Bus Inf Syst Eng 62(1):61–75 (2020)
https://doi.org/10.1007/s12599-018-0555-z

RESEARCH PAPER

Designing Business Analytics Solutions

A Model-Driven Approach

Soroosh Nalchigar · Eric Yu

Received: 8 December 2017 / Accepted: 8 June 2018 / Published online: 6 August 2018

Abstract The design and development of data analytics systems, as a new type of information systems, has proven to be complicated and challenging. Model based approaches from information systems engineering can potentially provide methods, techniques, and tools for facilitating and supporting such processes. The contribution of this paper is twofold. Firstly, it introduces a conceptual modeling framework for the design and development of advanced analytics systems. It illustrates the framework through a case and provides a sample methodological approach for using the framework. The paper demonstrates potential benefits of the framework for requirements elicitation, clarification, and design of analytical solutions. Secondly, the paper presents some observations and lessons learned from an application of the framework by an experienced practitioner not involved in the original development of the framework. The findings were then used to develop a set of guidelines for enhancing the understandability and effective usage of the framework.

Keywords Conceptual modeling · Requirements engineering · Business analytics · Machine learning · Data analytics

Accepted after 1 revision by Jelena Zdravkovic.

S. Nalchigar (✉)
Department of Computer Science, University of Toronto, 40 St George Street, Toronto, ON M5S 2E4, Canada
e-mail: soroosh@cs.toronto.edu

E. Yu
Faculty of Information, University of Toronto, Toronto, Canada
e-mail: eric.yu@utoronto.ca

1 Introduction

Data analytics is rapidly becoming an integral part of many types of business information systems (Bichler et al. 2017). Yet there are few systematic methods to guide the development of business analytics solutions. Despite rapid advances in algorithms and technologies, many organizations still struggle to identify how to use analytics to take advantage of their data to address business problems (LaValle et al. 2010; Ransbotham et al. 2016). Building data analytics solutions has proven to be challenging due to several inherent difficulties.

Elicitation and clarification of analytical requirements are difficult but critical steps in the development of advanced analytics systems (Kandogan et al. 2014). This is to a great extent due to the large conceptual gap between business stakeholders and analytics experts. The continuous and rapid growth of machine learning and analytics algorithms, technologies, and applications intensifies the gap. Studies show that the lack of understanding on how to use business analytics techniques is a leading barrier to effective design and implementation of these systems (LaValle et al. 2010). Moreover, analytics requirements often need to be clarified for both stakeholders and analytics teams. Data science projects include asking and experimenting with a series of (initially wrong) questions in order to improve, modify, refine, and eventually get to better questions, insights, and valuable decisions (Sullivan 2014).

Analytics requirements, once elicited, must eventually lead into system design, experimentations with, and implementation of machine learning algorithms. A large number of algorithms exist and more are being developed. *Designing analytics solutions* includes decisions on algorithms while taking into account numeric metrics as well as

non-functional requirements. Algorithm selection is a critical design decision that influences several aspects of the eventual analytics solution, such as understandability of results, scalability, memory, tolerance to noisy data, and missing values. Meeting these quality requirements can be crucial to the success of the system (Luca et al. 2016).

Monitoring the impact of analytics on the business requires the project team to define and agree on a set of metrics (Chandler et al. 2011; Davenport et al. 2012). Lack of such measures could result in evaluating the right analytics system based on a wrong set of metrics and business success criteria. On the other hand, early definition of these metrics is reported to be critical to the success of the business analytics initiative (Shanks et al. 2012).

Moreover, *aligning analytics systems and techniques with enterprise strategies* is critical for eventual success of the analytics initiatives (LaValle et al. 2010; Kohavi et al. 2004). Such alignment results in an ongoing understanding of enterprise objectives by the analytics team while securing continuous business support and executive sponsorship.

Machine learning and advanced analytics applications are new capabilities for many organizations. A *shortage of talent with deep expertise in statistics and machine learning* is reported to be an obstacle towards effective use of analytics (Manyika et al. 2011). To extract value from analytics, business managers and stakeholders need to know about machine learning algorithms and their potential applications (Yeomans 2015).

In other more established areas of information systems engineering, many of the above challenges have been addressed by using techniques from conceptual modeling. By constructing a conceptual representation of the application domain of an information system and describing its semantics, such techniques can offer substantial value in developing data analytics systems (Storey and Song 2017). Conceptual modeling can provide systematic ways for identifying stakeholders' strategic goals, decision processes, and analytical questions along with insights that are required from analytical solutions. These approaches would allow connecting requirements to analytics system design, making tradeoffs among alternative algorithms, reasoning, and ensuring satisfaction of non-functional requirements. They support communicating and documenting experiments with algorithms at early phases of projects. By constructing conceptual models, data science teams along with stakeholders elaborate on and refine business strategies, identify key performance indicators and agree on a set of metrics that can be monitored for analyzing the impact of analytics solution on business. Conceptual modeling can provide a systematic way of translating business questions into data analytics and mining problems by aligning business goals and analytics technologies. Lastly, design patterns and catalogues in the forms of conceptual models can be used to provide and communicate well-proven solutions to recurring business analytics problems.

Our earlier works have introduced a conceptual modeling framework to support the design of advanced analytics solutions (Nalchigar et al. 2016; Nalchigar and Yu 2018). The main contribution of this paper is to augment the framework by providing methodological steps for constructing models in the modeling views. Also, we uncovered limitations and potential improvements of the framework through a case study in which the framework was applied by a practicing professional who was not involved in the development of the framework. As a result of testing the framework, a number of guidelines have been developed to assist in the use of the framework. An earlier version of this paper contains a more complete discussion on the benefits of the framework (Nalchigar and Yu 2017).

This paper is organized as follows. Section 2 presents an overview of the framework. Section 3 presents the design catalogues. Section 4 shows benefits of the framework for requirements elicitation, clarification and design of analytics systems. Section 5 discusses observations from a participant applying the framework along with guidelines for using the framework. Section 6 provides the research method and threats to validity. Section 7 summarizes related works and highlights the contributions. The paper ends in Sect. 8 with conclusions and directions for future work.

2 A Conceptual Modeling Framework

2.1 Overview

Advanced analytics projects require collaborative effort among team members with specialized knowledge and skills covering three major areas of work: understanding the business, designing the analytics solution, and getting the datasets ready for training and deployment. While each area has its own focus, team members need to understand each other's work so as to be able to communicate and coordinate effectively to achieve project goals. Thus the modeling framework is organized into three sub-models (hereafter called views for simplicity): the *Business View*, the *Analytics Design View*, and the *Data Preparation View*. Organizing modeling concepts into these views serves as a means for dealing with the complexity of analytics solutions, enhancing collaboration and clarity, and for managing the diversity of skillsets and roles required in such projects. These views, while representing different aspects and serving different purposes, are linked to each other to bridge the gap between business goals, machine learning algorithms, and data stores.

The Business View aims to (1) facilitate the elicitation and clarification of analytics requirements in business contexts, (2) support analysis of those requirements (e.g., prioritization), and (3) ensure the alignment of business and analytics strategies. The main modeling elements are Actors, Strategic Goals, Indicators, Decision Goals, Question Goals, and Insights (see Fig. 1a).

Strategic Goals symbolize business objectives and strategies. *Indicators* represent numeric metrics that measure and monitor performance with regard to some objectives. *Decision Goals* represent situations where an *Actor* needs to select one option among a set of possibilities. They symbolize the decisions that are (or will be) supported by the analytics system. *Question Goals* represent the "needs-to-know" of the Actors during decision processes. For each Question Goal the *Type*, *Topic*, *Tense*, and *Frequency* attributes are specified. Question Type denotes the question phrase (what, who, when, where, why, how). Question topic captures the focus of analysis and reveals related parts of enterprise data stores for the problem at hand. Question Tense (past, present, future) represents the temporal aspect of the focus of the analysis. In many cases, specifying the tense facilitates finding a family of analytics techniques that is most relevant to the business needs. Question Frequency indicates how frequent the corresponding actors need an answer for the Question Goal. *Insight* elements characterize the kinds of patterns and findings that *answer* the Question Goals. For each insight element the *Type*, *Input*, *Output*, *Usage Frequency*, *Update Frequency*, and *Learning Period* attributes are defined. These attributes support translating the business questions into data mining problems.

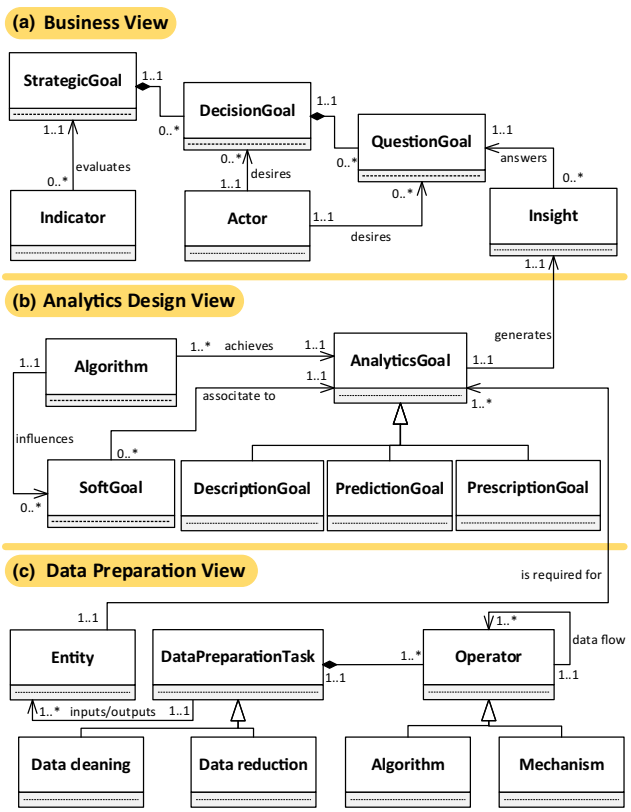


Fig. 1 Simplified metamodels for **a** business view, **b** analytics design view, and **c** data preparation view

The Analytics Design View aims to (1) support exploration of alternate approaches for the analytical problem at hand, (2) facilitate design of (machine learning) experiments and identifying trade-offs, and (3) support algorithm selection and monitoring their performance over time. The main modeling elements are Analytics Goals, Algorithms, Softgoals, and Influences (see Fig. 1b).

Analytics Goals capture the intention of the analysis to be performed over the datasets. Three types of analytics goals are distinguished. If the analytics aims to predict the value of a data attribute (i.e., a variable or data column), it is called a *Prediction Goal*. If the analytics aims to summarize and explain the dataset, it is called a *Description Goal*. If the analytics aims to find the optimal alternative given a set of options and criteria, it is called a *Prescription Goal*. The type of Analytics Goal can be derived from the type of Insight that is required to generate (from the Business View). Each Analytics Goal is then connected to its corresponding Insight element via the *generates* link. *Algorithms* are procedures and calculation steps that are needed to fulfill an Analytics Goal. They are connected to Analytics Goals through the *performs* link, showing a means-end relationship (Yu 2011). *Softgoals* represent quality requirements to be taken into account during design of the machine learning solution. *Influence Links* show how the Softgoals are satisfied through operationalization and design decisions. This view is connected to the previous modeling view through the *generates* links.

The Data Preparation View aims to (1) support the sharing and reuse of prepared data assets, (2) enhance data awareness among analytics users, and (3) ease data understanding by providing a reference for data engineers (who prepare datasets) on data preparation activities. The main modeling elements are Entities, Relationships Preparation Tasks, Operators, and Data Flows (see Fig. 1c).

Entities and their *Relationships* represent the raw data tables and their conceptual relationships. They also represent prepared datasets which are the eventual output of data preparation activities. The prepared datasets are connected to their corresponding Analytics Goals via the *is required for* link. *Data Preparation Task* represents the general task of preparing data for accomplishing some analytics goals. *Data Cleaning*, *Data Reduction*, Data Transformation, and Data Integration are types of preparation tasks. A Data Preparation Task consists of one or more *Operators* that

are linked via *Data Flows*. This view is connected to the previous modeling view through the *is required for* links.

2.2 Sample Usage Methodology

This section presents a sample methodology for constructing models in the three modeling views of the framework. The modeling steps are explained in a top-down fashion, starting from high level business strategies towards machine learning solution design and data preparation workflows. However, in practice, such models can be developed and complemented through bottom-up and/or hybrid approaches. Hence, the steps explained here are considered as a sample usage methodology. Furthermore, construction of the models in each view is meant to be led by different roles. Business View models can be built primarily by business analysts. Analytics Design View models can be constructed and updated mainly by data scientists (who create, implement, or apply machine learning algorithms). Data Preparation View models can be created primarily by database administrators and data engineers (who have a solid understanding of existing data assets, database design and queries in the business domain). In real-world projects, there can be variations and overlaps in such roles depending on the nature and complexity of the problem and structure of the project.

2.2.1 Constructing the Business View Model

Business View models are built iteratively with participation from business stakeholders, business analysts, and data scientists. Constructing such models involves understanding of a business in terms of its goals, their interrelationships, and metrics that the business use to monitor how effectively it is achieving business goals.

The modeling process starts with identifying the *Strategic Goals* and their *Influences*. Strategic Goals are refined into lower-level goals through *Decomposition Links*. After modeling Strategic Goals, performance *Indicators* are identified and linked to them. Next, the modeler identifies *Situations* and their *Influences*. Situations represent factors that can influence the achievement of strategic goals in a favorable or unfavorable way. They refer to partial state of affairs (partial model of the world) and can be internal or external to the business. The outcome of these initial steps is a Business Intelligence Model (BIM) instance for the business domain under consideration. Further details about the BIM language, including examples, a sample methodology, and a case study, can be found in (Horkoff et al. 2014; Barone et al. 2012).

To proceed towards analytics, the business analyst starts by asking the question of: what are the decision(s) that need to be made (and by whom) in order to achieve each business goal? In this step the business analyst works closely with business stakeholders to identify the key users of analytics solution, their work processes and decisions that they are responsible for. The output of this step is an extended BIM model with Strategic Goals decomposed into one or more *Decision Goals*.

The next step is facilitated by asking: what would the decision maker(s) need to know during the decision processes? Each Decision Goal thus leads to one or more *Question Goals,* each of which can be further refined into more detailed Questions Goals. At the most detailed level of refinement, each Question Goal represents a set of requirements covering a certain aspect of the analytics solution. The Business Questions Catalogue (introduced later in Sect. 3.1) provides a wide collection of common Question Goals that can support the modeling task in this step.

For each Question Goal at the lowest level of refinement, *Insight* elements are then specified. The Insight elements respond to the question: what kinds of answers are needed for the Question Goals to be satisfied? The *Type* attribute of Insight points to the types of analytics (e.g., predictive model, logical rules) to be performed, and from there to the relevant algorithms and techniques for mining the datasets. For effective modeling of Insights, the business analyst and data scientist need to have a good understanding of the business questions on one hand, and an understanding of different kinds of machine learning tasks and analytics models on the other hand.

2.2.2 Constructing the Analytics Design View Model

Constructing the Analytics Design View models starts with specifying the top level *Analytics Goals* that the system would achieve. First, for each Insight element from the Business View model, an Analytics Goal at the highest level of the Analytics Design View model is specified. Analytics Goals are connected to their corresponding Insight element via the *generates* link. Towards this end, the data scientist can start by asking the question of what kind of analytics (descriptive, predictive, or prescriptive) would be appropriate to generate the Insight element under consideration? Next, the Analytics Goals are decomposed into more specific lower level goals depending on the nature and shape of the available data on one hand, and nature of the problem on the other hand.

In the next step, for each Analytics Goal, a set of *Algorithms* that can fulfill such goal are modeled. To model Algorithms, one can start by asking the question of what Algorithm(s) exist for fulfilling the Analytics Goal at hand? The choice of algorithms in the model is a design decision that is affected by the shape, size, and format of the dataset at hand. For example, given a classification type of

Analytics Goal, if the input variables are categorical, Naïve Bayes algorithm can be a good candidate. On the other hand, the choice of algorithms can imply certain data preparation steps (such as removal of missing values, normalization of numerical features) to be taken into account while constructing the Data Preparation View model. For example, data normalization is a critical step to be taken into account when using a distance-based mining algorithm (e.g., *k*-Nearest Neighbor). An understanding of the data is required for this step to be performed effectively. Moreover, a good understanding of different kinds of machine learning algorithms and analytics approaches is required in this step. The Algorithm Catalogue (see Sect. 3.2) provides a wide collection of (common) machine learning algorithms categorized by the types of Analytics Goals. Context elements in the catalogue are used to help decide which algorithms are suitable given the characteristics of the dataset.

In the next step, the criteria for making design decisions and algorithm selection are modeled in terms of *Softgoals* and *Indicators*. To identify Softgoals, one can start by asking the question of what are the quality attributes or non-functional requirements (NFRs) that need to be satisfied from the point of view of the users? To identify Indicators, one can start by asking the question of what numeric metrics would be used to compare and evaluate the algorithms? In collaboration with stakeholders, the data scientist defines and obtains agreements on (upper and lower) threshold values for Indicators (e.g., minimum required accuracy for predictive models, maximum execution time for an algorithm). Also, Softgoals are refined and the Influence links among them are modeled. The Algorithms Catalogue provides a wide collection of Softgoals and Indicators that are relevant and common for various types of Analytics Goals.

The next step focuses on modeling the *Influence Links* from Algorithms to Softgoals. In this step, the data scientist (here in the role of modeler) can perform existing analysis techniques over the (goal) model to find Algorithms that make critical Softgoals unachievable. By removing such Algorithms from the model, the modeler can prune the space of alternatives early in the design phase of the project and thus reduce the number of experiments to be conducted. Towards this end, a complete and accurate modeling of Softgoals in the previous step is essential. The Algorithms Catalogue provides labeled Influence Links among Algorithms and Softgoals, representing the knowledge on how well the Algorithm is expected to perform with respect to various Softgoals.

The last step focuses on modeling the Influence Links from Algorithms to Indicators. In this step, the selected algorithms are tested on the prepared dataset(s) and the values for Indicators are calculated. These values are modeled in terms of numeric labels for Influence Links from Algorithms towards Indicators. This step includes setting and tuning the parameters for algorithms. The data scientist keeps track of the choice of these parameters. The modeling artifact and values for the indicators are presented and discussed with business users. Based on observations and experimental results, the modeler may reconsider and update the labels for contribution links towards softgoals in the previous step. Also, given such findings, the data scientist may consider experimentation with additional algorithms and update the Analytics Design View model. At the end of this step, the algorithms are ranked and design decisions are finalized.

2.2.3 Constructing the Data Preparation View Model

To construct a Data Preparation View model, one starts by acquiring an understanding of existing data models and by selecting portions of data files and schema that are relevant or needed for the data analytics solution. Towards this end, one can start by asking the question of what kind(s) of data are actually needed for delivering the results and answering the Question Goals at hand? The Question Goals (and their topics) and Insight elements (including their learning period and update frequency attributes) in the Business View model should be understood and referred to during this step. Data Preparation View models are built primarily by database administrators and data engineers with participation from data scientists. Entity Relationship Diagrams or data warehouse schema models and other documentation can be used in this step. Visualization, and initial descriptive/statistical analyses may be performed at this step by data scientists to understand the shape, size, and type of the data at hand and to verify the quality and meaning of data attributes.

In the next step, the focus is to define the prepared dataset and attributes on which the algorithm(s) would be executed. By specifying prepared datasets, the data scientist in collaboration with database admins and engineers specify the required output of data preparation steps. Towards that, one can start asking the question of what data attributes (i.e., features), in what format, and aggregation level are needed for the Question Goals under consideration? This includes decisions on the attributes, data types, aggregation levels, and selection of records (filtering). Also, feature selection analyses and correlation tests may be performed by the data scientist to exclude/include certain attributes. Given such findings, the project team may reconsider the input datasets and revise the model from the previous step.

In the next step, the focus is to decide and design the flow of *Data Preparation Tasks* that transform the input data into the prepared datasets. Towards this end, one can start by asking the question of what (sequence of)

integration, cleaning, aggregation, filtering and other data preparations are needed for transforming the raw data tables into the prepared data tables? Findings from data understanding is a critical input to this step. Database administrators and data engineers and data scientists work together while taking into account data quality and treatment aspects. This includes decisions on how to deal with noise and outlier values, treat imbalanced dataset, address missing values, use sampling methods, derive and construct new attributes, change data types, among others. Data Preparation Catalogue (introduced in Sect. 3.3) provides techniques and algorithms for performing various data preprocessing tasks which can be referred to while performing this step.

2.3 An Illustration

In this section, through an illustrative case, we provide examples of primitive concepts and explain sample steps for constructing such models. These are explained using an illustrative case of a shopping mobile app company. The company offers a variety of products to its users via in-app purchases. It aims to improve its market share and net profit by focusing on user retention and loyalty offers. The stakeholders are interested in using machine learning and advanced analytics solutions to support a wide range of decisions about their marketing campaigns and reward programs. The company's data stores include user demographics, their activities within the app, and their online purchases.

Constructing Business View models start with elicitation of Strategic Goals, relevant Indicators, Situations and their Influences. In Fig. 2 (top portion), `Improve customer retention` and `Achieve high performance through email campaigns` are examples of Strategic Goals. `Click through rate (%)` and `Conversion rate` are examples of an indicator. Also, `Low switching costs to customers` is an example of a situation.

The modeling activity continues by elicitation of Decision Goals and decomposing them into Question Goals. In Fig. 2, `decision on content of the emails` is an example of a Decision Goal. It shows that in order to `Achieve high performance through email campaigns`, the corresponding actor[1] needs to make the `Decision on content of the emails` to be sent to the target users. Also, `What are the most relevant products for each user group?` is an example of a Question Goal. It shows that in order to make the `Decision on content of the emails`, the corresponding actor needs to know the products that are more relevant for each group/cluster of users.

[1] Due to space limitations, actors are not shown in Fig. 2. See Nalchigar and Yu (2018) for instantiations of this element.

Next, Insight elements are linked to Question Goals through the *answers* links. In Fig. 2, `User-Product Association Rule Model` is an example of an Insight. It symbolizes a set of `Logical rules` (e.g., Canadian users with an age between x and y are likely to buy product z), which answer the question of `What are the most relevant products for each user group?` At run-time, this Insight requires `User demographics` data as input, in order to generate a list of `Product(s)` as the answer to the question. This Insight is used on a `Weekly` basis and the rules are mined from the dataset with a `60 months` time interval. Figure 2 contains more examples of each modeling concept from the Business View.

Constructing Analytics Design View models start with modeling Analytics Goals for each Insight element from the Business View model. In Fig. 2 (middle portion), `Predict user churn` is an example of a Prediction Goal which is decomposed into `Classification of user profiles and purchases`. Also, `Describe user behaviour` is an example a descriptive analytics intention, which is further decomposed into the goal `Discover patterns in user purchases`.

The modeling activity continues with defining the algorithms and the criteria that are used for comparison and monitoring their performance. Towards this, the Algorithms, Indicators and Softgoals are modeled. Figure 2 shows that `Apriori`, `ECLAT`, and `FP-Growth` are alternative algorithms for achieving the pattern discovery goal. Also `Accuracy` and `Sensitivity` are examples of an Indicator, while `Speed of learning` and `Tolerance to missing values` are instances of Softgoals.

Modeling this view is followed by specifying how the algorithm selection criteria are influenced by alternative algorithms. Towards that, Influence Links from Algorithms to Softgoals and Indicators are created and their labels are populated. In our example, the Influence Link from the algorithm `Apriori` towards the Softgoal `Speed of learning` shows that this algorithm will `Hurt` (–) achievement of that softgoal. Also, the link from `FP-Growth` to the indicator `% of redundant rules` shows that the algorithm will result on the value of `0.17` for that indicator, determined through experiments. More examples of each modeling concept in the Analytics Design View can be found in Fig. 2.

Constructing Data Preparation View models start with an understanding of existing data tables, attributes, and relationships. Figure 2 (bottom portion) shows that for each `User`, demographics data such as `Age` and `Gender` are captured. This is followed by specifying the outcome of data preparation activities which is a (set of) dataset(s) ready to be analyzed/mined by algorithms. In our example, `Demographic Product` and `Churn`

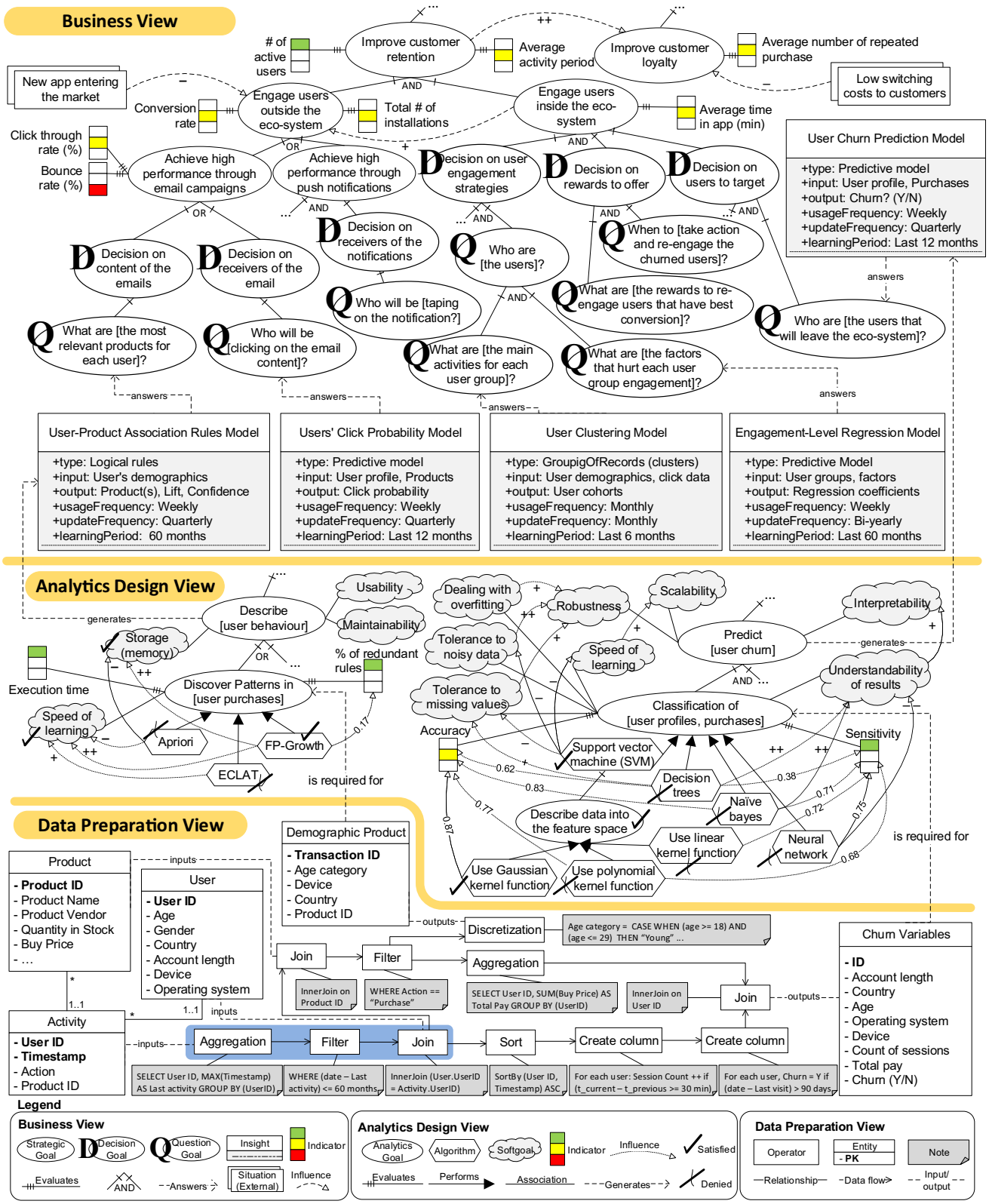


Fig. 2 Fragments of the three modeling views for the shopping mobile app company. Due to space limitation, the analytics design view and the data preparation view are showing the solution for only two (out of eight) question goals in the business view

Variables represent prepared data tables that are linked to the previous view via the *is required for* links.

The modeling activity continues with designing the workflow and operations that are needed to extract and transform the raw datasets into the prepared data tables. In Fig. 2, the blue-shaded area in the Data Preparation View shows an example of a Data Reduction task. It shows that the system excludes those users who have not done any shopping or other activities for more than 5 years. Also, Create column and Join are examples of operators. Operators are linked by Data Flows to represent the sequence and dependencies.

In this step, modelers can use the Note elements to attach clarifications and details to each Operation in the model. For example the Note For each user, Churn = Y if (date – Last visit) > 90 days associated with a Create column operator shows that a new data column is created and its value is Y if the corresponding user has been inactive for more than 3 months. More examples of concepts in the Data Preparation View can be found in the bottom portion of Fig. 2.

3 Design Catalogues

Creating and revising models in the three modeling views requires knowledge about business objectives and decisions, machine learning algorithms and techniques, as well as data preprocessing and cleaning approaches. An important component of the framework is a set of design catalogues that provide such knowledge required for modeling activities in the three views. The catalogues organize and represent a body of analytics know-how knowledge to be used and referred to during requirements analysis and design of analytics solutions. They provide proven solutions to common and recurring analytics problems in the form of conceptual models. Three kinds of catalogues are distinguished in the framework.

3.1 Business Questions Catalogue

The goal of this catalogue is to represent a wide range of business questions that can be answered with machine learning and analytics solutions. While constructing Business View models, business analysts, stakeholders, and analytics experts can use this catalogue to browse through an organized set of Question Goals based on their Type and Tense. Within each category, a wide range of instances exist where each instance is mapped to a specific analytics goal. For example, the two question goals of Who will be [taping on the notification?] and Who will be [clicking on the email content]? (from Fig. 2) are listed under the category of *Who* and *Future*, and both are mapped to *Prediction Goal*. In this way, the catalogue bridges the gap between business questions and analytics techniques. Figure 3a shows more examples of this catalogue.

3.2 Algorithms Catalogue

Effective design of analytics systems requires experimentation with and selection of machine learning algorithms. This catalogues codifies the know-how on analytics techniques and algorithms. In particular, it represents different machine learning Algorithms that are applicable for a given Analytics Goal. The catalogue also represents well-known Indicators (i.e., metrics) for evaluation and comparison of those algorithms. For each Analytics Goal, the catalogue also provide relevant Softgoals (i.e., quality requirements) whose lack of consideration can become major issues later in the project life-cycle. Moreover, it encodes the knowledge on how each Algorithm is known to influence meeting those Softgoals. Figure 3b depicts a fragment of this catalogue. For example, in this catalogue Local outlier factor (LOF) and k-NN global anomaly detection are among algorithms for performing Unsupervised anomaly detection. Fast computation time and High accuracy are among quality requirement to be considered.

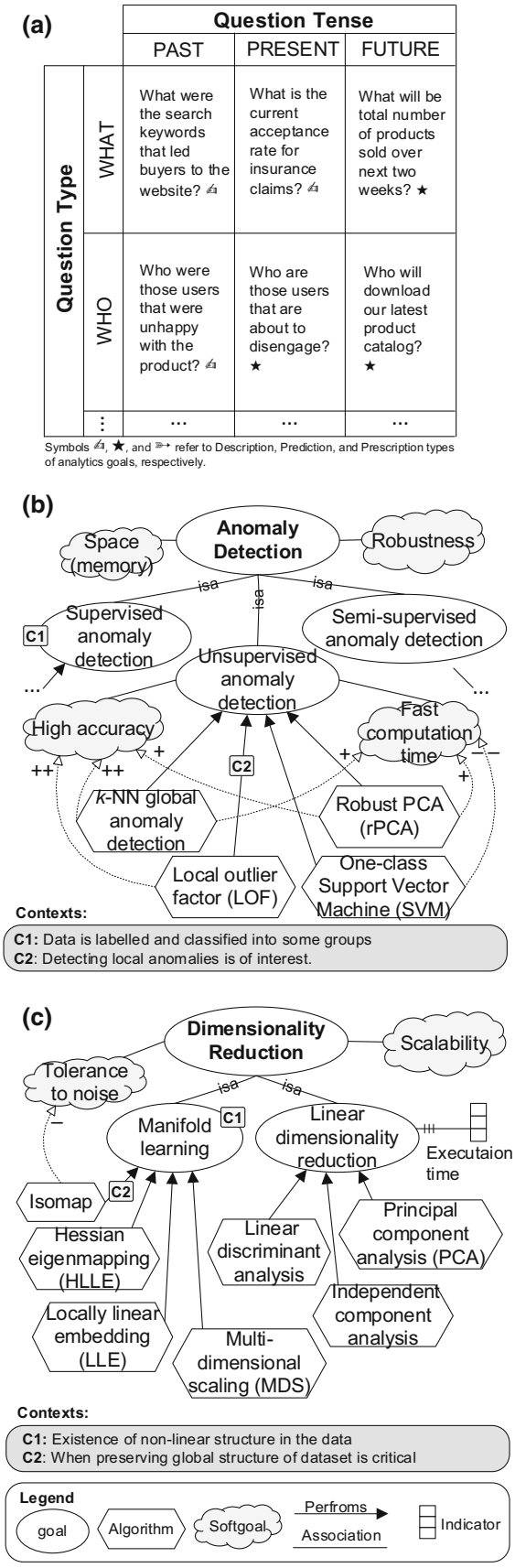
3.3 Data Preparation Catalogue

This catalogue has a similar structure to the Algorithms Catalogue, but representing the specialized know-how for data preparation. This catalogue helps developers find known methods for addressing data preparation tasks such as data cleaning and data value normalization. Figure 3c shows a portion of this catalogue. It shows that Linear discriminant analysis and Principal component analysis (PCA) are among the different ways of performing Linear dimensionality reduction.

Due to space limitations, the metamodels and content of these catalogues are not discussed here. Readers are referred to (Nalchigar and Yu 2018; Nalchigar et al. 2016) for more details.

4 Illustration of the Benefits of Applying the Framework

The Introduction section of the paper briefly discussed some of the ways in which a model-driven approach can support the development of data analytics solutions. In this section, we discuss two of those benefits in detail using examples from the shopping mobile app case. A full



◀**Fig. 3** Fragments of **a** Business questions catalogue; **b** Algorithms catalogue; and **c** Data preparation techniques catalogue. Nalchigar and Yu (2018) includes more content and examples of each catalogue

description of the benefits along with examples can be found in Nalchigar and Yu (2017).

4.1 Eliciting and Clarifying Analytics Requirements

Figure 2 shows a fragment of the Business View model for the shopping mobile app company. It shows that the company aims to `Engage users inside the ecosystem` as one of its strategic goals. The model shows that there exist a set of performance indicators such as `Average time in app (min)` to monitor how well it is doing with respect to its goals.

Achieving Strategic Goals requires business stakeholders to make critical decisions. For example, in order to `Engage users inside the ecosystem`, one needs to make the `Decision on user engagement strategies`, among others. In order to make decisions, business stakeholders need to know the answer(s) to some questions. For example, in order to make the `Decision on user engagement strategies`, the corresponding actor needs to know `Who are the users?` (a broad question that includes ambiguities). Towards answering that question, the actor needs to know `What are the main online activities of each user groups?` and also `What are the factors that hurt each user group's engagement?` The model shows that by having a `User Clustering Model` one can answer the former question. This insight receives `User demographics` and their `click data` as input and generates `User cohorts`, which answers the question of `What are the main online activities of each user group?`

Characterizing the business in terms of strategies, decisions, analytical questions and insights is a critical step towards effective design and implementation of analytics systems. Understanding business strategies helps stakeholders and project team to justify why they are performing the analytics work. In the framework, this is represented as *Strategic Goals*, such as `Engage users inside the ecosystem`. Without taking strategy into account, the project team and stakeholders would not know the *why* behind analytics initiatives. Understanding business decisions results in discovering areas that need support from analytics solutions and data-driven initiatives. In the framework this is captured in terms of *Decision Goals*, such as `Decision on user engagement strategies` in Fig. 2. This modeling element ensures the connection between analytics solution and organizational decision processes. Moreover, it facilitates linking

analytics-driven insights into actions and leveraging the analytics findings in business operations and decisions.

Eliciting business questions results in discovering the focus of the analytics project and the issues that it is intended to inform. In the framework, this is represented in terms of *Question Goals*, such as `Who are the users?` By modeling Question Goals, one is indeed eliciting the needs-to-know of stakeholders towards their decisions, which will result in performing the right analysis for the right user. Moreover, confirming the Question Goals with stakeholders support the process of understanding and communicating analytics findings, once they are generated.

By refining business questions into sub-questions, one can discuss and resolve early ambiguities that are raised by business stakeholders. In the framework, this is represented in terms of *Decomposition Links* that break a Question Goal into sub-goals. For example, in Fig. 2, the question goal of `Who are the users?` is refined into sub-questions. In addition, Question Goals are analyzed in terms of *Type*, *Topic*, *Tense*, and *Frequency*. Specifying these attributes for each question goal assists in arriving at a set of clear and accurate requirements in addition to enhancing the communication and understanding between developers and stakeholders.

Understanding analytical insights help characterizing the type of findings that are required for answering the business questions. In the framework, this is represented in terms of *Insights*, such as `User Clustering Model`. This allows specification of the actual outcome of the machine learning algorithms. By modeling the desired outcome, indeed the project team reveals the (group of) analytics techniques to be used for the problem at hand. Insight elements are modeled in terms of *Type*, *Input*, *Output*, *Usage Frequency*, *Update Frequency* and *Learning Period* (See Fig. 2). During the process of modeling, by refining question goals into sub-questions and thereafter specifying the insights, one can clarify the analytics requirements, reduce ambiguities, while having the stakeholders involved in the process.

4.2 Deriving Analytics Solution Design

The middle section of Fig. 2 shows part of an Analytics Design View model for the shopping mobile app case. On the right side, the model shows the Analytics Goal of `Predict user churn`. Towards that goal, the analytics solution needs to achieve the `Classification of user profiles and purchases`. The model shows that there are several alternative algorithms that can perform the Classification Goal, such as `Support Vector Machine (SVM)`, `Decision Trees`, `Naïve Bayes`, and `Neural Networks`. These Algorithms are evaluated with regard to some numeric metrics such as `Accuracy` and `Sensitivity`. The model also shows that Softgoals such as `Tolerance to missing values`, and `Tolerance to noisy data` are considered while designing the system. The model also represents how each algorithm would influence the metrics (numeric labels) and the softgoals (qualitative labels). For example, use of `Neural Network` would result in the value of for `0.75` for `Sensitivity` while it would `Break` (–) the softgoal of `Understandability of results`. The model shows that the selected Algorithm is `Support Vector Machine (SVM)` with the `Use Gaussian kernel function`.[2]

At design time, by knowing the desired types of outputs, one can find the kinds of analytics techniques that need to be performed. In the framework, this is captured through *Insight* elements, their *Type*, *Analytics Goals*, and *generates* links. The Insight type specifies what kinds of machine learning output would be required for the business question at hand. The type of Insight, once clarified, reveals the category of machine learning algorithms that can be used for the requirements at hand. For example, in Fig. 2, the insight `User Churn Prediction Model` with the `Predictive Model` type, suggests the need for predictive analytics (i.e., prediction goal). In Fig. 2, this is represented in terms of the prediction goal of `Predict [user churn]`.

The type of Analytics Goal, once revealed, suggests a relevant set of alternative Algorithms for the problem at hand. The Algorithm Catalogue (see Sect. 3.2) supports this step. The project team can browse through it to derive the design of the analytics system. In Fig. 2 the prediction goal is decomposed into the `Classification of user profiles and purchases` which can be performed by alternative algorithms.[3] Designing analytics systems include making decisions on algorithms with respect to criteria. In the framework, those criteria are modeled in terms of *Softgoals* and *Indicators*. Soft-goals, their Influence, analytics Indicators along with their priorities will be used for making design decisions. Lack of such considerations can result in an implementation where critical quality requirements are not satisfied.

5 Discussion

We demonstrated different ways in which the modeling framework can be used in an illustrative case. We

[2] Assuming that the `Accuracy` metric has the highest priority among the metrics and softgoals.

[3] Due to space limitations, the model in Fig. 2 is showing only one of the classification goals. There can be several classification models for predicting user churn each with different prediction periods and time intervals.

presented instances of models in three modeling views and described some of the analyses that they can enable. Our earlier works (Nalchigar et al. 2016; Nalchigar and Yu 2017, 2018) also provide more cases and demonstrate other usage settings. Such illustrations serve as preliminary validation of the framework and suggest that it can have a positive impact in the requirements analysis and design of analytics solutions.

Aside from serving as potential use cases of the framework, the case studies helped us to receive feedback and learn about some limitations and potential improvements to the framework such as the following:

- From a meta-model design perspective, the only link or conceptual relationship between two Decision Goals is the Decomposition Link. The meta-model can be extended to accommodate other kinds of links among decisions (e.g., sequence, trigger, and influence). While this can enable new types of analyses, it requires further research and considerations from organizational decision theory.
- In the Analytics Design View, those Indicators that are attached to the same Analytics Goal are treated equally. We encountered situations where Indicators can have different degrees of importance and also can be conflicting. This requires the framework to capture importance and priorities of the Indicators and Softgoals. While this can increase expressiveness of the framework and ease algorithm selection, the models may become more complex and harder to learn and use.
- We also found that modelers might mix goals with meta-goals (goals about goals) all in the same diagram.
- In the course of the case studies, we identified that each goal (e.g., to increase x) is naturally paired with an implicit decisions (e.g., decision on how to increase x).
- The modeler may have difficulty in finding appropriate wording to concisely and accurately express Question Goals and Decision Goals. Meaningful naming of these elements is essential for arriving at a set of accurate and precise analytical requirements; since they reveal the type of required analytics.
- We found that the semantics of catalogues need to be clarified and that guidelines for creating and extending catalogues are also needed.

These findings and observations motivated us to create a set of guidelines for applying the framework. The guidelines aim to enhance the usability, correctness, and understandability of models in the three modeling views and to improve the overall consistency and effectiveness of the framework. In addition to the observations above, two other sources of information were also used to develop the guidelines: (1) lessons learned from an ongoing project where the framework is being tested and models are discussed with real business stakeholders; and (2) experience of authors in the area of goal-oriented modeling techniques augmented with benchmarks from existing goal-oriented catalogues (such as guidelines for *i** modeling). The guidelines are grouped into different categories according to the concern they address. *Elicitation Guidelines* aim to facilitate elicitation of various modeling elements in the three views. Two illustrative example of guidelines in this category are:

- *Elicitation of Question Goal Topics*. In the presence of data warehouse schemas (e.g., snowflake schema), the topics of Question Goals can be extracted from the measures in the fact table and (part of) its associated dimensions. Figure 4 provides an example of this guideline for elicitation of two Question Goals.

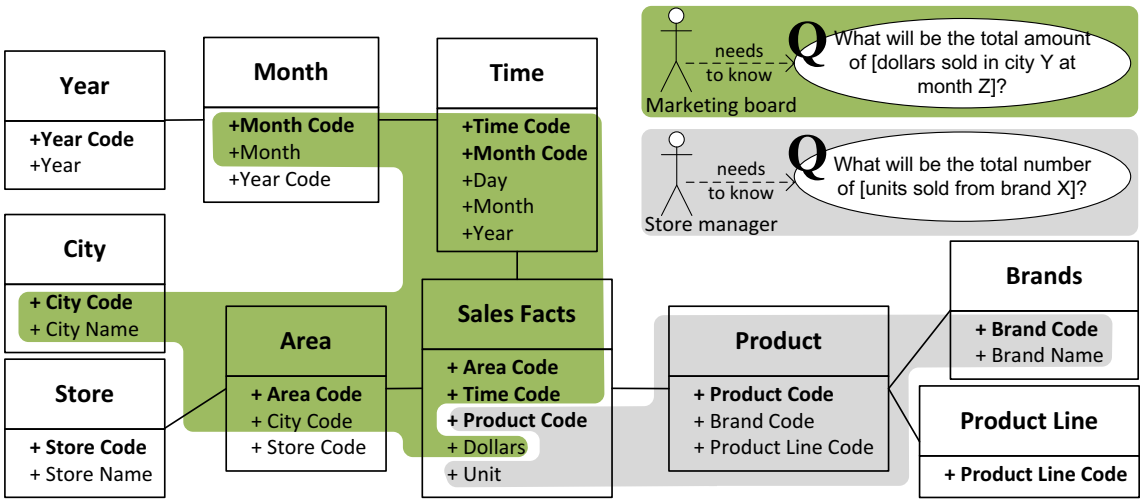


Fig. 4 Two examples showing how data warehouse schemes can facilitate identification of topics in question goals

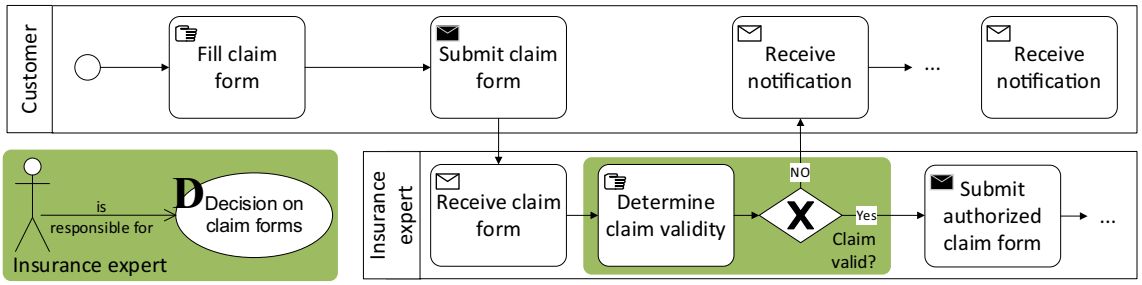


Fig. 5 An example showing how BPMN process models can facilitate elicitation of decision goals and corresponding actor(s)

- *Elicitation of Decision Goals.* Decisions are made by actors (humans or software agents) at all levels of an enterprise. They are choice points within execution of some ongoing process. If process models are available (e.g., BPMN models), Decision Goals can be elicited from them (e.g., activities that are just before a diverging Gateway). Figure 5 shows an example of this guideline in an insurance claim approval process.

Syntax Guidelines aim to improve the syntactical correctness of the models and correct usage of different modeling elements. Examples of guidelines in this category are:

- *Direction of Decomposition Links between Strategic Goals and Decision Goals.* Direction of decomposition links between Strategic Goals and Decision Goals should be only from Strategic Goals towards Decision Goals and not the other way round. In this way, the modeler specifies what decisions need to be made as part of achieving the Strategic Goal at hand. Figure 6a depicts examples of this guideline.
- *Source and Destination of Influence Links.* Influence Links are not allowed from Decision Goals and Question Goals. Influence Links should only be used to represent the influence of Strategic Goals on Strategic Goals, of Situations on Strategic goals, and of Situations on other Situations. Figure 6b shows a possible case of wrong use of Influence Links.

Due to space limitations, other categories of guidelines (e.g., *Naming Guidelines*), are not discussed here.

6 Research Method and Limitations

In our previous works (Nalchigar et al. 2016; Nalchigar and Yu 2018), the framework was tested in three cases by the authors (creators of the framework) playing the role of modelers. The primary focus of validation in those papers was to examine if the framework can express and communicate some abstractions of real analytical systems. In this paper, the models were initially created and analyzed by an independent participant who had work experience as a data scientist in addition to some experience in conceptual modeling and goal-oriented requirements engineering. The case and its models were developed from two main sources: (1) a collection of analytics case studies and white papers retrieved from Internet, and (2) the authors' and participant's collected experience from real data mining projects. The participant was not involved in the development of the framework and hence this collaboration

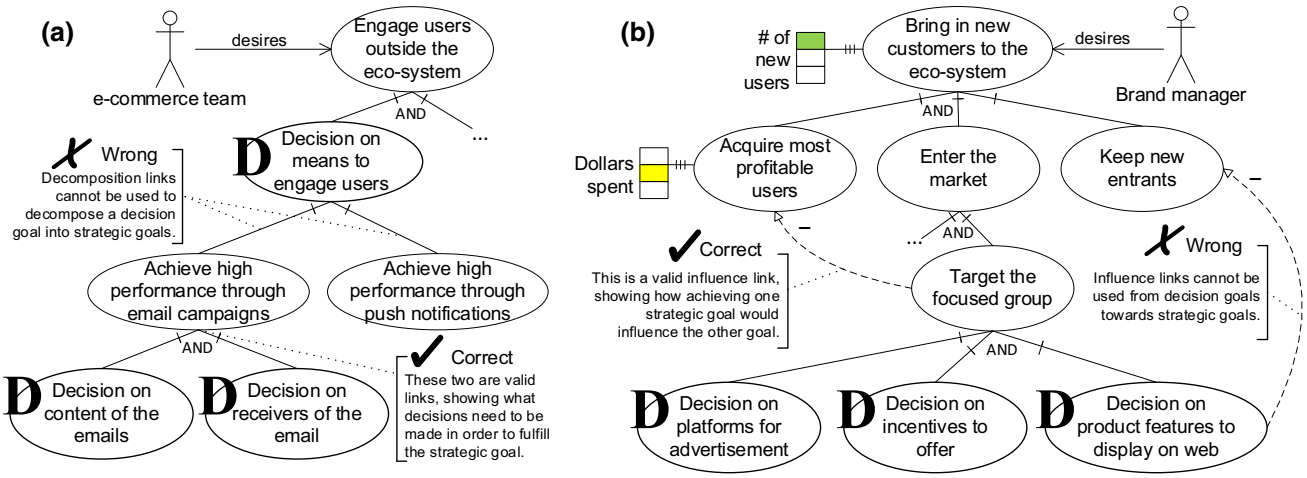


Fig. 6 Examples of some possible correct and wrong uses of decomposition links (**a**) and influence links (**b**)

allowed us to observe and record the difficulties that one would face during construction of conceptual models. Such observations were used to find and discuss limitations of the framework, and also to develop a set of guidelines whose examples were discussed in the previous section. These can be seen as part of the demonstration and evaluation activities of a design science cycle with some implications for the design and development step (Peffers et al. 2007).

Several factors can impact the validity of the findings and limit the generalizability of observations in this paper. *First,* while the testing of the framework was conducted initially by a participant who was not involved in the development of the framework, the authors subsequently assisted the participant in revising the models during several weekly meetings. The modeling was performed by the participant as part of an individual studies course supervised by one of the authors. The content of models were modified and syntactical issues were resolved during those meetings and after.

Second, the case studies in this paper did not involve any real business stakeholder(s). As a result, the findings in this paper are mostly reported in the form of *potentials* which need further validations. However, the involvement of the participant with some years of data science job experience helped us to make business questions and analytics solutions closer to reality. In addition, we tried to enrich the content of models by searching and reviewing multiple case studies and white papers.

Third, the benefits and limitations that were discussed are by no means comprehensive. The study involved only one participant and the findings in the paper mostly relate to only two (out of three) modeling views. We believe that there are more benefits and limitations associated with the framework that need further validations. For example, there are some expected benefits form the Data Preparation View (such as the re-use of the prepared data assets within enterprise) which we were not able to show in this paper. This was mainly due to lack of detailed-enough information on what data is being captured by business organizations, and how their data schema looks like. The Data Preparation View model in Fig. 2 was created mostly based on assumptions and examples obtained from the public domain.

7 Related Work

While modeling techniques have been proposed to assist in several areas related to the design of analytics solutions, we are not aware of any systematic framework that provides model-based support to connect all stages from goal-based requirements to analytics design to data preparation. We briefly review and compare related work in several areas.

7.1 Conceptual Modeling for Data Warehouses

Some works focus on modeling the requirements for data warehouses. Prakash and Gosain (2008) propose the goal-decision-information (GDI) model for analyzing data warehouse requirements. They develop a decision requirements metamodels (Prakash et al. 2010) and use informational scenarios (Prakash et al. 2004) to elicit data warehouse requirements. Giorgini et al. (2005) proposes a goal-oriented approach to requirement analysis of data warehouses, based on the Tropos methodology. Gosain and Bhati (2016) review the existing goal-oriented approaches for requirements phase of data warehouse development. The framework in this paper is different in the sense that it focuses on advanced analytics and machine learning solutions.

7.2 Conceptual Modeling for Business Intelligence (BI)

These works propose modeling approaches for developing BI solutions. The Business Intelligence Model (BIM) language represents enterprise in term of strategies, processes, indicators and more to bridge the gap between business and data (Horkoff et al. 2014). Barone et al. (2012) show usage of the BIM language for modeling the requirements of business intelligence system in healthcare domain. The framework in this paper extends the BIM language by introducing new concepts (such as Question Goals, Decision Goals, Insights, Algorithms, and Operators) and design catalogues to support development of advanced analytics solutions.

7.3 Data Mining Ontologies

Some works propose formal ontologies to support users during data mining and knowledge discovery processes (Ristoski and Paulheim 2016). Serban et al. (2013) provides a survey of intelligent assistants for the KDD (Knowledge Discovery in Databases) analysis process. Such ontologies do not capture concepts relevant to business requirement such as Actors, Strategic Goals, Softgoals, and Influences.

7.4 Information Systems Research on Analytics

Data analytics has increasingly attracted the interest of information systems (IS) research community (Agarwal and Dhar 2014; Abbasi et al. 2016). An important part of this body of literature focuses on the usage and impact of analytics on the organization and society. For example, Seddon et al. (2017) study the process (analyze–insight–decision–action) through which business analytics creates

business value. Sharma et al. (2014) provides a research agenda for understanding the relationship between business analytics, decision making processes, and organizational performance. These contributions are in terms of managerial principles and guidelines, towards theories. There is a lack of modeling approaches for analysis and design of data analytics solutions.

7.5 Existing Tools

A number of (commercial and open-source) software and platforms exist for performing analytics, including IBM Watson Analytics, Microsoft Azure ML, SAS, RapidMiner, etc. While they speed up the data preparation and experimentation with algorithms, they do not support business and requirements aspect of analytics solutions.

7.6 Data Mining Process Models

These models, such as CRISP-DM model, provide process models and methods for conducting data analytics projects. Mariscal et al. (2010) provide a survey and a comparison of such models. These works do not provide any modeling language for requirement analysis and design of analytics solution.

8 Conclusions and Future Work

Modeling offers effective ways to conceptualize, analyze, design, and develop information systems. Advanced analytics solutions, as an emerging and integral part of business information systems, have not taken advantage of such approaches. This paper proposed a modeling framework for requirements analysis and design of such systems. The framework consist of three modeling views and was presented through a sample methodology that describes how models are created in each view. The framework also includes a set of design catalogues to support the modeling. Using a case, we illustrated how the framework can support requirements elicitation, clarification and design aspects of business analytics solutions. Observations and findings from an application of the framework by a participant were presented and used to extend the framework with guidelines. Future work includes testing and improving the usefulness, usability, and learnability of the notation and method through empirical studies that involve real stakeholders. Such studies will serve as further validation and evaluation activities of design science research approach. We are also interested in investigating how the framework can be adapted as part of the process for designing off-the-shelf analytics tools. Moreover, we plan to develop tools that support construction of models as well as navigation and search through the catalogues.

Acknowledgements We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Abbasi A, Sarker S, Chiang RH (2016) Big data research in information systems: toward an inclusive research agenda. J Assoc Inf Syst 17(2):1–32

Agarwal R, Dhar V (2014) Editorial – big data, data science, and analytics: the opportunity and challenge for IS research. Inf Syst Res 25(3):443–448

Barone D, Topaloglou T, Mylopoulos J (2012) Business intelligence modeling in action: a hospital case study. In: Ralyté J et al (eds) 24th international conference on advanced information systems engineering. Springer, Heidelberg, pp 502–517

Bichler M, Heinzl A, van der Aalst WMP (2017) Business analytics and data science: once again? Bus Inf Syst Eng 59(2):77–79

Chandler N, Hostmann B, Rayner N, Herschel G (2011) Gartner's business analytics framework. https://www.gartner.com/imagesrv/summits/docs/na/business-intelligence/gartners_business_analytics__219420.pdf. Accessed 19 July 2018

Davenport TH, Bensoussan BE, Fleisher CS (2012) The complete guide to business analytics (collection). FT Press, Upper Saddle River

Giorgini P, Rizzi S, Garzetti M (2005) Goal-oriented requirement analysis for data warehouse design. In: Proceedings of the 8th ACM international workshop on data warehousing and OLAP, pp 47–56

Gosain A, Bhati R (2016) Goal oriented approaches in data warehouse requirements engineering: a review. In: Unal A et al (eds) Smart trends in information technology and computer communications: first international conference. Springer, Singapore, pp 244–253

Horkoff J et al (2014) Strategic business modeling: representation and reasoning. Softw Syst Model 13(3):1015–1041

Kandogan E, Balakrishnan A, Haber EM, Pierce JS (2014) From data to insight: work practices of analysts in the enterprise. IEEE Comput Graph Appl 34(5):42–50

Kohavi R, Mason L, Parekh R, Zheng Z (2004) Lessons and challenges from mining retail e-commerce data. Mach Learn 57(1–2):83–113

LaValle S, Hopkins MS, Lesser E, Shockley R, Kruschwitz N (2010) Analytics: the new path to value. MIT Sloan Manag Rev 52(1):1–25

Luca M, Kleinberg J, Mullainathan S (2016) Algorithms need managers, too. Harv Bus Rev 94(1):20

Manyika J et al (2011) Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute, New York

Mariscal G, Marban O, Fernandez C (2010) A survey of data mining and knowledge discovery process models and methodologies. Knowl Eng Rev 25(2):137–166

Nalchigar S, Yu E (2017) Conceptual modeling for business analytics: a framework and potential benefits. In: IEEE 19th conference on business informatics, vol 01, pp 369–378. https://doi.org/10.1016/j.datak.2018.04.006

Nalchigar S, Yu E (2018) Business-driven data analytics: a conceptual modeling framework. Data Knowl Eng. https://doi.org/10.1016/j.datak.2018.04.006

Nalchigar S, Yu E, Ramani R (2016) A conceptual modeling framework for business analytics. 35th international conference on conceptual modeling. Springer, Heidelberg, pp 35–49

Peffers K, Tuunanen T, Rothenberger MA, Chatterjee S (2007) A design science research methodology for information systems research. J Manag Inf Syst 24(3):45–77

Prakash N, Gosain A (2008) An approach to engineering the requirements of data warehouses. Requir Eng 13(1):49–72

Prakash N, Singh Y, Gosain A (2004) Informational scenarios for data warehouse requirements elicitation. In: Atzeni P et al (eds) Conceptual modeling – ER 2004, LNCS vol 3288, pp 205–216

Prakash N, Prakash D, Gupta D (2010) Decisions and decision requirements for data warehouse systems. CAiSE Forum 72:92–107

Ransbotham S, Kiron D, Prentice PK (2016) Beyond the hype: the hard work behind analytics success. MIT Sloan Management Review, March 2016

Ristoski P, Paulheim H (2016) Semantic web in data mining and knowledge discovery: a comprehensive survey. Web Semant Sci Serv Agents World Wide Web 36(Supplement C):1–22

Seddon PB, Constantinidis D, Tamm T, Dod H (2017) How does business analytics contribute to business value? Inf Syst J 27(3):237–269

Serban F, Vanschoren J, Kietz JU, Bernstein A (2013) A survey of intelligent assistants for data analysis. ACM Comput Surv 45(3):31:1–31:35

Shanks GG, Bekmamedova N, Willcocks LP (2012) Business analytics: enabling strategic alignment and organisational transformation. In: Proceeding of 20th European conference on information systems

Sharma R, Mithas S, Kankanhalli A (2014) Transforming decision-making processes: a research agenda for understanding the impact of business analytics on organisations. Europ J Inf Syst 23(4):433–441

Storey VC, Song IY (2017) Big data technologies and management: what conceptual modeling can do. Data Knowl Eng 108:50–67

Sullivan J (2014) Get the right data scientists asking the 'wrong' questions. Harv Bus Rev. https://hbr.org/2014/03/get-the-right-data-scientists-asking-the-wrong-questions. Accessed 19 July 2018

Yeomans M (2015) What every manager should know about machine learning. Harv Bus Rev 93(7)

Yu E (2011) Modelling strategic relationships for process reengineering. In: Yu E et al (eds) Social modeling for requirements engineering. MIT Press, Cambridge, pp 11–152