CRM Analytics Framework

Joseph P Bigus * bigus@us.ibm.com
Upendra Chitnis * chitnis@us.ibm.com
Prasad M Deshpande † prasdesh@in.ibm.com
Ramakrishnan Kannan ‡ rkrishnan@in.ibm.com
Mukesh K Mohania † mkmukesh@in.ibm.com
Sumit Negi † sumitneg@in.ibm.com
Deepak P † deepak.s.p@in.ibm.com
Edwin Pednault * pednault@us.ibm.com
Soujanya Soni † soujanya.soni@in.ibm.com
Bipen K Telkar § bitelkar@in.ibm.com
Brian F White * bfwhite@us.ibm.com

Abstract

Implementing a CRM Analytics solution for a business involves many steps including data extraction, populating the extracted data into a warehouse, and running an appropriate mining algorithm. We propose a CRM Analytics Framework that provides an end-to-end framework for developing and deploying pre-packaged predictive modeling business solutions, intended to help in reducing the time and effort required for building the application. Standardization and metadata-driven development are used in the solution; this makes the framework accessible to non-experts. We describe our framework that makes use of industry standard software products and present a case study of its application in the financial domain.

1 Introduction.

Organizations collect vast amounts of data about their customers, relationships with the customers and their interactions. Analytics on such customer interaction/relationship data[1] can provide customer segmentation groupings (for example, at its simplest, dividing customers into those most and least likely to repurchase a product); profitability analysis (which customers lead to the most profit over time); personalization (the ability to market to individual customers); event monitoring (for example, when a customer reaches a certain dollar volume of purchases); what-if scenarios (how likely is a customer or customer category that bought one product to buy a similar one); and predictive modeling (for example, comparing various product development plans in terms of likely future success given the customer knowledge base). This kind of analysis leads to better and more productive customer relations in terms of sales and service.

However, data that is useful to perform such analyses is typically spread out across data sources. This data needs to be first integrated and imported into a standard form, typically in a data warehouse. Depending on the specific business problem, this data is further processed using OLAP operations to select the relevant dimensions at the appropriate level of aggregation. Data mining analysis such as clustering, classification, regression or market basket analysis is applied to this data to get the business insights of interest. Finally, the results of the data mining operations need to viewed in various ways using reporting tools to enable the business user to understand the results and make decisions. Thus companies who want to develop knowledge discovery and data mining applications must form teams of highly-skilled specialists in data modeling, data preparation, ETL, OLAP, data mining, and business intelligence reporting. These projects turn out be big budget multi-year projects requiring a lot of investment in terms of time and money.

We address this problem by proposing a CRM Analytics Framework that provides an end-to-end framework for developing and deploying pre-packaged predictive modeling business solutions. The goal is to make predictive modeling accessible to non-experts for specific business problems by embedding best practices into plug-in modules utilizing the principles of metadata driven development and standardization. A metadata driven approach enables defining the metadata once and transforming it automatically for the different stages, thus reducing manual effort. The principle of standardization aims to achieve the maximum reuse of a developed predictive modeling solution to multiple customers; we exploit the similarity between requirements of different customers within a same vertical industry (e.g., banking) and employ a single standardized data model for each vertical. Once the actual customer databases are mapped to the standard, the

15th International Conference on Management of Data COMAD 2009, Mysore, India, December 9–12, 2009

*IBM Research - Watson
†IBM Research - India
‡IBM Software Group, India
§IBM Global Business Services, India
[1] http://searchcrm.techtarget.com/

pre-packaged solution can be used directly for each customer with minimal changes.

Developing a packaged solution using the framework is then done by identifying the industry data model, defining the pre-processing steps following which the mining tool and the reporting tool are configured to package the reports for the problem at hand. Such a packaged solution can be deployed by identifying the source data and mapping it to the data model used by the package; the reports can be readily deployed therein.

2 Data Mining Flow

2.1 Current Process

Data Understanding and Acquisition

Once the business problem (e.g., attrition prediction, targeted marketing) is finalized, the current process starts with identifying the data sources relevant to building a model for the problem. Since only a subset of attributes from each data source may be relevant, the relevance of attributes from the identified data sources are then assessed. Now, a data warehouse is defined based on the chosen entities and attributes. This involves defining the dimensions, hierarchies on the dimensions, measures and the associated star schema to store the data. The final step requires mapping the enterprise data sources to the data warehouse definition, so that data can be populated in the warehouse by defining ETL scripts that would load the data into the star schema.

Data Preparation

The star schema for the data warehouse may have dimensions with hierarchies. For example, the hierarchy for the $Time$ dimension would include Day, $Month$, $Quarter$ and $Year$, representing various levels at which the data can be viewed. Although the data is often stored at the most detailed level of the hierarchy, aggregate levels may be more relevant to the prediction task. Further, most data mining engines expect data to be in a single relational table. The data preparation step is the bridge that converts the star schema data in the warehouse to the single table data required for modeling. This would involve joining the fact table with the dimension tables and aggregating the data to the level required for prediction.

Modeling and Evaluation

Having prepared the data in a manner as expected by the mining engine of choice, the mining engine is now invoked with the required parameters (e.g., table name, name of attribute to be predicted etc.). The mining algorithm is chosen from one or many of classes such as association rule mining, clustering or classification. It is not unusual to have a user who may want to compare the performance of multiple clustering algorithms on the same data. The learnt model is then applied on any available test data to assess the quality; this evaluation may lead the user to new choices for the mining operator or algorithm, leading to an iterative process leading to an eventual choice.

Data Understanding and Acquisition	30%
Data Preparation	40%
Modeling and Evaluation	15%
Deployment and Reporting	15%

Table 1: Effort Distribution

Deployment and Reporting

Now, the model is deployed on the real data in the warehouse to get predictions for the attribute of interest. The results could then be viewed using an interactive reporting software (that usually have views aiding easy visual analysis) prior to taking business decisions.

2.2 Transformed Process

The guiding line for transforming the process is the effort distribution as shown in Table 1; this was estimated based on discussions with practitioners in IBM. The complexity of the data forces a large fraction of the time to be spent in the first two stages. Besides, personnel working in these stages must have strong business skills to identify the relevant attributes and database skills to write the ETL scripts. We address these issues by using the principles of standardization and meta-data driven development.

Standardization

Within a specific vertical industry such as banking, the diversity of requirements, data models, and business problems is limited. For example, banking data often comprises of transactions and customer demographics. Standardization increases the reusability of the components to enable greater reuse among customers within the industry. Once the data model is standardized, all the subsequent steps such as data preparation, building the models, and reporting are readily applicable without any adaptation. To deploy a packaged solution in a customer environment, it is merely necessary to map the available data to the standardized model; this is often easy since data sources are very similar within an industry. This reduces the time required to deploy data mining solutions for new organizations. Figure 1 shows the various components of a packaged analytics solution based on IBM industry standard data models, Infosphere data preparation operations and Cognos reports.

Meta-data Driven Development

Each of the stages in the transformed process could still be handled by different software (e.g., Cognos and

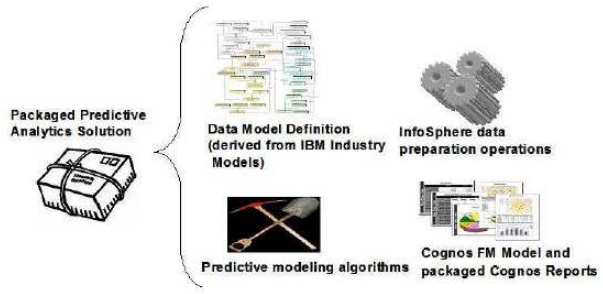

Figure 1: Industry Specific Packaged Solutions

InfoSphere as shown in Figure 1; this flexibility enables one to leverage the capabilities of the various products. However, invoking each of these APIs involves supplying them with parameters such as location of the data and configuration parameters for algorithms. Having employed a standardized data model, it is often possible to programmatically generate such parameters that serve as metadata. This metadata-driven development would substantially reduce the effort as we will see in the subsequent sections.

3 CRM Analytics Framework

3.1 Standard Data Models

Our framework is dependent on standard data models; such models, as mentioned earlier, lend reusability across various customers within the same industry. The key challenges that we seek to overcome by usage of standard data models are that of *maintaining a consistent understanding of the data model across personnel of varied backgrounds* and *providing interoperability among the different products that are involved in the process.*

Practitioners within IBM have defined standard data models such as the Banking, Telecom and Retail Data Warehouses [6] for the respective industries. Providers of standard data model providers will typically provide a tool to browse the model and create an implementation model from the standard data model. For example, Enterprise Model Extender(EME) is such a tool that enables visualization of BDW.

Scoping is the practitioner term used to refer to identifying a subset of attributes from a standard data model to be used in the implementation. Such scoping is aided by a tool such as EME. EME then generates the OLAP and physical data model based on the chosen dimensions and measures in the scope. The OLAP model generated from this stage will be used in the data preparation and the reporting stage discussed in Section 3.2 and Section 3.4 respectively. The OLAP definition is deployed to the data warehousing server such as IBM Cognos that would generate the ETL to populate the warehouse from the operational databases. It is important to note that there are no standards for the OLAP definition; however, tools from the same vendor such as EME and Cognos understand what each other expect. In more general scenarios, customers can develop a bridge to transform the OLAP model generated by the standard data models tool to the definition format required for their data warehousing server.

3.2 Data Preparation

The effort required in creating the data view with suitable derived fields makes this the most time-consuming stage in the process. The different steps involved in the current practice include *assessing the relevance* of the available data for the problem at hand, *identifying data fields* for usage in the mining schema and *building ETL scripts* that transform the data to the mining schema. The above sequences of steps are often repeated if the subsequent step of model creation does not yield a satisfactory model. For example, let us assume that a *customer attrition* problem, results in false negatives. On analysis, the reason of the false negatives may be due to not choosing an attribute such as *existence of declined cash withdrawal transactions* in the data preparation step. This causes the algorithm to miss out on all those customers who may churn due to not being able to withdraw cash. Upon discovery of such discrepancies, the user may choose to revisit the data preparation step and make necessary modifications.

The key assumptions that enable us to semi-automate the data preparation steps without significant loss in flexibility are outlined below:

1. **Existence of Standard Data Models:** We assume the existence of industry specific standard data models (e.g., the *Banking Data Warehouse* in Section 3.1). Such standard data models have a pre-defined schema and documentation thus reducing the chances of various different interpretations of the same field.

2. **Fixed Set of Finite Business Problems:** Business problems often come from a finite space and are generic. For example, *customer profitability analysis* and *customer value prediction* are only slightly different business problems that may possibly be not different from the perspective of attributes that would be used for both. Our assumption of a very small finite space of generic business problems is hence well-justified.

3. **Well-known Attribute Relevance Information:** We assume that the relevance of an attribute in a chosen industry model to the business problem at hand is well-known. This relevance information is often re-discovered in the current practice for every instance of having to do an analytics task. Such specific assessments could be

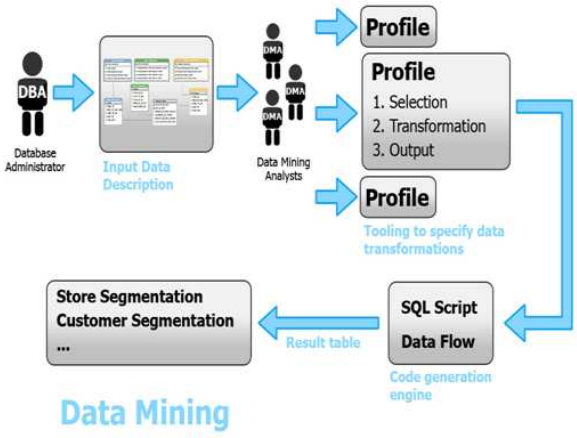

Figure 2: New method steps of aggregation and transformation of data residing in relational databases

captured and documented as meta-data; CRM-AF relies on such meta-data to automate the data preparation step.

In the context of these assumptions outlined above, it is easy to see that a semi-automatic data preparation process automatically falls out. We outline the transformed process (Ref. Figure 2) as below.

1. **Data Understanding:** This process is completely eliminated since we rely on standard data models and metadata derived from their documentation and assume that the user chooses from among a finite list of business problems.

2. **Identifying Data Fields:** Now, the metadata on attribute relevance is used to identify the data fields to be used for modeling. The system also identifies ways of generating derived attributes and suggests them to the Data Mining Analysts(DMA). The DMAs then validate it and use the tools for data transformation to specify the transformations. This step, hence, is only semi-automatic and requires intervention of Data Mining Analysts. *This is the only step requiring manual intervention.*

3. **Building ETL Scripts:** Our data preparation engine uses the identified schema and knowledge of the industry data model to generate ETL scripts for the required transformation.

3.3 Analytics Engine

The main component of the *Analytics Engine* is the *store* of various predictive modeling algorithms. This may be seen as a meta-analytics engine since it is connected to various Data Mining Engines(these Data Mining Engines are mostly commercial software such as *DB2 Intelligent Miner*, *SAS*, *SSPS* etc) through adapters and leverages the functionality of these engines. To abstract the algorithmic challenges from a naive user, the engine also contains configuration information. A typical simple use case of the Analytics Engine may be described as a sequence of steps such as *invoking the analytics engine*, *identifying the relevant algorithms* and *executing them.*

Invocation

The invocation phase allows for various styles of invocation. Apart from having to specify mandatory parameters such as *business problem type* and *variable to be predicted*, it allows for certain other options. Predictive modeling algorithms are often able to provide various satellite data (such as confidence scores) in addition to the predicted class name. Based on these capabilities, a reporting studio may be able to display results in varying fashions (e.g., ranked list, top-k etc). We refer to such alternate modes of presenting/perusing results generically as *evaluation criteria*. The invocation method allows the user to specify an evaluation criterion based on which a subset of applicable algorithms may be identified. If the invocation is done using the *preferred* mode, the preferred algorithm for the problem, evaluation criterion combination is run. For those users who want to obtain results of multiple algorithms for comparison, the invocation could be done using the *try-all* mode which would cause all applicable algorithms to be run.

Identification

This phase uses the input information to select the appropriate algorithms to be executed. In a very simple case, the analytics engine chooses either the preferred or all of the algorithms from the store that support the chosen evaluation criterion and the business problem. Another filtering step is to filter out algorithms that are not supported in any of the data mining engines connected to the CRM-AF. For example, an algorithm such as SVM may not be applicable on data with categorical attributes. The result is a set of one or more algorithms (from the store) that would be executed.

Execution

The execution phase executes all those algorithms as chosen by the identification phase using the appropriate data mining engines. If multiple algorithms are to be executed, (for example, in the case of the *try-all* mode) it chooses the data mining engines in such a way that response time is minimized. For example, if two algorithms are chosen and one of them can run on SAS and SPSS, whereas the other can only run on SPSS, the choice would be to schedule the former on SAS and the latter on SPSS (this maximizes parallel execution, assuming that the CRM-AF has access to only one instance of each of these engines).

3.4 Reporting

The final stage of predictive analysis application is reporting. The business user would like to analyze the results of the predictive modeling in various ways to aid in making decisions. This analysis can be done by creating OLAP style reports on the data in warehouse. The attributes predicted by the models are populated back into the warehouse so that they can be part of the reports. In a meta-data driven reporting paradigm, reporting would be a two step process involving *building of a data warehouse* and *generation of reports*.

Building a Data Warehouse

In our meta-data driven approach, we first *define a meta-data model that specifies the format of the data warehouse*, by scoping from a industry standard model such as the BDW model as described in Section 3.1. This warehouse is extended to include the columns predicted by the model using the Analytics Engine described in Section 3.3. Building a warehouse is essential since direct analysis of data from the operational database is usually cumbersome. Data is structured for transactional processing; hence queries tend to be complex and require joins across many tables. In addition, queries that summarize large data volumes will impact transactional system performance.

Generating Reports

Once the data warehouse is prepared, reporting engines could be used for generating various reports. The reporting engines takes the data warehouse(with OLAP model and populated data) and reporting meta data to generate reports. The business users can perform various OLAP operations such a drill up/down, pivot/unpivot, slide etc over the predicted output.

4 A Case Study for Banking Industry

In this section, we describe the application of our framework for developing an end-to-end pre-packaged predictive modeling business solution for Direct Mailing Campaign of Potential Customers. The dataset for this case study is motivated out of COIL dataset [13]. This data set contains information on customers of an insurance company. The data consists of 86 variables and includes product usage data and socio-demographic data derived from zip area codes. The end to end development has the following key steps.

- *A star schema from BDW standard model*: After mapping the COIL data fields to the BDW model, a star schema was scoped out of it.
- *A Mining transformation profile metadata*: A transformation profile was then generated; this helps to transform the data from the star schema to a format that the mining algorithms expect (typically, a relational table). This is depicted in Figure 3.

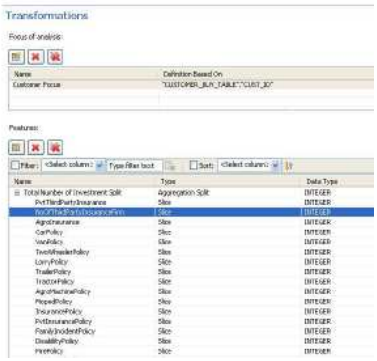

Figure 3: A mining transformation profile

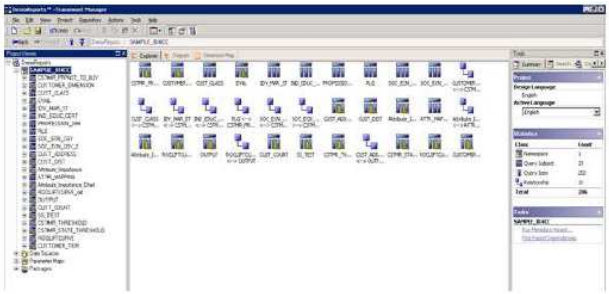

Figure 4: A meta data model in data warehouse

- *Configuring CRM Analytics Framework against an algorithm*: Since Transform Regression [8] is known to work well for direct mailing campaign selections, it was pre-configured as the algorithm of choice for the chosen problem.
- *Defining the meta data model*: The star schema was input to IBM Cognos for use as a meta data model so that reports can be built using it. The meta data model in a warehouse can be visualized as shown in Figure 4.
- *Standard reports*: Now, pre-packaged reports were built on the meta data model; a sample report is in Figure 5. This report may be run over the predicted data to display the probabilities of positive response for a direct mailing campaign; drill-down operations are also supported by Cognos.

Once the pre-packaged solution is developed, it can be deployed for any organization who needs a solution for a direct mailing campaign. The only additional step is to map the actual customer operational data sources to the data warehouse model in Cognos. Cognos then generates the ETL scripts for populating the data warehouse. The customer can thus have an out of the box predictive modeling solution deployed and running with a quick turn around time.

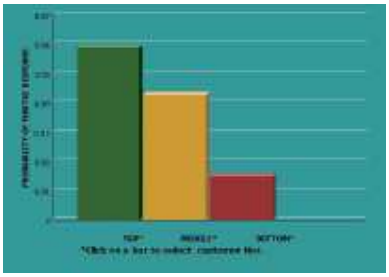

Figure 5: A Sample Report from Cognos

5 Background

Analytics on customer data has been a subject of interest over many years now [3]. In particular, analytical techniques have been shown to be useful for a variety of applications such as targeted marketing [1, 2], de-duplication of customers [4], customer value estimation [14] and customer segmentation [10].

CRM-AF is an attempt to automate the CRM analytics process to create a framework that would require minimal re-configuration before deployment, at the same time ensuring maximum utility for a broad set of scenarios. The various phases, those of data understanding, preparation, modeling, evaluation and reporting are all individually very well studied. Optimizing and implementing efficient ETL scripts is a non-trivial task [12, 11]; this is indicative of the complexity involved in automating ETL generation. Modeling for CRM analytics is often done using predictive modeling techniques that build a model out of historical training data and are used to predicting future behavior [9]. Popular techniques for predictive modeling include SVMs [5], classification and regression etc [7]. Popular algorithms are often provided as part of analytics toolkits such as SAS[2], SPSS[3], DB2 Intelligent Miner[4] and R[5]. Such toolkits often are very specialized to certain domains; for example, SPSS was primary targeted for analytics issues in social sciences[6]. Current businesses often have a very broad set of requirements of predictive analytics functionality that is impractical to be satisfied by any one single predictive modeling engine. For example, SAS has very powerful text mining capabilities[7] whereas DB2 Intelligent Miner is much more focused towards relational data. Intelligent Miner can seamlessly work on data hosted on the DB2 relational database management system, whereas relational data has to be exported into a SAS format if SAS functionality is to be applied on it. CRM-AF comes in handy in such situations since it can integrate with multiple data mining engines and thus could provide the user with the union of functionalities across multiple data mining engines.

[2] http://www.sas.com/

[3] http://www.spss.com/

[4] http://www.ibm.com/software/data/intelli-mine/

[5] http://www.r-project.org

[6] http://en.wikipedia.org/wiki/SPSS

[7] http://www.sas.com/technologies/analytics/datamining/-textminer/index.html

6 Conclusions.

In this paper, we have described a CRM Analytics Framework that provides an end-to-end framework for developing and deploying pre-packaged predictive modeling business solutions, thus reducing the time and effort required for developing the application. The framework provides tooling for various stages of solution development including data warehouse design, loading using ETL, data preparation, mining and reporting. The metadata is transferred automatically from one stage to the next, thus eliminating the need for manually creating the metadata required for each stage. By using industry standard data models and packaging various parts of the solution, it is possible to create pre-packaged solutions that can be easily deployed in different customer environments with minimal customizations. As proof of concept, we have built a packaged solution for the financial domain that can be deployed in banking and insurance companies.

References

[1] N. Abe, N. K. Verma, C. Apté, and R. Schroko. Cross channel optimized marketing by reinforcement learning. In *KDD*, 2004.

[2] C. Apté, E. Bibelnieks, R. Natarajan, E. P. D. Pednault, F. Tipu, D. Campbell, and B. Nelson. Segmentation-based modeling for advanced targeted marketing. In *KDD*, 2001.

[3] C. Apté, B. Liu, E. P. D. Pednault, and P. Smyth. Business applications of data mining. *Commun. ACM*, 45(8):49–53, 2002.

[4] J. Basak and S. Goyal. Cross-channel customer mapping. In R. Y. Lee, editor, *ACIS-ICIS*, pages 119–126. IEEE Computer Society, 2008.

[5] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[6] R. Kimball, M. Ross, W. Thornthwaite, J. Mundy, and B. Becker. *The Data Warehouse Lifecycle Toolkit, 2nd Edition*. Wiley, 2008.

[7] S. Morishita. On classification and regression. In *Discovery Science*, pages 40–57, 1998.

[8] E. P. D. Pednault. Transform regression and the kolmogorov superposition theorem. *SDM*, April 20-22, 2006.

[9] S. K. Z. Pintelas. Supervised machine learning: a review of classification techniques. *Artificial Intelligence Review*, 26:159–190, 2006.

[10] B. Saglam, F. S. Salman, S. Sayin, and M. Turkay. A mixed-integer programming approach to the clustering problem with an application in customer segmentation. *European Journal of Operational Research*, 173(3):866–879, September 2006.

[11] T. K. Sellis. Formal specification and optimization of etl scenarios. In *DOLAP*, pages 1–2, 2006.

[12] V. Tziovara, P. Vassiliadis, and A. Simitsis. Deciding the physical implementation of etl workflows. In *DOLAP*, pages 49–56, 2007.

[13] P. van der Putten and M. van Someren. Coil challenge 2000: The insurance company case. *Sentient Machine Research, Amsterdam. Also a Leiden Institute of Advanced Computer Science Technical Report*, June 2000.

[14] P. Verhoef and A. Donkers. Predicting customer potential value: an application in the insurance industry. Research paper, ERIM, Jan. 2001.