Third, it would profile the segments in dimensions required for retention campaign re-design.

Using expert reflection-in-action, we evaluated the operating and strategic outcome for the Telco, from the project that was formulated using CRISP-DM as the project methodology. The research findings were that the project based on CRISP-DM would be limited in its executibility and strategic impact. This would severely restrict the competitive advantage realisable from the project.

Our research identified six key limitations of CRISP-DM in the SPC environment:

- diagnostic technique for defining the project's business goals or business deliverables. This is about defining the required informational and marketing components required for the strategic initiative;
- introduction of new business and analytics subject matter expertise into the project environment. This relates to increasing the understanding of the business problem and its possible solutions through new marketing and data mining subject matter expertise;
- mapping technique between the project's business deliverables and the supporting data mining plan. This is about assuring that the data analytics best support the project's business deliverables;
- knowledge management activities required by SPC for assessing discovered information against business deliverables, environmental and circumstantial

factors, for adapting the information, and for developing competitive, executable business solutions;

- monitor and control of business and data mining solutions over time for effectiveness and efficiency; and
- a number of soft project and business solution implementation issues.

The main *research goal, which* flowed from the above finding, was to develop a new, more potent data mining project methodology for the SPC environment. In developing this methodology, the researcher used concepts from the Business, Knowledge Discovery, and Data Mining literature, also drawing on his previous corporate management experience and MBA qualification. The researcher called the new method *Strategic Analytics Method* (SAM).

Essentially SAM is the integration of data analytics project methodology and a proven SPC tool, which is known as Strategic Planning Method (SPM). SPM is a generic decision-making process designed for producing competitive outcomes under conditions of uncertainty and limited resources. SPM is widely used in various guises by business, software engineering, the military, and many other applications.

SAM presents a major departure from CRISP-DM's data centricity, to a project centered on the project's business deliverables. SAM is targeted at data miners and data analysts working in a commercial environment, and at business intelligence practitioners.

Practically SAM contributes the following to data mining projects methodology:

- moving the focus from data-related activities to business deliverables;
- insights about the restrictive impact of the pre-project *status quo* on the results of the project, the dimensions of the status quo which must be defined into a business problem, and how to achieve that definition;
- technique for injecting new business and analytics subject matter into the stale business environment, to enable competitive breakthrough;
- technique for developing business deliverables or goals for the project, which will be competitive. This includes considering the new subject matter, and overcoming the restrictions presented by the current understanding of the *status quo*;
- mapping technique between the project's business deliverables and the data mining plan, which assures the data mining outputs optimally supporting the attainment of the business deliverables;

- technique for assessing discovered information for its relevance to the business deliverables;
- knowledge management activities for developing the discovered information into competitive business solutions which are executable under the organisation's limited resources and limiting circumstances;
- substantial qualitative and quantitative technique for developing monitor and control plans for both the analytics and the business solution;
- activities, which pro-actively manage soft issues before they impact on the project negatively. For instance, we reframe data preparation activities as a process, which gradually reduces project risk associated with the data. This offers more understandable and acceptable justification to the business audience about this resource-intensive part of data mining projects;
- insights for distinguishing between *iteration* and *repetition* of activities on advanced SPC projects, and technique for knowing when to start and stop iterating, or repeating. This distinction provides contextual vocabulary for communicating with the business about required project effort.

The research validates SAM on the same Telco ABC problem, which was used for evaluating CRISP-DM. The validation came through being able to formulate a project using SAM in which we:

- assisted Telco ABC in breaking through their limited pre-project marketing perceptions and expectations, to formulate business deliverables based on new marketing and analytics subject matter, which constituted competitiveness in customer retention management;
- formulated and executed a data mining project which produced the information required by the business deliverables;
- improved the Telco's calculation of the extent of the problem;
- developed knowledge from the discovered information which complemented applicable new marketing subject matter;
- developed the knowledge into a competitive retention management solution executable under the Telco's limiting circumstances and limited campaign resources. We presented the solution as new marketing objectives and strategies, and developed these into a retention campaign strategy with various key components;
- developed a comprehensive monitor and control plan for the campaigns and the operationalised data analytics solution;
- quantified the project ROI as about 187 times the investment.

May 2006 m_vanrooyen@hotmail.com / +61-402-032-059

1 Chapter 1

In this chapter, we first describe the context of the research problem, and how we identified the research problem. We then define the research problem, and our goals with it. The research objectives, which support those goals, are formulated, and their scope beaconed.

We then define the research methods – or strategies – which we follow in attaining the research objectives. The research outcomes are described in terms of targeted audiences. Last, we present the logic for the chapter sequence in the thesis, and a chapter summary.

1.1 Context of the research problem

Goal-orientated organisations with limited resources are finding it increasingly difficult to generate competitive advantage from an ongoing optimisation of traditional resources e.g. material and financial resources (Porter 2002, p.36). It is generally accepted that a phase of economic development is unraveling, in which competitive advantage will increasingly depend on innovative business decision-making (Ruthven 2002). The decision-making will be aimed at innovatively solving problems and meeting opportunities. It becomes evident that, since the aim is concerned about *doing* something about problems or opportunities, the value of the decision making will greatly depend on its executibility as solutions (Meltzer 2000, p.1) (Grant 1996, p.375).

The business process which unites the decision-making with the competitiveness and the executibility of those solutions, is called Business Intelligence (BI) (Liu 2003, p.429) (Pyle 2004, pp.54ff, 165, 662) . An important aspect of BI is that it optimizes the decisions, and their executibility, for limiting internal and external organizational circumstances. Such optimisation is often expressed in terms of paradigm shift (Pyle 2004, pp.13, 18).

Executable solutions need to be executed to be effective. Because organizations operate in a dynamic environment, the executed solutions further have to be maintained effectiveness and efficiency over time. The business process which combines BI with the execution and maintenance of the solutions, is called the Strategic Planning Cycle (SPC) (Pearce and Robinson 2004).

The input to the SPC is information. Information on its own, however, is not executable. One main function of the SPC process therefore, is transforming the information into executable solutions. This transformation is achieved via a sequence of perceptive, interpretive, comparative, and decisive cognitive activities. SPC practitioners complement their cognition with technical analytics tools, which generate and manipulate information.

An example of increasing SPC activity, is understanding customer behavior and intention. The purpose of SPC there is formulating business solutions, which realise the business potential of those behaviors and intentions (King 2003, p.1). Much of the information for this application of SPC is resident in the organisation's data, where it is spread across disparate data domains of finance, marketing, etc. Because of the volume of the data – in the realm of petabytes in some cases - and the scope of the data, this information remains hidden from human observation and even from discovery by classical analytical methodologies.

Data mining techniques were developed in the area of machine learning (ML) (Hastie, Tibshirani et al. 2001) (Berthold and Hand 2003), where they are successfully applied toward solving structured problems. Examples of data mining applications in a structured automated environment are the monitoring of parameter measurements of chemical manufacturing processes, and the automatic adjustment of the process to stay within the required parameter measures. Further examples are the automatic adjustment of robot movements on motor vehicle assembly lines to compensate for the wear-down of the welding rod, and the monitoring of on-line buyer responses in the e-commerce environment. An example of the last is an online book retailer automating the changing of the sales rating of books according to buyer behavior.

The application of data mining is described as effectively *...learning from data.* (Hastie, Tibshirani et al. 2001, p.vii), or Knowledge Discovery from Databases (KDD). The *potential* for using data mining techniques as a learning tool in the BI environment has been recognised for a number of years (Ganti, Gehrke et al. 1999). In fact, data mining algorithms have already been adapted for this environment, and successfully used for the discovery of information in massive commercial data bases (Fayyad, Piatetsky-Shapiro et al. 1996) (Han and Kamber 2001, pp. xix, 4) (Levinson 2000).

Data mining is also being applied in the SPC environment. Examples of this application is manufacture process debugging, fraud detection and AML (anti-money laundering) in financial markets, and behavioral propensity modeling in telecommunications, banking and insurance (Rud 2001). Retail and banking use data mining for identifying cross-sell and up-sell opportunities, for segmenting customers on value and behavior, and for profiling those segments for positioning of the marketing campaigns.

In this thesis, we refer to the application of data mining in the whole SPC. The successful transfer of data mining into the SPC environment has been possible due to similarities between the ML and SPC data mining environments. The main similarities are:

- the existence of a real-life problem or opportunity which is reflected in data; and
- expert reflection about:
  - the root indicators of the problem or opportunity;
  - how to overcome the problem or meet the opportunity, with a solution which addresses root indicators; and
  - keeping the solution relevant to changes in the problem or solution over time lapsed.

There are also differences between the ML and SPC environments, which have complicated the use of data mining in the SPC environment. We summarise these differences in Table 1.1:

Difference	ML environment	SPC environment
Problem structure:	Relatively structured and technical.	Relatively unstructured and non-technical (Gibson, Ivanchevich et al. 1991, p.576).
Problem boundaries:	Defined	Indefinite
Overlap in expertise:	Substantial overlap in the data mining and problem domain experts' quantitative skills.	Relatively insubstantial overlap in the data mining and business domain experts' quantitative skills.

Difference	ML environment	SPC environment
Nature of solution:	Technical and methodical, therefore often automatable.	Combination of technology and human activity, relatively less methodical, therefore not readily automatable.
Size of problem data domain:	In the order of hundreds to ten thousands of observations.	In the order of ten thousands to millions of observations.
Problem diagnostic technique:	○ Advanced in formulating the real-world problem; ○ Advanced in formulating the accompanying data mining problem; ○ Advanced in identifying key discovered informational components that require monitoring over time.	○ Advanced in formulating the real-world problem; ○ Relatively less advanced in defining the accompanying data mining problem; ○ Relatively less advanced in identifying key discovered informational components that require monitoring over time.
Problem response to solution:	Certainty about the problem's response to its technical solution, due to the problem's structured and definite nature.	Relatively less certainty about the problem's response to its business solution, due to the problem's unstructured and indefinite nature.
Control of solution for ongoing relevance over time:	Very effective and mostly automated updating of the solution in a changing problem environment.	Very effective, mostly non-automatic updating of the solution in a changing problem environment.

Table 1.1: Differences between SPC and ML data mining environments

Data mining project methodology is a tool for assuring the integration of data mining projects with business goals (Pyle 1999, p.10) (Liu 2003, p.436). Recognising this, and the potential BI and SPC benefits from data mining, and the complexity of introducing data mining into this new environment, the data mining community developed two data mining methodologies more than six years before the time of this writing in 2005.

The one methodology was *SAS Data Mining Projects Methodology* (=SDMPM) (SAS Institute 2000). This methodology is proprietary intellectual property, and its review by the researcher therefore excluded from publication in this thesis. At the time of writing in 2005, the researcher had published his review of SDMPM separately (Van Rooyen 2004).

The other methodology which was developed, was CRISP-DM (Chapman, Clinton et al. 1999-2000). CRISP-DM is the most comprehensive *published* data mining project methodology (Grossman, Hornick et al. 2003, p.458). For this reason we used our review of CRISP-DM as the basis for this research.

The stated purpose of both methodologies is to provide utility for assuring the usefulness of the outcome of data mining projects to the business community (SAS Institute 2000, p.xi) (Chapman, Clinton et al. 1999-2000, p.3). CRISP-DM presents itself as a standard for data mining projects which is *...sufficiently mature to be adopted as a key part of their business processes...*(Chapman, Clinton et al. 1999-2000, p.3).

Despite the development of the above data mining project methodologies, there have been ongoing expressions of the need for better integration data mining and business goals in the SPC environment. Shortly after the publication of CRISP-DM an industry analyst cautioned about the completeness of CRISP-DM for this purpose (Linden 2000). The most striking formulation of this problem the researcher found, was the need for '…a codified, competent data analysis strategy for the voluminous data environment, which formally collaborates what both the analyst and the computer is good at.' (Liu 2003, pp.436, 443). More recent sources have confirmed the ongoing need (Fayyad 2004) (Pyle 2004) (Han 2004).

The researcher has corporate project management experience in implementing various types of supporting technologies of business solutions. In these environments, project methodologies always played a vital role in assuring successful project outcomes. In light of the aforesaid need for better integration between the business needs and the data mining technology, the candidate was particularly interested in the utility of the data mining project methodologies for their stated purpose in the business environment.

We identify the research *context* and its elements, expressing them visually in Figure 1.1:

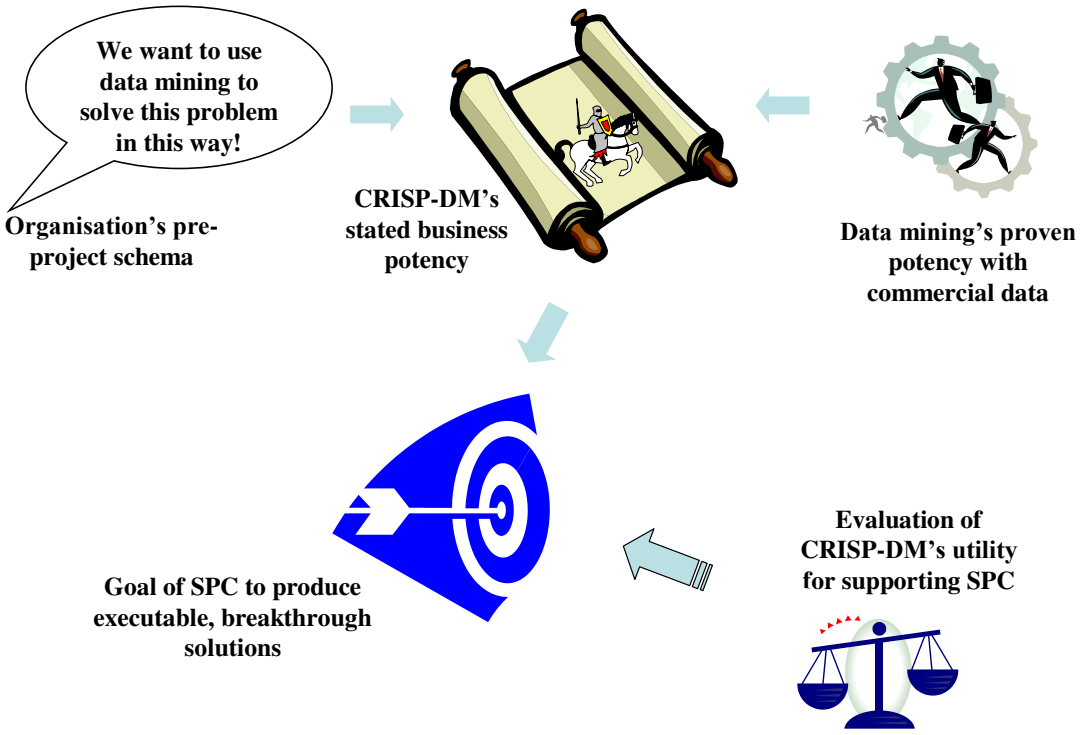


Figure 1.1: Research context

We describe the contextual elements:

- in the call-out symbol – the organisation's *preconceptions* about a business problem or opportunity, and its *expectations* about solving the problem or meeting the opportunity, using data mining technology and a good data mining project methodology. We further refine this contextual element in Chapter two into a *pre-project schema* of expectations;
- in the cog-and-man symbol – data mining's proven potency with commercial data;
- the target symbol – the goal of SPC of producing breakthrough solutions considering limiting organizational circumstances, and effectively and efficiently executing those solutions in a changing environment;
- in the scroll-with-knight - CRISP-DM's claimed business utility.

1.2 Research problem, goal, objectives, scope, and significance

1.2.1 Research problem defined

An evaluation of the utility of CRISP-DM for formulating and executing an effective data mining project in the SPC environment.

The researcher formulates six evaluation criteria against which CRISP-DM is evaluated for utility. These criteria are:

- diagnostic technique for defining the project's business goals;
- the introduction of new business domain knowledge about the problem / opportunity, and its solution;
- bridging technique between the project's business goals and the data mining plan;
- corporate knowledge management activities associated with SPC;
- monitor and control; and
- a number of soft issues.

The analysis finds that CRISP-DM presents limited utility compared to all these criteria. The strategic implications of this limitation for business, are the loss of long-term differentiation in their chosen market (Porter 2002, p.36), and a reduction in profitability (Rud 2001). The strategic implication for data mining is difficulty in acceptance as an SPC support tool. The limitation of data mining project methodology therefore is a research problem worth solving.

1.2.2 Research goal and objectives

The main research *goal* of this thesis is to develop a new data mining project methodology, which has true potency in the SPC environment. This methodology would have a business '…process and system-centric view of KDD…effective in solving real-world problems.' (Kargupta, Joshi et al. 2005, p.ix). The researcher formulated the following *research objectives* for supporting the research goal:

- evaluate the utility of CRISP-DM against the six criteria of utility; and

- develop a new data mining project methodology, which overcomes the SPC limitations of CRISP-DM.

We strive to understand utility for SPC better theoretically, and to use that insight to improve the effectiveness of a technology for SPC (McBurney and White 2004, p.8). In this understanding and improvement, we depend on insights we developed from a number of research domains. We call this new methodology *Strategic Analytics Method*, or SAM for short. The last research objective is to:

- establish the tractability of SAM in the real world, by demonstrating its potency on actual commercially generated data, in an organization where there were limiting preconceptions and circumstances about the data mining project.

We used real-world data about a retention management problem from an Australian mobile telecommunication company. For confidentiality reasons we call that organisation *Telco ABC*.

1.2.3 Research scope

The scope of this research relates to three domains. They are Business, Databases and Information Management, and Knowledge Discovery and Data Mining. The research within these domains included:

- from Business - Psychology, Corporate Strategic Management, Marketing Management, Telecommunications Analytics, Risk Management, Total Quality Management (TQM), Enterprise Resource Planning Methodology (ERP), Project Management, Operations Management, Data Management, Information Management, Knowledge Management, and CRM industry whitepapers;
- from Knowledge Discovery and Data Mining - Statistical Learning, Inference, Data Mining, Concept Drift, Data Mining Project Methodology, Knowledge Discovery in Data Databases (KDD), and data mining software vendors' online help documentation and course material.

The scope of our field research is the application of SAM to Telco ABC's mobile retention management problem. We use SAM to discover information about Telco ABC's retention problem and a hypothesised business solution. We then cast that discovered information into a classical marketing Segmentation-Targeting-Positioning (STP) framework, and develop executable knowledge from that information.

The executable knowledge takes the form of potential retention management campaign offers. The campaign offers are segmented and prioritised by a demand function, a psychographic measure, a measure of commercial value, a risk management feature, and a feature, which represents the main churn cause. The scope of the knowledge therefore extends past simply predicting who the potential churners may be, to also developing executable knowledge about behavioral and value segments. This establishes a forward linkage (Zikmund 2003, p.59) into the data mining project, of segmenting and profiling which traditionally would have been achieved with non-data mining methods.

1.3 Research methods

In this section, we describe the methods we used in this research, and their outcomes and benefits for data mining.

1.3.1 Description of methods

The *setting of our research* is industry-based business research, where the *goal of research* is improving the decision-making, through developing a definite course of action with an identified problem (Zikmund 2003, pp.5, 8, 61, 73).

Given the research setting and goal, we chose the generic research strategy of *Participatory Action Research* (Denzin and Lincoln 2003), specifically that of *action science.* Action science combines the main characteristics of action research, which are a study of the situation, and an evaluation of what needs to be achieved about the situation. It subsequently strives to remove the barriers toward achievement – with a rigorous professional interpretation and enactment of the situation, exposing the gap between a hypothesis (or theory) and a problem (Denzin and Lincoln 2003, pp.340, 342) as a process of the *design* of a preferred solution. The approach has also been termed *expert reflection-in-action* (Schön 1995, p.46).

A characteristic of our research setting is that we do not have control over the situation of the study (McBurney and White 2004, p.214). Our approach is similar to that of the practitioner's reflection-in-action, where practitioners reflect on their intuitive and tacit knowledge in the midst of action, using this capacity to cope with unique, uncertain, and conflicted situations in practice (Schön 1995, pp.9, 16).

Philosophically, we present it as an iterative, Hegelian conversational triangle between the expert's *hypothesised* solution for a problem, the reality dimensions of the problem

that impact on the desirability of the solution (*the antithesis*), and the gradual reframing by the expert of hypothesis and antithesis, resulting in the best *preferred synthesised solution* under the circumstances (Schön 1995, p.45) (Denzin and Lincoln 2003, pp.347ff). Paraphrasing its proponent's words, it is a conversational spiral, which strives to make the situation conform, through reforming the situation and the hypothesis about the problem (Schön 1995, p.151). Essentially the conversation consists of:

- considering the nature of a complex problem situation (peculiarities, complexity, and uncertainty, and the objective that needs to be achieved by the solution (Schön 1995, pp.129ff);
- conceptually developing a perceived solution for the problem based on an expert hypothesis about the problem;
- evaluating the workability and the implications of the proposed solution against an affirmation schema, which is based on the desirability of the perceived consequences on the problem situation AND attainment of the objective, and the intention of those consequences;
- the evaluation takes the form of *move testing* (Schön 1995, p.146), a process of investigating reality in order to transform it, transforming reality in order to investigate it (Denzin and Lincoln 2003, p.377), or changing the situation to understand it, and understanding it through changing it (Schön 1995, p.132);
- where the perceived consequences are undesirable, reframing the problem or the hypothesis conceptually through iterating through other potential solutions (Schön 1995, pp.55, 63, 79, 94);
- until affirmation is attained against the desirability of the consequences within, the situation and the attainment of the objective (Schön 1995, p.153).

The research draws on the candidate's corporate management experience in marketing and operations strategy formulation and execution, business expertise from an MBA degree, proven concepts from the previously identified research and business domains, and proven or accepted concepts in the data mining literature, to formulate propositions around the six evaluation criteria and data mining methodology. The proposed Strategic Analytics Methodology is a result of the researcher's experience amalgamated with an

inductive and inferential reflection-in-action process *in a practical industry setting*. (Schön 1995, p.130; Denzin and Lincoln 2003, p.377; Zikmund 2003, pp.43, 47).

We then establish validity for the construct that SAM is a *useful* data mining project methodology in the commercial data environment. We establish *internal validity* for this construct by drawing on established concepts in their various fields during the expert reflection-in-action argumentation (Schön 1995, p.24) (Gorman and Clayton 2003, pp.57, 60).

External validity is attained from successfully applying the new methodology in a data mining project in a real-life SPC setting (McBurney and White 2004, pp.169, 172) (Schön 1995, p.141) (Zikmund 2003, p.273). There we demonstrate how using SAM results in shifting limiting preconceptions about what was achievable from a data mining project. We discover information which results in a new understanding of a retention management problem. We also design a solution for retention management which contains objectively optimal and novel marketing knowledge, and which is executable in that practical setting. We support the external validity, by calculating the potential dollars that can be saved by the organisation in their retention management following this methodology (McBurney and White 2004, p.169, 172).

The external validity is further supported *generalising* the business solution (Gorman and Clayton 2003, p.61). The retention management solution we develop is transferable into any retention management environment, irrespective of the industry.

Internal and external validity is further strengthened by the fact that the researcher has had a successful corporate career in planning and executing business operations solutions (Schön 1995, p.138) (Hammersby 2004, p.138) (Gorman and Clayton 2003, p.63).

Linking external validity to credibility of the researcher's interpretations with the relevant audience, the external validity will ultimately be established, by consensus within the data mining and SPC communities, about the usefulness of our contribution, or not (Schön 1995, pp.141, 151; Gorman and Clayton 2003, p.63).

Our *experimental rigor* is established by adhering to the conversation until affirmation has been reached i.e. the solution is adequate for reaching the objective (Schön 1995, pp.136, 141, 151) with no undesirable side-effects.

The *experimental medium* of reflection-in-action is cognition (Schön 1995, p. 157). The experimentation is deemed successful when we can demonstrate the success of the cognitively formed evolution *through action* (Denzin and Lincoln 2003, p.381). Our *experimental control* therefore resides in a comparison of the actionability – or executibility – of the results achieved by CRISP-DM, with those achieved SAM (McBurney and White 2004, p.337).

Irrespective of its application environment, there are constants in reflection-in-action (Schön 1995, pp.270ff.):

- the professional language and experiential repertoire of the expert which allows the expression of the professional's evaluative perceptions of the situation;
- the cognitive experimental medium, which allows the researcher to manipulate or suspend some of the factors, that would be either too costly or risky to test in the complex real world of the problem;
- appreciative systems for problem setting, evaluation of inquiry, and reflective conversation;
- overarching theories for making sense of phenomena, interconnecting the reflective episodes;
- the framing of the in-situ researcher within an institutional role and tasks.

The *role of the researcher* in reflection-in-action experimentation is transactional; shaping the situation, while at the same time being shaped by it (Schön 1995, p.150). This is in contrast to classical, positivist experimentation, where the researcher's stance toward the problem is that of a disengaged observer. The hypothesis in this action research domain is therefore concerned with transforming the situation, more than it is with understanding the situation (Bryman and Bell 2003, pp.314ff).

We draw on the ethnographic research vocabulary to describe situation and the role of the researcher. *Access* to the setting was by introduction of a leading data mining software vendor, to an organisation, which wanted to experiment with a modeling approach to their retention management. The researcher spent 9 months on site at the industrial partner in an *overt research capacity*, which means it was known that he was a researcher. The *key informants* were two Customer Relationship Management analysts as well as the retention manager (the owner of the business problem). One of the

analysts had the ethnographic role of *gatekeeper*, controlling the flow of information to and from the researcher.

During these 9 months on site, the *researcher's role* was participant-as-observer; that is as researcher being part of the productive processes of the organisation. This period is broken down into:

- an initial 7 months phase, where he was part of a team constructing a mining data mart;
- a subsequent 2 month period, during which he built a predictive model for the churn event.

Concurrent with the last 3 months, the *researcher's role* was observer-as-participant – i.e. no participation in the process of production but still interactive with the team – regarding retention management knowledge and practice issues. This last role allowed the researcher to formulate ideas about practical issues and knowledge required, for a breakthrough solution in their retention management.

Detailed notes were recorded at all stages, and also reports about data mining experimentation that were generated using SAS Enterprise Miner's Reporter Node, which are consulted in the reflective process. The researcher terminated his field-based research when he considered that the data was sufficient for basing his research on.

We visually present the operationalising of our research goals in Figure 1.2:

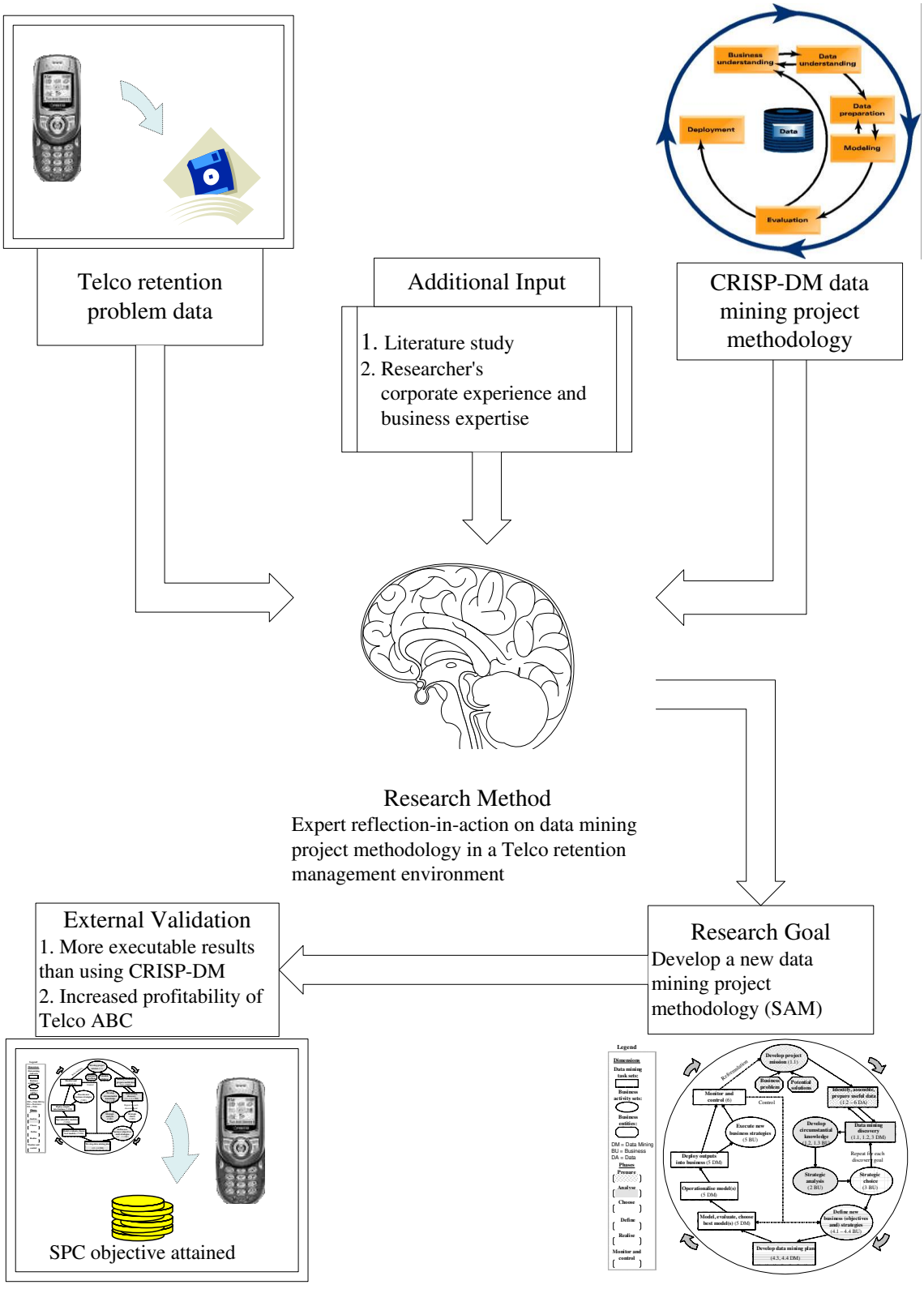


Figure 1.2: Operationalising the research goals

1.3.2 Research outcomes and benefits

1.3.2.1 For commercial practitioners

- An improved understanding of the practical and theoretical issues which affect data mining project methodology in the SPC environment;
- an improved data mining project methodology which assures the production of executable SPC project results.

1.3.2.2 For software vendors and consultants

The availability of a useful project roadmap, with its supporting vocabulary, with which to promote the strategic value of data mining in the SPC environment.

1.3.2.3 For the industry partner

A business solution for solving their retention problem using a data mining approach and new marketing subject matter. The estimated commercial benefit of this approach is in excess of AUD $10m per annum.

1.4 Thesis structure

We have divided the thesis into five parts. Each part encapsulates the relevant thesis chapter(s). The arrows between the rectangles show the sequential logic in the thesis structure.

Chapter 1 sets the context of the research problem, defines the research problem, and explains the research method followed. The literature study part has two chapters. Chapter 2 introduces supporting concepts that form the basis for our reflection on the state of data mining methodology. Chapter 2 ends with a brief overview of CRISP-DM as a data mining project method. Chapter 3 presents an in-depth literature study on how the automated data mining environment defines a data mining target concept, and how and what should be monitored about such a target concept. We argue that - based on similarities between the two application domains – that both the principles and practice of *concept drift monitoring* are transferable into the SPC environment.

Chapters 4 and 5 form the reflection-in-action. In Chapter 4, we reflect on the utility of CRISP-DM, for producing a solution for Telco ABC's retention problem. We conclude there that CRISP-DM is deficient in utility, and cannot produce a solution for Telco ABC. In Chapter 5, we develop SAM, which is a data mining project methodology,

which overcomes the deficiencies of CRISP-DM. This includes a transfer of the principles and practice of concept drift into SAM, and therefore into the SPC environment.

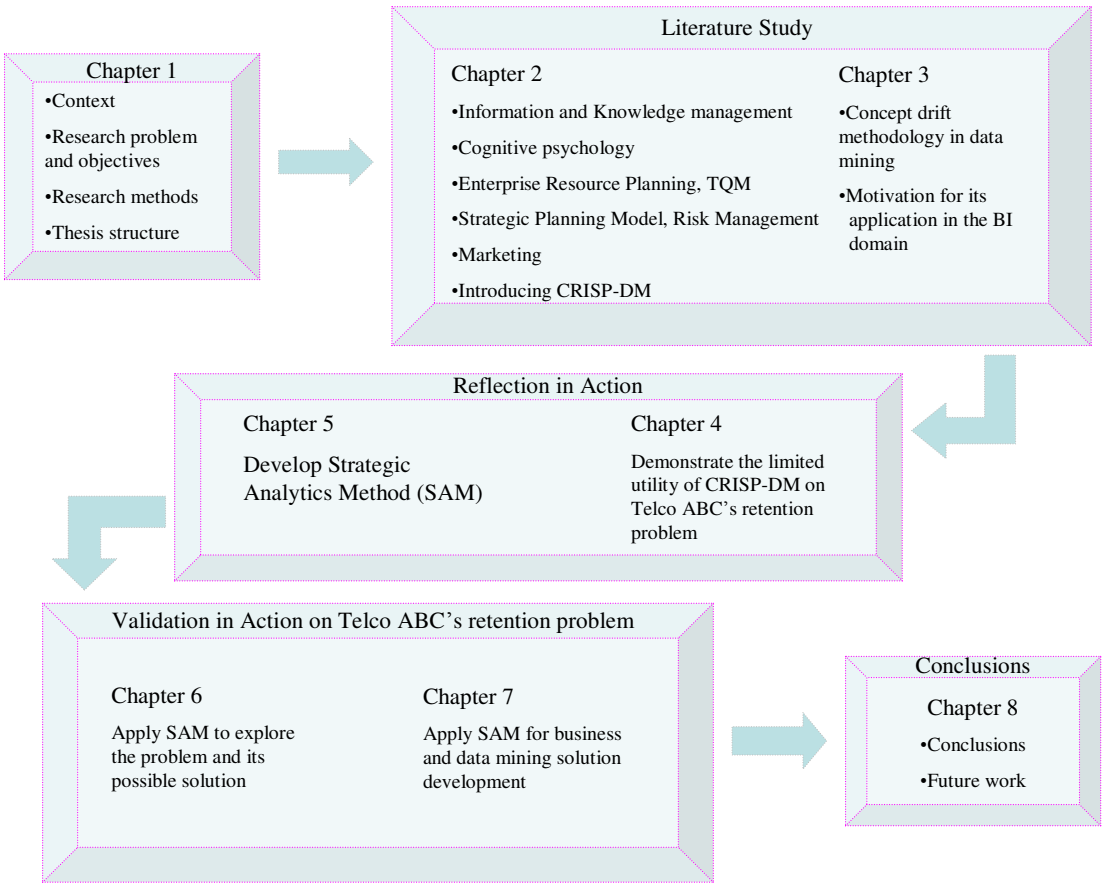


Figure 1.3: Thesis structure

The next part of the thesis structure is the validation in action of SAM. In Chapters 6 and 7, we establish the external validity of SAM in a commercial setting, by applying it to Telco ABC's retention problem. We establish the validity by demonstrating the tractability of SAM on real commercial data, by discovering relevant information about Telco ABC's retention problem and a possible solution, and by developing executable business and data mining solutions. This includes an application of concept drift principles for monitoring changes in discovered information and developed knowledge over time. In Chapter 8, we present the conclusions about this research, and identify opportunities for further investigation.

1.5 Chapter summary

In this chapter, we described the context of the research problem as the slow uptake of data mining as analytics methodology in the corporate Strategic Planning Cycle environment. In this environment, the competitiveness and executibility of results are of paramount importance. We added that organisational breakthrough - or paradigm shift – was a major component of competitiveness. Data mining methodologies have been developed for this application of data mining, but the researcher hypothesised that they lacked utility in assuring competitiveness and executibility of results, and that this lack was a contributing factor to the slow uptake.

Based on this the researcher formulated the research problem about evaluating the utility of the most commonly used methodology – CRISP-DM – for formulating and executing an effective data mining project in the SPC environment. We then defined three research goals. First was evaluating CRISP-DM against six criteria of utility in the SPC environment. Second was to develop a data mining methodology, which overcomes the limitations of existing methodologies. Third was to authenticate that new methodology on a challenging, real-life Telco retention problem.

The scope of the research was identified as Business, Databases and Information Management, Knowledge Discovery and Management, and Data Mining. We also identified a number of relevant sub-domains within this scope. The choice of domains was to:

- allow the formulation of a representative list of evaluation criteria for CRISP-DM in the business environment;
- provide the *materiel* for the evaluation of CRISP-DM against the criteria; and to
- provide the material for the design of a data mining project methodology with full utility in the Strategic Planning Cycle environment.

We described our research method and explained the choice of it. The research approach is *Participatory Action Research*, using the method of *action science*. Within this method we used an experimental technique called *expert-reflection-in-action*. This approach was chosen because of the limited control the researcher had over the research environment. We described the research outcomes and identified benefits for

commercial data mining practitioners, data mining software vendors, for consultants, and for the Telco industry partner.

We ended the chapter by describing the thesis as having five parts. Chapter 1 contains the context, research problem and methods. A literature study makes up the second part, and contains Chapters 2 and 3. The third part is the reflection-in-action, consisting of Chapters 4 and 5. In Chapter 4 we evaluate CRISP-DM against six criteria of utility. In Chapter 5, we reframe data mining project methodology and contribute a data mining project methodology, which improves on the identified deficiencies, and adds utility in further important areas.

The third part of the thesis is the validation of the new methodology in Chapters 6 and 7, using the Telco ABC business problem and the data reflecting it. The last part consists of Chapter 8, where we present our research conclusions, and identify opportunities for future research levering off our work.

In the following two chapters we present the literature study. The purpose of the literature study is to support our choice of evaluation criteria of CRISP-DM. They will also provide the material we use when developing SAM.

2 CHAPTER 2 – Contextual literature research

In this section, we first introduce knowledge from the previously mentioned research and business domains. The purpose with this knowledge is supporting our defining of the evaluation criteria of CRISP-DM, and to provide materiel we draw on when designing SAM.

The chapter also contains a section where we introduce our industrial Telco partner's existing paradigm about their business problem and its ineffective solution, and about the benefits they are expecting from applying data mining to their problem and its solution. We need detail about this paradigm when we reflect-in-action on CRISP-DM's ability to facilitate paradigm shift for our Telco partner.

We further introduce new marketing subject matter expertise. We need this knowledge for a number reasons. First, we need subject matter expertise for our reflection-in-action about the utility of CRISP-DM in producing breakthrough in a marketing SPC environment. The second is for later demonstrating SAM's capability in introducing new subject matter expertise into the SPC environment. The third is to produce a breakthrough solution for our industrial partner's nagging retention management problem.

Last, we offer a brief introduction of the CRISP-DM data mining project methodology. It serves as a brief orientation of our object of evaluation. The reader not familiar with CRISP-DM, should consult the source document for a more in-depth understanding of CRISP-DM.

2.1 Information and Knowledge Management

From the Data Management, Information Management, Knowledge Management, and business literature we introduce the following practical concepts:

- *Data* are the non-mentally recorded measurements about events, their connections and relationships (1999, p.2). The *origin* of data is measurement and recording. An example of a tool for the recording of data, is a database. An example of a measuring device, is a temperature sensor in a production plant;
- *Information* is a signal that reduces uncertainty about a state or event (Lucas 2000, p.26). Signals are measurable, and either present or absent (Pyle 1999, pp.406ff.).

The presence or absence and measurability of this informational signal, means its origin is *discovery.* Data mining is a tool for discovering patterns of information (Ganti, Gehrke et al. 1999) (SAS Institute online b, p.3). Despite the common use of the term *Knowledge Discovery in Databases* (KDD) in the data mining literature e.g. (Han and Kamber 2001, pp.5ff.), there is sufficient evidence in the data mining literature supporting the view that what data mining discovers, resides at the level of information (Hastie, Tibshirani et al. 2001) (Pyle 1999, p.23 and Chapter 11). At least one major data mining project methodology recognises this too (SAS Institute 2000). A leading industry expert has stated that *data mining algorithms are 'knowledge free' ... meaning in real applications (they) lack even the very basic 'common sense reasoning' needed to recover even from simple situations* (Fayyad 2004, p.3). Information on its own does not have executibility, and without executibility, it does not have business value (Meltzer 2000, p.1); therefore:

- *Knowledge* adds to the informational signal the insights and understanding about the causes and implications of the information (Kogut and Zander 1992) (Whitten, Bentley et al. 2004, p.27). Knowledge is the executable *know-how* of the information – the *…which actions produce which results, and how and when to take them…* to produce results (Zikmund 2003, p.21) (Pyle 1999, p.2). Knowledge is *created* by a process of human cognition (Gibson, Ivanchevich et al. 1991; Schön 1995; Lucas 2000, pp.26ff.). In business, the development of knowledge takes place formally in knowledge developing business activities, which we will describe later. Developing knowledge from relevant information, constitutes the executibility of that information (Grant 1996, p.375). In business, executable knowledge constitutes a solution for a problem or an opportunity. In business, value is derived from *relevance* and *executibility;*

- *Intelligence* refers to the novelty and exclusivity dimensions of discovered information (Lucas 2000, p.35); it is a nuance of information, which depends on *what we think others know* or do not know about information, which is known to us.

Knowledge management has *three phases* (Takeuchi 1998). They are creation, knowledge dissemination within the organisation, and reflection of the knowledge in the organisation's products, services, and systems. The knowledge management model of Ferran-Urdaneta (1999) also has three stages:

- knowledge creation;
- knowledge legitimisation;
- knowledge sharing.

The updating of the corporate strategy is a main tactic for knowledge sharing and legitimisation within the organisation (Sveiby 2001), and more than ten years before that date, complete books were published in the Information Management domain, on the subject of the relationship between information, knowledge, and corporate strategy e.g. (Ward, Griffiths et al. 1990).

Nonaka (1994) identified two *types of knowledge*:

- *Explicit* – that which relates to facts and understanding which tend to be widely recognised by domain practitioners (Gibson, Ivanchevich et al. 1991, p.572). This is also known as *book* knowledge;
- *Tacit* - also called *heuristic* (Gibson, Ivanchevich et al. 1991, p.572). This experiential or judgemental knowledge is related to implementing explicit knowledge, and is therefore also known as domain *mastery*.

The intention to create and use knowledge is an *enabling condition* for information discovery and knowledge development (Nonaka and Takeuchi 1995, p.74), and progressive companies will reflect this intention in their mission statement in their corporate strategy (Von Krogh, Ichijo et al. 2000) and (Slembek 2003, p.32). *Subjective requirements* for developing knowledge are cognitive talents and skills, and some idea of what is relevant - or not - based on existing and new domain knowledge. The role of new knowledge as a *catalyst* for achieving paradigm shift has been recognised for at least a decade in the literature e.g. Wikstrom and Normann (1994).

The most effective way for developing organisational knowledge is in *cross-functional teams* (Nonaka 1994; Ferran-Urdaneta 1999; Nunamaker 1999) and collaboration is central to the effectiveness of the knowledge development process (Slembek 2003, p.21). An effective tactic for increasing both an organisation's technical and domain knowledge, is *importing* it from outside the organisation (Leonard-Barton 1995) (Slembek 2003, p.34).

2.2 Cognitive Psychology Theory

According to the Information and Knowledge management literature, data mining does not discover knowledge, but information, and knowledge is developed from information through human cognition. In the SPC environment data mining therefore relies on a human engineering and a business domain expert for its *configuration*, the *interpretation* of its outputs, and for *decision making* about the use of those outputs. We draw on insights from Cognitive Psychology about the influence of human psychology and thinking processes, on the effectiveness of the information discovery and knowledge development.

Two sequential psycho-cognitive processes are involved in information discovery and knowledge development. We present these two processes in Figure 2.1: Perception and cognition. The first process is the *Perception process*, and it consists of two stages; an *Awareness response* and *Perceptive meaning appraisal*. The second process is the *Cognition process*. The *Cognition process* has one stage called *Cognition*.

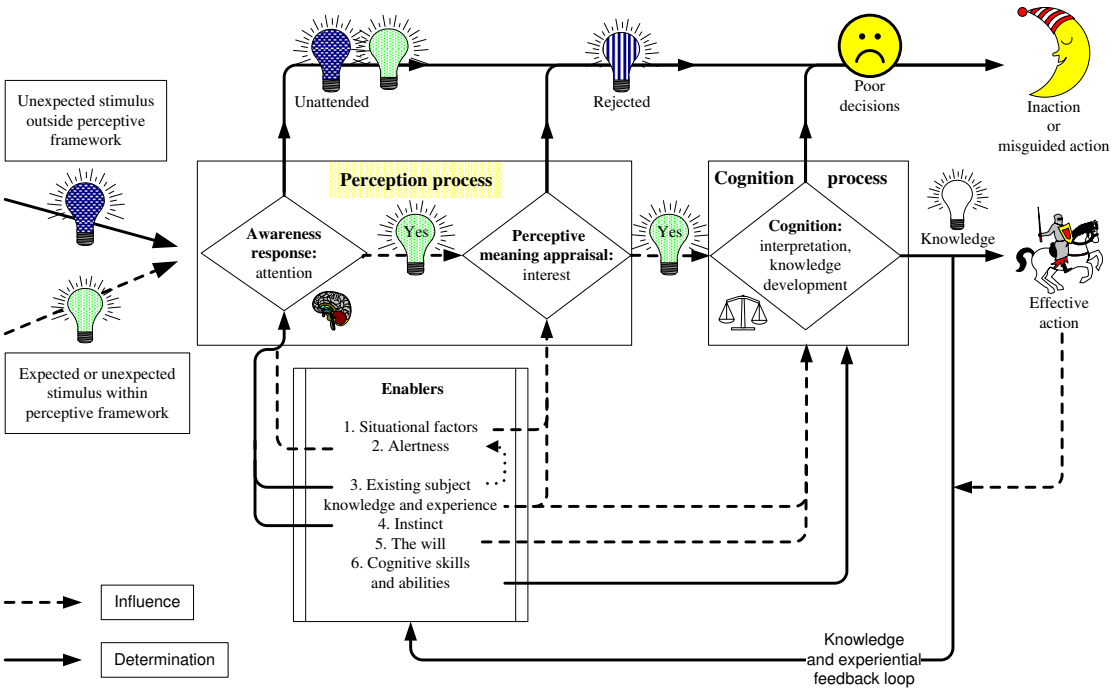


Figure 2.1: Perception and cognition

2.2.1 Perception process

In this section we present relevant insights about the two stages of the *Perception process*. The *Perception process* is indicated by the dot-filled rectangle in Figure 2.1.

2.2.1.1 Awareness response stage

We represent the *Awareness response* stage with a diamond shape, labelled accordingly in Figure 2.1. In *Awareness response* our minds process incoming stimuli for attention. Incoming stimuli reach the brain through the senses. In Figure 2.1 we present incoming stimuli by the two light icons and their arrows leading into the *Perception process.* Incoming stimuli fall into two categories. The first category is those stimuli which fall outside our perceptive framework, and are therefore unexpected. The second category is stimuli which fall inside our perceptive framework, and are either expected or unexpected.

In the *Awareness response* stage the brain gives an orienting response of *attention*. The response may be positive or negative, as represented by the icons labelled *Yes*, or *Unattended*, leading from the *Awareness response* stage (Goleman 1998, pp.42, 44, 48). This awareness response is *triggered* by a hormonal reaction in the lower regions of the human brain, specifically in the hippocampus (Goleman 1998, pp.42, 47). Of particular interest, is that this low-level reaction *determines* the level of response associated with all stimuli – the so-called *cognitive static* - which we experience as *attention* (Goleman 1998, p.41). In Figure 2.1 we indicate *determination* with solid lines.

Stimuli which fall outside our perceptive framework, are always *determined* to be unattended. They therefore always enter *Awareness response* determinedly, and exit it determinedly through the path labelled *Unattended*. Exiting this way results in Inaction. Of particular importance is that we cannot experience *inattention* of a stimulus of any type.

Stimuli which fall within our perceptive framework, may be attentively perceived or not – irrespective of their unexpectedness. They therefore enter *Awareness response* as an influence. Dependent on whether they get our attention or not, they exit through either the *Yes* or *Unattended* path.

We have no control over *Awareness response's* determination of attention, even when a stimulus falls within our perceptive framework. However, our attention is *enabled* by the listed factors in Figure 2.1:

- *alertness* influences the *Awareness response* by sensitising us for incoming stimuli within our perceptive framework. The higher our alertness, the more

sensitive we are for giving attention to arriving stimuli within our perceptive framework- whether we are expecting them or not. Consider however, that alertness does not enable attention to stimuli which fall outside our perceptive framework;

- *existing subject knowledge and experience* determine the *Awareness response*, through setting the *perceptive bounds* (Lucas 2000, p.29). Low levels of *subject matter knowledge or experience* limit our perceptive bounds, increasing the incidence of stimuli that fall outside our perceptive framework. Our mind does not recognise such stimuli, resulting in them being *Unattended*. Inattention is an undisturbed state of mind - and not an experience – which means we *remain unaware that we have missed* an important stimulus. Such unawareness is a disturbing part of our psychological make-up. In the professional setting, it is a major cause of *paradigm lock*, where an organisation is unable to recognise problems or opportunities as they arise. In the data mining environment, they limit the recognition of potentially relevant information.
- increasing the *subject knowledge and experience* widens the perceptive boundaries. This increases the chance of stimuli getting an attentive *Awareness response*. In the data mining environment, higher levels of subject *knowledge and experience* improve the chances of what the data mining literature refers to, as *the discovery of previously unknown knowledge, patterns etc.*

  Further, *existing subject knowledge and experience* influence the *Awareness response*, through increasing our state of alertness;

- *instinct* determines the *Awareness response*, and the effect of *instinct* is therefore beyond our control. The reason is that the mechanism of the response is hormonal, seated in the limbic brain. Attention is subconscious, and no rational meaning proper is attached to the stimulus (Gibson, Ivanchevich et al. 1991, pp.64ff.).
- *awareness response* is subjective and varies between individuals. It also varies for the same individual over time because of the influence of our moods. A stimulus which has elicited attention, is passed through the *Yes* path to the *Perceptive meaning appraisal* stage of the *Perception process*.

2.2.1.2 Perceptive meaning appraisal stage

In Figure 2.1 we represent *Perceptive meaning appraisal* as the second diamond shape within the *Perceptive process.* In the *Perceptive meaning appraisal* stage*,* attentive stimuli are more consciously appraised for meaning to the situation at hand (Gibson, Ivanchevich et al. 1991, pp.64ff.). Depending on the meaning we attach to the stimulus, we experience this appraisal as *interest* or *no interest*. If a stimulus is uninteresting, it is rejected, resulting in *Inaction*. Because *Perceptive meaning appraisal* resides in the conscious mind, we experience the rejection of a stimulus too. Interesting stimuli are passed down the *Yes* path to the *Cognitive process*.

The *purpose* of *Perceptive meaning appraisal* is psychological. It is to cast stimuli into an acceptable experiential and psychological framework. As beings we favour as interesting those stimuli that best fit the 'existing picture', leaving us *psycho-socially comfortable.* This means that *Perceptive meaning appraisal* is selectively *biased* toward optimising our existing experiential and psychological framework (Gibson, Ivanchevich et al. 1991, p.69). This bias is uncontrollable and situational, which means that it also affects our professional life. One result of this in the professional environment, is as *paradigm lock*.

In Figure 2.1 we listed the enablers of *Perceptive meaning appraisal*. We now discuss them in more detail:

1. *situational factors* influence the meaning appraisal restrictively or progressively. In the case of the SPC application of data mining, influential *situational factors* may be:

    1.1. the professional *attitudes* of the data miner, the business subject matter expert (SME), and of the organisational leaders toward the problem at hand;

    1.2. the *perceptions* of the SME and organisational leaders, improvements which may result from organisational change; and

    1.3. the *existing* formal paradigm (strategy, structure, systems, resource lock-in etc.), and the informal paradigm (culture, politics etc.) on the implementation of change, based on discovered information;

3. *existing subject knowledge and experience* influences the appraisal of a new stimulus for interest. Low levels of *existing subject knowledge and experience* framework reduce the chance of a new stimulus being rejected.

   Perceptive meaning appraisal resides in the conscious mind, driving our tendency toward psychological closure. If we are uncertain in our appraisal, psychological closure makes us search for more subject knowledge in order to attain closure.

   The *influence* of *existing* and *new subject knowledge and experience* when using data mining in the SPC environment now becomes apparent. Information which is discovered by the data mining, is a stimulus of the *Perception process*. If a discovery leaves us feeling uncomfortable – because it does not fit into our existing framework – we may perceive it as uninteresting *despite its relevance or importance to the professional task at hand.* That means that the discovery process can be terminated even before we start developing new knowledge in the *Cognition process*. Increasing the levels of *subject knowledge and experience*, enables the data miner and SME to better recognise the relevance of discoveries, and to push through to cognitive knowledge development.

   A further influence of increased *subject knowledge and experience*, is that it enables the organisation to break through situational restrictions, overcoming paradigm lock.

2.2.2 Cognition process

Cognition is the next stage of our mental involvement in the data mining process. In this section we present relevant insights about cognition. The *Cognition process* is indicated by the stripe-filled rectangle in Figure 2.1. Within there we represent the *Cognition* with a diamond shape labelled accordingly.

In this stage we interpret information, and develop knowledge from it. Here we *consciously manipulate information* through a *process* of interpreting the information for relevance, evaluating alternative explanations and solutions, and then selecting those explanations and solutions that will produce the most suitable results under the circumstances (Gibson, Ivanchevich et al. 1991; Lucas 2000).

Poor knowledge development results in *Poor decisions*, leading to either *Inaction or misguided action*. Good knowledge development results in executable *Knowledge* and *Effective action*.

We recall from McBurney et al. (2004, p.381) that cognition is the experimental medium of expert reflection-in-action. The *influencing enablers* of cognition are:

3. *existing subject knowledge and experience*, providing:

   3.1. the professional language and experiential repertoire which allows our evaluative perceptions of the situation;

   3.2. the appreciative systems for problem setting, evaluation of inquiry, and reflective conversation about problem solving;

   3.3. and overarching theories for making sense of phenomena, interconnecting the reflective episodes;

5. the human *will* to meet the goal (driven by aspiration), and to change direction if that is required for meeting the goal.

   Poor decisions are caused by a lack of *subject knowledge* or *experience*, or a lack of *cognitive skills or abilities*, or a lack of *will* power.

6. The *determining enablers* of cognition are our *cognitive skills and abilities*, which include logic and creativity.

2.3 Strategic Planning Model

The Strategic Planning Model (SPM) is an iterative planning process. It is widely used in open and unstructured problem environments to formulate achievable outcomes. In that environment, there are constraints on achieving outcomes, and the constraints take a number of forms e.g. limited resources, or external or internal issues which are out of control and negatively affect the pursuit of goals. This characteristic makes it very suitable for application in the corporate environment, because organisations function in open, unstructured, uncontrollable environments, have limited resources at their disposal. The SPM facilitates the selection of the best solution *under the circumstances*, for the pursuit of a goal. The facilitation by the model to proceed with the best solution *despite the limiting circumstances,* makes it a *Strategic* Planning Model.

An important area of organisational application of the SPM, is corporate strategy formulation (Pearce and Robinson 1991; Pearce and Robinson 2004; Pyle 2004). Here it is used to clarify and focus the organisation's thinking and planning about their business in their chosen market, and about how they conduct that business operationally. These

results are expressed in the objectives and strategies hierarchies. In this strategic corporate application, SPM is applied between every one to five years, depending on how fast the organisation's leaders perceive their market and operating environment to be changing. SPM is also popular with the military in its strategic application.

SPM is versatile enough to be used with varying levels of detail, and a frequent corporate *tactical* application is the Management Information Systems Development Lifecycle (Post-Anderson 2002, pp.458ff.) (Boyce and Blair 2003, Chapter 8). The more tactical the application of SPM, the shorter the planning horizon. SPM is therefore often used corporately in a project environment. Examples are Enterprise Resource Planning (ERP) and Customer Relationship Management (CRM) software vendors, who have embedded SPM in their software implementation projects. There, it is used to focus the technology on the business's goals, and also for implementing strategic and operational breakthrough, which was made possible from implementing the technology. The military also use SPM tactically during battle.

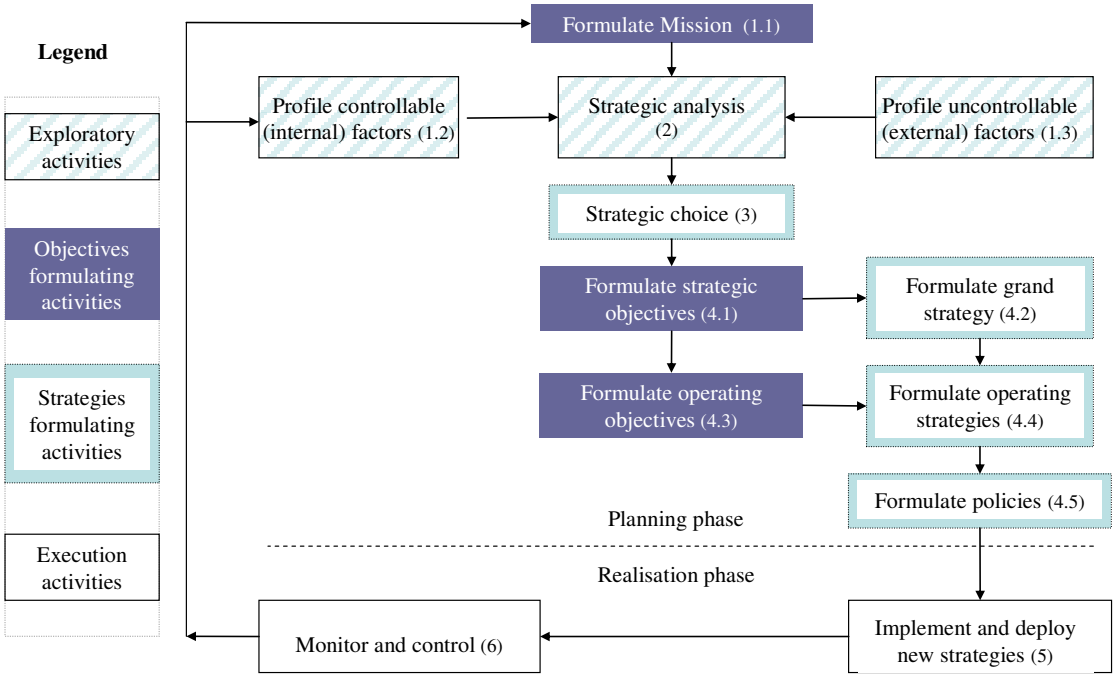
We introduce SPM, because we use it when evaluating the utility of CRISP-DM. Further, we draw on it when developing SAM. Irrespective of the strategic or tactical purpose of a planning cycle, business generically refers to any planning cycle which uses SPM, as a *Strategic Planning Cycle*, or SPC. In this thesis we follow suit, using *SPC* irrespective of any time horizon involved, or the strategic / tactical nature of a business goal.

For more than a decade now, the major proponent of SPM in the Corporate Strategy literature has been Pearce and Robinson (2004) and (1991). We draw on this source for insights about the use of SPM during SPC.

SPM is a logically and temporally sequenced framework of planning activities. We present the model Figure 2.2: Strategic Planning Model. Arrows indicate the sequential dependency between the activities. The three solid dark rectangles represent an *objectives formulating* activity hierarchy. The three striped rectangles represent *exploratory activities*. Generally controllability / uncontrollability is related to the organisationally internal or external location of a factor. The four framed rectangles represent *strategies formulating* activity hierarchy. The remaining two plain rectangles represent *execution activities*. A *Planning phase* precedes a *Realisation phase*. The

broken horisontal line separates the two phases. The numbering within the rectangles is referenced in the discussion below.

In this thesis, we refer to the party or organisation using SPM, as the *planning entity.*



(Adapted from Pearce and Robinson 2004)

Figure 2.2: Strategic Planning Model

2.3.1 Mission and circumstantial profile

1. There are three inputs into SPM:

   1.1. the *Mission* is the expression of the broad objective, which the *Planning entity* is in pursuit of using SPM. It expresses a *desired* destination for which there has not been a detailed consideration about the *possibility* of achieving it. That means the *restrictions* on achieving it have not yet been identified, and the *ways* of achieving the *Mission* have not yet been developed. The *Mission* could also express a complex problem, which needs to be solved, as we will see in a later example from the application of the SPM in a technology environment. The *Mission* is the first objective in the objectives hierarchy;

   Commercial example of a *Mission* is Land Rover wanting to profitably market a short-wheel base version of their Defender series in Australia;

Military example of a *Mission* is to capture a strategic port intact from a defending enemy;

1.2. *Profile controllable factors* is concerned with profiling the effects of all factors under the planning entity's control, on reaching the *Mission.* These factors include the resources at the disposal of the organisation. This activity profiles all the alternative solutions in terms of their resource requirements and the benefit they add, towards achieving the *Mission.* If the planning entity has commercial goals, the *resource requirements* of each solution is quantified as monetary *cost*, which reflect the quantity and quality of resources that will be utilised or consumed by that solution.

Likewise, the benefit of each solution toward achieving the *Mission*, is quantified in monetary terms, which reflects the *benefits* of each solution. The *benefit* is usually expressed in *revenue*. If the application of the SPM is at a conceptual level only, then the cost and benefits may be expressed qualitatively as weaknesses and strengths.

There remains the step of calculating the *desired outcome* from the controllable factors profile. This step entails deducting the *costs* from the *revenue* to arrive at a *profit*, or *desired outcome*. This calculation does not yet consider the impact of the *Uncontrollable factors* on attaining the *Mission*.

Commercial example (continued) – two scenarios for attaining the *Mission* are profiled by Land Rover. The first is assembling vehicles locally from imported components. The second is importing fully assembled vehicles from England. Land Rover calculates the potential *profit* from the local assembly or import scenarios respectively, as $5000 per vehicle for local assembly, and $10 000 per vehicle if fully imported. This profit is based on a *desired* selling price of $40 000. This profit determines the *desirable outcome* for Land Rover;

Military example (continued) - the *cost* to the attacking party is in ships, aircraft, and infantry. The cost they are willing to pay is quantified as 10 troops, a helicopter and a fighter bomber and their four pilots, but no ships. The *benefit* to the attacking party, is obtaining facilities through which to deploy reinforcements needed in another strategic front. However, the port facilities have to remain 95% intact during the attack, if it has to have strategic value to

them. They estimate that if they can surprise the defenders, damage to port facilities will be minimal, and the targeted cost in lives and equipment will be achieved. This determines the *desirable outcome* for the attacking party;

1.3. the *Uncontrollable factors* profile is concerned first with identifying those factors, which are *uncontrollable* by the planning entity. Second, the profiling identifies and quantifies the *contributory* or *detracting nature* of each *uncontrollable factor* on attaining the *Mission.* Once again, if the *planning entity* has commercial goals, the ideal is to quantify the *detraction* or *contribution* in monetary terms. In more conceptual applications, these effects may be expressed qualitatively as threats and opportunities. The output is the *Uncontrollable factors* profile;

Commercial example (continued) – Land Rover estimates that a major competitor, Mitsubishi Motors of Japan, will respond to the Land Rover operation, by importing fully assembled short-wheel base Pajeros. The estimated retail cost of the Pajeros will be $6000 dollars less than the desired-for retail price of the short-wheel base Defender;

Military example (continued) – the *Uncontrollable factors* profile is the attacking party's intelligence about land, air and sea resources under the defending party's control, their geographic deployment in and around of the port, and their state of alertness. Intelligence also has determined that the defending party has also rigged demolition explosives at key port installations, and will detonate these if they are attacked.

The *Uncontrollable* and *Controllable factors* profiles are often referred to as the *organisational circumstances*.

2.3.2 Strategic analysis

2. In this activity we develop the alternative *possibilities* for attaining the *Mission*. This is an overlay of the *desirable outcome* and the *Uncontrollable factors* profile.

Commercial example (continued) – in practice, Land Rover will have to reduce their selling price if Mitsubishi starts competing locally. The worst case scenario is that Land Rover will have to reduce their selling price by a full $6 000 to sell the targeted volumes. Under that scenario, the Mitsubishi response would be *detractive*

to the full value of $6 000. In this case Land Rover will still profit $4000 if selling fully imported vehicles (the $10 000 desired minus Mitsubishi's $6000 discount), or will make a $1000 loss per unit if manufactured locally (the $5000 less a $6000 discount). This sets the range of worst *possible outcomes* for Land Rover.

The best-case scenarios, are that the market will perceive the value-proposition of the Defender over the Pajero to be in the range of $3 000 to $6 000. In this case, the uncontrollable competition by Mitsubishi will be limited to a range of $3 000 to zero on Land Rover's *desired outcome*. The range for the *best possible outcome* then is a profit of between $4 000 and $6 000 per unit sold on fully imported vehicles.

Military example (continued) – Based on their estimates about surprising the enemy, the attacker now creates a range of worst to best scenarios, for the uncontrollable factors to realise themselves i.e. of enemy response. From these, they further estimate a range of *possible outcomes*. For instance the attacker estimates that if there is no surprise, the attack *at worst* would cost them attacker 100 troop lives, three helicopters and fighter bombers each, their twelve pilots, and a submarine with 50 crew. Additionally, the defending party will destroy the port facilities, rendering it useless. The attacker's intelligence suggests, however, that there is a good chance of isolating the demolition explosives in a clandestine operation. At the same time, the clandestine operation could silently sabotage the anti-aircraft capabilities of the defender. The *best possible outcome* calculated on total success of the clandestine operation, then is the loss of the 10 troops, no aircraft or pilots, but unfortunately the submarine and its crew will not be saved.

2.3.3 Strategic choice

3. In *Strategic choice* we first rank all the *possible* outcomes from worse to best. Next, we choose the best *possible* outcome. This *best possible* then has considered all the resources at the planning entity's disposal, their control over those factors they could control, and also the impact of all factors not under their control – in attaining the *Mission*. We call this output the entity *Strategic choice*.

   Commercial example (continued) – arranging the scenarios described from bad to worse, it becomes apparent that the only way in which Land Rover can be certain of a profit, is by importing fully assembled vehicles. This becomes their *Strategic choice* under the circumstances;

Military example (continued) considering all the scenarios described above, the attacker realises that they will not be able to surprise the defender with a full-force attack. Second, they realise that under the circumstances at best, they will incur a higher cost than the *desired*. In view of the importance of securing the port, their *Strategic choice* then becomes to first launch the clandestine operation, and only if that is reported in as successful, to proceed with the main attack.

The *Strategic choice* is at the top of the strategies hierarchy supporting the attainment of the *Mission*. The *Strategic choice* is a critical stage in the SPM; it completes the exploratory planning, and refocuses the *Planning phase* on planning for executibility. The executibility is required for the transition into the *Realisation phase*. Put in other words, *Strategic choice* constitutes the *bearing* for achieving *strategic paradigm shift*.

2.3.4 New Strategic and Operating objectives and strategies

4. The objectives and strategies hierarchies are now developed for the *Strategic choice.* The hierarchy is supportive, first because the framework supports the achievement of the *Mission*, and second because the strategies support the objectives. The objectives constitute the *what* the organisation needs to achieve in various key areas, and the strategies the *know-how* – or knowledge in Knowledge Management terminology - of achieving those *objectives*. Further, the framework is hierarchical, because of the increasing detail in the content of both the objectives and strategies, as it cascades from one level down to the next. The first level of the hierarchy contains the *Strategic objectives* and their supporting *Grand strategy*:

    4.1. *Strategic objectives* are the conceptual, long-term *what,* or objectives, which need to be attained in support of the *Strategic choice*. We now discontinue the previous commercial and military examples, on favour of presenting Telco ABC's objectives and strategies in a later section*;*

    4.2. the *Grand strategy* is the level in the plan hierarchy that supports the *Strategic objectives*, by expressing the *know-how* about how to attain them. This level represents the transition from a *Planning phase*, to a *Realisation phase* of the plans. Of importance for our application, is that any new *Strategic objectives* and *Grand strategy* represent new knowledge, and constitutes *strategic paradigm shift.*

The second level of the hierarchy is the *Operating objectives* and their supporting *Operating strategies*. This level of the plan hierarchy is where the *Strategic objectives* and the *Grand strategy* are converted into business function operating level details. They bring the planning entity one step closer to the *Realisation phase*;

4.3. *Operating objectives* present the *practical what* the organisation needs to achieve with the *Strategic objectives;*

4.4. *Operating strategies and tactics* support the *Operating objectives*, and contain the most detailed, critical and practical *know-how* about achieving the objectives hierarchy. Because of their detail, they contain measurable *Success criteria*, which form the basis for the later activity *Monitor and control*. Of importance in our application, is that any changes to the *Operating objectives* and the *Operating strategies*, define *operational paradigm shift*;

4.5. in situations where the execution in the *Realisation* stage is complex, *Policies* are the rules and guidelines that for the execution of the *Operating strategies* at the daily effort level. *Policies* represent the end of the strategies hierarchy.

2.3.5 Implement and deploy new strategies (Execute)

5. *Implementation and deployment* is the main execution activity of SPM. It supports the objective and strategy hierarchies through expressing the *with what* of the *know-how* expressed in the operating strategies. *Implementation* is the allocation and deployment of the resources, which the organisation requires to execute the strategies. Implementation and deployment constitute the *execution* of the solution (Pearce and Robinson 2004, Section III).

   The level of detail, which is required by the SPM from this activity, is dependent on the complexity of the strategies that were formulated above. In the commercial application, the requirements are complex, and *Implementation and deployment* is usually developed using McKinsey's 7-S framework. McKinsey's 7-S has been in use for decades (Peters and Waterman 1982, p.11). The 7-S framework expresses the interlinked application of seven implementation tools. The tools are:

   5.1. Strategy - the hierarchy of objectives and strategies;

5.2. Organisational Structure – the organisation of flows of information, authority, responsibility, and accountability;

5.3. Systems – all the systems and organisation uses for its operating and administrative purposes e.g. IT systems, procedures, ERP systems, plant, strategy etc;

5.4. Shared Values – which is the organisation's culture and formal and informal motivational structures and systems;

5.5. Style of leadership – proactive, reactive, dynamic etc.;

5.6. Management Staff Skills – the managing of the organisation;

5.7. Employee Task Skills – the skills of staff for doing their productive work.

These seven elements are combined and configured for implementation. The *…deployment* sub-activity is the *going live* of the *Implementation and deployment*, in other words, when the implementation becomes productive. The *Implementation and deployment* business activity set, can also be referred to as the *Execution* business activity set.

2.3.6 Monitor and control

6. The activities of *Monitor and control* assure the reliability of the *strategies* in supporting the *objectives* hierarchy, and therefore attaining the *Mission*. This assurance is achieved through measuring the performance of the *Operating strategies* against the measurement criteria contained in the *Operating objectives*. When the performance is not satisfactory, the *Strategic choice* may be changed to one of the other possible solutions, the objectives and strategies hierarchy modified accordingly, and the execution also modified accordingly. In extreme failure, the organisation may undergo another planning cycle using SPM, to identify new possible solutions, from which the *Strategic choice* can be made.

Because we will refer to them on our evaluation of CRISP-DM later, we present the relevant sections of Telco ABC's objectives and strategies hierarchy, *as we have cast them into an SPM format*. The example is incomplete since it does not fully demonstrate the activities of the *Strategic analysis and choice*, *Implementation and deployment*, or *Monitor and control*. For our purposes, we are more interested in the content of Telco

ABC's *objectives and strategies hierarchy*. The objectives and strategies constitute Telco ABC's current paradigm about their retention problem.

We set the convention for the remainder of this thesis, of referencing SPM according to the activity numbers in Figure 2.2. For instance *SPM 4.2* refers to the *Grand strategy*.

2.4 Strategic Planning Model as a Knowledge Management tool

In this section, we present our refinement of the three-phase Knowledge Management model of Ferran-Urdaneta (1999), in light of our work experience with SPM. Second, we interpret SPM as a knowledge development process. We motivate the refinement and interpretation, because we consider SPM as a practical tool for hypothesis testing and for creating executable knowledge – business solutions – from information *under limiting organisational circumstances*. We draw on this in the design of SAM in a later chapter.

The model of Ferran-Urdaneta (1999) has three phases of knowledge management. We add a fourth stage to the end of the Ferran-Urdaneta model, and fuse it with the SPM activities as follows:

1. knowledge creation stage:

    1.1. SPM 1.1 – *Mission* is our existing hypotheses, which is based on our preconceptions and ideas. The hypotheses are untested for any of the organisation's data, commercial, or operational circumstances;

    1.2. SPM 1.2, 1.3 – these are cognitive activities which *develop* the knowledge, which we *estimate* is required for developing and testing the hypotheses. The development is based in the infusion of new information and knowledge. Because we are operating in an open environment, where there are limiting and extenuating circumstances, we further *profile* this knowledge for the effects of any organisational circumstances;

    1.3. SPM 2 – is a *knowledge analysing* activity. Here we *test* the created knowledge in the hypotheses, for its realism under the organisational circumstances. Since the aim is to progress on to a solution, we may reframe hypotheses until they are supported under the organisational circumstances. If they can not be reframed for support, we have to reject them;

2. knowledge legitimisation stage:

    2.1. SPM 3 – is a *knowledge interpreting* activity, where we *legitimise* the hypothetical knowledge, by *selecting* the best hypotheses for achieving the *Mission* under the organisational circumstances;

    2.2. SPM 4 – are *developmental* knowledge activities, which legitimise the knowledge in the selected hypotheses by developing its *executibility*. The executibility of knowledge resides in the *Strategies hierarchy*. To be of any value in the SPC environment, information therefore has to be developed into new *Operating objectives* and *Operating strategies*. Since the strategies and objectives hierarchy are the organisation's paradigm, the development of any new objectives and strategies, constitute an *executable paradigm shift*. Knowledge which is not executable, remains illegitimate, and no paradigm shift has occurred;

3. knowledge sharing stage – the value of business knowledge lies in it being executable. We consider that in the corporate environment, the execution of knowledge, is the execution of the new *Strategies* which were formulated. That sharing takes the form of training and giving them the tools for its execution (Takeuchi 1998). In the SPM framework, we achieve this sharing through the 7-S elements. SPM 5 therefore is the *executive* activity for the sharing of knowledge;

4. monitor and control – we refine the knowledge management model, by adding this fourth stage. In a changing environment, we assure the ongoing legitimacy and executibility of knowledge over time, through *analytical, interpretive*, and *executive* monitoring and controlling activities. This is similar to SPM 6.

The above demonstrates how SPM forms a practical model for developing information from knowledge, once the organisation has entered an SPC project environment. We also conclude from the above, that in order to proceed from one stage to the next, we need to have successfully constituted the stage we are in. For instance, one can't continue developing knowledge based on a hypothesis which has been rejected.

Our reference to this section in the remainder of the thesis, will be in the form of *KM par. no.* For instance, *KM 4* refers to the knowledge legitimising stage (par. 4) above.

2.5 A practical pre-project schema

In the first chapter, we introduced the concept of an organisation's *pre-project schema*. In this section we present the detail of that schema. The *pre-project schema* is a separate entity from the organisation's existing *paradigm*. The schema is our rationally unqualified, psychological embodiment of the knowledge management framework above. We refer to the pre-project schema as *SCHEMA* in the remainder of this thesis.

The SCHEMA is an organisation's pre-project *expectation* about how a project will assist it from escaping its existing paradigm. The schema has three dimensions, and five stages. The three *dimensions* are:

1. our *preconceptions* about a business problem or opportunity; and
2. our *expectations* on the SPC project about solving the problem or meeting the opportunity; and
3. our *intended application* of supporting technology – like data mining – during the project for realising those expectations.

Ironically, despite all our good preconceptions, expectations, and intentions, SCHEMA in itself can inhibit the potential for paradigm shift. This is because SCHEMA forms part of our psychological make-up, and we saw earlier how our psychological make-up can impact on SPC projects. We therefore include in our evaluation of CRISP-DM, its ability in overcoming limiting SCHEMAs.

SCHEMA has six stages:

1. the business decision-maker has a hypothesis that there is a problem or an opportunity, and the goal is to seek information with which to support, reframe, or reject this hypothesis. The use of technology is for discovering information, which confirms or refutes the existence and extent of a suspected problem or opportunity, or helps with reframing. Hypothesis support or successfull reframing is followed by;
2. the business decision-maker has a hypothesis about the cause(s) of the problem. The goal of the decision-maker is to support, reframe, or refute the hypothesis about the causes of the problem or opportunity, which - through inference - is developed into knowledge about the problem or opportunity (Hastie, Tibshirani et al. 2001, p.99). The use of technology is to discover information about the main components of the problem for inference. These first two stages explore the *business problem* or

opportunity. Following successful hypothesis support or reframing about the causes of the problem, the decision-maker;

3. cognitively formulates a hypothesis - or hypotheses - about solving the problem or about realising the opportunity (Schön 1995). Successful hypothesis formulation is followed by the use of technology for discovering information, which the decision-maker develops into knowledge, which supports or refutes the hypotheses about the solution or response. This stage explores the *business solution* for the problem or opportunity. Following successful hypothesis support about the solution, the decision-maker;
4. cognitively develops the knowledge which supported the hypothesis into an executable solution. The use of technology is for assuring that the knowledge is executable under the organisation's particular commercial and operational circumstances. Once an executable solution has been constituted, the decision-maker will want to;
5. execute the solution. The use of technology here, is supporting the ongoing execution of the solution;
6. the use of the technology for monitoring and controlling about the problem and its solution.

Note the conditional progression from one stage to the next. In practice, we will define a sequence of data mining goals in support of this (Reinartz 1999, p.20).

We will reference this *schema* in this thesis as *SCHEMA par. no.*

2.6 Relevant sections of Telco ABC's Corporate Strategy

2.6.1 Mission (SPM 1.1)

The *Mission* (SPM 1.1) (sometimes referred to as the company *Vision*) is the company's stated perception about itself in terms of its industry and service, primary market and principal technology, and survival[i] (Pearce and Robinson 1991, p.56). We will find in the survival clause, high-level reference to the organisation's awareness of the need to innovate, either by intention, or sometimes negatively expressed in terms of necessity.

[i] Other elements are Philosophy, Self-concept, and Public Image, but we'll keep the critically relevant elements above.

In the mission, the level of abstraction is so high, that measurement criteria are not formulatable.

Telco ABC's Mission: *We want to be the leading provider of:*

- *a mobile telephony service for consumers who have a preference for simple technology and are in the low average revenue per user (ARPU) bracket by industry standards (primary market);*
- *who are geographically concentrated within urban boundaries (geographic area);*
- *through a cost-effective mobile phone network infrastructure (technology);*
- *and want to generate further profitable growth in that low revenue market, and protect our existing market share there, both through some innovative acquisition and retention marketing, and through innovative improvements to the cost efficiency of our mobile phone network (survival).* Note the evidence here of Telco ABC's *intention to change* (Nonaka 1994, p.74).

2.6.2 Controllable factors profile *(SPM 1.2)*

In the corporate strategy environment, this is referred to as the *Company profile.* It is the state of the company's understanding about its own competitive strengths and weaknesses by critical analysis. Because this is a fairly high-level application of the SPM on a qualitative problem, the strengths and weaknesses are expressed qualitatively.

Telco ABC's *Company profile* is:

- *our mobile network has been fully amortised, and it now has the cheapest network to operate and maintain* – strength;
- *we have a very effective in-house call centre infrastructure and skills-base in place* – strength;
- *our network does not have the geographic cover of our competitors'* – weakness;
- *we do not have the resources of our bigger competitors* – weakness;
- *we have above industry loss of existing customers to competitors (churn)* – weakness.

2.6.3 Uncontrollable factors profile *(SPM 1.3)*

In the corporate strategy application, the *Uncontrollable factors* are known as the *External analysis*. The company qualitatively assesses them as either threats or opportunities, which contribute or detract from attaining the *Mission*. The assessment spans from the more remote influences (that which all companies operate in e.g. political and economic environments), through the closer industry environment (that which their industry operates in e.g. suppliers' power, legislation about completion etc.), to the closest external environment (that which the company itself operates in e.g. competitors, labour, suppliers, customers).

Telco ABC's *External analysis* is:

- *we are a small player in our industry* – threat;
- *the industry is mature, making it difficult to find new growth opportunities* – threat and opportunity at the same time;
- *because of market maturity, competing we will aggressively target our customers through acquisition marketing campaigns. We estimate our churn rate to be 3.5 percent per month* – threat;
- *our competitors are beginning to use data mining technology for managing their own voluntary churn problems, making it more difficult for us to poach customer's from them* – threat;
- *as the market matures, there is a fragmentation of consumer behavior regarding mobile telecommunications. On the one hand, there is emerging adaptation of new generation mobile features e.g. content, image transfer, internet access etc. On the other hand, there are the non-adapters, who consider mobile telephony is a basic commodity, and are not moving past the use of traditional features like voice and sms* – threat and opportunity.

2.6.4 Strategic analysis (SPM 2)

For our purposes, we are less concerned with how Telco ABC developed their *desired* and *possible* solutions to reach their *Strategic choice,* and more concerned with the *content* of the *Strategic choice* entity. In the SPC application of the SPM, the *possible* is generally accepted as a quadrant of generic strategies. This approach is particularly suitable when the profiling was qualitative. These four *possibilities* are *Turnaround*,

Expand, *Defend*, and *Diversify,* and they are presented as the four blue rectangles in Figure 2.3: Telco ABC's Strategic choice.

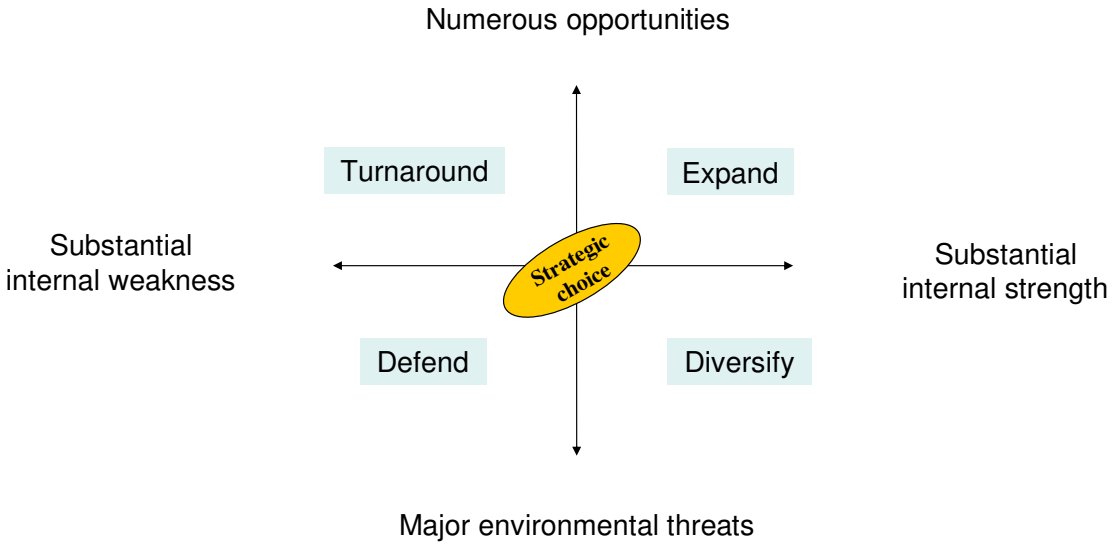


FIGURE 2.3: Telco ABC's Strategic choice

2.6.5 Strategic choice (SPM 3)

We present Telco ABC's Strategic choice in with the shaded ellipse in Figure 2.3. The **Strategic choice** means *We'll innovatively and profitably grow the cost sensitive commoditised segment of the market we already are established in, while we will innovatively defend the profitable sections of our existing market share from poaching by competitors.* Telco ABC's retention problem lies in the Defend quadrant, and the organisation seeks innovatively to defend.

2.6.6 Strategic objectives (SPM 4.1)

In the SPC application of SPM, the *Strategic objectives* are the conceptual, long-term objectives for Profitability, Productivity, Competitive position, Technological leadership, Employee development and relations, and Public responsibility. Even though they are long-term objectives, they are specific enough to have measurable

criteria, and to be motivational. We are particularly interested in the competitive position.

Telco ABC's long-term competitive *Strategic objectives* are:

- profitability – *improve the profitability of the organisation by 5% p.a. each over the next two years*;
- productivity – *(a) reduce the average cost per call of our mobile network by 5% per year over the next two years (b) reduce the call centre cost of our per customer contact by 20% per annum over the next two years*;
- competitive position – *reduce above industry average churn rates to more acceptable levels*;
- technological leadership – *we will not make any capital investment in our network over the next two years to enhance its technological status or to improve its geographic coverage.* This means they will follow in their chosen market.

2.6.7 Grand strategy (SPM 4.2)

From the above *Strategic objectives* Telco ABC have chosen generic grand strategies of *Focus* and *Cost leadership*.

Telco ABC's relevant *Grand strategy* details:

- *we will build our competitive position in the mobile phone market through market retention campaigns.* This *Grand strategy* is called *Focus*.

2.6.8 Operating objectives (SPM 4.3)

They are broken down by the different business functions, e.g. marketing, finance, operations. They are specific within their business function (e.g. marketing, finance, operations) and have measurable success criteria. Telco's relevant *Operating objectives* are:

- Operations function - *reduce the average call duration of our customer call centre by 30% per year for both acquisition and retention calls out, over the next 3 years*;

- Retention marketing function - *reduce our customer churn from our existing estimate 3.5% per month, to between 1% and 1.3% per month within six months, and maintain it there.*

In practice retention objectives get a higher priority than acquisition objectives, because retention is a much more efficient spend of the marketing dollar than acquisition – by as much as 100% (Lenskold 2003, p.2).

2.6.9 Operating strategies (SPM 4.4)

Telco ABC's relevant *Operating strategies* are:

- *we will start using data mining to identify the high-risk potential churners on a monthly basis;*
- *we will target market retention campaigns at customers who have 30 days or less remaining on their existing service plan;*
- *we will segment the targeted customers into six need groupings based on the age demographic parameter:*
  - *16 – 21 years = peer group;*
  - *22 - 30 years = personal aspiration;*
  - *31 – 35 = coupling settlers;*
  - *36 - 50 = family ties;*
  - *51 – 59 = empty nesters;*
  - *60 and over = retired;*
- the *campaign offer* for each group will consist of:
  - *a service plan from our current age-based lifestyle plans portfolio which matches the customer's age;* Some plans include a matching handset;
  - *if that plan does not include a new handset, we will offer a replacement handset at a discounted price, to those customers who have a handset model with reported reliability or reception problems;*
- *we will design the retention campaign offer in-house in our marketing department, and implement them in-house though inserts with the monthly bill mail-out, and with calls from our call center.*

We are not concerned with any of Telco ABC's *Policies* in our research.

2.6.10 Implement and deploy new strategies (Execute) (SPM 5)

The *Implementation and deployment* of the retention management is through a Retention Manager. The Retention Manager is dependent on information, which is supplied by a team of analysts, which is based on his specified needs. The Retention Manager manages the design of the retention campaigns through an in-house team, and coordinates the deployment of these with the organisation's in-house call centre and mail-out teams.

2.6.11 Monitor and control (SPM 6)

The Retention Manager also monitors the success of the retention campaigns, and controls through making modifications to the existing retention campaigns, or through designing new campaigns.

From this section on SPM, we realise that an organisation's objectives and strategies hierarchy, is the cerebral column of the organisation – it keeps the issues apart, while holding them together in the way which the strategic planners have been best able to conceptualise about the business and its interaction with its chosen market. Looking at the SPM from a Knowledge Management perspective, it is the legitimised, condensed and shared knowledge and conscience base of the organisation (Ferran-Urdaneta 1999). As such it *represents the organisation's existing paradigm* (Sveiby 2001), and its formal expressions about its *intentions* about innovation (Von Krogh, Ichijo et al. 2000) (Slembek 2003, p.32).

2.7 Total Quality Management

We present Total Quality Management (TQM), which is a design philosophy that practices *forward linkage of the utility for purpose* of a product or project. *Utility for purpose* is the *relevance of design* for the *purpose*. The *purpose* is the service or product function, which must be achieved. The *forward linkage* moves the attainment of utility forward, from expensive experimentation or from trial-and-error approaches during the productive stages of the manufacturing / project, into the more cost efficient *design stage* of the product / project. The approach is especially valuable in circumstances where experimentation is risky strategically, or very expensive in practice. In other words, it is used in environments where once things go into the productive stage, they

have to happen right *every time,* and *from the first time* (Dobler, Burt et al. 1990, pp. 122, 145, 381) (Pèrigord 1990, pp.106ff.). *TQM's utility* is assured utility for purpose, through effectively defined project deliverables, locked in during the design phase of any project.

A typical manufacturing application, for instance, is automobile design, where it is very costly to physically experiment to find the most suitable design for a specific application vehicle, and as much as possible has to be achieved about utility for that specific purpose, during the vehicle's design phase. With the benefit of hindsight TQM is a perfectly sensible approach to the design of a product, process, or service; however, it was quite a breakthrough when it was first propagated.

TQM can also be used when the deliverables are intangible. An application of TQM in which we are particularly interested for our purposes, is in the SPC. SPM establishes forward linkage of design for purpose in a number of stages:

- first, as the *Strategic choice* (SPM 3) of the *best possible hypothetical solution* under the limiting circumstances, before we spend effort on activities (SPM 4.1 – 4.4) developing executibility for knowledge, which in the first place is inexecutable under organisational circumstances;
- second, as the formulation of the new executable objectives and strategies (SPM 4.1 - 4.4), before the *Execution* activity (SPM 5);
- third, in a project for the implementing of a technology, the *Operating objectives and strategies* constitute the technology's *utility for purpose.*

TQM has the further advantage of being a formal communication protocol, which establishes agreement about the *project deliverables* between the different expert collaborating parties. Such agreement is of particular value for proactively managing disruptive political and resource change management issues, which may arise during SPM's *Realisaton phase*. These change management issues are often referred to as *soft issues*.

The better the designer's subject matter knowledge, the better will be his/her understanding of the *purpose* of design, and the more relevant will be the *utility* of the design (Dobler, Burt et al. 1990, p.409).

2.8 Lessons from the ERP environment

The candidate has qualifications in the Enterprise Resource Planning (ERP) industry's leading SAP R/3 solutions implementation methodology, called ASAP (SAP 2000). ERP has proven very successful over the last two decades, in supporting organisational paradigm shift, which was not possible before without ERP. We attribute the success of ERP to a number of reasons:

- the ERP *approach* is infusing *executable* subject matter knowledge into stale business problem situations. Even though the knowledge is proven, its novelty to the implementing organisation helps with solving previously insurmountable problems, or with meeting opportunities, which were outside the organisation's previous abilities. The new subject matter knowledge of ERP is best-practice business processes. This knowledge is categorised into a number of business functions e.g. finance, supply chain, CRM etc.;
- ERP *software* provides the platform for:
  - *introducing* the knowledge into the organisation; and for
  - *legitimising* the knowledge through providing a technological execution platform (software processes with checks and balances) for executing the knowledge; and for
  - *distributing* that knowledge to the organisational execution points, through the organisation's IT network. This is the *sharing* of knowledge;
- even though the solutions which ERP bring, are legitimate, shareable, and best-of-breed, they have to be *adapted* for the implementing organisation's unique circumstances. ERP projects use extremely sophisticated and integrated project methodologies for adapting both the organisation, and the introduced knowledge, for the organisation's circumstances. These methodologies use the principles of project management, SPM and TQM, collaborative expert teamwork, and mapping techniques for:
  - *analysing the nature of the business problem* or opportunity in light of the new knowledge;

- *defining the possibilities* with solving the problem or realising the opportunity, using the technology platform and the new solution, for the organisation's circumstances;
- *defining project goals* which support the execution of the possible solution;
- *defining and executing any reconfiguration* of the software for realising the project goals. This is aligns the technology plan with the project goals;
- *aligning* the organisational 7-S dimensions to optimise the execution of the solution, assuring the optimal benefits of the solution to the organisation. This includes proactive change management activities, and collaborative, interdisciplinary expert teamwork in the project plan;

❖ developing monitor and control plans and facilities, for assuring the ongoing quality of the executed solution under changing circumstances.

The return on investment (ROI) on the technology was realised threefold:

- reduction in cost from increased operational efficiency and elimination of redundancy;
- increased revenue from realising new opportunities;
- strategic competitive advantage from improved capability in meeting future opportunities and problems.

2.8.1 ERP and data mining compared

During our literature study about data mining, Knowledge Management, and TQM, we recognised *differences* between the *utility* of ERP and data mining technologies, *similarities* between the *purpose* of their projects, and *differences* between the *contents of the project methodologies*. After evaluating the similarities and differences, we developed a firm view, that data mining project methodologies can be greatly benefited, by learning from ERP project methodologies.

One *similarity* was the *purpose* of both ERP and data mining *projects*, to produce business solutions through introducing new knowledge, in a way which was not possible before the technology. The *utility* of the ERP *technology* for lies in providing a practical platform for executing the knowledge management activities. However, this utility alone, proved insufficient in delivering successful business outcomes. The failure

in the early days of ERP to produce successful business outcomes is proof of this. The reason was later diagnosed as deficient project methodologies. Once the ERP industry developed project methodologies which had *KM utility in their own right*, the business outcomes of ERP projects remarkably improved.

We saw in earlier sections, that data mining can discover information, and can also be used for developing knowledge from that information. Similar to ERP, data mining therefore has the potential for *introducing* new subject matter knowledge into the organisation. However, since data mining is also only a technology – like ERP – data mining is also dependent on the knowledge management utility of its project methodology. For this reason we have included utility about knowledge management in our evaluation of CRISP-DM as a project methodology.

The last difference between the two technologies is about the sustainability of their respective ROIs. With ERP projects, the competitive advantage gradually erodes as your competitors also gain access to the generic domain knowledge. With data mining, the potential for developing organisationally unique knowledge remains as long as the organisation's data is different from their competitors. This sustainability of ROI from data mining, offers a strong incentive to optimise data mining project methodologies for the business outcomes they produce.

2.8.2 An example of an ERP project

We present a simplified example of the application of the SPM and TQM components of an ERP project methodology. This example demonstrates the versatility of the SPM being applied to a technical problem in an SPC environment.

2.8.2.1 Mission (SPM 1.1)

The *Mission* in our ERP project is to achieve a *Company A*'s supply chain objectives, using ERP technology as an enabler. From a KM perspective, the hypothesis is that ERP will solve the problem (KM 1.1). These supply chain objectives had been set in the last round of corporate planning. The business problem is that the strategies for attaining these objectives are failing. These strategies had also been set during the last round of corporate planning.

Company A's supply chain *Strategic objectives* (SPM 4.1) are:

- Profitability - *we will increase the organisation's profitability by 2% p.a. over the next 5 years*;
- Productivity - *we will investigate ways of improving the productivity of our supply chain function*;
- Competitive position – *we will investigate ways of improving the order delivery component of our customer service experience*;
- Technological leadership – *we will investigate options for developing technological leadership in our supply chain function*;
- Employee development and relations – *we will maintain the good relationship we have with our labour unions*.

Company A's failing supply chain *Grand strategy* (SPM 4.2) is:

- *using our own in-house developed supply chain software solution called* Homegrown.

Company A's supply chain *Operating objectives* are (SPM 4.3):

- *we will reduce costs in our supply chain operation by 15% per annum over the next two years each, which equates to an 1.5% increase in overall company profits*;
- *while at the same time reducing customer delivery complaints by 30% per annum over the next five years*;

This represents the organisation's current *paradigm*, or the currently legitimised knowledge.

2.8.2.2 Controllable factors profile *(SPM 1.2)*

The *Controllable factors* are the existing solution, and any alternative solutions which may be on offer. The profile of the existing *Homegrown* as a solution is, captured in Company A's existing supply chain *Operating strategies*:

- *we use business processes A in* Homegrown *for goods receipt*;
- *we use business process B in* Homegrown *for order processing*;
- *in addition, we use business process C in* Homegrown *for processing delivery of orders*.

The analysis of *Homegrown's* processes A, B, and C highlights the following:

- process A lacks integrity checks on the receipt of goods, resulting in goods being receipted incorrectly, labelled incorrectly, stored incorrectly, and incorrect goods being sent out. This results in costly order reworks, goods returns, and customer complaints;
- the process B does not allow us to enter an order receipt date and time when we process an incoming order, leaving us unable to prioritise order processing by date and time received. Customer are complaining that orders are not delivered in the same sequence as what they were placed with us;
- process C has no facility to track deliveries dates and times after they have been dispatched, resulting in customers not knowing when their goods will be delivered. This results in customers over-ordering to ensure they do not run out of stock, and then complaining when the extra stock arrives too.

Each of a number of ERP vendors offers an alternative solution to *Homegrown*, based on best-practice subject matter knowledge. Here we introduce the new knowledge (KM 1.2). To maintain the tractability of the example, we will include only the alternative solutions of ERP *Vendor A*:

- Process X is for the receipt of goods, and covers the same functional area as *Homegrown's* process A. Process X left unmodified, will meet all *Company A*'s goods receipt operational needs, plus bring improvements, plus it will resolve the known problem with *Homegrown's* process A;
- Process Y is for processing of incoming orders, and covers the same functional area as *Homegrown's* process B. Process Y left unmodified, will meet all *Company A*'s order processing needs, plus it will resolve the problem with *Homegrown's* process B;
- Process Z is for processing of outgoing orders, and covers the same functional area as *Homegrown's* process C. Process Z needs some modification to meet all *Company A*'s operational needs, but will then bring improvement in some other areas, and will resolve the problem with *Homegrown's* process C;
- Processes Y and Z can be further modified to reduce staff numbers in the supply chain function by a quantity of 10.

2.8.2.3 Uncontrollable factors profile *(SPM 1.3)*

Company A is not achieving either of the operating objectives using *Homegrown*. Homegrown is causing losses of 15% of supply chain costs - 1.5% of the company's overall profits – due to inefficiencies.

Further, the *Company A* is suffering customer churn due to deteriorating services associated with *Homegrown*. These losses are equal to 20% of supply chain costs, or 2% of the company's profits per year. In total *Homegrown* then is responsible for losses of 35% per year, the 15% and 20% added together.

Company A has determined that the programming language within *Homegrown* has become redundant. For this reason, it will be prohibitively expensive to modify Homegrown for overcoming its existing deficiencies.

Were *Vendor A* to make the additional modifications to processes Y and Z for reducing personnel numbers, *Company A* estimates that the union will not accept such redundancies. *Company A* estimates that they will incur labour unrest that will disrupt supply chain services for two weeks. This will result in them having to shut production down for one week, because they can only carry one week's inventory.

2.8.2.4 Strategic analysis (SPM 2) (KM 1.3)

Solution A is persisting with *Homegrown*.

Vendor A's unmodified processes X and Y, and the first modification to process Z, will improve net profit by 3%. This is calculated as a 15% saving in supply chain costs (1.5% in company profits) through attaining the objectives, plus a 5% saving in supply chain costs (0.5% contribution to company profits) because of the improvements to other areas. There is an additional contribution of 10% of supply chain costs (1% of company profits) from increased margins on supply chain services because of other improvements. We name this *possible Solution B*.

If the second modification is done to process Z – the one which will reduce staff numbers – *Company A* calculates an additional saving of another 5% in supply chain costs (0.5% in company profits) due to reduction in staff numbers. However, the costs of pursuing this option with a recalcitrant union will *detract* from the desirability of this solution. *Company A* calculates the value *detracted* as -15% of supply chain costs (-1.5% of annual profit); the net effect of the staff savings and loss of production being -

10% of supply chain costs (1% of company profit). To get the *possibility* of this second modified solution, they have to offset this -10% (-1%, of profit) with the 30% (3% of profit) *net profit value* in the paragraph above, to get an overall *net profit value* of 20% of supply chain costs (2% of company profit) for this second modification to process Z. We name this *possible Solution C*.

The three possible solutions, and their net profit value therefore are:

Solution A: -3.5%

Solution B: 3%

Solution C: 2%

2.8.2.5 Strategic choice (SPM 3)(KM 2.1)

Since *Solution B* has the biggest net profit value, the *Strategic choice* is made for *Solution B.* This choice establishes the *bearing* for strategic paradigm shift.

2.8.2.6 New Strategic and Operating strategies and objectives (SPM 4.1 – 4.4)(KM2.2)

Developing the new *Strategic choice* entity into new *Strategic objectives* and a *Grand strategy*, *establishes* the strategic paradigm shift.

2.8.2.7 Company A's new Strategic objectives *(SPM 4.1)*

- Profitability – *we will increase the organisation's profitability by 2.5% p.a. over the next 5 years.* The extra 0.5% over the previous 2%, is made possible by the choice of *Solution B*. This 0.5% profitability is a quantitative *measure* of the constituted strategic paradigm shift;
- Productivity – *we will pursue identified productivity improvements in our supply chain function before the beginning of our next financial year.* The paradigm shift is from *investigation* to *pursuit*, plus the fact that there now is a definite time scale identified for reaching the objective;
- Competitive position – *we will improve the order delivery systems from the beginning of the next financial year*. The paradigm shift is that the improvement will now proceed, and that a time frame has been identified for it;
- Technological leadership – *we will implement new supply technology in our supply chain function before the start of our next financial year.* The paradigm

shift is that a decision has been made to proceed, and a time frame set for achieving it;

- Employee development and relations – *we will maintain the good relationship we have with our labour unions.* There is no paradigm shift on this objective;

2.8.2.8 Company A's new Grand strategy *(SPM 4.2)*

We develop the new supporting *Grand strategy* for the *Strategic choice* entity:

- *We will replace* Homegrown *as our supply chain solution.* Strategic paradigm shift is established in the replacement of Homegrown by something new;
- *We will deploy (go live) with* Vendor A's Solution B *at the start of our next financial year;*
- *We will use* Vendor A's *implementation methodology for implementing* Solution B*;*
- *We will jointly resource the implementation team.*

The last three strategies were not in the old *Grand strategy,* and constitute strategic paradigm shift.

2.8.2.9 Company A's new Operating objectives *(SPM 4.3)*

Company A defines their new *Operating objectives* as:

- *We will reduce costs in our supply chain operation by 30% p.a. in the next financial year, which equates to a 3% increase in overall company profits;*
- *We will achieve a 90% reduction in customer complaints by the middle of the first financial year.*

The improvements from this paradigm shift are:

- an increase in targeted savings from 15% p.a. over two years, to a 30% saving p.a. in the first year. This means that *Vendor A's Solution B* will gain *Company A* a full year in time in realising a total 30% saving in their supply function;
- a virtual elimination of customer complaints from the beginning of the next financial year, compared to a 30% improvement target p.a. over 5 years.

These potential benefits have to be *realised* through new *Operating strategies*, which we now define.

2.8.2.10 Company A's new Operating strategies *(SPM 4.4)*

Company A defines its new *Operating strategies*:

- *we will use* Vendor A's *unmodified processes X and Y to replace* Homegrown's *processes A and B;*
- *we will use* Vendor A's *process Z, modified in ways 1, 2 and 3 to replace* Homegrown's *process C;*
- *we will remodel our operations which are currently supported by processes A, B and C, to dovetail with the best practise of business processes X, Y, and Z;*
- *we resource the implementation team with 60% vendor contractors and 40% internal staff secondments onto the project, and they will all be under direction of the vendor. However, our Operations Manager has a veto rights in all decisions.*

The operational paradigm shift is:

- the replacement of all the *Homegrown* processes with *Vendor A's;*
- the remodelling of the operations to optimise the integration of the new solution.

We present the application of the SPM on the ERP problem visually in Figure 2.4:

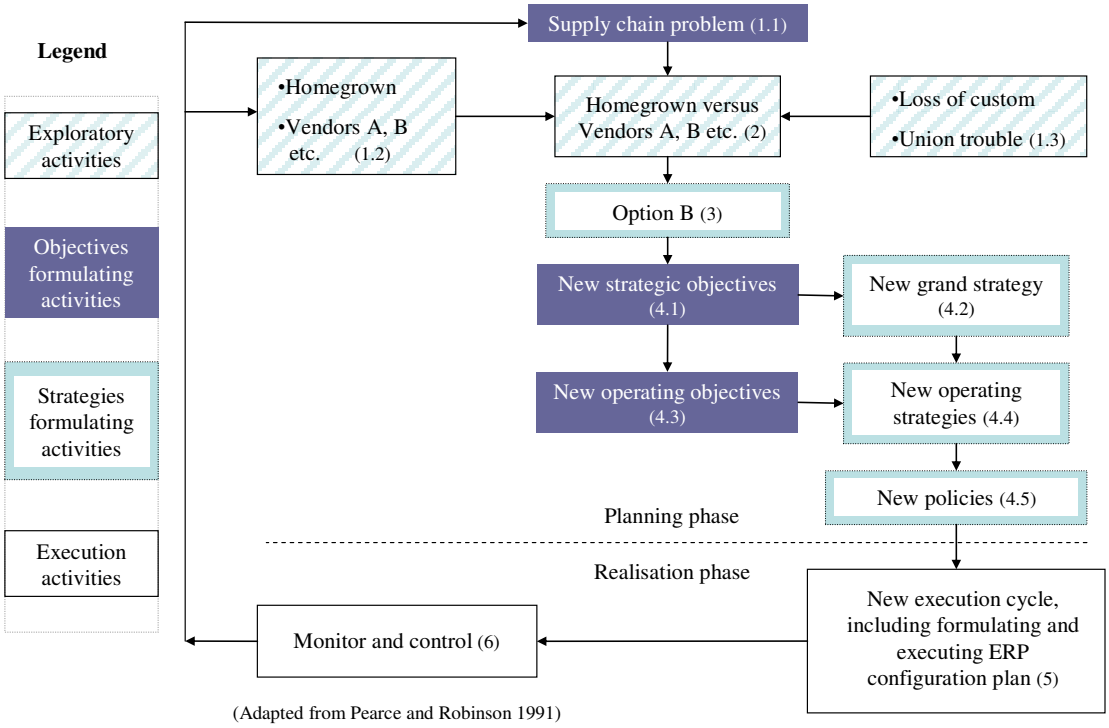


Figure 2.4: Application of SPM in an ERP project

2.8.2.11 Implement and deploy new strategies *(SPM 5)(KM 3)*

New *Implement and deploy* activities are required for *realising* the operational paradigm shift. We indicate this by the new content in the SPM 5 rectangle in Figure 2.4. This includes an ERP project plan. That plan includes reconfiguration of the ERP processes that support *Company A's* new operating *strategies*. This ERP project plan is what should be called the *data mining plan* in the data mining methodologies.

The *mapping technique* between the business objectives and strategies, and the ERP technology plan, is making the *Operating strategies* the *objectives* of the ERP project plan. The ERP objectives then become:

- to implement process X unmodified;
- to implement process Y unmodified;
- to modify process Z in way 1, 2 and 3 and implement.

The *Technology objectives* are then realised through *Technology strategies*. Technology strategies will be developed for making the required software changes to module Z.

Note from Figure 2.4 how that in projects like this, we first define the business solution which is executable by the organisation – under the organisation's circumstances - before proceeding with the formulation of the technology plan in *Implement and deploy new strategies*. This is a TQM *forward linkage of utility of purpose* of the technology, within the SPC project environment.

The *Implementation and deployment* activities for the supply chain function will further entail a matching reconfiguration of the 7-S elements, to embed the new subject matter knowledge injected by the project.

2.8.3 Role of new knowledge and collaborative expert teamwork

The new knowledge, which is injected into the stale situation, presents the new possibilities for the organisation, and guides the defining and execution of the project. The influence of the new knowledge starts with the defining of the *Mission*.

ERP solutions require *substantial subject matter knowledge* from both the technical and business experts involved in the project. We saw in the section on Knowledge Management, how expert collaboration is a very effective tool for harnessing new knowledge within an organisation. Expert collaboration functions as an expert-reflective-in-action conversation during all the stages of an ERP project.

2.8.4 Role of Strategic Planning Model in the ERP solution

- The application within SPM of TQM's forward linkage of *utility for purpose* to the design stage, assures both the utility of the *project goals* and of the *technology goals*. This assures that the technology plan directly supports the project goals, producing acceptable business outcomes;
- SPM further assures a lock-in of project scope. This prevents project creep, or the abandonment of the project at a later stage, when it is discovered to be aimless. This lock-in also allows the planning for, and commitment of limited organisational resources on the project;
- the SPM establishes binding agreement of *what* needs to be achieved, and *how* that is achieved, between all the parties with political interest in the project. It further serves as a communication instrument to those in the organisation, which are not directly involved in the project. As such, it is a technique for proactively managing hidden soft issues, before they manifest themselves during the *Implementation and*

deployment activity. Examples of such soft issues are misunderstanding about practical issues, or political hedging between the parties, at the point where the technology and the business integrate;

- the SPM sets the agenda for the expert collaborative teamwork;
- SPM helps with distinguishing between three sequential stages in SPC projects:
  - the defining of the *project goals* which establishes the business deliverables;
  - the defining of the *utility* of the technology, required by the business deliverables; and
  - the *execution* of the business solution which realises the business deliverables.

  The benefits of this distinction for our research purposes, is the realisation that it is not the application of a technology, which brings the paradigm shift. Technology only serves as a *carrier* of newly injected subject matter knowledge, and as an *enabler* for sharing that knowledge;
- the measurement criteria of the reformulated objectives, provide the anchor points for the *Monitor and control* activity.

2.9 Risk Management

Risk Management essentially is the science of identifying the chance of an unwanted event, in order to eliminate it, or to mitigate or isolate the impact of the unwanted event if it cannot be eliminated. We draw on the Australian and New Zealand Risk Standard AS/NZS 4360:1999, which is a generic, best-practice framework used across a wide spectrum of private and public industries for managing risk.

Risk Management's benefit is the objective prioritisation of limited resources that are available for risk management, so that the objective is achieved in the most cost efficient way. The link between risk management and our data mining on Telco ABC's retention problem is that both have the purpose of reducing uncertainty about the churn event. The data mining reduces risk through discovering information about the churners (who they are and when it will happen), and the risk management through giving us a formalised perspective on managing uncertainty or unwontedness. Risk management offers good perspectives on managing retention as an unwanted and uncertain event.

The data mining then becomes a primary tool for discovering information that can be used to develop knowledge for the managing of the retention problem.

A risk management has the following four components:

1. *risk event* is the event which is unwanted, or about which we want to reduce uncertainty; in the case of our Telco ABC it is voluntary customer churn;
2. *event likelihood* is the likelihood (or chance) of the event occurring within a time. With our customer retention problem, it is the likelihood of each customer churning over the time window of our prediction period of 2 – 90 days. In our research, we obtain the *event likelihood* from output produced by our modelling. The feature containing the output is called *p_t1*;
3. *event impact,* where the impact is the *magnitude of the event were it to occur*. In the case of Telco ABC's customer churn, the magnitude of the event is the profit that is lost with each customer leaving. In our industrial Telco setting, the *cost* associated with each customer is not available for publication, so we are unable to calculate profit. Instead of *profit*, we will set our event impact equal to the Telco industry standard measure of the last three months average *revenue* per user, called *ARPU*. In the case of our retention problem, we want to be able to prioritise the retention resources toward the highest impact churners by measure of *ARPU*;
4. *event risk level* = *event likelihood* * *event impact*. In practice, this product is used to prioritise the time resource between *risk events*, so that the *risk event* with the highest *event risk level*, always gets the first attention. In the case of Telco ABC's retention problem, that would equate to *risk level* = *p_t1* * *ARPU*.

Risk defined, we will innovate in our segmentation, by including either a combined *event risk level*, and / or one or both of its constituents i.e. *event likelihood* and *event impact* This could be of value when prioritisating resources during the retention campaigns. The framework is presented visually in Figure 2.5:

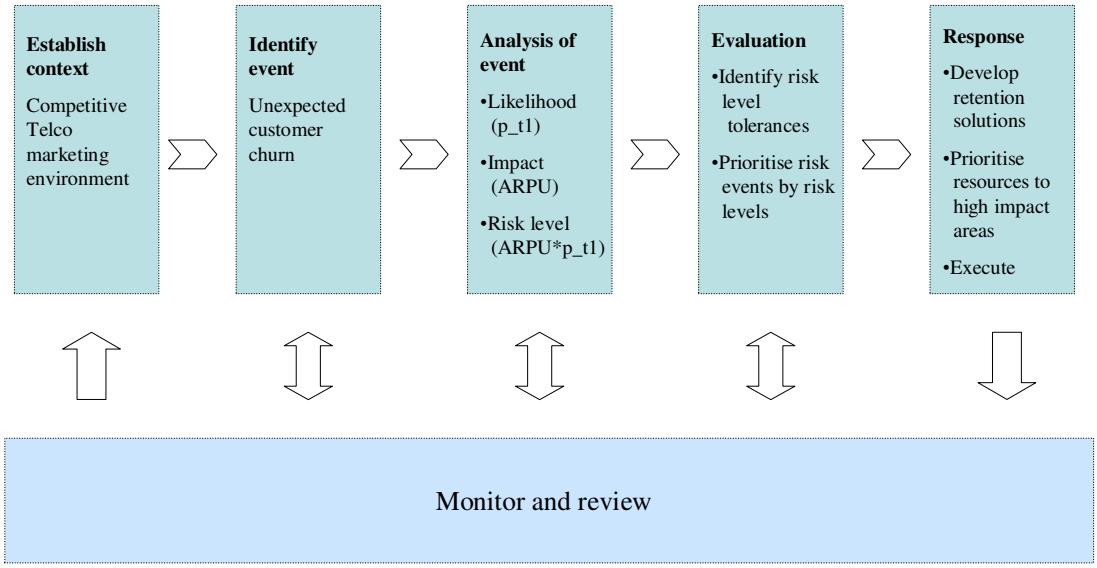


Figure 2.5: A retention Risk Management framework for Telco ABC

The *context* of our event is the competitive telecommunications market, where industry average voluntary churn rates are estimated to be about 2.5% per month. That is the average industry event likelihood. The *determination* of Telco ABC's event *likelihood* within this project - using data mining - will allow us later to calculate their event *risk level*. Based on these two criteria, we then later identify an appropriate *risk tolerance* for Telco ABC regarding their retention management problem. This means that those customers of Telco ABC we find to have a *risk level* which lies outside the *risk tolerance*, will be targeted by a retention solution.

2.10 New marketing subject matter

Our research was in the mobile telecommunications industry, and the business problem was that of unexpected, voluntary customer churn. It is a problem of serious negative economic consequences for the industry – mounting into the 10's of million dollars per annum *per service provider* - calculated with Formula 1 as:

$$\left[\frac{Q_{churn}}{Q_{acquired}} \bullet M\right] + \left[\frac{Q_{churn}}{::Q_{total}} \times CRM\right] + R_{churn} \qquad \text{Formula 1}$$

where *over a measurement* period:

- Q_{churn} is the quantity of customers who churned;
- $Q_{acquired}$ is the number of customers that were acquired through marketing;
- M is marketing expense in dollars;
- $::Q_{total}$ is the arithmetic mean of the number of customers;
- CRM is the dollar amount spent on Customer Relationship Management (CRM);
- R_{churn} is the lost revenue from churned customers.

In other words:

- money that has been spent on acquiring customers who churned in the end;
- money spent on CRM maintaining customers who churned in the end; and
- lost future revenue from the churned customers.

A cost we do not include in the formula, is the loss of long-term strategic differentiation in the market place (Porter 2002, p.36). Using the following industry standard figures for a Telco an example calculation in USD is:

$$\left[\frac{300000}{400000} \bullet USD30\right] + \left[\frac{300000}{1000000} \bullet USD20\right] + \left[USD300 \bullet 300000\right]$$

$$= \left[0.75 \bullet USD12000000\right] + \left[0.3 \bullet USD20000000\right] + \left[USD90000000\right]$$

$$= \left[USD9000000\right] + \left[USD6000000\right] + \left[USD90000000\right]$$

$$= USD105000000$$

where the calculation period is one year:

Q_{churn} = 30% (or 2.5% per month of the average number of customers);

$Q_{acquired}$ = 40% of the average number of customers;

M = USD 30-00 per customer;

:: Q_{total} = 1 000 000 customers average;

CRM = USD20-00 per customer;

R_{churn} = USD 300-00 per customer

We recall that in the SPC environment, the purpose with any problem extends past understanding its extent and nature, to include its solution (Meltzer 2000, p.1). The desire of Telco ABC to remedy the problem is known from their corporate objectives and strategies hierarchy. We also identified in their strategy, the intent to remediate the problem innovatively. Because our business problem is a marketing one, we turned to the marketing literature:

- to find guidance about what constitutes executibility in marketing;
- to identify new marketing domain knowledge which we can inject into the stale retention problem;
- to learn about recent technical developments in the application of data mining techniques in market segmentation, upon which we can draw in or modeling approach.

The classic *STP* principle of market acquisition provides our executibility requirements (Kotler 1988) (Kotler 2002). We introduce the principle here, and develop the theme for our situation:

- Segment – group marketing objects together who are similar by measures about parameters of commercial interest;
- Target – choose by criteria of commercial relevance, the segments most attractive commercially;
- Position – develop marketing plans, and campaign offers and action plans, with which to market to the targeted segments.

The above elements are sequenced for *acquisition* marketing, in other words marketing where you are trying to convince mobile phone users, to use your service; in our case, our problem is a market *retention* one, where the aim is to convince *existing* customers *not to stop* using your service. For a retention problem, we therefore modify the classic principle to *TSP* – Target, Segment, and Position. We want to identify and target those

with the highest probability of becoming a churner, segment them for campaign design, and then position our retention marketing to the segments.

2.10.1 Target

Targeting requires known identity over a certain time. Practically, we need to:

- discover the posterior probability of each customer to become a voluntary churner;
- within a time window which is practical for the organisation to complete the campaigning within;
- in addition, identify those potential churners that will have the biggest impact on the profitability objectives.

We can achieve steps 1 and 2 with a probabilistic classifier, using behavioral data of a 92-day period leading up to actual churn. Such a model has the added benefit, that a domain expert can infer the root causes from its parameters (Hastie, Tibshirani et al. 2001, p.99).

The second element of the retention targeting, is the impact of the potential churners on organisational profitability. We know from Telco ABC' operating strategies, that they are interested in changing their retention management, to better focus on those potential churners that will contribute toward the profitability objective. In order to achieve this, we innovate by drawing on industry sources for knowledge about *value segmentation* (Griffin 2003) (Mattison 1999, p.124) (Meltzer 2000, p.21). We find there the principle of *Recency-Frequency-Monetary* (RFM) (Badgett, Connor et al. 2003, p.2) or 'RFM-A', where the A is for average (DataPlus Millennium 2001). This has also been described as the *billed services behavior* (Slepian 2003, p.2), and its purpose is to identify segments *containing customers or prospects with high profit potential…* (Thearling 2003). In more detail, *RFM-A* then is:

- Recency – time lapsed since the last transaction with the customer, or the time period over which the measurement was made;
- Frequency – how many transactions the customer has made over the same selected period;

- Monetary and Monetary-A – the dollar profit, or revenue, for the customer over the selected period of time; acronymously referred to as *ARPU* (Average Revenue Per User) by the telecommunications industry, calculated over a 12-week period (Mattison 1999, chapter 9), (Costa Dr. 2001, pp.14, 22, 30ff.).

ARPU is available in our data as a variable called *Rev53*. In light of this principle, the logical solution in the case of Telco ABC is to target for retention only those potential churners that have an *ARPU* above the Telco ABC average.

There is also evidence in the literature, for targeting only those potential churners, which meet certain profitability criteria (Badgett, Connor et al. 2003).

2.10.2 Segment

To learn about the second component of executibility, we consulted an authoritative literature source about recent innovations in the application of advanced models to market segmentation problems (Wedel and Kamakura 2000). The intention was to learn about the application of techniques for segmentation. However, the source also provided new marketing domain knowledge, which we injected into Telco ABC's problem.

The maxim of marketing is *meeting the customer's specific need*. However, there are problems with applying this maxim when there are many customers; it will be impractical and prohibitively expensive, to develop a product or service which uniquely meets each customer's needs. Marketers call the strategy for dealing with this problem, *market segmentation*. It entails segmenting the market into groups with sufficiently homogenous needs, to allow for the development of group-tailored products or services that meet the homogenised need.

Philip Kotler (Kotler 2002), the father of market segmentation (Mazanec and Strasser 2000, p.12), defined *market segmentation* is '…the task of breaking the …. market … into segments that share common properties.' (Kotler 1988, p.69) and those properties can be behaviors, identifying characteristics.' (Best 2000, p.105). The *assumption* underlying clustering is that '… consumers differ from each other, but only to a certain extent, so that they can be grouped into relatively homogenous groups or segments.' (Wedel and Kamakura 2000, pp.325, 329).

More technically, market segmentation is '… an exploratory and descriptive method of data analysis where a group of objects … is partitioned into subgroups based upon some

measure of proximity (similarity, dissimilarity or distance) between those individual objects.' (Woods 2003b, p.1). An intuitive explanation of clustering is that it finds similarities between observations within the data, with reference to defined columns. This could be contrasted with Factor Analysis, where Factor Analysis finds similarities between columns of data, with reference to the observations (Diekhoff 1992, p.361).

Marketers infer the homogenised customer needs from those measures of proximity. Subsequently market segmentation is the discovery of groups of people with an homogeneous need and other properties of commercial interest or *managerial relevance* (Wedel and Kamakura 2000, pp.3, 296) (Westphal and Blaxton 1998, p.xv) (Best 2000, p.109) (Woods 2003b). This benefits marketers, by allowing them to target their campaign offers to those homogenised groups.

The measures of proximity, or properties, or features upon which segmentation can be based when we use data mining clustering techniques for the segmentation, are collectively referred to as the *clustering / segmentation base* (Wedel and Kamakura 2000, pp.5, 43, 189). These features can be based on customer characteristics, behavior etc. Using a *combination of multiple measures* of commercial interest for market segmentation, is an established practice (Kotler 1988, p.69).

In important market segmentation source presents recent innovations in the choice of the clustering base. This innovation is best understood from an evolutionary perspective on the history of this choice (Wedel and Kamakura 2000):

- Segmenting by *demographics* – the assumption being that customers of the same demographics have the same needs. This proved ineffective, with improving social research methods eventually disproving this assumption. Intuitively, there may be differing needs within segments of the same demographic characteristic / profile;
- Segmentation by *behavior* – the assumption for this approach, is that people with the same behavior, have the same needs. This phase subsequently reached maturity when marketers realized that people with the same needs, do not necessarily behave in the same way. This is especially true when maturing markets are proliferated by products and services competing to meet the same needs, while offering different behavior patterns to their users while the needs are being met;

- Segmentation by *combining needs and the intention to fill the need in a specific way* – is the current innovative approach, which recognises different consumer behavior under similar needs conditions. The benefit of this approach is that marketers can refine their marketing with an estimation of the chance that the customer will respond favorably to their specific marketing offer over a competitor's offer. We seek to employ this innovation in our data mining solution for Telco ABC's retention problem.

2.10.3 Position

In practical marketing terms, profiling is the formalized understanding (GhostMiner 2002) about the segmented target in terms of factors under the organization's control, which are partly used for the formulation of the *P* of *STP,* or *Positioning*. The segment profile constitutes actionable information about the tangible components a customer's experience with Telco ABC (handset, plan type), upon which the retention marketing campaign offer can be based. Put another way it is the *levers* within the company's control that can be *pulled* to influence the target. The profiling provides executibility to two of the *Merchandise* component, of what is known in the Telco industry campaign terminology as the '5M's' (Mattison 1999, p.239). We present a brief summary of the '5M's', to place profiling contextually:

- *Market* – this is those customers the company will direct its marketing efforts to;
- *Margin* – the net profit margin from the customer over a selected period (usually at least for the duration of the new plan you are going to offer them) after marketing and ongoing operational costs. Where a Telco does not have visibility about *Margin*, *ARPU* is often used as a proxy;
- *Merchandise* – the product or service that will be offered e.g. an upgraded handset, or a more suitable plan etc. These are the experiential components of the customer's relationship with the company, which the company can manipulate and modify, for a retention offer to the customer. (There may be experiential components which the company can or does not want to modify, like the customer service level);
- *Medium* – the communication medium that will be used to communicate the message e.g. TV advertisement, or a printed brochure posted with the monthly

bill, or a personal phone call from their in-house call centre. This is often referred to as *Channel*;

- *Message* – relates to conceptual elements of the 'approach' to each market segment, e.g. its objective (to prevent the customer from leaving), and the emotional and cognitive mood of the message e.g. aggression v subtlety, fear v reason;

The Clustering node of SAS Enterprise Miner, calculates the frequency percentages of all the categorical features. This technology is referred to by (Wedel and Kamakura 2000, p.33) as the ability to ... *identify segments and simultaneously profile them with consumer descriptors.* These statistics are available in table format for use in any merchandise profiling we want to do for retention offer design.

There is ample evidence in the marketing literature, that a change in a customer's need *after he/she has become a customer*, results in customer dissatisfaction (Engel, Blackwell et al. 1995) (Goncalves 1998) (Kurz and Clow 1998) (Rosa 2002) (Yassael 1998) (Lovelock 2000). This is dissatisfaction is particularly relevant in the Telco industry. Here, becoming a customer usually locks you into a service contract and handset combination, which is matched to a particular use pattern. When your use – and therefore need – changed during that contract period, you experience dissatisfaction. To make matters worse, the industry penalises customers who want to move to a plan which better suits their new need during the initial contract period.

According to those sources, customer dissatisfaction may also arise from other causes than a change of need. Such causes may be poor customer service, frustration with a faulty handset, poor reception, network faults, poor customer service etc. One purpose with the content of the offer then, is offering the best possible combination of elements that meet the customer's (changed) need (Costa Dr. 2001, pp.28, 29).

A number of things need therefore to be achieved in profiling need. The first is an overlay of the customer's existing usage behavior and their existing merchandise (plan type and handset type). The domain expert strategically analyses the overlay, and identifies any discrepancies between the customer's demand function, and the utility of the merchandise. So for instance, if a customer has a high demand function, but:

- has a handset which is not suitable for high use; or

- is on a plan which does not have a rate structure which is suitable for high use;

then a discrepancy is identified. The essence of the new offer in the Telco industry, would be to rectify such a discrepancy, by offering a plan and handset combination that is better suited to the customer's demand function (SAS Institute 2004) (Emagine 2003). In more detail, a customer can be offered:

- a plan with a call cost structure which is optimal to the customer's demand function - *price personalization* (Compton 2001);
- more loyal customers should respond better to a new plan with a longer duration and vice versa - *service personalization* (Compton 2001);
- people, who make either long calls, or many calls, are offered a more convenient handset type for frequent use, than those who have different call behavior - *product personalization* (Compton 2001).

Another purpose with the offer, is addressing any other cause of customer dissatisfaction. The profiling then needs to include profiling root cause(s) of the problem. The profiling of root cause should be done on a segmented basis. There is evidence in the literature (Hastie, Tibshirani et al. 2001, p.99), that the effects in the model, can be used for inference about root cause of a problem. Another approach, could be to do analyse the results from Principal Components Analysis within each segment (Apley 2003).

By the nature of our approach, the data about the customers who are identified as *potential churner*, will have good signal about root cause. This signal will be present in the six retention segments. This opens up the possibility, that the new retention offers of Telco ABC could include addressing the root cause for potential churn within each segment.

2.10.4 Algorithmic innovations: PROMIX

Wedel and Kamakura developed a two-stage segmenting algorithm for the *acquisition marketing* environment, called GLIMMIX (Wedel and Kamakura 2000, pp.XIX-XXII, 19, 106ff.). They describe it as a *finite mixture model*. It is *mixed* because it uses a combination of class and interval variables, and *finite* because it assumes linearity in the data structure. The GLIMMIX algorithm:

- uses an R-sqr linear method to determine which features will offer optimal clustering; and then
- executes the clustering using a k-means algorithm, after an operator has chosen a commercial suitable base from those statistically significant variables.

Because our retention problem basically is a marketing problem, we evaluated GLIMMIX for application in Telco ABC's problem environment. We found the concept relevant to our problem, but insufficient for solving our problem. We will now explain this:

- GLIMMIX was tested on small, narrow data sets. Such data sets are significantly different to the one, which contained Telco ABC's retention problem; it is wide and large, and therefore contains a lot of variability. We are therefore firstly uncertain about the technical sufficiently of GLIMMIX in our data;
- GLIMMIX was tested on data, which contained linear structure, by design. The problem structure within Telco ABC's data is non-linear, because it is a two-class, probabilistic problem. An algorithm which learns a linear function, is unsuitable for our data;
- our problem is a *market retention* problem, and not a market acquisition one. In *market acquisition*, the main purpose with segmentation is to determine how many segments there are, and then how are they different segment basically. This is known as an unsupervised classification problem. In market acquisition, is only one round of segmentation (Wedel and Kamakura 2000, pp.315, 316). With a retention problem like ours, we have two rounds of segmenting the data. The first time, we know how many segments we want, and this is known a supervised classification problem. We want two segments; one segment for *probable churner* and one segment for *non-probable churner*. Our second round of segmentation is an unsupervised classification problem, just like the acquisition problem.

GLIMMIX will not fair well on a non-linear problem, and does not provision for the first round of supervised classification, as is required by a retention problem. Therefore, GLIMMIX was not suitable for our problem. However, GLIMMIX did give us knowledge for developing our own algorithm, more suitable to our problem.

In GLIMMIX terminology, our algorithm is a *five-staged finitely mixed model*. It is *finite* because – even though there is non-linearity in our data - we still model a linear *but more suitable* binomial function in the first round. This function is the log of the odds (Woods 2003a). In the second round, we also model a linear function, minimising the cubic cluster criterion, using k-means. Our algorithm is *mixed* in its use of features, but in a different sense to GLIMMIX. In the first round, we depend on a combination of class and interval features, but only after having discretised the interval variables. In the first round, we therefore use only discreet features. In the second round we use only interval features:

- discretise all interval features. There is evidence that discretising interval variables produce more accurate results in a Bayesian modeling environment (Woods 2003a) (SAS Institute online c) (Hastie, Tibshirani et al. 2001);
- select the features most influential in determining class membership, using an χ^2 algorithm. χ^2 is more suitable in a two-class problem like ours, than R^2 (SAS Institute online d) (Woods 2003a);
- build a supervised Bayesian classifier using LogReg to achieve the two-class separation (Mattison 1999, chapter 12), and for identifying the most important features for inference about class membership;
- select those observations, which fall outside the risk threshold. The risk threshold was introduced in the section on Risk management; then
- apply supervised k-means to a managerially relevant segmentation base. It is a supervised application of k-means, because we ask for a predetermined number of segments, which fit with organisation's retention management strategy. It is also supervised, because the segmentation base is determined by managerial relevance.

We call this algorithm PROMIX. PROMIX is generically applicable to retention management problems, where the underlying dynamics are behavioral. The name derives from the probabilistic nature of the retention problem, and the combined use of only discrete features in the first round and only interval features in the next round. It constitutes a so-called *chain of models* (Wedel and Kamakura 2000, p.245), or what CRISP-DM refers to as the *combination of problems* required to solve the business problem.

The parameters, which are adjusted for managerial relevance, are:

- the number of classes included in the problem. There may for instance be three classes in the problem, not two;
- features of political importance can be forced into the model despite not having been selected as statistically important;
- the number of segments asked for in the second round. This is determined by practical marketing considerations;
- the choice of base features in the second round, for managerial relevance.

In our Telco ABC application, we ask for six segments in the second round, to fit the organisation's existing retention management strategy. We select the base features in an innovative way:

- innovatively combine *Recency* and *Frequency* with one variable, using the sum of mobile voice phone calls that were made over the last three months (Costa Dr. 2001, p.11). This constitutes the *demand function* of marketing (Wedel and Kamakura 2000, p.26) (Badgett, Connor et al. 2003, p.1), from which we will infer that the more calls someone made over this period, the higher their need for mobile voice telephony. We don't factor in WAP and SMS, to keep our demonstration simplified;
- there is authority for considering customer loyalty in retention management (Badgett, Connor et al. 2003, pp.3, 17). Customer loyalty can be defined as '…the customer behaviour of a sustained profit stream without the need for incremental marketing investment.' Traditionally the value of knowledge about customer loyalty lay in calculating the relative value of incentives offered to customers in order to retain them. We include in our clustering base a derived feature, which expresses customer loyalty to Telco ABC. This expresses the customer's attitude toward the company's service (Levy 2001, pp.1, 2). We innovate however, and infer from this feature a psychographic measure of a customer's *intention to respond positively* to a retention offer (Wedel and Kamakura 2000, pp.7, 14, 17) (Mazanec and Strasser 2000, p.15). Further there is evidence in the industry literature that this attitudinal dimension is an indicator of the relative duration of a new contract that could be offered to a potential

churner in a retention campaign offer – the higher the loyalty index, the longer we expect them to renew their commitment for (Compton 2001). Our loyalty index can also be considered as a representation of post-purchase behavior, where those customers with high loyalty had a very positive post-purchase experience, and vice versa. We refer to (Engel, Blackwell et al. 1995; Schiffman and Kanuk 1997; Goncalves 1998; Kurz and Clow 1998; Yassael 1998; Lovelock 2000);

- we include *ARPU* in the segmentation base because it represents the *RFM-A*, or *Value* function;
- we include *p_t1* in the segment base, since this represents the *Event likelihood* require for the Risk management-driven allocation of retention campaign resources. Since we have not found this approach in the retention management literature we have reviewed, we consider this an innovation in the choice of a segmentation base.

Applying k-means clustering for the segmentation - instead of Telco ABC's current OLAP segmentation techniques – we obtain non-prejudiced, a-priori segmentation results, albeit relevant to management and strategy. This technique better represents the natural structure within the data (SAS Institute online e, p.1) (Mazanec and Strasser 2000, p.17). This allows management to develop a less prejudiced understanding about their retention segments, leading to the design of more relevant campaign offers.

A further benefit to Telco ABC of using k-means, is that it enables the segmentation along all the dimensions of the chosen base *simultaneously*, in a way that gives a standardised *comparative* profile between the clusters. The comparison enables the better relative distinction during the campaign offer design, and better *prioritisation* of resources during execution of the campaigns; neither of which is possible using conventional analytical approaches for segmenting.

We add six generic criteria for successful market segmentation from (Wedel and Kamakura 2000, p.4) and other sources. With this we conclude about the new domain knowledge, which is relevant to the executibility of segmentation:

- distinct – the segments should be characteristically distinct or separable by measures about their base variables;

- substantial – be substantial enough to provide a net profit after campaign cost per individual;
- accessible – they should be within reach promotionally and logistically (Meltzer 2000, p.13);
- differentially responsive or homogenous (Simoff 2003, p.4) – there should be an expectation of a response to a campaign that was tailored for that segment;
- Strategic fit – the segments should fit with the Corporate Strategic Framework. (Interpretable and usable (Simoff 2003, p.11));
- Stability – at least for the completion of one identify-and-respond cycle. The commercial importance of the dynamic aspects of segments are mentioned in the literature (Mattison 1999, p.201), and this becomes one of the concepts we will monitor technically for drift.

The debate about determining the ideal cluster size commercially has not been resolved. Commercially you want your clusters distinct enough to warrant the development of a unique strategy, and big enough to warrant the special attention (Wedel and Kamakura 2000, pp.60, 91).

2.10.5 Concluding about new domain knowledge

From the above it becomes apparent that the solution of the problem, has been brought closer by the introduction of new subject matter into the problem. This new knowledge has laid the foundation for better understanding the extent and nature of the Telco ABC's business problem, and what will constitute the executibility of a new solution, even before we have done an hour of mining. We demonstrate in later chapters, how this newly introduced knowledge, helps with the defining of the project goals, and of the data mining plan.

In TQM terminology, and in a paradigm lock situation, the new knowledge helps with defining the new *purpose* within the *utility for purpose concept* for the SPC project. We will demonstrate how a useful project methodology, directs the data mining toward discovering information relevant to the new solution, and how such a methodology should develop that information into related knowledge For our purposes, we are less concerned with how Telco ABC developed their *desired* and *possible* solutions to reach their *Strategic choice,* and more concerned with the *content* of the *Strategic choice*

entity. In the SPC application of the SPM, the *possible* is generally accepted as a quadrant of generic strategies. This approach is particularly suitable when the profiling was qualitative. These four *possibilities* are *Turnaround*, *Expand*, *Defend*, and *Diversify,* and they are presented as the four blue rectangles in Figure 2.3: Telco ABC's Strategic choice., which constitutes an executable solution.

2.11 CRISP-DM data mining standard for Business Intelligence application

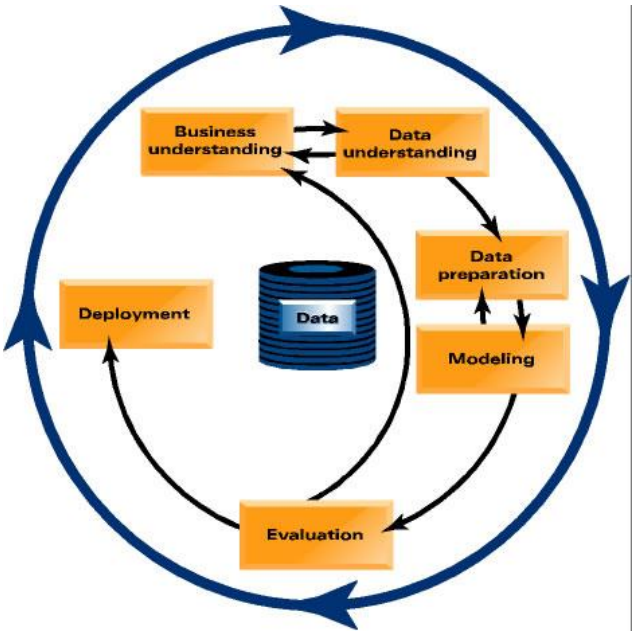


Figure 2.6: CRISP-DM

About 5 years ago from the time of writing, two data mining methodologies were created. They are CRISP-DM (Chapman, Clinton et al. 1999-2000) and SAS Data Mining Projects Methodology (SAS Institute 2000). The purpose of these methodologies is providing project management utility (Pyle 1999, p.10) (SAS Institute 2000, p.xi) (Chapman, Clinton et al. 1999-2000, p.3). In this section, we offer a brief introduction to CRISP-DM. We visually present CRISP-DM in Figure 2.6. The process flow of the methodology is clock-wise, and the counter-clock arrows indicate the iteration.

2.11.1 Business understanding

In this step, CRISP-DM defines the business problem and the project objectives about the business problem (Chapman, Clinton et al. 1999-2000, pp.13-19 and 35-41). From

this, a data mining problem is crafted, and a preliminary project plan which contains the technical data mining objectives.

2.11.2 Data understanding

The first step within *Data understanding* is identifying and collecting the data that is suspected of meeting the project's objectives (Chapman, Clinton et al. 1999-2000, pp.20-22, 43-46). In most commercial applications, the goal is to obtain as rich a data picture as possible about the business problem, and such data is usually resident in a number of disparate databases that reflect business and operational processes. At this stage, there is uncertainty about where specifically among all this data to explore for the formulation of a hypothesis about association in the data with the target. The strategy for overcoming this uncertainty, is to either consult the metadata to identify the data's origins, nature, operational intent, and commercial meaning and application; or if no metadata is available, then it has to be generated.

Next follows the exploration of that data under guidance of the metadata, for discovering associations between individual or aggregated features of the data, with the data mining target. This is exploring for 'knowledge you know you don't know' (Westphal and Blaxton 1998, p.65) - meaning the exploration is driven by suspicion and domain experience - and that information will be discovered using the right tools correctly. This results in the formulation of a hypothesis about which data the targeted event may be reflected in, and potential relationship between the data and the target. Once the hypothetically relevant data has been identified, its quality is verified for completeness, errors, missing values etc.

2.11.3 Data preparation

In *Data preparation* the data is prepared for further hypothesis formulating exploration about the target, and the features that may support the discovery of knowledge about that target (Chapman, Clinton et al. 1999-2000, pp.23-25, 48-52). Interesting data subsets are identified, and further investigation strengthens the hypotheses about hidden knowledge, resulting in the selection of data sets. The selected data from the different domains are then integrated into one comprehensive view in order to access it with the data mining tool.

There follows the final transformations of the data, addressing issues identified in previous phases; sanitizing by replacing missing values etc., constructing derived features from the data to lift out dimensions or context, and formatting the data to the requirements of the data mining tool which will be used for knowledge discovery. A round of feature elimination can be included, where features that carry the same signal about the target, are omitted or grouped, and those with an insufficient association with the target, dropped. A signal within data can be defined as the predictable relationship between the variation in the value of the input features, and the variation of the target feature(s) (SAS Institute online b).

2.11.4 Modeling

There follows a stage of experimenting with the purpose of selecting the best-performing, relevant modeling technique for the identified data mining problem (Chapman, Clinton et al. 1999-2000, pp.27-29, 53-58). So for instance, if it is a classification problem, then experimentation could be with differently configured decision trees, neural nets, and perhaps a logistics regression algorithm. The different models are assessed on certain criteria (e.g. accuracy, information preserved, specificity, lift) which test their ability to generalise - practically this test is their performance against data not seen before - where the best model is the one that best generalises. Some software even allows the ensembling of various models, which theoretically should improve the generalisation. The requirement of training, evaluating and testing the model against unseen data, means the application of various sampling and data partition techniques too (SAS Institute 1998).

2.11.5 Evaluation

Models are then compared on business criteria, determining which one(s) best meet(s) the business objectives, and are most understandable commercially (Chapman, Clinton et al. 1999-2000, pp.30-31, 57-59). There could also be some evaluation of the results within a test environment - if that luxury is available - before a final model selection is made. There should also be a re-assessment of the potential return on the data mining dollar invested in the identified problem, compared to the potential return on investment on new problems that may arisen in the meantime, or even investment some more interesting aspect of the existing problem which exploration to date may have unearthed.

2.11.6 Deployment

This is where the plan is developed for the implementation of the data mining program and its results into the business, e.g. of the logistics of the data mining, including the communication of the outputs to the business for designing the business solution. Plans are developed for the monitoring of the ongoing technical relevance of the outputs, and maintenance of the model when required. (Chapman, Clinton et al. 1999-2000, pp.32-33, 60-62).

2.12 Chapter summary

In this chapter we established that in order for information to be useful in the SPC environment, it needs to be developed into competitive, executable knowledge. We established that the process of knowledge development had three phases, and contributed our own, fourth phase for purposes of evaluating CRISP-DM and developing SAM. We learned that an effective way of introducing new subject matter knowledge into an organization, is importing it from outside. This applies to both business and analytics knowledge.

We further learned how that the human mind is central to the effective use of data mining for information discovery, through recognizing the relevance of the information, and for developing that information into knowledge. We established that we can improve the effectiveness of information discovery and knowledge development, by enabling the human mind with new subject matter expertise. We then contributed that injecting new subject matter expertise into the SPC data mining project environment, is of paramount importance for producing organizational breakthrough or paradigm shift.

We saw in an earlier chapter that the SPC environment is more unstructured than the automated data mining environment, bringing with it uncertainty about what constitutes breakthrough and executibility, and also about the organisation's operating circumstances. In this chapter we then introduced a well-established planning tool for producing breakthrough and executibility in the uncertain SPC environment. That tool is the Strategic Planning Model. We contributed the view that the success of SPM is partly due to it being a practical knowledge management process. We successfully argued that this feature makes SPM an ideal basis for a data mining project methodology aimed at the SPC environment. We demonstrated how other breakthrough technologies – like ERP and CRM – had been using SPM for this purpose since the 90's decade.

One of the outputs of SPM is an organisational strategic framework, consisting of objectives and strategies. We explained how that an organisation's existing strategic framework, constitutes its existing paradigm about a problem and its solution. We then successful argued that paradigm shift can be described in terms of changes to the corporate objectives and strategies hierarchies. We also contributed a pre-project schema, with which to express an organisation's unchallenged and untested preconceptions, assumptions and expectations, at the outset of SPC projects.

We also argued successfully for introducing into the SPC data mining environment, key concepts from Total Quality Management and Risk Management, and linked them to the SPC project environment. TQM is a design approach which assures the suitability of an outcome for its designated purpose. Risk Management is a proven approach to managing risk under conditions of uncertainty.

We defined the existing paradigm of Telco ABC about their retention problem and its existing solution, basing it on existing objectives and strategies. We successfully defined what constituted new marketing and data mining subject matter, which would bring about competitiveness for Telco ABC, and would also greatly enhance the executibility of the project results. The competitiveness was specifically linked to behavioral and value segmentation. We also contributed a novel data mining algorithm for the retention marketing environment which supports the classical marketing Segment-Target-Position approach.

In the following Chapter 3 we introduce from the technical data mining literature, key principles, practice and process of *concept drift*. Concept drift is used in the automated data mining environment for monitoring problems which change over time, and for automating the control of their solutions.

3 Chapter 3 - Concept drift detection methodology in data mining

In this chapter, we present a study of the concept drift literature. We aim to:

- learn what constitutes a data mining concept;
- understand how a concept changes over time;
- understand the implications of concept drift in a changing environment;
- learn about key principles and practice for detecting and responding to concept drift, and its process;
- determine if concept drift is relevant to the SPC application of data mining; and if it is relevant;
- determine how we could integrate it into the SPC data mining environment.

3.1 Context of concept drift

In the concept drift literature, *concept* generally relates to the *measures about relevant information about a data mining target*, which was discovered using data mining. *Relevance* relates to the utility of the information, in developing an automatically executable solution for a problem (Widmer and Kubat 1993). *Concept* is therefore used within the context of a solution for a problem associated with the data mining target, and we know that *solution* has a strong notion about *attainability* or *executibility*. The relevance of the discovered information is assured through formulating a technical data mining *target*, which relates to a well-defined real-life problem.

The coining of the phrase *concept drift* is attributed (Widmer and Kubat 1996) to Schlimmer and Granger publishing their STAGGER algorithm 1986. Its use there was to describe changes within discovered information about a data mining target. The study of *concept drift* then is about methodologies for tracking changes to that *concept* or *information*, to assure the *ongoing relevance* of the solution.

Helmhold (1994, p.3) offers the following technical description of a concept: A tracking problem consists of a set (or domain) X and a family H of {0,1}-valued functions defined on X, called the target class. A {0,1} valued function defined on X is called a concept. Klinkenberg (2001) gives a probabilistic notation for concept drift as the

difference between batches of the sample distribution of P(x,y); specifically the difference in sample distribution between $P_i(x,y)$ and $P_{i+1}(x,y)$ where $_i$ = the batch number.

What was interesting in our literature study of the SPC data mining literature, was that we did not come across one instance of similar use of the terms *concept* or *concept drift* (Han and Kamber 2001, p.179). Further, we were hard pressed to find any terminology in the SPC literature, which is used to express the same phenomenon. Not even the monitoring sections of the two data mining methodologies we reviewed, changed our view on the SPC literature (Chapman, Clinton et al. 1999-2000, Deployment section) (SAS Institute 2000, Implement in Production and Review sections).

Classical statistics postulate that data are measurements about events in the real world, and that these measurements are reflected in the data population features (Pyle 1999, p.2) (Levin 1987). Changes in the real world will reflect as changed values in the data population's features. This idea is reflected in the data mining literature on detecting concept drift where it is described as data 'population drift' (Harries and Horn 1996) (Klinkerberg 2001) (Kelly, Hand et al. 1999). Harries and Horn (1995, p.3) link to the classical approach, saying that *concept drift* is caused by changes in the underlying, real world, which reflect in the data over time.

Some authors (Widmer and Kubat 1993) express this as a distinction between *real* concept drift, and *virtual* concept drift. *Real* drift is what can be observed in the physical world, and *virtual* drift is what is observable in data about the real world. An example of this distinction in the types of *drift* can be found in a change in fashion. If there is a change in fashion from wearing pink to wearing red, then in the real world we can observe *real* concept drift as the change in the way people dress. If there is data about this fashion change, then we can *virtually* observe the fashion change as chances in the data values about color.

Batched data mining is dependent on sampling practices, and any changes in the data population will be reflected in samples that are taken for data mining. In the event of online processing, the drift is represented in the subtly changing data values in the incoming streams of data over time (Westphal and Blaxton 1998). In either batch or online mode, the change will then manifest as a drifting concept or target. Klapper-

Rybicka et al. (2001) describe this as the *temporally extension of features* (my ordering of their words) with reference to changing data values over time.

Some authors use terminology about *context*, as a cause for concept drift (Michalski 1987) (Widmer and Kubat 1993) (Widmer and Kubat 1996). In the real world, a concept is dependent on context, and a change in context => change in concept => concept drift. Harries and Sammut (1998) also, uses contextual terminology to describe the cause for drift, when they refer to 'potentially unstable underlying phenomena' which is hidden in the data as hidden context, but which can be discovered over time lapsed.

These changes in the distribution statistics of populations affect the probabilities of class membership (Kelly, Hand et al. 1999, p.368). Kelly et al. offer a *probabilistic description* of the cause of concept drift:

- the class priors: $p(c_i)$, i = 1,2…n where n is the target class size;
- the conditional distributions of the classes may alter over time: $p(x|c_i)$. Changes here tend to affect the error rate of the prediction or the identification of the concept;
- the posterior distribution of the classes may change: $p(c_i|x)$, affecting the relevance of the model.

Changes in the data affect the target and the concepts developed about it, through any or all of these three probability dimensions. If the real world event changes that affect target concept are sufficiently represented in the data, then their influence on the target concept is through $p(c_i)$ and $p(x|c_i)$. However, if the real world event changes that affect the target concept are not sufficiently represented in the data, then the target is differentially affected apart from $p(c_i)$ or $p(x|i)$. Referring to the previous example of the two fashion colors; if the reason red became the new fashion, was because red garments became more *affordable* than pink ones, and there is nothing in your data set to reflect this *cost* dimension, then the influence on the target is through $p(c_i|x)$.

Automating the discovery of a concept, and keeping track of its drift, requires the presence of the right conditions for discovery e.g. (Helmhold and Long 1994). The first condition is that the change is detectible by the algorithm – in other words, that there is some component of the change that can be measured. The second is that the rate of

change is gradual or constant (Long 1999, p.1) – *...that consecutive pairs of distributions have small total variation distance.* – and therefore falls within the limits or boundaries of what the algorithm can detect. Third, that the results of the measurements are *comparable over time*. The basic elements of concept drift therefore are *time lapsed* (Jacobs and Blockeel 2002) and *change in measured value* (Kohane and Haimowitz 1993, p.1).

Detecting concept drift becomes more difficult with increased complexity of the hypothesis (Helmhold and Long 1994), or with an increase in the quantity of principal features that influence the concept. A second significant problem is a low prior probability of the target $p(c_i)$, where if it is so low that it borders on the detectible, the algorithm may struggle with learning it. The last is big problem for online analysis. Thirdly, the level to which external events are actually featured in your data (the $p(c_i \mid x)$ above) affect the success of detecting the concept drift.

3.1.1 Practical applications of concept drift

There is variety of applications of concept drift technology, and the most popular is the automatic classification of text (Lanquillon 1999) (Widyantoro, Ioerger et al. 1999) – also known as information filtering. It is well established in the on-line domain of e-Commerce, where for instance the classification of documents or books needs to be updated to reflect the changing tastes of customers.

Another application in e-Commerce is the monitoring of the browsing activity of users of an online service, where the operator of the website is interested in identifying changing trends in user browsing (Hulten, Spencer et al. 2001) (Jacobs and Blockeel 2002).

One of the most established applications is in trend spotting of buyer behavior in the retail business, where programs are run over data to identify deviations from established buyer patterns (Agrawal and Psaila 1995). There are also applications in the computer hardware monitoring environment, where for instance the ISP's apply automated monitoring and re-distribution of workloads over computer networks.

3.1.2 The implications of a drifting concept

The problem and its solution are in the domain of the real world. We saw in the preceding section, that a drifting concept means that events in the real world have

changed. This means that the problem has changed too. We saw in an earlier section that the ongoing relevance of a solution is of paramount importance. The implication of a drifting concept therefore, means that the solution needs to be updated to remain relevant for the changing problem.

If concept drift is not detected, or not responded too, over time it will result in a reduction of the effectiveness of the solution. For instance, an online bookseller may rely on buyer feedback for the rating of books, in order to make recommendations about these books to future customers. Such an online book vendor may rely on an automated data mining algorithm, to process the buyer feedback and to make book recommendations based on that. Buyer tastes change over, and new books on the same topic come on the market. This means that over time the ratings of a book by the online seller's customers will change. If the algorithm has no mechanism for detecting this change, the relevance of the recommendation of the books will deteriorate over time (Stanley 2000).

In order to overcome this negative effect of drifting concepts, engineers in the automated application environment of data mining, over time developed methodologies for detecting concept drift.

3.2 Concept drift detection methodology

In this part, we present some of the main research about concept drift methodology. We present the literature study chronologically, and do not consider it as exhaustive. Despite this, the literature study was very influential on my thinking about data mining project methodology. The review will focus on the components of a concept detection framework as they emerge from the literature study. At that point, we will generically summarize such a detection framework, with the purpose of learning from it.

The detection of concept drift is dependent on an algorithm of choice depending on the nature of the data or the problem. In the same way, the algorithm we develop will be determined by what is suitable to our industrial research environment.

We make the distinction between the batch and on-line approaches to detecting concept drift in the literature:

- *batch processing* is examining the data in a static state at set points in time – either as a whole population, or as samples that were taken at points in time. The methodology can detect concept drift between batches;
- in *incremental or online processing,* we continually examine the data in a stream without sampling. Concept drift methodology can also detect such online drift over time. This approach is also referred to as *active* (Agrawal and Psaila 1995).

In a later section, we will discuss some of the data, which was used in the experiments in the literature.

According to Widmer and Kubat (1996), the earliest published study in concept drift methodology was that of **Schlimmer and Granger (1986)**, resulting in an algorithm called STAGGER. Its mechanism for judging when to discard old information, and therefore learn a new one – the heuristic - was based on Bayesian measures of logic (logical sufficiency and logical necessity). An existing concept had decayed once it fell below a pre-set measurement threshold. STAGGER was effective, but was prone to over-learn as time lapsed. A second limitation was that it could not recognize recurring concepts because of a lack of memory. Although STAGGER was a breakthrough, the over-learning and lack of memory meant it had limited practical use.

In the same reference, credit is given to **Kuh, Petsche et al. (1991)** for identifying the *rate of drift* as a factor determining its learnability. Also referenced in this source, is **Krizakova and Kubat (1992)**, who receives credit for having developed the memory based learning heuristic. It was based on a decaying weight over time lapsed; a learning strategy of *forgetting* by first-in-first-out (FIFO).

Jacobs and Nowlan (1991) presented a methodology for learning *local concepts* within data, using neural networks. They first clustered the data by certain chosen base features, and then trained an expert network within each cluster to discover the local concept.

Salganicoff (1993) developed a weight decay ('forgetting' approach) to track the changing concept in his DARLING algorithm. DARLING measures the concept by a minimum probability value for accurately predicting a binary outcome. A minimum threshold is set, below which the concept is forgotten. The decay of the weights is not time-based, but density based on a nearest neighbor criterion; the concept persists until it can be replaced by new evidence from the same local area in the data (p.7). The big

advantage of this weighted approach over time-linked decay methods is that a concept will persist in time until it becomes irrelevant. This effectively did away with the FIFO approach. This approach could be advantageous in situations where step changes in the concept take place in an online environment, in which event time-based methods will be limited in their response by the resilience of the existing concept. Darling uses CART (Breiman, Friedman et al. 1984) and ID-3 (Quinlan 1984) as the detecting algorithms. The role of concept 'memory' as part of the learning mechanism was well established in the literature by this date.

The contribution of **Helmhold and Long** (1994), was a breakthrough in concept drift detection. The key to their success was the development of a measure for concept drift itself, and not for measures of the underlying data. The measure is P(f_t not equal f_{t+1}). t = target and f = function, therefore the measure is the probability of disagreement between subsequent target functions. Their work established that the more complex the target class hypothesis is – i.e. the bigger the quantity of independent features - the more difficult it is to track concept drift. Their algorithm assumes slow and gradual concept drift, and an upper limit above which the algorithm looses track of change.

Widmer and Kubat (1993) developed their FLORA 1-3 range of algorithms for detecting concept drift. He used artificially generated, experimental data. The data was a combination of four variables (size, color, shape, time), in what may be defined a simulated online situation. They build on the learning strategy of *forgetting*, and the algorithms use a time window adjusting heuristic for the *forgetting* rate – essentially a variable FIFO. What was insightful here, is that *truth* (my term) or *falsity* of the drift in the concept, is determined by how representative the data is of the real-world drift. The practical implication of this for data mining applications where solving a complex problem is the purpose, is that the data should capture as much subject matter context as possible.

Widmer (1994) and **Widmer and Kubat** (1996) subsequently developed FLORA 4, to overcome a limitation of FLORA 1-3 in distinguishing between random variance (noise) in the data, and actual concept drift. FLORA 4 achieves this by adding to the window adjusting heuristic, a test of statistical confidence for the conceptual hypothesis. They use the IB3 algorithm of Aha, Kibler et al. (1991) for the detection. Further, FLORA 4 recognizes the complexity of the hypothesis as a factor that affects the

predictive accuracy; the more features in the data, and the faster the rate of change in data values, the more complex the hypothesis. However, the data that was used to test FLORA 4 had 2^{10} possible hypotheses, which is relatively small compared to today's commercial data sets.

Harries and Horn (1995) used C4.5 (Quinlan 1993) to experiment in time-series prediction on financial markets data, in a high frequency batch environment. The aim of their study was to minimize the effect of noise and concept drift on predictive accuracy, in an environment where there was time limitation on retraining of the algorithm. Their work is relevant under conditions where regular model retraining is onerous due to resource constraints. However, they succeeded in demonstrated the application of concept drift methodology on commercially generated data. The financial market data they experimented on is in many respects similar to that of the telecommunications industry. It has high noise levels, is temporally ordered with no reliable frequency for the behavioral indicators, and the rate of change of the concept varies over time.

Agrawal and Psaila (1995) contributed research by detecting concept drift in an online commercial environment. They base their framework on a decision tree that mines a database for statistically interesting rules. As rules are discovered, they are added to a historic rules base with the statistics about their confidence and support. When the confidence or support of existing rules change, the rules base is updated. They presented a trigger mechanism that activates a notification to the observer when any changes of interest occur in the rules, or rule parameters, from their data. The intention is for the solution to be adapted to remain relevant. Their framework forms a useful basis for the description of concept drift in terms of *commercial interestingness*. There seems to be some discrepancy between their title and introduction - which clearly refers to online data mining - and the experimentation section, which describes experimentation in batch mode. The data they use in their experimentation, were artificially generated events, which they call 'shapes'.

Harries and Horn (1996) insight is that the 'concept' is context-based, and the context can be *visible* (*direct* as they call it) or *hidden*. Quite often changes in the world the data is modeled on, are reflected visibly in one or more data features, called the *environmental attribute(s)*. Other authors call them *indicator variables* (Ramsey and Grefenstette 1993, p.1). Now, in the event that the concept drift is *hidden*, it can be

discovered by other means than FLORA1-5's windowing approach. That methodology is unsupervised clustering of the data *en masse*, and then investigating the clusters over time lapsed, to identify the hidden drifting concepts. They call the time lapsed an *environmental attribute*, and such drifted concepts *local concepts*. We comment that in the case of massive commercial data sets, the context change quite often *is* spread over multiple features, and that their work presents an insightful approach. They developed an algorithm called SPLICE for discovering the *time attribute* required, for optimal segmentation of the data. The algorithm facilitates the tracking of changed concepts between the segments over time lapsed. The data used in this experimentation was artificially generated.

Domingos (1997) work is in the arena of feature selection for lazy learners, demonstrating the benefits of introducing context into the feature selection stage of modeling, by first clustering the data by a nearest neighbor method.

Robnik-Sikonja and Kononenko (1996) reference the work of Widmer and Kubat (1996) and Domingos (1997), offering a refinement in dealing with context. They formulate the cause for myopia in algorithms that do not recognize the influence of context, as the assumption of the independence of features – when in the real-world features may not always be independent. First clustering then captures the dependence, after which they perform regression (in their case using a tree) within the cluster to discover the local concept. There are substantial experimental design tips; when building the model, keep the parameters constant between different times to enable the comparison of the concept between times.

Freund and Mansour (1997) challenged the assumption of those approaches where the drift is small enough over time for a hypothesis to remain valid for a 'long time', as they call it. Instead, they present an algorithm, which works under conditions where the rate of change is rapid, *and* the change is the result of prior and / or posterior influences. In order to capture this probabilistic dimension, they deploy a weighted error of the hypothesis as a heuristic. There are two restrictions; first, that the rate of change is linear, and second that there is an upper bound on the error. Their experimentation is on data with few input features, and where the rate of change is linear. It leaves an awareness of the prudence of experimenting with the input features to discovery the nature of the relationship between the independent features and the dependent, before

any modeling is done, and by including transformed features in feature selection stage (Hastie, Tibshirani et al. 2001, p.127).

Chakrabarti, Sarawagi et al. (1998) researched a slightly different problem than concept drift; that of automating the discovery of *interestingness.* The relevance of their work is in introducing the use of informational bit coding as a measure of the concept. This work provides insight that the comparative measurement of concepts – as opposed to absolute measurement - is quite often sufficiently useful for commercial problem solving.

Harries and Sammut (1998) worked on the difficulty that batch learners have with detecting hidden changes in context over time. The causes for this difficulty can be twofold; the first is where the batch learner assumes that 'the training set is homogenous' – which I think was meant to read *the training sets are homogenous between batches in time*. The problem here is that the need to retrain the algorithm on new data, as it becomes available, is not recognized. This explanation is very plausible for their time, when the state of the software user interfaces, and limitation on computational power, made regular updating of a learner a very burdensome activity.

The second cause of difficulty for batch learners to detect hidden changes in context over time, is given that the changes in the context are hidden – they do not reflect in (an) indicator feature(s). This strengthened insight that the hidden context can be spread over multiple features, and that data mining is a suitable technology for investigating this. Consequently, their approach is to investigate the data *en masse* using a second-generation SPLICE. This splices the data into temporal clusters, using changing measures about a pre-set error as an heuristic (pp.7-8). Following the clustering of data they then use a tree as a context sensitive feature selection tool, learn the local concept, and use that model for prediction within the time cluster.

Computational resources now make it possible each time new data becomes available, to use unsupervised clustering to discover hidden concept in a multi-featured data environment, and to retrain the model within each cluster. In commercial situations, the clustering schedule may need to coincide with calendar-based commercial routines, like the completion of billing postings, or the updating of data warehouses. SPLICE is related to FLORA, in drawing from a history of previous instances to verify a concept,

albeit in a batch environment. The data used in this experimentation was the STAGGER dataset.

Long (1999) follows on the research of Kelly, Hand et al. (1999), who related concept drift to a change in the probability of the hypothesis. Long makes an important contribution by quantifying the rate of statistical drift that an agnostic concept drift detection algorithm can tolerate, while maintaining a selected level of accuracy. Long's reference to the term 'agnostic' in the automated data mining environment, is in opposition to the requirement for gnostic interpretation in the SPC environment, where human interpretation and judgment is required for determining if the concept has drifted sufficiently, to update a solution.

The research path from Klinkenberg through **Lanquillon** (1999), uses overlap of statistical confidences, related to changes in a moving average of standard deviation, as a measure of concept drift for a binary classification problem. The overlap is described as a 'density function', and the less the overlap (and consequently density) the more the concept has changed. Lanquillon (1999) has a section on classifying the types of concept drift (*change* as he calls it), which could be of use in a SPC application. Their types of concept drift are:

- new topics arise;
- existing topics disappear;
- existing topics change in content.

We add another category to theirs, which is important in a SPC environment, where resources have to be committed to reaching objectives for fixed periods:

- existing topics unchanged.

Widyantoro, Ioerger et al. (1999) developed a 3-descriptor model for the online information filtering domain, which learns detects concept drift based a parallel two-window approach. They use very simple model classification accuracy as the measure of drift.

The work of **Bartlett, Ben-David Shai et al.** (2000) is based on the probabilistic relationship between consecutive targets, similarly to Helmhold and Long (1994) and Long (1999). Bartlett et al. develop a probabilistic measure into a '…general condition on switching frequency…' (p.5). They also develop an algorithm, which is generic

enough for detecting concept drift with batch or online data presentation, and both gradual and arbitrary change of the target. However, they maintain the restriction of bounds on scope and time, within which the drift must fall. Their work provides a solution for detecting concept drift in commercial environments. There, the target can undergo both periods of sustained and gradual drift, or step changes due to external influence of competitor activity. E.g. in the mobile Telecommunication industry, constant numbers of customers may leave you for a competing service provider as their mobile phone plans expire; or they may leave you in a sudden surge when a competitor successfully launches a plan that attracts big numbers of your customers over a short time. We comment that there is no reference in this article to experimentation in support of the algorithm.

Klinkerberg and Joachims (2000)'s experimentation follow the school of Widmer and Kubat (1996) and Lanquillon (1999) with fixed or adaptive time window size. They present an algorithm based on Support Vector Machines, successfully calculating the ideal adaptive window size, and adjusting it to the current extent of concept drift. Their application of a leave-one-out technique to calculate the ideal window size in concept drift is novel. Leave-one-out works by discarding irrelevant information (Lunts and Brailovskiy 1967). Their algorithm gives good results where concept drift has to be detected automatically. Further benefits are:

- less manual parameterization of the algorithm to optimise window sizes; and
- allowing earlier detection of the concept drift, therefore giving more time in which to adjust for avoiding misclassification.

The success of the prediction is measured as classification error and precision / specificity - both measures are already employed in the SPC mining environment. Their algorithm adds a third measure *recall*, which is the probability of the algorithm recognizing relevance. *Recall* plays a similar role to human memory in concept drift detection in the SPC data mining environment.

In a following article **Klinkerberg** (2001) continues the experiment using the support vector machine with adjustable window size, minimising the algorithm's generalisation error on new data. Lanquillon (1999) offers a criticism of the SVM method: because SVM tries to find a good representation of the *boundary* between classes, they are not good at detecting sudden, step changes in the *content* of the classes.

A next reference that used information theoretical measures in its heuristic was **Lau, ter Hofstede et al.** (2000). They base their approach on the work of Alchourron, Gardenfors et al. (1985), who represents epistemic entrenchment (belief inertia) by an entrenchment rank – a form of weighting. The algorithm assumes, like so many now, that the change is consistent and minimal. This assumption is tenuous in the open-system commercial problem environment. The semantic relationships among information objects have to be engineered into their algorithm's memory by a subject matter expert. This is a further restrictive condition of what is supposed to be an automated mechanism in a structured problem environment. What is relevant to the SPC data mining environment however, is that subject matter expert's interpretation should form part of the heuristics for detecting concept drift, and for deciding to respond to it or not.

The work of **Alchourron, Gardenfors et al.** (1985) provides a very relevant classification for types of 'belief state transition' (= concept drift): expansion, contraction, revision, which is suitable for use in a SPC environment.

An article that stands out in my mind is **Stanley's** "Learning Concept Drift with a Committee of Decision Trees." (Stanley 2000). This is the first experimentation that moves away from the time window approach, and he does that using a committee of decision trees, the so-called Concept Drift Committee (CDC) algorithm. They have an hypothesis-based approach to concept drift. Here, a group of decision trees was all trained on data from different times of the monitored domain, each representing a different hypothesis about the concept at a different time. The trees then form a committee, which casts a weighted vote, where the weight is determined by the accuracy of that tree in predicting the target on a previous batch of data. In comparative experimentation with the FLORA family of algorithms, it consistently outperforms FLORA, both under conditions of constant or step drift. The reason they postulate, is that the Stanley approach is less influenced by variance between samples, because the decision is a weighted average of the variance over time.

Additionally, its success is not dependent on first discovering change, and adjusting the learning accordingly by adjusting the time window. Rather, the concept is updated through retraining anyhow, irrespective of whether it has changed or not. This approach is therefore sensitive at detecting the concept drift when it happens. It is suitable for

detecting both gradual and instant concept drift (compare (Widmer and Kubat 1996)). This article offered big hope for using the commercial KDD data mining infrastructure in detecting and describing concept drift without a custom-designed algorithm, without having to first know whether drift has occurred or not. In the commercial environment, we can develop relevant measures about our concept irrespective of what algorithm we use, and interpret changes in those measures between batches, as concept drift.

We comment on prominent research by **Breiman** (2001) who experiments with a similar concept as Stanley, which he calls *Random Forests*. The data set used in this experimentation consisted of 1000 observations and 1000 features, which is classified as sparse data. *Random Forests* aimed at overcoming a problem in sparse data sets. There, no single feature on its own has enough substance on which to split nodes. It starts by building a forest of trees, each tree splitting nodes within randomly selected features. The output of the trees is then combined, voting for the most popular class within input vectors. Some industry practitioners (Woods 2003a) propagate a simpler solution for learning from sparse data than *Random Forests*. That solution is a well-configured Neural Network.

Hulten, Spencer et al. (2001) use a tree-based algorithm too, which they modify to optimize detection of fast, incremental change in the concept, in an online environment. It remains a time window approach, and a fixed one at that, where the size of the time window is a function of the number of features and the depth of the tree. What makes this suitable for an online environment, is the speed with which the tree is updated. This is achieved by maintaining within the node, not only the node statistics, but also the statistics of the subsequent leaves; a leaf is updated only when its accuracy falls below a threshold at determined by comparisons done at the node level. A limitation of their algorithm is that it only works for character value targets, and will not be suitable for detecting step changes. They do their experimentation with artificially introduced drift.

One school of thought uses the preservation of Information Theoretical measures between data inputs and target outputs, with which to detect concept drift. We mentioned Leo Breiman's Random Forests approach above, (Dietterich 1998), Freund and Saphire's Adaboost (1996), and Bauer and Kohavi (1999).

We reviewed the work of **Klapper-Rybicka, Schraudolph et al.** (2001). They built a concept drift detection algorithm based on recurrent neural networks, using the neural

net's hidden layer to replicate memory with weighted decay. Used this way, the network has the ability to detect regularity in input features by maximizing the preservation of information theoretical measures of inputs through the network to the network output. In the event where data is ordered sequentially, their techniques are used to divide the data into temporally homogenous clusters that optimize the preservation of information. They base their algorithm, Binary Information Gain Optimisation (BINGO) on a single layer, sigmoid-activated-with-logistic-output-function network, where the information preservation is measured as binary bits. Using the ability of the network to detect regularities, BINGO *...seeks to identify independent dichotomies in the data*, effectively grouping data with the most signal preservation together.

Nonparametric Entropy Optimisation (NEO) inversely uses the ability of the network to group similar inputs together, by dividing the data into temporal clusters that obtain the lowest information entropy (= loss of information). In NEO entropy is linked to a density function. The sophistication of the approach is in the network not being very dependent on changes in the variance of the underlying population distributions, to detect any shift in the concept.

Klapper-Rybicka, Schraudolph et al. (2001) provide valuable insights into concept drift detection. No matter how sophisticated the approach, there remains the need for some form of memory. Additionally, where there is concept drift over time, the same level of drift can occur over time segments of different length. In the case of our proposed research, where the time segments will be fixed at monthly, the implication is to consider that we will be detecting average monthly drift, and that there may have been nuances of drift in subsets of the month, that are not detected by our batch method.

Jacobs and Blockeel (2002) used classification trees with time attributes as one of the features, to predict a sequence of targeted events. Their work creates awareness about elapsed time as a driver of hidden context within data, enabling the discovery of concept drift between timed intervals. Our problem is the inverse of what Jacobs et al. solved. In our case, we do not want to predict a sequence of events. Rather, we use the relationships between timed events – in the form of usage behavior over time – as independent input when modeling the target event. Further, in our case we combine the interaction between static features (= where the values do not change over time) with dynamic features, working with a much wider representation of the problem. The more

commercial literature refers to this as the combination of *descriptive* and *transactional* features (Westphal and Blaxton 1998). An example of a dynamic feature in our data is an averaged revenue indicator over a set time, while examples of static features are the handset type or plan the customer is on. It is this combination of static and dynamic events, which enable us to use the commercial KDD data mining infrastructure for developing our concept about the target.

The work of Dorian **Pyle** (1999) also defines feature selection in terms of Entropy Optimization. He calls feature selection the ranking of input-output channels according to information content they carry about the output. **Rendell and Ragavan (1993)** use *...average information entropy...* over all relevant attributes.

3.2.1 Summary of concept drift in the automated solution environment

The literature on concept drift detection concentrates on the technical elements, which are required for the ongoing updating of a solution, and the practices of automating the inner workings of these elements where possible. We summarise our understanding of the elements of concept drift detection in Figure 3.1:

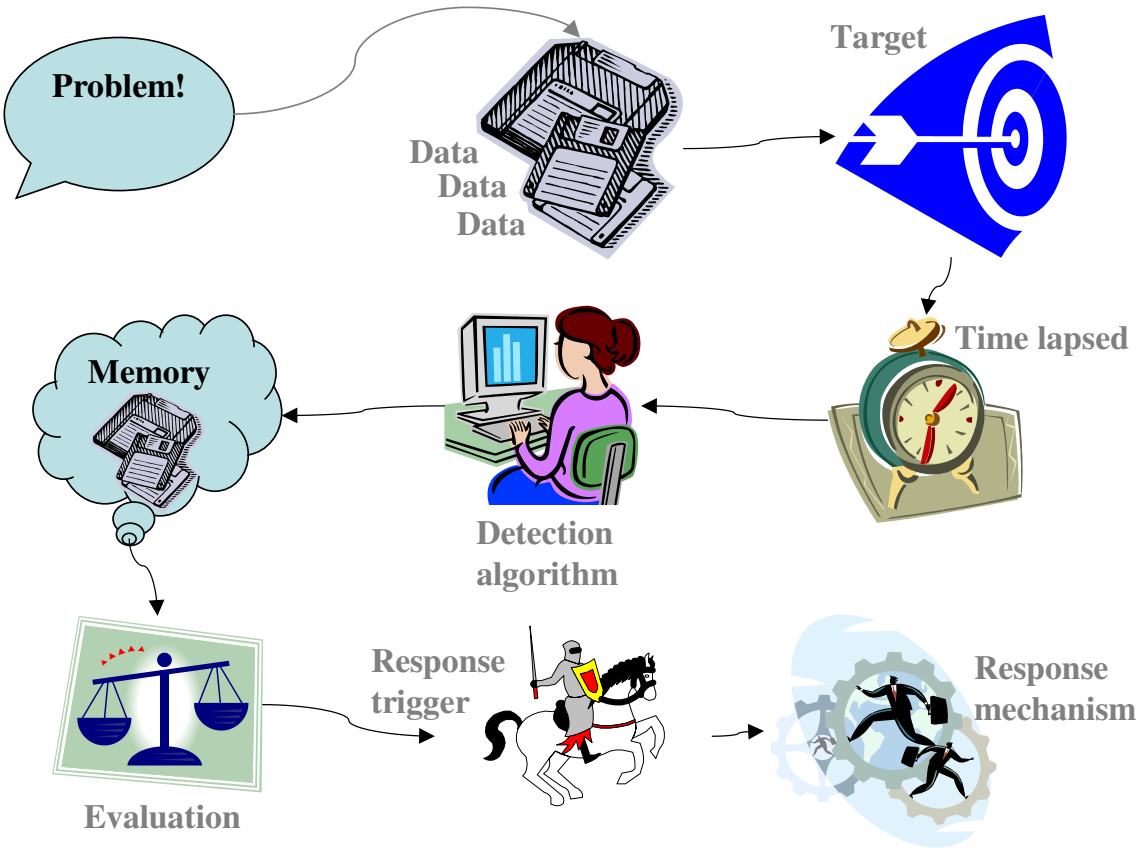


Figure 3.1: Elements of concept drift detection

- a real-life problem that is 'interesting' enough to justify resources for solving it;
- accessible data representation the problem, its context, and its solution;
- a definition of the concept upon which the ongoing relevance of the solution depends, and a representation of that concept in the data by a data mining target;
- time lapsed as an environmental factor, whether continuous or batched. In our example, time lapsed is continuous;
- an algorithm incorporating a heuristic principle, which discovers information about the target and therefore the concept;

- memory in which an established concept resides, enabling the comparison of concept measures over time lapsed;
- a response trigger which activates a response mechanism when the concept has drifted out of bounds;
- a response mechanism, which adjusts the solution for the drifted concept, keeping the solution relevant.

3.3 Making the case for using concept drift in Strategic Planning Cycle

We consider that there is a similar need in the SPC environment, for assuring the ongoing relevance of solutions. This is evident from *SPM 6*. Despite the proven effectiveness of concept drift methodology for keeping a solution relevant within the automated data mining environment, the researcher was struck by the absence in the commercial data mining literature, of reference to the concept drift literature. This silence is particularly striking within the data mining project methodologies.

Because of the above, the researcher makes the case for integrating the principles and methodologies on concept drift, into any data mining methodology which is aimed at the SPC environment. Bear in mind that the *mechanising and automating* of concept drift methodology in the SPC environment, is not a subject within the scope of this thesis.

There are a number of similarities between the two application domains of data mining, which make the case attractive:

- there is a problem worth solving;
- relevant data are available about the problem and its solution;
- concepts are formed about the problem and solution;
- concepts are represented with a data mining target;
- data mining algorithms are used as the enabling technology;
- time lapsed as an environmental factor is present;
- there is memory. In the event of the SPC environment, the memory resides in the human mind and in score code which contains mathematical models;

- there is a common need for assuring the ongoing relevance of solutions;
- there are periodic reviews of key aspects of the problem – the concepts - and an evaluation made if those have changed substantially enough, to update the solution. In the SPC environment, this review and evaluation may be cognitive. If a concept has drifted then
- key aspects of the solution are modified or updated to accommodate the change in concept. In the SPC environment, this modification or update may also be done cognitively.

Note that the similarities between the two domains include all of the *elements* of the concept drift methodology. The *similarities therefore mitigate for* the applicability of concept drift methodology in the SPC environment.

We now consider the potential effect on our case of a number of *differences* between the automated and SPC applications:

- we saw in the introductory chapter, that in the SPC data mining environment, the *problem is often more complex* than in the classical environment. We further saw that the SPC problem environment is more open than the classical problem environment. The problem complexity makes the identification of the main effects in the problem, and their degree of influence more complex. We do not however, consider this a drawback on using concept drift in the SPC environment. The reason is, that the uptake of the use of machine learning algorithms in the SPC environment, has proven that neither problem complexity nor problem boundary are obstacles to data mining in the SPC environment;
- *solutions* in the SPC environments are often *more complex*, and more fluidly defined, than in the automated environment. This makes the identification and parameterization of concepts *about the solution* more complex in the SPC environment. We find however, that solution complexity has not preventing businesses from successfully developing solutions before the arrival of data mining, and from keeping them relevant. We can overcome this issue practically in the commercial environment, by using a sequence of algorithms, each on a separate dimension of the complex concept;

- there is more *uncertainty* about the suitability of the data to the problem and solution at hand in the SPC environment. Well established data preparation techniques, however, are proving effective in overcoming this hurdle in the SPC environment. The arrival of data warehousing further ameliorates this problem, because of the process of data selection and preparation which accompanies these projects. This higher uncertainty therefore can not be considered as an obstacle to our case;
- in the automated environment, there is a high requirement on the immediate relevance of a solution. This requires mechanizing and automating that process as much as possible. It is more *difficult to mechanise and automate* the concept drift methodology in the SPC environment, because of the higher complexities described just above. However, in the SPC environment, the immediacy requirement of the solution's relevance is not as high as in the automated environment. Maintaining solution relevance in the SPC environment is more dependent on applying the principles and methodologies of concept drift, than on the level of mechanizing and automating of those principles and methodologies. This means greater difficulty in mechanization and automation are not obstacles to our argument.

We have now also established that the *differences* between the two data mining domains do not impede our case for introducing concept drift principles and methodologies into the SPC data mining domain.

The question remains as how to transfer the principles of concept drift methodology, into the SPC environment. We saw earlier that there is a knowledge management process at the core of SPM. We believe this knowledge management process offers a suitable opportunity for integrating concept drift into the SPC environment. We believe that this integration can be achieved by viewing the concept drift process as a knowledge management process. Once such a view is established, the way for integrating concept drift practices and methodologies into the SPC environment, is by embedding concept drift principles and methodologies within the relevant commercial knowledge management activities.

3.3.1 Concept drift as a Knowledge Management process

In this section we establish the view that the concept drift process can be considered a knowledge management process. We refer to Figure 3.2 in this section. The paragraph numbering in the discussion following Figure 3.2 follows the numbering of the 9 points in the figure.

1. The example problem worth solving is found in the body panel assembly line of a car manufacturer. Body panels are held in place by robots, and a robot welds them together. Fluctuations in the temperature of the welding rod tip at the welding contact point, cause the problem. The problem is, too low temperature results in insufficient bonding between the metal sheets, while too high temperature results in burn-through of the metal sheets, and therefore also insufficient bonding;
2. At this point there are three knowledge management activities, which are neither automatable nor mechanisable with current technology:
    2.1. a production engineer *hypothesizes* that what is causing the problem of insufficient bonding on the production line, is either too low or too high temperature at the point of contact. The engineer knows that if his hypothesis about the cause were correct, then the solution lies in controlling the power voltage between the rod and the sheet metal. He *hypothesizes* that part of the solution is to control the temperature range in real time, and that this can be automated;
    2.2. based on this hypothesis, he *identifies hypothetically relevant data* about the problem and its solution. In our example, this is a feature called *Temperature* and a number of other features. *Temperature* contains temperature values at the contact point at one second intervals. The other features are a time stamp, and features from which they can derive whether bonding took place or not at the time of a temperature measurement; (continued after Figure 3.2);

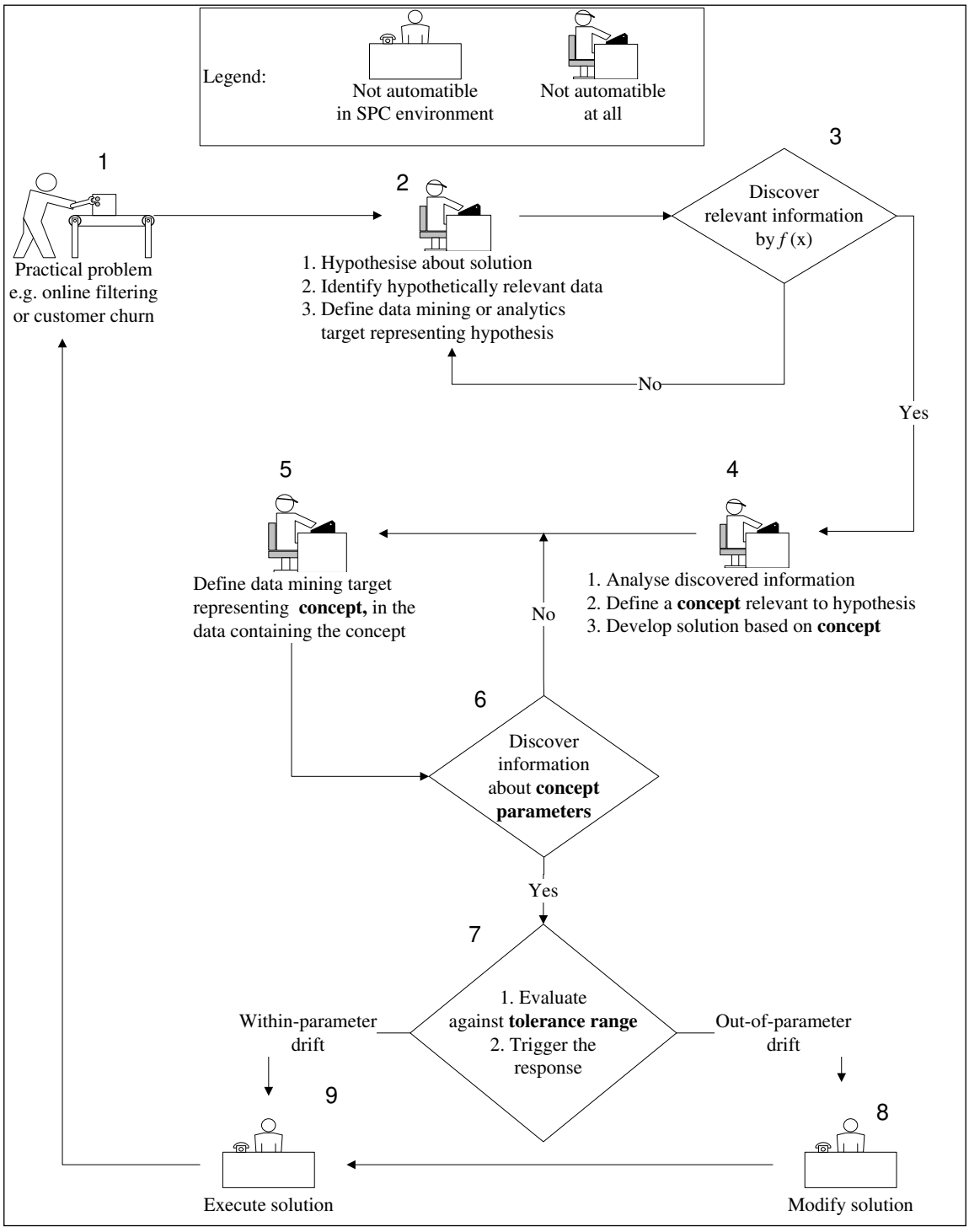


Figure 3.2: Concept drift as an KM tool

2.3. through expert collaboration with their data mining engineer, they *define a data mining target* relevant to the hypothesis. In our example, the target is a label they create, which contains binary values for *bonding* and *non-bonding*. The data mining engineer derives the values of the target feature, from these other features plus the time stamp;

3. using an automated learning algorithm with a suitable probabilistic learning function, *they discover relevant information* about the bonding target event. The information is the effect score of the different values of *Temperature*. This task is fully automatable in both environments. If no relevant information is discovered, then there is iteration back to the previous activities;

4. the production and data mining engineers complete non-automatable and non-mechanisable activities:

    4.1. they *analyse the discovered information*, making inference about the cause of the problem. In our example, the effect score from the Logistic Regression confirms the engineer's hypothesis, that temperature is the single biggest effect in the problem. Their ANOVA of the temperature values for times of bonding and non-bonding, discloses the temperature ranges, within which the bonding occurs;

    4.2. they define a *concept* relevant to the hypothesis. First, the concept *consists of* (a) *data feature*(s) in the relevant data. In our example, it is the data feature *Temperature* that underlies the effect. Second, the concept *consists of an acceptable parameter range* for measured values, within which the problem remains solved. The measures may be statistical, informational, probabilistic etc. In our example, the parameters of the solution is *1000– 1050 Centigrade*. This is the range their ANOVA showed, there always was bonding. The complete concept they identify for their solution then is *Temperature between 1000 – 1050 Centigrade*;

    4.3. they develop a *solution* for the problem based on the concept. The engineer develops a solution to control the temperature at the contact point within that specified range. This is achieved through automating the control of the electrical voltage through the welding rod and the metal. The engineer knows by how much voltage to increase or reduce the energy, for each Centigrade fluctuation in temperature. The data mining engineer *mechanises the memory* and *trigger mechanism*, through programming the range of tolerance and the Centigrade-to-volts conversion formula, and residing them in the random access *memory* (RAM) of a PC. The adjustment of voltage is accomplished through a *response mechanism*, which in our example is a Programmed Logic Control

unit (PLC). The PCL *automatically* responds to an incoming signal, which tells it by how much to adjust the voltage. The response is triggered by an *automated trigger mechanism*, which calculates the required response and sends the signal to the PLC;

5. another activity follows which is neither automatable nor mechanisable. The data mining engineer defines a *data mining target, which* represents the concept, in the data containing the concept. In our example the target is detecting the incoming value stream of the feature *Temperature*;

6. the data mining engineer selects an algorithm, which is suitable for monitoring such an incoming stream of interval data, and *automates the mining of* that data stream. In our case, the mining is done in the online mode. The automation is implemented on the RAM of the same PC used above;

7. further mechanisable and automatable tasks follow:

   7.1. the data mining engineer further modifies that algorithm continually to *evaluate* whether the values of the incoming data value stream, are falling inside the concept's tolerance parameters, or outside. This is all *mechanized and automated*, happening in the same PC's RAM. The algorithm compares the values of the data stream to the preprogrammed range of tolerance, which resides in the RAM. The data mining engineer further modifies the algorithm to send any temperature values which fall outside the range of tolerance for more than three seconds, to the trigger program;

   7.2. this trigger program is the mechanical and *automatic trigger mechanism*. Using the incoming data values and the preprogrammed Centigrade-to-voltage conversion, the trigger mechanism calculates the required voltage adjustment. The trigger mechanism will then automatically send a trigger message to the PLC for the amount of upward or downward voltage adjustment to make;

8. *an automatic adjustment to the solution,* keeping it relevant. This task is not automatable in the SPC environment with current technology. The PLC receives the signal containing the instructions for an adjustment, and accordingly adjusts the voltage;

9. in our example, as long as the discovered temperature range falls within the tolerance parameters, the PLC will control the voltage to a pre-set optimal default. This is an automatic execution of the ongoing solution. This task is not automatable in the SPC environment.

We have now established the view that the concept drift process in the classical data mining environment is an integrated knowledge management process. We depend on this view for integrating the principles and practices of concept drift in our new methodology SAM.

3.4 Chapter summary

In this chapter we identified a *concept* as the components of a real-life problem and its solution, which we need to maintain for relevance over time. We further saw that these *concepts* are reflected in relevant data about the problem and solution. As these concepts change in real life, they are reflected in the data as *concept drift*. The concept drift manifests itself either as changed values, or event probabilities. We established how in automated environments, data mining is successfully used to detect concept drift and to respond to it. We introduced the process, key principles and practices for this automation as described in the technical data mining literature.

We established in a previous chapter that in the SPC environment, there also is a need for:

- monitoring the understanding about a business problem; and
- monitoring and controlling its solution for effectiveness and efficiency over time lapsed.

We successfully argued that there was sufficient similarity between the uses of data mining in the automated and SCP environments, to draw on the principles, practice and process of *concept drift* for monitoring and controlling in the SPC environment. We successfully argued that *concept drift* be viewed as a technical version of the corporate knowledge management process. This opened the way for contributing, that key principles, practice, and process of *concept drift*, can be introduced into the SPC project environment, through amalgamating them with the knowledge management process within SPM. Since we have already argued that SPM should be a core component of our

reframing of data mining methodology, we have prepared the way for integrating *concept drift* methodology into our reframing of data mining project methodology.

We further supported the view, that to be of optimum value to the organisation, the outputs of monitoring and control have to be integrated back into the corporate knowledge management process. This applies to both business and technical outputs.

In the following Chapter 4 we evaluate CRISP-DM against the formulated utility criteria.

4 CHAPTER 4 – Evaluation of CRISP-DM

CRISP-DM presents itself as data mining process standard ... *sufficiently mature to be adopted as a key part of their business processes*... (Chapman, Clinton et al. 1999-2000, p.3). In this chapter we reflect on the utility of CRISP-DM as a data mining project methodology in the SPC environment. The evaluation is particularly biased toward contribution in solving the business problem (Grossman, Hornick et al. 2003, p.458). The knowledge for the evaluation, and the industrial setting, were presented in the first three chapters of this thesis. Though an evaluation of SDMPM falls outside the scope of this thesis, we mention that with the exception of the distinction between information and knowledge, SDMPM fared no better than CRISP-DM in a separate, previous evaluation (Van Rooyen 2004)

We divided the *utility criterion* for the evaluation of CRISP-DM into six components. They are:

1. diagnostic technique for defining the project's business deliverables;
2. mapping technique between the business deliverables and the data mining plan;
3. introducing new business subject matter expertise into a stale problem and solution environment for enabling competitive breakthrough;
4. knowledge developing activities;
5. developing a monitor and control plan; and
6. consideration about important soft issues.

In the next sections of this chapter, we first present the findings of our *reflection-in-*action on CRISP-DM about these six criteria. Then we *reflect* about how to remedy any shortfalls in utility. In the following chapter we consider these remedies in designing SAM. In chapters following the design of SAM, we will verify the remedies through *move testing* on Telco ABC's retention problem.

When we do use the term *discovering knowledge* - or similar - in this chapter, it is in conformance with its use in CRISP-DM. The reader should consider our earlier point, that strictly speaking it is information which is discovered, not knowledge. In this section, all page references, which are not accompanied by the author and date, are to

CRISP-DM. For example p.42 is a reference to (Chapman, Clinton et al. 1999-2000, p.42).

4.1 Diagnostic Technique for Defining Business Deliverables

Considering the principles of TQM, even a project should have *utility for purpose.* In the event of an SPC project, the *purpose* of the project is producing a competitive solution. That purpose is expressed in the project's *business goals* or *business deliverables*. The project methodology is the tool for defining and achieving the project's *utility for purpose* (Liu 2003, p.436). We gave an example in an earlier chapter, of how other business technologies depend on their project methodology for defining the project's business deliverables at the outset. The importance of defining the business deliverables at the outset, becomes even more evident when we consider that in complex SPC situations, there can be multiple business deliverables. This often results in a chain of supporting models (Wedel and Kamakura 2000, p.245).

In this section, we evaluate the *utility* of CRISP-DM in formulating the business purpose with the project – the business deliverables or business goals. The evaluation considers diagnostic technique or tools within CRISP-DM, which will result in:

- understanding the organisation's business problem and pre-project schema;
- analysing, interpreting and developing the schema's or business problem's components into competitive goals;
- converting those goals into project goals (Chatfield and Johnson 2000, p.2) (Westphal and Blaxton 1998, Chapter 3) (Pyle 1999, pp.12-21).

In order to evaluate against this criterion, we first introduce Telco ABC's five-stage *pre-project schema*:

1. Telco ABC knew about the existence of the problem. They had certainty that the extent of the problem had been overstated, but uncertainty about its true extent in terms of the quantity of voluntary churners every month. The *expectation* for this stage then, was determining the true extent of the business problem in quantities of actual churners. The *strategies* which had been formulated for achieving this *expectation* were:

    1.1. through interdepartmental discussion *...sorting out for once and for all the definition of a voluntary churner...* within the organisation; and

1.2. through:

1.2.1. improving the understanding about business processes in the data during the data selection stage for data mining; and

1.2.2. recreating the software script which defined actual past churners by the new definition, determining the true extent of the problem;

This would give a best estimate of the ongoing extent of the problem;

2. there was anecdotal evidence that the single biggest cause for voluntary churn was unhappiness with the existing handset. The cause for unhappiness was anecdotally linked to handset models, which were known for reception or reliability problems. The organisation's project leader has a cautious approach to making inference about root cause from any model, which was not part of a classical experimental design. Because of this, *no expectation was set on the project for learning root cause from the model's effects*, *and for forming any hypothesis* about understanding the problem's root cause. The Retention Manager however, had an *expectation* that the model's effects would confirm the anecdote about root cause;

3. there was awareness in the organisation that the current retention management solution was ineffective in addressing the problem of the cause, and awareness that data mining could be used for supporting the development of an effective solution. However, this awareness was limited to using the technology only for predicting who potential churners would be during a coming time window. Their hypothesis about a solution was limited to having more lead-time in which to execute the existing retention management solution, once the identity of the potential churners was known. Their awareness did not include innovations in technical market segmentation, supported by data mining techniques, to overcome the deficiencies of the existing segmentation approach. They consequently had *expectations* on the project for improving their ineffective solution to the retention problem. There was *no expectation* about forming any new hypothesis about the solution from new domain knowledge or further data mining. This situation is described in the Knowledge Management literature as low prior knowledge about best-practice (Cohen and Levinthal 1990; Gupta and Vijay 2000; Chen 2004). We describe this situation as one of paradigm lock, determined by the use of traditional analytics technology and stale subject matter expertise. Because of the above,

4. *no expectations* were formulated on the project for developing a competitively new solution;

5. there was an *expectation* that the data mining could support the execution of the existing retention solution. The *strategies* for realising this *expectation* were:

   5.1. using a neural network for learning the profile of actual past churners;

   5.2. scoring the customer data base monthly to identify the potential churners; and

   5.3. analysing those potential churners with traditional analytics, which were directed at executing the existing retention solutions presented in Chapter two; and

   5.4. monitoring the response of the problem to the solution using traditional analytics.

   The *expectation* was that the predictive approach would allow more time for executing the traditional segmentation analytics, and therefore for execution of the existing solution.

It should be apparent that our industrial setting formed a good setting for evaluating the utility of CRISP-DM's in defining the business deliverables.

4.1.1 Analysis of CRISP-DM for diagnostic technique

We researched CRISP-DM for a diagnostic technique we could use for formulating the above into competitive business deliverables.

First, CRISP-DM recognises the need for the *project* to produce results which best support the business's goals, and has a section with activities for achieving this (Chapman, Clinton et al. 1999-2000, Business understanding section). This *Business understanding* task set contains a subtask called *Determine data mining goals.* Here, CRISP-DM links the definition of the data mining project to the *business understanding* of a business problem, which is determined by the *business objectives/goals* e.g. (pp. 17, 18, 40), or even *business questions* (p.40), or the *business success criteria* e.g. (pp.17, 36, 69). Second, CRISP-DM recognises that the output from the project needs to be evaluated against these. There is a whole section for *Evaluation* dedicated to executing this recognition.

We needed more detail about what these business objectives / goals, questions, or success criteria refer to, and if there is any interpretation, analysis, or development associated with them. We researched CRISP-DM for guidance. The results are:

- on (p.29) we find the wording *Whereas the data mining engineer judges the success of the application of modeling and discovery techniques more technically, he contacts business analysts and domain experts later in order to discuss the data mining results in the business context*. Unfortunately *...business context...* is not explained. Further, the wording does not describe the defining of business goals at the outset of the project, referring to a *post hoc* project activity;
- on (p.60) the monitoring of the data mining results are against *…the results or its benefits*. The substance of benefits is not explained. Further, we know from SPM that monitoring is the last activity of a project, and therefore of no value for formulating project goals;
- (p.64) there is a statement *The data mining goals state the results in the project that enable the achievement of the business objectives*. Unfortunately how the business objectives are formulated in the first instance is not elaborated upon;
- on (p.72) a reference of interest was *Usually, the data mining project involves a combination of different problem types, which together solve the business problem.*' Upon consideration, this reference does not improve our understanding about the content of *business problem*, nor how it was formulated in the first instance;
- in the *Deployment* section, CRISP-DM distinguishes between the use of data mining for supporting business decision-making, and the use of data mining as a tool for executing the business solution. However, the *Deployment* section e.g. (pp. 60ff.) does not offer any useful insights on goal formation. It is preoccupied with technology implementing issues, offering no insight about diagnosis at the outset of the project;
- we investigated a reference in CRISP-DM to incidentally discovered information. This is discovered information, which may not necessarily be related to the original business problem, *...but might also unveil additional challenges, information, or hints for future directions.* (pp.30, 57). We were hoping to infer about the formulation of project goals here from a possible description of the link between the

business and this incidentally discovered knowledge. There was not enough for any successful inference;

- a particular eye-catching reference in the graphic representation of CRISP-DM, is the dependency arrow between the *Evaluation* task set and the *Business Understanding* task set (p.13), We researched this section of the document, to see if *Evaluation* offers any diagnostics about the business problem:
  - during the *Evaluation* task set (p.57), the data mining model is evaluated against the *business objectives* or the *business criteria*, but no detail is given about how to formulate them;
  - on (p.57) the data miner *...assesses the degree to which the model meets the business objectives and seeks to determine if there is some business reason why this model is deficient.* The potential cues here were *...degree...* and *...reason...*, but what to measure the degree of, or to base the reasoning on, is not given; there is no diagnostic value here;

We reflect that we could perhaps make inference from references in CRISP-DM about the *utility* of data mining's output. In our opinion the references are:

- establishing *useful* outcome to the project from business point of view (pp.17, 36);
- *useful* insights into a customer relationship (pp.17, 36);
- *users'* needs and expectations (p.36);
- characteristics of the model or information that may be *useful* for the future (pp.56, 58).

These references however are of no help for improving our understanding about diagnosing the business problem, and developing business deliverables about it.

We continued the research in CRISP-DM with the following results:

- the following synonyms do not appear in the document at all; *novate*, *innovate*, *innovation*, *paradigm* or *paradigm shift*, *revolution**, *breakthrough,* nor any form of *renew**;
- *change* is not used once within the context of the benefits of the outputs of the project;

- no form of *advance** is used within the context of the project's outputs;
- *impact** is used twice in context of outputs:
  - with reference to the *departments* that are impacted by the data mining (p.35). This offers no insight about diagnosing the business problem; and
  - with reference to outputs meeting the mining's technical goals (p.58). The technical goals are different goals to the SPC goals we are researching here;
- *transform* is used once (p.74) within the context of an exploratory data technique, associated with developing *insights*, in that case about transforming disloyal customers to loyal ones, and its obvious link is to the *data* during the exploratory phase. It is remarkable that this reference in the document, to what it is that data mining is meant to contribute to in the BI environment, is found in an appendix on problem types. Sadly it offers no diagnosis about the business problem;
- *knowledge* - which we earlier proved constitutes the executibility or actionability of information - is used three times within the context of output (pp.14, 60, 77). We found nothing which suggests that a diagnosis of the business problem took place in the first instance;
- *information* is used in the context of outputs three times, and *novel* twice:
  - for assessing or evaluating the *novelty or usefulness* of output (pp.57, 58), but what the usefulness is about is not described;
  - about the summarisation of discovered hidden knowledge (p.72) for the formation of hypotheses. What the hypotheses are formulated about is not described, nor is diagnostic technique offered for their formulation;
- the commercial relevance of *discovered* segments (pp.72, 73) is linked to the segmentation base, but no diagnosis is offered about establishing the segmentation base;
- the results of the data mining are:
  - *assessed* against technical data mining success criteria (pp.29, 30, 56). This gives no insight into the business needs; and

- *evaluated* against the business objectives (pp.30, 31, 57) and business success criteria (pp.31, 57, 58). Again, no diagnosis is offered about establishing these in the project;

❖ finally CRISP-DM then references its Appendix V.2 for *data mining* problem type descriptions. There is no diagnostic technique, which we could transpose to the understanding about the business problem.

There is also no evidence in the text of CRISP-DM, for any importation of pre-formulated SPC goals into the project. Such goals would be either:

- the organisation's *existing paradigm* about the problem within their Corporate Strategy; or
- the organisation's equivalent to our *pre-project schema*.

From the above it seems clear to the researcher, that CRISP-DM intends the formulation of clear business objectives / goals, or business success criteria or business questions etc. *within* the opening activities of the data mining project. This confirms the researcher's corporate experience, where a business comes to any project methodology, in the expectation of finding support for formulating the SPC goals for the project.

Since it appears to be the intention of CRISP-DM for these *objectives / goals* etc. to be developed during its *Business understanding* activity set, one would expect to find substantial diagnostic, goal-formulating project activities in CIRSP-DM's *Business understanding*. In the event where no pre-project schema existed, these diagnostic activities would then *formulate a pre-project schema* from the existing paradigm. This schema would contain the *expectations* of the organisation on the project, about a new solution, made possible by the use of the supporting technology. Following that - or in the event where a pre-project schema was pre-existent - these diagnostic activities would *develop the schema into business deliverables*, which are achievable through using data mining as the enabling technology.

4.1.2 Research findings

We find nothing in the above research, which we can justifiably interpret as diagnostic technique for either *creating a pre-project schema*, or *SPC goals* or *business deliverables*, or even for importing them if they were pre-existent (Pyle 2004a, Rule 1). We find that CRISP-DM does not offer utility for unlocking an existing paradigm.

This finding is compounded by the fact that CRISP-DM does not clearly distinguish between the *business deliverables* of the project, and the *technical data mining goals, which* support those. CRISP-DM treats them as equal, circularly defining the *data mining goals* as *…the intended outputs of the project that enable the achievement of the business objectives* pp.40, 64. CRISP-DM makes it clear that the data mining outputs are technical in nature e.g. accuracy measures, lift values pp.18, 40, 41, confirming our finding that it does not provision for output against SPC goals.

We further find that CRISP-DM does not communicate to the business community, the versatility of data mining in business decision-support across a spectrum of goals. The above stages of the *pre-project schema* are an example of a possible goal spectrum. This non-communication detracts from the methodologies' utility as a communication instrument, and as a project methodology. The author believes that this is one of the causes for the low uptake rate of the technology by business decision-makers.

4.1.3 Reflection-in-action

We here demonstrate the effect of the above on Telco ABC's project. There are two scenarios upon which to reflect. The first is where Telco ABC's project's *business deliverables* are taken directly from their existing objectives hierarchy, without having been diagnosed first. There we had the business's goal from the operating retention objective:

- *to reduce our customer's churn from our estimated 3.5% per month, to between 1% and 1.3% per month within six months, and maintain it there.*

Undiagnosed, this becomes an SPC data mining project goal:

- *to determine whether we have reduced our churn from 3.5% per annum to between 1% and 1.3% per annum, and if that is being maintained.*

This would produce an unnecessary project where:

- a data mining project is launched to determine what can be solved by simple SQL queries, and sum and division functions on a hand held calculator; further
- knowing if this goal is being achieved or not, does not constitute a solution. It measures a solution, which should have been defined as a business deliverable in the first instance.

In the event where the project's goals are based on the diagnostically unassessed *pre-project schema*, the SPC goals for the project become – *and are limited to* – Telco ABC's expectations in their *pre-project schema*. Undiagnosed these expectations become:

- *to determine the true extent of the business problem in quantities of actual churners* (SCHEMA 1); and
- *to know the identity of potential churners in advance* (SCHEMA 3).

Achieving the first goal would be attainable *without* a data mining project, through the strategies described in the SCHEMA 1 before. Knowing the true extent of the problem would certainly constitute an improvement over the *status ante*, but not because of any utility of CRISP-DM.

We reflect that attaining the second goal will also be an improvement over the *status ante.* Note also, how this goal is attainable without any further diagnostic evaluation, and therefore without using CRISP-DM. A classifying neural network will give Telco ABC the known identities of customers who have a p-score of >0.5 of belonging to the class *potential churner*.

The reader will be in temptation now to argue, that since these results will bring about an improvement which meets Telco ABC's expectations, it appears that the project will have attained its goals. The researcher's counter argument to that, is that even though the results meet the organisation's expectations, our measure of results is an objective and competitive one. There still have been no diagnoses which would test this attainment of existing goals against innovative market segmentation developments, or against what Telco ABC's competitors are attaining in this same domain.

Telco ABC's project results therefore remain sub-optimal in a number of areas:

- first, it is sub-optimal technically – non-linear models generated by neural networks are not good at maintaining their generalisability over time lapsed. This means that the model would have to be rebuild comparatively more often than a linear model like that produced by regression;
- second, when we compare this newly discovered information against the STP requirements of a marketing solution, it becomes apparent that knowing the

number of past actual churners, or the identity of potential future churners, does not constitute executibility for this marketing problem at all;

- third, they will have lost an opportunity to infer root cause of the problem, because the effects produced by a neural network are un-interpretable commercially;
- fourth, they will still be depending on an ineffective retention management solution. One of the reasons the retention management solution is failing, is because it is based on demographic segmentation, and not behavioural segmentation. Improving the targeting in a more proactive fashion does not overcome the segmentation problem. Depending on the utility of CRISP-DM Telco ABC would have lost an opportunity to formulate hypotheses about new segmentation, and to have supported that solution with data mining.

This is an unsatisfactory outcome for an organisation requiring a new solution to a nagging business problem.

4.1.4 Evaluative reflection and reframing

In this section, we reflect about how to overcome the limiting SPC outcomes from Telco ABC's project. We reflect about ways to overcome the limitation:

1) we need to shift their paradigm by:
   a) reframing Telco ABC's business problem, in a way which maintains their valuable existing *expectations*; and
   b) extending their *preconceptions* about what is possible from the project past their current limited preconceptions; and
   c) extending their *expectations* to include understanding the problem cause better, to develop a new retention management solution to replace the failing one, and to use data mining as a tool for supporting novel retention management solution;
2) provide a tool for bringing about the required paradigm shift.

Based on concepts introduced in the literature study in Chapter 2, we hypothesise that both these requirements can be met by using a data mining project methodology which:

- ❖ injects new subject matter expertise into the stale business problem; and
- ❖ has sufficiently diagnostic activities for Telco ABC to:

- evaluate their business problem in light of the new domain knowledge; and
- reframe their preconceptions about what is possible; and
- develop new expectations based on that reframing; and
- develop competitive business deliverables.

We hypothesise that we can incorporate this functionality in SAM.

Moving forward, we have to determine if CRISP-DM offers any utility about this requirement for injecting new subject matter expertise into the stale problem situation. We evaluate CRISP-DM against this criterion in the following section.

4.2 Introducing new subject matter expertise

This is a distinct concept from learning about other data mining solutions for similar problems. It is also distinct from technical myopia, which is defining the problem in terms of what is familiar in the data (Pyle 2004a, Rule 2). Our point in this section is about *not* defining the business problem *in terms of what is familiar to the business*. It is about recognising new subject matter expertise as the new repertoire of understanding and knowledge creation (Pyle 1999, p19), and formulating breakthrough business deliverables for the SPC project. Applied to our Telco setting, the SME would be new marketing subject matter expertise about segmenting and targeting.

4.2.1 Analysis of CRISP-DM for introducing new subject matter

We systematically researched CRISP-DM for any recognition of the need of introducing new subject matter expertise - and once introduced - for any technique about developing breakthrough business deliverables for the project. From a TQM perspective, the introduction and development should take place during any SPC goal formulation stage – or its equivalent in the case of CRISP-DM.

An electronic scan of CRISP-DM for the use of the word *knowledge* returned 23 instances. In some cases, the reference is *post* the formulation of any goals, where they will be of limited use to the SPC project. In other cases, the references precede the task set *Determine data mining goals*, where the introduction potentially would be of value to the SPC project. We analysed the following instances in our evaluation against this criterion:

- (p.13) - *knowledge* is used with reference to understanding the business's *requirements* on the project, and converting that knowledge into a *data mining project*. There is no indication here about any *novelty* associated with the knowledge, or for harnessing it to formulate breakthrough business deliverables;
- (p.18) - the context is that of establishing a business terminology glossary of the business knowledge available to the project. The context suggests capturing the business knowledge status quo only. This reference contains no requirement for novelty nor technique for harnessing the knowledge in goal formulation;
- (p.22) - CRISP-DM expresses a dependence on business knowledge for making up the data quality report, but there is nothing to suggest focusing the data evaluation toward novelty associated with the business knowledge. As it stands, this reference seems to fall into the category of defining the business problem in terms of what is familiar to the business in the data;
- (p.29) – here the data miner interprets the model according to his domain knowledge. There is nothing here suggesting a novelty dimension to the evaluation;
- (pp.37, 39) - has five instances of subject matter expertise within the context of the assessment of resources for the project. All these references are concerned with identifying and assessing knowledge sources and their availability. There is nothing here we interpret as a requirement for novelty, or as technique about harnessing novelty;
- (p.47) - there is a reference to the checking of assumptions, about what may be represented in the data, against knowledge. If we interpret this reference as referring to business knowledge, there is no indication of the need for novelty in the knowledge;
- (p.50) - relates to the creation of data features to represent important background knowledge, which is not represented in the data. There is nothing to suggest that the background knowledge which is being pursued in this activity, is novel;
- (p.51) – refers to the creation of new records with which to represent new knowledge. We interpret this as a post-data mining knowledge management activity, recording any new information discovered by information. This

instance offers no utility in formulating breakthrough business goals for the SPC project to begin with;

- (pp.56, 58) – this reference is about evaluating the discovered knowledge from the data mining project, against the existing knowledge base, to determine the usefulness of the discovered knowledge. The evaluation criterion for usefulness is not given, and there is nothing contextual from which to infer an element of novelty as a criterion. We therefore interpret this instance as propagating the existing knowledge base and paradigm;
- (p.60) – refers to deciding how *discovered knowledge* should be deployed to its constituents. There is nothing here we could interpret as a requirement that the discovered information was in support of newly introduced subject matter;
- (p.73) - in an appendix on business problems, there is reference to the value of the analyst's prior knowledge, for determining manually or semi-automatically the relevance of data structure to a business problem. There is nothing in this instance suggesting novelty of that prior knowledge. Further, the next sentence is unhelpful, saying that the data mining technology will automatically discover *unsuspected* structures in the data. Logically we can not equate *expectation* with *novelty*. Further, considering that new prior knowledge is required to improve our perceptive boundaries toward perceiving unexpected stimuli, we concur with a prominent author that this instance constitutes a reliance on the technology *…to reveal all*. (Pyle 2004a, Rule 4);
- (p.77) – refers to discovered knowledge, which is post hoc defining any project goals. There is nothing in the context suggesting that the discovered knowledge supports new SME in the first place.

We expanded the scope of the search within the CRISP-DM documentation, to include a systematic search for the incidence of the use of the term *business.* Within those search results, we narrowed the investigated down to those instances which in our opinion are concerned with the *what and how* of the business – objectives and strategies - which need to be understood for formulating business goals / deliverables for the project. We could not find one instance from which to develop insight about a requirement for *renewal* of business objectives and / or strategies, *before* formulating data mining objectives.

4.2.2 Research findings

We found no instances in the CRISP-DM documentation, which we can justifiably interpret as recognition *of the contribution* of newly introduced subject matter expertise toward formulating breakthrough goals for an SPC project. We also found nothing we could interpret as a *technique for introducing* such required novelty into the project environment. We evaluate that CRISP-DM does not offer utility about introducing new subject matter expertise into the project environment.

4.2.3 Reflection-in-action

We reflect that without injecting new subject matter expertise about marketing segmentation and targeting, Telco ABC will remain locked into their ineffective retention management solution. This is an undesirable outcome for an organisation wishing to develop a competitive solution.

4.2.4 Evaluative reflection and reframing

The way for overcoming the unwanted outcome is to inject the required new SME into Telco ABC's project environment. Since they are depending on a data mining project methodology for producing breakthrough, we need to offer them a data mining project methodology which offers utility against this criterion.

We hypothesise that if we include utility in SAM against this criterion, we will be able to produce a competitively advantageous business solution for Telco ABC. We move-test this hypothesis in Chapters 5 to 7 of this thesis.

4.3 Mapping technique between business deliverables and data mining plan

In the opening chapters, we established from the literature the *utility for purpose* of the data mining technology. The *utility* of data mining is in discovering information, and its *purpose* is supporting SPC projects achieve breakthrough (Levinson 2000). In the ERP and CRM industries, the *utility for purpose* of the technology is achieved through strategically aligning the technology plan to the business deliverables of the project. Those industries achieve the alignment through a practical *mapping technique*.

The mapping techniques first define the support – or *purpose* - which is required by each business deliverable from the technology. Second, they define the technical plan –

the objectives and strategies – which constitutes the *utility* required for attaining the purpose. In the event of data mining, the mapping technique should first define the *purposeful* support required by each business deliverable from the data mining. Next, it needs to define the data mining plan which constitutes the *utility* for attaining that support (Cooley 2003, p.608). By virtue of its nature as a plan, any *data mining plan* therefore should be described in terms of its *objectives (or goals)* and *strategies*.

In the case of CRISP-DM, the mapping is required between CRISP-DM's *Business understanding* and the formulation of its data mining plan. This chapter describes research, which we directed at the *utility* of CRISP-DM about such a mapping technique. We researched how CRISP-DM goes about *mapping the data mining plan to the SPC goals or purpose*, in a way which best supports the SPC project's goals or purpose (Cooley 2003, p.608). We were particularly interested in how CRISP-DM:

- formulates data mining *objectives* which target the *pursuit of the relevant information*, in support of the business deliverables of the SPC project; and
- formulates data mining *strategies* which *support* the data mining objectives.

4.3.1 Analysis of CRISP-DM for mapping technique

We first researched the CRISP-DM documentation for linkage between the data mining goals, and what CRISP-DM considers the project's SPC goals. The research commenced with an analysis of CRISP-DM's *Business understanding section*, subtask *Determine data mining goals*:

- (pp.18, 40) – the business goals become the data mining goals. We consider that the mapping technique here is simply making the business goals the data mining goals;
- (pp.18, 40, 41) – there is reference to the intended data mining outputs that enable the achievement of the business objectives. The *intended outputs* are technical, with no description of how to link them to the business deliverables;
- (pp.18, 41) – we find loosely formulated criteria for successful data mining outcomes. The following sentences clarify that the criteria are technical measures of the quality of the model, like lift etc. There is no description of how to link these criteria to the business deliverables. We interpret this reference as

an incorrect assumption that quality of modelling equates to support of the project's business deliverables;

- (p.19) – we find a task description for producing the project's (technical) plan, as *Describe the intended plan for achieving the data mining goals, and thereby achieving the business goals*. We perceive an evident link here between the data mining goals and the business goals. However, it falls short of constituting a mapping technique;
- (p.40) – there is an activity for translating the *business questions* to data mining goals. Here is a link from the business questions to the data mining goals. Again, no technique is offered for the translating;
- (p.40) – another activity is to specify the data mining problem type. We can interpret this as recognising the need for mapping a data mining *goal* to a data mining *strategy*. However, the mapping technique between the business goals and the data mining goals is not described;
- (p.41) – there is an unhelpful warning that the data mining success criteria is (are!) different than (from!) the business success criteria;
- (p.42) – generating the data mining plan is described as *Put all identified goals and selected techniques together into a coherent procedure that solves the business questions and meets the business success criteria.* Apart from mistaking a procedure for a plan, no mapping technique is offered between the data mining goals and the SPC goals.

Two instances in CRISP-DM hint of substance in mapping technique. The first example on (p.77) refers to the adjustment of pricing, based on information, which was discovered using data mining. The second example on (p.40) refers to the discovering of information about market segmentation. Both pricing and segmentation are marketing strategies. These two instances are of interest because they link the *goals* of the data mining to business *strategies*, and not to business *goals*.

In search of further clues, we did a programmatic search of CRISP-DM for the words *objective*, *goal*, *plan*, and *strateg**. These search words represent key components of any plan.

4.3.2 Research findings

We found in CRISP-DM diversity about how to map the *SPC project's* business deliverables or goals to *data mining* goals, and about formulating data mining goals from there. The mapping is achieved through; *making*, *translating*, *intending*, *defining criteria*, *describing the plan* etc. We find the seeds for *nuancing* a mapping technique in the two instances referencing business strategies. None of these references however, offer substance about mapping technique. We found nothing in the results from the plan keywords search contradicting the finding.

4.3.3 Reflection in action

Based on our corporate experience, our hypothesis is that the mapping technique we found in CRISP-DM, has limited utility in formulating supporting data mining goals from the SPC goals. In order to support this hypothesis, we have to *reflect-in-action* about formulating data mining goals from diagnosed SPC goals for the Telco project. To make this reflection possible, we now present fully developed SPC goals for the Telco ABC problem. For our reflective purpose here, we are less concerned about how we arrived at these SPC goals, than about their content. We do mention however, that we did base the diagnoses and goals on the literature review in Chapter 2.

The developed *SPC goals* for Telco ABC then are:

1. the goal about the existence of the problem, is *to determine the extent of the problem*;
2. *to learn about root causes of the problem* and *to develop a hypothesis about the nature of the problem*;
3. *to remedy the retention problem with a new retention management solution.* Since we are moving into the solution space, this would include *developing hypotheses about how to competitively remedy the problem*. The hypotheses should consider:
    - 3.1. *knowing the identity of the potential churners in advance, and within a time window, to gain a time advantage and to focus on the most-at risk customers first*;
    - 3.2. *segmenting those most-at risk customers on the basis of their need for a mobile phone, risk to churn, their value to the organisation, and their chance of responding positively to a retention campaign*; and

3.3. *addressing problem root cause on a within segment basis*;

4. *to execute this new solution on a monthly basis using data mining as a supporting technology*.

Executing *CRISP-DM's* mapping technique – which is making the SPC goal(s) the technology goal(s) - we now reflect on the usefulness of the subsequently defined technology goal(s) in supporting the above SPC goals.

The first-stage technology goal becomes *to determine the extent of the problem*. This is a supportive goal, attainable through the non-data mining strategies of the *pre-project schema* - discussions, reverse engineering business processes etc. We therefore carry those strategies over into the data mining project, where they will support this first SPC goal. CRISP-DM's bridging technique therefore is useful in supporting the first SPC goal.

The *second-stage* technology goal becomes *learn about the root causes of the problem.* This goal supports its SPC goal. We can formulate a supporting technology strategy for attaining this goal, based purely on technical data mining knowledge, which is *build a Logistic Regression model for discovering information about root cause, and make inference from the model effects*. Executing this data mining strategy will directly support the second-level SPC goals.

CRISP-DM's bridging technique therefore is useful in defining the purpose for the supporting technology in the second, or analytical stage of the SPC project. This is because during the analytical stage, the attainment of the SPC goals is determined more by what is possible technologically, than what is possible from the business solution (Patrick 2004). This means that, based on *technological* considerations about the greater suitability of a Logistic Regression for this type of problem, we may amend the strategy of the *pre-project schema* about using a Neural Network.

We now reflect on the data mining goal we could formulate from the *third-stage* SPC goal using CRISP-DM's mapping technique. That mapping technique was to make the business goals the data mining goals. Accordingly we formulate the data mining goal *to remedy the retention problem with a new retention management solution.*

We reflect that this data mining goal is inexecutable – we are unable to formulate data mining strategies which formulate it at face value. The reason is that data mining

algorithms are devoid of subject matter expertise, and therefore can't discover a solution for the retention management problem. Further, the business objectives are also devoid of the SME about executing those objectives. We evaluate that CRISP-DM's mapping technique of simply making an SPC goal the data mining goal, is ineffective once the project reaches the business solution development stage.

This deficiency in CRISP-DM perpetuates perceptions within the sceptical factions of the business community, that it is too complex to map the business needs into a data mining problem. Of even greater concern, is that a business, which depends on CRISP-DM for defining the purpose of the data mining in the solution space, may end up with disappointing results from the project.

4.3.4 Evaluative reflection and reframing

The mapping technique documented in CRISP-DM is sufficient for exploratory data mining, but breaks down when we enter the solution space. The reason for the ineffectiveness lies in the distinction between an *objective* and a *strategy* we made in previous sections. We hypothesise that we can overcome this shortcoming with the following mapping technique as follows:

- in the first two problem exploring and analytical stages of the SPC goal sequence, the mapping is from the SPC *goals* to the data mining *goals,* and then developing data mining strategies which support the data mining goal(s) or objectives;
- in the subsequent business solution exploring and operationalising phases, the mapping is from the SPC *strategies* to the data mining *goals*, and developing data mining strategies, which support the data mining goal(s) or objectives.

We find then, that the link between the SPC goal(s) and data mining goal(s) are dependent on the stage of the SPC project, but mapping the data mining *strategies* is independent of project stage.

We move test by applying our modified mapping technique to the *third-stage* of the SPC project. There, the SPC goal depended on having information about three new concepts we introduced from the marketing SME. These three concepts are none less than *SPC strategies* for solving the retention management problem. We therefore have to define data mining goals for discovering information for developing hypotheses about the business *solution*.

We now map our data mining goals to these three concepts:

- *to identify the potential churners within a time window;*
- *to identify the behavioural, demand, value and risk segment membership of the most at risk customers; and*
- *to identify churn root cause on a within segment basis.*

We are now able to formulate a number of executable data mining *strategies* in support of these three goals:

- *build a Logistic Regression model with the actual churn events of the last three months as the target;*
- *score the customer database with this model;*
- *sort the customer database by their propensity for becoming a potential churner, and select the potential churners for further attention by the* risk tolerance factor*;*
- *cluster that part of the customer database into six segments, by the four base features which represent the segmentation base (*ARPU, SUMC, p_t1, and customer loyalty*);*
- *profile each segment for root cause, using the Effect with the highest association to the target, and the clustering algorithm's OLAP output;*
- *profile each segment for their executable elements in the channel dimension.*

We evaluate that this modification to the mapping technique is useful during all phases of the SPC project. We move test this by incorporating this mapping technique in SAM, and validating it on Telco ABC's data in the chapters following on SAM.

4.4 Knowledge management activities

This criterion is about what one author has called *metastatistics* – the practical business inference actually drawn from the discovered information (Liu 2003, p.437). We saw in chapter two, that hypothesis formulation and testing cannot be attained technologically in the SPC environment. Similarly, developing executable knowledge cannot be achieved technologically. Both hypothesis formulation and knowledge development are part of the Knowledge Management process, for which the tool is human cognition. We

demonstrated in the section preceding, that we were unable to define a data mining goal for developing hypotheses about the nature of the problem, and about the new retention management solution. At best, we were able to define data mining goals and strategies, which resulted in the discovery of information, which can be used in cognitive hypothesis formulation and testing, and solution development.

In this fourth criterion, we reflect on the utility of CRISP-DM for cognitive knowledge management. This knowledge management process is dependent on the relevant information, which we discovered with data mining. This knowledge developing is required in order to attain to the master SPC goal, which is an objectively optimal, novel and executable retention management solution.

4.4.1 Analysis of CRISP-DM for knowledge management activities

We saw in Chapter 2 how an ERP project with SPC goals executed an interwoven chain of technical and cognitive knowledge developing activities. That was achieved using the SPM, because SPM contains a Knowledge Management process. Any data mining methodology aimed at the SPC environment, should similarly have KM utility.

In Figure 4.1 we visually express our experiential insights about the knowledge outputs, which are required from such a data mining project methodology. These outputs are in the triangle, and sequenced bottom-to-top, incrementally adding knowledge required for producing the project's business deliverables. These knowledge outputs are:

- formulated hypotheses about the existence of a problem or opportunity, its scope, and its solution;
- selected data relevant to the business deliverables;
- discovered information relevant to the business deliverables, which will be used for hypothesis testing;
- modified hypotheses for the limiting organisational circumstances;
- developed, executable solutions for the business problem or opportunity, best suited to the organisation's limiting circumstances.

In the right-hand column of Figure 1, we have generically summarised the project activities which are included in CRISP-DM, and their sequence. Within there we have underlined those activities which we classify as potentially knowledge formulating.

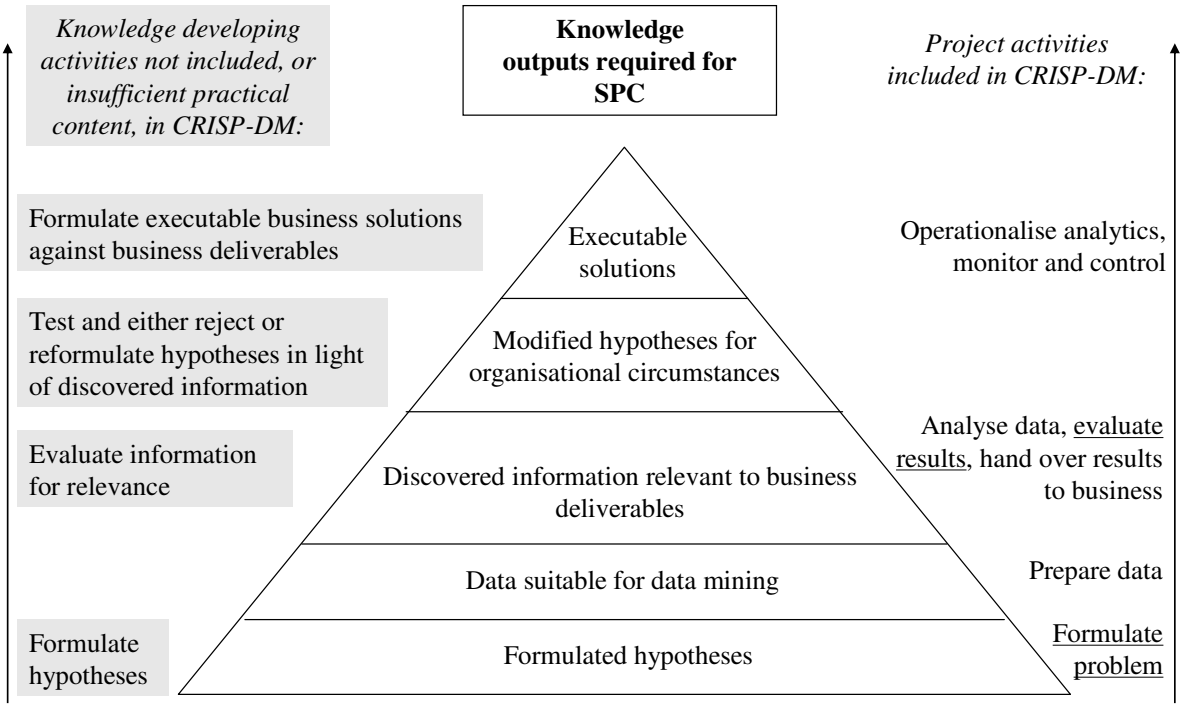


Figure 4.1: Required Knowledge Management activities

We have already seen that the *Formulate problem* activities of both methodologies have insufficient practical content about formulating hypotheses for basing business deliverables on. This practical content is required in the SPC environment for taking the practitioner from the unstructured pre-project environment of ideas and suspicions, into the structured project environment with business deliverables. Because the research found *Formulate problem* lacking, we display the words *Formulate hypotheses* in a greyed rectangle in the left-hand column of Figure 1, under the heading *Knowledge developing activities not included, or insufficient practical content, in CRISP-DM.*

The next CRISP-DM project activity which we consider, is *Evaluate results* (Chapman, Clinton et al. 1999-2000, Evaluation). In that section (p.57) the output of the project is defined in terms of *findings*, and these findings related back to the *business questions*. We have already determined that in CRISP-DM the *business questions* etc. do not include a consideration about an executable business solution. We did not find sufficient practical content for supporting the required SPC knowledge management, for instance a tool like an information score card or equivalent, for evaluating information against

the business deliverables. We reflect this in the left-hand column with the label *Evaluate information for relevance*.

A decades-long authoritative source on the business planning process (Pearce and Robinson 1991; Pearce and Robinson 2004) makes it clear that methodologies used in the SPC environment, are incomplete if they do not include developing new knowledge from insights, adapting that knowledge for limiting organisational circumstances, and finally developing executable business solutions. CRISP-DM terminates with the handing over of business reports, the operationalising of the models, and activities which express the need for developing monitoring and maintenance plans (Chapman, Clinton et al. 1999-2000, Deployment section). We researched CRISP-DM's *Deployment* section for the remaining knowledge management activities. The subsection *Plan deployment* talks about:

- having a *strategy* for deploying the data mining's results into the business (p.60);
- the strategy has *steps*;
- the activities within this task set are:
  - *summarising* the deployable results;
  - *distinguishing* between the various *…knowledge or information…* results. There is evidence that CRISP-DM considers *knowledge* and *information* as a synonymous of what is discovered by data mining (Van Rooyen 2004);
  - *monitoring* the use of the results or their benefits;
  - technically *deploying* the results within the *organisation's systems*; and
  - *identifying* potential pitfalls in the deployment.

We found here recognition for a *strategy* for dealing with the results of data mining within the business, including specific steps for that. However, there is nothing in the *activities* of CRISP-DM's *Deployment* section, which we consider of value for modifying or testing hypothesis for organisational circumstances, or for executing business solutions. The activities described by CRISP-DM are all directed at purely processing the informational outputs of the data mining, for *reporting* them to the business. That much is clear from the content of the report, which we find in the subsection *Produce final report*. We reflect these findings in the top two grey labels in

Figure 4.1; *Test and either reject or reformulate hypotheses in light of discovered information* and *Formulate executable business solutions against business deliverables.*

4.4.2 Research findings

It seems that at best, CRISP-DM considers that the data mining activities will only bring the relevant information to the attention of the organisation through reporting it. We find that the activities included in CRISP-DM have insufficient content and technique about KM for the SPC environment. One source describes this state of affairs as *the minimisation of interaction between the data miners and the business people (Pyle 2004a, Rule 8).*

This finding reflects some of the data mining and business analytics literature, where we have found a mentality of *hand-over of the results of the data mining to the business owners of the problem.* That mentality considers the business solution development as falling outside the domain of the data mining project, and in the domain of traditional analytical methods. An example of this mentality in the Telecoms analytics literature is (Mattison 1999, pp. 38, 39, 41, 43, 44); there the predicted concept is handed over to the marketing analyst, for developing actionability (knowledge) using traditional analytical methods.

4.4.3 Reflection in action

We reflect on the outcome of this state on the achievement of our developed staged SPC goals for Telco ABC:

- we will have difficulty in formulating hypotheses about the extent of the business problem;
- we see difficulty with the attainment of the goal to learn about the root cause(s) of the problem too. Such learning requires the cognitive manipulation of discovered information into hypotheses about the reasons for that root cause. So for instance if we discover information that *Handset type* is the effect with the biggest association with the target, we still have to infer that it indeed can be considered a root cause in the real world. After that, we have to analyse cognitively why *Handset type* may be root cause. That analysis may include making the mental connection with service records that certain Handset types indeed have technical problems, or poor reception etc. It is only after this has

been achieved, that we can start hypothesising about how to address root cause, by for instance replacing handsets with known problems;

- in the third stage, we need KM activities for formulating hypotheses about how to address the problem. We need to hypothesise about the information we need to discover to support the new SME. When that information becomes available, we need to manipulate it mentally to support or refute the hypothesis. In addition, if it supports the hypothesis, we need to develop the hypothesis into new business objectives and strategies, constituting paradigm shift. Without KM activities like these, the discovered information will remain inexecutable, not constituting a business solution.

We recognise that at the time of its publication, CRISP-DM brought substantial improvement on the *status ante*. The content of CRISP-DM strongly suggests that the intended improvement, was to direct the use of data mining in exploratory SPC analysis. We reflect however, that CRISP-DM presents itself as a standard generic data mining process, with the purpose of proving *…that data mining was sufficiently mature to be adopted as a key part of their business processes…* (Chapman, Clinton et al. 1999-2000, p.3). We interpret *…business processes…* as those business processes, which utilise the proven utility of data mining, for contributing toward breakthrough in the competitive commercial environment. In this environment breakthrough *must* extent past the understanding of the nature of the problem, to include also breakthrough *in solution*. We reflect that CRISP-DM is incomplete against this criterion of utility. The effects of this are:

- naïve users may think that they can depend on the data mining tool itself to reveal all (Pyle 2004a, Rule 4), and will be disappointed in attaining their business deliverables;
- experienced practitioners will continue to revert to tools outside of CRISP-DM for achieving a competitive SPC outcome;
- inexperienced practitioners in the use of data mining, are unable to distinguish between the technology and its project methodology, and will perceive an incapacity of data mining to contribute toward business deliverables.

We consider that this criterion may be trifle to the experienced practitioner, who may argue that it is obvious what KM activities are required during what stage of a SPC

project. We reply to that with the rhetorical question *Then why are these activities then not included in CRISP-DM?* We support our point by pointing to the difficulties the ERP and CRM technologies initially had, in incorporating KM in their project methodologies.

4.4.4 Evaluative reflection and reframing

We hypothesise that it is possible to develop a data mining project methodology, which includes the KM activities, which are required for the SPC environment. We believe that the key to achieving this is integrating SPM with data mining project methodology. We demonstrate this approach later with SAM, and validating it in later chapters.

4.5 Monitor and control plan

Monitor and control are important concepts in the business literature (Kotler 2002; Pearce and Robinson 2004). They are also important in data mining, and we dedicated Chapter 3 to them. Their importance derives from the need to assure the ongoing relevance of solutions under changing circumstances. In the SPC environment, there are aspects of the business problem, business solution, and the supporting analytics which require monitoring and control.

Analytically we need to monitor and control *concept drift*. Business components which require monitoring and controlling are *strategic creep*, the financial return on investment (*ROI*) from the project, and the *results the business solution is having* on the business problem or opportunity.

The changes we monitor for are caused by a natural tendency of events to regress to the mean (Pyle 2004, pp.511ff.), and in the SPC environment also by the influence of competitors in the market.

Control then is adjusting the business operations and / or analytics in response to the monitoring of parameters about key performance indicators (KPI), to keep the operations or analytics within acceptable KPI parameters.

Concept drift methodology provides the tools upon which to build an effective technical *Monitor and control* plan. In the next section, we present our research of CRISP-DM for substance relating to *concept drift methodology*.

4.5.1 Analysis of CRISP-DM for a monitor and control plan

We researched CRISP-DM with special attention to the section called *Plan monitoring and maintenance*. This is a subtask within the *Deployment* task set of CRISP-DM:

- the *Deployment* task set itself mentions *... check for dynamic aspects (i.e. what things could change in the environment.)* (p.61). We interpret this as referring to monitoring for manifest contextual creep in the data, similar as to what is described in the concept drift literature. No guidance is offered about technical monitoring method, nor how to identify these dynamic aspects;
- we found *...will the business objectives ... change over time?* in the *Deployment* task set (p.61). We think a generous interpretation of this, is as a veiled reference to concept drift which is *non-manifest in the data,* and which is reflected as strategic creep in the organisation's objectives and strategies hierarchy over time. That creep would have been generated by the organisation's cyclic application of the SPM, or its equivalent. If our interpretation is correct, then CRISP-DM has identified the *need* for monitoring this drift, but omitted any guidance on how to *do* the monitoring technically;
- the same section on (p.60) states the monitoring of technical aspects of the model. However, CRISP-DM does not offer any guidance on the technical *how* of monitoring changes of such measures e.g. changes in error measures, predictive accuracies etc.;
- in the *Deployment* task set (p.60), the *...use or benefits…*of data mining are monitored. We interpret this reference as identifying the need for a control mechanism for the ongoing ROI from the data mining project, against an ROI, which correctly should have been calculation at the outset of the project. Financial management methodology would be of use in monitoring this aspect of the model, but CRISP-DM does not reference such methodology;
- in the Deployment task set (p.60) there is reference to a *maintenance strategy and process*. We equate the maintenance as the *…control* portion of any *Monitor and control* plan. CRISP-DM offers no insight that any maintenance / control have to be preceded by an effective *Monitoring…* part of the strategy. That results in guidance being offered neither about *what* to monitor, nor about monitoring *methodology*.

4.5.2 Research findings

Our research found that CRISP-DM *recognises the need* to monitor for *concept drift*, *strategic creep*, and the project's *ROI*. Further, we found that CRISP-DM recognises that this need is met through a monitoring and control *plan*, and that the monitoring and controlling *is a time-driven process*. Practically however, we found that CRISP-DM's content does not accomplished any more than *statements* of recognition. There is no technical substance, about which we can say will result in the formulation of an effective *Monitor and control* plan. Most importantly, there is total silence on the need for monitoring the response of the business problem or opportunity against the business solution for it.

4.5.3 Reflection in action

Both methodologies rely heavily on the assumption that a monitor and control plan will be realised. Experienced BI practitioners and business administrators view this omission with concern, and this does not further the cause of data mining, in becoming BI's analytics support of choice.

In the case of Telco ABC's project, the effects of no technical monitoring may be:

- loosing track of the root causes of the churn problem over time. The effect of this on the business solution, is that the retention campaign offer may seek to address irrelevant causes of the problem;
- not noticing that the accuracy of the predictive model is deteriorating over time. The effect of this on the business solution, is that the retention campaigns may start targeting customers who are not necessarily the most at risk of churning;
- not noticing that the natural segment groupings within the data are changing over time. The effect of this on the business solution is the segmentation of the campaign offers becoming irrelevant over time.

Generally we are concerned about not leaving the business community ill at ease about the data mining community's appreciation of the importance of monitoring and control on business administration.

4.5.4 Evaluative reflection and reframing

We reflect that the above unwanted outcomes can be avoided, by an analytics project methodology having utility against this sixth criterion. The researcher is of the opinion that the utility of any analytics methodology will be enhanced, by drawing on concepts, vocabulary, and practice about monitoring and control, from both the business and the *concept drift* literature. A useful *Monitor and control* section should also include caveats about interpreting the measurements of the monitored parameters, in open problem environments, where not all the influencing factors are captured in the data. The SPC data mining environment falls into this category.

We further reflect that it is possible to include this utility into such a methodology, and we demonstrate that by designing an effective *Monitor and control* activity in SAM.

4.6 General and soft issues

In this section, we offer practical insights first on some issues, which have the potential to affect negatively the outcome of any project in any environment. Business systems authors e.g. (Checkland 1981) typically refer to these as *change management issues*. They present an invisible frontier of practical and political human-behavioural problems, which usually only manifest themselves during the implementation and deployment stages of projects. The best strategy for dealing with them is to include activities within the project plan, which pro-actively manage them.

We also offer comment on some other issues, whose potential negative effect on the project's outcome, can be avoided simply through good practice.

4.6.1 General issues

4.6.1.1 Artificial separation of model and information evaluation

The presentation of the *technical assessment* of the models, and the *evaluation of their outputs for supporting the business deliverables*, is artificial (pp.28.ff). The evaluation of the outputs of the model against business criteria and the technical assessment of models, should be done on and logic, not on or logic. A model either meets *both* technical and business requirements, or not. A model cannot be approved if it has done well against one requirement, but not against the other.

4.6.1.2 Consideration for TQM principles

The statement on (p.30) in the CRISP-DM's *Evaluation* task set, which describes evaluation of the models as finding *...some business reason why this model is deficient.* is a disregard of TQM's principle of *forward linkage* of utility for purpose. Pyle's Rule 1 for failure at data mining projects applies: *If you're aiming for failure, you don't need to consider how the results will be applied until your results have already been generated and delivered.* (Pyle 2004a).

During the *Evaluation* stage of CRISP-DM, there is reference to the evaluation of the project for its continuation or not (e.g. p.31). This constitutes the provisioning for a possible mid-stream abandonment of a project, and is contrary to the purpose of applying a strategic planning regime in the first place.

Good practice is *identifying at the outset* those projects, which are worth implementing, because of the benefits they will bring to the organisation. This also is a disregard of the TQM principle of *forward linkage of purpose* at the outset of any project, as opposed to toward its end.

4.6.1.3 About data mining discovering knowledge

CRISP-DM's vocabulary about what the data mining in the project produces is *discovered knowledge*. This is contradicts the separation between the two we saw in Chapter 2.

A further result from such terminology is to lead to a perception of a professional threat among the less informed business community. The perceived threat is from a technology, which develops knowledge, potentially replacing their professional cognitive skills.

4.6.1.4 Lack of content about the open business environment

There is no content in CRISP-DM about the unstructured nature of business problems, or about the open system nature of the BI application environment, and uncontrollable influences in that environment. This uncontrollability adds a degree of uncertainty to the response of the problem to the solution. This means that in the BI environment, the impact of the business solution on the business problem is *influence* and not *cause*. We believe some nuancing in this regard will help users of the technology, to better distinguish between a failure in the business solution due to causes beyond their control,

and due to having built the business solution on poor data mining results in the first instance.

4.6.1.5 Lack of feedback loop

There is a lack of content that *Monitoring and control* is part of a feedback loop, which ends with the update of the Corporate Knowledge Base. That updating is to reflect changed internal or external influences and factors. The same comment applies to the reporting of discovered knowledge about other potentially interesting problems or opportunities for the organisation. Further, the Corporate Strategy needs to be updated to institutionalise each round of paradigm shift, which has resulted from the project; the new knowledge needs to be legitimised. The current attitude of CRISP-DM seems to that the data mining project happens at arms length from contributing to both the Knowledge Base and the Corporate Strategy.

4.6.2 Soft issues – the impact of the human factor

There is a lack of recognition in CRISP-DM of the *influence* of the human element within the psycho-cognitive information processing and knowledge development processes. This manifests in a lack of content about the enablers of the human factor in the data mining environment. A practical view of enablers that should be considered in a project methodology follows.

4.6.2.1 Role of subject matter expertise in discovery

Entrenched, *existing subject matter expertise* has a limiting effect on the perceptive parameters during the human perceptive process. This phenomenon is known as paradigm lock. The effect of paradigm lock during information discovery is that it renders us incapable of recognising relevant incoming stimuli. In the data mining environment, those stimuli are the discovered information. Paradigm lock is overcome by the updating of subject matter expertise, of those involved in the project. It follows that the results of even exploratory data mining can be improved by introducing new subject matter expertise into the data mining environment.

4.6.2.2 Role of collaborative teamwork

A characteristic of data mining projects in the SPC environment is the different specialisation of the technical and business domain experts involved in the project. We saw in chapter two how collaborative expert teamwork has been recognised as an

effective tool in the knowledge management environment. The truth is, that the more disparate the specialities involved in a project, the bigger the imperative for such teamwork in producing results with the project.

CRISP-DM is silent about the influence of expert collaborative teamwork. We refer to Pyle's Rule for failure nr. 8: *A skilled miner can easily explore a data set and discover all the interesting, insightful, and useful relationships without any interaction, guidance, or involvement from the business managers.* (Pyle 2004a).

4.6.2.3 Role of professional circumstances

SPC projects are processes, consisting of a sequential chain of techno-cognitive activities. Psycho-cognitive factors are at play during each of the cognitive activities. Generating successful output from such an integrated process, is conditional on more that just the successful execution of the technical activities. It is also dependent on the successful execution of the psycho-cognitive activities in the process. The overall result can be no better than what the weakest link allows. The psycho-cognitive activities need to be recognised as weakest link within this integrated process, and proactively managed. The effects of *circumstantial factors* on the psycho-cognitive mechanisms of information discovery and knowledge development are not recognised. They and their effects are:

- *removing* organisational situational barriers, which subtly and unawares influence perceptive appraisal. Such barriers are personal political agendas, behavioural inertia, and dysfunctional organisational culture etc. This is typically achieved through:
  - visionary and supportive organisational leadership, supporting the development of information into breakthrough knowledge, and the application of that knowledge to its full potential by the organisation;
  - data miners improving their business subject expertise, so that they can better communicate the benefits of what they are discovering to the organisation, in language beneficial to organisational knowledge;
  - changing organisational structure and job descriptions and reward structures to entrench such improvements;

- *optimising* the application of logical and creative cognitive skills and abilities within the organisation. In addition, where those individuals involved in the project are deficient in these, their *upskilling*. (Yes, logic and creativity can be learnt!)

4.7 Chapter summary

This chapter has contributed key research insights about the current state of data mining project methodology. We evaluated the utility of CRISP-DM as an SPC data mining project methodology, in producing competitive, executable results. The evaluation is against the following six key criteria:

1. diagnostic technique for defining the project's business goals or deliverables;
2. mapping technique between the business deliverables and the data mining plan;
3. introducing new business subject matter expertise into a stale problem and solution environment for enabling competitive breakthrough;
4. knowledge developing activities for formulating, developing, testing, and selecting hypotheses, and also for developing executibility required by the business;
5. developing a monitor and control plan; and
6. consideration about important soft issues.

We started the evaluation against each criterion by describing the key utility components required by the criterion. Consequently we researched the CRISP-DM documentation for those key components, and presented what we found. We then applied the components we did find in CRISP-DM to Telco ABC retention management problem. Next, we evaluated the results of the application, by reflecting-in-action on the competitive and executable nature of the result producible for Telco ABC.

Our research established that CRISP-DM lacked utility in all six dimensions. Since the Telco problem and data was typical of the SPC data mining environment, we concluded that CRISP-DM had limited potency in producing competitive, executable results, in the SPC environment. This supported our hypothesis that the state of data mining project methodology was contributing to the slow uptake of data mining in that environment.

We then concluded that we could remedy the state of data mining project methodology, by reframing it for each of the six criteria. Specifically:

- having established that CRISP-DM lacks in diagnostic technique for defining the business deliverables for the project, we identified that reframing data mining project methodology to inject new SME, and adding diagnostic techniques to develop business deliverables from that new SME, would offer the required utility against this criterion. We proposed incorporating this reframing into our proposed Strategic Analytics Methodology;
- having established that CRISP-DM has insufficient utility for introducing new subject matter expertise into the SPC project environment, we identified that reframing data mining project methodology in a way which includes opportunity and technique for introducing such new SME, will add utility against this criterion. We proposed incorporating this reframing in SAM;
- having established that CRISP-DM has a mapping technique which is useful when exploring the problem space, but is not useful once we enter the solution space, we developed a reframing of the mapping technique, which would offer the required utility against this criterion in the solution space too. We proposed incorporating this improved mapping technique in SAM;
- having established that CRISP-DM offered insufficient utility in knowledge management for the SPC project environment, not progressing past the reporting of discovered information to the business, we proposed reframing data mining project methodology against this criterion, by embedding SPM into a data mining project methodology. We proposed to accomplish this in SAM;
- we established that CRISP-DM recognises the importance of a monitor and control plan, but falls short of offering utility in the SPC environment. This affected monitor and control of both the business and technical dimensions. We proposed that data mining project methodology could be improved, by incorporating key concepts from the *concept drift* and business literature. We further proposed to demonstrate this approach in our design of SAM;
- we also identified a number of soft issues which impact on the SPC analytics environment, but under-represented in CRISP-DM. Important soft issues are the degree and nature of expert collaboration at a team level, and overcoming limiting psycho-cognitive factors at the individual expert level. Of paramount importance is the way in which CRISP-DM presents data preparation as a

technical task set only, while not also explaining how that task reduced project risk, and how it strategically supports the business deliverables. We made suggestions for optimising data mining project methodology for these and other soft factors, and proposed incorporating those suggestions in SAM.

In the next chapter, we present our data mining methodology called Strategic Analytics Methodology (SAM). SAM contains our substantial reframing of data mining project methodology for the SPC environment.

5 Chapter 5 – Developing Strategic Analytics Methodology

5.1 Supporting strategic alignment

We saw earlier that business has entered a development era, where success is not assured any more by simply adapting to changing circumstances. Now, *competitive* advantage is required for survival and growth. Competitive advantage comes from developing value faster and better than your competitors can. Value is measured in money at the highest level. Value is generated in one of three generic ways:

- increasing price;
- reducing cost; and
- increasing volume.

The things which enable the creation of value, are called the *value drivers*. The strategic hierarchy we described in Chapter 2 defines how we harness the value drivers for the business. Anything in the business which supports the development and execution of the strategic hierarchy, we therefore call value adding or *strategic* (Pyle 2004, pp.33ff.). We can therefore say that any data mining project methodology which supports in the above way, is a *strategically aligned* analytics methodology, or *strategic analytics methodology*.

We saw in earlier chapters, that data mining projects have the potential for supporting competitive advantage. We saw in the previous chapter, that CRISP-DM – as a data mining project methodology, lacks this support, because it falls short of utility in a number of criteria which are vital in the SPC cycle. This chapter builds on Chapter 4, reframing data mining project methodology from a *data-centric* approach, to a *strategic* or *business-centric* approach. We call this new methodology Strategic Analytics Methodology (SAM). We hypothesise that SAM will be suited to the SPC environment because it progressively support strategic development through data mining projects. It does this through enabling:

- ❖ breakthrough in the *perception* of business opportunities, threats, problems etc.;
- ❖ breakthrough in *cognition* or understanding about:

- the nature of the problem, opportunity etc.; and
- how we can resolve the problem, meet the threat, or realize the opportunity;

❖ harnessing the breakthrough in new organisational objectives and strategies, *constituting strategic paradigm shift*;

❖ realising the objectives through the execution of their supportive strategies, *constituting operational paradigm shift*;

❖ *assuring* the ongoing relevance of the solution under changing circumstances over time lapsed.

In following chapters, we move test the hypothesis that SAM offers SPC utility, by applying it to the data of Telco ABC's retention problem and producing an executable solution. We now present SAM in FIGURE 5.1:

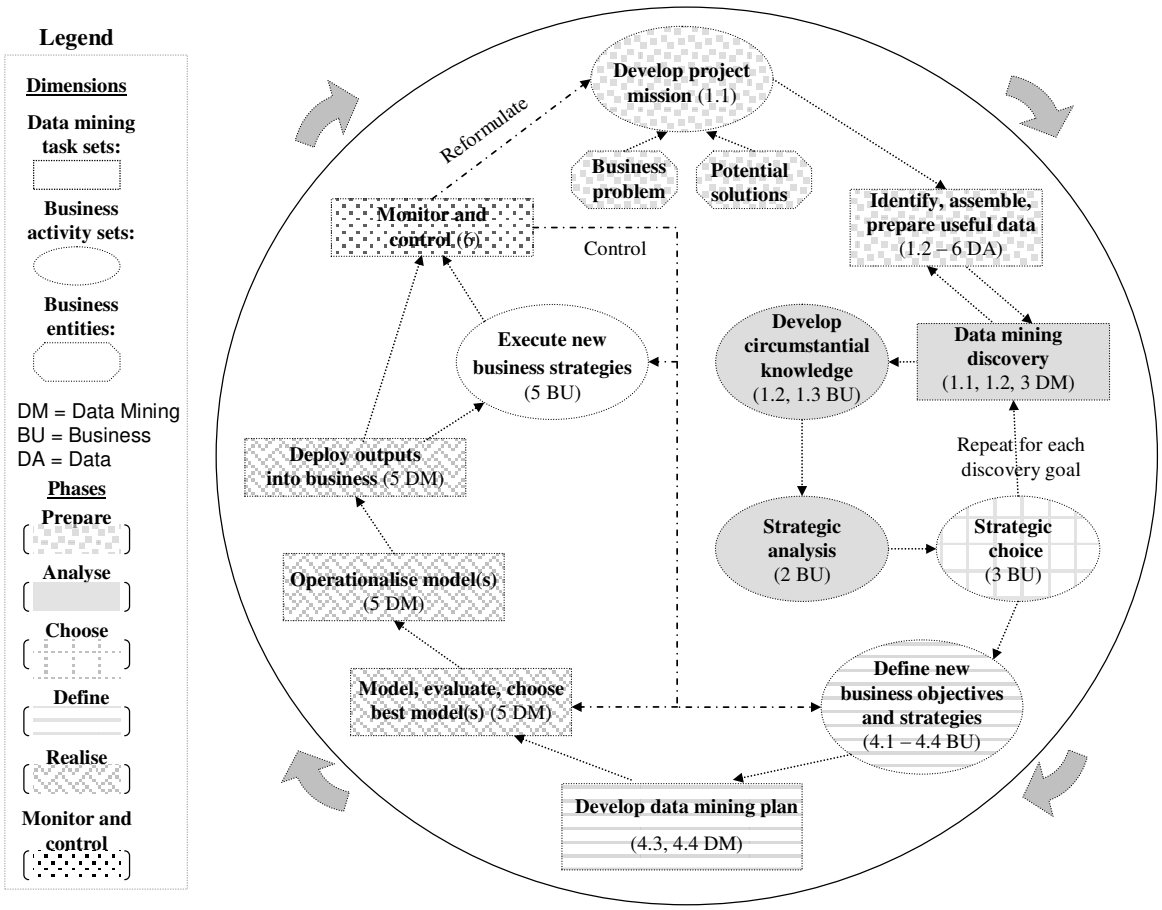


Figure 5.1: Strategic Analytics Methodology

The departure from a *data-centric* model to a *business-centric* model, is visible from:

- the absence at the center of Figure 5.1 of any data icon, in contrast to CRISP-DM's data icon at its center of (Chapman, Clinton et al. 1999-2000, Figure 2 p.13);
- the embedding within SAM of the six *Business activity sets* from SPM which create business value. They are *Develop project mission*, *Develop circumstantial knowledge*, *Strategic analysis*, *Strategic choice*, *Define new (objectives and) strategies*, and *Execute new business strategies*.

SAM accommodates four roles for data mining in the SPC environment:

- an *exploratory, analytical* application in assisting hypothesis testing about the nature of a problem or opportunity, and about the potential solutions for it;
- a *directed, analytical* application in determining and developing possible business solutions for the problem or opportunity;
- a *productive* application in supporting the *realisation* of the business solution;
- a *monitor and control* application in assuring the relevance of the business and technical solutions over time.

5.2 Dimensions of Strategic Analytics Methodology

We name the individual shapes within the SAM diagram the *elements* of SAM. It is the combining and sequencing of these that provide the strategic support. These elements are from three dimensions, which are *Data mining task sets* (rectangles), *Business activity sets* (ellipses), and *Business entities* (octagons). We represent these dimensions in the legend to Figure 5.1.

We have arranged the entities within a circle. Arrows along the outside perimeter indicate a clockwise process flow, which is similar to that of CRISP-DM. The two *Business entities* are located just below the position of the 12^{th} hour, and the arrows leading from them denote them as the starting point for the project.

In CRISP-DM, the only business activity set is *Business understanding*, the remaining task sets are data mining task sets. CRISP-DM does not present any *Business entities*.

5.3 The role of SPM in SAM

Any analytics methodology should consider the organisational circumstantial constraints (Liu 2003, p.436). We explained the utility of SPM in Chapter 2 as a tool for

producing breakthrough which is executable under the organisation's particular circumstantial constraints. For this reason we use SPM in SAM.

We indicate this integration of SPM in the data mining environment, with labels within the elements which reflect their SPM origin. The *numeric* portion of the (label) follows the numbering of the SPM elements we used before. We give the meaning of the *alphabetic* component of the label in the legend. So for instance, the label *(1.1)* in the task set *Develop project mission* means this is an application of SPM 1.1 for *both* business and technology purposes. The label *(4.1 – 4.4 BU)* in the activity *Define new business objectives and strategies*, means that we are applying SPM 4.1 - 4.4 on a business component of the project.

The two business entities of SAM are not labeled as SPM, since they are not included in SPM. *Business problem* in SAM has its origin in the operating environment of the business. *Potential solutions* is in neither SPM nor CRISP-DM, and originates from the literature study in Chapter 2. We introduce it into SAM as an enabler, representing newly injected business domain knowledge.

We also introduced the permeating enabler in the literature study called *Expert collaboration.* This enabler is practiced in all the activities and task sets of SAM. A later section gives the detail on this practice. Displaying a permeating enabler like *Expert collaboration* in Figure 5.1, will unnecessary clutter Figure 5.1. For this reason, we have not included a visual element for *Expert collaboration* in Figure 5.1.

5.4 Strategic progression

In SAM, the execution of the various business activities and data mining tasks, progresses the project strategically. This means that just as in SPM, progressing to the next entity in SAM is dependent on first having met the objectives of the current entity. We indicate the direction of the progression in SAM with directional arrows.

Because SAM is an integration of a number of dimensions, there are numerous inter-dimensional dependencies. They are between the:

- two business entities *Business problem* and *Potential solutions*, and the business activity set *Develop project mission*; between the
- business activity set *Develop project mission* and the data mining task set *Identify, assemble prepare useful data*;

- data mining task set *Data mining discovery* and the business activity set *Develop circumstantial knowledge*;
- the business activity *Strategic choice* and the data mining task set *Data mining discovery*;
- business activity set *Define new business objectives and strategies* and the data mining task set *Develop data mining plan*;
- data mining task set *Deploy outputs into business* and the business activity set *Execute new business strategies*;
- business activity set *Execute new business strategies* and the data mining task set *Monitor and control;* and
- data mining task set *Monitor and control* and the business activity *Develop project mission.*

Dependency between elements is familiar from CRISP-DM, where dependencies are also based on the achievement of goals. We saw in Chapter 4 that the goals in CRISP-DM were mostly technical and not business related. SAM expands CRISP-DM's concept of task / activity interdependency, by adding the business strategy components as dependencies.

In SAM, we distinguish *iteration* through an activity or task, from the *repetition* of the task or activity in the following way:

- under conditions of *uncertainty* about the effectiveness of the activity or task in achieving objectives or goals, we *iterate* through elements to develop effectiveness until we have achieved an objective or goal;
- under conditions of *certainty* about the effectiveness of the activity or task in achieving objectives or goals, we *repeat* elements to produce different objectives or goals.

The two data mining task sets *Identify, assemble, prepare useful data* and *Data mining discovery* are executed in an unexplored environment. This means there is uncertainty how to attaining their respective objectives. Under these circumstances, *iteration* may be required between those activity sets to achieve a certain objective. The arrow between these two activity sets therefore indicates *iteration*.

There is a repeat loop between the business activity set *Strategic choice* and the data mining task set *Data mining discovery*, which we have indicated with the arrow market *Repeat for each DM goal* in Figure 5.1. We need to repeat *Data mining discovery, Develop circumstantial knowledge, Strategic analysis,* and *Strategic choice* for each different business objective. We may however, also have to *iterate* here until we attain one specific goal.

There is a project mission *Reformulate* loop between the data mining task set *Monitor and control* and the business activity set *Develop project mission.* When the monitor and control measures fail in keeping the business and data mining solutions relevant to the *Business problem*, a redefinition of the project mission is required. Such redefinition is a *repeatious*.

CRISP-DM assumes a *project mission* reformulation (Chapman, Clinton et al. 1999-2000, Figure 2, p.13), and does not contain anything explicit about the project mission.

5.5 SAM as a Knowledge Management cycle

In Chapter 2, we identified the four KM stages and the KM activities, which we consider of importance in the SPC project environment. We saw in Chapter 2 that we can overlay SPM with KM, and that the SPC could be viewed as a knowledge management process.

We saw in Chapter 4, how CRISP-DM does not include the KM activities which are required by the SPC environment. After considering the effect of that on the possible results from a project using CRISP-DM, we reflected that a data mining project methodology should include complete KM utility.

We present in Figure 5.2 our reframing of the KM utility of data mining project methodology as it is contained in SAM:

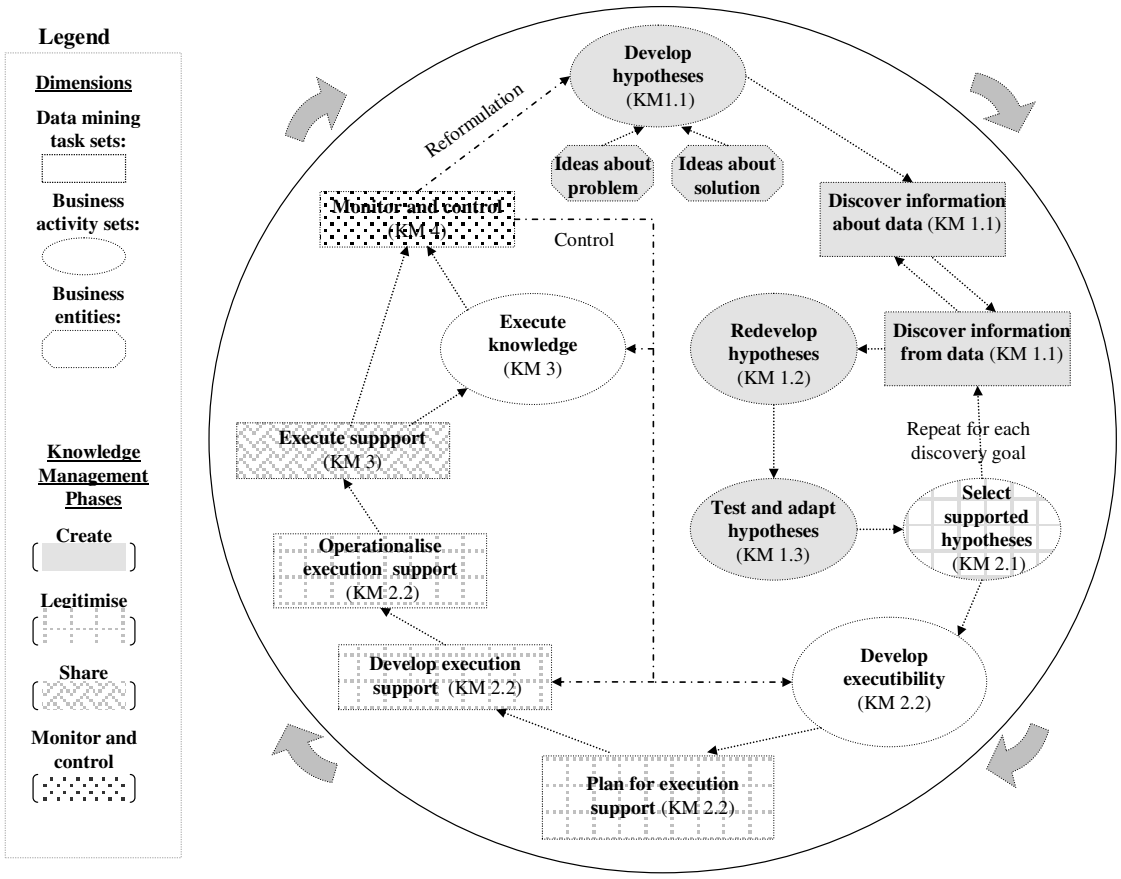


Figure 5.2: Knowledge management in SAM

Figure 5.2 maintains the lay-out of Figure 5.1. However, we have replaced the descriptive labels of the SAM elements with labels relating to their KM role within SAM's KM process. In the legend of Figure 5.2, we describe the four phases of our modified KM process, as we have incorporated it in SAM. They are the familiar *Create*, *Legitimise*, *Share*, and *Monitor and control* we saw in Chapter 2.

Note how the KN phases – which we have given distinct background fills - interweave between the different SAM elements, effectively tying together the activities with the SAM data mining project methodology. We next discuss the KM phases within SAM, and the function of each of the SAM elements.

5.5.1 Create

KM's *Create* phase starts with *Business problem* and *Potential solutions*, and ends with SAM's business activity set *Strategic analysis*:

- *Business problem* and *Potential solutions* are the preconceptions and ideas underlying our hypotheses. They are pre-existing input into the KM process, and are not numbered;
- in *Develop project mission,* we develop hypotheses about the nature and extent of the business problem or opportunity, its root causes, its solution, and the operationalising of the solution. In these we depend on new SME at hand. Based on those hypotheses, we set the KM goals for the project (Pyle 2003, p.368). We therefore labeled this business activity in Figure 5.2 with *Develop hypotheses (KM 1.1).* The attainment of those goals is dependent on the following refuting, supporting, or reframing of the hypotheses, using data mining as the supporting technology and human cognition as the main tool;
- in *Identify, assemble, prepare useful data* we discover information *about* the data, and develop untested knowledge about that data, for developing the hypotheses underlying the SPC goals. Because the hypotheses about the data are yet untested by data mining, we labeled this data mining task set with in Figure 5.2 with *Discover information about data (KM 1.1)*;
- in *Data mining discovery* we discover the information *from* the data, which is required for redeveloping and testing the hypotheses underlying the SPC goals, under the organisation's data, commercial, and operating circumstances. We are still *contributing toward* the development of knowledge which is untested for organizational circumstances, and label this data mining task set with *Discover information from data (KM 1.1)*;
- in *Develop circumstantial knowledge,* we develop all the discovered information into their respective, yet untested knowledge profiles, which we estimate are required for hypotheses testing under the organisation's circumstances. We therefore label this business activity set with *Redevelop hypotheses (KM 1.2)*;
- in *Strategic analysis* we *analyse* the respective knowledge profiles, and support, or reframe the hypotheses under the organisation's data, commercial and operational circumstances. This is the hypotheses testing activity of SAM, and we label it with *Test and adapt hypotheses (KM 1.3).*

5.5.2 Legitimise

KM's *Legitimise* phase in SAM starts with the business activity *Strategic choice*, and includes the following business activities and data mining tasks:

- in *Strategic choice* we *interpret* the hypotheses testing and reframing, and either *reject* a hypotheses, or *select* the best-supported or reframed hypotheses under the circumstances. This selection is a legitimisation of the executibility of the knowledge within the hypotheses under the organisation's circumstances. Knowledge, which is not executable under the organisation's circumstances, cannot be legitimised. We have therefore given it the label *Select supported hypotheses (KM 2.1)*;
- in *Define new business objectives and strategies* we further legitimise the selected knowledge, through now constituting its executibility for the organisation's unique circumstances. We labeled this business activity with *Develop executibility (KM 2.2)*;
- in *Develop data mining plan,* we plan the use of data mining for supporting the executibility of the legitimised knowledge. Since we are still supporting executibility, we label this data mining task set *Plan for execution support (KM 2.2)*;
- in *Model, evaluate, choose best model(s)* we develop that technical executibility support. We label this data mining task set with *Develop execution support (KM 2.2)*;
- in *Operationalise model(s)* we support the knowledge legitimisation, by operationalising the technological executibility support. This is the last data mining task set with the label *Operationalise execution support KM 2.2*.

5.5.3 Share

KM's *Share* phase spans the data mining task set *Deploy outputs into business*, and the business activity set *Execute new business strategies*:

- in *Deploy outputs into business*, SAM assists with the sharing of knowledge, through executing the data mining support into the business systems and structures who depend on the information. The data mining has therefore

become one of the systems within in the 7-S framework. We therefore label this data mining activity with *Execute support (KM 3)*;

- SAM's *Execute new business strategies* shares knowledge through executing the business solution, which is based on the knowledge. We therefore labeled it with *Execute knowledge (KM 3).*

5.5.4 Monitor and control

KM's *Monitor and control (KM4)* phase resides in the SAM data mining task set with the same name. This task set contains analytical, interpretive, and executive activities, and the information upon which they are based, assuring the ongoing legitimacy of the knowledge over time lapsed.

From the above we see that SAM allows for the development of a data mining project, which supports all the phases of the SPC KM process. In doing so, it overcomes a major limitation of CRISP-DM in this regard. We validate this claim in a later chapter.

5.6 Strategic Planning Cycle phases of SAM

The point of departure in the SPC data mining project environment, is total uncertainty about even the potential for solving the business problem or meeting the business opportunity, by mining the organisation's data. The reader will recall that Risk Management is primarily concerned with *reducing uncertainty about unwanted events*. Risk Management reduces uncertainty through *managing* risk. By analogy of Risk Management, in the project management environment we strive to manage the risk which stands in the way of attaining the *wanted event* or SPC goals / business deliverables. In the SPC data mining project environment, we use SMP as a progressive risk management tool.

The SPM we presented in Chapter 2 had two phases, being *Planning,* and *Realisation*. In the uncertain project management environment, we require more visibility about progress than what is afforded by just two phases. To improve visibility about progress toward the project's goals, we have given SAM six phases. They are *Prepare*, *Analyse*, *Choose, Define*, *Realise*, and *Monitor and control*.

We gave each phase a distinctive background fill in Figure 5.1, and reflect this in the legend of Figure 5.1. Because SAM is a combination of SPM business activity sets and technical data mining task sets, the phases span the boundaries between the various

SAM elements. The names of the phases express the grouped roles of their entities in the strategic value chain. We discuss the role of each entity within its phase in detail in a following section of this chapter.

Summarising, SAM develops certainty about attaining a project's goals under conditions of organisational uncertainty, and visibility on the progress to those goals.

5.6.1 Prepare

Prepare spans the two entities *Business problem* and *Potential solutions,* one business activity set *Develop project mission,* and one data mining task set *Identify, assemble, prepare useful data.* We have given these four entities the same confetti fill.

In the *Prepare* phase, we define the *business deliverables or SPC goals* of the project, and determine the *potential for attaining them* through preparing the organisation's data for data mining discovery. Spoken from a TQM perspective, the *potential utility of the data for the business purpose* is defined in this phase.

5.6.2 Analyse

The *Analyse* phase spans the data mining task set *Data mining discovery*, and the business activity sets *Develop circumstantial knowledge,* and *Strategic analysis.* We have given these elements a smooth fill in Figure 5.1. This is where cognitive discovery and the development of insight resides (Pyle 1999, p.23). The strategic support of the project during the Analyse phase is:

- developing *potential* into a number of *possibilities* for supporting the project's SPC goals with data mining, given the organisation's circumstances;
- *quantifying the* possible monetary, strategic, and operational *benefits* of the SPC goal support scenarios.

This establishes a best-case worst-case scenario for the project's support of the business, given the organisation's circumstances. From a Project Risk Management perspective, at the end of this phase we removed some of the uncertainty about the definition of the *wanted event*, the business deliverables. We have also determined a range for the project's *event impact* or it is ROI. There still is total uncertainty as to *event likelihood* or the executibility of the business deliverables.

5.6.3 Choose

The phase *Choose* is contained in the one business activity set *Strategic choice*, *with* the distinct grid fill. The strategic support in this phase is:

- identifying the best possibilities for supporting the business under the organisation's circumstances. This eliminates the remaining uncertainty about defining the SPC goals; further
- locking in the best possibilities this way, has the benefit of reducing uncertainty about the *impact* of the project in dollars and cents.
- giving the project manager confidence in justifying the continued allocation of resources toward pursuing the following project phases.

Uncertainty still remains as to the *event likelihood* or execution of the business deliverables.

Spoken in TQM terminology, *Choose* constitutes the *possible utility for purpose* of the project, at the most forward link possible, before any further effort is spent in the project pursuing inexecutable SPC goals.

5.6.4 Define

The *Define* phase of SAM spans two elements. The first is the business activity set *Define new business objectives and strategies,* and the other is the data mining task set *Develop data mining plan.* We give them the horisontally lined fill to indicate their phase membership. The strategic support in this phase, is harnessing the developed knowledge into a novel, executable business solution, and aligning the data mining technology to support this execution. We achieve this harnessing through:

- developing a new business objectives and strategies hierarchy; and
- defining the productive data mining project, which supports the realisation of the new business objectives and strategies.

From a Project Risk Management perspective, the *Define* phase reduces risk by:

- establishing how the organisation will realise the *wanted event*; and by
- reducing uncertainty about harvesting the benefits (*event impact)* of the project.

Spoken in TQM terminology, the *Define* phase constitutes the *actual utility for purpose* of the project, at a time before the organisation allocates resources for the execution of an unworkable business solution.

5.6.5 Realise

Realise spans four entities. They are the three data mining task sets - *Model, evaluate, choose best model(s)*, *Operationalise model(s),* and *Deploy outputs into business* - and the business activity set *Execute new business strategies.* We indicate the phase membership of these entities by giving them a common weaved background. Together they function as SPM 5 *Implement and deploy new strategies (Execute)*, and we have labeled them in Figure 5.1 with the *5.* This means that SAM recognises that data mining is another tool in the organisation's 7-S business strategy implementation toolbox.

The strategic support in *Realise*, is realising the business solution and harnessing the returns from the project. This comes through executing the business solution and its supporting data mining project.

Executing the business solution and its supporting data mining project, is the ultimate reduction of uncertainty about the *likelihood* of the *wanted event*. However, uncertainty remains as to the effectiveness and efficiency of the execution over time, or the *event impact*. That uncertainty is removed in the next phase of SAM.

5.6.6 Monitor and control

This phase is contained in the data mining task set *Monitor and control*, which we have given a speckled fill. The strategic support of *Monitor and control*, is ensuring the ongoing relevance of the data mining and business solutions over time.

From a Project Risk Management perspective, this phase reduces uncertainty about the effectiveness and efficiency of the solutions over time lapsed, or the *event impact*. According to the TQM approach it measures the effectiveness and efficiency of the *utility*.

5.7 SAM as a reframing of CRISP-DM

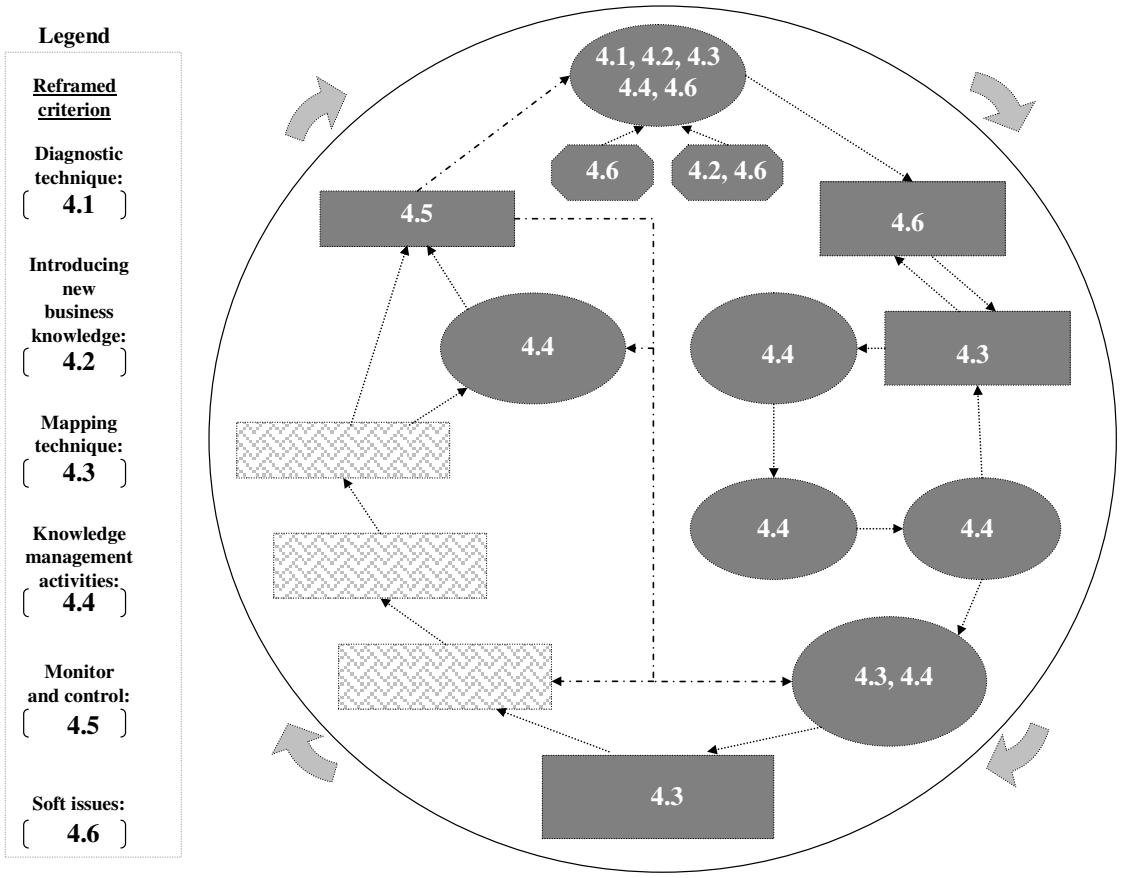


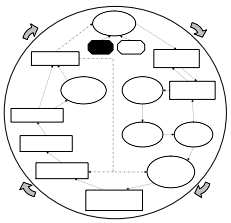
Figure 5.3: SAM's reframing of CRISP-DM

In Chapter 4, we reflected about reframing CRISP-DM to improve its SPC utility against six criteria. We said there that we would constitute the required reframing in SAM. In this section we visually represent the reframing in Figure 5.3. The section following the figure contains the detailed discussion of the reframing. The names of the entities in Figure 5.3 can be corroborated with Figure 5.1. The numeric labels of the entities in Figure 5.3 refer to their matching section in Chapter 4.

5.8 Functioning of the SAM elements

In this section, we discuss the content and function of each SAM element and its strategic support of the business deliverables. This section offers substance on how SAM reframes data mining methodology along the six main utility criteria we listed in Chapter 4. There also is more detail on some of the other issues discussed in Chapter 4.

5.8.1 Business problem



5.8.1.1 What *Business problem* is

In Chapter 4, we reflected that CRISP-DM needs reframing about soft issues. In *Business problem* we start with this reframing, and have labeled Business problem with the matching Chapter 4 section 4.6. It is the *uncovering of the framework within the existing business model* (Pyle 2004, p.128).

At the lowest level of definition, the business problem may be an *unchallenged* and simple *feeling of uneasiness* (Pyle 2004, p.63) by the owner that *something is not right* or that *something is happening*. This can be particularly true if the *Business problem* is an emerging commercial or operational threat or opportunity. By its nature such a feeling would be prejudiced. In other situations the business problem may consist of a defined pre-project schema. We have already seen how such a schema consists of untested perceptions and expectations. By their nature they are prejudiced. We demonstrated in our ERP supply chain example, that the business problem may also be defined by an organisation's existing *objectives and strategies hierarchy*, or the *existing paradigm*.

We here reframe CRISP-DM for the soft issue of the *perceptions* about the project *status quo*. Particularly we make the organisation aware that any approach to a project always is prejudiced and preconceived. These prejudices and preconceptions have a limiting effect on the breakthrough potential of a project. In the project environment, as in life situations, we are faced with the irony that our preconceptions and prejudices limit the potential for breakthrough that we can perceive.

In the case of Telco ABC, we saw how that even a well-defined *solution* might be part of the problem. That ineffective solution was also captured by their existing retention management strategies. In other cases the business problem may be a combination of the above manifestations, each manifestation located within a business unit. The provision within SAM for considering *failing strategies* as part of the *Business problem*, further is a reframing of CRISP-DM's defining the business problem in terms of *Business objectives* and *Business success criteria* only. This reframing in SAM, extends the value of data mining project methodology in the SPC environment past merely *improving our understanding* about a problem or opportunity, to also include

developing the solution for the better understood problem or opportunity. This better reflects the multiple analytical goals of SPC projects (Berthold and Hand 2003, pp.2ff.).

SAM retains the *objectives* of the existing paradigm as part of the problem, or even as the whole problem. Further, where no *objectives* exist, then stage one and two of the *pre-project schema* express this problem dimension. This makes SAM useful for SPC applications, where the improved understanding about the problem or opportunity, must first result in the defining of new objectives about it, before we can even think about a solution for it.

Business problem further contains the preconceptions about the magnitude of the problem in dollar terms. In *Business problem* we make the organisation aware that they have to objectively quantify the ROI of the project, before entering the execution phase of the project. *Business problem* constitutes a forward linkage of TQM's *utility for purpose*. This is a big reframing of CRISP-DM's provision for mid-stream abandonment of a data mining project in its *Evaluation* task set, when discovery is made about the project's ROI.

5.8.1.2 *Business problem's* strategic purpose

From a TQM perspective, *Business problem* formally expresses the potential opportunity for the project. This should result in a reduction in misunderstandings and resistance to change in later phases, and have a motivational effect on the experts.

From a Risk Management perspective, *Business problem* is the *unwanted event*, about which there exists uncertainty about its *event likelihood*, and about how they can be eliminated, reduced, or isolated. The *event impact* is the project's ROI, and in the case of Telco ABC we calculate it using Formula 1.

5.8.1.3 How *Business problem* supports

Business problem supports by raising the awareness of key individuals within the organization, about acknowledging the limiting affect of the *status quo*. One author has described this step as framing a *model of preconceived notions* (Pyle 2004, p.126), devoting a whole chapter to it (Pyle 2004, Chapter 5).

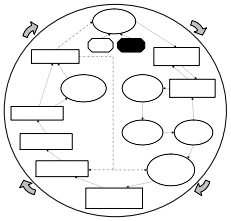
If there is no pre-project schema, *Business problem* achieves its purpose by defining a *pre-project schema*. If a *pre-project schema* is extant, we then generate an *understanding* among key individuals about the common and differential

preconceptions and expectations within the organization, on what the project should deliver. We do not seek agreement which eliminates this diversity at this stage.

We achieve this defining and understanding through applying facilitating techniques like interviews, group plays and informal brain storming. This brings all the preconceptions into the open for common digestion and understanding. Understanding typically is required by the project power brokers or sponsors e.g. CEO, CIO, CFO, the project leader(s), the subject matter experts, and the analysts.

Note that we are not yet injecting new SME into the project environment. We are simply casting preconceptions, expectations etc. into a 5-dimensional pre-project schema framework. The structure of the pre-project schema and its common understanding, lays the foundation for the formulation later of agreed to business deliverables. For this reason, we recommend that the facilitation be directed toward the pre-project schema structure.

5.8.2 Potential solutions



5.8.2.1 What *Potential solutions* is

In Chapter 4, we reflected that the reframing of CRISP-DM should include the injection of new SME into the business problem. *Potential solutions* is the entity within SAM, which constitutes this reframing. We have therefore labeled *Potential solutions* with *4.2* in Figure 5.3. We move test this reframing later on Telco ABC's retention problem.

In KM terms, *Potential solutions* is the *materiel* which enables formulating the *hypotheses* with breakthrough potential, about the causes of the problem or opportunity, and about the business solutions for them (Cooley 2003, p.611).

SAM recognises the injection of six types of business knowledge into the project. The first two types of knowledge are propagated by (Nonaka 1994), and the others are of our own making:

- *explicit* - facts and concepts which are widely recognised by domain experts to produce breakthrough on problems similar or related to the *Business problem*;
- *tacit* - or experiential and judgmental – business knowledge relating to implementations of solutions, which have produced breakthrough on problems similar or related to the *Business problem*;

- *theoretical academic hypothesis* about solutions which have produced breakthrough on problems similar or related to the *Business problem;*
- *new knowledge by analogy or proxy* to the *Business problem*, acceptable where none of the other types of knowledge is available. In the event where the organisation's managers suspected that the limited ability of the traditional analytics function also contributes to the business problem, the organization may also consider introducing:
- knowledge about *data mining*; Further, the organisation may consider introducing
- knowledge about data mining project methodology - like SAM - to enhance the advantage from a SPC project.

SAM is not biased as to the internal or external origin of any new knowledge. For that matter, the knowledge could have been developed through in-house *collaborative expert cognition,* or through the experts' own academic or professional development.

SAM has some requirements on the *nature* of the injected knowledge:

- it must be *novel* to the organisation implementing SAM, and must have *fundamental depth* about *cause and effect.* This is to enable the unlocking of the existing paradigm;
- it must produce sufficient *overlap* between the business SME of the analysts and the subject matter expert to enable expert collaboration.

By analogy of the ERP industry's maturing experience, we prognose that as the application of data mining in the SPC application matures, there will ultimately be a convergence of both subject matter expertise and technical data mining expertise, in individuals of extra-ordinary competence.

5.8.2.2 *Potential solutions's* strategic purpose

Its first purpose is assuring that the existing paradigm is unlocked. Further, from a soft issues perspective, the purpose of *Potential solutions* is reducing the risk associated with change management. This is a reframing of CRISP-DM for soft issues, and we reflect that in the label *4.6* in Figure 5.3.

Viewed from a TQM perspective, *Potential solutions* is the *potential utility* of the business solution, which could result in the attainment of the projects SPC goals. From a Project Risk Management perspective, *Potential solutions* constitute the *potential* remedial action the organisation can undertake, to reduce, eliminate, or isolate the risk associated with the *unwanted* event *Business problem.*

5.8.2.3 How *Potential solutions* supports

From TQM and KM perspectives, the *Potential solutions* presents to the stale problem situation the *purpose*, and the *potential utility for purpose,* in understanding the business problem or opportunity, and in formulating a competitive solution for it. The *utility* resides in the cause and effect *know-how* of the knowledge, which is contained in the knowledge. The *purpose* resides in giving *directional guidance* to the KM process. *Potential solutions* is effective in:

- in SAM's *Create* KM phase:
  - providing the *materiel, which* helps with the formulation of the hypotheses. This happens in the business activity set *Develop project mission*. There, it also assists with purging and updating the expectations and strategies within the *pre-project schema*, resulting in the optimal understanding of *Business problem*, and the development of optimal SPC goals for the project*;*
  - establishing the *relevance* of data in the data mining task set *Identify, assemble, prepare useful data;*
  - directing the *Data mining discovery* towards the discovery of information which supports the hypotheses, and helping with the recognition of that information once it has been discovered;
  - directing the development of knowledge in *Develop circumstantial knowledge*, and the hypothesis testing in *Strategic analysis*;
- in SAM's *Legitimise* KM phase:
  - enabling the selection of the best supported hypothesis in *Strategic choice*;
  - guiding the development of executibility in *Define new business objectives and strategies* as an objectively optimal, novel and executable business solution;

- defining a data mining plan in *Develop data mining plan* which supports the execution of the new business solution;

❖ in SAM's *Share* KM phase, assuring that quality of the execution of the business solution in *Execute new business strategies* and *Deploy outputs into business*;

❖ in SAM's *Monitor and control* KM phase, identifying the key concepts to monitor and control.

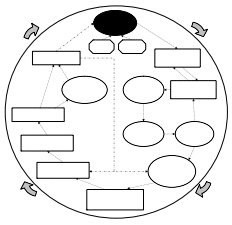
Human perception and cognition are central to all the KM and technical activities in SAM. We explained in Chapter 2, how increased levels of knowledge enhance the perceptive and cognitive processes, allowing the unlocking of an existing paradigm. New SME then presents a new *pattern template* allowing for breakthrough (Pyle 2004, p.39).

Our corporate project management experience is that new knowledge proactively ameliorates change management problems:

- by ensuring the knowledge overlap required for expert communication and collaboration; further by
- serving to motivate the experts through increasing the Pygmalion effect we explained before.

This results in creative synergy between the experts for breaking through existing perceptive and cognitive boundaries of the *Business problem*.

5.8.3 Develop project mission



5.8.3.1 What *Develop project mission* is

In Chapter 4, we reflected that the reframing of CRISP-DM should include:

- improved diagnostic technique for formulating the project's SPC goals;
- the injection of new domain knowledge;
- bridging technique;
- knowledge management activities;
- soft issues; and
- monitor and control.

Develop project mission is an entity within SAM, which constitutes reframing against all these criteria. We will move test this reframing later on Telco ABC's retention problem. In this section, we will explain at various opportunities how the relevant reframing is achieved.

Corporate Mission Component	Concept contained in the component	SAM Project Mission Component
Primary market and service	The *what* being pursued by the organisation or	The project's *binding* SPC goals or business deliverables
Technology	The *how* of that pursuit by the organisation or project	The *desired* strategies for attaining the SPC goals. Strategies cover the use of data mining as the analytical tool, and the use of a data mining project methodology e.g. SAM
Geographic region	*Where* the need resides	Which function(s) in the organisation is (are) affected and therefore involved in the project
Survival	The *when* or temporal scope for the mission	The operational *duration* and *frequency* of the outputs produced by the project

Table 5.1: Components of the project mission

Develop project mission first is the SAM business activity, in which we formulate the hypotheses which are required by SAM's KM process. It contains reframing about the KM activities, and we therefore labeled it *4.4* in Figure 5.3.

Second, *Develop project mission* produces the *mission* for the SAM project. We have labeled it *1.1* in Figure 5.1 to match it with SPM's *Mission* (SPM 1.1). It is an expression of the SPC requirements on the project (Pyle 1999, p.29). We identify the key technology application factors in the business problem (Liu 2003, p.437). It is a *mission* and *not a plan*, since we have not yet factored in any business and technology

circumstances and their effect on the project. We referred before to what these limiting circumstances may be.

The *project mission* could represent all five of the KM dimensions, as we saw them in the pre-project schema. In each dimension present, there are four components, which we explain in Table 5.1 by proxy of the dimensions of SPM's *Mission* (SPM 1.1)

The SPC goals in the *project mission* are *binding*, while the strategies with the supporting technology are *desired*. The *possible* strategies with the supporting technology are first *estimated* later, when formulating the *data mining mission* in the task set *Data mining discovery*, and the real, possible strategies *confirmed* by the exploratory mining there.

5.8.3.2 *Develop project mission's* strategic purpose

We saw in Chapter 1, how that the 5 stages of the *pre-project schema* presents a template of all the KM dimensions, which need to be covered by any SPC project, to produce an executable business solution. The purpose of *Develop project mission* is to develop each dimension of the *pre-project schema* - which was developed in *Business problem* above – into its four mission components. This happens under the influence of the SME introduced in by *Potential solutions*, and takes us from the unchallenged project *status quo*, to defined business deliverables.

The second purpose of *Develop project mission,* is to set the binding SPC goals for the project, which express the pursuit of understanding about the problem / opportunity, and the perceived objectively optimal, novel, and executable business solution for the business problem or opportunity. From a project TQM perspective, this is the *forward linkage of the purpose* with the project. Complementary, *Develop project mission* is the *forward linkage of utility* for that purpose. Here the *utility* is the desired strategies within the *project mission*, for attaining the SPC goals. The SPC goals are the *desired* within *Potential solutions* we want to achieve with the project.

Behaviorally, the purpose of *Develop project mission* is to establish *agreement* between the collaborating experts, about all four mission components in each dimension. This first has a motivational function. Further, such agreement avoids misunderstandings or even abandonment at later stages of the project. This is a reframing of CRISP-DM in the soft issues criterion, therefore the *4.6* label in Figure 5.3.

5.8.3.3 How *Develop project mission* supports

Develop project mission meets its purpose primarily through the application of a diagnostic technique for formulating the SPC goals within the *project mission*. A key feature of this diagnostic technique is introducing the new SME which is required for developing the *pro-project schema*. This is a reframing of CRISP-DM in both the diagnostic technique and new SME criteria, and we therefore labeled it with *4.1* and *4.2* in Figure 5.3.

The first important difference between the resulting *project mission* and the *pre-project schema*, is that the SPC goals in the *project mission* are binding while those in the *pre-project schema* are not. Second, the technology strategies for pursuing the SPC goals, may also have to be updated in the *project mission* to reflect a more desirable use of the supporting technology, now that the data mining engineer has brought his or her knowledge to the project. Third, the *project mission* has two additional components compared to *pre-project schema* – the *where* and *when*.

The fourth difference is that the while the *pre-project schema* does not constitute expert agreement, the project mission does. In *Develop project mission* we have to develop the various understandings accompanying the pre-project schema, into converged agreement about the business deliverables. The last difference is that the agreement and business deliverables must be permeated by the new technical and SME introduced by *Potential solutions*.

The following diagnostic technique is applied to each of the five levels (SCHEMA 1-5) of the *pre-project schema*. The singular *hypothesis* can be replaced with the plural *hypotheses* at any level:

2 *overlay* the *hypothesis* which underlies the *goal* of the *pre-project schema* with the hypothesis represented by new SME, and with technical data mining expertise (Pyle 2004, p.62);

3 *analyse* the overlay using *expert reflection-in-action* techniques, with the purpose of diagnosing any need for reframing that *hypothesis,* better to reflect the hypothesis of *Potential solutions* and data mining. Where the need for reframing is diagnosed;

4 *reframe* and *reformulate* that *hypothesis* to better reflect *Potential solutions'* insights about the business problem / opportunity and solution, and the possibilities presented by data mining as analytical tool;

5 *analyse* the new hypothesis – using *expert reflection in action* – and *formulate* a SPC goal which best supports that hypothesis;

6 repeated the above diagnostic technique for each of the remaining stages of the KM requirements within the *pre-project schema*. The extent of *any* SPC project is determined by the number of stages for which we formulate SPC goal(s).

We have to bear in mind that the potential use of data mining technology is not to the exclusion of other analytical techniques. We saw in Chapter 4 how strategies consisting either of traditional analytical techniques, or non-analytical techniques might suffice in the earlier stages of the project. This is the point in defining the project, where the organisation needs to decide whether to pursue some or all of the business deliverables with technologies other than data mining. This constitutes TQM's forward linkage of the *utility* of the supporting technology.

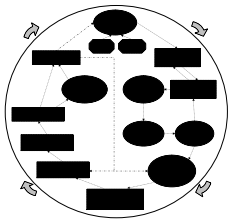
This further means that some technology strategies can be executed even while developing the *project mission*. The example of this with in Telco ABC problem, is where we define the churn event and date during *Develop project mission.* In other cases it means we execute strategies during *Identify, prepare, assemble useful data.* The example for this in the Telco ABC project, is where we do the reverse engineering of the churn processing protocol, and the identification of which business segment we want to target, during the data preparation.

The reframing of data mining project methodology in the above way, allows for the defining of SPC goals, which better reflect an organisation's needs from a SPC data mining project. It is a useful technique for defining a project with SPC goals anywhere in the spectrum of KM requirements, from simply exploring the problem / opportunity space, right through to developing a competitive solution.

From the above, it becomes apparent that further this diagnostic technique is a major tool for establishing situational understanding and agreement between experts, about what the organisation wants from the data mining project.

The diagnostic we describe above is not exclusive of other diagnostic techniques or tools; rather, it sets a minimum standard for diagnosis. We anticipate situations which could require advanced diagnostic techniques like those propagated in the business literature and by Pyle (2004, pp.154, 534ff).

5.8.4 Expert collaboration



5.8.4.1 What *Expert collaboration* is

Expert collaboration is a reframing of CRISP-DM in the soft issues criterion. This introduces into the data mining project environment, an enabler we presented in the literature study. It is not represented in any of the figures, since it is a permeating condition. Were it presented in Figure 5.3, we would have labeled it *4.6.*

Collaboration refers to the *style* in which experts exercise their individual professional efforts for attaining the SPC goals. The requirement for such a style becomes apparent, when we contemplate the complexity of the interweavedness of the experts' individual effort in all the entities within SAM. The business subject matter expert's professional effort is exercised in:

- the KM *Create* phase as:
  - driving the optimal defining of the *Business problem*;
  - introducing new domain business domain knowledge in *Potential solutions*, and mastering relevant data mining concepts;
  - driving the developing of an optimal *project mission* through hypothesising and goal formulation;
  - contributing toward the identification of the relevant data in *Identify, prepare, assemble useful* data;
  - assisting with the recognition and interpreting of relevant information in *Data mining discovery*;
  - knowledge development and hypothesis testing in the remaining business activities;
- the KM *Legitimise* phase as:
  - identifying the best supported hypotheses and selecting them;

- developing them into executibility business solutions;
- assisting with the application of the bridging technique in *Define new business objectives and strategies*;

- the KM *Share* phase, using the deployed information to design the execution of the business solution in *Execute new business strategies*, and to oversee that execution;
- the KM *Monitor and control* phase:
  - identifying the relevant business concept which require monitoring and control; and
  - contributing to the development of the *monitor and control plan*.

The data mining expert's professional effort is exercised in:

- SAM's *Prepare* phase:
  - contributing toward the expectations about the *Business problem*;
  - mastering the relevant SME concepts in *Potential solutions*;
  - contributing to the hypothesising and goal formulation in *Develop project mission*;
  - developing strategies for supporting the SPC goals in *Develop project mission;*
  - determining the *utility* of the data for supporting the *project mission*. We see later that *utility* includes *relevance*;
- SAM's *Analyse* phase:
  - developing the technical data mining approach in *Data mining discovery* and direct the mining to best support the discovery of information which supports the *project mission*, and controlling that support;
  - assuring that the knowledge development in and hypothesis testing best reflect the discovered information, and assisting the analysis of the tests with further analytics;
- SAM's *Choose* phase, presenting the technical dimensions of the legitimisation and assisting with the choice;
- SAM's *Define* phase:

- contributing to the executibility requirements of the business solution in *Define new business objectives and strategies,* and identifying the link upon which to base the mapping to the technology plan;
- applying the mapping technique to formulate the best possible data mining plan;

❖ SAM's *Realise* phase, executing the data mining task sets in a way which best supports the knowledge sharing requirements of *Execute new business strategies;*

❖ SAM's *Monitor and control* phase:

- recognising the relevant business concept which require monitoring and control;
- identifying the data mining informational concepts which support those; and
- developing the *monitor and control plan*.

The above lists present a complex picture of the interweavedness of the experts' individual effort. This complexity needs to be managed in order to produce results. A collaborative style of cooperation is an effective tool for managing such complexity.

5.8.4.2 *Expert collaboration's* strategic purpose

Expert collaboration adds business value in a number of ways in SAM. In TQM, terminology first, *Expert collaboration* assures the systematic execution of the *forward linked utility of purpose* of each SAM entity, before proceeding to the next entity.

Expert collaboration is a behavioral risk management technique, reducing risk of resistance to change, which is often associated with projects that span more than one function within the organisation. Motivation is another behavioral purpose. We explain this last statement in the next section.

5.8.4.3 How *Expert collaboration* supports

The TQM assurance of the systematic execution, is achieved through *agreement* about the sufficiency of the output of each task or activity in supporting the *project mission*, before proceeding to the next task or activity.

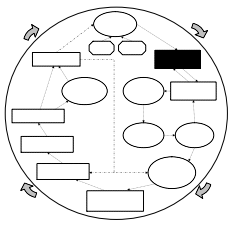
The collaborative *style* unlocks synergy between the experts, which results in creativity, and the consequent unlocking of existing paradigms. Collaborative style is achieved through the experts combined *reflection-in-action* during each SAM entity to:

- stimulate each other's awareness response where a relevant signal falls outside the perceptive bounds of the other expert;
- enhance each other's conscious meaning appraisal of interest where one expert fails to make the meaningful appraisal;
- complement each other's cognitive blind spots and cognitive weaknesses during cognition.

The *behavioral* purposes are achieved in the following ways:

1. reducing the behavioral risk of resistance to change, through communication between the experts. The candidate has seen this approach prove itself in the corporate change management environment;
2. *accountability* within the domain of expertise, refers to the professional implications for an expert, if he/she fails or succeeds in attaining the SAM Mission;
3. *mutual support* is the concern each expert has about the other expert's effort, agreement, and professional commitment;
4. an *act of the will* toward personal commitment;
5. the *Pygmalion effect* (Gibson, Ivanchevich et al. 1991, p.134) generated from the above four points, assures the achievement of the *motivational* purpose.

5.8.5 Identify, assemble, prepare useful data



5.8.5.1 What *Identify, assemble, prepare useful data* is

Identify, assemble, prepare useful data is the task set of SAM, where we do everything about, and to the data, which is required to make it suitable to, and accessible for, the data mining.

This is the first entity of SAM which produces technical output. This task set completes the *Prepare* phase of SAM. We use SPM as a tool *within* this task, and have therefore given this task set the label *1.2 – 6 DA* in Figure 5.1.

We did not identify specific issues in Chapter 4 about CRISP-DM relating to data preparation. In SAM however, we are adding in business-valued dependencies between the elements of the project methodology. This was in addition to CRISP-DM's technical

dependency between tasks. In this section, we modify the data preparation task for this business-valued dependency.

The technicalities of data preparation and assembly are well established in the data mining literature, and we are at pain not to repeat those here. Examples are (Pyle 1999), (Westphal and Blaxton 1998, Section II) (Groth 1998, Chapter 2) (Han and Kamber 2001, Chapter 3).

This is a lengthy section of SAM, and we justify that as follows:

- first, it is generally accepted in the data mining literature, that establishing the usefulness of the data takes up to 70% of project cost (Liu 2003, p. 440). Because of the time consuming nature of this task, there is evidence in the literature that a business case is required for justifying the amount of time spent on this task set in mining projects (Pyle 2003, p.374) (Pyle 1999, p.89). Our approach supports the business case, first by assuring the business audience of the progressive business value added in this activity, and second by casting that into a familiar TQM framework of *forward linkage* of *utility for purpose;*
- second, it is important to assure that the initial *Data mining discovery*, is done on data for which the data assembly and preparation strategies are practical on a regular, productive mode. This allows for the economic sustainability of solutions within the *Realise* phase of SAM;
- third, some of the project's non-data mining strategies may need to be executed in this exploratory data environment.

5.8.5.2 *Identify, assemble, prepare useful data's* support

The TQM purpose is to *define the utility* of the data, for data mining toward the project's SPC purpose. The Project Risk Management purpose is to *reduce the uncertainty* about achieving the business deliverables, *by constituting the utility* of the data for that purposed data mining.

From an SPM perspective, *Identify, assemble, prepare useful data's* purpose is to establish the *possibility* within the data *materiel*, for supporting the *project mission*. *Utility* and *possibility* therefore have the same purpose, and technical and hypothetical dimensions. This means that *utility* and *possibility* are the same concept in this

environment, just viewed from different perspectives. For that reason, we use the term *utility* in this section as inclusive of *possibility*.

In SAM, we identify four criteria for the data's *utility*:

1. the *relevance* of the data, which is their:
    1.1. *business relevance* for supporting the KM activities, as they reflect in the SPC goals. This is data which are generated by the operational processes of the business function where the *Business problem* resides (Huan, Yu et al. 2003, p.410), or are relevant to *Potential solutions*. This data is usually found spread over a number of disparate data domains within the organization (Mozer, Wolniewicz et al. 1999, p.3);
    1.2. *potential technical signal*, which indicates the potential for discovering *business relevant* information from that data;
2. *assembleability* of that data into an optimal flat table form, which is:
    2.1. *accessibility* to the data mining algorithms during *Data mining discovery* and later in the *Realise phase* of the project; and
    2.2. economical *sustainability* during the *Realisation* phase of the project;
3. *technical desirability* for mining toward the SPC goals. Technical desirability is concerned with:
    3.1. the elimination of noise from the data;
    3.2. the optimisation of the technical signal within the data e.g. of inter- and intra-domain interactions between features (Pyle 2003, p.367) (Westphal and Blaxton 1998, p.51);
    3.3. the creation of labels if there are supervised data mining problems;
    3.4. reducing dimensionality (Han and Kamber 2001, Chapter 3) (Huan, Yu et al. 2003);
4. *proven utility* of the data during the mining toward attaining the SAM project's mission.

5.8.5.3 How *Identify, assemble, prepare useful data* supports

Identify, assemble, prepare useful data supports through applying SPM activities and technical data preparation tasks. The activities and tasks are specifically designed, and optimally sequenced according to the TQM principle *forward linkage of utility for purpose*, to simultaneously prepare the data for the data mining (i.e. increase its utility for purpose) (Pyle 2003, p.370), and to reduce the project's risk of not attaining *project mission*.

We commence the application of SPM activities within this task set, by formulating four *Operating objectives* from the above four data utility criteria. These objectives are:

1. *establish the relevance* of the data for supporting the *project mission* through data mining;
2. *establish the assembleability* of the data;
3. *establish the technical desirability* (SPM 2) of the data for data mining;
4. *prove the data's utility* for supporting the *project mission* through data mining. This is the *technical possibility* (SPM 3) with the data.

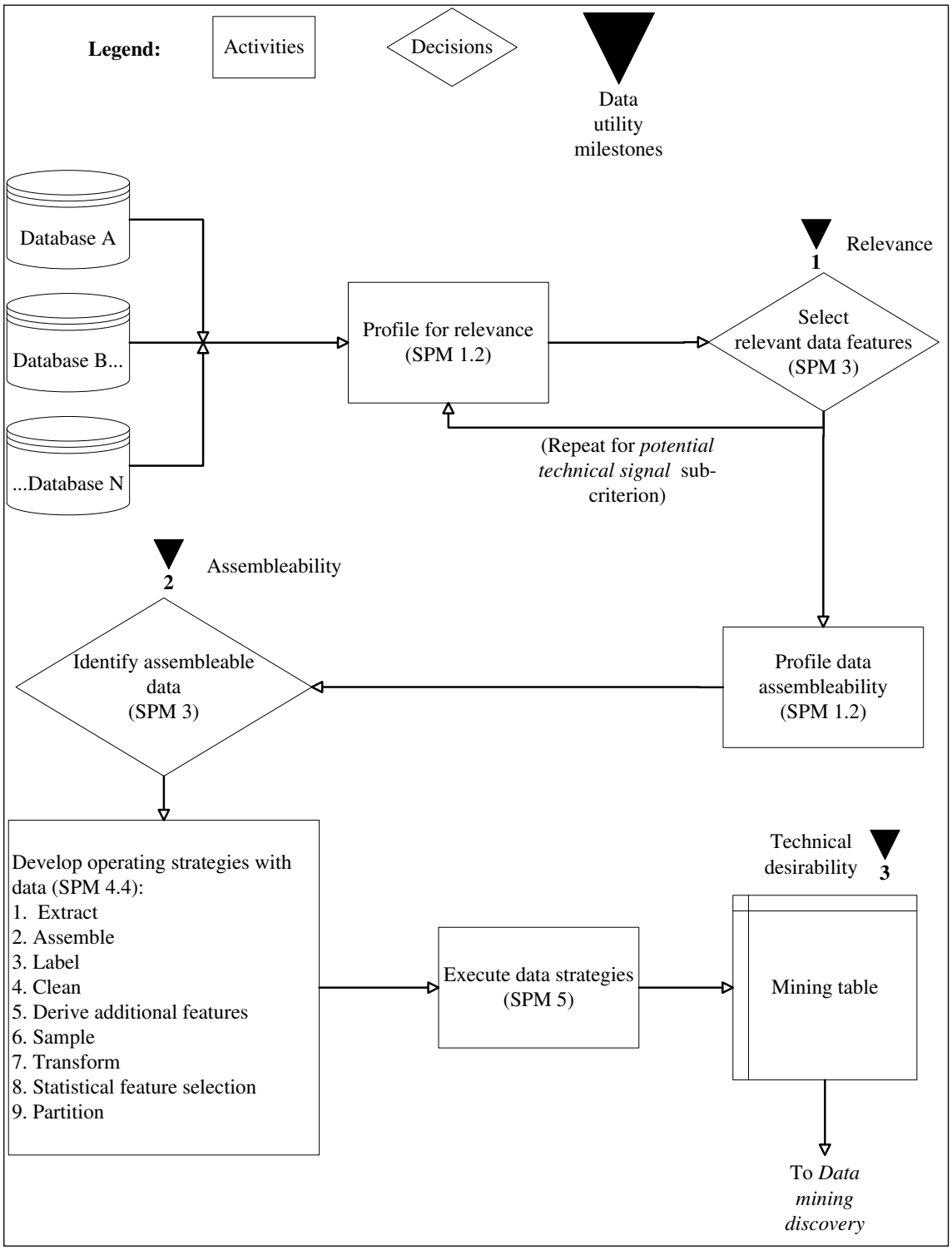


Figure 5.4: Data preparation in SAM

We aim *Identify, assemble, prepare useful data* at attaining the first three objectives. We aim the following data mining task set *Data mining discovery* at attaining the fourth objective.

There is a repeat of SPM activities, and some SPM activities are also modified for an application in *Identify, assemble, prepare useful data*. We next give an overview of this modification and application of SPM activities in this task set. Figure 5.4 will form the basis for that discussion. The overview is followed by an in-depth discussion of the various tasks and decisions within *Identify, assemble, prepare useful data*. In Figure 5.4, we have renamed the SPM activities to reflect their data preparation application, and we have retained their link with SPM through SPM labels (SPM 1.2, 3 etc.).

The legend of Figure 5.4 shows the tasks and decisions of *Identify, assemble, prepare useful data.* There also are milestones for each of the three objectives we need to achieve. The three database elements at the start of the figure, represent an organisation's disparate data. At this point in a project, there is total uncertainty about which data meet any of the utility criteria.

5.8.5.3.1 Application of SPM

There are two applications of a *modified* SPM *cycle*, in which we also *modify* the *criteria* being evaluated. The first application is to attain the first objective – *establish the relevance* – and the second application is to attain the second objective, which is *establish assembleability.*

We first discuss the modification of the SPM *cycle*. We recall that SPM has a Hegelian triangle. That triangle consists of the thesis *Profile controllable factors* (SPM 1.2), the antithesis *Profile uncontrollable factors* (SPM 1.3), and the synthesis between the two in *Strategic analysis* (SPM 2) and *Strategic choice* (SPM 3). Now, considering the purpose of evaluating the data in the project, we are not interested in the *effect* of any *controllable* or *uncontrollable* factors on the degree to which the data meet the evaluation criteria. We are simply interested in the *degree* to which the data *meet a criterion or not*, and the cost of any ad hoc improvement we could make to the degree of meeting the criterion. We therefore modify the SPM cycle in its two applications here, by excluding the synthesis. Further, because there is no synthesis, we can also exclude the antithesis. Therefore, once we have established the thesis, we can proceed directly to the *Strategic choice*.

We present this modified application of SPM in *Identify, assemble, prepare useful data* in Figure 5.4:

- in the pursuit of the first objective we apply:

- SPM 1.2 in the activity *Profile for relevance (SPM 1.2)*;
- and SPM 3 in the decision *Select relevant data features (SPM 3)*. Note how there are no activities labeled *SPM 1.3* or *SPM 2* before the first milestone;
- we once repeat this activity and decision for a second criterion, shown with a backward looping arrow in Figure 5.4 and marked accordingly;

❖ in pursuit of the second objective we apply:

- SPM 1.2 in the activity *Profile data assembleability (SPM 1.2)*;
- and SPM 3 in the decision *Identify assembleable data (SPM 3)*. Note how there are no activities labeled *SPM 1.2* or *SPM 2* before the second milestone.

The further modify and apply SPM's evaluation criteria in as follows:

❖ in pursuit of the first objective, replacing the *controllability* criterion with:

- the *business relevance* sub-criterion 1.1 above in the first completion of the task *Profile for relevance (SPM 1.2)*; and with
- the *potential technical signal* sub-criterion 1.2 above during the second completion of the task *Profile for relevance (SPM 1.2);*

❖ in pursuit of the second objective, replacing the *controllability* criterion within the activity *Profile data assembleability* with the *assembleability* criterion 2 above of the data.

In the *Select relevant data features (SPM 3)* decision, we limit the assessment of any criterion to a *qualitative expert assessment* only of the *degree* to which the data meets the criterion. There is no quantified *cost impact* or *contributory impact* of the data in dollar terms, as one would expect from an unmodified application of SPM's *Strategic analysis* (SPM 2) on a commercial problem. The first reason is that the cost of the data first is a sunk cost, which was factored into ROI calculations of the data infrastructure before, and it is unacceptable accounting practice to accrue a cost twice. The second reason is that the current state of the data is a given, having been driven by operational requirements; currently organisations will not spend money to change operational data requirements, for analytical purposes.

In the *Identify assembleable data (SPM 3)* decision, we need to include actual costing associated with assembling the data, since this is one of the incremental costs of the

project to the organisation. There is a basic cost associated with the assembleability of any data feature, which decreases as its degree of *assembleability* improves. The degree of assembleability is determined by how difficult – and therefore expensive - it is to include the data feature in the assembled data. We cannot, however, calculate profit associated with data preparation activities in SPM 2 and SPM 3, because the profit from the project can only be established in the applications of SPM 2 and 3 in the main body of SAM.

When – and only if – the second *assembleability* objective has been met, do we pursue the third objective; *establish* the data's *technical desirability* for data mining. We achieve this third objective, through formulating a number of technical strategies, and executing them. The strategy formulation takes place in *Develop operating strategies with data* – which is an application of SPM 4.4. We *Execute data strategies* as an application of SPM 5. That attains the third objective, and results in the required format of the data for mining. In Figure 5.4 that is shown as *Mining table.*

We pursue the fourth objective - *prove the data's utility* – in the SAM data mining task *Data mining discovery.*

We conclude the overview, by commenting that in the business community the use of the term *iteration* in projects, is perceived as an *unnecessary* repetition of something, which was not planned well for, at its outset. This implies negative risk and cost consequences from the project for the business. The business community therefore views the *iterative* descriptions in the data mining literature of this major project task set, with suspicion. Our approach in addressing these concerns about the business community, is to:

- reiterate our previous distinction between *iteration* and *repeat* of the same task or activity. We *repeat* for with a *different decision criterion* each time. In *Identify, assemble, prepare useful data* we repeat for the two sub-criteria of the data's *relevance;*
- accept the unavoidability of *iteration* in the exploratory phases of projects (Liu 2003, pp.436, 437), while at the same time explaining the support produced by each iteration. The business community can relate to the concept of cumulative support.

We now discuss in detail the activities and decisions within *Identify, assemble, prepare useful data*.

5.8.5.3.2 Establish data's relevance

5.8.5.3.2.1 First application of *Profile for relevance (SPM 1.2)*

The first application of SPM is a cognitive activity set for attaining the first data utility objective. The activities of this activity set are:

- audit the organisation's metadata from the perspective of domain knowledge, for supporting the attainment of the *project mission* (Pyle 1999, p.138). Metadata is the organisational knowledge base about the information technology systems. It can also be termed *data about data,* and contains technical information about the origins of the data, the path they have followed through the data infrastructure, the business and operational processes they represent, and their commercial meanings (Date 2000, p.70).
- profile - by expert reflection-in-action - the audit, determining the business relevance of each data feature for supporting the project mission;
- rate the features qualitatively on a scale of uncertain/low/medium/high according to the degree to which they hypothetically support the *project mission*;

5.8.5.3.2.2 First application of *Select relevant data features (SPM 3)*

The first application of *Select relevant data features (SPM 3)* is selecting those data features, which have the best *business relevance*. The decision is based on subject matter and domain analytics expertise.

5.8.5.3.2.3 Second application of *Profile for relevance (SPM 1.2)*

- extract five or ten percent samples of the features, which were selected above. During the extraction activities, survey the extraction process (Pyle 1999, pp.113, 138) and note any issues which may affect the *assembleability* of the data. These issues may relate to IT resource issues, database connectivity, database access times, data dynamics, data keying issues;
- generate statistical measures of dispersion, measures of central tendency, of cardinality, and of missing values for those sampled features;

- analyse and interpret these basic statistical parameters of each feature for the *potential* that a feature *may* carry *potential technical signal*. So for instance if there are many missing values, or low cardinality, the feature may not carry much technical signal;
- rate each feature – by expert assessment - according to the degree to which it supports the utility sub-criterion of *potential technical signal*, on a scale of uncertain/low/medium/high;

5.8.5.3.2.4 Second application of *Select data features (SPM 3)*

Select those data features, which have the best *potential technical signal*. The decision is based on domain expert assessment.

5.8.5.3.3 Establish data's assembleability

5.8.5.3.3.1 Profile data assembleability (SPM 1.2)

The application of SPM is a qualitative, cognitive activity set for attaining the second objective. The activities of *Profile data assembleability (SPM 1.2)* are:

- audit the data assembleability survey which was created above during the sample extracting;
- profile the audit by expert *reflection-in-*action, determining the regular assembleability of each feature in a productive mode in the *Realise* phase of SAM. The profile includes a quantification of the cost associated such regular assembly of each feature;
- sort the features ascending by their cost of assembly.

5.8.5.3.3.2 Identify assembleable data (SPM 3)

The decision *Identify assembleable data* is an unstructured combination of:

- decisively eliminating those data features which fall outside the organisation's estimated budged for *assembleability*; and
- domain expertise about which features should be included despite apparent high cost, because of incremental benefit added to the project ROI.

In SAM, we provisioned for a repeat of any of the *decisions* of *Identify, assemble, prepare useful data,* should we find in *Data mining discovery* that a decision needs to be

reviewed. The iteration arrow between these two task sets in Figure 5.1 indicates that review.

Note that in *Identify, assemble, prepare useful data,* there is an ongoing assaying of the data (Pyle 1999, p.125) about each utility criterion. This should be augmented to the metadata.

5.8.5.3.4 Establish data's technical desirability

5.8.5.3.4.1 Develop Operating strategies for data (SPM 4.4)

In this task set, we extract and assemble the *relevant*, and *assembleable* data into one table, and pursue the third objective – *establish technical desirability*. We achieve all of this through developing a number of *Operating strategies* and executing them in the next activity. The developing of these strategies is an application of SPM 4.

We have already commented on the completeness of the technical content of these strategies in the literature. In SAM the technical strategies with the data are:

- *Extract* is the developing of the code, which was used for the exploratory sampling in *Profile for relevance (SPM 1.2)*, into a productive version(s) for getting the data out of the disparate location, into staging tables. This code will have to consider accessibility issues which were profiled during *Profile data assembleability (SPM 1.2)*;
- *Assemble* is the developing of code which assembles the staged data into one flat table, which will later be mined in both *Data mining discovery* and in the *Realisation* phase of the project;
- *Label* is the creation of a target variable in the event of supervised data mining. The label is required for the classifier to learn (Cooley 2003, p.609) (Gehrke 2003, p.4);
- *Clean* is:
  - eliminating collinearity. When there are too many features that carry the same signal, the model parameter estimates become unstable (SAS Institute online b, p.8) (Hastie, Tibshirani et al. 2001, pp.53, 59) (Diekhoff 1992, p.276);
  - replacing of existing values with other values may be required for reasons of consistency, understandability, for parsimony under conditions of high

cardinality (Cooley 2003, p.608), for overcoming limitations of some algorithms (SAS Institute online b, pp.6ff.);

- imputing missing values is usually required because of algorithmic constraints and for unlocking signal in the data;
- filtering outliers, because some algorithms are sensitive to outliers (Pyle 2003, p.378);

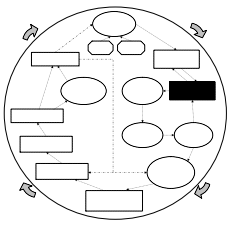
- *Derive additional features.* We saw substantial evidence in the chapter on concept drift, that hidden context – and therefore hidden signal - can be exposed through deriving additional features in the data (SAS Institute 2003a, p.2-33). This is also known as *feature extraction* through some functional mapping (Huan, Yu et al. 2003, p.411). Examples of derived features are spline functions (Steinberg 2003);
- *Sample.* There are numerous sampling techniques documented in the literature e.g. (SAS Institute 1998). The choice of a sampling technique depends among other things, on the nature of the problem, the amount of data available, and prior probabilities;
- *Transform* data is for instance the discretising of interval independent features if the dependent feature is categorical (Woods 2003a), and the transformation of the distribution of an independent interval feature to better match the distribution of a dependent interval feature (SAS Institute online b);
- *Statistical feature selection* is identifying those independent features which have the biggest influence on the value of the dependent feature, by a threshold level of some measure of statistical significance e.g. χ^2 or R^2 (Huan, Yu et al. 2003, pp.412ff.) (SAS Institute online d). The number of features in a model can have up to a quadratic effect on modeling resource requirements (SAS Institute online b, pp.16ff.), and further make the model difficult to understand. The purpose of this strategy is to build a model which is not costly to maintain, not too complex to understand, and does not place excessive demand on the IT infrastructure;
- *Partition* the data for data mining is well documented in the literature (Agrawal and Psaila 1995) (Pyle 1999) (Pyle 2003) (SAS Institute online b, p.44) (SAS Institute 1998, p.21). SAM follows the data partitioning denomination of training, validation, and test data of Bishop (Bishop 1995).

5.8.5.3.4.2 Execute data strategies (SPM 5)

In order to attain the third objective, we have successfully to execute the technical *Operating strategies* with the data. This activity is the execution of those *Operating strategies*, and therefore labeled SPM 5. Considering the output of *Identify, assemble, prepare useful data* from an SPM perspective, attaining the third objective with the data establishes the *technical desirability,* of the data.

There now remains the meeting of the fourth objective – *prove the data's utility* for supporting the *project mission*. This objective can only be attained through actually discovering useful information with data mining in the data. We can therefore only judge that we have attained this objective, at the end of SAM's data mining task set *Data mining discovery.* When we do meet this objective, we will also have established the *possibility* of what we can achieve with the data, under the technical data and data mining circumstances.

5.8.6 Data mining discovery



5.8.6.1 What *Data mining discovery* is

In Chapter 4, we hypothesised that data mining methodology can be improved, by formulating a mapping technique between the SPC goals, and the data mining plan. In this section, we offer reframing in this regard, and indicate that with the label *4.3* in Figure 5.3. We also offer reframing in the presentation of iteration in data mining, by casting iteration in a progressive, business valued framework. Further, we introduce novelty about confidence level thresholds for the results of some of the data mining. We propose further innovation in activities for evaluating the discovered information for relevance to the business deliverables.

Data mining discovery is the task set of SAM, where we *define* the key technology application factors, we identified in *Develop project mission*, and *match* any models to those factors (Liu 2003, p.437). It can also be termed the exploration of the problem and solution spaces (Pyle 1999, p.21) in the organisation's data. It uses *expert reflection-in-action* (Schön 1995) and modified SPM activities, for evaluating the extent to which discovered information meets *confidence* and *relevance* measures. We indicate the use of SPM activities 1.1, 1.2, and 3 with the *(1.1, 1.2, 3 DM)* part of the label of *Data mining discovery* in Figure 5.1.

This task set itself does not evaluate the business utility of the discovered information, but integrates it with an initial, evaluative looping of the discovered information through the business activities loop *Develop circumstantial knowledge, Strategic analysis,* and *Strategic choice.*

5.8.6.2 *Data mining discovery's* support

Data mining discovery firstly *discovers* the information which is required for the creation of knowledge in the *Create* phase of SAM's KM cycle.

Second, it evaluates the data mining's technical output for meeting confidence measure thresholds, as jointly determined by the expert team.

Third, it attains the fourth objective of *Identify, assemble, prepare useful data* – which was *proving the utility of the data* in supporting the *project mission* through data mining. This is the same as its TQM purpose, which is establishing the *technical possibility* with the *utility for purpose* with the data, as far forward in the project as possible.

From a Project Risk Management perspective, it provides the *materiel from the data* which potentially could be used in realising the business deliverables.

5.8.6.3 How *Data mining discovery* supports

Data mining discovery supports through a combination of two technical task sets with the three SPM activities.

5.8.6.3.1 Formulate data mining mission (SPM 1.1)

This is the application of the SPM 1.1 activity. We reflect this in the …*1.1*... part of the label of *Data mining discovery (SPM 1.1, 1.2, 3 DM)* in Figure 5.1.

Now that we know the data better through having prepared it for mining, we can develop the *project mission* into a *data mining mission* or a technical charter, for the project. The *data mining mission* contains the technical data mining goals, and strategies for attaining those goals – just like a plan. It is important to consider however, that even though the *data mining mission* has the elements of a plan, it remains a *mission*. This is because both the strategies and objectives are untested under the organisation's data circumstances.

In formulating this *data mining mission*, we consider that the SPC goals of the *project mission* are binding. That means that the data mining mission, must support those SPC goals in an optimal manner. We assure that optimality, by applying the mapping technique which we introduced in Chapter 4. Since SAM will ultimately stand alone as a document, we repeat the mapping technique here:

- in the first two *problem / opportunity* exploratory or analytical phases of the project, the technique is:
  - developing the SPC *goal(s)* into data mining *goal(s)*; and
  - developing data mining strategies which will attain the data mining goal(s);
- in the subsequent business *solution* exploring and operationalising project phases, the technique is:
  - developing the SPC *strategies* into data mining *goal(s)*; and
  - developing data mining strategies, which will attain the data mining goals.

We recall that the SPC goal of the *fourth* level (SCHEMA 4), is not concerned with the discovery of information, but with developing the business solution. We therefore limit the formulation of goals in the data mining mission, to those which relate to the discovery of information. They are the *first three* SPC goals. The developing of the business solution – SCHEMA 4 - takes place in the KM activities which follow *Data mining discovery.* The development of the SCHEMA 5 goals and strategies, take place in *Develop data mining plan.*

The *sophistication* of the data mining strategies could span anything from a simple non-technical investigation (like inter-departmental communication), through non-statistically significant techniques (like conditional queries), through modeling for statistically significant inferential purposes, to a chain of models for solving multiple tasks associated with one goal (Reinartz 1999, p.3), to experimental design for proving cause. Since the data mining strategies within the *data mining mission* are *unproven* for supporting their SPC goals under the data circumstances, we permit *iterative experimentation* for strategy development, until we do support a SPC goal.

5.8.6.3.2 Determine acceptable confidence measure thresholds

This exploratory environment within *Data mining discovery* means that there are two things we want to be confident about when executing these *strategies*. The first is their *effectiveness* in discovering information which supports the SPC goals, under the data circumstances. The second, is to test the *duplicatibility* of these strategies under the data circumstances.

One component of assessing the model's *effectiveness*, is a *qualitative* assessment of the discovered information by the business domain expert's assessment, for its support of the project's *business goals*. Additionally, and depending on the algorithm used, there are also *technical* measures for assessing this effectiveness. In the event of a classifier, accuracy is a technical measure of *effectiveness*. In the event of an unsupervised classifier like k-means, we base our confidence in the *effectiveness* on the amount of natural data structure, which was captured. The measure for this is R^2. In the event of predicting a quantity, effectiveness for instance in measured on mean squared error.

Effectiveness is a measure of the improvement of identifying the targeted event or its value, over what is possible with a random, non-modeled approach. The minimum level of acceptable *effectiveness* depends on the business domain, but in the CRM domain of our Telco ABC's retention problem, the minimum cut-off is about 70% accuracy.

The cut-off for R^2 – the amount of natural structure we want to capture in the data – is subjective, but in the business domain we would accept 0.7 as enough confidence to base decision upon (Woods 2003b). For mean squared error one would expect the range not to exceed the estimated range of project ROI.

We base the assessment of our confidence about the *duplicatibility* of the results of the modeling, on statistical *measures of confidence* about the model. The *measures of confidence* express our confidence in rejecting the null-hypothesis, that the model discovered the information purely by chance, and that the results therefore are not repeatable. Duplication means that we will discover the same information in the same data, applying the same modeling technique with the same settings in future. Put in other words, these measures are used to determine the degree to which the data is *modelable*, which was the fourth data preparation objective.

Examples of *confidence measures* for *duplication* are the p-values of the null-Hypothesis in probabilistic problems, and the F-statistic in value problems (Levin 1987, Chapter 10) (Woods 2003a) (SAS Institute online a).

Depending on the problem at hand, we set *cut-off levels* about *confidence measures*, below which we are not confident in the *duplicatibility* of our modeling results any more. In the medical environment, the *confidence measure levels* may be set at 99% (Hastie, Tibshirani et al. 2001). In the business environment, these *confidence measure levels* are set at lower levels, depending on the business domain. Acceptable *confidence measure levels* on commercially generated data can be as low as 95% (Woods 2003a).

For reasons of project economics, we present novelty about model effectiveness and duplication measures in Table 5.2. We motivate this novelty as follows. The SPC goals in the *project mission*, follow the KM and *pre-project schema* structure we developed in Chapter 2. In that structure, the first two stages were concerned with exploring the *business problem / opportunity* space, and the third was concerned with exploring the *business solution* space. We also saw there, that we could only progress on to exploring the *solution* space, once we have legitimised our understanding about the extent and nature of the *problem / opportunity*. In SAM, the legitimisation is achieved through hypothesis support in *Develop circumstantial knowledge, Strategic analysis*, and *Strategic choice.* Such a legitimisation about the problem space requires that the discovered information be supported by both acceptably high *effectiveness* and *duplicatibility* (e.g. rejecting the null hypothesis).

	Effectiveness accuracy measures	**Duplication confidence measures**
Business problem environment:	Fixed	Fixed
Business solution environment:	Pliable	Fixed

Table 5.2: Model measures

In the case of the third stage goals about the *business solution* however, there is one more legitimisation test after hypothesis support. That test is developing the *executibility* of the solution under the business's commercial and operational circumstances. We apply this test of the fourth stage SPC goal in *Define new business objectives and strategies*.

Developing that executibility may require some adjustments to the third-level SPC goals and data mining goals and strategies which attain that, to reflect the *possibilities* of execution under the organisation's data circumstances. This adjustment may introduce more variance into the later project results, than pliable model *effectiveness* measures. Retaining the high *duplicatibility* means that we can put in effort later during the SAM *Realise* phase to improve *lift*, after we have profiled any solutions for the organisation's circumstances.

We saw before how a typical *mission* has four components in each dimension. Above we covered the first two components *what* (goal(s) and *how* (strategies). The *where* component of the *data mining mission* are the data features in the mining table. The fourth dimension of the *data mining mission* - *when* – is a carry-over from the *project mission*.

5.8.6.3.3 Execute *data mining mission*

We now execute the data mining mission. This execution includes the first three sub-tasks below, as well as the external business activities in the loop we labeled *Repeat for each DM goal* in Figure 5.1. If there are multidimensional goals, the execution of the *data mining mission* and the loop are *repeated* for each goal.

5.8.6.3.3.1 Execute strategies

In this first technical subtask, we execute each of the strategies of the *data mining mission* with sufficient diligence to meet their respective *confidence* and *effectiveness* requirements. The sequence starts with the first dimension goal in the *data mining mission*. For instance, if there was a first dimension goal about understanding the existence of a problem better, then we start there. Alternatively, if there were no goals about the *problem / opportunity* space, then we start with the goals about the *solution*.

In this exploratory environment, the strategies in *data mining mission* are points of departure for experimentation. We now experiment with those strategies, and where

they fail, we experiment to discover other strategies. This maintains the forward momentum of the project. We will refer to these *data mining strategy findings* later in SAM's task set *Develop data mining plan*.

5.8.6.3.3.2 Profile technical results (SPM 1.2)

This second subtask is an application of SPM 1.2. This is reflected in the *…1.2…* part of the label of *Data mining discovery* in Figure 5.1. We have once again removed the antithesis and synthesis, proceeding directly from SPM 1.2 to SPM 3. This modification is similar to the one done in *Identify, assemble, prepare useful data* above, and for the same reasons. We have now replaced the profiling criterion *controllability* with the criteria *confidence measures* and *effectiveness.*

Using expert *reflection-in-action*, we *profile* the technical outputs from the data mining for their:

- measures about *statistical confidence;* and their
- *effectiveness*.

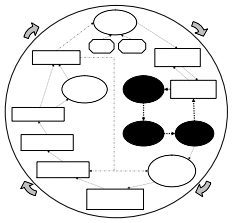
This is in preparation for determining if the discovered information meets the confidence and effectiveness thresholds in the following subtask.

5.8.6.3.3.3 Confirm technical success (SPM 3)

In this subtask, we identify those data mining results, which meet the *confidence* and *effectiveness* thresholds. We do this through applying SPM's *Strategic choice,* which is SPM 3 and *expert reflection-in-action* and an information score-card. We use *reflection-in-action*, because in this exploratory environment, some pliability is required in the decision. This pliability is especially important during the decision about the results about the *solution goals*, where the effectiveness requirement may be relaxed.

When a strategy has not produced passable results, we may iterate through the above subtasks, or even through *Identify, assemble, prepare useful data* and the *Data mining mission* formulation task above. When a strategy has produced passable results, we pass the associated discovered information through to the KM loop, where we determine if the information matched the expectations (Pyle 2004, p.260).

5.8.6.3.4 Evaluating the discovered information using the Knowledge Management loop



The KM loop contains the KM activities, which are required for validating the discovered information against their SPC goals in the *project mission*. Note that the evaluation is against the SPC goals, and not against the *data mining mission goals*. It is also a high-level *evaluation* of discovered information for relevance or not, as against the detailed redevelopment of hypothesis in a later looping through this knowledge management loop. In this looping we:

- receive the discovered information and develop that into untested knowledge, which is required for hypothesis testing under the organisation's commercial and operational circumstances. This happens in *Develop circumstantial knowledge*; then we
- test the hypotheses from *Develop project mission* for their supporting the SPC goal about the problem / opportunity or solution. The purpose with the project is to progress toward a solution, and we allow reframing of the hypotheses in order to get support. The support happens in *Strategic analysis;* then we
- legitimise the hypotheses for their SPC goal support through acts of evaluation and selection. The use of the mapping technique also assures that the data mining goal will support the hypothesis.

When a third-stage goal - which relates to the *business solution* - has been attained, the project proceeds to SAM's *Define* phase and the *Define new business objectives and strategies* business activity.

Once the discovered information has been validated for relevance in this high-level application of the knowledge management loop, we proceed to a more in-depth application of the activities within the loop, for developing the executibility of the knowledge.

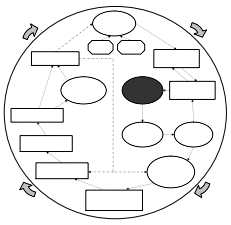
5.8.6.3.5 Iteration

In the event where we cannot legitimise a hypothesis, or find that a data mining strategy does produce the required information, we may *iterate* until we discover the required information and legitimise or reframe the hypotheses. The iteration is either in

Data mining discovery, or may include *Identify, assemble, prepare useful* data. The *purpose* of the iteration is to support the *binding* SPC goals through experimenting in the data.

If iteration still is not productive, then the causes may be insufficient planning skills or ability in the organisation, or a lack of domain knowledge, or that data mining is not a suitable approach to the problem. The organisation may decide to up-skill, or to abandon the project using data mining.

5.8.7 Develop circumstantial knowledge



In Chapter 4, we reflected on reframing data mining methodology, to include KM activities. We explained earlier how SAM's business activity *Develop project mission* was the first of the KM activities, which constitute this reframing. The SAM business activities starting with *Develop circumstantial knowledge* through to *Execute new business strategies* constitute the balance of this reframing, giving SAM potency in practical inference. These activities cover the first three KM phases (Figure 5.2). We express their reframing of Chapter 4 in the *4.4* labels of Figure 5.3. Apart from the formal reframing of the hypothesis which underlie the business deliverables for the newly discovered information, this knowledge management loop also constitute the evaluative activities of the *relevance* of the discovered information, as we described it in *Data mining discovery*..

5.8.7.1 What *Develop circumstantial knowledge* is

From the KM cycle within SAM, *Develop circumstantial knowledge* is the first business *knowledge developing* activity in the KM *Create* phase of SAM. It is the first business activity of the *Analyse* phase of SAM. *Develop circumstantial knowledge* is the unaltered application of the SPM activities *Profile controllable factors* (SPM 1.2) and *Profile uncontrollable factors* (SPM 1.3), on the information which was strategically selected in *Data mining discovery*. It creates the operational recipes for the contextual situation (Pyle 2004, p.97), or. We have accordingly labeled *Develop circumstantial knowledge* in Figure 5.1 as *(1.2, 1.3 BU).*

If there are alternative, non-data mining based solutions, which are competing with the data mining-based solution, then they are profiled here too.

5.8.7.2 *Develop circumstantial knowledge's* support

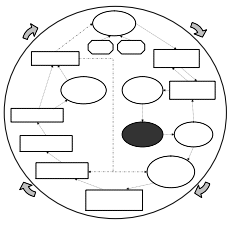
The KN purpose of the business activity *Develop circumstantial knowledge* then, is creating the knowledge profiles from the discovered information, which are required for hypothesis testing about the support of the data mining results for the whole range of SPC goals. From a TQM perspective, it supports the defining of the *possible utility for purpose* of the *business solution*, and of the *possible understanding* of the *problem*. Its Project Risk Management purpose is to profile the understanding of the *problem / opportunity* or the *wanted event* for later actioning.

5.8.7.3 How *Develop circumstantial knowledge* supports

Develop circumstantial knowledge supports by applying SPM 1.2 and 1.3 to the discovered information, and developing the discovered information into untested knowledge profiles. We saw in Chapter 2 that knowledge is know-how, and here we develop that know-how from the information, which is required for testing the hypotheses, which underlay the SPC goals. The organisation's commercial and operational circumstances bring with them controllable and uncontrollable factors, which affect the knowledge we are developing, for this particular organisational environment. The profiles then are the developed knowledge, after it has been adapted for the specific situation.

This activity is repeated for the information, which was discovered for *each* different *SPC goal*. When we develop the profiles, we therefore need to consider that there are different controllable and uncontrollable factors at play for each goal. So for instance, during profiling knowledge for hypothesis testing about the causes of Telco ABC's retention management problem, an *uncontrollable factor* called *competitor acquisition marketing activity* may have a big impact on the knowledge profile.

5.8.8 Strategic analysis



5.8.8.1 What *Strategic analysis* is

This is the last activity of the *Analyse* phase of SAM. It is last of SAM's KM phase *Create*. It is the unaltered application of the SPM's *Strategic analysis* activity (SPM 2), reflected in the *...2 (BU)...* part of the numeric label.

5.8.8.2 *Strategic analysis'* strategic support

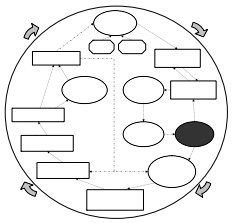
From an SPM perspective, is determining the *desired* (SPM 2.1.1) and the *possible* (SPM 2.1.2) business solutions under the organisations commercial and operational circumstances. In TQM terminology, *Strategic analysis* constitutes the *possible understanding* of the *cause of the problem*, or the *possible utility for purpose* of the *business solution*. It adds business value in the KM process, through *testing the hypotheses* about the whole range of SPC goals. As a Project Risk Management activity, it removes uncertainty about the possible constitution of the *wanted event*, and about *event impact*.

5.8.8.3 How *Strategic analysis* supports

Through an unaltered application of *Strategic analysis* (SPM 2) and *expert reflection in action*, we test the validity each of the existing hypotheses, which underlie the SPC goals, for their SPC goals support *in the light of* the newly developed knowledge profile.

Since the purpose of the data mining is to introduce novelty into the organisation, we may use the new knowledge to *update* or *reframe* a hypothesis to better support a SPC goal.

5.8.9 Strategic choice



5.8.9.1 What *Strategic choice* is

Strategic choice is the *Choice* activity of SAM's SPC phases. It is an application of SPM 3, therefore the *…3…* in its label in Figure 5.1. It is the first activity in SAM's KM phase *Legitimise*. It is the choosing between alternatives within the uncertain business environment (Pyle 2004, p.128).

5.8.9.2 *Strategic choice's* strategic support

Where we have developed multiple hypotheses about one goal, strategic support is firstly to identify and legitimise the *best* hypothesis for attaining the SPC goals. If there was just one hypothesis about a goal, then this activity evaluates and decides whether that hypothesis sufficiently supports the SPC goal. For that reason, it *Strategic choice* has the *KM 2.1* in its label in Figure 5.2.

As an SPM business activity, the purpose is choosing the *best possible* hypothesis, for *understanding* the *business problem*, and for *constituting* a competitive *business*

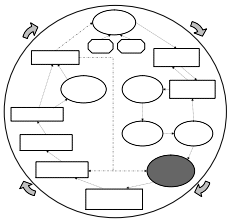
solution valid under the organisation's limiting commercial and operational circumstances.

As a Project Risk Management activity, it eliminates the remaining uncertainty about the *impact* of the *wanted event*, and about the constitution of the *wanted event*. Its TQM business value-add is establishing the forward link of the *possible utility for purpose* of the business solution.

5.8.9.3 How *Strategic choice* supports

Through an unaltered application of *Strategic choice* (SPM 3) and *expert reflection-in-action*, the project team *evaluates* the (updated or reframed) hypotheses against their SPC goals, and decide if they support their SPC goals or not. They then choose the best-supported hypotheses, and reject the rest, legitimising the knowledge to be used in solution design.

5.8.10 Define new business objectives and strategies



When we have legitimized the knowledge required for attaining the SPC goals, the project proceeds to the *Define* phase. *Define new business objectives and strategies* is the first entity in this phase.

5.8.10.1 What *Define new business objectives and strategies* is

Its input is the legitimised knowledge about the *business solution*. It is the second business activity within the KM phase *Legitimise*, and is an unaltered application of SPM's objectives and strategies formulating activities. We therefore labeled it *4.1 – 4.4 BU* in Figure 5.1 with reference to SPM 4. Also, it contains reframing of CRISP-DM about the mapping technique, and has been labeled *4.3* in Figure 5.3. Further, because of its KM function, it reframes CRISP-DM in the KM dimension, therefore the label 4.4 in Figure 5.3.

5.8.10.2 *Define new business objectives and strategies'* support

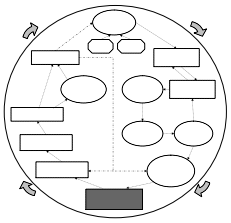
Its KM purpose is *formulating the executibility* of the new business solution as new business objectives and strategies (Pyle 1999, p.29). Its TQM purpose is *establishing* the *utility* of the business solution. That means it reduces uncertainty of project risk associated with the *wanted event's likelihood*, and reduces uncertainty about harvesting its positive *event impact*.

5.8.10.3 How *Define new business objectives and strategies* supports

Define new business objectives and strategies adds business value through defining the *Strategic objectives* and *Grand strategies* for the business solution, and then developing them into their respective business functional *Operating objectives* and *Operating strategies*. This in an unaltered application of SPM 4. These objectives and strategies support the knowledge, which was legitimised by *Strategic choice*.

Further, because it contains the *Operating objectives and strategies* hierarchy of the business solution, it later becomes the *link* for applying the *mapping technique* between the SPC goals and the data mining plan.

5.8.11 Develop data mining plan



5.8.11.1 What *Develop data mining plan* is

We formulated a *data mining mission* in *Data mining discovery*, because of uncertainty about what was possible with data mining under the organisation's data circumstances. The execution of that *mission* removed that uncertainty and established the *technically possibilities* with the data mining. However, there remained the uncertainty about what information the organisation would find useful under its limiting commercial and operational circumstances. This uncertainty still prevented us from formulating a data mining plan in the true sense of what a plan is – objectives and strategies for supporting what is executable by the business. We removed that uncertainty when we legitimised the hypotheses for supporting their respective SPC goals in the activities above.

Develop data mining plan therefore is the point in SAM where we formulate the plan for the technology, for which we know there are no more limiting circumstances. This plan takes the form of the *Operating objectives and strategies* of the technical data mining plan. We indicate this with the *(SPM 4.3 4.4 DM)* part of the label in Figure 5.1. We consider that the *Strategic objectives* (SPM 4.1) and *Grand strategies* (SPM 4.2) within the business solution, are the same for the technology, and we do not have to redevelop them for the technology plan.

We also find here the application of the mapping technique, which is a reframing of Chapter 4's section 4.3. We reflect this in its *4.3* label in Figure 5.3. It is the first data mining task set of SAM's *Define* phase. It is the activity in the KM cycle's *Legitimise* phase, where we plan the technology's support for the legitimising, and it is labeled accordingly in Figure 5.2.

Apart from focusing the following data mining task set *Model, evaluate, choose best model(s)*, *Develop data mining plan* also provides the impetus and focus for the data mining task sets *Operationalise model(s), Deploy outputs into business,* and *Monitor and control.*

5.8.11.2 *Develop data mining plan'* support

Its purpose in SAM as a data mining activity is to define the technical data mining project, which *supports* the execution of the new business solution and paradigm. As a KM activity, it develops the technological support for the legitimisation of the developed knowledge. The Project Risk Management value is defining part of *how* the organisation will realise the *wanted event*, also reducing uncertainty about the *event*

likelihood. It adds TQM value, by defining the *utility of the technology* in realising the business solution.

5.8.11.3 How *Develop data mining plan* supports

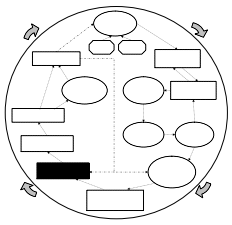
This activity supports through defining the data mining plan in a technical application of SPM 4.3 and 4.4, and in using the mapping technique.

The new business objectives and strategies which we developed in *Define new objectives and strategies*, are the reference for mapping the data mining goals and strategies. We then develop data mining strategies, which support the *objectives* in the *data mining plan*.

A best-case scenario is that the *data mining plan's* objectives and strategies are very similar to those which were formulated in the *data mining mission's* third and fourth goal levels. In the worst case, substantial reframing took place of the hypothesis about the solution above, and the *data mining plan* has to be substantially different from the *data mining mission*.

At this stage of certainty about the data mining, the only room we allow for exploratory data preparation or data mining activities, are those which will improve the modeling techniques we selected in *Data mining discovery*. This is also the point where we incorporate any required parsimony about the data i.e. we focus our technical strategies on using those data features which have been found statistically relevant.

5.8.12 Model, evaluate, choose best model(s)



We now enter the *Realise* phase of SAM. We have broken the data mining tasks of this phase into three, to demystify the data mining activities of this phase for the business community. We are specifically concerned that the business community understands that the deployment of the data mining results is an additional step to operationalising the models. This distinction will particularly be of value in large projects, where resources may need to be justified to operationalise models and / or to deploy their outputs into the business.

The distinction also assists with understanding the cumulative nature of the technology's supports of knowledge legitimisation.

5.8.12.1 What *Model, evaluate, choose best model(s)* is

This is the first technical activity of SAM's *Realise* phase. It is also one of the KM phase *Legitimise's* technical activities. We here improve the models to better fit the data and business deliverables (Pyle 2004, pp.427ff.).

5.8.12.2 *Model, evaluate, choose best model(s)'s* support

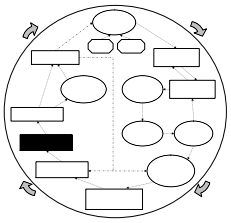
Its strategic support is optimising the resolution and stability of the discovered information for use by the business. There is no purpose relating to the relevance of discovered information, since that was established in the hypothesis legitimising loop following *Data mining discovery.* From a TQM perspective, it enhances the *utility* of the business solution toward supporting the SPC goals, through increasing the *quality* of the discovered information.

5.8.12.3 How *Model, evaluate, choose best model(s)* supports

This activity supports through optimising the effectiveness of the information we discovered in *Data mining discovery.* We achieve this through experimenting with data preparation and data mining tactics. That means – for instance – that we chose k-means clustering for segmentation in *Data mining discovery,* then we may experiment with improving the clustering results with a tactic like *hierarchical* k-means clustering, or with clustering algorithm like a Self Organising Map. Alternatively, if we used a Decision Tree as a classifier earlier, then now may experiment with an alternative classifier like an Artificial Neural Net.

We complete the activity by evaluating the improvements in information quality, and by choosing the best models based on this evaluation.

5.8.13 Operationalise model(s)



5.8.13.1 What *Operationalise model(s)* is

Operationalise model(s) is the second technical task set of the *Realise* phase of SAM. It is also a task within the KM *Legitimise* phase. It is partly similar to CRISP-DM's *Deployment* task set.

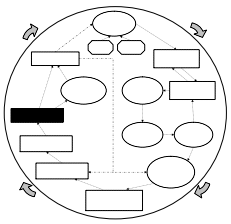
5.8.13.2 *Operationalise model(s)'* support

Its purpose is to operationalise the data mining models economically in a productive mode, supporting the business solution which depends on it.

5.8.13.3 How *Operationalise model(s)* supports

It adds business value through developing the software and hardware infrastructure, which is required for either automating the data preparation, database scoring, and the deployment of those results into the business solution environment, and minimising ongoing human involvement in the process.

5.8.14 Deploy outputs into business



5.8.14.1 What *Deploy outputs into business* is

This task is similar to CRISP-DM's *Deployment* task set. This is the last of the technical activities of SAM's *Realise* phase. It is the only technical task, which falls within the KM *Share* phase.

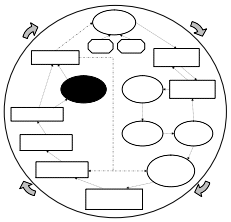
5.8.14.2 *Deploy outputs into business's* support

Its purpose is to share the information, which is discovered with the improved and selected model(s), with the organisational structures responsible for the business solution execution. From a TQM perspective, it *supports* the realisation of the *utility for purpose* of the SPC project.

5.8.14.3 How *Deploy outputs into business* supports

First, through manipulating the discovered information into a format which is most useful for the business users in their business solution development and execution. Second, by transferring that information into the information platforms of that business function.

5.8.15 Execute new business strategies



5.8.15.1 What *Execute new business strategies* is

This is the last of the reframing of data mining methodology for KM activities, and we therefore labeled it *4.4* in Figure 5.3. At the same time, it is the only non-technical activity of KM's Share phase. In Figure 5.2 it is labeled *KM 3*. It is the business activity which executes the business solution, and is therefore *5 BU* in Figure 5.1.

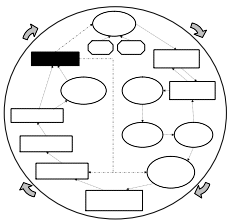
5.8.15.2 *Execute new business strategies'* support

As an KM business activity, it shares the knowledge among those within the organisation, who are responsible for the design and execution of the new business solution. Its purpose in SAM is providing assurance about the execution of the business solution, which CRISP-DM does not have. As Project Risk Management activity, it realises the *wanted event.* Uncertainty remains about the dollar benefit of the *event impact*, or project ROI. Its TQM function is *realising the utility for purpose* of the business solution.

5.8.15.3 How *Execute new business strategies* supports

Execute new business strategies is an unaltered application of SPM 5. Here, we execute the new business solution based on the now-relevant high-quality information, which we discovered in *Model, evaluate, choose best model(s)*. In this context, the data mining technology is one of the most important systems among the organisation's 7-S tools.

5.8.16 Monitor and control



5.8.16.1 What *Monitor and control* is

In Chapter 4, section 4.5 we reflected that data mining methodology needed reframing about *Monitor and control.* This technical task set *Monitor and control* constitutes our reframing in that regard. We labeled it *4.5* in Figure 5.3 to reflect this reframing. It is based on SPM's *Monitor and control* function. It is the only task set of both KM and SAM's *Monitor and control* phases.

5.8.16.2 *Monitor and control's* support

Its main purpose is a TQM one; assuring the ongoing relevance of the data mining and business solutions under the changing circumstances of an open system environment (Pyle 2004, pp.511ff.). It has a Project Risk Management purpose of removing uncertainty about the positive *event impact* of the *wanted event*; it monitors ROI over time. Its KM purpose is assuring the ongoing relevance of the knowledge underlying the business solution. From an SPM perspective, it enables the ongoing administration of the business solution under changing organisational circumstances.

5.8.16.3 How *Monitor and control* adds business value

It adds business value through producing an effective *Monitor and control* plan. The creation of the plan draws on concepts from *business administration*, *concept drift methodology*, and *statistical process control* (SPC2). We have already referenced business administration in the form of SPM, and concept drift methodology in Chapter 3. References for SPC2 are for instance (Borror 2003) (Dobler, Burt et al. 1990, pp.392-396).

We want to monitor and control two things:

- the identified informational *concepts*, which were of importance in the understanding of the problem / opportunity, and the design of the business solution. The concepts, which were important in the problem understanding, were identified in the KM legitimising activities of SAM. The concepts which are of importance for the solution, are those which are deployed into the organisation in *Deploy outputs into business;*
- the quantified business measures of success for the project.

Next, we identify the *measures* about those *concepts*, which we need to monitor over time:

- measures about the models which discover the *concepts* are for instance:
  - *confidence* - e.g. p_scores and F-statistic if the model was regression, R^2 and convergence of the clustering criterion if the model was clustering etc. In the event of Association Rule Discovery, we may add *confidence* and *support*;
  - *effectiveness*;

- *changes* in the statistical significance of features in the models; and
- features disappearing and new ones appearing;

❖ measures about business success for instance are:

- campaign *response rates*;
- *costs* associated with the business and data mining activities in the *Realise* phase of SAM;
- savings and revenue improvements from the business solution, when overlain with the costs give us ROI.

As with SPC2 we define *upper and lower bounds* for these measures, which we consider during monitoring. We develop *response strategies* for each measure about each concept, were it to rise above its acceptable bound or to fall below it. We also design a monitoring schedule, which is suitable for each concept. So for instance we would want to know campaign response rates after every campaign, but perhaps check model confidence and lift only once in six months.

Over time, we monitor the measures about each concept and evaluate if they are within or without bounds, and respond according to the developed strategy. The trigger mechanism is when the measure about a concept crosses the bounds of acceptance. The memory is the recorded measurements over time. The response mechanism is defined in the strategy, and may be automated or require human intervention.

We *execute the response strategy*. In the event of responding to a technical measure, we may need to retrain a model to improve its *effectiveness*. That can be done through repeating the data mining activities in the *Realise* phase of SAM. We indicate this control loop with the arrow between *Monitor and control* and the beginning of the *Realise* phase.

In the case of responding to a business measure, we may have to chance a *tactical* dimension of a campaign. An example could be discontinuing an incentive on a handset to upgrade. We indicate that with the control loop leading back into *Execute business strategies*.

In other cases, we may find that an ineffective campaign needs to be discontinued, or its design changed. In that case it amounts to a redesign of a business strategy in *Define*

new business objectives and strategies. We indicate this control loop with the arrow between *Monitor and control* and *Define new business objectives and strategies*.

When control fails, we need to go back into an exploratory mode. That exploration will follow the path indicated by the arrow marked *Reformulate.*

5.9 Chapter summary

This chapter contributes a data mining project methodology for the SPC environment with improved utility for delivering executable, competitive advantage. We have called this new methodology Strategic Analytics Methodology, or SAM for short. SAM is suitable for use in exploratory, directed, productive, or monitoring modes. SAM can be used with success by experienced or inexperienced organizations, in the application of advanced analytics and analytics project plans.

SAM builds on the best components of CRISP-DM, substantially reframes others, and offers innovation about data preparation. In SAM, we retained CRISP-DM's layout of a sequential, clock-wise activity and task flow, with defined interdependencies between the tasks and activities. The biggest departure from CRISP-DM, is replacing CRISP-DM's data-centric approach with a business-centric one.

We achieve this replacement by embedding the Strategic Planning Model into SAM, by way of a number of business activities, which were not present in CRISP-DM. The new business activities result in a novel sequencing of business activities and technical tasks, and in redefined dependencies between them. The reformulated data preparation tasks also reflect our business-centric approach.

The business activities introduce the very important Knowledge Management functionality required by data mining projects in the SPC environment. The KM activities define the business deliverables, focus the analytics toward supporting them, and develop the discovered information into knowledge. This knowledge best supports the business deliverables, and is also executable under the organisation's circumstances. We visualised the embedded KM process in Figure 5.2.

We further support the business-centric nature of SAM, by breaking the methodology into six sequential project phases, and giving them names which describe the project's progression toward the business deliverables. This increases visibility of the project's strategic progression compared to the visibility offered by CRISP-DM's phases. We

visualized SAM in Figure 5.1 for understanding and reference, maintaining a clear distinction between its business activities and technical task sets, their sequencing, and project phases.

We used descriptive language which both the business and business analyst communities can relate to, describing how activities and tasks progressively add business value to the project. Further, the detail about each business activity and technical task is presented in three ways; first what it is in terms of a business activity or analytical task, second the strategic support it gives to the project, and third the task descriptions for generating that strategic support. At the same time, we retained a sufficient technicality in the language to enable the business subject matter experts to accomplish their activities, and the analytical experts to accomplish theirs. So for instance, data preparation is described in terms of reducing uncertainty about the data for achieving the business deliverables. This casts the data preparation into language, which demystifies and justifies for the business community this time intensive phase of most data mining projects. At the same time we describe the data preparation as improving its desirability for applying algorithms to it, an approach the data miners can relate to.

We distinguish for both communities between the *iteration* and *repeat* of a task or activity. We *iterate* when we experiment with alternative strategies for achieving an elusive goal. We *repeat* for attaining sequential goals. We define project-relevant measures for knowing when to start and stop the iteration or repetition.

We further support SAM's business centeredness, by incorporating principles and practice from TQM in both business activities and technical tasks. TQM is used in the Operations Management environment for optimising design of products or services. We incorporated TQM in SAM through detail about defining key business and technical outcomes, and through identifying the most appropriate junctures within the project for this.

We further support SAM's business core through incorporating principles and practice from Project Management. Project Management inverts the principles and practice of Risk Management, progressively reducing uncertainty about wanted project deliverables. We devised each business activity and technical task to contribute toward

the project's business deliverables. We also develop checks and balances for assuring this progress.

We successfully reframed CRISP-DM for all six evaluation criteria in Chapter 4. We visualised this reframing of CRISP-DM by SAM in Figure 5.3, and described it in detail. In summary we:

- developed a diagnostic technique for taking the organisation from an unstructured pre-project environment, to the formulating of competitive business deliverables or goals;
- established the best time at which to inject new subject matter expertise into the project environment, and present techniques for how to accomplish that;
- devised a mapping technique between the business deliverables and the data mining goals which assures optimal support for the business deliverables;
- successfully incorporated the required knowledge management activities;
- developed a useful monitor and control task with technique; and
- incorporated soft activities like expert collaboration to pro-actively manage change management issues.

In SAM we also constituted a technical data mining project innovation; a substantial monitor and control task set with potency for developing an effective monitor and control plan. The plan covers both the business and analytical solutions. This task set is based on key principles, practice and process of the technical *concept drift* and business literature *corpi*. It describes the key business and analytics dimensions of the project, which require monitoring, and offer technical detail on how to monitor them. We also give detail about when and how to exercise tactical or strategic control in each of these dimension.

We provided a further project innovation in SAM. We provided technique for the evaluation of the discovered information against the business deliverables or goals. That technique is an initial high-level looping of the information through the knowledge managing activities *Develop circumstantial knowledge, Strategic analysis,* and *Strategic choice*. We execute this looping from within the task set *Data mining discovery*. This looping precedes the more diligent execution of that same loop in its own right, for an

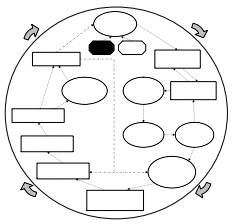
in-depth development of that information into executable knowledge, after having established the relevance of the information first.

In the following two chapters, we validate SAM in Telco ABC's retention management problem. Chapter 6 presents the validation from the start of SAM up to the end of *Data mining discovery*. Chapter 7 continues from *Develop circumstantial knowledge*, ending with the monitor and control plan.

6 Chapter 6 – Apply SAM for discovery

In this chapter, we apply SAM to the Telco ABC retention management problem. The purpose is to move-test our reframing of data mining project methodology in a practical SPC environment, establishing external validity for our argument. We facilitate the discourse with a navigational aid. The aid is the SAM diagram, with the activity or task under discussion filled in black.

6.1 Business problem



We explained in Chapter 1 that the business problem is a retention management problem in the mobile telecommunications industry.

6.1.1 The existing paradigm

Initial discussions with key people within Telco ABC, revealed that:

- there was uncertainty as to what constituted the voluntary and involuntary components of the churn event; therefore
- there was uncertainty as to the true extent of the churn problem in customer numbers and dollar figures; and
- there was uncertainty which business segment should be targeted by this new modeling approach to the problem.

We turned to the various elements in Telco ABC's objectives and strategies hierarchy, which gives a description of the problem in terms of their existing paradigm. From Chapter 2, the existing problem retention management objectives and strategies were:

1. the marketing function *Operating objective* to reduce the churn of the consumer customers from 3.5% per month to between 1% and 1.3% per month and maintain that level;
2. the marketing function *Operating strategies*:
   2.1. we *target* market retention campaigns at customers in the consumer market segment, who have 30 days or less remaining on their existing service plan;
   2.2. we *segment* the targeted customers into six need groupings based on the age demographic:

2.2.1. 16 – 21 years = peer group;

2.2.2. 22 - 30 years = personal aspiration;

2.2.3. 31 – 35 = coupling settlers;

2.2.4. 36 - 50 = family ties;

2.2.5. 51 – 59 = empty nesters;

2.2.6. 60 and over = retired;

2.3. the *campaign offer* for each group consist of:

2.3.1. a service plan from our current age-based lifestyle plans portfolio, which matches the customer's age. Some plans include a matching handset;

2.3.2. if that plan does not include a new handset, we offer a replacement handset at a discounted price, to those customers who have a handset model with reported reliability problems;

2.4. *design* the retention campaign offer in-house in our marketing department, and implement them in-house though inserts with the monthly bill mail-out, and with calls from our call center.

It was apparent that there was conflict between the preconceptions about the business problem, and the targeting of the current retention management programs. There also was uncertainty about the size of the problem.

6.1.2 Pre-project schema

We distilled the following *pre-project schema* for Telco ABC. This captures their preconceptions about the business problem, discloses their expectations on the project, and how they expected to realise those expectations through the project:

6. Telco ABC knew about the existence of the problem. They had certainty that the extent of the problem as 3.5% per month had been overstated, but uncertainty about its true extent in terms of the quantity of churners every month. This is complicated by an outstanding definition of the voluntary and involuntary aspects of the churn. The *expectation* for this stage then, was defining the churn event, determining its true extent in quantities of actual churners, and its true financial impact. The *strategies* which had been formulated for achieving this *expectation* were:

6.1. interdepartmental discussion *...sorting out for once and for all the definition of a voluntary churner...* within the organisation;

6.2. improving the understanding about the business processes and protocol for processing the churn event during the data selection stage for data mining; and

6.3. recreating the software script which defined actual past churners by the new definition, and labelling the data;

This would give a best estimate of the ongoing extent of the problem;

7. there was anecdotal evidence that the single biggest cause for voluntary churn was unhappiness with the existing handset. The cause for unhappiness was anecdotally linked to handset models, which were known for reception or reliability problems. The organisation has a cautious approach to making inference about root cause from any model, which was not part of a classical experimental design. Because of this, *no expectation was set on the project for learning root cause from the model's effects*, *and for forming any hypothesis* about understanding the problem's root cause. Unofficially though, the Retention Manager had an *expectation*, that the model's effects would confirm the anecdote about root cause;

8. there was awareness in the organisation that the current retention management solution was failing in its purpose, and awareness that data mining could also be used for supporting the development of a solution for a problem. However, this awareness was limited to using the technology for predicting in advance, who potential churners would be. Their hypothesis about solution was limited to having more lead-time in which to execute the existing retention management solution, once the identity of the potential churners was known. Their awareness did not include innovations in technical market segmentation, which could be used through data mining, to produce a solution which overcomes the deficiencies of the existing segmentation approach.

   They consequently had *expectations* on the project about identifying potential churners in advance, and for that somehow to improve the failing solution to the retention problem. There was *no expectation* on the project forming any new hypothesis about the solution from new domain knowledge or further data mining. We describe this situation as one of paradigm lock determined by the use of existing

technology and stale business domain knowledge. The *strategy* for realising the *expectation* was:

8.1. using a neural network to learn the profile of past actual churners.

Because of the above,

9. *no expectations* were formulated for the data mining project to cognitively develop a new executable solution;

10. there was an *expectation* that the data mining could support the execution of the existing retention solution. The *strategies* for realising this *expectation* were:

    10.1. scoring the customer data base monthly to identify the potential churners; and

    10.2. analysing those potential churners with traditional analytics, which were directed at executing the existing retention solutions presented in Chapter two.

    This would allow time to do the traditional analytics on those identified churners, for the monthly execution of a traditional solution.

6.1.3 Estimating the economic magnitude of Business problem

Based on the above, Telco ABC estimated the annual magnitude of the retention problem using the Formula 1 from Chapter 2:

$$\left[\frac{Q_{churn}}{Q_{acquired}} \bullet M\right] + \left[\frac{Q_{churn}}{:: Q_{total}} \times CRM\right] + R_{churn}$$

$$= \left[\frac{0.42m}{0.45m} \bullet AUD50\right] + \left[\frac{0.42m}{1m} \bullet AUD25\right] + \left[AUD400 \bullet 0.42m\right]$$

$$= \left[0.93 \bullet AUD22m\right] + \left[0.42 \bullet AUD25m\right] + \left[AUD168m\right]$$

$$= \left[AUD20.46m\right] + \left[AUD10.5m\right] + \left[AUD168m\right]$$

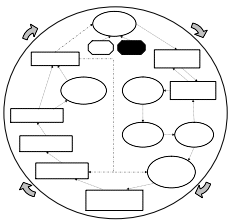
$$= AUD198.96m$$

where:

- Q_{churn} = 42% (or 3.5% per month of the average number of customers);
- $Q_{acquired}$= 45% of the average number of customers;

- M = AUD 50-00 per customer (Emagine 2003);
- $:: Q_{total}$ = 1 million customers average;
- CRM = AUD25-00 per customer;
- R_{churn} = AUD 400-00 per customer

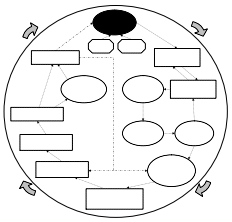
6.2 Potential solutions



We covered potential solutions as the TSP approach in the literature study in Chapter 2.

6.3 Develop project mission

6.3.1 SPC goals and strategies



In this section, we apply the diagnostic technique by SCHEMA level and formulate the SPC goals and their supporting strategies. We number the SPC goals or business deliverables according to their SCHEMA origination level number; for instance SPC 1 is derived from SCHEMA 1 and so on. Note that *SPC 1* etc. are goals, and should not be confused with *SPM 1* etc., which are entities of SAM.

6.3.1.1 Understanding the extent of the problem (SCHEMA 1)

Telco ABC is certain about the *existence* of the retention management problem. This has been confirmed before, using traditional analytics. They are aware that industry sources state the average monthly mobile telecommunication churn rate in the 2.5% region. Telco ABC therefore is *uncertain* that their estimate of 3.5% is realistic.

Further, one of the causes of the high estimate is the differing hypotheses within the organisation about what constitutes a churner. The defining of the churn event proved to be an interdepartmental issue, which needed resolution. In order to formulate a project goal for SCHEMA 1, we had to get agreement on this vital component of SCHEMA 1 within the *pre-project schema* - the definition of churn.

We evaluated the strategies within *pre-project schema* SCHEMA 1, and found that we should – and since it was a non-technical strategy, could - execute the interdepartmental discussion at this forward point in the project. The execution of the second and third strategies of the *pre-project schema's* SCHEMA 1, would have to wait until we had the

data in an assembled form. We would therefore execute them in *Identify, prepare, assemble useful data.*

6.3.1.1.1 Execute the first pre-project schema's SCHEMA 1 strategy

Four departments were involved directly or indirectly in the project, and each of these departments had a distinct perception about the definition of the churn event. The involvement and perceptions of each department were:

- Finance – the involvement of the Finance department was the provisioning of the final bill of the voluntary churned customer, after the churn event. Their definition of churn was when a customer paid the last outstanding bill. The data for the churn event was the date of payment. In some cases, *voluntary* churners did not pay their last bill, in which case Finance had to pursue them in the same way for the outstanding bill, as in the case of the *bad debtors*. In addition, in financial circles, bad debtors and fraudsters are sometimes known as *involuntary* churners. Because of this similar experience with the two different types of defaulters, and the overlapping use of the term *churner*, Finance did not distinguish between the voluntary and involuntary nature of the churn (Costa Dr. 2001, p.7). Their expectation on the retention management, was that the retention management would target bill payment defaulters overall, which would result in a significant reduction in their outstanding bill management burden. Finance was indirectly involved in this project, because their Billing database hypothetically was a source of data, which would contain signal about behavior in the lead-up to the churn event;
- Call center operations had very direct, and manifold, involvement in the project. First, the Call center operations have operational responsibility for the business protocol of processing the churn event. The Call center operators field incoming phone call from the churner, which notifies the company of a decision to leave. Their protocol when receiving such a call was to populate the call reason field, with a call reason code indicating *Churn*. Further, the Call center operator would then cancel the customer's contract while the customer was on the line, which the CRM system would date stamp. This cancellation then generated alerts to the Finance department, to make up a final bill for the churned customer, which the customer would receive as their last monthly bill. The CRM system also alerted

the Operations department, to cancel the customer's access to the mobile phone network. The cancellation was affected by Operations within 72 hours of the customer's call. It is apparent that the CRM database hypothetically contains operational data, which should carry signal about our *Business problem*. Second, the Call center operation is the main retention marketing channel; they therefore execute the retention campaigns, which are designed by the Marketing department. This CRM system is their tool, which assists them with executing the retention management campaigns. As such, the CRM system is where the data mining's results are deployed within the organisation. In light of all this, the definition of the Call center for the churn event was the receipt of the phone call from the customer announcing their intended departure. The date for the churn event was the CRM system's date stamp for the phone call. Their distinction between a voluntary churner, and an involuntary churner was clear; a *voluntary* churner phoned to announce an intention to leave, and an *involuntary* churner was when finance notified the Call center staff to deactivate a customer, because of outstanding bill payment, fraud etc.

- Operations – the Operations department manage the transmission infrastructure. From their perspective, the churn event was when they affected the cancellation of the churned customer's access to the transmission infrastructure. Their database, called CDR, contained a wealth of usage behavior, relevant to our *Business problem* and *Potential solutions.* Operations weren't too concerned about the distinction between voluntary and involuntary, because they got their notification of cancellation through the CRM system anyhow;
- Marketing – the involvement of the Marketing department firstly was to initiate this project. Second, Marketing was to design the retention campaigns, and manage their execution through the Call center department. From a Marketing perspective, there was a clear distinction between *voluntary* and *involuntary* churn, and that the project was aimed at *voluntary* churn only. Their definition of *voluntary* churn was when the *customer* made the decision to churn, and *involuntary* was when the organisation made the decision to terminate the relationship. The most practical defining for *voluntary* churn, was when the customer made the phone call to Telco ABC's call center, to notify of a decision to stop using their service.

After discussions with all the respective parties about their definitions of churn, and their expectations from the project, we reached agreement that:

- we were targeting *voluntary* churn only; and
- the date for the *voluntary churn event,* was when the customer called the Call center, to cancel his/her service. The date of the churn event therefore would be the CRM system's date-stamped phone call, with the call reason code for *Churn*;

We were therefore able to execute the first strategy of the *pre-project schema's* SCHEMA 1 at this point in the project. This would allow us to better evaluate Telco ABC's hypothesis about the extent of the churn event, and formulate more focused SPC goals in the *project mission*.

6.3.1.1.2 SPC 1 goal

Telco ABC objectives and strategies relating to churn did not refer to which of Telco ABC's business segments, we would target with the project. These business segments are *Consumer*, *Small-to-medium enterprise* (SME) and *Corporate*. Since the retention management is targeted, this omission from the *pre-project schema* was part of the *Business problem*, and needed resolution. We decided to include this aspect in SPC 1 goal. The SPC 1 goal becomes:

- determine the *extent* of the problem in both numbers and dollars within each of Telco ABC's market segments.

6.3.1.1.3 SPC 1 strategies

The strategy 1.1 of the *pre-project schema* had been executed successfully in the discussions above. That can now fall away. We then carry over the remaining strategies from SCHEMA 1 in the *pre-project schema*. As we said in Chapter 4, we did not need data mining to meet the SPC 1 goals:

- improving the understanding about business processes in the data during the data preparation stage of the project;
- recreating the software script which defined actual past churners by the new definition, and labelling the data accordingly;
- using traditional analytics to determine the *extent* of the problem by business segment;

- selecting the segment which gives the biggest positive retention management *event impact* from the modelling approach, in terms of retained income.

6.3.1.2 Understand the causes of retention problem (SCHEMA 2)

We saw in the data mining literature that we could make inference about problem root cause from the effects within a model. There were divided expectations on the project about learning about root cause. Further, there was insufficient consideration about how knowing about root cause can help with solving the problem on a segmented basis.

6.3.1.2.1 SPC 2 goals

Based on the importance of root cause analysis we saw in the literature for overcoming customer dissatisfaction, we formulate SPC goals for this second stage, to

- identify the main causes of voluntary churn and their relative impact; and to
- reframe our hypothesis about churn root cause if necessary;

6.3.1.2.2 SPC 2 strategies

Following these goals, we formulate their supporting strategies. We identify the main causes of churn through:

- *making inference* from a data mining model, which has learned information about the root causes.

We reframe the hypothesis through:

- *analysing* the existing hypothesis about root cause in light of the inference, and reframe using expert *reflection-in-action*.

6.3.1.3 Investigating a new solution (SCHEMA 3)

The reason why Telco ABC initiated the project, was that the results from their existing retention management programs, were generating below industry campaign response rates of about 25%. They considered that as a failure of the strategies. These strategies were presented as the existing paradigm earlier in this chapter.

In the third level of Telco ABC's *pre-project schema,* we found their existing hypothesis about how the data mining project would benefit the retention management program. That hypothesis based any benefit on potentially knowing the identity of the potential churners, and knowing it in advance. The benefit to the solution would then

come in executing the existing retention management strategies toward those customers, before they actually churned.

According to our diagnostic technique, we have to overlay that hypothesis (which includes the existing strategies) with the hypothesis of *Potential solutions.* We are familiar with those hypotheses from Chapter 2. Following the overlay, we have to analyse the existing hypotheses in light of the new hypothesis, and reformulate the hypotheses. We now describe this.

6.3.1.3.1 Analysis of existing Grand strategy

The *Grand strategy* to conduct retention management based on segmented market retention campaigns is accepted best practice. There was nothing in *Potential solutions* to suggest a need for improvement of the *Grand strategy*.

6.3.1.3.2 Analysis of Marketing *Operating strategy* for retention targeting

The Marketing *Operating objective* for targeting, bases the identification of the potential churners on one factor; the customer's contract expiry date. A one-factor approach is too simplistic for resolving a complex retention targeting problem. The reason we say this, is that churn is a complex problem, and there are multiple factors involved in complex problems. Such a simplistic approach as what Telco ABC is using for targeting is resulting in low accuracy in identifying potential churners. This low accuracy is one of the causes for the low retention campaign response rate, because the retention campaigns are targeting the wrong customers.

We hypothesise, that the first step to improve the response to the retention marketing is to improve the accuracy of the targeting. We achieve this through more accurately identifying the potential churners. A good way of improving the accuracy of identifying potential churners is to use an analytical approach, which can simultaneously consider a number of factors. Using data mining to build a model for identifying the potential churners provides such a solution.

We identify a further aspect of the targeting, which can be improved. The current identification of potential churners does not provide a long enough time window for the event, within which to response with retention campaigns. We hypothesise that increasing the time window for response, places Telco ABC in a position to respond

more timely with the retention campaigns. This further enhances campaign response rates. The increase of a time window is possible using data mining.

Telco ABC's current approach to targeting is unranked. By this we mean, that if a customer has a contract expiry date less than one month away, they are a potential churner. Otherwise, they are not. This means, that there are no grading of the potential of someone becoming a churner. This makes it impossible to identify those customers, with the greatest chance of churning, toward whom to prioritise retention campaigns resources. Put in Risk Management terminology, it makes it difficult to prioritise resources for the management of an *unwanted event*. We hypothesise that the non-prioritisation of retention management resources is a contributing factor to the low response rates.

We hypothesise, that a better approach to an unstructured and complex problem like this, would be a probabilistic approach. This is where customers are ranked on a continuous scale, for their *propensity* to become a churner. This propensity also functions as a Risk Management *event likelihood* factor. This nuancing, allows for the identification of the higher risk customers, which makes the prioritising of resources toward those customers plausible. Such prioritising gives better campaign response rates, while also improving the efficiency of the retention marketing dollar spent. Using data mining algorithms, which produce such continuous ranking of propensity here, helps with overcoming the deterministic targeting approach.

The interpretation of this analysis is that Telco ABC's existing hypothesis about the benefits from knowing the identity of potential churners in advance, and directing the retention program to ward them in time, is supported in part. It needs to be modified to replace the current Marketing *operating objective* about targeting customers with 30 days or less, with a targeting strategy, which rather targets potential churners based on a multitude of factors.

6.3.1.3.3 SPC 3 goal - targeting

Based on the above, the first SPC 3 goal we define is to:

- investigate the possibilities of improving the retention management targeting.

6.3.1.3.4 SPC 3 strategies – targeting

We improve the targeting through:

- identifying potential voluntary churn in advance, considering a rich combination of factors; and
- allowing for the graded prioritisation of retention management resources over a monthly response time.

6.3.1.3.5 Analysis of Marketing *Operating strategy* for segmentation

Marketing is based on meeting customers' needs. Telco ABC's existing segmentation base is demographics. We saw in *Potential solutions* that using demographics to determine needs is an indirect approach, which has fallen into disrepute among marketers. The reason for this is demographics do not reflect the underlying needs accurately enough. We hypothesise that this is the reason why Telco ABC's segmentation strategy, which is based on demographics, is not producing sufficient retention management results. We hypothesise that replacing this demographic segmentation base with a multi-dimensional behavioral one, and basing the profiling for campaign offers on that, should improve the response rate Telco ABC can get from their retention campaigns.

We saw from the literature study, that recent developments in marketing segmentation, is to base the segmentation on direct *behavioral* observation. It needs to be based on the demand function (Cooley 2003). The benefit of this, is that it is a more direct and accurate method of determining the needs of the customer. Retention offers, which closer match the needs of the customers, result in a higher response rate. In the case of the mobile Telco industry, these needs are for the different mobile phone service types e.g. voice calls, sms etc. Telco ABC has the data, which reflects customer's usage behavior in voice, sms, wap etc. We therefore hypothesise, that a first step toward improving the retention segmentation then, would be to use some of this customer behavioral data, for segmentation. This can be done using data mining techniques.

Customer loyalty is another element of *Potential solutions, which* is not included in Telco ABC's current segmentation strategy. We hypothesise that combining those behavioral elements with an indicator about a customer's *intention* to fill a need in a specific way, further enhances the segmentation base. That is because intention is an indication of the propensity that a customer may respond positively to a retention offer based on that need. The benefit of this, is that is allows for the prioritisation of retention resources toward those customers who are most likely to respond, improving the

response rate per retention dollar spent. We have a feature in the data about Telco ABC's customers, from which to identify their intention to respond positively. This intentional factor can be included in the segmentation base for this project.

Potential solutions also showed how market segmentation is complemented by *value segmentation* using RFM-A. The benefit of including this in the segmentation is that it allows for further prioritisation of retention campaign resources, allowing high retained income per retention dollar spent. We hypothesise that if we can actually create a feature in our data, which reflects a segment's value, and include it in our segmentation base, we greatly enhance the retention segmentation of Telco ABC.

The segmentation technique used by Telco ABC is a-priori. This means that the segment membership of potential churners is predetermined by human assumptions about the structure within the data. Such segmentation results in unnatural groupings of customers, and we hypothesise that this is a further contributor to the low campaign response rates. We hypothesise that we can further improve the campaign response rates by segmenting the potential churners according to the natural structure within the data, and then basing the campaign design on the characteristics of those segments. This is called unsupervised segmentation, and we can do that with data mining.

We innovated by also including a Risk Management component in *Potential solutions.* We believe that the campaign response rates can further be improved, by again considering the *event likelihood* of each potential customer in the segmentation. We hypothesise that including *event likelihood* in the segmentation base, allows for the prioritisation of retention resources toward the highest risk segments, within those customers who fall outside Telco ABC's *risk tolerance*, further improving the response per retention management dollar spent.

The segmentation technique currently used by Telco ABC has a further limitation. It does not present the segment profiles in a comparative format. We hypothesise that this is making it very difficult to prioritise the retention resources between the retention segments. With data mining, we can produce output from the segmentation, which presents the segment profiles in a relatively comparable format. The presentation is visual as well as quantitative. This achieves two things for the segmentation:

- provide for a better relative distinction between the campaign offers to each segment;

- improve the effectiveness and efficiency of the retention manager when prioritising retention resources between the segments.

6.3.1.3.6 SPC 3 goals - segmenting

Based on the above, we formulate the SCHEMA 3 SPC goal for Telco ABC, to investigate the possibilities of:

- replacing the demographic segmentation base of the targeted potential voluntary churners with a multidimensional behavioral base; and
- developing comparative segmented profiles.

6.3.1.3.7 SPC 3 strategies – segmenting

The strategies for improving the segmenting are:

- basing the segmentation on a demand function, intention to respond positively to a campaign offer, monetary value, and risk of becoming a churner;
- following natural partitions in the data; and
- establishing facilities for effective and efficient inter-segment comparative profiling.

6.3.1.3.8 SPC 3 goal – root cause profiling

We saw earlier that Telco ABC's retention management programs were failing in their purpose. The strategies about *campaign offer* are part of that failing solution. Despite this, there were no expectations in the third level *pre-project schema* concerning profiling and campaign offer design.

We know from the literature study, that addressing the root cause of the problem is a good way of preventing future customer dissatisfaction. Telco ABC's second strategy for campaign design, is aimed at addressing a hypothesised root cause of Handset Type. The approach is good in principle, but needs improvement.

Currently it is based on the *assumption* that phones are the root cause of the churn problem. The assumption is based on anecdote and dealer reports, with no hard evidence for the assumption. We can improve on this approach, by first learning from the model effects if indeed phone is a root cause. If we can indeed confirm that Handset Type is a major root cause by, then we can further learn from the breakdown of the effect labels, which phones exactly are the problem types, and rank then accordingly.

This ranking allows us to prioritise resources toward replacing the phones, which are causing most of the problem, where they are found.

Apart from being more effective, there is a further benefit to the organisation in this quantitatively supported approach. The anecdotal nature of the evidence about root cause, has frustrated efforts by the Retention Manager, to obtain resources for systematically addressing this main cause of churn. This has resulted in sub-optimal retention campaign offers, which do not address one of the major causes for churn. This could be another way of explaining the persisting high rate of churn. If the retention manager can get good proof for his anecdote, then it is easier to justify the resources required in this area.

In order to optimise any potential business solution, we therefore formulate SPC goals, which cover profiling for campaign design.

6.3.1.3.9 SPC 3 goal – root cause profiling

The SPC goal we formulate for profiling is to:

- improve root cause profiling for retention campaign offer design.

6.3.1.3.10 SPC 3 strategies – root cause profiling

The supporting strategies for that SPC goal are:

- profiling root cause on an intra- and inter-segment comparable basis.

6.3.1.3.11 Marketing Operating strategies for campaign offer execution

The strategy to design the retention campaigns in-house, and implement those through mail-outs and calls, is in order.

We now have a layered and sequential SPC goal structure (Reinartz 1999) within the *project mission*, which constitutes a *chain of models* (Wedel and Kamakura 2000, p.245) (Berthold and Hand 2003, Introduction). This greatly improves the actionable business intelligence, which the project delivers (Meltzer 2000, p.1).

6.3.1.4 Solution development (SCHEMA 4)

We saw in the *pre-project schema*, that their was no expectation about developing a new solution, based on any improved understanding of the extent and nature of the retention management problem, and which uses any discovered information about the newly introduced domain knowledge. In light of the above reformulation of elements of

the *pre-project schema's* SCHEMA 3, we need to formulate SPC goals and strategies in the *project mission*, which indicate the pursuit of a new retention management solution.

6.3.1.4.1 SPC 4 goals

The SPC goal for the fourth level is to:

- develop a new retention management solution, which will be competitive, given Telco ABC's circumstances.

6.3.1.4.2 SPC 4 strategies

The strategies for supporting this goal are:

1. developing the *executibility* of the business solution through:
   1.1. using any newly developed understanding about the extent and nature of the retention management problem (SCHEMA 1 and 2);
   1.2. using any newly developed possibilities about solving the retention management problem (SCHEMA 3);
   1.3. using SAM or a business solution design tool (e.g. SPM), which factors in the organisation's commercial and operational circumstances; and then
2. formulating integrated new retention management objectives and strategies about targeting, segmenting, and profiling, which *express the executibility*.

6.3.1.5 Solution support (SCHEMA 5)

There is an expectation in SCHEMA 5 that data mining technology can support the execution of a retention solution. However, the retention solution was the existing one. We have already formulated a SPC goal to formulate a new solution. It is possible to use data mining to support any new solution on an ongoing basis, and we now reformulate the level 5 goals.

6.3.1.5.1 SPC 5 goals

The goal is to:

- use data mining to support the deployment of the new retention solution on an operational basis.

6.3.1.5.2 SPC 5 strategies

The supporting strategies are:

1. operationalising the models of SCHEMA 3 on a required monthly basis; and
2. deploying the discovered information into the marketing business function platforms in a timely manner each time; and
3. monitor and control the effectiveness of the solution and the supporting analytics over time.

6.3.2 The departments affected

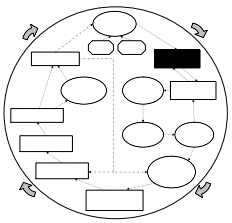
From the above, it becomes apparent that the following departments are affected, and how:

- Marketing – by having more relevant information on which to base their retention management Targeting, Segmenting, and Profiling;
- Call center - having more focused campaigns to execute, and therefore being in a better position to reduce their operating costs; and
- Analytics – designing the data mining plan, which supports this SAM project, and operationalising it. They will also be responsible for preparing the data on which the data mining project will depend.

6.3.3 Frequency and duration

We estimated a substantial economic magnitude for Telco ABC's *Business problem* above. That calculation will be revised downward later in the project, but will still offer substantial annual economic benefit were Telco ABC to improve their response rates using the new solution. At the same time, Telco ABC's *Strategic choice* to defend its existing market share, expresses an organisational commitment in principal to pursue the retention of churners. In light of the potential annual economic benefit of a SAM project, and this commitment to defend its market share, we evaluate that the *duration* of the project mission as *annual*, possible with an annual review of continuance. This evaluation is supported by the *Grand strategy* nr. two, which is the maintaining of the existing customers with retention campaigns.

6.4 Identify, assemble, prepare useful data



6.4.1 Profile for relevance

In this section, we determine which data are relevant for our project. We first profile and select the data which meets the sub-criterion *business relevance.* Then, we profile and select the data, which meets the *potential technical signal* sub-criterion.

6.4.1.1 Profile for business relevance

This section covers the activities we completed during the first application of Figure 5.4's *Profile for relevance (SPM1.2).*

The data of Telco ABC was distributed across three disparate databases. Each database is dedicated to a business or operational process; being Billing and Credit (=BC), Call Data Repository (= CallDR), and Customer Relations Management (= CRM). This is presented in Figure 6.1:

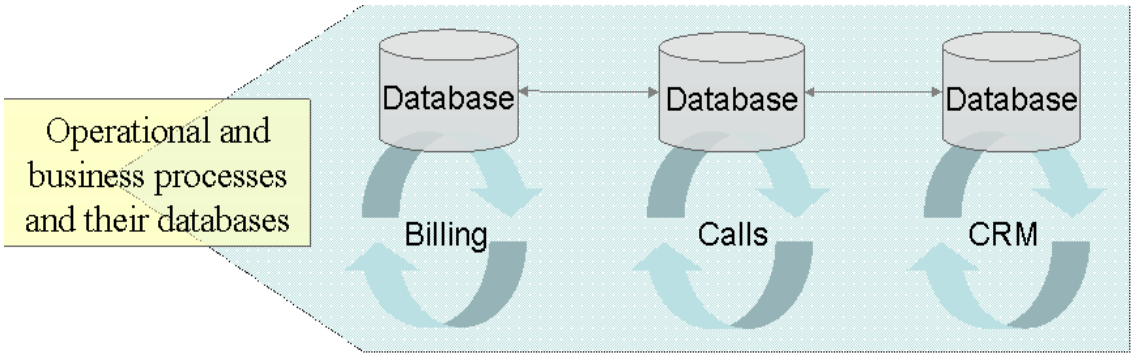


Figure 6.1: The three data domains within Telco ABC

There literally were thousands of features in the data, representing different levels of granularity about a multitude of business processes, demographics etc. Our challenge was to select from all this diversity, an ensemble of features, which were relevant to our SPC goals.

We needed data for:

- labeling the data for the churn event;
- training a classifier which distinguishes between the churn event and the non-event over the selected time window; and

- features, which could support hypothesis testing about the problem and the proposed solution.

Following evidence that change in behavior carries strong signals about impending churn (Lovelock 2000, p.151) (Steinberg 2003), we wanted a substantial behavioral component in our selected data. Considering that churn was a complex problem, there may be more factors past the behavioral, which affect or indicate impending churn. We investigated data from the following dimensions:

- Service use behavior:
  - call quantities and duration of the different types of calls (voice, sms, wap, voicemail, roaming etc.), going back 12 weeks;
  - points of call origin and termination;
  - roaming on other networks;
  - ratios of on-net to off-net calls made;
- billings for the above use:
  - discrete and summarized billing figures, and derived billing ratios;
  - recurring and non-recurring charges;
- demographics:
  - biographic;
  - economic;
  - residential;
  - social;
- psychographic:
  - loyalty history with the company;
  - mobility index of mobile phone use;
- product and service experience:
  - account provisioning details;
  - plan details re type, age, and remaining time on contract etc.;

- handset details incl. type, age, ownership etc.;

- credit information:
  - payment history (excluding bad debtors);
  - method of payment etc.;
  - identification provided;
- details of contact with the customer service center or touch point (Cooley 2003, p.598):
  - number of calls in last 3 months;
  - time since last call etc.;
  - reason codes from conversations with the call center;
  - number of goodwill credits and how recent;
- service experience:
  - number of faults in last three months and how recent;
  - number of dropped calls in last three months and how recent;
  - system reason codes for faults;
  - fault durations;
- marketing channel the customer originally was recruited through.

Our practical approach was to investigate the metadata. We encountered a typical problem which data miners experience in the commercial environment, which is that the metadata is outdated, incomplete, or even non-existent. The solution we implemented for overcoming this problem was a combination of:

- interviews with owners of business processes;
- input from domain experts;
- working through different versions of metadata from the disparate databases;
- working through metadata that had been compiled for an existing analytics data mart; and
- reverse engineering business processes from data.

Our metadata audit identified data 846 features, which we then profiled against the sub-criterion *business relevance*. We gave each of these 846 features a business relevance *score* of irrelevant/low/medium/high. We further hypothesized about later creating 40 additional data features based on the behavioral data, including spline ratios, which could contain signal about impending churn. The derived features were:

- count and duration about types of incoming and outgoing calls, within and between the two categories for both mobile and fixed numbers;
- billing ratios to other billing portions and the whole bill;
- sums of wap volumes and sms volumes;
- spline calculations (change of activity over time);
- on-net and off-net activity.

6.4.1.2 Select business relevant data

Our metadata was in Excel format. In this first application of Figure 5.4's *Select relevant data (SPM 3),* we sorted the metadata alphabetically by the column which had the rating. There were 400 features, which had a rating of high, and we selected all of them for further evaluation.

6.4.1.3 Profile for potential technical signal

In this section, we describe the activities we completed in the second application of Figure 5.4's *Profile for relevance (SPM1.2)*

We extracted 10% samples of the 400 *business relevant* features. We also derived the 40 ratios and spline features during the extracts, giving 440 features now.

Issues and experience associated with making these sample extracts are added to the augmented metadata report. They will be relied on during the next iteration through SPM, which identifies data assembleability.

We then applied SAS `proc univariate` and `proc freq` procedures to those features. These procedures report the measures of dispersion, measures of central tendency, and missing values of each feature. We analysed the results of these procedures. Examples of results were features that measured services other than mobile calls, which were between 0 and 30% populated.

An example of interpreting the analysis was that if only 30% of customers had this service, these features might not carry a signal about churn. We rated each of the 440 features against the sub-criterion *potential technical signal*, using the same rating scale as for the previous sub-criterion. For example, we rated all these features which were only 30% populated as *uncertain*.

Contrastingly, the feature, which measured contact with the Call center in the last three months, was also populated at less than 30%. There is evidence however (Cooley 2003, p.598), that recent contact with the customer service center - a touch point event - could be an important indicator of an impending decision to churn. In this case, we decided to score this feature with *high* potential signal about the hypothesis.

6.4.1.4 Select data with potential technical signal

Here we describe the activities we completed in the second application of Figure 5.4's *Select relevant data (SPM 3).*

We sorted the metadata by the column, which had this second score. One hundred and forty features had a *high* score against this criterion, and we selected them all. This reduced the number of features from 886 of interest, to 140 of *relevance*. The remaining features were deleted from the metadata spreadsheet, and this reduced metadata spreadsheet saved with a new name.

6.4.2 Profile data assembleability

This section gives a description of our application of *Profile data assembleability (SPM 1.2)* and *Identify assembleable data (SPM 3)* within SAM's task set *Identify, assemble, prepare useful data.*

We identified in *Develop project mission* that Telco ABC will want to go into regular monthly production with the data mining support. We therefore had to profile the selected features, for insurmountable issues regarding their regular monthly extraction, and assembly. This is in order to avoid futile effort in developing inexecutable strategies for data preparation.

Recall that we had extracted 10% samples of the *hypothetically relevant* features from the three operational databases, for their statistical exploration. During the extraction of those samples, we had identified a number of issues, which had to be profiled for their

effect on the *assembleability* of the data. At that time, these issues were noted against the relevant features on our metadata spreadsheet.

6.4.2.1 Profile data assembleability

In light of the identified issues, we profiled the features for meeting the assembleability criterion, by scoring the features on the scale of uncertain / low / medium / high. Features, which had no *assembleability* issues attached to them, were allocated a score of *high* by default.

6.4.2.1.1 Database access times

With database access time, we refer to the time of the day, which is most suitable for running the complete extracts against a database. Having 10% samples drawn on an ad hoc basis during the working day, was not a problem for the owners of any of those three databases. However, doing a full population extract of our tables, consisting of hundreds of thousands of observations, posed a problem for the owners of CallDR. Their server could not support that level of query activity between 6am and 10pm. We would be limited to the time window between 10pm and 6am for any regular, substantial data extraction.

6.4.2.1.2 Data dynamics

This term refers to the different update dynamics of feature values between the different databases, and even between tables within the same database. As an example, the static, attributive customer demographic data is entered into the CRM system at the time of establishing the service agreement with the customer, and then once a year or bi-yearly, when that agreement is renewed. On the other hand, event data (Cooley 2003, pp.598-600), like data about faults, calls to the service center etc are updated incidentally.

A typical situation with service providers (e.g. telephones, energy, banks), is that there are customer groups who are on four different weekly billing cycles *within the financial month*. Subsequently the features in BC, which have the billing values, are updated from CallDR weekly, but only for those customers who are in that week's billing cycle. However, the features with the credit history are updated once a month for everyone. This issue would need consideration when doing regular extracts.

The dynamics of CallDR are by the second. That means any extract will be updated for the most current behavior.

6.4.2.1.3 Synchronisation between BC and CallDR

A particularly complex and interesting data synchronicity issue came up between CallDR and the BC domains. Because of the different update dynamics between the two databases, any extract strategy will always result in asynchronicity between the call behavior data from CallDR, and the billing for those calls from BC. At best, the asynchronicity will be for a week, and at worst for three weeks. We considered that as long as the asynchronicity was constant always the same for future extracts, this would not pose a problem for data assembleability.

6.4.2.1.4 Incomplete behavioral data for churners in their month of churn

This problem relates to the fact that a churner's access to the network, was cut off within a day or two of announcing their churn to Telco ABC. Since the customer could churn at any time during the financial month, this cut-off could occur at any time during the financial month. This cut-off reflects in the churner's behavioral data as a sudden discontinuance of any activity. This is in contrast to the behavioral data for a *non-churner*, which would show a continuance of activity during a financial month. The issue here, was that such a fall-off in behavioral activity for a churner, constitutes an after the event record of the churn event. Modeling on such data would teach us the profile of someone who had already churned, and not the profile of someone who was *about to churn*.

6.4.2.1.5 Keying issues

The assembleability criterion also includes keying issues between the data. We identified a number of keying problems, mostly between the three data domains, but also within them. Examples of these are:

- in CRM data are recorded at both account and phone number levels, but the keying between tables is on the account field;
- in BC data are recorded at account and phone number levels, but the keying between tables is on phone number and account number;
- in CallDR, data are recorded and keyed at phone number level.

Once again, the metadata proved insufficient for profiling these, especially on the keys between domains, but experimentation with tools, and discussions with the IT people

gave us comfort that none of these would be insurmountable for creating one single mining table.

6.4.2.1.6 Analytics server availability

The above issues affected the extraction of data from their domains, and their keying. An issue, which we now profile, is the analytics server's availability for executing these tasks of extraction, staging, and assembly of the data on an ongoing basis. The running of the data extraction, and assembly scripts would place high demand on the analytics server, which during the working day was being used for traditional analytics. Fortunately, there is facility within SAS to schedule such tasks to run during off-peak times, so we profiled the analytics server during such times as *high*.

All considered the team formed some good ideas at this stage for overcoming all the extraction, keying, and assembly issues above. This allowed us to score all 140 features with a *high* for against the *assembleability* criterion.

6.4.2.2 Identify assembleable data

Because all our selected data also ended up with *high* scores for their assembleability profile, our activity for their identification was to update the metadata spreadsheet accordingly.

6.4.3 Develop technical desirability

By this stage, we have:

- identified those features of *business relevance*;
- eliminated those features which don't seem to have *potential technical signal*;
- determined that there are no insurmountable issues relating to automating the extraction, and assembly of the above features.

We now describe the *Develop operating strategies with data (SPM 4.4)* and *Execute data strategies (SPM 5)* in Figure 5.4. These activities constitute the *technical desirability* of the data for data mining.

6.4.3.1 Extract

The strategy we formulated for extracting the 140 features we had identified above was to do the extraction on the fourth day of the new financial month, at 10 pm. We would

extract full populations of the features, while at the same time overcoming all of the extraction issues we identified in the previous section:

- the three operational databases would be accessed during their quiet time;
- the difference in dynamics would be overcome, by simply extracting all features every month. As long as the extracts are done on the same billing cycle week every month, the problem with the 4 billing cycles is also surmounted in this way;
- the asynchronicity between BC and CallDR data, would give us resolution about behavior which would be lost if we synchronised; unsynchronized we obtain behavior over a calendar month (from CallDR), *and* over a financial month (from BC);
- this would utilise the analytics server at a time when it would not disrupt the analysts' daily activities.

We overcome the problem of the incomplete most recent month's behavioral data for churners, by extracting an additional 30 days' behavioral history for churners. We use that extra month's history to overcome this problem during the staging of the data.

About 20 different extraction programs in BASE SAS code contained these extraction strategies. The execution of the extraction code resulted in about 20 different tables residing in a staging directory, where they awaited the execution of the assembly strategies. These extraction strategies become part of the data mining plan, which is formulated in *Develop data mining plan.*

6.4.3.2 Data assembly

The strategies for assembly are captured in about 20 A4 pages. The software code with the data assembly strategies, keyed the extracted tables through a number of staged steps. The specific strategy of dealing with the incomplete churner's behavioral data in their month of churn was to replace their incomplete behavioral data for their month of churn, with the previous month's complete behavioral data – a form of relative time-based event aggregation (Cooley 2003, p.606). WE also included in the staging code the rotating of the behavioral data into one observation per phone number, which is required by all data mining software (SAS Institute 2003c, pp.7-45ff.) (SAS Institute online b, p.2).

The assembly code executed the final assembly of the 140 features into one flat table, ready for executing the following data preparation strategies.

The value of expert collaboration with a very experienced analyst was very evident from the creation of the assembly code, and it took five months to obtain a debugged flat mining table. That table is referenced in the data mining project plan as Customer Data Repository$_{\text{month t.}}$ We visually present the data preparation path in Figure 6.2:

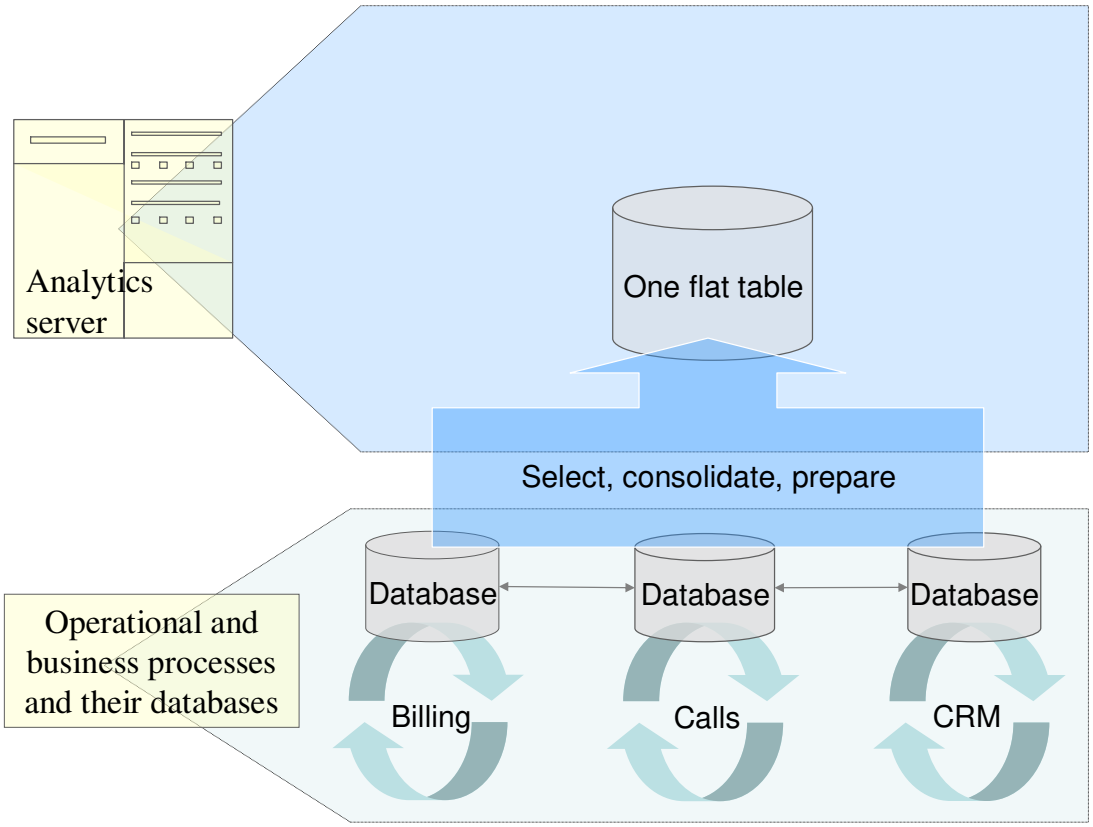


Figure 6.2: Identify, assemble, prepare data

6.4.3.3 Label the data for the churn event

This is the creation of a single feature, which distinguishes the actual past voluntary churner from the non-churner, for the building of the predictive and inferential model. In order to label the data for the above agreed definition of a *voluntary* consumer churner, we had to execute the first two SCHEMA 1 *project mission* strategies. Having the assembled data, we were at the ideal point in the project to execute these two strategies.

6.4.3.3.1 Execute project mission strategies one and two for SPC 1

These two strategies were:

- improving our understanding of the business process and protocol for processing the voluntary churn event; and
- recreating the existing software script for this updated understanding and for the earlier agreement about the voluntary nature of the churn, and labeling the data accordingly.

We executed the first strategy through conditionally exploring the assembled table. We executed the second strategy through studying the existing code for identifying churners.

6.4.3.3.2 Findings

We found that:

- the application of call reason codes by the Call center operators was inconsistent, and therefore unreliable for defining churn. We could not depend on the reason code *Churn* to identify the churn event;
- the Call center operators were using the churn processing protocols in the CRM system, for also processing handset or plan upgrades. These upgrades would then also show up as an actual churn event; and
- the existing code for identifying churners did not distinguish between a *voluntary* and *involuntary* churner.

The findings confirmed the hypothesis that the current estimate of 3.5% was inflated.

6.4.3.3.3 Response

We eliminated the *involuntary* component of churn with code which *drops* bad debtors and known fraudsters from the assembled data. We then recreated the code for identifying churners, considering that:

- we have to exclude the handset and plan upgrades;
- we can't use the label *Churn* in the call reason field; and
- we have to consider the agreed *dating* of the voluntary churn event.

Using this new code, we created a binary label in the data which best reflects *voluntary* churn. We called the label *T* and its values are '1' for a voluntary churner, otherwise '0'.

After removing the involuntary churn, and excluding the upgrades, the *voluntary* churn rate was only 2.5%. The 1 percentage point difference in churn rates (3.5% minus 2.5%) quantifies the inflation within the Telco ABC's estimate of their churn rate. This estimate had included all Telco ABC's market segments.

Revising our previous calculations for the 1% point change, we now recalculate the annual magnitude of the voluntary churn in all segments:

$$\left[\frac{Q_{churn}}{Q_{acquired}} \bullet M\right] + \left[\frac{Q_{churn}}{::Q_{total}} \times CRM\right] + R_{churn}$$

$$= \left[\frac{0.3m}{0.45m} \bullet AUD50\right] + \left[\frac{0.3m}{1m} \bullet AUD25\right] + \left[AUD400 \bullet 0.3m\right]$$

$$= \left[0.67 \bullet AUD22m\right] + \left[0.3 \bullet AUD25m\right] + \left[AUD120m\right]$$

$$= \left[AUD14.67m\right] + \left[AUD7.5m\right] + \left[AUD120m\right]$$

$$= AUD142m$$

6.4.3.3.4 Execute project mission strategy three for SPC 1

We now were in a position to execute the third SPC 1 strategy of the *project mission*, which was identifying the business segment, which would give the best return by retained income. We executed the third strategy, through writing code for a cross-tab of the churn event x business segment, and executing the code on a 20% random sample of the mining table. We then summed the ARPU within each segment.

6.4.3.3.5 Findings

The information we discovered executing the third strategy, was the portions of the ARPU, which resides in each business segment. We developed the knowledge that 80% of the ARPU of *voluntary* churn, resided in the *consumer* business segment. Further, the churn rate in this segment is 1.5%. When we divide the 1.5% of the *consumer voluntary* churn with the 2.5% of *overall voluntary* churn, we found that 60% of the incidence of *voluntary churn* resided with the consumer market.

6.4.3.3.6 Execute project mission strategy four for SPC 1

We now needed agreement on which of the three business segments will be targeted with the new modeling approach. That was the fourth strategy of SPC 1 in the *project mission*.

After team discussions with the Marketing department and the Retention Manager, we had agreement to focus on *voluntary* churn in the *consumer* segment, where 80% of the lost revenue resides in 60% of the incidence. We had to adjust the AUD142 m. by a factor of 0.80, to obtain the annual magnitude of the problem in the *consumer* segment. After adjustment, the magnitude is AUD113 m.

Since the 1.5% constituted an average churn rate, we decided to base the *risk tolerance* figure on this, making the *risk tolerance* figure 1.5%. That constituted agreement, that we would target the 1.5% most-at-risk part of the database in the retention management campaigns each month.

6.4.3.3.7 Response

We applied code to the assembled table, which dropped all observations, which did not belong to Telco ABC's *consumer* segment.

6.4.3.4 Clean

Our strategies for cleaning the data encompassed imputing missing values, replacing existing values with more meaningful values, eliminating collinearity in the data, and filtering outliers.

6.4.3.4.1 Eliminate collinearity

We wanted to avoid the trap of multicollinearity in our regression model. Our strategy was using R^2 correlation matrix for identifying collinearity (SAS Institute 2003a, pp.4-3ff.). We used `proc corr Spearman` in BASE SAS to generate the correlation matrix. The choice for the Spearman option over Pearson is because the Spearman test is rank ordered. This is suitable for calculating correlation when the data consists of both interval and categorical measures.

We then identified those features which were correlated 0.65 and higher. The choice for which of the correlated features to retain and which to drop, was considered for keeping our spread of dimensions, which we wanted to maintain.

We retained 54 interval and 16 categorical features, retaining 70 features in the table.

6.4.3.4.2 Replacement

Since some of our imputation strategies depended on data values in other variables, we first formulated and executed the strategies for replacing existing values. Examples of replacement strategies were:

- the categorical variable PLN (which is the service plan name) had in excess of 150 values. Left untreated, such cardinality was going to cause problems in two ways. The first was an IT resource problem, which was going to manifest itself during the building of the Regression model. This is because LogReg transposes each value of a categorical variable into a new dummy variable, and each of these dummy variables is then regressed as a factor. The number of variables (including these dummy variables) in LogReg, has a quadratic effect on resource requirements. We therefore had to reduce the IT reduce demand;
- the second problem was that of *operationalising of the levels* referred to before; plan is one of the tangibles of the retention campaign offer. Any profiling, which may be done on plan at a later stage, will result in an unworkable level of detail. We therefore had to reduce the number of values for plan.

The replacement strategy for PLN was to decimate the values into about 15 categories.

Further replacement strategies resulted in:

- geographic dispersion being reduced to one third of the original values;
- residential status types being halved, including eliminating multiple spellings of the same values;

A further replacement strategy was required to overcome a limitation of SAS Enterprise Miner's Regression node. That node limits the length of fields to 16 characters. This means values, which may be spelled the same over the first 16 characters, are read as the same value. The strategy was to replace duplicate values within the first 16 characters, with short two or three character codes.

6.4.3.4.3 Imputation of missing values

The imputation of missing values is required for practical purposes foremost. Both LogReg and k-means Clustering algorithms - which we plan to use - ignore observation

with missing values (SAS Institute online b, p.9). This may result in models of lower accuracy than what is otherwise possible. We used three imputation strategies, each dependent on the complexity of the imputation problem:

- define a general default replacement value;
- specify a unique or principled value (Pyle 2003, p.376);
- identify a value dependent on a combination of other, non-missing values of that same observation, based on observations where the value is present. This is achieved using decision tree rules.

Because we planned to use the third strategy too, we did replacement before imputation.

We further made provision for the handling of future unknown values of categorical variables (Pyle 2003, p.383). We expected the score code to encounter unknown values for plan type and handset type in future data, during the *Realisation* phase. The origin of such 'unknown' values would be new plans and handset types, which had been sold in the two months since the model had been trained. The strategy for coping with this is through setting some options in the Replacement node of Enterprise Miner.

We decided to exclude from our modeling, indicator variables which are populated wherever a missing value had been imputed. Usually these features are included to detect associations between missing data, and the target. This is a similar concept to Pyle's Missing Value Check Model (Pyle 2003, p.373). The reason we left them out, was that we expected at the time that we had enough signal in the data without this category.

We demonstrate the intricacies and interweavedness of our data, and their implications for replacement and imputation, with some examples given in Table 6.1:

Name	Status	Model Role	Imputation Method	Replace <	With Value (<)
LOYP	use	Input	tree imputation with surrogates	.	.
TOUT	use	Input	Set value - 99	1	0
TOUQ	use	Input	default constant - 0	.	.

Table 6.1: Replacement and imputation

Note in Table 6.1, a tree imputation method with LOYP. One of the dimensions of similarity that the tree was authorized to consider, was DemB, which contains values of customers' marital status. DemB however, had a discrepancy in its values, with two different spellings of the value 'De Facto'. If such a discrepancy were left in DemB, the imputation decision tree would have a separate rule for each spelling of 'De Facto', resulting in different imputation values for what essentially is the same marital status. The most practical way of avoiding such issues, is first to do the replacements in the data, then the imputations of missing values.

The values used for default imputation of missing values were '0' for interval variables, and '#' for character variables. The example TOUQ in Table 6.1 is a numeric variable where we imputed a missing value with a default value. The reason is, TOUQ carries the count of contacts a customer has had with the call center in the last 92 days (three calendar months); a missing value here simply means that there was no contact, which is '0'.

The imputation strategy pays off later, when TOUQ is identified by χ^2 as having a strong enough association with the target, as to be included in the 12 variables that are passed on to LogReg.

The imputation strategy for missing values of TouT was to replace them with a set value of 99. The reason is, that when a feature represents time lapsed since an event in the last 92 days, a missing value can't be imputed with '0', because we've just seen that '0' means 'on the same day'. Missing values could also not be imputed with any other number in the range of 1 – 92, because that would distort the actual values in that range. Since TouT has a numerical format, we also cannot impute with a distinctive symbol - like '#' - for instance. The remaining choice was to impute with a value which falls outside the range of 0 – 92, hence 99. Further, there is a sufficient gap between '92' and '99' to make it possible to easily place '99' in it in its own bin at a later stage.

Table 6.2 displays the creation of a replacement value 'DF', with which both other spellings of 'De Facto' and 'De-facto' are replaced in the data:

Name	Replace Value
DEMB	De Facto=DF, De-facto=DF

Table 6.2: Replacement values

6.4.3.4.4 Filtering outliers

Name	Commercial Meaning	Range included or excluded
LOYR	Indicator for time left on existing contract	Excluded about 0.1% at both ends of the distribution, eliminating some evidently erroneous values
LOYP	Loyalty indicator	Excluded about 0.1% at both ends of the distribution, eliminating some evidently erroneous values
REV53	ARPU	Included all values between '0' and max minus 0.4%, to eliminate what appeared as either exceptional active consumer call behavior, or commercial accounts that slipped past our definition of consumer
SUMC	Sum of voice calls in the last 3 months	Included all values between '0' and max minus 0.4%, to eliminate what appeared as either exceptional active consumer call behavior, or commercial accounts that slipped past our definition of consumer

Table 6.3: Outlier filtering

There were two reasons to filter some outliers. First is that the k-means clustering algorithm – which we use for creating the features SegMas and DistMas, and for the later retention segmentation of the predicted churners - is sensitive to outliers (Simoff 2003, p.41) We applied filtering strategies to four interval variables, as displayed in Table 6.3.

6.4.3.5 Create additional features

We decided to derive seven additional features, which were also presented to variable selection later. Six of these features did not make it past the variable select by χ^2.

Four of these derived features were ratios, which were not predictive:

- monthly mobile charges / total phone bill;
- monthly sms calls to each of four competing carriers / the sum of sms calls to all carriers;

- calls from each of week eleven weeks (weeks 12 to 2) / week 1 calls. These are called spline calculations;
- calls made to on-net numbers / calls made to off-net numbers.

The two other derived features, which were not predictive:

- sum of dropped calls over the last 3 months;
- sum of sms over the last 3 months;

The derived feature, which proved an association with the target event by later χ^2 variable selection later, was:

- SumC, which was the sum of voice calls, made over the last three months.

After creating these seven additional features, there are 77 features in $CDR_{modplus\ t}$; 61 interval, and 16 categorical.

6.4.3.6 Segmentation of $CDR_{modplus\ t}$

In this data preparation step, we add two more features to our data. There is evidence in the literature that cluster membership, and even the distance from the cluster mean, may have some association with the values of the dependent variable (SAS Institute online e, p.1) and (Berson, Smith et al. 2000, pp.161ff.). It is a good technique for discovering competing patterns within the data (Westphal and Blaxton 1998, p.xiv). Wedel et al. explain that clustering captures homogeneity within local data areas (Wedel and Kamakura 2000, p.306).

We created these two new features from a 20% sample of $CDR_{modplus\ t}$, using hierarchical clustering, and a clustering base of three interval variables, LoyP, SumC, and Rev53. The new features are"

- DistMas - the distance of each observation from its cluster center;
- SegMas - the segment membership of the observation;

This technique groups observations by similarity of a profile, which was previously unknown (Meltzer 2000, p.9).

The 20% sample gave 42 000 observations, and there is evidence that this is a minimum number from which to produce robust clustering results (Woods 2003b, p.144). The number of clusters was set at ten, and I set the option for seed replacement as 'full',

which optimizes the discovery of distinct profiles from the data. The three clustering variables are each measured on a different scale:

- Rev53 is measured in dollars;
- SumC as a count of events;
- LoyP in a unit of time lapsed.

We standardized to equalize the influence of each feature's variability on the clustering (Woods 2003a, pp. 2, 16ff.). In the standardisation, we did not subtract the mean (Simoff 2003).

The strategy for clustering was a sequential combination of non-overlapping, disjoint clustering followed by hierarchical clustering. The first stage is by k-means, which is the partitioning of observations in such a way as to minimize the Euclidian distance between the standardised variable value of the observation and the standardised variable mean for a cluster (Simoff 2003, pp.15, 35). The algorithm runs `proc fastclus` to achieve the disjoint clustering. Then it switches to `proc cluster,` where I selected the Ward technique. This technique preserves the most information from the previous step (Woods 2003b, p.22), since it considers the sum of squares between two clusters ADDED UP OVER ALL OBSERVATIONS, and merges clusters to minimise the within-cluster sum of squares over the previous generation (SAS Institute online e, p.8) `Proc cluster` iterates until meeting the Least Squares clustering criterion cut-off is met, obtaining a more meaningful overall clustering (Simoff 2003, p.27). HOWEVER, the new cluster centers that were calculated by `proc cluster` are submitted to `proc fastclus` for one final iteration (Woods 2003b, p.27), which generates the final cluster representation.

Statistics for Variables

Variable	Total STD	Within STD	R-Square	RSQ/(1-RSQ)
SumC	1.00000	0.45417	0.793740	3.848251
Rev53	1.00000	0.43253	0.812926	4.345480
LoyP	1.00000	0.42486	0.819501	4.540212
OVER-ALL	1.00000	0.43736	0.808723	4.228007

Pseudo F Statistic = 111808.7

Approximate Expected Over-All R-Squared = 0.78457

Table 6.4: Clustering base statistics

Table 6.4 gives the measure of:

- how much of the structure of data was captured by the clustering (Simoff 2003, p.11). This is the Over-All R-Squared (in Table H) of 0.80. This is a good outcome for basing commercial decisions upon (Woods 2003b, p.8). Put another way, it measures the success of the clustering basis (independent variables), in determining the dependent variable SegMas (Ray 1997, pp.110ff.);
- the relative contribution of the components of the clustering base toward determining cluster membership. The contribution of the base features was almost equal, which is a desirably smooth outcome.

We know from the later quantitative variable selection, that segment membership (SegMas) has a sufficient association with the dependent variable p_t1 to be selected for inclusion in the modeling. DistMas will be rejected by the execution of the quantitative variable selection strategy.

Adding these two extra features, there now are 79 features in $CDR_{modplus\ t}$; 62 interval and 17 categorical.

6.4.3.7 Sampling

We are dealing with a rare event (SAS Institute 1998, pp.15, 19), which in the Telco industry typically is in the order of 2 – 3% per month (Mozer, Wolniewicz et al. 1999, p.1). Under such circumstances, the most suitable sampling technique is stratified, equal proportion random sampling (Levin 1987, pp.280ff.) – also known as 'enriched sampling' or over-representation (SAS Institute online b, pp.16, 17, 20, 37). We based the stratification on the response variable (SAS Institute 1998, p.25) (Steinberg 2003). This increases the prior probability of the event to a level where the signal can be more effectively extracted by the modeling. However, it also introduces bias into the posterior probabilities when they are overestimated. This bias is then compensated for, by adjusting the prior probability in the target profiler, to reflect the actual proportion between the strata in the population. The adjustment is through moving the intercept down the x-axis again. The sample size was 10% of $CDR_{modplus\ t}$, and the sample file we called Modeling Data Set (MDS_t)

6.4.3.8 Dealing with data Distribution issues – final data transforms

The label feature *T* we created earlier, is a categorical variable with a binomial distribution (Woods 2003a, p.3). Figure 6.3 gives an example of such a binomial distribution.

Under such circumstances, the dependence on the Central Limit Theorem for making statistical inference becomes trivial (Levin 1987. pp.294ff.), and we therefore do not strive no maximise the normality of the distributions of any of our interval variables. Rather, there is evidence that a strategy of grouping independent interval variables in the presence of a categorical response variable enhances the predictive outcome (SAS Institute online c) (Woods 2003a, pp.42ff.). The strategy is to group the values of the independent interval variables into equi-depth bins or equi-width buckets (Han and Kamber 2001, pp.110, 125ff.). This strategy has an additional advantage, of unlocking any non-linearity, which may be in the relationship between the independent and dependent variables (SAS Institute online d, p.10).

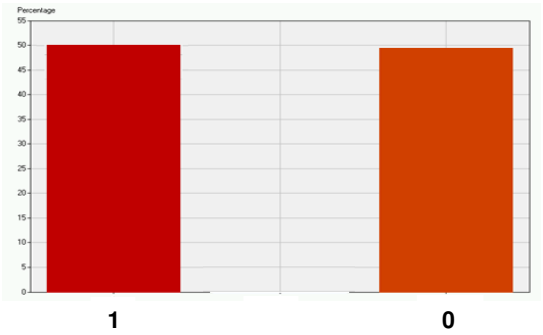


Figure 6.3: Binomial distribution of the label feature

The strategies subsequently were

- grouping all independent interval variables using a technique that optimizes the groups for association with the target. This approach takes into account informational patters between that input variable, and the label (Pyle 2003, pp.379ff.);
- in some instances, that technique did not return any meaningful grouping, in which cases we then grouped by bucketing into quintiles;
- grouping using both optimal binning and quintiles in the case of LoyR.

By example of LoyP, we present its non-transformed (standardised) value distribution in Figure 6.4. The feature had been standardised using the z-score method:

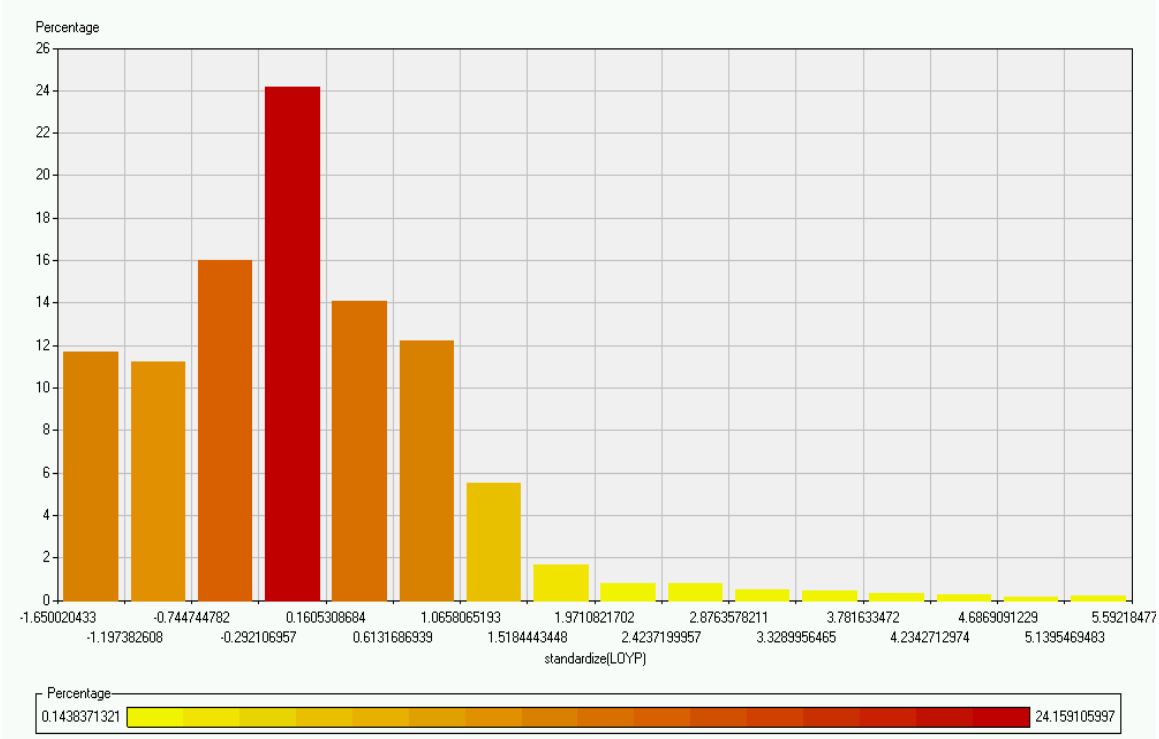


Figure 6.4: Distribution of LoyP (standardised)

Figure 6.5 presents LoyP after the optimal binning to T. The bin boundary values have been hidden for confidentiality reasons:

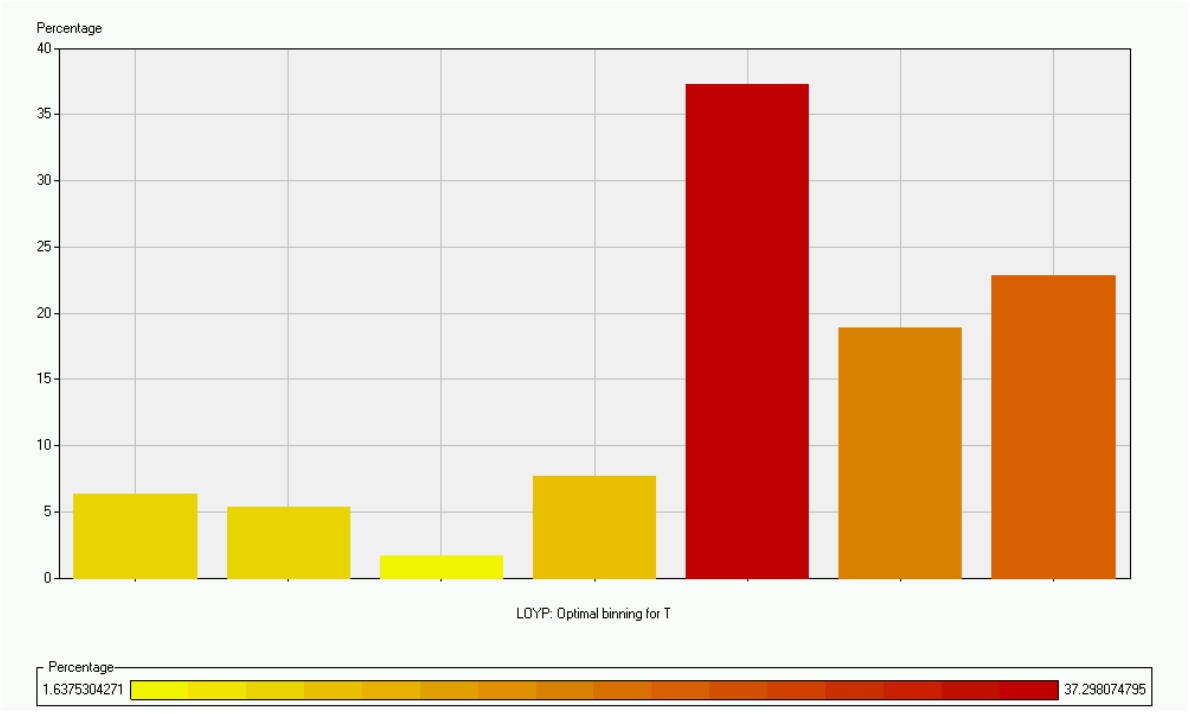


Figure 6.5: Distribution of LoyP after binning

When we binned and interval feature, we dropped the non-transformed features from the data, except in the case of the three features we were going to base our retention segmentation on. These three features were required for the demand and response functions of the retention segmentation. The plan was to use k-means clustering for retention segmentation, which is dependent on interval features.

At the same time, retaining these features overcame the difficulty of *operationalising of the levels* (Wedel and Kamakura 2000, p.298). What this refers to, is that although the grouping of the data optimises the predictive accuracy, the binned variables make the design of a response impractical. The output from the k-means segmentation, then gives the average value of each retention segment for these non-transformed features. Such average values are practical for retention campaign design and execution.

None of these three non-transformed features made it past statistical variable selection later, because their bins were more predictive. We therefore manually excluded them from the model input.

At this point, 82 features remain in the data; 65 binned interval features, and 17 categorical features, including the churn event label.

6.4.3.9 Statistical feature selection

Building a model with 82 features will place overwhelming demand on IT resources, and will produce a model which is costly to maintain, and too complex to understand. We therefore needed a strategy for discovering the statistical influential independent variables (Ray 1997, p.41). We list here a number of strategies we considered for achieving this, and motivate our choice:

- R^2 – is a linear measure of any relationship between the independent variables and the dependent (SAS Institute online d, pp.9ff.). The segmentation of the table earlier had resulted in capturing only 80% of the data structure by R^2. From this, we inferred that there would be some non-linearity in the data, in relationships between the independent variables and the dependent variable. Although R^2 can be used with a binary target, it measures linear relationships. We therefore rejected R^2 as a strategy;
- stepwise regression – is very effective for variable selection when there are ordinal variables (SAS Institute online d, p.1), but we had only three ordinal

variables in the data. This method is not very efficient when there are many features present (Kolyshkina, Petocz et al. 2003, p.230) and (SAS Institute 2003a, p.4-4). For these reasons using regression was rejected as a strategy;

- χ^2 criterion – is very efficient and effective for feature selection with a binary target, especially at 95% confidence for rejecting H_0. It also detects non-linearity in the relationships between the independent and the dependent features (Connor-Linton 2003). We selected this method. The execution of the strategy is creating dummy variables for each value level of categorical variables, and binning interval variables to 50 bins. Then drawing 2x2 contingency tables from these dummy variables and bins. A decision tree – targeted at the churn label - is then built to obtain node statistics for the dummy and binned variables, and those variables that pass the test of significance, are selected.

Twelve features pass the above significance test, and they are presented in Figure 6.6:

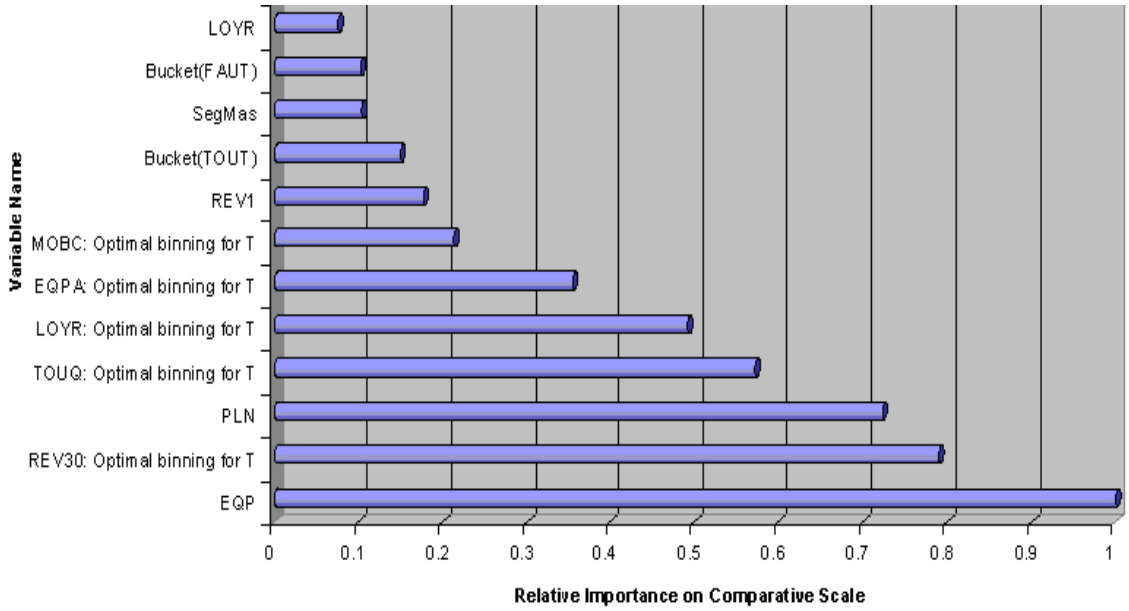


Figure 6.6: Relative feature significance by χ^2

At this point, we manually intervened by including DemD – a demographic indicator of wealth – to make provision for company anecdote that this feature was a contributor to churn.

There is evidence (Apte, Bibelnieks et al. 2001) that the accuracy of propensity models, are improved if the data is clustered, and then a regression model is built within each cluster. We did not have the resources to experiment to this extent. However, we

substituted by creating the SegMas feature, and including that in the modelling. Note that SegMas - which was created from the clustering of the master table $CDR_{modplus\ t,}$ was significant. DistMas did not pass this significance test.

We are also fortunate that the two main features, which will be of use for retention segment profile development (EQP and PLN), are ranked in the top three by association with the dependent variable. The measures of the features, which are passed through for modeling, are displayed in the middle column of Table 6.5:

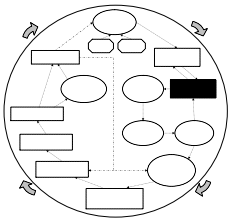
Variable label	Variable Measure	Presence in Number of Tree Nodes
EQP	Categorical	9
REV30: Optimal binning for T	Categorical (Binned interval)	2
PLN	Categorical	3
TOUQ: Optimal binning for T	Categorical (Binned interval)	3
LOYR: Optimal binning for T	Categorical (Binned interval)	2
EQPA: Optimal binning for T	Categorical (Binned interval)	2
MOBC: Optimal binning for T	Categorical (Binned interval)	3
REV1	Ordinal	1
Bucket(TOUT)	Categorical (Binned interval)	1
SegMas	Categorical	1
Bucket(FAUT)	Categorical (Binned interval)	1
DemD	Categorical	1

Table 6.5: Selected variable importance

6.4.3.10 Data partitioning

We partitioned MDS_t randomly into $Training_t$ and $Validation_t$ sets, to enable 'split-sample' validation of the propensity model to be created. We equally stratified the partition on the churn label *T,* so that each partition contained equal numbers of churners and non-churners. The split was 70/30 $Training_t$ / $Validation_t$.

6.5 Data mining discovery



6.5.1 Develop data mining mission

In *Develop project mission* we successfully executed the first strategy pertaining to the SPC 1 *expectation*. Despite that, that *expectation* remained fully unmet, and we carried it – and its remaining strategies - over into the *project mission* as the SPC 1 goals, and their four supporting strategies. We then successfully executed the four strategies during *Identify, prepare, assemble useful data.*

At this point then, the SPC 2 and higher-level SPC goals of the *project mission* are unattained. In this section, we develop the *data mining mission, which* pertains to these outstanding SPC goals. We follow this up with an execution of the strategies, and then with the KM activities associated with each. We repeat this pattern of presentation for each of the remaining SPC goals.

6.5.1.1 Data mining goal for SPC 2

Based on the data preparation activities, we now know that we have sufficient data for building a model for making inference about root cause. We link the data mining goal to the SPC goal, and apply the bridging technique to base the data mining goal on the SPC goal. The data mining goal we formulate is to:

- discover information about the root causes of the voluntary consumer churn problem, and the relative impact of the root cause elements on the churn event.

6.5.1.2 Data mining strategies for SPC 2

The supporting strategies we formulate are:

- building a classifying model which distinguishes in the data between the profile of a voluntary consumer churner and a non-churner; and
- making inference from the model effects about the elements of root cause, and their relative impact.

This places a condition on the classification model. It should be of a type which produces parameters, which are transparent and understandable enough, to base inference on (Wedel and Kamakura 2000, pp.22, 101) (Hastie, Tibshirani et al. 2001, p.99). This disqualifies the Artificial Neural Network, which we found in the SPC 2,

since that algorithm does not produce transparent effect scores. In stead, we choose the Logistic Regression.

6.5.1.3 Determine the where for SPC 2

The data we use for inference is the mining table. At this stage we do not know which features are the important ones, that being the purpose of discovering through the modeling.

6.5.1.4 Set confidence levels for SPC 2

The *confidence measure* cut-off level we set for this model is 95% for the null hypothesis. The lower limit for *lift* we set at five times.

6.5.1.5 Data mining goals for SPC 3 - targeting

We link the data mining goal to the SPC strategies, and apply the bridging technique to base the data mining goals on the SPC strategies. The data mining goals we formulate for improving the retention management targeting are to:

- discover multi-dimensional information which identifies the chance of a consumer voluntarily becoming a churner within at most the next three months;
- identify the most at risk voluntary churners.

6.5.1.6 Data mining strategies for SPC 3 - targeting

The supporting data mining strategies for the SHEMA 3 data mining goal above are:

- building a Logistic Regression model using three months' voluntary consumer churners' and non-churners' behavioral history and other static features, learning the profile of actual past voluntary consumer churners; and
- scoring the consumer database with the model (Westphal and Blaxton 1998, pp.xv, 71) (SAS Institute online b, pp.1, 7); and
- sorting the consumer database descending by their probability score from the model for becoming a voluntary churner; and
- targeting the top 1.5% of that sorted consumer database for further attention.

We discounted the use of an Artificial Neural Network for reasons of the previous section. Here in SPC 3, we also considered the use of a Decision Tree as the classifying algorithm. The Decision Tree however, produces the p_scores in ranges, and not

continuously as the Logistic Regression does (SAS Institute online b, p.8) (Ridgeway 2003). Trees also require a third test set for true performance (Hastie, Tibshirani et al. 2001, p.221). Creating three, smaller partitions in our already scarce data, may actually reduce the accuracy of a tree, which we have built such small data sets. For this reason, we discount the use of the Decision Tree in SPC 3.

Logistic Regression (LogReg) is very suitable for modeling a binary target (Woods 2003a) (Kolyshkina, Petocz et al. 2003, pp.229). LogReg also produces a linear model, which tends to be more robust on noisy commercial data (Hastie, Tibshirani et al. 2001, pp.80, 96). This extends the model's useful lifetime, which is an important consideration in the commercial environment.

6.5.1.7 Data mining goal for SPC 3 – segmenting

Linking to the SPC strategies, and using the technique of making those the data mining goals, we formulate the goals to:

- discover the natural four-dimensional segment membership of each targeted potential churner; and to
- develop a facility, which allows for visual and quantitative inter-segment comparative profiling.

6.5.1.8 Data mining strategies for SPC 3 – segmenting

The strategies we formulate for supporting these goals are:

- segmenting the targeted top 1.5% of the consumer database with an unsupervised k-means classifier, using as segmentation base the data features *ARPU* (Average three month Revenue Per User), *P_T1* (the probability score generated by the Logistic Regression), *LOYP* (a derived feature of intention to respond), and *SUMC* (the three month average number of mobile voice calls); and
- summarising in table form the segments' quantitative measures of proximity and dissimilarity on a comparable standardised basis; and
- visualising that standardised summary.

6.5.1.9 Data mining goals for SPC 3 – root cause profiling

We link to the SPC strategies in the *project mission*, and apply the bridging technique to formulate the data mining goals. They are to:

- discover the frequency of each label of the main root cause effect within each segment; and to
- present that in intra- and inter-segment visual and quantitative comparable formats.

6.5.1.10 Data mining strategies for SPC 3 – root cause profiling

The data mining strategies, which support these goals, are:

- using the segmented label frequencies output tables from the clustering node; and
- developing a tool, which presents those frequencies as intra- and inter-segment percentages of root cause effect.

6.5.1.11 Determine the where for SPC 3

The *where* for SPC 3 strategies is:

- the flat mining table we originally created; which is
- statistically reduced to the most associated features with the churn event;
- the consumer database of one month which is scored with the model and sorted descending by propensity of becoming a voluntary churner; and then
- the top % of that database which is selected by Telco ABC's *risk tolerance* (Chapter two);
- the four data features which are used for the segmentation base; and
- the data feature, which underlies the model effect most, associated with the churn event.

6.5.1.12 Set the confidence levels for SPC 3

The model for inference of root cause in SPC 2, and the predictive model in SPC 3, is the same. We therefore carry over the *confidence measure* cut-off and the *lift* requirement from SPC 2.

The second model, which we use in SPC 3, is k-means clustering. Confidence there is a more fluid concept as with supervised classification, and determined by a number of measures. The first is the degree of latent data structure, which has been captured. The measure for that is $R^{2.}$ This is a linear measure, and the shortfall between this figure and 100 percent, is the result of non-linearity in the data, and sub-optimality in the tuning of the clustering algorithm. We feel confident if we have captured a 70% overall R^2.

The second measure of confidence with clustering is the number of distinct profiles, which were created from the data. We measure this as the distinct number of nearest clusters among the clusters. The higher the number of distinct nearest clusters among the clusters, the more confident we are that the cluster profiles are distinct. In following Telco ABC's 6-segment strategy, we create six segments. We will be confident about this measure, if there are at least three distinct nearest clusters listed in this measure.

A third measure is of confidence with clustering, is that the clustering result has converged to the *convergence criterion*. This is a measure of the improvement, which is made in the objective (below), every time the algorithm iterates. Convergence means that one more iteration will not significantly improve the results. We feel confident if the algorithm has attained convergence against this criterion.

The objective of the clustering algorithm is to minimise the *Root-Mean-Square Standard Deviation* while maximising the *Distance to Nearest Cluster*. A combined measure of the success of the simultaneous achievement of these two goals by the algorithm, is *Fisher's Criterion* (Heckel 2003, p.100ff.). *Fisher's Criterion* is the ratio *Distance to Nearest Cluster* / [the sum of the *Root-Mean-Square Standard Deviations* of two nearest clusters)] The higher the value of this criterion, the greater the informational gain from clustering the data. A *Fisher's criterion* therefore of greater than one becomes our fourth measure of confidence about the clustering.

Last, our confidence in the results of the clustering is influenced by the *relative importance of segment base features*, in determining the structure within the data. One would prefer to see a spread of importance over all the base features, rather than one or two features dominating.

The segmental frequencies of the root cause label in the campaign offer profiling, are the results of simple addition, and do not require statistical confidence measures.

6.5.1.13 Comment on SPC 4 and 5

We place a reminder that the goal and strategy about new solution development in the project mission, are non-data mining in nature. This SPC strategy is executed in the business activities, which follow on *Data mining discovery.*

The goals and strategies for the SPC 5 – the operationalising of the models supporting the new business solution – takes place in *Develop data mining plan.*

6.5.1.14 On executing data mining mission for SPC 2

The data mining goals above for SPC 2 and 3 are different. SPC goals 2 and 3 also have separately worded data mining strategies. However, when we compare the mechanics of these strategies, we find that the strategy for SPC 2 is identical to the first strategy for SPC 3's targeting; we will be executing both those strategies in the same Logistic Regression model.

For this reason, we include the execution of the *data mining mission's* SPC 2 strategy with the execution of the first strategy of the *data mining mission's* SPC 3. We do, however, execute their KM loops separately. This means that the reader will find the headings associated with the profiling of the technical results and the knowledge management, in the following section of SPC 3.

6.5.2 Execute data mining mission for SPC 2 and 3

6.5.2.1 Execute SPC 3 strategy one and SPC 2 strategy

6.5.2.1.1 Build the classifier

We built a LogReg model, using the stepwise regression method. The inputs were the statistically significant features we had identified during the data preparation in MDS_t. We described the split of MDS_t into training ($Training_t$) and validation ($Validation_t$) sets on a 70/30 ratio before. The model assessment was against validation, and we assessed for profit maximization (SAS Institute online b, pp.38ff.). The profit matrix was a profit value of '1' for accurately classifying a churner, and a profit value of '0' for accurately classifying a non-churner. This biases the profit toward correct classification of the churn event, through weighting the calculation of the Betas in favour of the churn event, moving the equal decision boundary closer toward the churn event, resulting in a better identification of potential churners (Ridgeway 2003, pp.163, 170).

We followed the analytic approach in training error optimisation, having selected to optimise Schwarze's Bayesian criterion (Hastie, Tibshirani et al. 2001, p.203). We set the prior probability of being a churner to the true prior probability, which was 1.5%.

The model produced code, which could be used to score the consumer database - called $CDR_{month\ t}$, giving each observation a p_value for becoming a voluntary churner in the next three months.

6.5.2.1.2 Profile technical results - classifier

6.5.2.1.2.1 Effectiveness

The accuracy of the LogReg modeling on the validation data set was a misclassification rate of 47.5%. This translates into an accuracy of about 52.5% on $Validation_t$. At first glance, this accuracy rate is low, but it has to be understood in light of the low prior probabilities, which we set; they affect the intercept in the model, which determines the number of observations, which are classified as an event. Of importance, is the *lift, which* we calculate from this result. The *lift* is displayed in Figure 6.7:

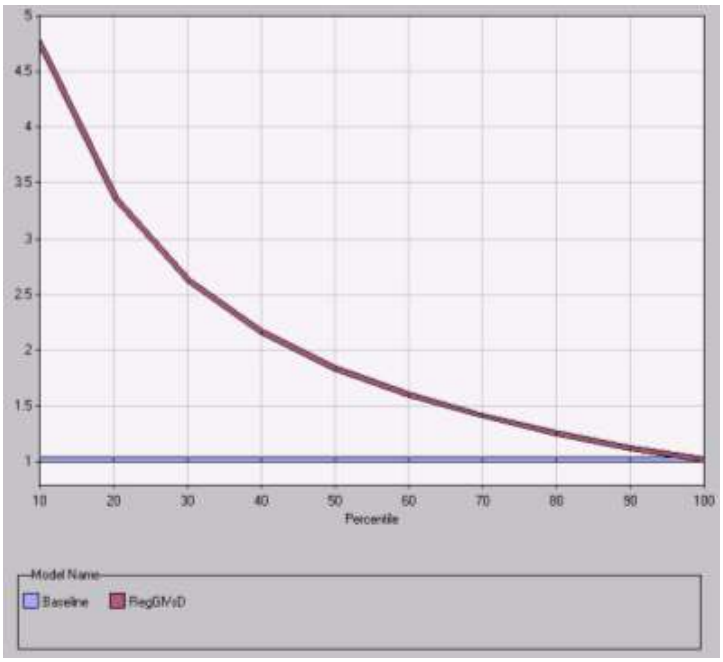


Figure 6.7: Lift value on $Validation_t$

Reading from the X-axis of Figure 6.7, we see that the lift is 4.75, in the top 10% (Y-axis in Figure 6.7) of a sorted $Validation_t$ data set. This lift means that in the top 10% of the validation set, we get almost a five times improvement over a pure chance approach. These results were a bit disappointing, and we tried to improve it including two-way

interactions in the modeling. The inclusion of these interactions was not successful, due to hardware resource constraints.

6.5.2.1.2.2 Confidence

We find our confidence measure about rejecting the null hypothesis, in the model's Type III Analysis of Effects. These we present in Table 6.6.

The *Pr>Chi-Square* column gives the confidence in the null Hypothesis that all the betas in the model = '0'. The way to convert these numbers into our confidence about the repeatability of the results is to deduct them from the number 1. Therefore, in the case of Effect A, we are 1 - 0.0001 = 0.9999, or 99.99% confident in rejecting the null hypothesis, or that the results are repeatable.

Type III Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > Chi-Square
A (cat.)	27	510	0.00010
B (cat)	11	365	0.00010
C	26	246	0.00010
SegMas	9	126	0.00010
D	5	21	0.00006
E	6	121	0.00010
F	6	195	0.00010
G	7	362	0.00010
H	3	19	0.00002
I	3	40	0.00010
J	4	128	0.00010
K (cat)	6	33	0.00010

Table 6.6: Type III analysis of model effects

6.5.2.1.3 Confirm technical success – classifier

6.5.2.1.3.1 Effectiveness

Note from Figure 6.8's X-axis, that the cumulative response in the top 20% of the database is 70% (Figure 6.8's Y-axis). Relating this back to lift, it means that in the top

20% of the validation data set, the cumulative lift is seven. It means that considered over 20% of the validation data, our *lift* value is at the upper boundary of the typical commercial lift range of 5 – 7 times. This means that somewhere in the range between the top 10% and top 20% of observations, we do meet the minimum lift of five times.

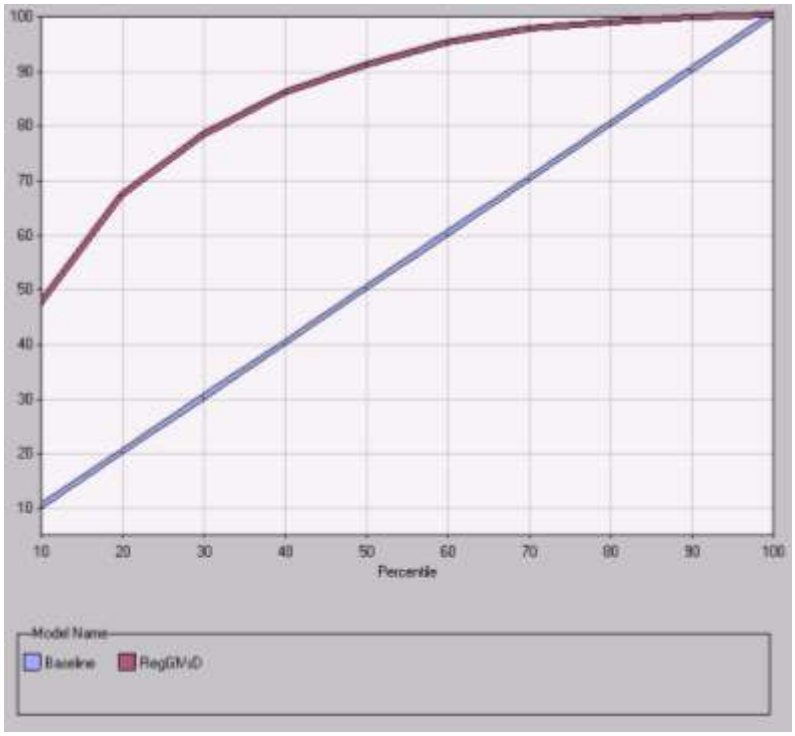


Figure 6.8: Cumulative lift value on Validation$_t$

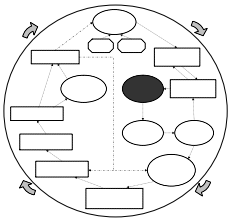
Based on this, we accept the lift of 4.75 in the first 10% of the database as providing enough confidence for inference. Since we introduced some leniency about *lift* when exploring the solution space, we have no problem accepting this *lift* as sufficient for exploration of the solution space. The decision to accept the *lift* is made easier by the fact that we have some options in mind about improving the lift of this model during *Model, evaluate, choose best model* in the *Realise* phase of SAM.

6.5.2.1.3.2 Confidence

We do not calculate the individual confidence scores for each effect in the model, but at a glance, we see that the 99% applies to all of the effects in the model. This surpasses the 95% lower confidence limit we set before, and we accept the *repeatability* of the model with confidence, both for inference about root cause, and about support for solution design.

6.5.2.2 Knowledge development looping for SPC 2 – root cause

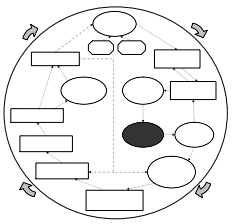
6.5.2.2.1 Develop circumstantial knowledge



The information we discovered is the Wald χ^2 of the effects in the model. This is the total estimate score for an effect divided by its standardised variance (Woods 2003a, p.9). This is a z score for the effects (Hastie, Tibshirani et al. 2001, p.263). It expresses the relative importance of that effect in the model, in distinguishing between a voluntary churner and a non-churner.

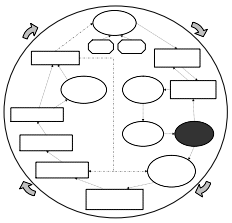
Effect A has the single highest Wald score, making it the most important effect in the model, is Effect A (cat.). Effect A is the categorical feature for Handset Type. Its values are the labels of the different Handset Types, and we see from the degrees of freedom, that there are 28 different Handset Types. We also have the effect scores of each of the individual Handset Types in the model, but we interpret and analyse them later in the third SPC 3 goal about profiling root cause.

6.5.2.2.2 Strategic analysis



Effect A is 40% more important than its nearest rival Effect B, calculated as 510 / 365 = 1.40. Effect A also has 70% of the impact of its nearest two rivals Effects B and G, calculated as 510 / (365 + 362) = 0.70. For having to make a business decisions about addressing a single root cause in the campaign offers later, we accept Handset Type as root cause.

6.5.2.2.3 Strategic choice



Recognising Handset Type as the single most important root cause of voluntary churn, has supported the hypothesis about problem phone types causing voluntary churn. We have met the SPC 2 SPC goal about identifying the main root cause of voluntary churn, within a feature which can be used in campaign offer design.

Having also attained the SPC 1 SPC goal in *Identify, assemble, prepare useful* data, we have completed our learning about the problem space. We can now pursue the remaining SPC 3 goals.

6.5.2.3 Execute SPC 3 strategy two, three, and four – targeting

6.5.2.3.1 Score the consumer database

The LogReg model generates code, which contains the model that distinguishes between the voluntary consumer churner and the non-churner. We applied that score code to $CDR_{month\ t+1}$, which was the most recent consumer database available at the time. The score code applies all the transforms to the consumer database, which we did to the data in its preparation, before scoring it. The resultant scored table we call SCR_{t+1}. An interval feature called *P_T1* is added by the score code.

6.5.2.3.2 Profile technical results - scoring

LogReg was 33% accurate on $CDR_{month\ t+1}$, which was 18% less than its accuracy on $Validation_t$. We tested the LogReg on four more $CDR_{month\ n}$ data sets, and its accuracy was between 33% and 43%. This equates to a lift between 3.3 – 4.3 times in the first 10% of $CDR_{month\ n}$.

From the researcher's experience, one looses a number of percentage points in accuracy between validation and score data sets, but this loss was excessive, and falls below our threshold for lift of five. Since we are in the solution exploring space, we can be lenient about this lift, but only if we think we can improve on it later. We would prefer however, to have a temporary solution for maintaining our confidence, even now during SPC 3, in order to retain the momentum with the project.

We explore the possibility of improving the lift first. We consider that our data has an average square error on the $Validation_t$ data set of 17.5%. This parameter estimates the error or variance in the data (SAS Institute online b, pp.42, 43), and therefore their noise. This *uncontrollable variation* is variance in the outcome, also referred to as the variance of the target around its mean (Hastie, Tibshirani et al. 2001, p.197). There is nothing we can do about this, as it is a given in the data.

This error figure is double the rate of banking data the candidate has worked with, and about 65% more than that which the candidate has encountered in supply chain data. The high noise negatively influences the accuracy of the model, in that the model also learns the noise. The event we are trying to predict, has an occurrence of less than 2% in the data, therefore the signal to noise ratio in our data is about 1:8, being the 1.5% churn

rate divided by the 17.5% ASE above. Any model, which has learned a low signal ratio, struggles for accuracy in correctly predicting such a rare event.

There are a number of techniques for possibly improving the performance of the model, and therefore the confidence in the SPC 3 results:

1. add interactions to increase its complexity. This reduces the bias of the model, but can be offset by the additional variance it adds (Hastie, Tibshirani et al. 2001, pp.197ff.). This is a resource intensive option, which we may have to try on other infrastructure;
2. add additional features in the data, which expose any non-linearity, which the Logistic Regression may not be detecting. This is a time intensive option;
3. increase the proportion of non-churners in the training and validation data from 50% to 70%, giving a 30/70 proportion of churners / non-churners, instead of the 50/50 proportion we used. This greater number of non-churn observations would capture more of the variance about the non-event, which could result in an improvement in model performance. Given the noise in the data, we don't think this option will solve the problem;
4. use the existing 50/50 sample proportion based on the target label in a bootstrap technique. Build a number of models on each sample, and then aggregate their results into one score code (SAS Institute online b, pp.43ff.). We believe that this strategy would work, since it captures substantially more of the variance of the non-event, reducing the sampling error (Hastie, Tibshirani et al. 2001, pp.197ff.). However, this option would be very resource intensive;
5. partition the data into training, validation, and test sets, and then choosing the model, which fared best on test. The chosen model would be the one with the lowest bias (SAS Institute online b, p.44). We believe however, that our problem lies in not having captured enough variance about the non-event, and this tactic would not overcome that problem. This approach would further reduce the number of observations in each data set, making the approach self-defeating;
6. another approach - under the existing resource constraints - would be to:
    6.1. use the false but equal 50/50 prior probabilities in the data now, and adjust the intercept at the end for the true prior probabilities;

6.2. change the model assessment from profit maximisation on validation, to accuracy on validation training error optimisation criterion;

6.3. standardise the interval features; and

6.4. apply more aggressive outlier replacing strategies to the interval features; and

7. use a combination of some of the strategies above.

We feel confident that one, or a combination of the above strategies, will succeed in improving lift in the *Realise* phase of SAM.

We now turn our attention to a temporary solution for the low lift on the $CDR_{month\ n\ datasets}$. These lift results only affect our confidence in the model's *effectiveness*, in that portion of the results in $\mathbf{CDR_{month\ t+1\ or}}$ $CDR_{month\ n}$, where the model was inaccurate (Ridgeway 2003). We explain this using the following two concepts – resolution and calibration:

1. *resolution* is the ability to separate the classes at the decision boundary in the data (Ridgeway 2003, pp.164ff.). This is the traditional *sensitivity* factor upon which medical confidence is based, which is visualised as an ROC chart. Basing our confidence on this factor alone, overlooks the following;

2. *calibration* is the ability to assign meaningful p-values to the outcomes, or ranking observations by their chance of being a churner or non-churner in our case. Logistic Regression is particularly effective at producing good estimates of the p-value for *linearly* determined events (Ridgeway 2003, pp.163, 164, 169). When there is some *non-linearity* in the data – as we suspect is the case with ours – the logistic function is less effective at calibration in the area of the decision boundary, which may explain the low accuracy. In areas away from the decision boundary however, the effectiveness at calibration resumes (Ridgeway 2003, Figure 7.2).

What this means for us, is that if we limit our exploration of the solution space SPC 3 to those observations, which were correctly classified in SCR_{t+1}, we can be 100% confident of those results about the solution space. We can extend that *lift* confidence in effectiveness into a bigger portion of the database in *Model, evaluate, choose best model(s)*.

This means that we need to modify the fourth strategy of SPC 3 for targeting, not to investigate the top 1.5% of the sorted database. We now only take the top *n* observations from the database, which we correctly classified with this model. Practically, we round that accuracy over all the $CDR_{month\ n}$ data sets to 35%, and only take about the top 5000 observations from the scored production SCR_n data sets for further analysis.

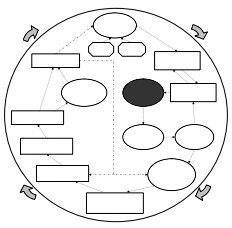
6.5.2.3.3 Confirm technical success - scoring

Given the above possibilities for improving the accuracy of LogReg, and the temporary strategy for maintaining our confidence, we confirm the technical success of the LogReg for purposes of exploring the SPC 3. Before we can proceed to the KM activities for the information discovered here, we need to execute the remaining strategies for SPC 3 targeting.

6.5.2.3.4 Sort the scored table and identify the target

We executed the third and modified fourth strategies for targeting, with a descended sorting of SCR_{t+1}, by the feature *P_T1*. After the sorting, the observations with the highest p-score are at the top of SCR_{t+1}. We then selected for retention segmentation, the number of observations at the top of table SCR_{t+1} , which equals 35% of the actual number of churners. Because we knew the actual number of churners in SCR_{t+1}, that chosen number of observations was one thousand two hundred and sixty. We called the selected table $RetSeg_{month\ t+1}$. We already discussed the confidence level associated with this result above, and can now proceed to the KM loop, for determining if we supported the relevant SPC 3 SPC goals.

6.5.2.4 Knowledge development looping for SPC 3 – targeting

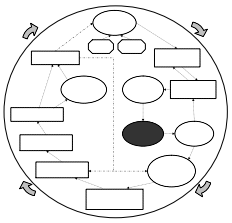


6.5.2.4.1 Develop circumstantial knowledge

The information we discovered when scoring the database, was the *p_value* given to each observation within the feature *P_T1*. These p_values range on a scale of almost ‘0’ to almost ‘1’. This value is the conditional probability of each consumer becoming a potential churner within the time window of the next 0 – 92 days.

The information we discovered when we selected the table $\text{RetSeg}_{\text{month } t+1}$, is those observations which we have confidence about the effectiveness of the LogReg model.

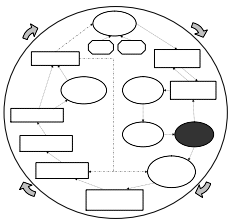
6.5.2.4.2 Strategic analysis



The higher the *p_value* of a consumer, the higher their chance of becoming a voluntary churner in the next three months. The observations in $\text{RetSeg}_{\text{month } t+1}$ in this exploration of the solution space, are those consumers, which Telco ABC will target with their retention management campaigns in the *Realise* phase of SAM.

We now know – at exploratory levels of confidence at least – the chance of each consumer in the consumer database becoming a voluntarily churner in the next three months. We also know that it is possible in future, to identify which consumers fall outside the *risk tolerance* figure, and whom we should target with retention management campaigns. We can also rank consumers *within* that group for the prioritisation of retention campaign resources.

6.5.2.4.3 Strategic choice



We have successfully developed a prototype data mining solution, which attains the two SPC 3 data mining goals for targeting. We have therefore attained the SPC goal of successfully investigating ways of improving targeting from Telco ABC's data, which considers the new business domain knowledge we introduced there. We can proceed on to executing the segmentation strategies of SPC 3.

6.5.2.5 Execute SPC 3 strategies – segmenting

6.5.2.5.1 Execute first SPC3 strategy - segmenting

We chose the segmentation base as follows:

- SUMC_6Z3 – this feature contains the standardised value of the sum of a customer's voice calls over the last three months. This feature represents the behavior-based demand function, which our segmentation hypothesis demands. In the discourse and tables that follow, we therefore call the column with SUMC_6Z3's segmentation values, *Demand function*. We limited the demand function to one – excluding demand functions about for instance SMS and WAP - to keep our later demonstration about designing a campaign offer, tractable;

- REV5_435 – this feature contains the values of the average three-month revenue per user – also called ARPU. As such, it represents the relative monetary attractiveness of each segment. We therefore named it *Value measure* ARPU in the following discourse and tables. It also expresses the *event impact* component of Risk Management;
- P_T1 – this feature contains the conditional probability of an observation becoming a voluntary churner or not. In Risk Management terminology, this feature is the *event likelihood* of the unwanted churn event, within the time window of 0 – 92 days. We therefore have named it *Event likelihood within time window* in the following discourse and tables;
- LOYP_RLV – this feature contains an interval value of a customer's tenancy with Telco ABC. There is evidence in all the literature, which we references about value-based segmentation, which loyal customers tend to remain loyal. This is ground for us to interpret past loyalty, as a psychographic measure of a customer's *intention* to respond positively to a well-designed campaign offer. We appropriately named it *Intention to respond positively*. Based on the evidence in the literature, this feature also enables the campaign design to pitch the duration of the new contract on offer correctly; the higher a customer's loyalty, the longer the duration of the contract can be in the retention campaign offer.

The values of these features we previously had standardised in the $CDR_{month\ t+1}$ table, using the standard deviation technique. We did that for confidentiality reasons. After standardisation, they have an overall mean of 0 in $CDR_{month\ t+1}$ and in SCR_{t+1}, and a range spread around that new mean. This mean is displayed in the last row of Table 6.7.

We did not standardise *P_T1* in $CDR_{month\ t+1}$ - as is evident from no value in the corresponding cell in Table 6.7 – because the feature did not exist in $CDR_{month\ t+1}$. *P_T1* is created later in SCR_{t+1}. We also do not standardise it in $\mathbf{RetSeg_{month\ t+1}}$, because its values already are within the standardised range of [0,1].

Segment base features' vital statistics				
	Demand function Sum of voice calls over 3 months	**Value measure ARPU**	**Event likelihood within time window P_T1**	**Intention to respond positively LOYP**
$RetSeg_{month\ t+1}$ range:	-0.7 - 14.7	-0.61 - 22.2	0.22 - 0.99	-2.0 - 9.5
$RetSeg_{month\ t+1}$ mean:	0.31	0.58	0.43	0.13
$CDR_{month\ t+1}$ mean:	0	0		0

Table 6.7: Segment base features vital statistics

We segmented $RetSeg_{month\ t+1\ using}$ k-means, experimenting with both hierarchical and non-hierarchical approaches. The non-hierarchical k-means – with a manual override asking for six segments only – provided sufficient results when we allowed enough iteration to reach convergence of the objective function.

6.5.2.5.2 Profile technical results

Technical segmentation measures ($RetSeg_{month\ t+1}$)					
Average intra-segment	0.11	**Overall R^2:**	72%	**Number of nearest segment profiles:**	4
Percentage of targeted consumers within segment:	Segment 1:	21%		**Fisher's Criterion: (Average inter-segment distinctiveness)**	1.88
	Segment 2:	13%			
	Segment 3:	6%			
	Segment 4:	5%			
	Segment 5:	1%			
	Segment 6:	55%			
Relative importance of segment base features:	LOYP_RLV:	0.40		**Convergence of optimisation function:**	Yes
	P_T1:	1.00			
	REV5_435:	0.40			
	SUMC_6Z3:	0.28			

Table 6.8: Technical segmentation measures (RetSegmonth t+1)

We base our profiling of the technical results on Table 6.8. The *Number of nearest segment profiles* is four. This means that we do not have any dominating segment in the data. The *relative importance of segment base features* is shared over all four, with no one feature dominating the structure.

The Fisher's Criterion is 1.88, which means we have almost doubled the information available about the natural partitions in the data.

The *Overall* R^2 of 72% surpasses our minimum confidence level of 70%. Further, the output of the algorithm stated that the convergence criterion had been met within the number of iterations we had specified.

We can also see that the algorithm converged on the optimisation function.

6.5.2.5.3 Confirm technical success

The clustering algorithm has met all our confidence measures. In order to determine if we have met the relevant SPC goal, we first need to execute the second and third SPC 3 strategies for segmenting.

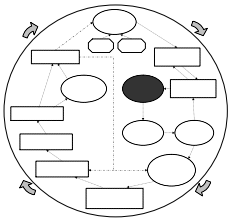
6.5.2.5.4 Execute second SPC 3 strategy - segmenting

The clustering node of SAS Enterprise Miner automatically creates the standardised quantitative summary for us. We present that as tables below in the KM loop. Since this output is produced from the results of the same algorithm as the segmenting itself, we have already established technical confidence. These tables should allow for the effective and efficient quantitative comparison between segments. We test this in the KM loop below, after which we know if we have met the data mining goal, and have supported the underlying SPC goal.

6.5.2.5.5 Execute third SPC 3 strategy – segmenting

The clustering node of SAS Enterprise Miner also visualises the quantitative comparative measures in a standardised means plot. We introduce an example of a standardised means plot in the KM loop. Since this output was produced from the results of the clustering algorithm, we accept those technical confidence measures as sufficient. We evaluate how these quantitative comparative measures, and their visualisation supports the SPC goal, in the KM loop below.

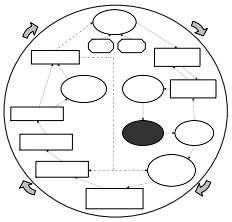
6.5.2.6 Knowledge development looping for SPC 3 – segmentation



6.5.2.6.1 Develop circumstantial knowledge – segment membership

The clustering algorithm adds a feature to the data set $\text{RetSeg}_{\text{month } t+1}$, which contains the segment membership of each observation to one of six segments, as a label. This is the discovered information.

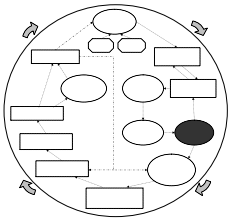
As proof of these membership labels, we refer to the percentages in the *Percentage of targeted consumers within segment* section in Table 6.8 above. There we see the 6 segment numbers, and for instance that 55% of all the observations in $\text{RetSeg}_{\text{month } t+1}$ belong to the sixth segment, while 5% belong to the fourth segment. The frequencies, on which those percentages are based, are in the second column of Table 6.9.



6.5.2.6.2 Strategic analysis – segment membership

The six segments match Telco ABC's current six-fold retention management segmentation. Having six *Retention segments* supports the existing marketing *Operating strategy,* which is based on six segments (Badgett, Connor et al. 2003, pp.15ff.). By keeping the number of segments the same as the current, makes it easier for the business to digest the novelty of the other new information.

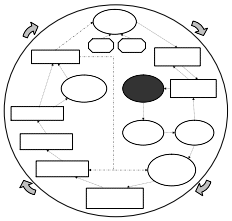
Further, these segments are based on the four features of hypothetical relevance we identified in the formulation of the first SPC 3 goal for segmenting.



6.5.2.6.3 Strategic choice – segment membership

We now have information and knowledge about the important 4-dimensional profile of each observation in $\text{RetSeg}_{\text{month } t+1}$, which we can use for improving the segmentation. We have therefore attained the first SPC 3 goal for discovering the segment membership of each observation by a multidimensional behavioral base.

6.5.2.6.4 Develop circumstantial knowledge – comparative segment profiling



In this discussion, we use membership to a behavioral segment as a basis for comparative behavioral expectation. We offered substantial literature evidence for this approach earlier from the marketing segmentation literature. We add evidence from our own research in support of this approach. Recall how we created the contextual feature SegMas, through clustering $CDR_{modplus\ t}$ on behavioral features during data preparation. We next saw in Table 6.6, that SegMas was selected as an effect in the LogReg model. That selection means that behavioral segment membership was indicative of actual past behavior, which was voluntary churn in our data. We subsequently used that model successfully to predict expected churn in other data. This is evidence that *in our data*, membership to behavioral segments can be used to base comparative behavioral expectations upon.

Table 6.9 contains the most important quantitative information for inter-segment comparison; the standardised retention segment profile values:

Standardised Retention Segment Profile Values ($RetSeg_{month\ t+1}$)				
Retention segment number	**Demand function Sum of voice calls over 3 months**	**Value measure ARPU**	**Event likelihood within time window P_T1**	**Intention to respond positively LOYP**
2	0.11	0.68	0.86	0.38
5	6.87	5.30	0.85	-0.52
1	0.08	0.44	0.54	-0.37
4	0.13	0.17	0.41	3.86
3	2.71	5.64	0.34	-0.71
6	0.07	-0.01	0.30	0.04

Table 6.9: Standardised retention segment profile values (RetSegmonth t+1)

Table 6.9 has been sorted descending by the fourth column. The first column has the segment number. The other four columns contain the standardised mean values of each base feature, within each segment. This standardisation makes possible *quantitative*

inter-segment profile comparison. We next offer an example of a quantitative inter-segment profile comparison.

The consumers in segments two and five have a similarly high chance of becoming voluntary churners over the time window; this is evident from their *P_T1* means of 0.86 and 0.85. However, notice that low *ARPU* value of segment two compared to the high value of segment 5; this means that even though these customers have the same chance of becoming churners, those in segment 5 have an almost 10 times higher *ARPU* value. The reason for the high *ARPU* value in segment 5 becomes evident, when we see that these customers make the most voice calls of all the segments. That is evident from their high mean in the second column of segment 5.

We make the example of visual assistance for quantitative profile comparison of segment 1 with the overall profile of the targeted potential churners in Figure 6.9:

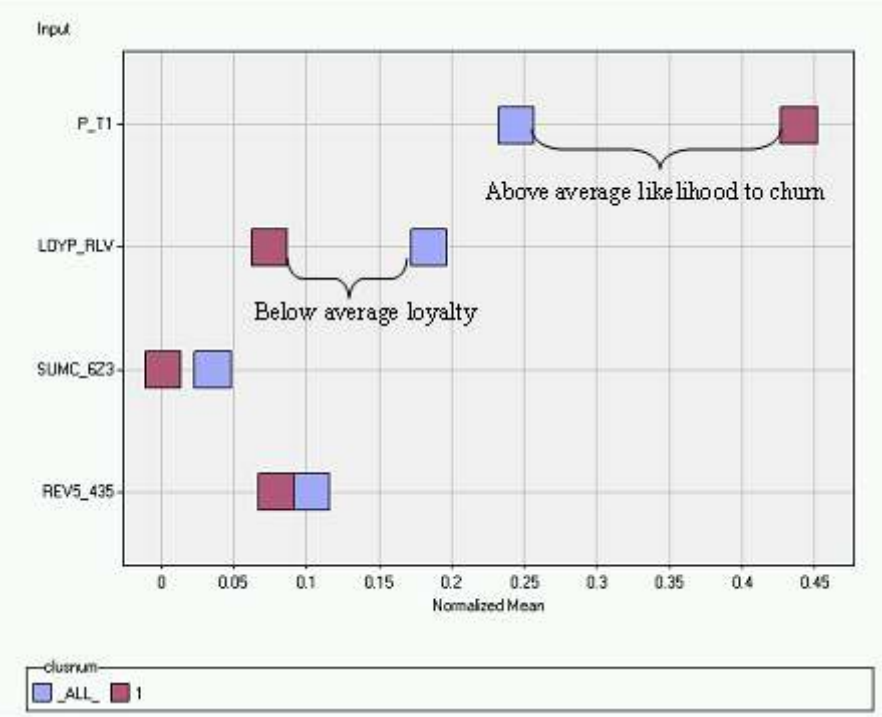


Figure 6.9: Standardised means plot segment 1

A plot like this is available for each segment. On the Y-axis, we have the four base features, and on the X-axis, the standardised mean of each feature. The dark squares are the overall standardised mean for all potential churners in $\mathsf{RetSeg}_{\text{month } t+1}$, while the lighter squares, show the standardised means for segment 1.

In the case of segment 1, we can see that the consumers in this segment:

- are almost twice as likely to become a churner during the time window, than the overall average of the targeted customers. This is evident from their standardised mean value of about 0.45, compared to the overall standardised mean for this feature of about 0.25;
- have less than half the intention to respond positively to a well-designed campaign offer than overall response. This is evident from their standardised mean of about 0.08 for the feature *LOYP*, compared to an overall means for *LOYP* of about 0.18;
- have a below average demand function, as is evident from the feature *SUMC;* and
- about average value to the company.

We can also do inter-segmentation comparison within $\text{RetSeg}_{\text{month } t+1}$ using these means plots. Figure 6.10 displays the profile of Segment 2 in $\text{RetSeg}_{\text{month } t+1}$:

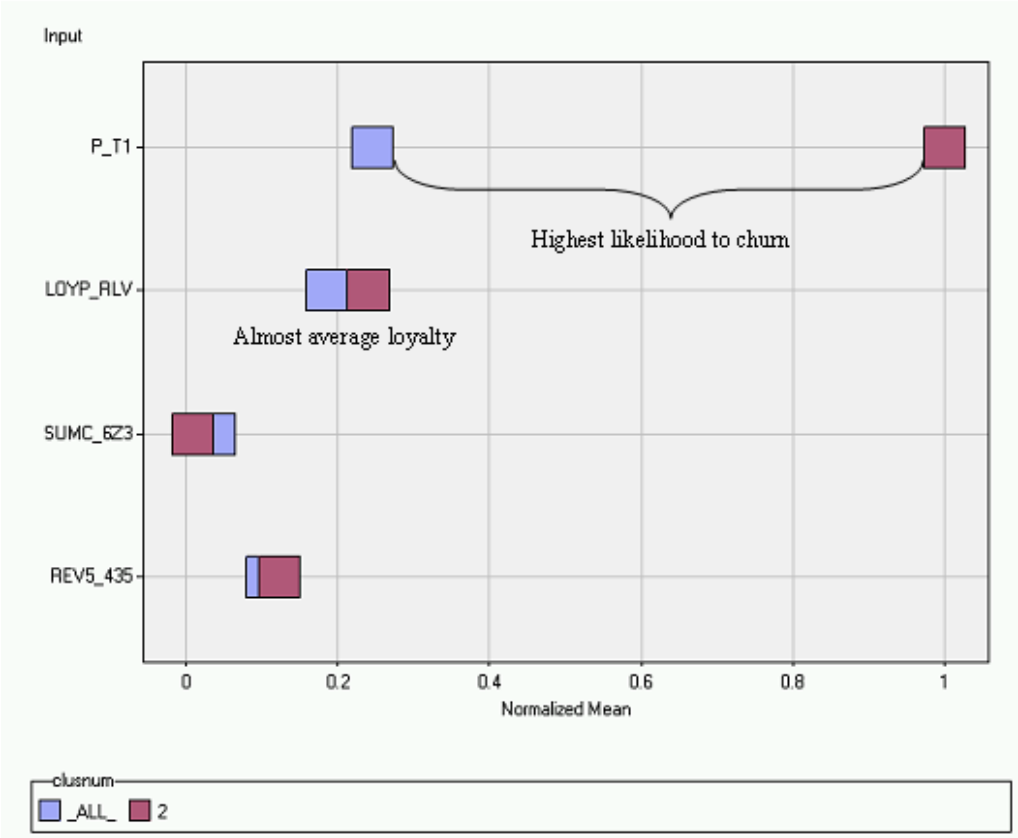


Figure 6.10: Standardised means plot segment 2

Notice that Segment 2 has almost four times the overall *Event likelihood* (*P_T1*) than Segment 1. We calculate this as segment 2's standardised mean of 1 divided by

Segment 1'a standardised mean of 0.25 = 4. Relative to Segment 1, it has twice the *Event likelihood*, which was 1 / 0.45 ~ 2. Similar comparisons can be made between the values of the other features in the two segments.

Apart from the all-important standardised means, we also use for comparative purposes the Frequency of cluster, Fisher's Criterion, the Nearest Cluster, the Root-Mean-Square Standard Deviation within each cluster, and Maximum distance from the cluster seed. We refer to Table 6.10 for this discussion, where the shaded headings contain the business meaning of all these technical measures of comparison, and the unshadowed row beneath them, the technical headings:

Overall Retention Segment Measures (RetSeg$_{month\ t+1}$)						
Retention segment number	**Number of customers in segment**	**Intra-segment homogeneity (smaller = more similar)**	**Span of intra-cluster homogeneity (smaller = better)**	**Nearest segment in profile**	**Distance to nearest segment (bigger=better)**	**Inter-segment distinctiveness (bigger = more distinct)**
CLUSTER	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	Fisher's Criterion
1	257	0.08	0.57	6	0.32	2.22
2	163	0.12	0.62	1	0.41	1.97
3	76	0.13	0.75	6	0.31	1.62
4	60	0.12	0.61	6	0.36	1.95
5	13	0.14	0.56	2	0.49	1.87
6	684	0.06	0.27	3	0.31	1.62
Total:	1253					

Table 6.10: Overall retention segment measures (RetSegmonth t+1)

The technical measure *Root-Mean-Square Standard Deviation* is the within-cluster variance *over all the base features*. This quantifies the diversity within each cluster. The smaller this number, the more similar the members are within that segment *over all the base features* when compared to members of other segments. The greater the similarity, the more homogenous we can expect their response to a campaign offer. For

comparative purposes, we see that segments 6 and 8 to have the most homogenous profiles within.

Considering that the cluster shapes are not spherical, *Maximum Distance from Cluster Seed* is the technical measure of how far the furthest observation is from the mean within that cluster. We therefore have named it *Span of intra-cluster* homogeneity. It is a measure of the most different response in that segment, compared to the overall response in that segment. When we consider these last two measures together, segment 6 should be the cluster with the most homogenous response overall, because it has the smallest number for both measures. Taken on its own, this measure helps us to understand whom those customers are who will never respond positively, even to a well-designed campaign offer.

Nearest cluster measures the *Nearest segment in profile* to a retention segment. This helps the organisation merge similar segments for retention management, if the organisation becomes hard-pressed for retention management resources. This measure is also of value for assuring distinctiveness between the campaign offers to different segments. This statement becomes clearer when combined with *Fisher's Criterion*.

Fisher's Criterion measures the *Inter-segment distinctiveness*. Combining the interpretation of the *Fisher's Criterion* with the *Nearest segment in profile*, helps with the segment merging decision. For instance, if the organisation were considering merging segment six with either segment 1 or 3 or 4, one would advise merging with segment 3. The reason is that of all three merging options, the *Fisher's Criterion* is the smallest between segments 6 and 3 – they are the least distinct of the three segments. After such a merging the distinctiveness of the campaign offers between the remaining segment 1 and the merged segments, should be more distinct than the campaign offer between the remaining segment 4 and the merged segments. That is evident from the bigger *Fisher's Criterion* between segments 1 and 6, than between segments 4 and 6.

The *Frequency of Cluster* measure the *Number of customers in* each *segment*. This helps with calculating the required time resource between retention campaigns, and with allocating time and money resources accordingly, or with prioritising limited resources. All else equal, segment 6 should consume the most retention management resources.

There is a relationship between the *Frequency of Cluster, RMS-SD,* and *Maximum Distance from Cluster Seed* measures of the segments. This relationship is a function of

the objective function of the clustering algorithm. We visualise this relationship in Figure 6.11:

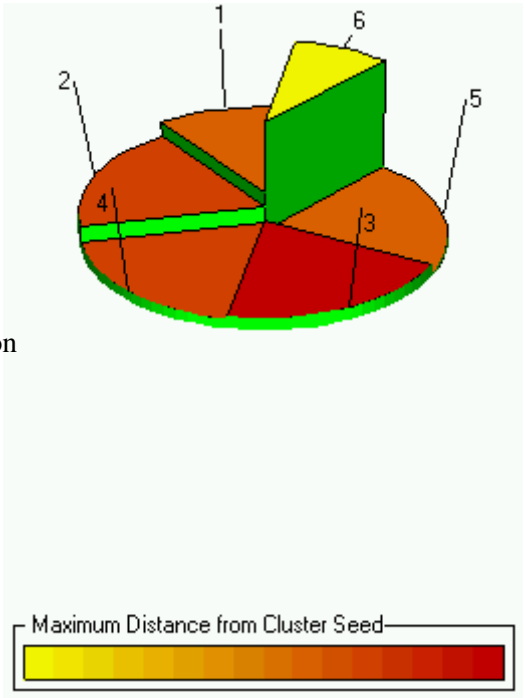
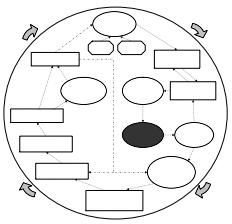


Figure 6.11: Relationship between 3 measures

The similar size of the slices in Figure 6.11 visualise the similar internal homogeneity of the clusters, with the exception of Segment 6. Segment 6 is even more homogenous internally than the others. In commercial data, there always is one big cluster of high homogeneity (Woods 2003b). Segment 6 contains about half of the potential churners, and taken at face value should consume about half of the attentive time of the Retention Manager.

These comparisons assist with the prioritising of campaign resources, toward those groups who similarly have a high *Event likelihood*, and possible also a similarly high *Value measure*. In the *Intention to respond positively* dimension, they help to set the organisations relative expectancy for segments, about positive responses to a campaign offer. It also helps to profile the length of the commitment in the campaign design.

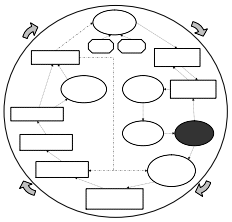
6.5.2.6.5 Strategic analysis – comparative segment profiling



The discussion and examples in the previous section demonstrated how that this ability to do quantitative and visual *intra- and inter-segment comparisons* among retention management segments, could improve campaign offer design. This improvement is both

in retention management resource prioritising, and in tailoring campaign designs to better match the segmented profiles of commercial interest about customers (Woods 2003a, p.11).

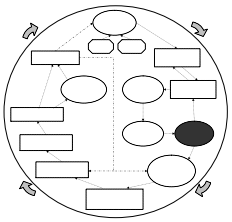
6.5.2.6.6 Strategic choice – comparative segment profiling



We now have information and knowledge for comparing the 4-dimensional segment profiles discovered in $RetSeg_{month\ t+1}$, and the tools with which effectively and efficiently to make the comparative profiles. We have therefore attained the second SPC 3 goal of effective and efficient intra- and inter-segment profiling.

We are now in a position to pursue the

6.5.2.6.7 Strategic choice - segmentation results



The strategic choice is to use the results of the segmentation for knowledge creation in SAM's business activity sets *Develop circumstantial knowledge* and *Strategic analysis.* We also choose to use the segments discovered, in the following root cause profiling.

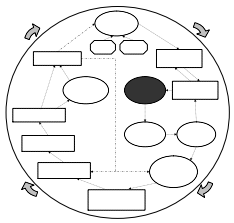
6.5.2.7 Execute SPC 3 strategies – root cause profiling

6.5.2.7.1 Execute strategies

We executed the first strategy using the segment profile statistics table, which the SAS Enterprise Miner Clustering node builds for each segment after it has segmented the data. It is effectively a cross-tab of the label values of all the categorical features by segment.

Drawing that out into MS Excel format, we then added the individual label effect scores from the LogReg's output. The calculations of percentages – and their visualization – we did in Excel. Since all the output was produced from the results of the LogReg and k-means algorithms – which we evaluated before for confidence - we have already established technical confidence in these results.

6.5.2.8 Knowledge development looping for SPC 3 – root cause profiling



6.5.2.8.1 Develop circumstantial knowledge

We present the discovered information, and the knowledge we developed about segmented root cause, in Table 6.11:

Feature A labels [A]	Label's occurrence within Segment 1 [B]	Label's contribution to total positive T-score within Segment 1 [C]	Label's relative impact within Segment 1 [D]	Label's impact portion within Segment 1 [E]	Label T-score [H]
#	19.84%	15.43%	3.06	36.76%	4.80
AA					
AB	0.39%	5.13%	0.02	0.24%	1.60
BA	2.33%	6.32%	0.15	1.77%	1.97
BB	2.33%	6.05%	0.14	1.69%	1.88
BC	3.89%	6.32%	0.25	2.95%	1.97
CA					-0.14
DB					-1.15
DC	21.40%	7.15%	1.53	18.36%	2.22
DD	1.56%	6.32%	0.10	1.18%	1.97
E	4.67%	7.15%	0.33	4.01%	2.22
GA					-0.50
GB	0.39%	1.35%	0.01	0.06%	0.42
H					
JA	0.39%	4.09%	0.02	0.19%	1.27
JB	6.61%	6.87%	0.45	5.45%	2.14
JC	4.67%	2.20%	0.10	1.23%	0.68
JD	5.84%	1.65%	0.10	1.16%	0.51
JF	17.10%	8.52%	1.46	17.49%	2.65
JH	6.61%	6.87%	0.45	5.45%	2.14
JJ	1.95%	8.59%	0.17	2.01%	2.67
Totals:	99.97%	100.00%	8.33	100.00%	

[I]Total positive T-score in segment 1:	31.10
[J] Total positive T-score in segment 2:	35.40
[K]Total positive T-score in segment 3:	33.25

Table 6.11: Root cause profile segment 1

The columns are labeled A – H from left to right to facilitate this discussion. Three more cells are labeled I – K in the bottom right corner of the table.

We discuss the *discovered information* first. Column A gives the disguised labels of Effect A. Each of these Handset Types makes an individual contribution toward the problem. This individual contribution is the label's T-score for each label, which was calculated by LogReg when it built the model. We display the T-score for each label, in column H. The T-score value column H of the first label # in column A, is 4.80. The T-score is the label's effect in the model, divided by its estimate error.

For profiling business 'cause', we distinguish between labels with positive T-scores, and those with negative T-scores. The labels with positive T-scores contribute toward the churn problem, and labels with negative T-scores, reduce the churn problem – so to speak. Their frequencies are omitted, and we have shaded three lines that contain negative T-scores.

Not all these labels are present in a segment in $\mathsf{RetSeg}_{\text{month } t+1}$. So for instance, column B of segment 1 is empty for the label AA, because the label AA is not present in segment 1. In contrast, the label = # is present, making up 19.84% of the present labels in segment 1.

We now introduce the *knowledge* we have developed for comparative profiling. Not all the labels occur in a segment, and the sums of the segments' T-scores therefore are different. For example, the sum of positive T-scores in segment 1, is 31.10 (cell I), for segment 2 it is 35.40 (cell J), and for segment 3 it is 33.25 (cell K).

For intra-segment profiling, we now apportion each label's T-score to the sum of positive T-scores of that segment. The apportionment of label #'s T-score in segment 1, is given in column C as 15.43%. We calculated this as 4.80 (column H) divided by 31.10 (cell I). Next, a label's value in column C has to be adjusted for its value in column B to get its *Relative impact within segment* in column D. Using the values for label = # in this way, we multiply the 19.84% with 15.43%, again multiplying that result with 100. The resulting value is 3.06 in column D.

The scale of column D is not easy to understand, so we convert it to a percentage in column E. That % value in column E now expresses the within-segment percentage, which allows for intra-segment profiling of root cause. The percentage value for label # - 36.7% - is the highest of all the figures in column E. This means label # has the single most impact toward root cause within segment 1.

Further, resources spent on those customers who have a label = #, give a relative return factor of 3.06, or eliminate 37% of the risk associated with root cause effect in segment 1. This is twice the relative return of 1.53 for label = DC, and 10 times the return of the 0.33 for label = E. We visualize this comparison Figure 6.12:

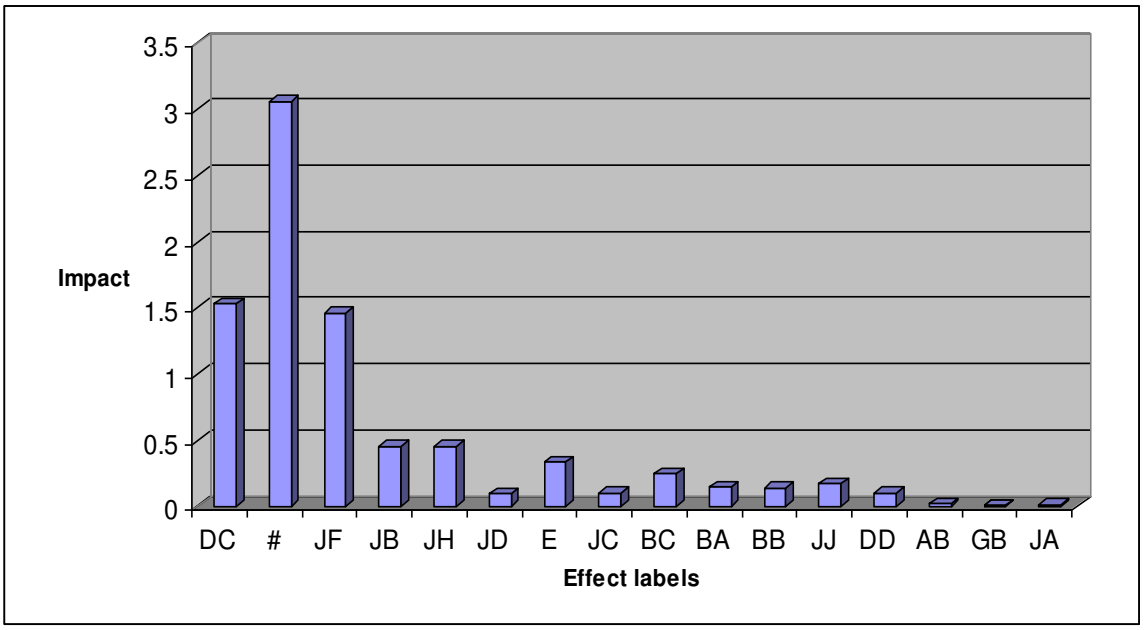
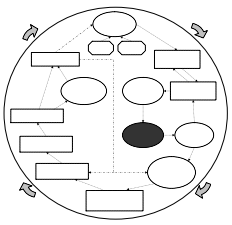


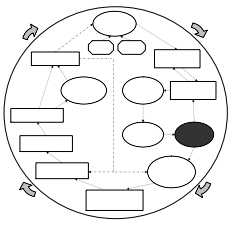
Figure 6.12: Handset Type's relative impact within Segment 1

6.5.2.8.2 Strategic analysis



Since Telco ABC has limited resources for addressing root cause in the campaigns, they now have knowledge with which to prioritise the allocation of those resources on an intra-segment basis. For instance, allocating resources toward label # in segment 1, addresses more than a third of the root cause *within that segment*. We know that we can develop this tool for inter-segment comparison of root cause too, and do that in SPC 4.

6.5.2.8.3 Strategic choice



From the above discussion, it becomes evident that we have successfully executed the two data mining strategies for root cause profiling. The knowledge we developed above fully supports the SPC goal for root cause profiling.

6.6 Chapter summary

We summarise this chapter in two dimensions; what we attained for Telco ABC with the project up to this point, and what we proved about SAM up to this point.

In the *Business problem* entity of the *Prepare* phase of the project, we helped Telco ABC define the *status quo* in a way which helped them better understand the limiting effect this was having on the outcome of their project. We did this by airing the feelings, pre-conceptions, and expectations of the key players in the project, documenting those, analysing their existing paradigm, and facilitating mutual understanding about where each person was coming from and where the organisation was at as an entity.

We then successfully defined an agreed-to mission for the project in *Develop project mission*. Here we used the SAM diagnostic technique for overlaying the status quo with new subject matter expertise, and formulating hypotheses in an expert collaborative way. The hypotheses covered all five dimensions of the project from problem understanding to solution support. We then developed the hypotheses into agreed-to business deliverables or SPC goals, each containing the four components of *what*, *how*, *where* and *when*.

The business deliverables constituted substantial competitive breakthrough for Telco ABC, through becoming proactive in targeting customers for retention management, and in focusing campaign design at key retention segment characteristics. We used SAM's mapping technique for defining the supporting high-level project strategies for the business deliverables. We also got consensus within the organisation about how to calculate the churn rate, and accurately quantified the true extent of the problem in dollar terms for the first time.

In *Identify, prepare, assemble useful data* we subsequently prepared their data by determining the relevant features, assembling them, and cleaning them into a desirable format for analytics. We achieved this using SAM's data preparation process and activities. We successfully demonstrated to key stakeholders, the reduction of project risk brought about by the data preparation activity. We also demonstrated how that this task set progressively supports the project's business deliverables.

In the *Analyse* phase of the project we developed a data mining mission for Telco ABC, which supported all five dimensions within the business deliverables. We achieved this

by again applying the mapping technique in SAM, to formulate data mining objectives, and then mapping those to data mining strategies. We then applied the strategies to the data, discovering information in support of each of the business deliverables. The discovered information was:

- the customers in the data base most at risk of becoming a voluntary churner over the 90-day time window;
- the membership of each of those customers to a value and behavioral segment;
- the profile of those segments by the key churn driver, handset type.

We proved the technical utility of the data *materiel*. We also evaluated the technical measures of *replicatibility* and *effectiveness* about this information, and found that – although they could be improved upon – they were certain enough to pursue knowledge development. We pursued knowledge development through an initial execution of the knowledge development loop in SAM - *Develop circumstantial knowledge, Strategic analysis*, and *Strategic choice*. We demonstrated that it was possible to formulate a high-level campaign approach for one of the segments, establishing the *relevance* of the discovered information to the competitive business deliverables, to be executed under the organisation's circumstances. This enabled us to formulate our confidence in committing the organisation to the *Define* and *Realise* phases of the project.

We established the utility of SAM in a number of dimensions. First, helping an innovating organisation overcome the limiting pre-project environment of feelings, preconceptions and uncertain expectations, establishing the potential for competitive breakthrough from the project. We accomplished this in *Business problem.*

In *Potential solutions*, we established the utility of SAM in injecting new subject matter expertise into the stale *status quo*. Further, we established SAM's utility in *Develop project mission* for creating a structured, agreed to project mission, with business deliverables that express competitive potential for the organisation. Further, SAM enabled us to develop the project mission into all four required mission components for each of the five dimensions of the business deliverables. SAM also enabled us to formulate high-level goals for the analytics technology, which support the business deliverables.

In *Identify, assemble, prepare useful data* we demonstrated the success of a business-centric approach to data preparation, the functionality of a process we designed for that, and of distinguishing between iteration and repetition. We proved utility in justifying for the organisation the resource utilization during this phase of the project.

We proved the success of *Data mining discovery* in mapping out a practical data mining plan, supportive of all dimensions of the business mission, and discovered information which had relevance to the business deliverables. This was done in a way, which gave us confidence in the technical quality of the information. We proved the technique for establishing the relevance of the discovered information.

We further proved the usefulness of the knowledge development activities in SAM for:

- developing the unstructured and unchallenged pre-project preconceptions and expectations into hypotheses about the problem and its solution;
- developing the business deliverables from those hypotheses;
- redeveloping the hypotheses underlying the business deliverables in the project mission in light of the newly discovered information;
- testing and adapting them for the organisation's particular circumstances; and
- choosing the best hypotheses under the circumstances; and

We demonstrated the importance of breaking analytics project entities out into separate business activities and analytics tasks, and the practicality of their sequence in SAM up to *Strategic choice*. We also made progress in supporting the logic of the project phases in SAM. We also demonstrated the benefits of nuancing the confidence measures in information, in order to retain commitment of resources for further progression with the project. We also proved that SAM allows for the parsimonious use of technology, through attaining some of the business deliverables in the project mission formulation stage with simple queries.

We have proven our success in reframing CRISP-DM for overcoming the limiting effect of soft issues. This reframing included expert collaboration, and introducing new subject matter expertise, for shifting the boundaries of understanding, agreement, and hypothesis formulation. We have demonstrated our success in reframing CRISP-DM for diagnostic technique for formulating competitive business deliverables.

In two instances, we established the success of our reframing of mapping technique for developing a supportive analytics plan – once in *Develop project mission* and once in *Data mining discovery.* We also demonstrated the success of our reframing of CRISP-DM about introducing new subject matter expertise into the stale *status quo*.

We proved that our added knowledge management activities are useful in defining competitive business deliverables from the unstructured and unchallenged pre-project environment, in incorporating new subject matter and formulating breakthrough hypotheses, in updating those hypotheses in light of discovered information, and in developing and test those hypotheses for supporting the executibility under the organisation's limiting circumstances. This knowledge management utility was a major reframing of CRISP-DM.

In the following chapter, we continue our application of SAM to the Telco problem. There we start with the Define phase of SAM, developing the executibility of the business solution.

7 Chapter 7 – Apply SAM for solution development

In the previous chapter, we demonstrated the utility of SAM in discovering information, and developing knowledge from the information, which supports the business deliverables, and is executable *under the organisation's circumstances*. In this chapter, we intend to demonstrate first SAM's utility in developing the executibility of that solution. Second, we want to prove SAM's utility for formulating an SPC project which supports the deployment of that business solution. Third, we demonstrate the utility of SAM's *Monitor and control* in assuring the relevance of the problem understanding, and of the data mining and business solutions, over time lapsed.

We proceed by pursuing the remaining SPC goals we formulated previously in *Formulate project mission*. The SPC 4 and 5 goals remain:

- the SPC 4 goal was to *develop a new retention management solution, which will be objectively optimal, novel and executable, given Telco ABC's circumstances*;
- the SPC 5 goal was *to use data mining to support the deployment of the new retention solution on an operational basis*.

The strategies for supporting the SPC 4 goal were:

3. developing the *executibility* of the business solution through:
   - 3.1. using any newly developed understanding about the extent and nature of the retention management problem (SPC 1 and 2);
   - 3.2. using any newly developed possibilities about solving the retention management problem (SPC 3);
   - 3.3. using SAM or a business solution design tool (e.g. SPM), which factors in the organisation's commercial and operational circumstances; and then
4. formulating integrated new retention management objectives and strategies about targeting, segmenting, and profiling, which *express the executibility*.

The strategies for supporting the SPC 5 goal were:

4. operationalising the models of SPC 3 on a required monthly basis; and
5. deploying the discovered information into the marketing business function platforms in a timely manner each time; and

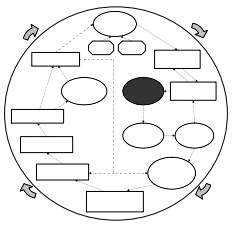
6. monitor and control the effectiveness of the solution and the analytics over time.

We will attain the SPC 4 goal first through iterating through the three business activity sets *Develop circumstantial knowledge, Strategic analysis,* and *Strategic choice.* When we have made a strategic choice, we proceed to *Define new business objectives and strategies* and develop the objectives and strategies which constitute the executable solution. This again takes us through the *Analyse* and *Choose* phases of SAM – and for the first time - across into the *Define* phase of SAM. From a SAM's KM perspective, we will create and legitimise the knowledge, which constitutes the competitive, executable business solution. Later, we execute the new business solution in *Execute knowledge.*

We pursue the SPC 5 SPC goal through defining the data mining technology solution, which will support the execution of that business solution, and by executing the data mining solution. This will take us through the four data mining task sets starting with *Develop data mining plan* through to *Deploy outputs into business*. This will take us through the KM phase *Legitimise*, and into the *Share* phase. We complete the SPC 5 goal by developing the monitor and control plan in *Monitor and control*.

We comment that in practice, this business knowledge development will be done through expert collaboration. There, the role of the data miner would be assuring the optimal use of the discovered information, and the subject matter expert would develop the knowledge into an executable solution. In the case of this project however, we did not have access to that subject matter expert. The candidate, drawing on his corporate experience and business knowledge, therefore did the business knowledge and executibility development in this thesis. Since the candidate is not a subject matter expert, he has restrained the knowledge and executibility development in the thesis, to what he estimates is sufficient for proving the utility of SAM as a SPC-supporting project methodology.

7.1 Knowledge development loop (Develop circumstantial knowledge, Strategic analysis, and Strategic choice)



7.1.1 Develop circumstantial knowledge

7.1.1.1 Execute SPC 4 strategy one

We execute this strategy through combining the knowledge we had developed during the execution of SPC 1, 2 and 3, in the embedded SPM business activities in SAM.

We saw in Chapter 2, how that executibility of a retention management solution, resides in the TSP principle – Targeting, Segmenting and Positioning. The discourse will follow this sequence.

7.1.1.1.1 Targeting

The executibility of the *Targeting* is knowing the identity of the most at-risk consumers, who fall outside the risk tolerance range. We will establish the *Targeting* executibility for an operational business solution, every time when we score the consumer data base with the LogReg model, and select the top 1.5% of observations from the scored table.

7.1.1.1.2 Segmenting

The executibility of the *Segmenting* is knowing the comparative profiles of the segments we have created from the targeted consumers.

The executibility of the *Profiling* is *knowing how these profiles compare* by their *Demand function*, *Value*, *Intention to respond positively, Event likelihood,* and root cause.

We begin by developing the comparative profiles of the targeted customers in the six segments. This will be required for designing two aspects of the campaign offer later.

7.1.1.1.2.1 Base segment profiles

We call this the base segment profiles, because it concerns the four segment base features of customers by *Demand function*, *Value* (Event impact), *Intention to respond positively,* and *Event likelihood*. The profiling of the segmented root cause follows later.

We have ordered the presentation of the profiles according to their *Event likelihood*, which is in following with Risk Management practice.

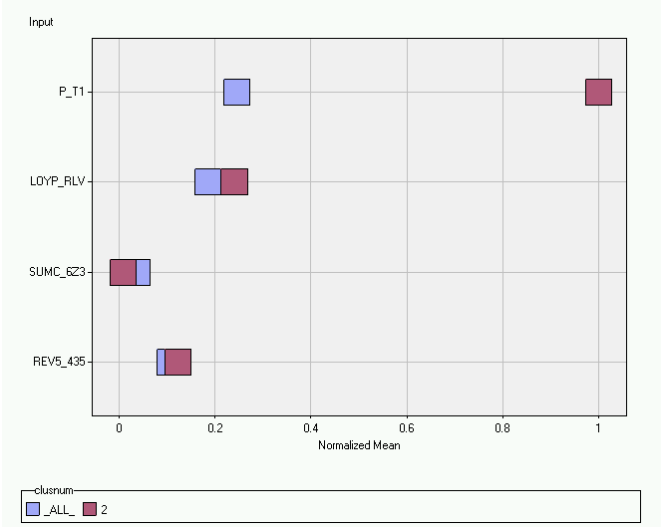


Figure 7.1: Segment 2 – Bulls

The profile of segment 2 is visualised in Figure 7.1. We called segment 2 *Bulls* because:

- they have the highest *Event likelihood* among the segments, and are about to crash through the gate to chase some cows; but
- they have above average *Intention to respond positively* – in reliable Taurean fashion, they will stay put if the cows are brought to them;
- they have average *Value* and *Demand function*.

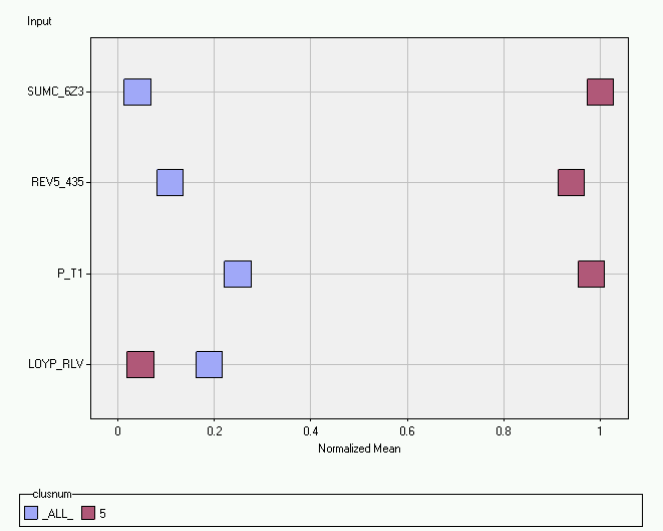


Figure 7.2: Segment 5 – Cash cows

The profile of segment 5 is visualised in Figure 7.2. We called segment 5 *Cash cows* because:

- they have high *Event likelihood* in comparison to the other segments - they are being lured by greener pasture; and
- they have below average *Intention to respond positively* – it will be difficult to green their existing pasture to retain them; and
- they have high *Value* and *Demand function* – therefore milk them while you can.

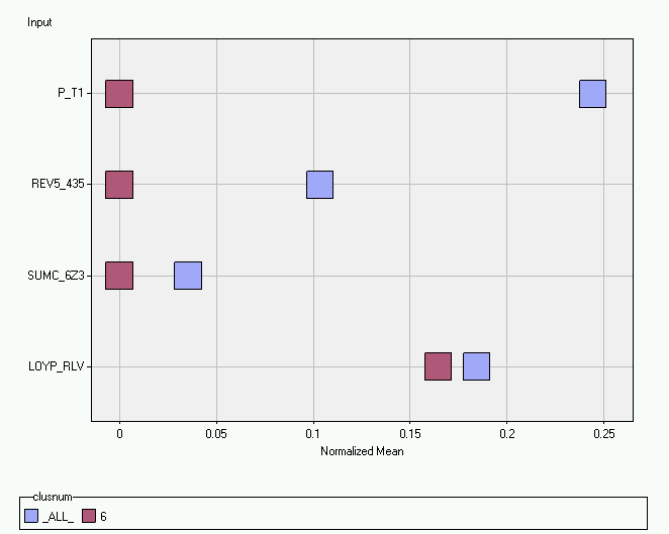


Figure 7.3: Segment 6: Stingy stodgies

We visualised the profile of segment 6 in Figure 7.2. We called segment 6 *Stingy stodgies* because:

- they have the lowest *Event likelihood* in comparison to the other segments - they will be the last to leave the pub; while
- they have substantially below average *Value* and *Demand function* – they don't spend much while they are in the pub; but
- they have about average *Intention to respond positively* – which means they will probably be back again in future.

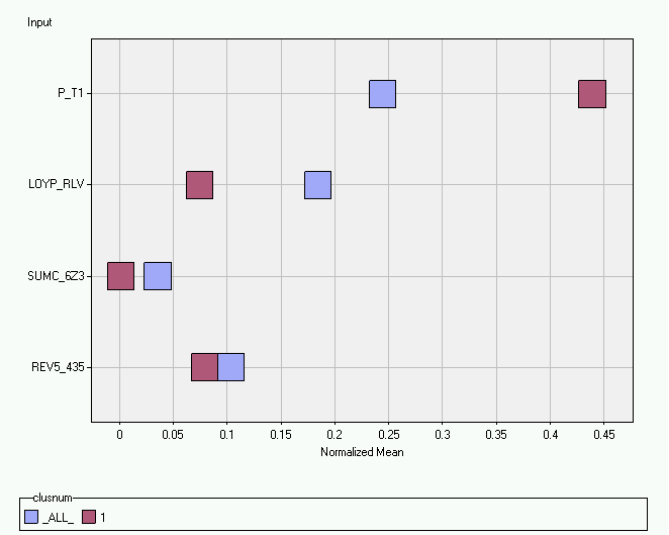


Figure 7.4: Segment 1: Disloyal dogs

Figure 7.4 is the profile of segment 1. We called segment 1 *Disloyal dogs* because:

- they have quite above average *Event likelihood* in comparison to the other segments - they are about to follow the *Cash cows* and the *Bulls* in leaving the farm; and
- they have below average *Value* and *Demand function* – they are worth less than the other farm animals while they are around; and;
- they have quite below average *Intention to respond positively* – they have less loyalty than the other farm animals.

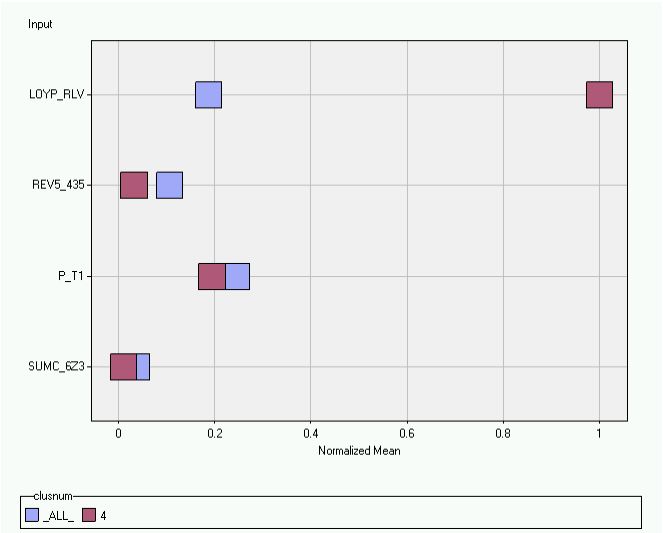


Figure 7.5: Segment 4 – Loyal friends

In Figure 7.5 we present the segment 4 profile. We called segment 4 *Loyal friends* because:

- they have below average *Event likelihood* in comparison to the other segments - they will only desert you after all your animals have left you; but
- they have very high *Intention to respond positively* – they will stay with you if you ask them not to desert you; and
- they have more *Value* and *Demand function* than the *Stingy stodgies* – they are worth more to you than the *Stingy stodgies*.

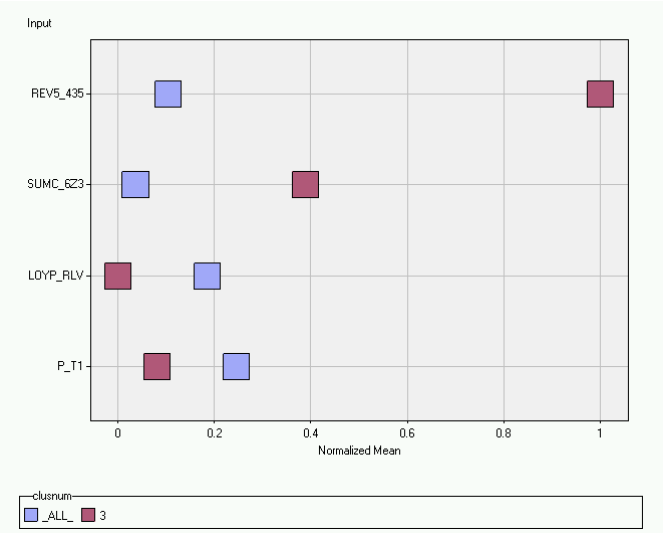


Figure 7.6: Segment 3 – Wealthy assertive friends

In Figure 7.6 we present the profile of segment 3. We called segment 3 *Wealthy assertive friends* because:

- they have an *Event likelihood* between the *Loyal friends* and the *Stingy stodgies*; but
- they less *Intention to respond positively* than the *Loyal friends* and the *Stingy stodgies* – by nature they are assertive in pursuing their own interests; and
- they have high *Value* and *Demand function* – they are worth a lot to you as friends.

This concludes the base profiling of the segments.

7.1.1.1.2.2 Root cause profile

We now enter the domain of inter-segment profiling for root cause. In this section, we refer back to Tables 6.8 and 6.11, and to Table 7.1.

Effect A overall impact profile in RetSegmonth $_{t+1}$ (Segments 1 - 3 displayed)						
Feature A labels [A]	Label's impact portion within Segment 1 [E1]	Label's impact portion within Segment 2 [E2]	Label's impact portion within Segment 3 [E3]	Label's portion of overall impact within RetSegmonth $_{t+1}$ (Segment 1) [O1]	Label's portion of overall impact within RetSegmonth $_{t+1}$ (Segment 2) [O2]	Label's portion of overall impact within RetSegmonth $_{t+1}$ (Segment 3) [O3]
#	36.76%	56.91%	4.38%	7.54%	7.40%	0.27%
AA			1.98%			0.12%
AB	0.24%		1.46%	0.05%		0.09%
BA	1.77%	4.84%	1.80%	0.36%	0.63%	0.11%
BB	1.69%	2.94%	3.42%	0.35%	0.38%	0.21%
BC	2.95%	6.16%	8.95%	0.61%	0.80%	0.54%
BD		0.13%			0.02%	
BE						
DA			3.50%			0.21%
DB						
DC	18.36%	8.45%	6.07%	3.76%	1.10%	0.37%
DD	1.18%	0.44%	3.58%	0.24%	0.06%	0.22%
E	4.01%	0.49%	12.13%	0.82%	0.06%	0.74%
FA		0.40%	1.64%		0.05%	0.10%
FB		0.87%			0.11%	
GA						
GB	0.06%	0.28%			0.04%	
H						
JA	0.19%	1.14%	2.31%	0.04%	0.15%	0.14%
JB	5.45%	4.78%	7.78%	1.12%	0.62%	0.47%
JC	1.23%	1.53%	7.45%	0.25%	0.20%	0.45%
JD	1.16%	0.46%	9.81%	0.24%	0.06%	0.59%
JF	17.49%	5.33%	12.06%	3.59%	0.69%	0.73%
JG		0.43%			0.06%	
JH	5.45%	3.83%	11.67%	1.12%	0.50%	0.71%
JJ	2.01%	0.59%		0.41%	0.08%	
T						
Totals :	100.00 %	100.00%	100.00%	20.51%	13.01%	6.07%

Table 7.1: Effect A overall impact profile in RetSegmonth $_{t+1}$

The *Total positive T-score in segment…* (cells I, J, K in Table 6.11) allow for inter-segment comparison of *overall* root cause effect. In the case of segments 1, 2 and 3, they contain about equal proportions of overall root cause effect. However, to allow true comparative profiling, these figures need to be scaled for the number of consumers in each segment. We achieve this scaling by multiplying the within segment % value - the column E of Table 6.11 – with $\text{RetSeg}_{\text{month t+1}}$'s *Segment's percentage of total number* from Table 6.8. We display the results of the multiplication in Table 7.1.

The inter-segment comparative values for segments 1 – 3, are given in columns O1 to O3. Segment 1's column E has been carried over from Table 6.11, and we now call it column E1. Columns E2 and E3 are previously undisplayed column E values for segments 2 and 3.

We now have the *intra-segment* comparative figures for segments 1 – 3 in the columns E_n, as well as their *inter-segment* comparative values in the columns O_n. So for instance, the 36.76% of label # in column E1 was multiplied by the 21.5% of segment 1 in Table 6.8. The product of 7.54% is displayed in column O1. Comparing the inter-segmental impact of label # between segments 1 – 3, we see that this Handset Type represents equal proportions of root cause effect in segment 1's column O1 and segment 2's column O2, but almost no impact in segment 3 (column O3).

To demonstrate the executibility of this *inter-segment* profiling, we use label # as an example. Allocating resources to eliminating label # in the three displayed segments in Table 7.1, will address about 15% of overall root cause; which is 7.54% in segment 1 (column O1), 7.40% in segment 2 (column O2), and a small portion in segment 3.

The percentages in the columns E_n sum to 100; this is apparent from the 100% in the last line of *Totals* in Table 7.1. The sums of the percentages in the columns O_n *do not* add up to 100% within the column. The importance of this second point, is that the sums of the columns O_n add up to 100% horisontally. This is not apparent from Table 7.1, because it only displays the values for three segments; the 20.15% (column O1), 13.01% (column O2), and 6.07% (column O3). If this table contained all six segments, then the *Totals* for all the columns O_n would have added up to 100%.

7.1.1.1.3 Positioning

The executibility of the *Positioning* lies in the design of the *content* of the targeted retention campaign offer for each segment, and the *sequencing of their execution*. The *content* is the way the company position itself toward the targeted customer in terms of the base profile *and* the root cause profile. The *sequencing* refers to a temporary resource prioritisation issue, in other words, which offers to execute first, second etc. In this section, we present some possible retention offer *content* and *sequencing*. The design of these offers, is based on the following three dimensions:

1. the behavioral and psychographic profiles;
2. incentivising combined with the root cause profiles for Effect A;
3. the prioritisation of the retention management resources within a Risk Management framework. Resources are call center time and financial resources for offering incentives.

7.1.1.1.3.1 Offer content in response to behavior and psychographics

Customers' needs may change after they have committed to a service plan. There is ample evidence in the marketing literature, that such change in a customer's need, may result in customer dissatisfaction (Engel, Blackwell et al. 1995) (Goncalves 1998) (Kurz and Clow 1998) (Rosa 2002) (Yassael 1998) (Lovelock 2000). Basing the offer on the current behavioral and psychographic measures, assures that the offer meets the customer's current needs, and removes any latent dissatisfaction and therefore a potential cause of churn.

Offer content which is matched to the behavioral and psychographic profile, has two dimensions:

- the first is the *duration of the agreement* being offered. We base this on the customer's measure about *Intention to respond positively.* The higher a customer's *Intention to respond positively*, the longer the duration of the agreement, which should be offered. Practically:
  - to those segments with a mean measure for this feature, offer an agreement which is similar in length to the one they are covered by (or where covered by if it has expired);

- to those segments where the measure is below the mean, offer an agreement similar in length to the one they are currently covered by, but as an incentive call center staff may negotiate to an agreement which is shorter in duration to the one they are now covered by;
- to those segments where the measure is above the mean, offer an agreement longer in duration to the one they are now covered by, but as an incentive may negotiate to one of similar length to the one they are now covered by;

- the second is a *plan* with a rate structure which best suits the customer's current call patterns. The components of rate structure are a combination of the quantity and duration of calls the customer makes. *Demand function* gives us the amount of calls. *Value* comprises of the flag fall component of call cost, and of the duration component of the call cost; *Value* therefore is a proxy for the duration of calls. Combining *Demand function* with *Value*, we can determine the plan with the best call rate structure for a customer's use. So for instance if a customer makes many short calls, offer a plan with a rate structure which is tailored for that behavior.

Practically, this part of the offer content can be based on the organization's existing plans repertoire, as long as we change the plan selection criteria for the new approach.

7.1.1.1.3.2 Offer content for addressing root cause

A third component of the offer is whether to address root cause or not, and how to present the solution to such a sensitive issue, to the customer in a way which he/she will perceive as attractive.

We believe the tactic for achieving this, is by casting any offer content about root cause, within an *incentive* framework. Because incentives have the characteristic of attractiveness, they have the further benefit of improving the response to the campaign offer. This makes sense when we consider that incentivising the offer is common practice in the telephony industry. Further, the industry has a well-established practice of incentivising with replacement handsets. In our case, we know that Handset Type is the major cause of churn.

Offering replacement handsets as incentive in some cases, then will serve a double purpose for Telco ABC:

- it will improve the customer's response to the campaign offer; and

- will support Risk Management practice, by removing a main contributory factor to *Event likelihood*.

The conditions for the handset being perceived as attractive by the customer, and for the handset removing a root cause of churn, are:

- it must be a handset which does not have any quality or reception problems associated with the customer's existing handset; and / or must
- be better suited to the customer's current use pattern than their existing handset; and
- must have current features.

Replacement handsets cost money, and we know that Telco ABC has limited resources for incentivising. Risk Management practice directs resources to those areas, which will give the biggest effect on reducing the *Event likelihood.* We therefore firstly want to prioritise the incentivising toward those potential churners, who have handsets who present the most risk.

Further, since incentivising is about making the offer more attractive, we should concentrate the limited incentivising resources, on those segments from which we expect lower response rates - even to a well-targeted campaign offer. These are the segments with the lowest mean value for the measure *Intention to respond positively*.

Combining the incentivising with addressing the root cause, we then incentivise those customers who have high positive T-score handsets, and belong to a segment with a low *Intention to respond positively*.

The availability of resources for incentivising, will determine how many segments – and potential churners we incentivise. The availability of resources may partly depend on the degree to which Telco ABC's can structure the incentive in a revenue-neutral way i.e. can we supply such replacement phones at cost, or at a loss, or at a profit. Depending the amount of resources available, we present the following possible prioritised approach to incentivising:

- in the event where Telco ABC has unlimited resources for incentivising – perhaps through incentivising in a revenue neutral or profitable way –they will offer a suitable replacement handset to each potential churner with a positive T-

scored handset, in all three segments which have a below mean value for *Intention to respond positively*. These three segments are segments 6, 1, and 3;

- in the event that Telco ABC has substantial funds for incentivising one big segment, they would offer the incentive to those potential churners in segment 6, which have positive T-score handsets. The reason is that segment 6 represents 55% of all the root cause. We know this figure, because even though we have not displayed column O6 in table 7.1, the *Totals* for those columns, are the same as the segment's proportion of segment frequency in Table 6.8. From Table 6.8, we know segment 6's proportion is 55%;
- in the event that the resources are not that substantial, then those potential churners with positive T-score handsets in a smaller segment should be incentivised. Of segments 1 or 3, we would incentivise such potential churners in segment 3. The reason is their *Intention to respond positively* is about the same, while the mean *Value* of segment 3, is higher than segment 1's mean *Value*; this means more future value per customer who responds positively in segment 3, than in segment 1;
- in the event that the resources are so limited, that the whole of segment 3 cannot be incentivised, then we refer to column E3 in Table 7.1, and incentivise those customers who currently have the phones with the biggest T-score within segment 3. The choice then is to incentivise those customers with phones labeled *E* and possibly *JF*. We base our following possible retention campaign offers, on this severely restricted resource scenario.

Incentivising needs to be combined with the authority for the call center staff, to use the incentive as a negotiating instrument, with which to elicit a positive response. The execution of the incentive is completed when the new handset is delivered to the potential churner.

7.1.1.1.3.3 Prioritisation of campaign execution between segments

We explained in chapter two, how Risk Management is used to prioritise limited resources, toward minimising the *event likelihood* and *event impact* of an *unwanted event*. Our *unwanted event* is potential voluntary churn. Our unwanted *event likelihood* is captured by the segment base feature *Event likelihood*. Our unwanted *event impact* is

captured by the segment base feature *Value*. Our *unwanted event* has a time dimension. Following Risk Management practice, that means that we have to sequence the execution of the retention campaigns between the segments over the time window.

The prioritisation is based on *Event likehood* and *Value*. The retention campaigns are executed first in those segments with a combined highest *Event likelihood* and *Value,* and last in those segments with a combined lowest *Event likelihood* and *Value.* This adds a fourth component to the offer of prioritising execution over time lapsed.

We also develop an operational benchmark for the *amount of time to be spent* on executing each segment's campaign offer. We know from earlier sections, that the campaigns will be executed monthly. The benchmark expresses the portion of retention campaign man-hours, which should be allocated to the execution of an offer within a segment. We do not know Telco ABC's total man-hours, which are available for retention campaign execution, meaning we cannot calculate the benchmarks in hours. We therefore calculate the benchmark as a proportion of the available man-hours. That proportion is equal to the proportion of a segment's frequency in $RetSeg_{month\ n}$. Those proportions are given in Table 6.8 *Segment's percentage of total number*. This adds a fifth component to the offer, that of benchmarking the retention management operations for productivity.

7.1.1.1.4 Retention campaign offers

We now offer executable knowledge about possible retention campaign offers based on the above. The offers are broken down into their five components, by segment.

7.1.1.1.4.1 Segment 5 – Cash cows

1. *execution priority* – execute the campaign to this segment first, because this segment has a very high *Event likelihood* and *Value*;
2. *agreement duration* – similar to the current agreement duration, but allow negotiation about a shorter duration as an incentive;
3. *plan* – rate structure suitable for many short calls;
4. *handset incentive* – none;
5. *operational benchmark* – 1% of retention management man-hours.

7.1.1.1.4.2 Segment 2 - Bulls

- *execution priority* – execute the campaign to this segment second, because this segment has a very high *Event likelihood* but average *Value*;
- *agreement duration* - similar to the current agreement duration, no negotiation allowed;
- *plan* – rate structure suitable for average number of calls and call duration;
- *handset incentive* – none;
- *operational benchmark* – 13% of retention management man-hours.

7.1.1.1.4.3 Segment 1 – Disloyal dogs

- *execution priority* – execute the campaign to this segment third, because this segment has a higher *Event likelihood* than segments 4 or 3, and the same measure of *Value* as segment 4. Segment 1 has a larger frequency than segment 4, which means the overall *Value* retained from segment 1 is higher than that of segment 4. The offer to this segment should therefore be executed before that to segments 4 or 3;
- *agreement duration* – similar to the current agreement duration, but allow negotiation about a shorter duration as an incentive;
- *plan* – rate structure suitable for average number of calls and average duration;
- *handset incentive* – none;
- *operational benchmark* – 20% of retention management man-hours.

7.1.1.1.4.4 Segment 4 – Loyal friends

- *execution priority* – execute the campaign to the segment fourth, because this segment has a higher *Event likelihood* than segment 3, but lower *Value* than segment 3. Despite segment 4's lower *Value*, it is a relatively small segment, which means it will be a fast execution before starting on segment 3. Fast execution will be supported by segment 4's significantly high *Intention to respond positively*;
- *agreement duration* – longer than their current agreement duration, but allow downward negotiation as an incentive;
- *plan* – rate structure suitable for average number of calls, or below average duration;
- *handset incentive* – none;
- *operational benchmark* – 5% of retention management man-hours.

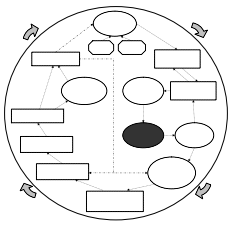
7.1.1.1.4.5 Segment 3 – Wealthy assertive friends

- *execution priority* – execute the campaign to this segment fifth. The execution of the offer to segment 3 precedes the execution of the offer to segment 6, because segment 3 it has both higher *Event likelihood* and *Value* than segment 6;
- *agreement duration* – similar to the current agreement duration, but allow downward negotiation as an incentive;
- *plan* – rate structure suitable for less calls than segment 5, but about double the call duration of segment 5;
- *handset incentive* – offer customers with handset types *E* and *JF,* replacement handset *X* , where handset *X* has no known quality, has superior reception, has the latest features, is easy to operate, and has good sound quality. Phone X's ease of operation is suitable for those who make an above average number of calls, its good sound quality is suitable for those making long calls. As an alternative, offer a handset type, which has a negative t-score in the model. The latest features will be attractive to the *Wealthy assertive friends*, who tend to be peer image conscious;
- *operational benchmark* – 6% of retention management man-hours.

7.1.1.1.4.6 Segment 6 – Stingy stodgies

- *execution priority* – execute the campaign to this segment last, because this segment has the lowest *Event likelihood* of all the segments;
- *agreement duration* – similar to the current agreement duration, but downward negotiation allowed as an incentive;
- *plan* – rate structure suitable for above average number of calls, and above average duration. Possibly the same as for segment 5;
- *handset incentive* – none;
- *operational benchmark* – 55% of monthly retention management man-hours.

We have now developed the *desirable* executibility of the retention management solution under the organisation's circumstances – at a planning level.



7.1.2 Strategic analysis

Historically Telco ABC has had an average 25% response rate with their existing retention solution. This means that they were about 25% successful at preventing voluntary churn. The dollar value of their existing retention solution therefore is 25% of the problem magnitude we calculated in *Identify, assemble, prepare useful data*, or 0.25 x AUD113 m. = AUD28.25 m. per year.

We estimate that using targeted campaign offers like the above, should improve retention campaign response rates to at least the industry benchmark range of 35% to 50%. The incremental dollar value of our retention approach compared to the existing approach, constitutes the *return* from this project. At 35% - 50% response rates, our approach would be worth between 113 m. x 0.35 or AUD39.7 m. and 113 m. x 0.5 or AUD56 m. At the lower response rate the incremental dollar value is our 39.7 m. minus the existing 28.25 m, or 14.4 m. At the higher response rate the incremental dollar value is our 56 m. minus the existing 28.25 m. or 27.75m. The *range of incremental revenue* from our project therefore is AUD14.4 m. to AUD27.75 m.

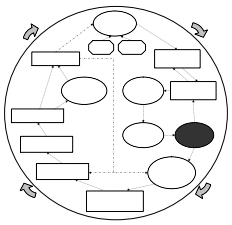
We now use this incremental revenue range to calculate the potential ROI of our project. To do this we need to calculate the incremental costs to Telco ABC from executing the new retention management approach. Our high-level description and calculation of the incremental costs are:

- Telco ABC could use their existing marketing talent after training key staff for *understanding* the propensity modeling and segmentation approach and output, and for *interpreting* that for campaign design. A one day course for 5 managers (Retention, Content, and Product Managers) would suffice. This one-off training could be done by an external expert at a *cost* of about $5 000;
- Telco ABC would use their existing 7-S execution toolbox. However, call centre staff may need training in negotiation skills, required by some of the segments. This training costs about $500 per person per day. A one-off, one day course would suffice. Assuming Telco ABC have 100 people in their save team requiring this training, the cost would be $50 000;
- maintaining the new analytics data mart would require not more than 2 working days per month for one senior analyst. At a lavish $400 per day, that would amount to $800 per month, or $7 200 per year;
- operationalising and maintaining the models would require another one day per retention management cycle. Assuming a retention management cycle of a calendar month, using the same senior analyst this would add $4 800 per year;
- monitoring and controlling the business components of project ROI, campaign response rates etc. would generate any incremental cost. This would simply replace monitoring and controlling the outgoing solution;
- monitoring and controlling the supporting analytics for concept drift, strategic creep etc. would require one day's training of the senior analyst by an expert at $5000. The monitoring and control would take up one day per cycle, which is an incremental $400 per day or $4 800 per year. The subtotal incremental cost here is $5 000 plus $4 800 or $9 800.

Adding up all the incremental cost gives a total figure of AUD 76 800 per year. We now calculate the Return on Investment as the low-end incremental dollar divided by the incremental cost, or $14.4m \div 76.8k$

$$\text{ROI} \approx 187.5 \text{ times}$$

7.1.3 Strategic choice



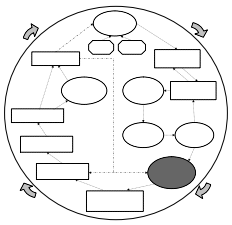
We evaluate that there are no pressing limiting circumstances, why the above solution is not workable in Telco ABC's circumstances, and we have therefore attained the first SPC 4 goal for developing the executibility of the business solution.

Further, the potential ROI from the project of about 187.5 times is substantial. This ROI presents a very attractive justification to continue with the project into the *Define* and *Realise* phases.

We recognise that our business evaluation of limiting organisational circumstances was not an in-depth one; such analysis is not the purpose of this discourse. The purpose of this section simply was to demonstrate the mechanism of SAM in producing executibility, which considers organisational circumstances. Were an in-depth business evaluation of the circumstances in this KM loop, to have found non-executibility of the business solution, then SAM offers utility in overcoming that by another iteration through the SPC 3 goals, or a reformulation of the project.

There remain the practical execution issues, which concern the dimensions of the 7-S implementation framework. In practice, where an organisation depends on SAM, those issues are addressed in the later SAM business activity set *Execute new business strategies.*

7.2 Define new business objectives and strategies



7.2.1 Execute SPC 4 SPC strategy two

Now we are in the business solution domain, where we will redefine business objectives and strategies. This is to develop the solution into a format for knowledge sharing, and for applying the mapping technique for developing the supporting data mining plan. The new objectives and strategies further define the paradigm shift, which we have achieved on Telco ABC's retention management approach, using SAM and data mining as tools.

7.2.1.1 New *Strategic objective*

The original *Strategic objective* about competitive position referred to above industry average churn rates. That strategy was influenced by knowledge, which the organisation

had developed through a previous application of SPM, specifically in SPM 1.2. We now have developed new knowledge using SAM, allowing us to reformulate the existing *Strategic objective* as:

- reduce voluntary churn rates to competitively acceptable industry levels.

Note how that SAM provides utility for expressing strategic change as a new *Strategic objective*.

7.2.1.2 New *Grand strategy*

The original *Grand strategy* remains unchanged. However, any reformulation would constitute *strategic paradigm shift*.

7.2.1.3 New *Operating objective*

We modify the original *Operating objective* for market retention to:

- halve the voluntary consumer churn from an established 1.5% to 0.75% within 3 months, and maintain that level.

Again, note how using SAM has allowed us to formulate strategic change, this time at an operational level. First, we have *redefined the success criterion* for this objective. We base the halving of the success rate on increased campaign response rates. Further, we *focus the objective* on the consumer business. In addition, we *purify the objective* of involuntary churn. Last, we *define the time* over which the results are measured to match the 3-month churn event window.

7.2.1.4 New *Operating strategies*

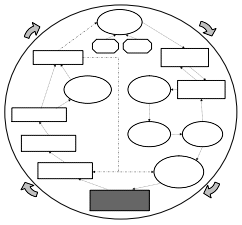
We modify the original *Operating strategies* for market retention to:

1. using data mining to identify potentially high-risk voluntary consumer churners every month, with 95% confidence that we are doing at least five times better than chance;
2. targeting the retention campaigns at those consumers who fall outside the 1.5% *risk tolerance*;
3. segmenting the targeted consumers simultaneously by their *call demand function*, their industry-use *ARPU*, their *event likelihood,* their psychographic measure of *intention to respond positively*, and their existing *Handset Type*, to

70% confidence about the true market structure and at least 1.5 times as much difference between the segments than within them;

4. developing retention campaign offers targeted at the profile of consumers by segment;
5. incentivising the offers with replacement problem Handset Types;
6. executing the campaigns for time and monetary resource constraints.

We retain the *Strategic objective* about the in-house campaign design unchanged. Note how SAM has allowed us to express operational paradigm shift with new *Operating strategies*.



7.3 Develop data mining plan

We are in a position now to define the components of the data mining plan. We now call the result a *plan* and not a mission, because for the first time we have certainty about the outcome from information discovery, and how we can attain that technically.

7.3.1 Data mining objectives

In this solution environment, we map the business's *Operating strategies* to the data mining objectives. These data mining objectives become to:

- identify every month the Telco ABC's consumers' risk of becoming a voluntary churner within the next three months, at a minimum 95% confidence in the model and a minimum 5 times lift;
- identify those consumers who fall outside the 1.5% *risk tolerance* of potential voluntary churner;
- discover the segment membership of those identified consumers with at least 0.70 R^2 and at least 1.5 times Fisher's Criterion overall;
- operationalise the data mining solution for minimum human involvement;
- deploy the above discovered information into the marketing department; and
- develop a monitor and control plan for the key informational concepts of the above.

7.3.2 Data mining strategies

In this section, we present the *Operating strategies* of the data mining plan. These strategies support the data mining objectives. In this section, *month* denotes a complete financial month. The financial month mostly ends on the first working day of a new calendar month, when the fourth and final weekly billing is run. The subscript symbol $_t$ denotes the most recent complete financial month $_t$, within which the most recent churn falls. That means that $month_{t+1}$ is the present financial month, which is also when we execute the strategies. We describe the strategies:

1. Extract into staging tables on the fourth working day of $month_{t+1}$ - from the Billing, Calls, and CRM databases:

   - the 10 statistically significant features;
   - the features required for creating the bad debt flag;
   - the features for distinguishing handset and plan upgrades from churn; and
   - the features required for distinguishing a consumer from a non-consumer;

   of financial $month_{t-1}$, $month_{t-2}$, $month_{t-3}$, $month_{t-4}$, and $month_{t-5}$. Notice how we allow an additional two months' potential churn signal for improving the signal in the data. Also create SUMC - which we also know is statistically significant - during the extraction. Assemble one flat table called Customer Data $Repository_{month\ t}$ (= $CDR_{month\ t}$), with 11 statistically significant modeling features i.e. 11 features of significance - and 1 000 000 observations;

2. the transactional (dynamic) data of churners only in $CDR_{month\ t}$ are incomplete, because the churners would have stopped transacting somewhere during the $month_t$. Bring each of $month_{t-1}$, $month_{t-2}$, $month_{t-3}$, $month_{t-4}$, and $month_{t-5}$ *churners' only* transactional data forward by one month in $CDR_{month\ t}$, overwriting each month's existing transactional records with that of the previous month. Call this table $CDR_{mod\ t}$ where the $_{mod}$ is for 'modified';

3. create a binary bad debt flag, and drop all observations which are bad debtors, or not 'Consumer' from $CDR_{mod\ t}$. Drop the bad debt indicator; over 985 000 observations remain in $CDR_{mod\ t}$;

4. create the binary label for each observation having become a churner or not in the previous calendar month, excluding service plan and handset upgrades from the churn event. We now have 13 features in the data, including the churn target label;
5. create $CDR_{month\ t-1}$, $CDR_{month\ t-2}$, $CDR_{month\ t-3}$, and $CDR_{month\ t-4}$, and prepare $CDR_{mod\ t-1}$, $CDR_{mod\ t-2}$, $CDR_{mod\ t-3}$, and $CDR_{mod\ t-4}$ in a similar way as above;
6. add to $CDR_{mod\ t}$ the observations of past *churners only* from tables $CDR_{mod\ t-1}$, $CDR_{mod\ t-2}$, $CDR_{mod\ t-3}$, and $CDR_{mod\ t-4}$, creating $CDR_{modplus\ t}$. We now have the data of non-churners of $month_t$, and of the actual churners of that month and of four more previous months;
7. open a project diagram in a data mining platform like SAS Enterprise Miner, and execute the value replacements and imputations, and replace or filter outliers. Note how we now allow for the replacement of outliers in addition to filtering them for improving the modelling. Further note there is no strategy for eliminating collinearity, because we have pre-knowledge about which features are collinear;
8. segment $CDR_{modplus\ t}$ (Van Everen 2002) by unsupervised k-means and Wards clustering on three selected features - *Rev53*, *LoyP*, and *SumC*, adding to $CDR_{modplus\ t}$ the feature *SegMas* for cluster membership. (Note we do not add the *DistMas* feature, because we know it is statistically insignificant). There now are 14 features in the data, including *SegMas*;
9. sample from $CDR_{modplus\ t}$ to form the Modeling Data Set (MDS_t). Sampling is stratified on the churn label, resulting in an equal number of observations of five months' churners and one month's non-churners;
10. do final data transforms to discretise interval variables;
11. partition MDS_t into $Training_t$ and $Validation_t$ sets on a ratio of 7/3, each partition evenly stratified on the churn label;
12. train a Logistic Regression model on $Training_t$ by stepwise effect selection to a significance by p-value of 0.05. Select the best model by its accuracy in classifying actual past churners within $Validation_t$;

When marketing are ready to design their first campaign in a future month $_{t+n}$:

13. repeat – after the first monthly billing cycle has been completed – above steps 1, 3, and also the removal of the handset and plan upgrades in step 4, creating $CDR_{month\ t+n}$ for that month;

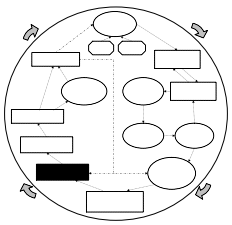
14. apply the score code from the LogReg to that newly created $CDR_{month\ t+n}$. The score code automatically repeats step 7 and 10. We call the scored data set $SCR_{month\ t+n}$;

15. sort $SCR_{month\ t+n}$ descending by the feature *P_T1*, and keep the top observations, which fall outside the risk tolerance of 1.5%. This results in the table Retention Segment $Month_{t+n}$ ($RetSeg_{month\ t+\ n}$), with about 15 000 observations;

16. hierarchically segment $RetSeg_{month\ t+n}$ using k-means and Ward's algorithm for a minimum number of clusters, which capture an overall R^2 and give a minimum Fisher's Criterion of 1.5. As clustering base use:

    - average three month revenue (*Rev5* or *ARPU*);
    - risk *event likelihood* (*P_T1*);
    - the sum of voice calls in the last three months (*SumC*);
    - the measure of customer loyalty (*LoyP*).

7.4 Model, evaluate, choose best model(s)



In accordance with SAM's allowance, we experimented with the strategies in the data mining plan, to optimise the quality of the two models for the plan, and to optimise the content of the discovered information for supporting the new business solution.

We comment that the strategies we designed in the formal Telco ABC data mining plan presented above, would be suitable for execution by a data mining engineer who is *resident within the organisation* at the time of executing *Model, evaluate, choose best model(s).* At the time the candidate executed *Model, evaluate, choose best model(s),* he was not resident in the organisation, and had to adapt the strategies for improving the models, to the available Telco ABC data *in his absence.*

We described in Chapter 6 of this thesis, how we had created a data set $RetSeg_{month\ t+1}$, which we used during the execution of the SPC 3 segmenting strategies in *Data mining discovery.* The available data to the candidate at the time of this experimentation, was a sample of the top *10%* by *P_T1* of the data set $SCR_{month\ t+1}$, about 98 000 observations

and not just the 5 000 observations of confidence we earlier targeted. Further, he had *five* of these 10% samples, being $\text{RetSeg}_{\text{month t-1}}$, $\text{RetSeg}_{\text{month t-2}}$, $\text{RetSeg}_{\text{month t-3}}$, $\text{RetSeg}_{\text{month t}}$, and $\text{RetSeg}_{\text{month t+1}}$. The data had been deidentified, and the interval variables standardised for confidentiality reasons. This data also contained a label of actual past churners.

We will now give a description of how we attained executed *Model, evaluate, select best model(s)* using these data sets.

7.4.1 Optimise the classifier

We substantially improved the accuracy and lift of the LogReg model. We:

- we emulated adding the extra 2 month's actual past churners from month_{t-4}, and month_{t-5}, by concatenating our five $\text{RetSeg}_{\text{month n data}}$ sets, and deduping them. The deduping lost a few hundred past churners, who were reappearing in each month's data, adding noise. One possible explanation for this repetition is that our definition of the past churn event is not 100% effective. We now had a data set with about 490 000 observations, or about 415 000 non-churners and about 75 000 past churners of five months, called $\text{RetSeg}_{\text{5 months}}$. The prior probability of actual churn in this data set was 18%, and not 1.5% as in the operational data;
- aggressively reduced the variance in $\text{RetSeg}_{\text{5 months}}$ by replacing outliers with suitable maximum values;
- rebinned the interval variables optimally to target;
- no data cleansing was required, since the data had been cleansed before as described in Chapter 6;
- we randomly sampled again 15 000 churners and 15 000 non-churners giving us the data set $\text{RetSegSample}_{\text{5 months}}$ with about 30 000 observations;
- we split $\text{RetSegSample}_{\text{5 months}}$ into 70% $\text{Training}_{\text{5 months}}$ and 30% $\text{Validation}_{\text{5 months}}$; Following this we
- rebuilt a stepwise LogReg model, asking for a model with significance of 95%, and choosing the most accurate model on $\text{Validation}_{\text{5 months}}$.

7.4.1.1 Evaluate and select

The new model was 70% accurate on Validation $_{5\ months}$. We visualise the confusion matrix in Figure 7.7:

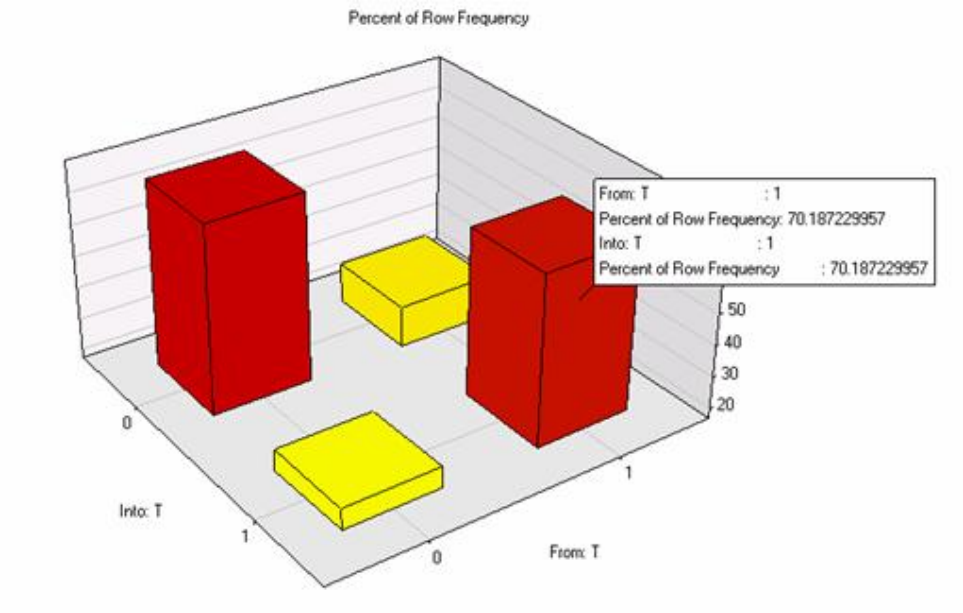


Figure 7.7: New classifier confusion matrix

The lift now on data with priors was 5.5, visualised in Figure 7.8:

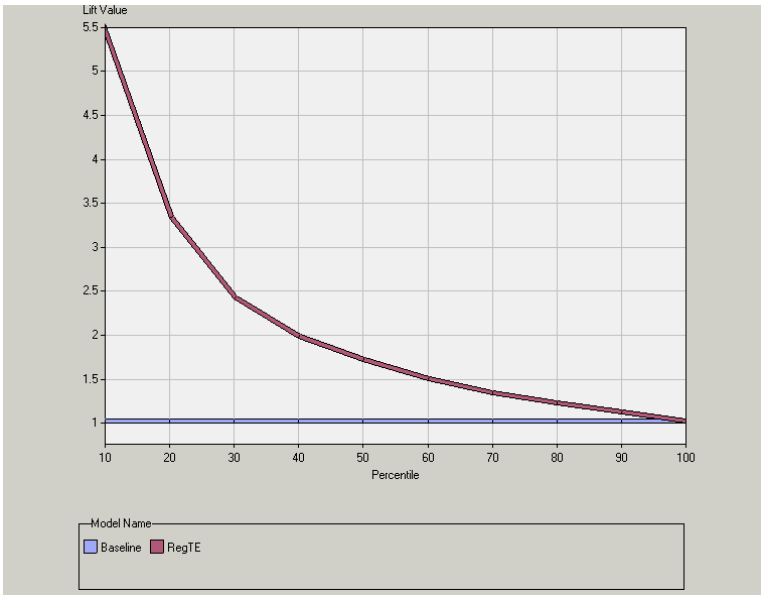


Figure 7.8: New classifier lift

The model also fares well in the category of ROC, visualised in Figure 7.9:

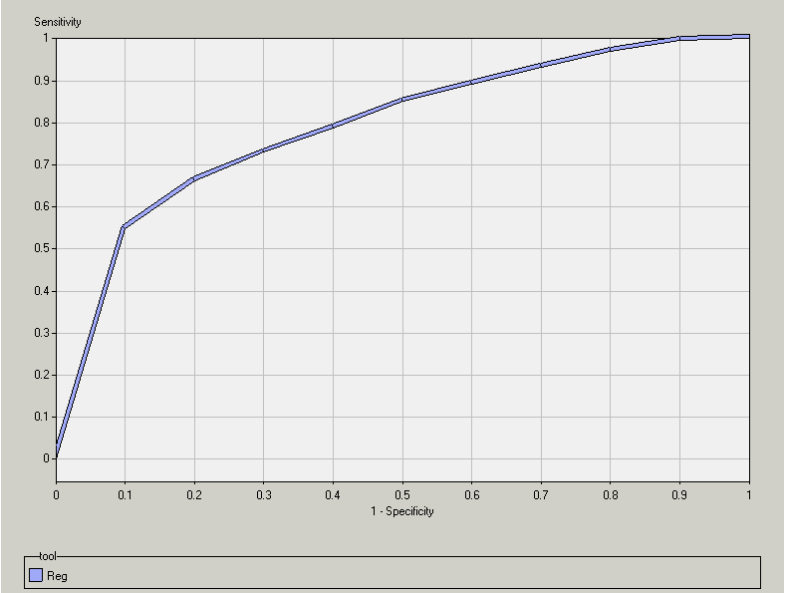


Figure 7.9: New classifier ROC

The confidence measure for repeatability – model significance – was in the 99% range, similar as the initial LogReg.

We tested the model by applying its score code to score each of the five available $RetSeg_{month\ n}$ data sets, in the process also creating five $SCR_{month\ n}$ data sets. The accuracy of the model on these five data sets was as follows:

- $RetSeg_{month\ t-3}$ – 50%;
- $RetSeg_{month\ t-2}$ – 79%;
- $RetSeg_{month\ t-1}$ – 73%;
- $RetSeg_{month\ t}$ – 63%;
- $RetSeg_{month\ t+1}$ – 91%.

There is obviously some variation between the months, but these test results are much improved over those in *Data mining discovery.* Considering the measures about the new model, we choose to operationalise it.

7.4.2 Re-target

As describe in the testing of the new LogReg model above, we had created five sequential monthly scored data sets. That emulated the scoring of the consumer database every month in steps 13 and 14 of the data mining strategies. We then sorted

these five $SCR_{month\ n}$ data sets descending by *P_T1* and retained the top 15 000 observations, emulating step 15 of the data mining strategies.

We named these five new data sets $RetSegFinal_{t\ n}$ (=$RSF_{t\ n}$ e.g. RSF_{t-1}). We evaluate that we have now have confidence in the targeting of all of the 1.5% of consumers, who fall outside the *risk tolerance*. We shall therefore proceed with the re-segmenting.

7.4.3 Re-segment

We clustered each of $RSF_{t\ n}$, this time using *hierarchical* clustering, which gives better results on larger data sets, than non-hierarchical just k-means. We initially asked for a minimum of six segments to match the existing quantity of six retention segments. We also used range as the standardisation technique. We also had replaced outliers in the features *ARPU* and *SumC*.

7.4.3.1 Evaluate and select

Overall Retention Segment Measures (RSF_t)					
Average intra-segment homogeneity:	0.14	**Overall R^2:**	71%	**Number of nearest segment profiles:**	5
Percentage of targeted consumers within segment:	1:	7%		**Average inter-segment distinctiveness (Fisher's Criterion):**	1.76
	2:	3%			
	3:	35%			
	4:	7%			
	5:	3%			
	6:	4%			
	7:	29%			
	8:	12%			
Relative importance of segment base features:	LOYP:	0.04			
	P_T1:	0.96			
	REV5:	0.64			
	SUMC:	1.00			

Table 7.2: Overall Retention Segment Measures (RSFt)

Neither a 6- nor a 7-segment approach could meet the confidence cut-off of 0.70 R^2. Asking for eight segments proved successful, and we present an example of the overall measures of confidence in the results we achieved in RSF_t in Table 7.2. We explain the need for eight segments, as there being more than twice the observations in the $RSF_{t\ n}$ data sets, than in the $RetSeg_{month\ n}$ data sets we encountered in *Data mining discovery.*

This means more clusters are required to capture the natural variance structure within the data.

Notice how on this data, our confidence by the overall Fisher's Criterion exceeds the 1.50 minimum we had set previously. The number of nearest segments now is compared to the eight segments, and this never fell below five. This is acceptable.

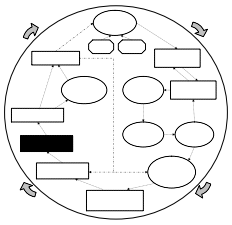
We observe that the *Relative importance of the segment base features* fluctuate between the months, with *P_T1* and *SUMC* always competing for leadership. *LOYP* always is the lowest, fluctuating between the 0.04 above and about 0.40 on two data sets. We will comment on this later in *Monitor and control*.

The different number of segments, result in *non-identical* segment profiles to the ones we had created in *Data mining discovery*. This is not in conflict with the new *Operating strategies*. The reader will recall that the new *Operating strategies* deliberately do not refer to the number of segments as the old strategies did, but to segmentation according to technical confidence measures. This is as an example of how the results of a SAM-driven project, should influence the business strategy *as suggested by the data*. The benefits of this approach will be improved distinction between the retention segments, with consequently better focused campaign offers resulting in improved campaign response.

The new segments are still profileable using similar technique and logic as before. Their profiles appear from one month to the next, allowing for some continuity in the campaign offers over time. In *Execute new business strategies* we will give an example of the profileability of these new retention segments, but to less detail than demonstrated before.

We are satisfied with the overall measures of segmentation quality, and will proceed with operationalising the segmentation model.

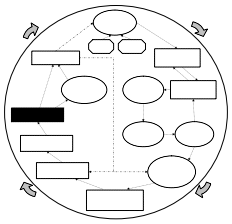
7.5 Operationalise model(s)



The model's operationalising strategies depend on the use of a project diagram, like that provided by SAS Enterprise Miner. In practice, we may use a scheduler to automate the execution of these strategies on for instance the fourth working day of every new financial month, after the last billing run of the previous financial month has been completed. The strategies were:

1. add a Code Node to the existing project diagram, and insert the BASE SAS code of step 13 into the Code Node. Link that Code Node to an Input Data Source Node;
2. set the data role in that Data Source Node as *Score*, and link this Data Source Node to the Score Node in the existing EM project;
3. link the LogReg Node to the Score Node;
4. insert another Code Node into the EM project after the Score Node, and in there place the code which generates the Excel tables required by marketing and for the Excel tables required later by *Monitor and control*;
5. use an Excel macro to extract the data tables from the SAS directories, and to generate the visualisation of the root cause profile;
6. when the information is required by marketing, run the Enterprise Miner nodes mentioned, and the Excel macro, and check the EM node and Excel logs for error messages.

Again, we do not intent to contribute to the wealth of literature available on this sub-topic, and these comments suffice for our purpose.



7.6 Deploy outputs into business

The deployment strategies were to provide to the marketing department:

- an Excel data table, which consists of the segment membership data feature of each of the 15 000 most at-risk customers, containing the phone id as the joining key to the CRM system;
- the quantitative and visual segment profiles by the four base features and Handset Type.

Here, we offer a deployment of the relevantly discovered information, to the business users who will be dependent on it in *Execute new business strategies.* Executing data mining strategy 23 above, provides an Excel table, which looks like the extract in Table 7.3:

Phone ID	Segment
54524	1
1548	5
356	8
44579	3
89488	4
68318	2
Etc.	Etc.

Table 7.3: Consumer segment membership

The *Phone ID* column contains the deidentified consumer phone number, which is the data key for joining the consumer's *Segment* membership to the CRM system. For instance, the consumer with phone number *356* belongs to segment *8*.

Table 7.4 contains the standardised means of the base features within each segment. These numbers are the *quantitative* segment profiles, upon which Telco ABC will base the profile descriptions, the prioritising of retention campaign resources, and the comparing of the segments over time.

The Retention Manager will do these segment descriptions (e.g. Dogs, Cows etc.) and comparisons, and the allocation of resources, every month in *Execute new business strategies.* In order to make his tasks easier, we have sorted the segments in Table 7.3 in the following way:

- descending by *P_T1*; then
- descending by *REV5*; and then
- descending by *LOYP*.

For example, in Table 7.4 segment 6 has the biggest mean value of *P_T1* and *REV5*, therefore it is at the top of the table. Note that segment 8 has a greater value for *P_T1* than segment 2, which does not display within the two decimal places here. Another possible logic for these tasks, could see the *LOYP* and *REV5* change position in the sort conditions.

Quantitative segment profiles					
Retention Segment number	**Number of customers in segment**	**Intention to positively respond**	**Value measure ARPU**	**Demand function Sum of voicecalls over 3 months**	**Event likelihood within time window**
CLUSTER	**Frequency of Cluster**	**LOYP**	**REV5**	**SUMC**	**P_T1 (sort by)**
6	114	-0.47	2.90	2.08	0.95
1	222	-0.47	1.40	0.42	0.94
7	861	-0.02	-0.48	-0.58	0.93
5	84	3.19	-0.33	-0.29	0.87
4	210	0.14	0.85	1.65	0.82
8	365	-0.12	0.36	0.23	0.81
2	96	0.00	2.71	1.91	0.81
3	1048	0.13	-0.56	-0.64	0.80

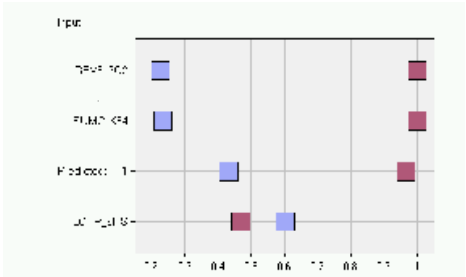
Table 7.4: Quantitative segment profiles

The other quantitative table which is deployed into the business, is Table 7.2 displayed before. In there we communicated our ongoing confidence in the technical results to the business as the *Overall* R^2 measure, and as the *Fisher's Criterion*. The relevance of the business solution over time depends on these concepts, and they are worthy of monitoring.

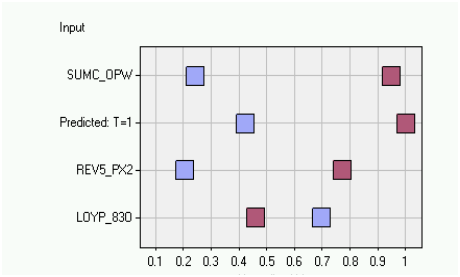
Table 7.2 also contains the *Percentage of targeted consumers within segment*, which is the number we base the allocation of retention management resources on. We base this on the *Frequency of Cluster* column in Table 7.4, as a proportion of the 15 000 targeted observations. This is worthy of monitoring over time.

Further, we deploy into the business the quantitative segmented root cause profiles. We gave examples of those in Tables 6.11 and 6.12.

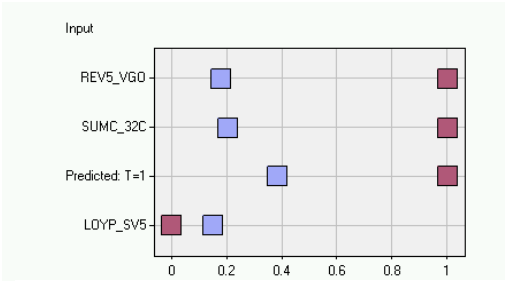
We saw examples of the visualisation of the segment profiles in an earlier section of this chapter, and an example of an intra-segment root cause profile in Figure 6:12. Following in Figure 7.10, we visualise the recurring monthly profile of the most at-risk segment, as we discovered it in our five data sets.



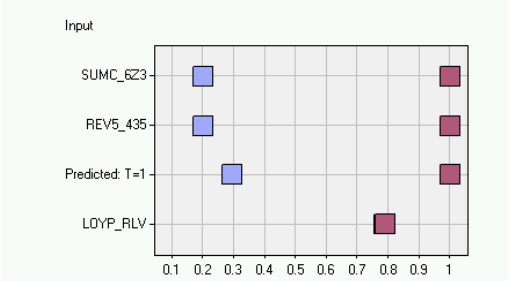
A: Segment 7 in month RSF_{t-3}



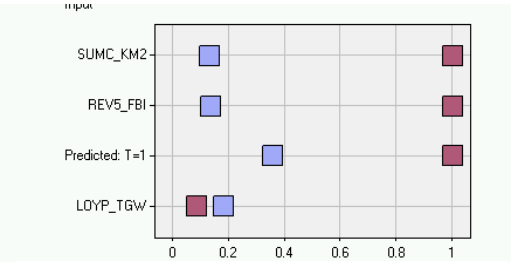
B: Segment 3 in month RSF_{t-2}



C: Segment 6 in month RSF_{t-1}



D: Segment 2 in month RSF_{t}

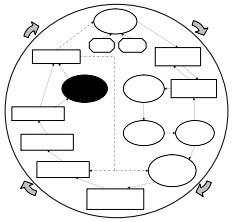


E: Segment 7 in month RSF_{t+1}

Figure 7.10 (A-E): Five monthly segment profile of most-at-risk segment

The business is particularly interested in recurring profiles from month to month, and Figure 7.10 demonstrates the consistency of the results in this regard, by example of one recurring profile. We consider that the order of appearance of the features in the Y-axis, is determined by their relative variance; thus in the first month the *Value* feature (*REV5*) had the most variance, placing it highest. In the second month, the *Demand function* feature (*SUMC*) had the most variance, placing *SUMC* highest. Overall however, we see the profile high *Demand*, high *Value*, high *Event likelihood*, and low *Intention to respond positively* occur every month.

7.7 Execute new business strategies



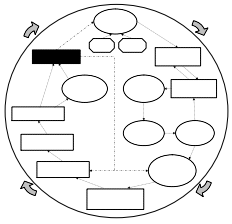
The first three new *Operating strategies* of the business are executed automatically every month through the operationalised data mining project diagram. In the case of Telco ABC, the first involvement of the business function in *Execute new business strategies* starts with new *Operating strategy* nr. 4.

The first thing the Retention Manager will do in executing that business strategy, will be to develop the comparative segment profiles of the four base features, using our sorted quantitative segment profile. The visualisation of those segment profiles we provided during *Deploy outputs into business* will assist in this task. Following that, he will profile root cause effect in the segment of choice, in a way, which makes best use of available resources for addressing root cause, as we saw possible in *Develop circumstantial knowledge.*

Following that, he will design the targeted campaign offer for each segment, and prioritise the execution resources according to the information we provided for this. For a first-time execution of the campaigns through the in-house call center, the call center staff may need to be trained in the logic of the new approach. Particular attention may be given to the conditions of discretionary negotiation, which may be allowed for about some offer dimensions. We gave examples earlier of such dimensions of negotiation, and their conditions.

Execution further will require the pre-emptive inventory management of targeted replacement handsets, so that there are no delays in deliveries after offers have been made to potential churners. The Handset Type quantitative profiles will assist with such inventory management, by converting the percentage figures there, back into quantities of handsets.

Finally, the call center staff will make the pre-emptive phone calls to the targeted consumers, and making the retention offer and assuring a positive outcome. Alternatively, offers can be mailed out with the bill of each targeted consumer.



7.8 Monitor and control

The data mining strategy we developed for monitoring and

control was to:

- identify the key concepts upon which the ongoing relevance of the new business solution depends, and develop the *Monitor and control* plan.

Here we present the monitor and control plan we developed. It is concerned with monitoring key concepts about:

- the understanding about the nature of the problem;
- the business solution for the problem; and
- the business success measures about the project.

The plan consists of objectives (the *what*), strategies (the *how*), a schedule (the *when*), and measurement criteria (by *how much*) which trigger any required control. We present the plan in tabled format in Tables 7.4, 7.5 and 7.6. Following the tables, we visualise some of the control results from our five data sets.

7.8.1 Problem understanding

The key concept about understanding the nature of the problem was the segmented profile of the root cause. We present the monitor and control plan in Table 7.5

Monitor and control problem understanding			
Objective	**Strategy**	**Timing**	**Measurement criterion**
Monitor Handset Type as root cause	When LogReg is rebuilt, check the effect scores	Every time Logreg is rebuilt for either predictive accuracy or because there are business indicators that root cause may have changed	Relative importance of Handset Type in model output, compared to other model effects
Control	Modify the problem understanding, and the business solution, and their profiling, to reflect the new root cause	When Handset Type has been superceded in the effect scores by a new effect	Updated hypotheses, and quantative and visual profiles

Table 7.5 Monitor and control problem understanding

7.8.2 Business solution relevance

Monitor and control business solution			
Objective	**Strategy**	**Timing**	**Measurement criterion**
Monitor classifier accuracy	Hold out a 5% sample from the targeted consumers, and don't make them a retention offer, so that they become actual churners. Apply the classifier to the consumer database, and obtain its accuracy on these actual churners	- Hold-out before the launch of the retention campaigns - Scoring on the fourth working day of the next financial month	Lift of 5 times
Control	Rebuild the classifier	When the lift falls below the required minimum level	Lift of at least five times, and confidence of at least 95%
Monitor confidence in segmentation	Check the R^2 and the Fisher's Criterion	Every month after having completed the clustering	R^2 of at least 0.70 and Fisher's Criterion of at least 1.5
Control	Rebuild the segmentation	When R^2 or the Fisher's criterion falls below the minimum level	R^2 of at least 0.70 and Fisher's Criterion of at least 1.5
Monitor required segmented campaign effort	Check the number of consumers in each retention segment	Every month after having completed the clustering	The number as a percentage of those falling outside the risk tolarance
Control	Proactive allocation of retention campaign resources	Monthly before the campaigns are launched	Call centre time, replacement handsets etc.
Monitor number of nearest segments	Check the number of nearest segments in the segmenting results	Every month after having completed the clustering	More than half the number of segments created

Objective	Strategy	Timing	Measurement criterion
Control	Rebuild the segmentation	When the number of nearest segments falls below the minimum level	More than half the number of segments created
Monitor relative feature importance	Check that there is a spread of effort between the base features *Value*, *Demand function*, *Event likelihood,* and *Intention to positively respond*	Every month after having completed the clustering	You do not want one or two of the four features to dominate the profile
Control	Change the feature standardisation and value replacement before segmentation	When less than three features show significant effort	You do not want one or two of the four features to dominate the profile

Table 7.6: Monitor and control business solution relevance

7.8.3 Project return on investment

Monitor and control project ROI			
Objective	**Strategy**	**Timing**	**Measurement criterion**
Monitor campaign response rates	Call centre operators to record responses, and managers to consolidate those per segment	As a month-end activity	Industry standards between 35% and 50%. Organisation should develop its own measure over time
Control	Change the campaign offers or rebuild the classifier	When the response rates fall below the industry range	As above
Monitor costs associated with campaign design and execution	Call centre and Retention Manager collaborate on costing	As a month-end activity	Compare to industry standard rates, and develop own standard over time
Control	Modify the campaign offers, or change the execution strategy	When costs fall outside of cost standards	As above
Monitor ROI from the project	Use formula 1, costs, ARPU etc. to calculate it	As a month-end activity	Compare to industry standards, and develop own standard over time
Control	Any of the above. If that fails, identify another project with a better ROI	When ROI falls outside of standard	As above

Table 7.7: Monitor project ROI

7.8.4 Visualising the monitoring

Visualising the monitoring results assists with their efficient interpretation. In figure 7.11, we visualise the change in the number of nearest segments over five months. Note from the Y-axis, how as time progresses (X-axis), there is an increase in the number of nearest segments. This means the resolution of our discovered information, and the knowledge we developed from it, has improved with the passage of time:

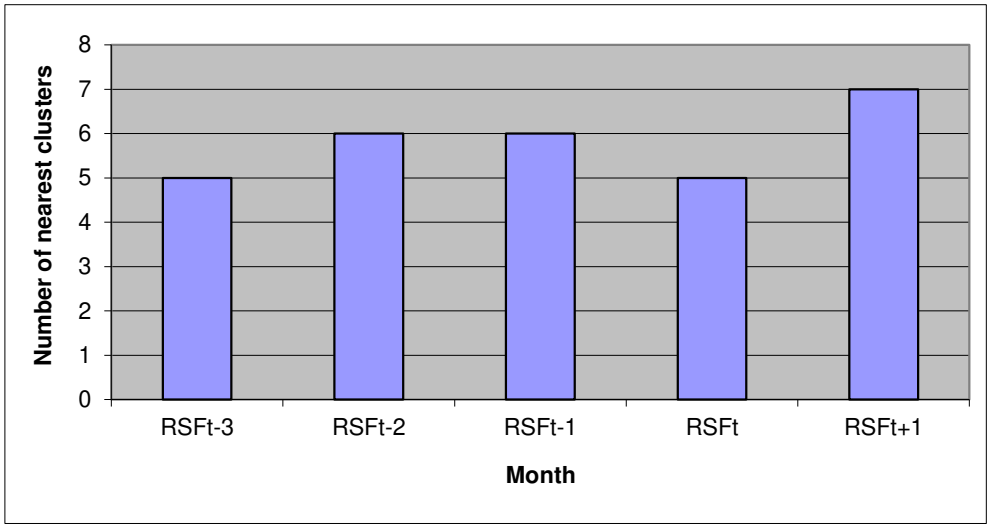


Figure 7.11: Change in number of nearest clusters

In Figure 7.12, we visualised the relative the change in relative base feature importance, in creating the segment structure over time:

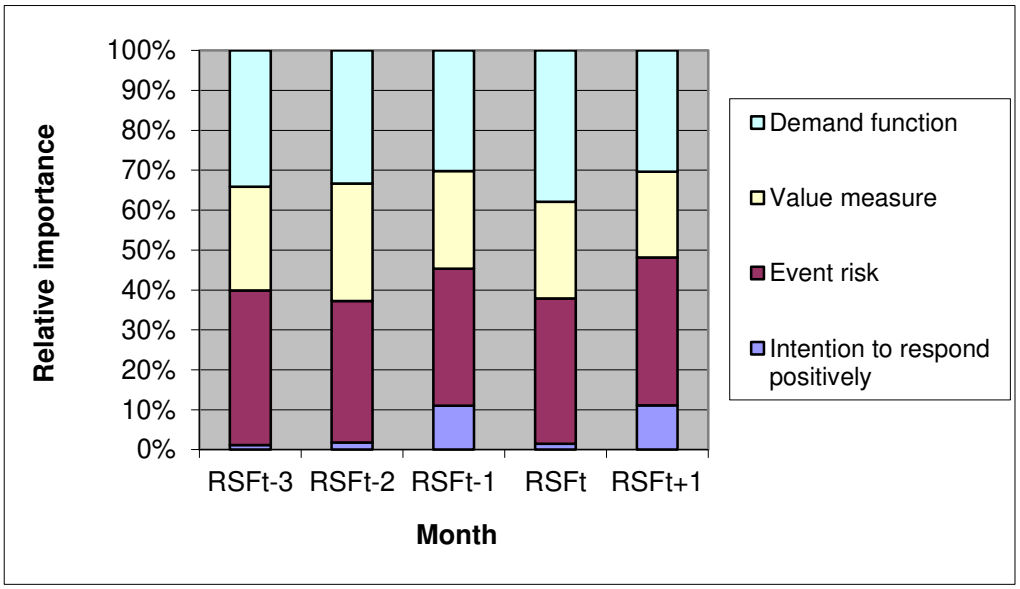


Figure 7.12: Change in relative feature importance over time

Notice how the relative importance of three of the features remains consistent over time. We would experiment with improving the stability of *Intention to respond positively*, if that does not stabilise in another three months. Alternatively, discovering the reason for that fluctuation could be the subject of a data mining project on its own.

In Figure 7.13 we have visualised the change our confidence measure, captured R^2, over time:

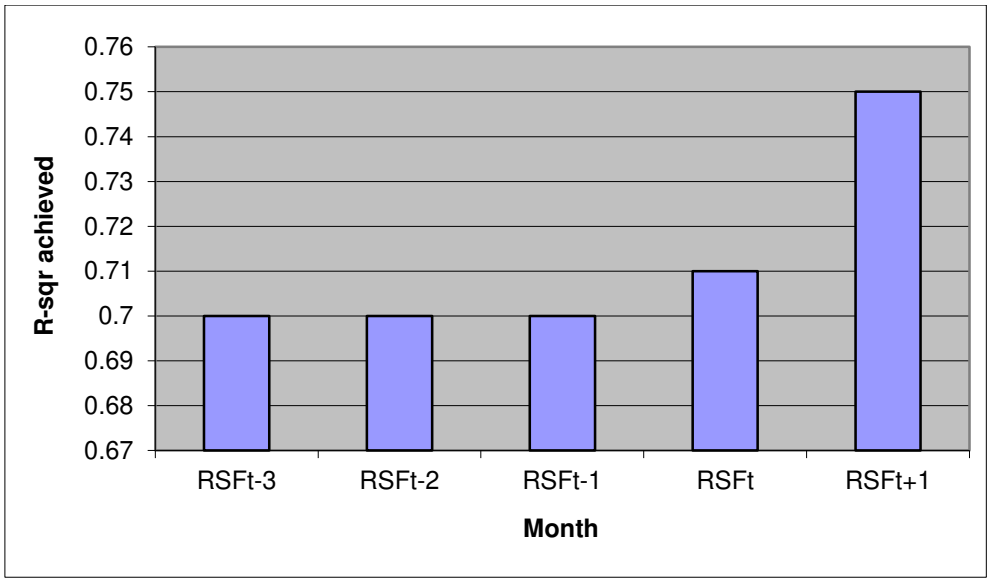


Figure 7.13: Change in captured R2 over time

We are pleased to note from the height of the bars against the values in the Y-axis, that our confidence in our segment structure increased over time lapsed.

We next visualised in Figure 7.14 the overall drift in segment profiles in five months:

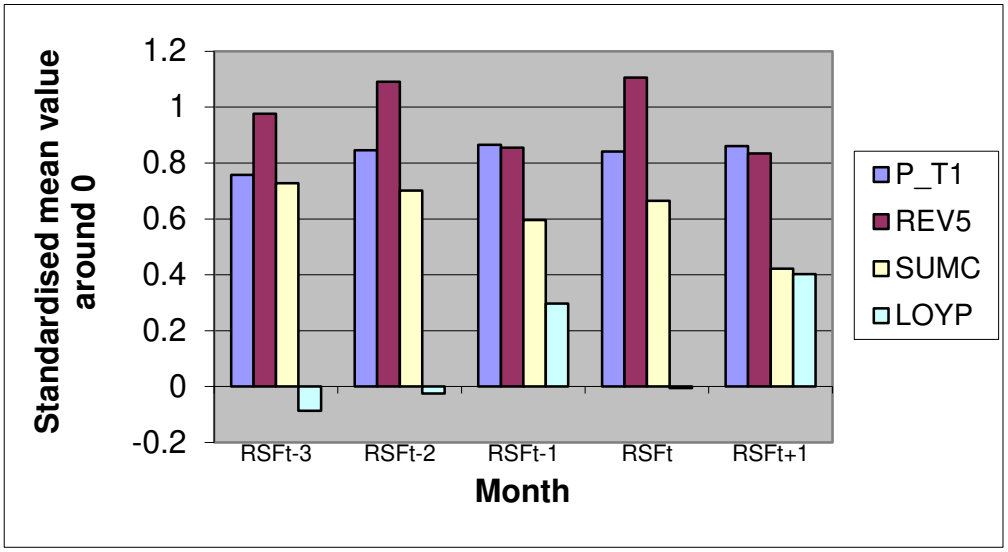


Figure 7.14: Drift in overall segment profiles

The zero line on the Y-axis, is the overall standardised mean of all consumers in a month. Figure 7.14 tells us that those consumers, who fall outside the 1.5% risk tolerance range, consistently are of above average *Demand function* and *Value*. This is evident from the above-zero values of *SUMC* and *REV3*. Further, there is a growing trend over the months, for loyal consumers to fall into this high-risk category. That is evident from the monthly increase in the value of *LOYP* against the Y-axis.

We next visualise in Figure 7.15, the change in overall Fisher's Criterion, or inter-segment distinctiveness:

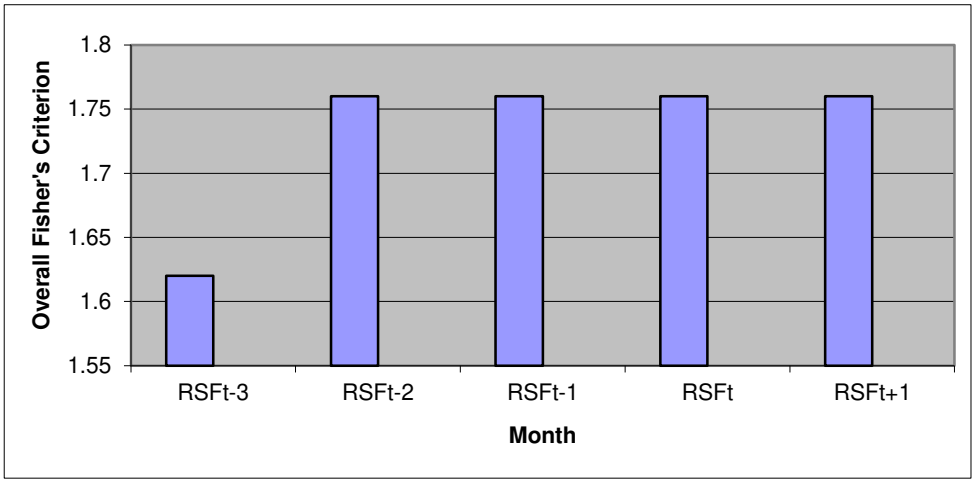


Figure 7.15: Change in overall inter-segment distinctiveness

We notice that in the first month the measure is out of tilt with the other months. We may attribute this to quirks in the RSF_{t-3} data, since that was the first version of the mining table. However, if these values fluctuate, we would investigate to find a cause.

In Figure 7.16, we visualise the change in the percentage of consumers within segments, over time lapsed:

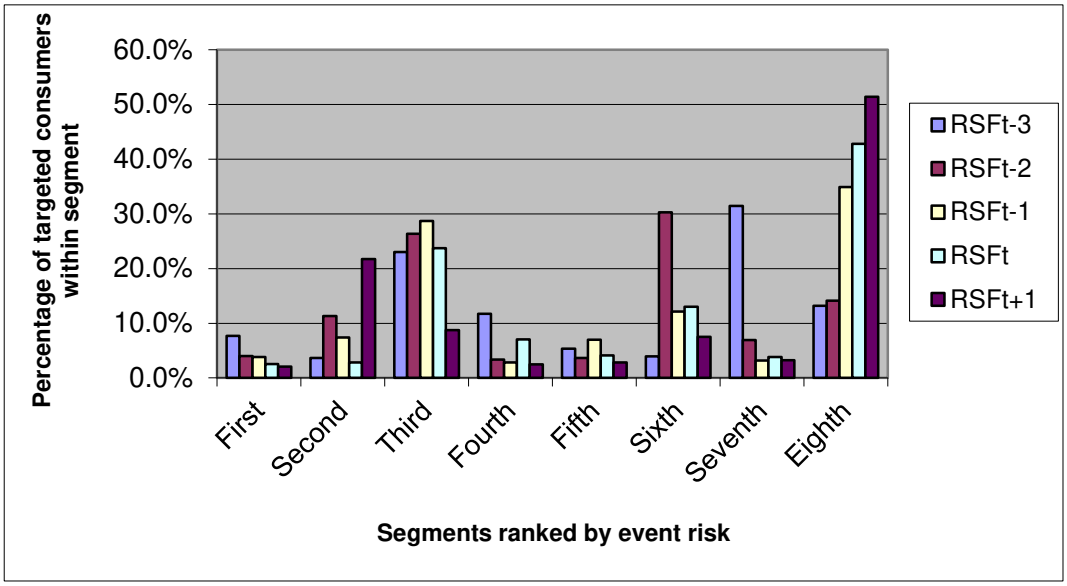


Figure 7.16: Change in percentage of consumers within a segment

Recall how the allocation of campaign execution resources was based on these figures. We see from Figure 7.16, that the eighth and third segments tend to consume most resources every month. Note however, how the sixth segment was an exception in RSF_{t-2}, and the seventh segment in month RSF_{t-3}.

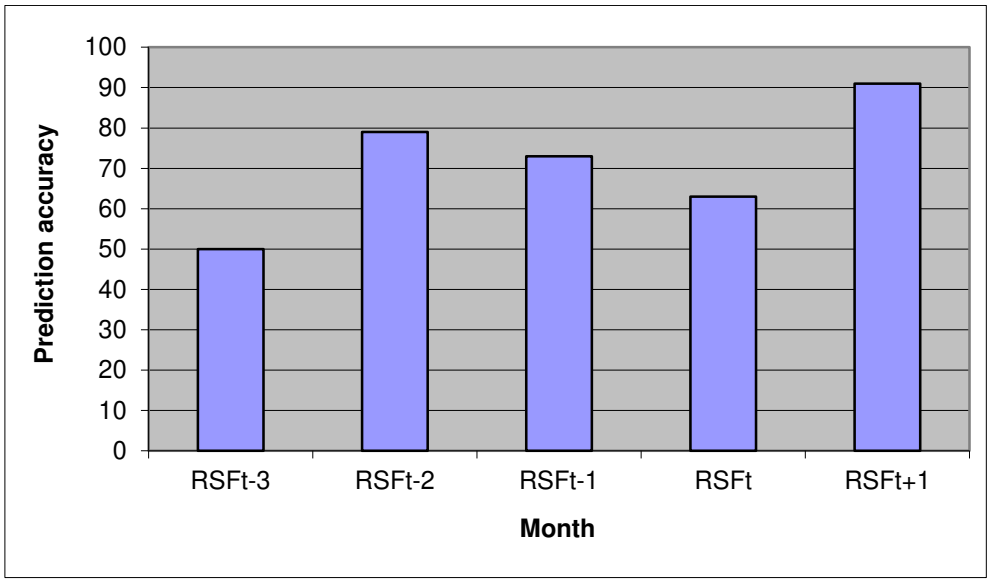


Figure 7.17: Classifier predictive accuracy over time

In Figure 7.17, we visualised the monitored results of the classifier's accuracy over time. We see that in the first month – RSF_{t-3} – the results were marginal. We may attribute this to the fact that this was the first mining table, and there may be some discrepancies in data values. The accuracy then increases to an acceptable 80% in RSF_{t-}

$_2$, after which drops off over the following two months. In RSF_{t+1}, the accuracy again increases. We could be witnessing a three-month cycle from RSF_{t-2} to RSF_t. In that case, the previous cycle ended in RSF_{t-3}, and the next cycle started in RSF_{t+1}. If there is such a cycle, then the classifier's predictive accuracy may gradually increase over time, until the concept starts to drift and accuracy decreases. The cause may be a trend within data we use in our analysis, which for some reason is only updated every three months to the systems where we extract our data.

This concludes the execution of SAM on Telco ABC's retention management problem data. In the following chapter, we offer our conclusions, future directions for further research, and comments.

7.9 Chapter summary

In this chapter we pursued two outstanding business deliverables for Telco ABC, as well as the validation of SAM from *Develop circumstantial knowledge* onward. This summary follows this two-pronged approach.

The two outstanding business deliverables were SPC goals 4 and 5:

- SPC 4 - was to *develop a new retention management solution, which will be objectively optimal, novel and executable, given Telco ABC's circumstances*;
- SPC 5 - was to *use data mining to support the deployment of the new retention solution on an operational basis*.

In *Develop circumstantial knowledge* we developed the executibility requirements of the retention management solution in the three classical marketing dimensions of Segment-Target-Position. We decided to only target the 1.5% most at-risk customers, to segment them by behavioral and value criteria, and to profile them in the product dimension for root cause. We then developed the positioning, as campaign responses for each segment consisting of product and service components. Using behavioral and value criteria we also prioritised time and money resources for campaign execution.

In *Strategic analysis* we calculated a incremental return of the solution over Telco ABC's existing solution in the range of AUD14.4 m. to AUD27.75 m. We calculated the costs associated with the project, and a project ROI of about 187 times the investment. These economics led to a *Strategic choice* that the solution presents a very attractive justification to continue with the project into the *Define* and *Realise* phases.

In *Define new business objectives and strategies* we defined the solution as new retention management objectives and strategies for Telco ABC. The objectives and strategies formalised and communicated the targeting, segmenting and positioning components of the solution. We compared them with Telco ABC's retention management objectives and strategies at the outset of the project, and defined the strategic and operational paradigm shift brought about by the project. At this point we attained the SPC 4 goal.

In *Develop data mining plan* we started to pursue the SPC 5 goal. We applied the mapping technique, and formulated the data mining objectives for supporting the new retention solution. We successful developed the data mining objectives into comprehensive data mining strategies for operationalising the data preparation and algorithms in campaign cycles, and how we would deploy the results to the marketing and CRM systems as reports and campaign lists.

We subsequently optimised in *Model, evaluate, choose best model(s)* the models we had produced in *Data mining discovery*, refining the effectiveness of our analytics results. We also executed the operationalising strategies in *Operationalise model(s)* and the strategies for results deployment in *Deploy results into business.* We commented on particular issues pertaining to Telco ABC's situation in *Execute new business strategies*, and pointed out that this activity is already well developed in the business literature.

In *Monitor and control* we execute the last step toward attaining the SPC 5 goal. Here we identified the key features of the business problem, the new solution, and the project returns, which require monitoring for effectiveness and efficiency. We also formulated control parameters for the analytics, and strategies for controlling within those parameters. The monitoring of the business solution – the campaign results - are covered, as this fell outside the scope of our research.

We now summarise the remaining validating of SAM in the Telco ABC environment. From the results obtained for Telco ABC in *Develop circumstantial knowledge, Strategic analysis,* and *Strategic choice* we supported the view that SAM offered utility in developing the detailed executibility requirements of a competitive solution. It achieves this through developing and testing the breakthrough hypotheses, which underlie the business deliverables, using the newly discovered information.

The results obtained in *Develop new business objectives and strategies* proved that SAM has utility in taking the executibility requirements, and formulating those into a new business solution. SAM supports expressing that solution in new business objectives and strategies, and the legitimisation and sharing of that solution throughout the organisation. It also offers utility in defining strategic and operational paradigm shift, which facilitates reporting at company board level.

From *Develop data mining plan*, *Model, evaluate, choose best model(s),* and *Operationalise model(s)*, we proved that SAM has potency in taking the new business decision, and formulating an operationalisable, supportive analytics plan for the SPC environment. This validates the mapping technique which we developed, and supports the view that a plan can only be formulated after uncertainty has been qualified and quantified.

The monitor and control plan we formulated using *Monitor and control,* established the utility of this task set in SAM for developing an effective monitor and control plan.

The results we achieved for Telco ABC in this section, particularly well demonstrated the visibility which SAM offers in the progress of the project toward realising the business deliverables.

8 CHAPTER 8. Conclusions and future research directions

8.1 Research contributions

We present our research contributions under the following five headings. Following these headings, we present ideas for future research, and then concluding remarks.

8.1.1 Research methods

The results obtained from this research have again proved the usefulness of the *action science research method*. Its usefulness is in providing robust validation for hypotheses in an industrial research setting, where the expense of classical experimentation would have been prohibitive.

Using *expert reflection-in-action* we designed a practical solution for a complex research problem, drawing on knowledge from disparate research domains. We therefore proved the potency of *expert reflection-in-action* as a knowledge synthesizing technique, where there is a diversity of considerations.

The hypothesis developing process in e*xpert reflection-in-action* is similar to the hypothesis developing knowledge management process at the heart of the Strategic Planning Model, which is embedded in SAM. We can therefore say that *expert reflection-in-action* is versatile and robust enough, to be of use in a solution targeted at an industrial audience.

8.1.2 Knowledge Discovery and Data Mining

8.1.1.1 Data mining project methodology

Our research found the commonly used data mining project methodology CRISP-DM deficient in producing competitive, executable solutions in the SPC environment. We therefore concluded that the state of existing data mining project methodology is one of the contributors to the slow uptake of data mining by the business community. In this conclusion we attained our first research goal.

We contributed a data mining methodology called Strategic Analytics Methodology (SAM). First, SAM is a significant research contribution because it builds on what was

useful in CRISP-DM, yet substantially *improves* the utility for delivering competitive, executable business solutions.

Second, SAM is significant because it is in a *useful format*, meeting the expressed need for *a codified, competent data mining strategy for the voluminous data environment, which formally collaborates what both the analyst and the technology is good at* (Liu 2003, pp.436, 443).

SAM is significant for a third reason; replacing CRISP-DM's data centric core with a business-centric core. This business core helps the business:

- evaluate the relevance of the discovered information to the business deliverables;
- defines the executibility dimensions of the business solution in terms of the new subject matter and the organizational circumstances;
- formulates the executibility dimensions into an executable business solution;
- contains activities which actually execute the solution; and
- monitors and controls the results of the solution over time.

All of these are things the *business is good at*. SAM therefore integrates what the *analyst, the technology, and the business* is good at, making it significant for the third reason.

The cost of the data preparation phase of projects is a major concern for business. SAM's business core also successfully reframed the data centric description of the data preparation stage of projects, into a business justification about reducing project risk, and progressively supporting the business deliverables.

We further contributed a new phasing of data mining projects, which offers improved visibility of the project's progress against the business deliverables.

Producing SAM attained our second research goal.

8.1.1.2 Technical

SAM also contributes significant knowledge for its more technical audience. First, we reframed for this audience the key concepts, principles and processes of *concept drift*, as an automated version of the knowledge management process. This will increase understanding of the working value of *concept drift*. This understanding will facilitate

the use of *concept drift* for monitoring and controlling by practitioners in the SPC environment.

Second, we proved how to import key concepts, practice and process from *concept drift* into the unstructured, less automated SPC environment, and use them for there for developing substance in monitoring and controlling. *Concept drift* helps with identifying the key components of both the business and analytics solutions which require monitoring, and helps with developing control strategies for those components.

Third, we provided a tool with technique for evaluating discovered information for relevance to business deliverables, and how to convert that information into useful knowledge. We used a knowledge management process, which combines collaborative *expert reflection-in-action* with technique for considering the impact of circumstantial knowledge. This further identified for the technical audience the critical soft skills they need to master, in order to facilitate the uptake of the technology in the SPC environment.

Fourth, we contributed an algorithm called *PROMIX*. This algorithm is provides a generic, industry-neutral recipe for analytics in the retention management situations, which supports the Segment-Target-Position sequence of classical marketing. It offers parameters which can be tuned in accordance with the industry's business drivers (e.g. product, channel, customer), and the business's particular situation.

8.1.3 Business intelligence practitioners

Our research also made useful contributions for the analysts and business subject matter experts who operate in the SPC environment. They may be employees of the organisation depending on the business deliverables, or advisors working for professional services organisations.

We identified important theoretical and practical analytics issues in the SPC analytics environment, and which require attention as inhibitors or enablers. Further, we contributed a practical roadmap about how to go about that attending. This will foster confidence in both communities to experiment with using data mining on SPC projects, ultimately promoting the uptake of the approach.

We further contributed, that the breakthrough and executibility which leads to competitive advantage, do not come from introducing a new technology into the

organisation. Rather, they come from introducing *new business subject matter* into the organisation, and using the new technology to support the development and execution of a solution which embodies that new business subject matter.

In SAM, we contributed a business advisory tool, with potency and technique, for departing from an organisation's limiting *status quo,* and delivering a competitive, executable solution, which has broken through those limitations. SAM supports the value proposition of professional business advisors, that they contribute incremental value over what is possible by the organisation, using internal resources only.

Our research further contributed to this audience, an advisory methodology for covering the whole spectrum of business questions asked in the SPC environment; from confirming the existence and extend of a problem, to understanding its key drivers, to developing a solution for it, to executing that solution, to monitoring and controlling the effect of the solution.

8.1.4 Business intelligence software vendors

We made two contributions for this audience. First, that similar knowledge management and technology alignment issues, as those experienced in the previous decade by the ERP solution industry and by the CRM solution industry until recently, are restricting the uptake of data mining technology in the SPC environment. We identified for data mining what those issues are.

Those industries overcame their limiting factors first through improving their project methodologies for aligning the technology to the business needs. Second, they overcame with programs which managed the impact of new subject matter and the technology on the organisation's knowledge. Similarly, we contributed a solution for overcoming the limitation of those factors on the uptake of data mining. However, due to the greater requirement for knowledge *exploration* in the SPC environment, the comparatively *greater impact* of data mining on the organizational knowledge, and the *higher frequency of projects* in the SPC environment, we embedded the knowledge management into the project methodology, to a greater extent than in the other industries.

Third, with SAM we have contributed a communication framework and tool, with which to explain and promote the strategic value of data mining to potential buyers of their software.

8.1.5 Telco industry and our industrial research partner

Using SAM we first developed quantified insights about the existence and extent of the problem. We determined that their problem was not out of hand, to the extent that they had thought, which encouraged them to continue with retention management activities. Second, we formulated a competitive, executable business solution, and an analytics plan for supporting the operationalising of the solution, and for maintaining the effectiveness and efficiency of the solution over time.

Third, we quantified their incremental benefit of the new solution over their existing solution upward of AUD14 m. These three things validated SAM in an industrial environment, attaining our third research goal.

We contributed for the Telco industry a practical roadmap for addressing a pressing competitive problem. It contributed to the industry valid ways to:

- overcome the limiting effect of the *status quo* on the breakthrough and execution potential from the project;
- introduce new subject matter for formulating competitive business deliverables for their problem;
- discover situational information with which to mold the new subject matter for competitiveness and executibility under their organisational circumstances;
- define the executibility components of the knowledge, formulate them into a business solution, and share that solution as a new strategic framework within the organisation;
- manage soft issues and legitimise the new solution through knowledge management;
- operationalise the analytics;
- use analytics to monitor and control key components of the problem and its business solution.

8.2 Ideas for future research

Our research made a significant contribution to a novel area of research in the Data Mining, Knowledge Discovery, and Business Strategy literature. However, we also recognise the limitations of our contributions. We have therefore formulated a number

of ideas for future research work - either by ourselves or by other researchers – which either emanate from the limitations of our research results, or from the new frontiers opened up by it:

- SAM requires validation in other industries than Telco. We perceive good opportunity for its immanent trialing in the Financial Services Industry;
- we also recommend that SAM be validated in environments, which are either more structured, or more unstructured, than the corporate SPC environment. An example of the first is the classical automated environment of data mining. An example of the second is physics, where there is an analysis opportunity for discovering hitherto elusive matter;
- we see an opportunity for improving the degree of validation supporting our work. SAM needs to be tested in an experimentally optimised, industrial research environment, where the its incremental contribution to the business solution is quantified;
- we think that we have not exhausted the utility components required by the SPC data mining environment. Research is required to confirm the criteria we identified, and perhaps to add to them;
- our work needs to be complemented with a software project plan, perhaps in MS Project format. This would be a useful tool for scoping out internal or professionally advised projects, for their time and cost dimensions. Having calculated the cost components of an SPC data mining project in this diligent way, will build confidence in the ROI calculations of the project, further facilitating the uptake of the technology. A project plan with reporting facility will further improve visibility into project progress, significantly reducing the actual and perceived risk associated with projects of this nature. Reducing risk in this way will increase the confidence of the target audiences in the approach;
- once a project plan has been developed, that should be complemented by in-depth change management activities for the project. This will practically reduce the risk to organisations of project failure, again increasing the confidence in taking up the technology and its approach;

- we perceive an opportunity for PROMIX to be tested in retention management problems in other industries; further
- we identify a need for composing generic algorithms for other customer lifetime management strategies apart from retention management, like up-sell, cross-sell, customer migration, time to event etc.;
- there is a need for applying the monitor and control plan to Telco ABC's execution of the business solution. This fell outside the scope of our research, would complement our research, and further benefit Telco ABC.

8.3 Concluding remarks

Both data mining and business are complex disciplines, each with their own artful and technical components. The expertise associated with each discipline, seems to have been contained within the discipline. Because of this containment, the competitive fusion of the two disciplines has been mystifying experts from both domains for some time.

We *trust* that in developing our rigorous multi-disciplinary methodology called SAM, our research has now demystified that fusion. We *hope* that this may gradually lead to confidence among business decision-makers, about the competitive potential presented by the data mining approach. We *believe* that such confidence would result in the championing of the data mining approach at the Boardroom level

9 Appendix A: Terms, abbreviations, acronyms

Term, abbreviation, acronym	Explanation
BI	Business Intelligence – the decision-making process responsible for the competitiveness and executibility of business solutions. Also see SPC below.
CRM	Customer Relationship Management - An approach to managing evolving relationships with customers over time. Also refers to software tools.
Diagnostic technique	Technique which us used for breaking a problem down into its various components, determines their roles, interactions and impact, then prioritises them for attention, and formulates the approach to overcoming the problem
Competitive advantage	Having a more effective and efficient solution for an industry problem, than what your competitors have (see effectiveness and efficiency)
Concept drift	Technical process, principles and practice used in the automated data mining environment, for monitoring and controlling problems and their solutions
Effectiveness	The degree to which you attain an objective
Efficiency	The quantity of resources you consume in the process of being effective
ERP	Enterprise Resource Planning – A systemised approach to optimising business processes, supported by software

Term, abbreviation, acronym	Explanation
KM	Knowledge Management – the process of developing hypotheses, testing them, adapting them for organisational circumstances, and selecting the best hypotheses, developing the components of their executibility, and defining those into a business solution in terms of objectives and strategies
Mapping technique	A technique which converts what the business needs into what a supporting technology needs to do toward meeting those needs
LogReg	Logistic Regression model
Paradigm	A pervasive way of thinking within an organisation. More formally, it is the organisation's existing knowledge structure, expressed by the objectives and strategies in its plans
Paradigm lock	A situation where an organisation is unable to change their knowledge content and/or structure
Paradigm shift	A deliberate, engineered change in the paradigm, expressed in updated objectives and strategies
Retention management	*…(sustaining) the existing profit stream from current customers, in effect decreasing the expected trends of customer defection.* (Lenskold 2003, p.3)
Schema	An unchallenged, untested collection of preconceptions and expectations about a project
Soft issues	Relating to non-technical factors impacting on projects and SPC, and which need to be proactively managed from threatening the supporting an outcome

Term, abbreviation, acronym	Explanation
SPC	Strategic Planning Cycle – the business process which unites BI with the execution of the solutions and their maintenance over time. Also see SPM
SPM	Strategic Planning Model – a process and tool used for driving and focusing the SPC
Strategic creep	Strategic creep – gradual and spontaneous changes to where organisational goals over time, often without the changes having been formally communicated to the analysts
Structured problem	A problem where its boundaries and components are known with certainty
Subject matter	A domain of application e.g. marketing or finance or analytics
Unstructured problem	A problem whose boundaries and components are not know with certainty
Voluntary churn	Used in the telecommunications industry to describe the apparent unexpected decision of an existing customer to discontinue his / her use of a telecommunication company's services, and to start using the services of a competing telecommunication company

10 Bibliography

Agrawal, R. and G. Psaila (1995). Active Data Mining. Proceedings of KDD95: First International Conference on Knowledge Discovery and Data Mining, Menlo Park, CA, AAAI Press: 3-8.

Aha, D. W., D. Kibler, et al. (1991). "Instance based learning." Machine Learning **6**(1): 37-66.

Alchourron, C., P. Gardenfors, et al. (1985). "On the logic of theory change: partial meet contraction and revision functions." Journal of Symbolic Logic **50**: 510-530.

Apley, D. W. (2003). Principal Components and Factor Analysis. The Handbook of Data Mining. N. Ye. London, Lawrence Erlbaum Associates**:** 193-212.

Apte, C., R. Bibelnieks, et al. (2001). Segmentation-Based Modeling for Advanced Targeted Marketing. Yorktown Heights, IBM Research Division, Thomas J. Watson Research Center, NY.

Badgett, M., W. Connor, et al. (2003). Driving an operational model that integrates customer segmentation with customer management. Somers, NY, IBM Global Services.

Bartlett, P. L., Ben-David Shai, et al. (2000). "Learning Changing Concepts by Exploiting the Structure of Change." Machine Learning **41**: 153-174.

Bauer, E. and R. Kohavi (1999). "An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants." Machine Learning **36**(1 / 2): 105-139.

Berson, A., S. Smith, et al. (2000). Building Data Mining Applications for CRM. New York, McGraw-Hill.

Berthold, M. and D. J. Hand, Eds. (2003). Intelligent Data Analysis - An Introduction, Springer.

Best, R. J. (2000). Market-based management : strategies for growing customer value and profitability. Upper Saddle River, Prentice Hall.

Bishop, C. M. (1995). Neural Networks for Pattern Recognition. Oxford, Oxford University Press.

Borror, C. M. (2003). Statistical Analysis of Normal and Abnormal Data. The Handbook of Data Mining. N. Ye. London, Lawrence Erlbaum Associates**:** 67-102.

Boyce, G. and B. Blair (2003). Accounting Information Systems. Sydney, McGraw-Hill.

Breiman, L. (2001). Random Forests. Berkeley, Statistics Department, University of California.

Breiman, L., J. H. Friedman, et al. (1984). Classification and regression trees. Belmont, CA, Wadsworth International.

Bryman, A. and E. Bell (2003). Business Research Methods. Oxford, Oxford University Press.

Chakrabarti, S., S. Sarawagi, et al. (1998). Mining surprising patterns using temporal description length. Twenty Fourth International Conference on Very Large Data Bases, New York, Morgan Kaufman: 606-617.

Chapman, P., J. Clinton, et al. (1999-2000). CRISP-DM 1.0: Cross Industry Standard Process for Data Mining. http://www.crisp-dm.org/CRISPWP-0800.pdf, CRISP-DM Consortium. **Accessed November 2003**.

Chatfield, C. S. and T. D. Johnson (2000). Step by Step Microsoft Project 2000. Redmond, Washington, Microsoft Press.

Checkland, P. (1981). Systems Thinking, Systems Practice. London, John Wiley & Sons.

Chen, C.-J. (2004). "The effects of knowledge attribute, alliance characteristics, and absorptive capacity on knowledge transfer performance." R&D Management **34**(June 2004): 311-321.

Cohen, W. M. and D. A. Levinthal (1990). "Absorptive Capacity: A New Perspective on Learning and Innovation." Administrative Science Quarterly **35**: 128-152.

Compton, J. (2001). Web Extra: Gartner's Five Categories of Personalization. www.destinationcrm.com/articles/default.asp?ArticleID=603, CRM Media. **Accessed November 2001**.

Connor-Linton, J. (2003). Chi Square Tutorial. http://www.georgetown.edu/faculty/ballc/webtools/web_chi_tut.html, Georgetown University Linguistics Department. **Accessed November 2003**.

Cooley, R. (2003). Mining Customer Relationship Management (CRM) Data. The Handbook of Data Mining. N. Ye. London, Lawrence Erlbaum Associates Inc.**:** 597-616.

Costa Dr., P. (2001). CRM Analytics for Telecommunications: the WAR Framework. http://bi.snu.ac.kr/KDMS/conference2002/Telco_CRM_Analytics.pdf, IBM Global Service. **Accessed November 2003**.

DataPlus Millennium (2001). RFM-A: Show me the Average! http://66.102.7.104/search?q=cache:LFI1VhKxBf0J:www.marketinganalyticsgroup.com/Library/White%2520Paper%2520-%2520Show%2520Me%2520The%2520Average.pdf++%22rfm-a:+show+me+the+average%22&hl=en, Marketing Analytics Group. **Accessed April 2003**.

Date, C. J. (2000). An Introduction to Database Systems. Reading, Massachusetts, Addison-Wesley Publishing Company.

Denzin, N. K. and Y. S. Lincoln (2003). Strategies of qualitative inquiry. Thousand Oaks, CA, Sage.

Diekhoff, G. (1992). Statistics for the social & behavioral sciences: univariate, bivariate, multivariate. Midwestern State University, Wm. C. Brown Publishers.

Dobler, D. W., D. N. Burt, et al. (1990). Purchasing and Materials Management: Text and Cases. New York, McGraw-Hill International Edition.

Domingos, P. (1997). Context-Sensitive Feature Selection for Lazy Learners. AI Review. **11:** 227-253.

Emagine (2003). Customer Retention NPV Model. www.emagine-intl.com, Emagine. **Accessed October 2003**.

Engel, J. F., R. D. Blackwell, et al. (1995). Consumer Behavior. Orlando, The Dryden Press.

Fayyad, U. (2004). "Editorial." SIKDD Explorations **5**(2).

Fayyad, U., G. Piatetsky-Shapiro, et al. (1996). "From Data Mining to Knowledge Discovery in Databases." AI Magazine **17(3)**(Fall 1996): 37-54.

Ferran-Urdaneta, C. (1999). Teams or Communities? Organizational Structures for Knowledge Management. Proceedings of the 1999 ACM SIGCPR Conference on Computer Personnel Research, New Orleans, ACM Press: 128-134.

Freund, Y. and Y. Mansour (1997). Learning under persistent drift. Computational learning theory: Third European Conference:

Ganti, V., J. Gehrke, et al. (1999). "Mining Very Large Databases." IEEE Computer Society Journal **32**(38): 38-45.

Gehrke, J. (2003). Decision Trees. The Handbook of Data Mining. N. Ye. London, Laurence Erlbaum Associates, Inc.**:** 3-24.

GhostMiner (2002). Customer Segmentation and Database Marketing in Financial Services. www.fqspl.com.pl. **Accessed February 2004**.

Gibson, J. L., J. M. Ivanchevich, et al. (1991). Organisations: Behavior, Structure, Processes. Boston, Irwin.

Goleman, D. (1998). Vital Lies, Simple Truths - the Psychology of Self-Deception. London, Bloomsbury.

Goncalves, K. P. (1998). Services Marketing: a strategic approach. Upper Saddle River, NJ, Prentice-Hall.

Gorman, G. E. and P. Clayton (2003). Qualitative research for the information professional: a practical handbook. London, Facet.

Grant, R. M. (1996). "Prospering in Dynamically-Competitive Environments: Organizational Capability as Knowledge Creation." Organization Science **7**(4): 375-387.

Griffin, J. (2003). Customer Segmentation: Divide and Prosper. iQ Magazine. www.cisco.com, Griffen Group.

Grossman, R., M. Hornick, et al. (2003). Emerging Standards and Interfaces. The Handbook of Data Mining. N. Ye. London, Lawrence Erlbaum Associates**:** 453-459.

Groth, R. (1998). Data Mining: A Hands-On Approach for Business Professionals. Upper Saddle River, NJ, Prentice Hall PTR.

Gupta, A. and G. Vijay (2000). "Knowledge Flows within MNCs." Strategic Management Journal **21**: 473-496.

Hammersby, M. (2004). Some reflections on ethnography and validity. Social research methods: a reader. C. Seale. London, Routledge**:** 241-245.

Han, J. (2004). Data Mining: An Introduction. The Eighth Pacific-Asia Conference on Knowledge Discovery and Data Mining School. S. Zhang. Sydney, University of Technology, Sydney.

Han, J. and M. Kamber (2001). Data Mining: Concepts and Techniques. San Francisco, Morgan Kauffman Publishers.

Harries, M. and K. Horn (1995). Detecting Concept Drift in Financial Time Series Prediction using Symbolic Machine Learning. Proceedings of the Eighth Australian Joint Conference on Artificial Intelligence, Singapore, World Scientific: 91-98.

Harries, M. and K. Horn (1996). Learning stable concepts in domains with hidden changes in context. 13th International Conference on Machine Learning, Bari, Italy: 106-122.

Harries, M. and C. Sammut (1998). "Extracting Hidden Context." Machine Learning **32**: 101-126.

Hastie, T., R. Tibshirani, et al. (2001). The Elements of Statistical Learning. New York, Heidelberg, Berlin, Springer-Verlag.

Heckel, M. (2003). Texture analysis via Data Mining. Australian Data Mining Workshop, Canberra, Australia, University of Technology Sydney:

Helmhold, D. P. and P. M. Long (1994). "Tracking Drifting Concepts By Minimizing Disagreements." Machine Learning **14**(1994): 27-45.

Huan, L., L. Yu, et al. (2003). Feature Extraction, Selection, and Construction. The Handbook of Data Mining. N. Ye. London, Lawrence Erlbaum Associates, Inc**:** 409-422.

Hulten, G., L. Spencer, et al. (2001). Mining Time-Changing Data Streams. KDD-01, San Francisco, CA:

Jacobs, N. and H. Blockeel (2002). Sequence Prediction with Mixed Order Markov Chains, Department of Computer Science, University of Leuven.

Jacobs, R. and S. Nowlan (1991). "Adaptive Mixtures of Local Experts." Neural Computation **3**: 79-87.

Kargupta, H., A. Joshi, et al., Eds. (2005). Data Mining: Next Generation Challenges and Future Directions. Cambridge, AAAI Press / The MIT Press.

Kelly, M., D. J. Hand, et al. (1999). The Impact of Changing Populations on Classifier Performance. Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 367-371.

King, R. (2003). Data Mining and CRM. www.crm2day.com, Contact Solutions. **Accessed September 2003**.

Klapper-Rybicka, M., N. N. Schraudolph, et al. (2001). Unsupervised Learning in Recurrent Neural Networks. Technical Report, IDSIA (Instituto Dalle Molle di Studi sull'Intelligenza Artificiale). **17**.

Klinkerberg, R. (2001). Using Labeled and Unlabeled Data to Learn Drifting Concepts. Workshop notes of the IJCAI-01 Workshop on Learning from Temporal and Spatial Data, Menlo Park, CA, USA: 16-24.

Klinkerberg, R. and T. Joachims (2000). Detecting Concept Drift with Support Vector Machines. In Proceedings of the Seventeenth International Conference on Machine Learning, San Francisco, Morgan Kauffman:

Kogut, B. and U. Zander (1992). "Knowledge of the Firm, Combinative Capabilities, and the Replication of Technology." Organization Science **3**: 383-397.

Kohane, I. S. and I. J. Haimowitz (1993). Hypothesis-Driven Data Abstraction with Trend Templates. Symposium on Computer Applications in Medical Care 1993. http://citeseer.ist.psu.edu/kohane93hypothesisdriven.html. **Accessed February 2003**.

Kolyshkina, I., P. Petocz, et al. (2003). Modeling Insurance Risk: A Comparison of Data Mining and Logistic Regression Approaches. Congress on Evolutionary Computation, Canberra, Australia, University of Technology, Sydney: 227-237.

Kotler, P. (1988). Marketing Management: Analysis, Planning, Implementation, and Control. Englewood Cliffs, New Jersey, Prentice Hall, Inc.

Kotler, P. (2002). Marketing Management: Analysis, Planning, Implementation, and Control. International, Prentice Hall.

Krizakova, I. and M. Kubat (1992). "FAVORIT: Concept Formation with Ageing of Knowledge." Pattern Recognition Letters **13**: 19-25.

Kuh, A., T. Petsche, et al. (1991). "Learning time-varying concepts." Advances in Neural Information Processing Systems **3**: 183-189.

Kurz, D. L. and K. E. Clow (1998). Services Marketing. New York, J. Wiley & Sons.

Lanquillon, C. (1999). Information Filtering in Changing Domains. International Joint Conference on Artificial Intelligence (IJCAI99), Stockholm, Sweden, August 1999: 41-48.

Lau, R., A. H. M. ter Hofstede, et al. (2000). Applying Maxi-adjustment to Adaptive Information Filtering Agents. The 8th International Workshop on Non-Monotonic Reasoning (NMR) 2000, Belief Change, Queensland University of Technology:

Lenskold, J. (2003). Retention Marketing Profitability. ROI Challenges Influencing the Retention Versus Acquisition Debate. www.lenskold.com, Lenskold Group. **Accessed May 2003**.

Leonard-Barton, D. (1995). Wellsprings of Knowledge: Building and Sustaining the Sources of Innovation. Boston, Harvard Business School Press.

Levin, R. I. (1987). Statistics for Management. Englewood Cliffs, New Jersey, Prentice-Hall, Inc.

Levinson, M. (2000). Slices of Lives. http://www.cio.com.au/index.php/id;63068691;fp;512;fpid;1455200604, CIO Magazine. **Accessed September 2005**.

Levy, D. J. (2001). Segmentation: Key to Efficient CRM. http://www.dmreview.com/article_sub.cfm?articleId=3957, DM Review Magazine. **Accessed September 2003**.

Linden, A. (2000). Free Methodology and Process Model for Data Mining Released. www.gartner.com/resources/92900/92961/92961.pdf, Gartner. **Accessed May 2005**.

Liu, X. (2003). Systems and Applications. Intelligent Data Analysis. M. Berthold and D. J. Hand. Heidelberg, Springer-Verlag**:** 429-442.

Long, P. M. (1999). "The Complexity of Learning According to Two Models of a Drifting Environment." Machine Learning **37**: 337-354.

Lovelock, C. H. (2000). Services Marketing: people, technology, strategy. Upper Saddle River, New Jersey, Prentice-Hall.

Lucas, J. H. C. (2000). Information Technology for Management. Boston, Irwin McGraw-Hill.

Lunts, A. and V. Brailovskiy (1967). "Evaluation of attributes obtained in statistical decision rules." Engineering Cybernetics **3**: 98-109.

Mattison, R. (1999). Winning telco customers using marketing databases. Boston, Artech House.

Mazanec, J. A. and H. Strasser (2000). A Nonparametric Approach to Perceptions-Based Market Segmentation: Foundations. New York, Springer-Verlag.

McBurney, D. H. and T. L. White (2004). Research Methods. CB, Thomson Wadsworth.

Meltzer, M. (2000). E-Mining Myth and Magic. http://businessintelligence.ittoolbox.com/browse.asp?c=BIPeerPublishing&r=%2Fpub%2FMM092500.pdf, www.businessintelligence.ittoolbox.com. **Accessed February 2003**.

Michalski, R. S. (1987). How to learn imprecise concepts: A method employing a two-tiered knowledge representation for learning. Procedures of the 4th International Workshop on Machine Learning, Morgan Kauffman: 50-58.

Mozer, M., R. Wolniewicz, et al. (1999). Churn Reduction in the Wireless Industry. NIPS 1999, Denver, Colorado, http://www.informatik.uni-trier.de: 935-941.

Nonaka, I. (1994). "A Dynamic Theory of Organizational Knowledge Creation." Organization Science **5**(1): 14-37.

Nonaka, I. and H. Takeuchi (1995). The Knowledge-Creating Company : How Japanese Companies Create the Dynamics of Innovation. New Yord, Oxford University Press.

Nunamaker, J. (1999). "Collaborative Computing: The next Millennium." Computer **32**(9): 66-71.

Patrick, J. (2004). The Scamseek Project Text Mining for Financial Scams on the Internet. Proceedings of the 3rd Australian Data Mining Conference, Cairns, Australia, University of Technology Sydney: 33-38.

Pearce, I. J. A. and J. R. B. Robinson (1991). Strategic Management: Formulation, Implementation, and Control. Boston, Irwin.

Pearce, I. J. A. and J. R. B. Robinson (2004). Strategic Management: Formulation, Implementation, and Control, McGraw-Hill.

Pèrigord, M. (1990). Achieving Total Quality Management: A Program for Action. International, Productivity Press.

Peters, T. J. J. and R. H. Waterman (1982). In Search of Excellence. New York, Harper & Row.

Porter, M. (2002). Absolutely Porter. Australian Financial Review Boss. **September**.

Post-Anderson (2002). Management Information Systems. Sydney, McGraw-Hill.

Pyle, D. (1999). Data Preparation for Data Mining. San Francisco, Morgan Kauffman Publishers.

Pyle, D. (2003). Preparing data for mining. The Handbook of Data Mining. N. Ye. London, Lawrence Erlbaum Associates, Inc**:** 366-391.

Pyle, D. (2004). Business Modeling and Data Mining. London, Morgan Kaufmann.

Pyle, D. (2004a). This Way Failure Lies. http://www.db2mag.com/story/showArticle.jhtml?articleID=17602328, DB2 Magazine. **Accessed December 2004**.

Quinlan, J. R. (1984). Learning efficient classification procedures and their application to chess and games. Machine Learning: An Artificial Intelligence Approach. R. S. Michalski, J. G. Carbonell and T. M. Mitchell. Berlin, Springer**:** 463-482.

Quinlan, J. R. (1993). C4.5: Programs for machine learning, Morgan Kauffman Inc.

Ramsey, C. L. and J. J. Grefenstette (1993). Case-based initialization of genetic algorithms. Proceedings of the Fifth International Conference on Genetic Algorithms, Morgan Kauffman: 84-91.

Ray, W. J. (1997). Methods toward a Science of Behavior and Experience. Pacific Grove, CA, Brooks/Cole Publishing Co.

Reinartz, T. (1999). Focusing Solutions for Data Mining: Analytical Studies and Experimental Results in Real-World Domains. Berlin, Springer.

Rendell, L. and H. Ragavan (1993). Improving the design of induction methods by analyzing algorithm functionality and data -base concept complexity. 13th International Joint Conference on Artificial Intelligence, Morgan Kauffman: 952-958.

Ridgeway, G. (2003). Strategies and Methods for Prediction. The Handbook of Data Mining. N. Ye. London, Lawrence Erlbaum Associates**:** 159-191.

Robnik-Sikonja, M. and I. Kononenko (1996). Context-sensitive attribute estimation in regression. http://citeseer.nj.nec.com/robnik-sikonja96contextsensitive.html, Faculty of Computer and Information Science, University of Ljubljana, Slovenia. **Accessed May 2003**.

Rosa, J. (2002). The Five Stages of Customer Relationships. http://dcrm.infotoday.com/articles/default.asp?ArticleID=1287, CRM Magazine. **Accessed April 2003**.

Rud, O. P. (2001). Data Mining Cookbook. New York, Wiley Computer Publishing.

Ruthven, P. (2002). Strategy Overview Presentation. SAS SUGA Conference September. Sydney.

Salganicoff, M. (1993). Density-Adaptive Learning and Forgetting. Proceedings of the Tenth International Conference on Machine Learning, Amherst, MA: 276-283.

SAP (2000). ASAP95 - Managing Organizational Change, SAP AG.

SAS Institute (1998). Data Mining and the Case for Sampling. Cary, NC, SAS Institute Inc.

SAS Institute (2000). SAS Data Mining Projects Methodology. Cary, NC, SAS Institute Inc.

SAS Institute (2003a). Applying Data Mining Techniques Using Enterprise Miner. Cary, NC, SAS Institute Inc.

SAS Institute (2003c). Advanced Programming (PROG2) Course Notes. Cary, NC, SAS Institute Inc.

SAS Institute (2004). SAS Customer Segmentation for Telecommunications. http://www.amdocs.com/documents/brochure-customerProfitability.pdf, SAS Institute. **Accessed June 2005**.

SAS Institute (online a). The Logistic Procedure. SAS Online Documentation, SAS Institute.

SAS Institute (online b). Predictive Modelling. SAS online documentation, SAS Institute.

SAS Institute (online c). Transforms Node. SAS online documentation, SAS Institute.

SAS Institute (online d). Variable Selection Node. SAS online documentation, SAS Institute Inc.

SAS Institute (online e). Clustering Node. SAS online documentation, SAS Institute.

Schiffman, L. G. and L. L. Kanuk (1997). Consumer Behavior. Upper Saddle River, New Jersey, Prentice-Hall Inc.

Schlimmer, J. C. and R. H. Granger (1986). Beyond incremental processing: Tracking concept drift. Proceedings of the Fifth National Conference on Artificial Intelligence, Morgan Kaufman: 502-507.

Schön, D. A. (1995). The Reflective Practitioner: How Professionals Think in Action. London, Ashgate Publishing Limited.

Simoff, S. J. (2003). Clustering.

Slembek, I. M. (2003). Evaluating and Improving Knowledge-Intensive Work Processes through the Applicaton of Information and Communication Technologies. Computer Sciences. Sydney, University of Technology Sydney.

Slepian, J. H. (2003). Finding Profit in Customer Behavior. CRM Magazine. http://dcrm.infotoday.com/articles/default.asp?ArticleID=3410, Destination CRM. **Accessed September 2003**.

Stanley, K. O. (2000). Learning Concept Drift with a Committee of Decision Trees, http://citeseer.nj.nec.com/482499.html, Machine Learning Research Group, Department of Computer Sciences, University of Texas, at Austin, Austin, TX 78712.

Steinberg, D. (2003). Telecommunication churn presentation. Presented at University of Technology, Sydney, Salford Systems.

Sveiby, K. E. (2001). "A Knowledge-Based Theory of the Firm to Guide in Strategy Formulation." Journal of Intellectual Capital **2**(4): 344-358.

Takeuchi, H. (1998). "Beyond Knowledge Management: Lessons from Japan." Online: http://www.sveiby.com/articles/LessonsJapan.htm.

Thearling, K. (2003). Increasing Customer Value by Integrating Data Mining and Campaign Management Software. http://www.crm2day.com/library/EpFkEEyAZAdwvuxtmS.php, www.crm2day.com. **Accessed June 2004**.

Van Everen, D. (2002). Customer Segmentation Strategies. http://www.line56.com/articles/default.asp?articleID=4055&TopicID=6, www.line56.com. **Accessed September 2005**.

Van Rooyen, M. (2004). An evaluation of the utility of two data mining project methodologies. Proceedings of the 3rd Australasian Data Mining Conference, Cairns, Australia, University of Technology Sydney: 85-97.

Von Krogh, G., K. Ichijo, et al. (2000). Enabling Knowledge Creation. New York, Oxford University Press.

Ward, J., P. Griffiths, et al. (1990). Strategic Planning for Information Systems. New York, John Wiley & Sons.

Wedel, M. and W. Kamakura (2000). Segmentation: Conceptual and Methodological Foundations. Boston, Kluwer Academic Publishers.

Westphal, C. and T. Blaxton (1998). Data Mining Solutions: Methods and Tools for Solving Real-World Problems. New York, Wiley Computer Publishing.

Whitten, J. L., L. D. Bentley, et al. (2004). Systems Analysis and Design Methods. International, McGraw-Hill.

Widmer, G. (1994). "Combining Robustness and Flexibility in Learning Drifting Concepts." Proceedings of the Eleventh European Conference on Artificial Intelligence (ECAI-94): 468-472.

Widmer, G. and M. Kubat (1993). Effective Learning in Dynamic Environments by Explicit Context Tracking. European Conference on Machine Learning, Springer-Verlag: 227-243.

Widmer, G. and M. Kubat (1996). "Learning in the Presence of Concept Drift and Hidden Contexts." Machine Learning **23**: 69-101.

Widyantoro, D. H., T. R. Ioerger, et al. (1999). An adaptive Algorithm for Learning Changes in User Interests. Proceedings of the Eighth International Conference on Information and Knowledge Management (CIKM'99), Kansas City, Missouri: 405-412.

Wikstrom, S. and R. Normann (1994). Knowledge and Value: A New Perspective on Corporate Transformation. London, Routledge.

Woods, T. (2003a). Propensity Scoring Models. Carey, NC, SAS Institute, copyrighted to StattApp Ltd.

Woods, T. (2003b). Segmentation of a Customer Database. Carey, NC, SAS Institute Inc., copyrighted to StatApp Ltd.

Yassael, H. (1998). Consumer Behavior and Marketing Action. Cincinnati, Ohio, South-Western College Publishing.

Zikmund, W. G. (2003). Business Research Methods. Various, Thomson South-Western.