

Extending UML for Modeling Data Mining Projects (DM-UML)

Óscar Marbán and Javier Segovia*

Polytechnic University of Madrid, Montegancedo, Spain

Abstract

Existing Data Mining process models propose one way or another of developing projects in a structured manner, trying to reduce their complexity through effective project management. It is well-known in any engineering environment that one of the management tasks that helps to reduce project problems is systematic project documentation, but few of the existing Data Mining processes propose their documentation. Furthermore, these few remark the need of producing documentation at each phase as an input for the next, but they don't show how to do it. On the other hand, in the literature there are examples of UML extensions for data mining projects, but they always focus on the model implementation side and fail to take into account the remainder of the process. In this paper, we present an extension of the UML modeling language for data mining projects (DM-UML) covering all the documentation needs for a project conforming to a standard process, namely CRISP-DM, ranging from business understanding to deployment. We also show an example of a real application of the proposed DM-UML modeling. The result of this approach is that, besides the advantages of having a standardized way of producing the documentation, it clearly constitutes a very useful and transparent tool for modeling and connecting the business understanding or modeling phase with the remainder of the project right through to deployment, as well as a way of facilitating the communication with the nontechnical stakeholders involved in the project, problems which have always been an open question in data mining.

Keywords: Data mining; Knowledge discovery; KDD; UML profile

Introduction

In practice, data mining projects are approached in an unstructured, ad hoc manner, and results are very dependent on the skills of the person(s) doing the job and on the tools they use [1-5]. For this reason, most data mining projects are beset by common development problems, including trouble defining project objectives that are achievable with the available data, effort focused on the data preparation phase, experimentation with data parameters and transformation in the data mining phase, lack of an approach and methodological support for project development, project resource management problems [1,2,6].

Some of the data mining project development problems can be reduced through effective project management [1,7]. One of the management tasks that help to reduce data mining project problems is systematic project documentation [1,8,9]. This is the focus of this paper. Becker and Ghedini [1] propose the systematic documentation of previous knowledge, experiments, data and results. The sheer number of files, experiments and results make it hard, in practice, to manage all this documentation, leading individual project members to adopt their own documentation strategy. To exploit documentation to the full, it should be a corporate resource, usable by all project members and created based on standards defined within the business [1]. This leads to effective management, planning and communication [7].

Although DM process models propose one way or another of developing projects in a structured manner, few propose their documentation. CRISPDM [10] proposes documenting some of the data mining project tasks, but at no point states *how* this is to be done. While it does specify what elements (documents) each task outputs, it does not indicate their format or specific content. On the other hand, one of the integral project processes¹ proposed by Marban et al. [11], defining a process model for developing data mining projects, is project documentation. Being a process model, however, again it does not say how to document the project. In sum, these papers propose what to do, but not how to document the project, a question that is linked to the application of a methodology within a process and not the actual

process. Becker and Ghedini [1], on the other hand, propose how to document the data mining project textually using a data mining project documentation software support tool.

Unfortunately, this paper describes the software tool but not the actual documentation process, structure and organization. Project documentation can be developed using a modelling language [12,13].

A modelling language is any artificial language that can be used to express information or knowledge or systems in a structure that is defined by a consistent set of rules. The rules are used to interpret the meaning of the components in the structure [14].

UML (Unified Modelling Language) [15] is one of the most widely used languages for defining and specifying document systems [16]. To model systems with certain specific needs, UML can be extended in two different ways:

- UML extension by means of a profile providing the stereotypes, tagged values and constraints needed in order to specify the peculiarities of the modelled system.
- Extension of the Meta Object Facility (MOF) [17], the modelling language from which UML is defined.

UML is used in computing and other branches of engineering to define all sorts of systems. For example, UML extensions have been defined for:

- Documenting web site development [18]
- Building web-based remote monitoring and fault diagnosis systems [19]

*Corresponding author: Javier Segovia, Informática faculty, Polytechnic University of Madrid, Montegancedo Campus s / n. 28660 Boadilla del Monte (Madrid) Spain, E-mail: fsegovia@fi.upm.es

Received July 03, 2013; Accepted September 16, 2013; Published September 30, 2013

Citation: Marbán Ó, Segovia J (2013) Extending UML for Modeling Data Mining Projects (DM-UML). J Inform Tech Softw Eng 3: 121. doi:10.4172/2165-7866.1000121

Copyright: © 2013 Marbán Ó, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

¹An integral process is a process that supports the other defined processes. It should be developed throughout the project for each of the processes enacted during project development.

- Defining real-time systems [20]
- Representing knowledge bases [21]
- Modelling systems engineering [22]
- Modelling physical systems in industrial engineering [23]
- Modelling data warehouse database design [24,25]
- Modelling CRM systems [26,27]

While, with a view to modelling data mining project analysis and design, no modelling language that supports all the data mining phases has yet been developed, research covering some parts of the data mining process has been undertaken.

Data mining tools include workflows based on nodes or visual elements that are used to gather the knowledge buried in the data. These workflows describe the process followed to gather the knowledge from these data, albeit confined to part corresponding to the implementation of the data mining model. Figure 1 shows an example of such a workflow for the Weka tool [28]. This workflow represents the process of building a data mining model, which is equivalent to implementing software application code.

Apart from the fact that they only represent part of the project implementation, the main problem with these workflows is that they are tool dependent. Even so, the generated models can be shared by tools supporting PMML [29]. PMML is an XML-based language for describing data mining models, again providing a language for representing data mining project implementation elements. PMML offers a standard means of defining data mining models, originally created for tool-to-tool data mining model exchange.

In the literature, there are examples of UML extensions for data mining projects. They always focus on the model implementation side and fail to take into account the remainder of the process. Accordingly, Zubcoff and Trujillo [30] present a UML profile for modelling association rules from data stored in a data warehouse. Xu and colleagues [31] present a UML profile for representing association rules for social networking data. Rizzi [32] also defines UML models for representing association rules. Zubcoff and Trujillo propose a UML profile for the conceptual representation of classification models [33] and clustering models [30]. In both cases, the data are assumed to be stored in a data warehouse.

Existing UML extensions for data mining propose artefacts for modelling some data mining project elements and results but not all of the important parts of the project. To mention but a few, they omit business, project requirements or problem analysis modelling. The main aim of this paper is to propose a profile-based UML extension that is usable in all the data mining project phases. This extension is based on CRISP-DM [10], as the *de facto* standard for data mining project development, and a broader process model for developing data mining projects that covers the phases omitted by CRISP-DM [11]. Whereas these and other process models propose what tasks to undertake and, in some cases, when to undertake them, they do not say how to model and document the task outputs, outputs that are inputs to other tasks. The conclusion is that elements and tools need to be provided to help to build a bridge between the inputs and outputs of each phase, a bridge that should be assembled by properly documenting the project.

Section 2 outlines the proposal (DM-UML or Data Mining Unified Modelling Language) based on how UML is used in other project types, exploiting the likenesses between some DM process phases and other projects. Section 4 presents a sample case study.

UML for Data Mining

As mentioned in the last section, although data mining project development process models and methodologies do state that project development should be documented, they do not say how to go about documentation. In this section, we present a way of using UML extension mechanisms to produce the technical documentation of the development project. These extension mechanisms will define the stereotypes required to add the data mining elements and thereby provide elements for developing the technical data mining project documentation. We will call this DM-UML (Data Mining Unified (Modelling Language)).

DM-UML models

The DM-UML elements can be taken directly from the UML 2.x definition [17] or can be added using extensions that UML 2.x provides. To define DM-UML we are going to use UML profiles as a language extension mechanism. The diagram in Figure 2 shows the models that DM-UML will include and their traceability. For example, the data mining use case model is obtained from the business goal model, and the data mining use case model shows how this model can be traced to the original business goal model.

The DM-UML models towards the top of Figure 2 are more closely related to the organization. The further down the models are in the hierarchy, the nearer they are to the data mining algorithms and tools used in the project. Traceability indicates that a change in a model will cause changes in the derived model or models. The dashed directed

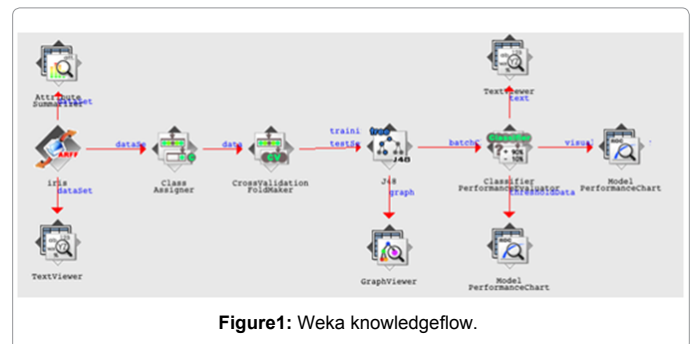


Figure1: Weka knowledgeflow.

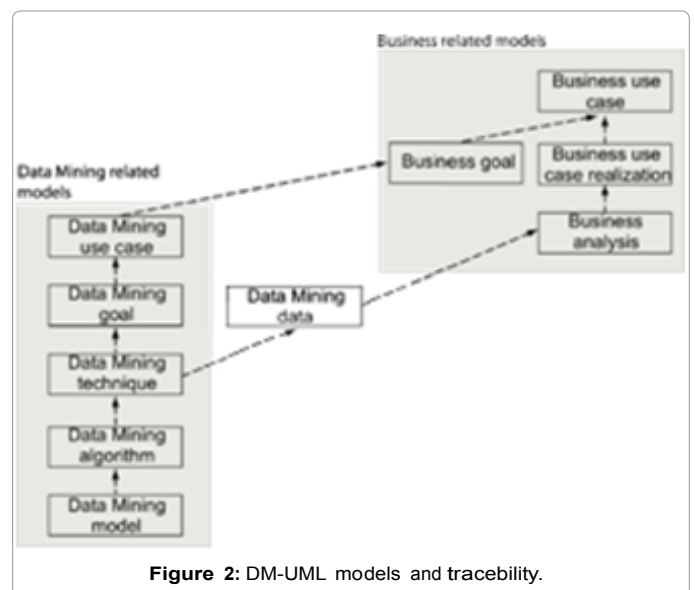


Figure 2: DM-UML models and traceability.

lines in Figure 2 show this traceability. This way, a change in the specification of the business uses case model affects how the business goal is specified.

As shown in Figure 2, a distinction can be made between two types of DMUML models. On the one hand, there are the models serving to represent the business, i.e. business-related models, and, on the other, the models proper to data mining, i.e. data mining-related models. The business-related models used in DM-UML are: business use case model, business use case realization model, business goal model and business analysis model. Business use cases are used to capture business requirements, and define a set of actions that a business performs to yield an observable result of value to a particular business actor. Business use case realization describes how business workers, business entities, and business events collaborate to perform a particular business use case. The business analysis model describes the realization of business use cases by interacting business workers and business entities. It is useful as an abstraction of how business workers and business entities need to be related and how they need to collaborate in order to perform the business use cases. Finally, a business goal is a requirement that must be satisfied by the business. Business goals describe the desired value of a particular measure at some future point in time and can therefore be used to plan and manage the activities of the business.

Existing UML models can be used to model the business [17] by adapting to the goals of the related data mining process phases and will not, therefore, be described in this paper. For a description of the elements of these models and the elements common to all UML models (notes, dependencies, relations and generalizations) [17, 34].

The business models and data mining models are linked by the data model or data mining data model. The first data model will be built from the data available in the organization for developing the project based on the business model. Later, this model will be refined and adapted to the data mining project needs (integration, derivation, data processing, inclusion of other data sources, etc.).

Additionally, we define the data mining models that are used for the technical part of the project: data mining use case model, data mining goal model, data mining technique model, data mining algorithm model and data mining models model. All the DM-UML models proper to data mining are new, they are the key objective of the paper and will be defined in the next section. To do this, we will make use of the UML extension mechanisms called “profiles” [35].

Data mining models

In this section we are going to define the DM-UML data mining models: data mining use case model, data mining goal model, data mining data model, data mining technique model, data mining algorithm model, data mining models model, as well as the elements required to be able to specify the data mining problem to be addressed.

The data mining use case is the foundation for the DM-UML models. The data mining use case describes the proposed functionality of the knowledge extracted by the data mining tasks as seen by the user in order to achieve a particular business goal. Potential users are company salespeople wanting to use the results of a data mining project to improve their performance, or the company’s CEO who wants to analyse in-house data to develop a new company strategy.

A data mining use case represents a discrete unit of interaction between a user and the knowledge. Different data mining use cases applied to the same information represent different ways of using or extracting knowledge for different or the same business goals. A data

mining use case is a single unit of meaningful work; for example, rank products based on their properties or sales, or develop a customer profile are both data mining use cases. Each data mining use case has a description that describes its functionality. A data mining use case may *include* another data mining use case’s functionality or *extend* another use case with its own behaviour. Data mining use cases are usually related to *actors*. An actor is typically a human that interacts with the knowledge to perform meaningful work, such as the CEO or the salespeople mentioned above, or the final consumer that buys a product from the company.

Data mining use case model: Apart from the data mining actors and use cases, such diagram-based models also include the data mining goals described in the next section. We use dependencies (dashed directed lines) to relate each of the elements appearing in these diagrams, as a change in the definition of any of the model elements will be propagated to the elements to which it is related. Table 1 shows the elements used in the data mining use cases. [36] presents the formal definition of these elements according to the profile-based UML extensions.

Data mining use cases are built from use cases and business goals. In principle, we will have a data mining use case for every combination of business use case and business goal. Some business goals are not specific and measurable enough to find supporting data mining use cases. These are typically strategic goals that need to be defined at less abstract levels. This definition will result in a hierarchy of business goals, where business goals must be traced from higher to lower level goals to produce a business goal hierarchy. As data mining use cases are created, we should look at whether the business goal to which it is related can be directly evaluated in business terms or whether the business use case has to be evaluated through another business goal appearing at higher levels of the business goal hierarchy.

Whether or not there is a data mining use case will depend on whether the business use case handles data that can be analysed and whether such an analysis can be used somehow to achieve the business goal. It is not uncommon either that the need and possibility of gathering data for data mining is detected within a business use case that does not handle sufficient data for data mining and collection is included as part of the project. The relation between business goals and data mining use cases (see description in [36]) is represented as a dependency (dashed directed line, see [36]) in the diagrams between the use case and the business goal, as shown in Figure 3.

The data mining use cases output knowledge. If knowledge is going to be used directly by an actor, the use case is directly associated with the actor (solid line, see [36]), as shown in Figure 4.

If on the other hand, it is going to be delivered as a document or integrated into an application, the knowledge is related to a document or data mining application, and these elements are related to the data mining actor, as shown in Figure 5.

Additionally, each data mining use case will have to have one or more associated data mining goals. The data mining goals will be used as a reference point for defining the validation of knowledge gathered from the data mining use cases. The different data mining elements to be built (data mining techniques, algorithms and models) will be gathered from the data mining goals.

Data mining goal model: The data mining goal model represents the data mining project requirements, that is, what is expected to be gained from the data mining project in terms of knowledge rather than business. Identifying a customer typology/profile, establishing



Figure 3: Business goal and data mining use case traceability.

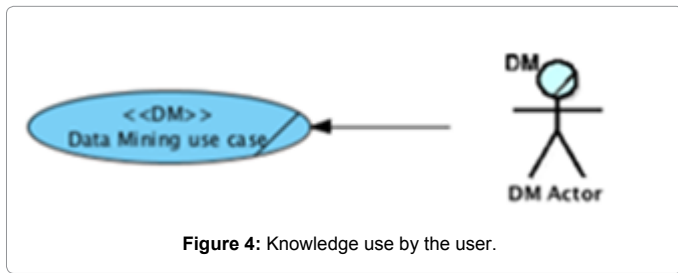


Figure 4: Knowledge use by the user.

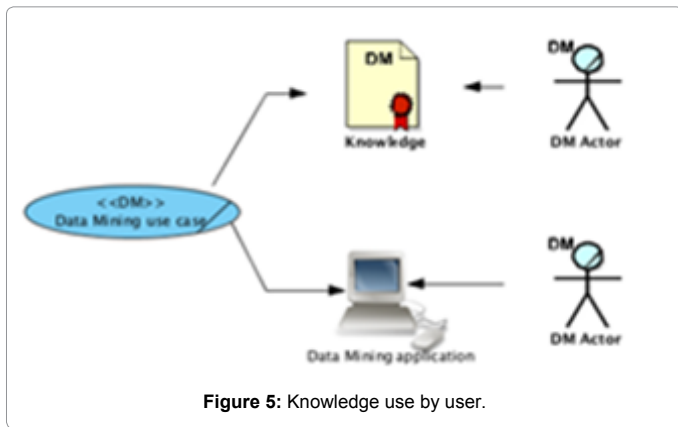


Figure 5: Knowledge use by user.

the commonest shopping basket associations or creating a descriptive model of the behaviour of a web user are all data mining goals. The data mining use case model explains how these data mining goals are used by actors to achieve business goals in a business use case.

The diagram of the data mining goal model will show all the data mining goals, depicting the data mining goal generalization hierarchies, if any, and their relationship to the use cases from which they derive. The data mining goal model elements are shown in Table 2. The same descriptions apply to these elements as in Table 1.

In [36] we present the formal definition of these elements according to profile-based UML extensions.

These elements are related by dependencies as shown in Figure 6. Each data mining use case has to have at least one associated data-mining goal, and several data mining use cases can tackle one data-mining objective.

Data mining data model: The data mining data model represents the sources of the available data for the project, with tables, columns, data types and data relations. This model is based directly on the UML definition for data models, but has been adapted by means of stereotypes to tailor the data models to the needs of data mining projects (data integration, transformation and derivation). This model represents the physical data model, that is, the structures to be stored in the data source. Table 3 shows the elements that appear in these diagrams. In [36] we present the formal UML profile-based definition of the elements shown in Table 3 and not previously defined in UML.

Data mining technique model: Data mining technique diagrams show the data mining techniques used to be able to achieve the data mining goals proposed in the data mining use cases. These diagrams show the data mining techniques and their possible input data related to the data mining goals that they achieve or help to achieve. They may also show the data sources that they use. Table 4 describes the elements used to create this model.

In [36] we present presents the UML profile-based formal definition of the elements listed in table 4 and not previously defined in UML. For example, Figure 7 represents the data mining technique to be applied to output a particular business goal.

There exist in the literature pre-defined UML profiles for some data mining techniques. For example, UML profiles for association were defined in [16,31], for classification in [33] and for clustering in [16]. To build these models, then, we will have a choice between using either the universal form defined in this paper or the pre-defined profiles, if any, for the data mining

Data mining algorithm model: The data mining algorithm model shows the data mining algorithms to be used to solve the problem. The algorithms to be used can be implemented in the data mining tool that is used in the project or can be developed ad hoc. Apart from representing the algorithms, these diagrams may also include the available data sources from where the data are to be taken. As the algorithms are directly derived from the data mining techniques, the data, if included, must match the data that appear, if any, in the respective data mining techniques diagram.

In [36] we present the UML profile-based formal definition of the elements shown in Table 5 and not previously defined in UML.

Figure 8 shows the alternative representations of the data mining

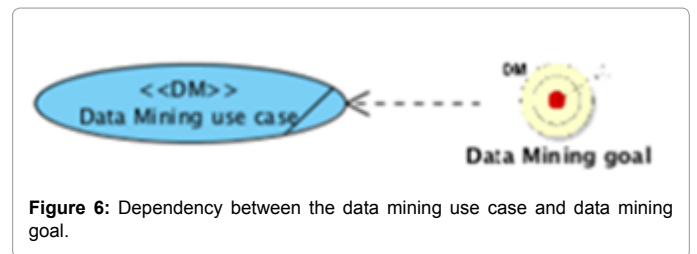


Figure 6: Dependency between the data mining use case and data mining goal.

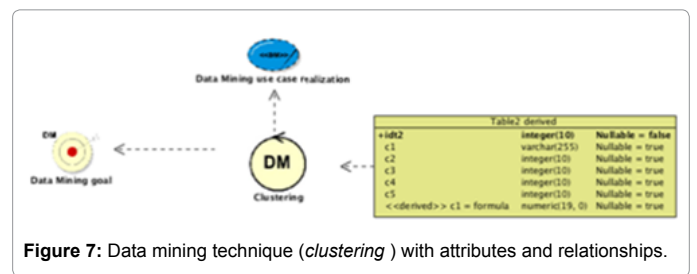


Figure 7: Data mining technique (clustering) with attributes and relationships.

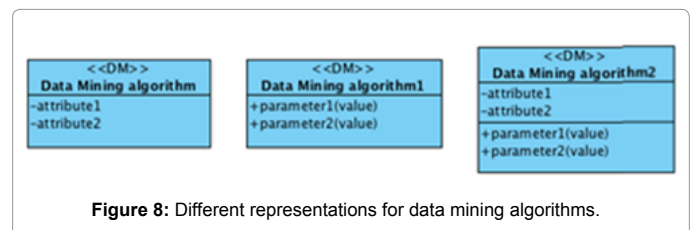


Figure 8: Different representations for data mining algorithms.



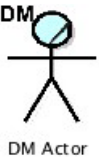
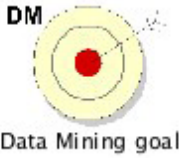



| DM-UML representation | Element | Description |
|---|----------------------------------|---|
|  Data Mining use case | Data mining use case | The data mining use case represents the output expected by the user from the viewpoint of data mining, i.e. what the user expects to be able to do with the knowledge gathered by the data mining system under development. Data mining use cases can be accompanied by a textual description or UML notes. |
|  Data Mining use case realization | Data mining use case realization | The data mining use case realization indicates how the system performs the data mining use case, i.e. describes how the system will gather and use the knowledge in the data mining use case. This element bears the same name as the data mining use case it performs. |
|  DM Actor | Data mining actor | A data mining actor represents the end user of the knowledge gathered from the data mining use case or cases related to the respective actor. |
|  Data Mining goal | Data mining goal | The data mining goal is a data mining requirement that the system must meet to add value to the knowledge gathered by the system using a use case. Data mining goals can be accompanied by a textual description or a UML note. |
|  Data Mining application | Data mining application | The data mining application represents the result of the data mining use case as a software application that makes use of the knowledge gathered in the data mining use case. |
|  Data Mining document | Data mining document | Data mining document represents the result of the data mining use case as a document containing the knowledge gathered, as either a list, interpreted knowledge, etc. |
|  Business goal | Business goal | A business goal is a requirement that the business must satisfy. (Defined in UML, see [MS09]) |

Table 1: DM-UML elements for data mining use case model.



| DM-UML representation | Element | Description |
|---|----------------------|-------------|
|  Data Mining use case | Data mining use case | See table 1 |
|  Data Mining goal | Data mining goal | See table 1 |

Table 2: DM-UML elements for data mining goal model.

final dataset from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order.

- **Modeling**

In this phase, various modeling techniques are selected and applied

and their parameters are calibrated to optimal values. Typically, there are several techniques for the same DM problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often necessary


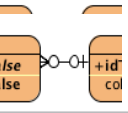
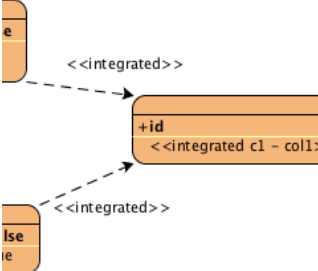

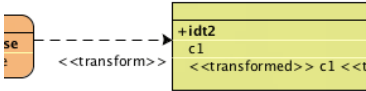
| DM-UML representation | Element | Description |
|---|--|--|
|  <pre><<Kind:vendor:version:location:user:password>> Data Source</pre> | Data source | Data source represents where the available project information is stored. It includes type, manufacturer, release and access mode. |
| <pre>Data table + column1 integer(10) Nullable = false column2 varchar(255) Nullable = true column3 date Nullable = false</pre> | Data table | Data table represents the tables in the data source, showing the columns, types and primary key (marked with +). |
|  | Data relationship | Data relationship represents the relationships between the different tables. Possible relationships are 0:n, 1:n, 0:1, 1:1. |
|  | Integration | If there is more than one data source, the data have to be integrated to prevent errors, like information duplication or inconsistent values. The integration element shows how they have been integrated. The supplementary documentation can discuss, unless obvious, what type of integration or transformation was carried out to output the target table. |
|  | Derived data | New data output from the original data in the data sources. This shows the formula for deriving new columns. The name of the table containing the new columns is the same as the original table, plus the word "derived". |
|  | Transformed data | Transformed data represents the change of format of some of the data to apply a specific data mining algorithm that calls for an alternative data format. The name of the table containing the transformed columns is the same as the original table, plus the word "transformed". |
| <pre>Table2 derived +id2 integer(10) Nullable = false c1 varchar(255) Nullable = true c2 integer(10) Nullable = true c3 integer(10) Nullable = true c4 integer(10) Nullable = true c5 integer(10) Nullable = true <<derived>> c1 = formula numeric(19, 0) Nullable = true</pre> | Modified data table (Derived data or transformed data) | Modified data table represents a set of data from a data derivation or transformation, specifying the fields that have been derived or transformed. |

Table 3: DM-UML elements for the data mining data model.

algorithm shown in Table 5. A data mining algorithm can also be represented by the input data it uses, the algorithm parameters and values, or both (data and parameters).

Data mining models model: This model shows which data mining models will be built and where they are stored (files) in the data mining tool used in the project. This should enable traceability, first, among the data mining algorithms and, second, among the models and files where they are stored in the data mining tool (strictly speaking, model work spaces and files). Additionally, the definition of the files will later enable the definition of the configuration elements that will take part in project configuration management. [36] presents the UML profile-based formal definition of the elements shown in Table 6 and not previously defined in UML. By way of an example, Figure 9 shows how a data mining model is related to the respective files in the data mining tool used.

CRISP-DM and DM-UML

CRISP-DM defines the phases that we have to do in a DM project.

CRISP-DM also defines for each phase the tasks and the deliverables for each task. CRISP-DM is divided in six phases (Figure 10). The phases are described in the following.

- **Business understanding**

This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a DM problem definition and a preliminary plan designed to achieve the objectives.

- **Data understanding**

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

- **Data preparation**

The data preparation phase covers all activities to construct the

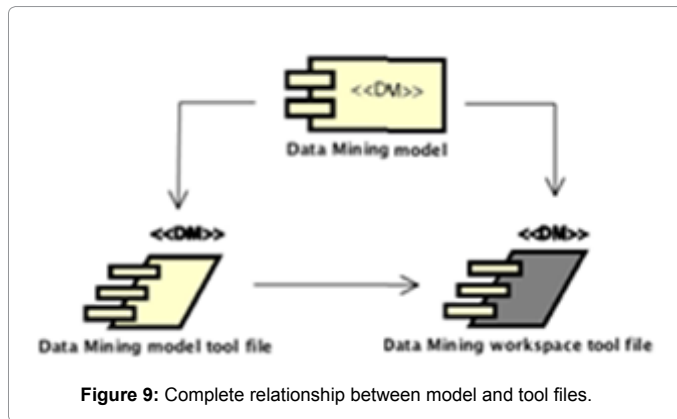


Figure 9: Complete relationship between model and tool files.

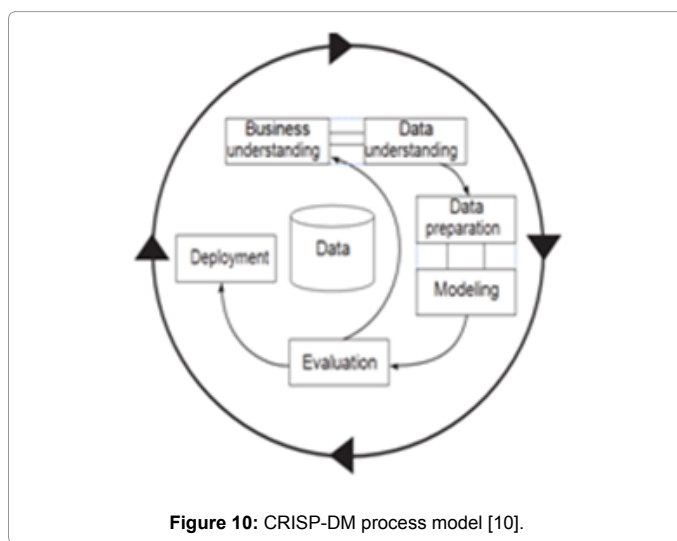


Figure 10: CRISP-DM process model [10].

- **Evaluation**

At this stage of the project a model (or models) will have been built that are of seemingly high quality from a data analysis perspective. Before proceeding to final model deployment, it is important to evaluate the model more thoroughly and review the steps taken to construct the model to be certain that it properly achieves the business objectives. At the end of this phase, a decision on the use of the DM results should be reached.

- **Deployment**

Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it.

We have chosen CRISP-DM because it is the “facto standard” to develop DM projects. Although, DM-UML could be used with any data mining methodology. Figure 11 shows CRISP-DM phases and which DM-UML models we can use in each phase.

In deeper detail, we can assign DM-UML models to CRISP-DM tasks (see Table 7).

Table 7 presents an outline of phases and generic tasks that CRISP-DM proposes to develop a DM project.

- **Business understanding**

Determine business objectives: To determine business objectives

we need to model the business. We could use the business use case, business use case realization, business analysis and business goal models to represent the main areas of the business in which data mining will be applied.

Determine data mining goals: Using the *business use case* and *business goal* models as input we could build the Data Mining use case, Data Mining goals models to obtain the data mining goals and success criteria of the data mining project.

- **Data understanding and Data preparation:** We propose the use of Data Mining *data* model to document the source data of the project. We could use Data Mining data model in each subtask of Data understanding or Data preparation tasks to model the transformations in the data.

- **Modeling**

Select modeling technique: We could use Data Mining use case and Data Mining goal models to obtain the Data Mining technique model to document the select modeling technique subtask of CRISP-DM.

Build model: This subtask could be document through Data Mining algorithm and *Data Mining model* models. Data Mining technique model should be used as input for developing the Data Mining algorithm model

- **Evaluation**

Evaluate results: The business use case, business goal, Data Mining goal, Data Mining use case and *Data Mining use case* models could be used to check project results.

- **Deployment**

In this phase we could not use any of the DM-UML models. If a software with embedded knowledge obtained through the data mining project has to be developed, software engineering techniques could be used.

Case Study

This section presents an example of the use of the UML extension

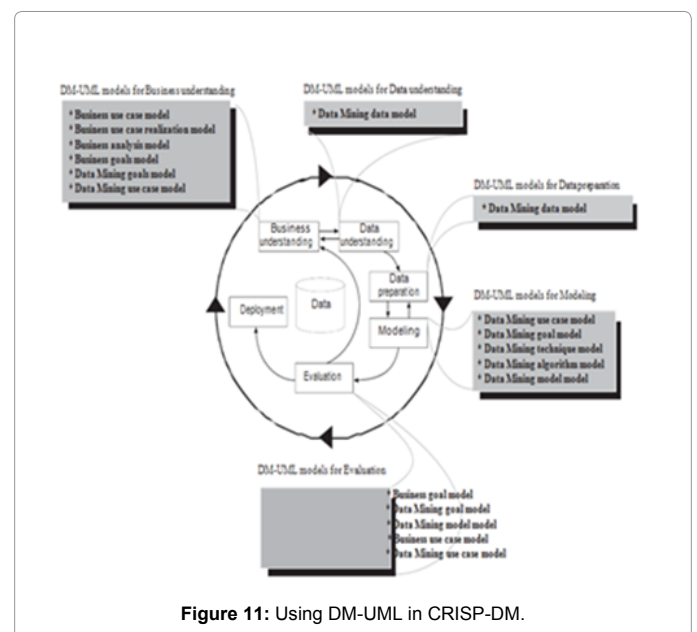


Figure 11: Using DM-UML in CRISP-DM.


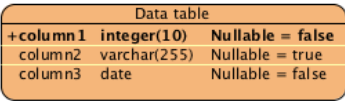
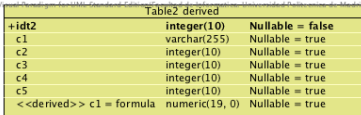

| DM-UML representation | Element | Description |
|---|--|---|
|  <p>Data Mining technique</p> | Data mining technique | The data mining technique element represents the data mining technique (clustering, neural networks, association, etc.) to be used to solve a particular problem in the data mining use case. It may also include the data on which the data mining technique is to be applied. |
|  <pre> Data table +column1 integer(10) Nullable = false column2 varchar(255) Nullable = true column3 date Nullable = false </pre> | Data table | See table 3 |
|  <pre> Table2 derived +id2 integer(10) Nullable = false c1 varchar(255) Nullable = true c2 integer(10) Nullable = true c3 integer(10) Nullable = true c4 integer(10) Nullable = true c5 integer(10) Nullable = true <<derived>> c1 = formula numeric(19, 0) Nullable = true </pre> | Modified data table (Derived data or transformed data) | See table 3 |
|  <p>Data Mining use case realization</p> | Data mining use case realization | See table 1 |

Table 4: DM-UML elements for the data mining technique model.



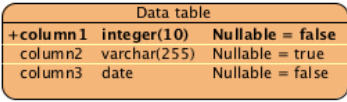
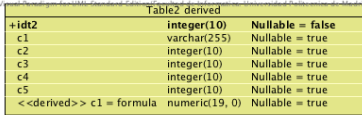
| DM-UML representation | Element | Description |
|---|----------------------------------|--|
|  <p>Data Mining algorithm</p> | Data mining algorithm | The data mining algorithm represents the elements to be used to build the data mining models that will output the knowledge. These elements really represent what the data mining tools (Clementine, Weka, etc.) are to use to gather knowledge. |
|  <p>Data Mining technique</p> | Data mining technique | See table 4 |
|  <pre> Data table +column1 integer(10) Nullable = false column2 varchar(255) Nullable = true column3 date Nullable = false </pre> | Data table | See table 3 |
|  <pre> Table2 derived +id2 integer(10) Nullable = false c1 varchar(255) Nullable = true c2 integer(10) Nullable = true c3 integer(10) Nullable = true c4 integer(10) Nullable = true c5 integer(10) Nullable = true <<derived>> c1 = formula numeric(19, 0) Nullable = true </pre> | Derived data or transformed data | See table 3 |

Table 5: DM-UML elements for the data mining algorithm model.

for data mining (DM-UML) presented in section 2. For the standard UML notation and business modelling notation, see [36].

For reasons of space, the example, which is a real project, is not developed in its entirety. It does, however, perfectly illustrate how DM-UML works as a communications bridge between all the project phases, constituting a suitable vehicle for documenting the project not only internally for project developers but also for project customers. It is especially interesting to see how phases with a business focus and KDD phases are connected. This is one of the major problems existing in DM projects [37-41,11].

Problem description

The example is based on a real project² conducted for a car brand, aiming, among other things, to increase the number of vehicle sales

and maximize customer profit throughout their life cycle as brand customers. The business description is usually given at a preliminary meeting with the customer. This meeting may have taken place at the customer's request or at the initiative of the data mining company to examine the possibility of offering a project to improve the customer's business.

The informal description is summarized below (including only the part of the business of interest for this problem). This is the description that will be used later for modelling.

The company has several ways of displaying products for potential customers to view. One is the traditional retail channel, where

²On the grounds of confidentiality, all references to or any elements identifying the company or brand name have been removed.





| ML representation | Element | Description |
|---|---------------------------------|--|
|  | Data mining algorithm | See table 5 |
|  | Data mining model | Data mining models represent the result of executing the selected data mining algorithm. Normally, the data mining tools store these models in files, which is what this element would represent. They are related to their source data mining algorithm by a traceability relationship. Additionally, this element can be documented with the information on the tool used to generate the model. |
|  | Data mining workspace tool file | To assure file traceability, the files generated by the data mining tool to create the data models also have to be referenced, specifying the name (and possibly the path) of the file containing the model. |
|  | Data mining model tool file | This element represents the generated model if it can be saved by the tool for later use. |

Table 6: DM-UML elements for data mining models model.

| Business understanding | Data understanding | Data preparation | Modeling | Evaluation | Deployment |
|-------------------------------|----------------------|------------------|----------------------------|----------------------|---------------------------------|
| Determine business objectives | Collect initial data | Select data | Select modeling techniques | Evaluate results | Plan deployment |
| Assess situation | Describe data | Clean data | Generate test design | Review process | Plan monitoring and maintenance |
| Determine DM objectives | Explore data | Construct data | Build model | Determine next steps | Produce final report |
| Produce project plan | Verify data quality | Integrate data | Assess model | | Review project |
| | | Format data | | | |

Table 7: CRISP-DM phases and tasks.

customers visit a dealer to take a look at the vehicles on show and, if they see a model they like, ask the dealer for additional information about the vehicle.

Another way of attracting customers is over the Internet. Specifically, the brand has a web portal where potential customers can virtually visualize the vehicles that are most likely to be of interest them and get detailed information about each one. Optionally they are invited to enter information about themselves and their interests (personal particulars like age or postcode, information about the vehicle they now own, how they learned out about the portal, etc.). The information about the actions taken by potential customers is recorded in a web log and can be studied.

On the other hand, the brand runs marketing campaigns through newspaper and radio advertising to publicize their vehicles or incentivize the purchase of a particular model. The result of these campaigns is quantified by salespeople asking customers visiting a dealer to ask for information about a model where they learned about the vehicle. The brand has two business goals: one is to increase the sale of new vehicles and vehicles in stock and the other is to maximize the profit earned on each customer individually either by offering vehicles with personalized options or by assuring that, when they change their car in the future, they buy the same brand again.

Business models

Business use case model: This phase aims to identify what parts of

the business are to be improved (business use cases) and what elements outside the business are involved (business actors). Analysing the problem description, we get the elements shown in Figure 12 (Figure 12 does not represent the entire business). The business actors are:

- “Customer”, which is a customer that has already purchased at least one vehicle or a potential customer
- “Business analyst”, which is the person or group that analyses what is going on in the business and decides what actions should be taken to achieve the business goals set by the company
- “Advertising agency”, which represents the advertising agency responsible for creating the advertising campaigns in the respective media according to the guidelines set by the business analyst.

The business use cases are divided into two groups, namely core and support use cases (see [36] for a description of the different business use case types).

- **Core business use cases:**

“Visit website”, which represents the action of displaying the brand catalogue on a website and viewing by a customer. When customers visit the website, customer data are gathered with customers’ consent as are data on the actions they take on the website.

“Sell a car”, which represents the action of selling a car at a dealer.

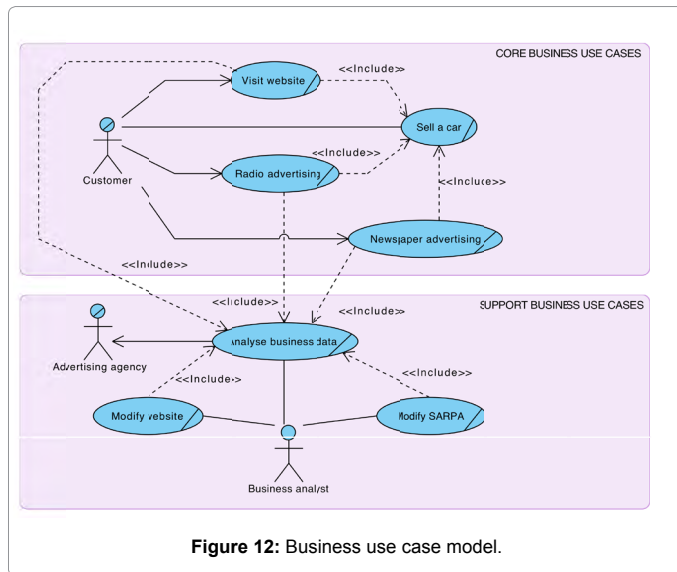


Figure 12: Business use case model.

“Newspaper advertising” and “Radio advertising”, which represents the actions of the brand advertising in the mass media (press or radio) and a potential customer receiving this advertising .

• **Support business use cases:**

“Analyse business data”, which represents the analysis of the available business data in the company to make the right decisions and manage to achieve the business goals.

“Modify website” represents the action of modifying the brand website according to the actions decided on by the business analyst after data analysis.

“Modify SARPA” represents the action of modifying the SARPA system that supports sellers with vehicle sales. This system is modified according to the recommendations decided on by the business analyst after data analysis.

Figure 12 uses lines to illustrate the relationship between the elements as described in [36]. The <<include>> dependency of the “Visit website”, “Radio advertising” and “Newspaper advertising” business use cases on the “Sell a car” business use case represents the fact that, once those use cases are complete, the “Sell a car” use case could also be executed as a result. For example, a potential customer visits the website, views the models and decides to buy one, thus completing the “Visit website” business use case. Then the customer will visit the dealer to buy the vehicle, thereby completing the “Sell a car” business use case. The “Sell a car” use case includes the “Visit website” business use case through the <<include>> dependency. The same applies to the <<include>> dependencies of the “Modify website” and “Modify SARPA” business use cases on the “Analyse business data” business use case.

Business goal model: Business use cases are justified if they can be associated with a business goal (build something, increase profits, improve productivity, reduce expenses, improve brand image, etc.), meaning that they have to be assigned to one or more business goals defined by the company. The rule is that each business use case must have at least one business goal, as, otherwise, it is irrelevant to the problem. The brand’s major business goal is to increase profits. This business goal has been divided into secondary goals, with the aim of increasing profits:

- **Increase Sales:** Increase the sale of new vehicles or vehicles in stock.
- **Increase Profits per customer:** Increase the profits earned per customer from both viewpoints: for sales throughout their life cycle as brand customers or a one-off vehicle sale with customized options³.

As Figure 13⁴ shows, the business goals were organized as a generalization [36]) where the overall goal is to “increase profits”. The other two are specific goals, each of which helps to achieve the overall goal. These specific goals can be further decomposed into much more specific goals, as shown in Figure 13.

Just as the business use cases are divided into core, support and management business use cases, the associated business goals will match the core, support and management business goals depending on the type of business use case with which they are associated. In this case study, the support and management business goals have to be associated, through the hierarchy, with a core business goal. Core business goals are the goals with business value.

As mentioned above, each business goal is associated with one or more business use cases depending on its characteristics, and each business use case is associated with at least one business goal. For example, as shown in Figure 13, the “Newspaper advertising” and “Radio advertising” use cases have the same business goals, the “Analyse business data” use case has multiple business goals, whereas the “Visit website” use case has only one business goal.

Figure 13 also shows the relationship between the business use cases and the business goals. Clearly, the union between these elements is a dependency, which means that the business use case depends on the business goal. Therefore any change in the business goal is likely to cause changes in the business use case. For example, the contents of radio adverts could be planned, designed and prepared differently if their goal were to change to “Sell more vehicles”.

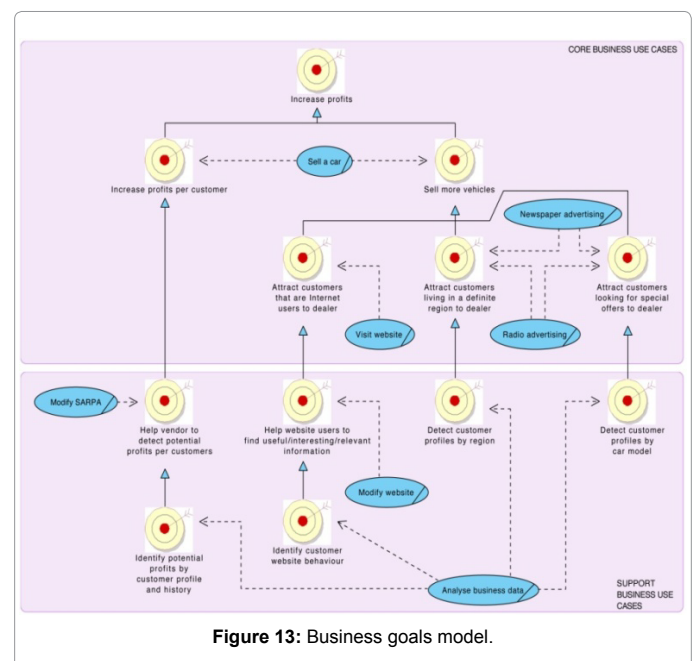


Figure 13: Business goals model.

³For simplicity’s sake, we opted not to further divide this subgoal, although it could have been split into two.

⁴The real business goals and interrelations have been simplified.

The data mining use cases will be derived from the business use cases and business goals as shown later. The business analysis diagrams are also derived from the business use case diagrams.

Business analysis model: Business analysis shows how different business elements interact to achieve a particular goal. This model includes the roles that business employees or business workers⁵ play. It also contains the business entities, which represent the objects that workers access, inspect, manipulate, produce, and so on. Entity objects provide the basis for sharing among workers participating in different use case realizations. Business entities range from abstract things, like data about a customer or product, or physical things, like a computer system or product. Finally, it also includes the generated business documents, such as a signed contract or sales invoice.

The business workers in the example are:

- “Business analyst”, who has responsibility for investigating business systems, identifying options for improving business systems and bridging the needs of the business with the use of IT.
- “Website”, which represents the website where the company displays its models to potential customers, and acts as a virtual dealer.
- “Vendor”, which is the salesperson that physically serves customers at a dealer.
- “SARPA”, a vehicle sales support computer system that salespeople use.

Business entities are the data that the company stores for use by the business workers (“Customer”, “Sale”, “Vehicle”), and any information generated by the Internet portal (“Weblog”, “Virtual Order”) or the seller-operated SARPA computer system (“Customer”, “”, “Sale”, “Vehicle”). On the other hand, the business documents are: advertising guidelines (“Guidelines for radio advertising” and “Guidelines for newspaper advertising”), guidelines for modifying systems (“Guidelines for SARPA modification” and “Guidelines for website modification”) and the vehicle sale invoice (“Invoice”). The potential data elements for use in the data mining project are gathered from the business entities that appear in the analysis diagrams (see Figure 14). Additionally, the business analysis model includes an organization unit element (IT department). This element represents the department within the organization responsible for implementing or modifying the web and SARPA.

Figure 14 shows the analysis of the seven business use cases from figure To build these diagrams we have to interview the business managers and analyse their work methods, as these diagrams represent the real operating procedures in the business. Below we detail the business use case analysis shown in Figure 14a-14g).

- **Newspaper advertising:** Before placing a newspaper advert, the business analyst examines the information available within the business regarding customers, sales and vehicles (Figure 14a). After studying the information, the business analyst creates the guidelines for newspaper advertising. This matches the Analyse Business Data business use case that is included in the Newspaper Advertising use case. Then the advertising guidelines are delivered to the advertising agency that creates the respective advert and publishes it in the press. The advert is seen by potential customers (customer) and the use case ends. If a potential customer decides to buy the vehicle in question

he or she will either visit the brand website, executing the Visit Website use case (Figure 14c), or visit a brand dealer, executing the Sell a Car use case (Figure 14d).

- **Radio advertising:** Exactly the same as the Newspaper Advertising use case, except that, in this case, the advert is placed in the radio medium (radio) (Figure 14b).
- **Visit website:** To set up the business website the business analyst examines the information available within the business about the customers, sales and vehicles, plus the user navigation data. Having studied the information, the business analyst will generate the website creation/modification guidelines that will be delivered to the IT department for it to create/improve the brand website, thereby executing the Modify Website use case. Potential customers (customer) will then visit the brand website (website) off their own bat or after having seen or heard a brand advert in the newspapers or on the radio. This website displays all the brand models (vehicle) and their technical characteristics. Customers that are interested in a particular model are given the option of configuring and viewing the model (engine, paint, wheels, finish, optional equipment,...) and saving the model in the “virtual dealer” (virtual order) to then visit a brand dealer and buy the vehicle, thereby executing the Sell a car use case (Figure 14c).

Customers may also choose neither to register their data on the website nor to place a virtual order and visit the brand dealer directly to buy the vehicle in question, again executing the Sell a car use case. The website registers the navigation process and actions taken by the potential customer in a weblog.

- **Sell a car:** To buy a car (sell by the brand) people (customers or potential customers) visit a dealer where they will be served by a brand salesperson (vendor). This salesperson will enter the customer’s data in the computer system (SARPA). The system will provide sales support based on the customer profile and customer preferences, which will possibly have been entered in a virtual order generated on the website, on the customer history with the brand, if any, and vehicles in stock, special offers, etc that the dealer has at the time. If there is a sale, the Invoice document will be generated. The computer system is developed and improved by the IT department (Modify SARPA use case) based on the guidelines received from the business analyst after studying the information available in the business about customers, sales and vehicles available from each dealer (Figure 14d).
- **Analyse business data:** This use case deals with the analysis of all the information available within the company. The business analyst will generate the guidelines for newspaper and radio advertising and the guidelines for modifying the website and SARPA system after analysing this information (virtual order, weblog, customer, sale and vehicle). These guidelines will be used in the respective business use cases as discussed in the above descriptions of the business use cases (Figure 14e).
- **Modify website:** As of the guidelines delivered by the business analyst in the Analyse Business Data use case, the IT department will make the changes to the website stated in the guidelines for modifying website document (Figure 14f).
- **Modify SARPA:** As of the guidelines delivered by the

⁵Business workers could also be systems already in place in the business.

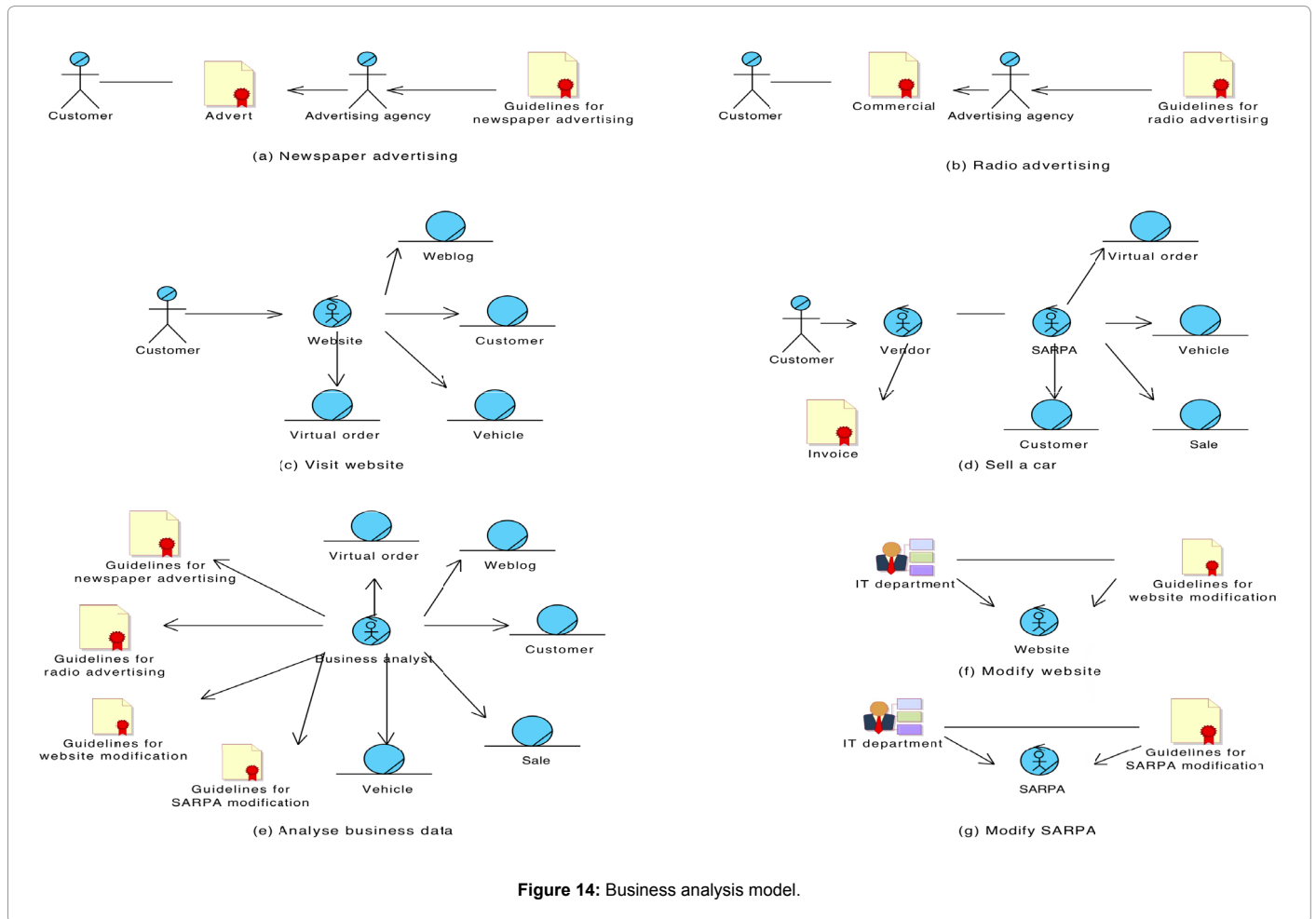


Figure 14: Business analysis model.

business analyst in the Analyse Business Data use case, the IT Department will make the changes to the SARPA system stated in the guidelines for modifying SARPA document (Figure 14g).

The *business use case realization* model has been omitted for abbreviating and for its simplicity. That model only joins the business analyse items with the corresponding business use case.

Data mining models

The next phase after analysing the business is to find out where the data mining project can come in to improve the business. The analysis diagrams indicate what data the business gathers for analysis, as well as where the results of the analysis can be used to achieve the business goals.

Data mining data models: The data mining project data model is derived from the business entities that represent the data available for the data mining project. Initially, it will only include the data source or the data source and the original tables. As the project advances, integrated, derived and/or transformed elements will appear [42], as will other sources of additional data available for the project. All this will constitute the data model to be used in the project. Figure 15 shows part of the initial data model of the example. The diagram illustrates the available customer data and where they are stored.

This model is built iteratively and incrementally throughout the project, as new data (new data sources, integrated, derived or

transformed attributes) are added or gathered as the project advances.

Figure 16 shows part of the final data model. It shows derived data (“age” attribute from “date of birth”, and “zip code” from “address”), and transformed data (sex attribute) that will be used by one of the data mining algorithms applied later. There is also a new attribute, “income”, which is added and gathered from a new data source provided by the Census Bureau, namely, the Population and Housing Censuses and Household Expenditure Surveys. After processing, this attribute offers an estimate of the average income per household in the area related to a zip code [43]. The income attribute is established by merging the customer table and the census-derived tables through the zip code attribute. This model was built after creating the Data Mining use case and Data Mining Goal models described in the following.

Data mining use case model: After analysing the business (business use cases, business goals, business analysis and data model), we create the DM-UML models representing part of the data mining project. This and the next task are perhaps the most critical parts of the DM process viewed from a business viewpoint and for which the literature provides less procedural support.

First, we have to build the data mining use cases model and then the data mining goals. The data mining use cases are obtained from the business use cases and goals. In principle, then, we would have a data mining use case for each business use case and business goal combination. Not all the business use cases are associated with a data mining use case. Whether or not there is a data mining use case

will depend on the business use case using data that can be analysed, and this analysis being used somehow to achieve the business goal. Neither is it unusual for the need and possibility of gathering data to be detected within a business use case using insufficient data to run data mining and for data collection to be added as another project use case. Figure 17⁶ shows the data mining use case model that indicates the relationship between the business use cases and goals and the data mining use cases. The name of the data mining use cases is formed from the name of the business goal from which they are derived plus the name or names of the related business use case, as shown in Figure 17. For example, the Analyse Business Data – Identify Potential Profits by Customer Profile and History data mining use case is obtained from the Analyse Business Data business use case and the Identify Potential Profits by Customer Profile and History business goal.

As the data mining use cases are created, we have to look at whether the business goal to which they are related can be evaluated directly in business terms or whether the business use case has to be assessed through another business goal appearing at higher levels of the business goals hierarchy. In this case, it would have to be added to the data mining use case together with its associated business use case. The DM use cases in this study (Figure 17) are related to support business use goals. When assigning goals to some of these data mining use cases, however, we would have to plot their relationship to core business use goals and their associated use case, as these are likely to be the goals that make the data mining use case worthwhile in business terms and also determine how the data mining use case should be executed.

The aim of the “Detect Customer Profiles by Region: Analyse Business Data Media Advertising” data mining use case, for example, is to search profiles by region not in the broad sense, but in the very regions where the radio or press campaigns are to be run, as this use case can be traced to the core business goals. This not only determines the place and scope of the region to be studied in the databases, but

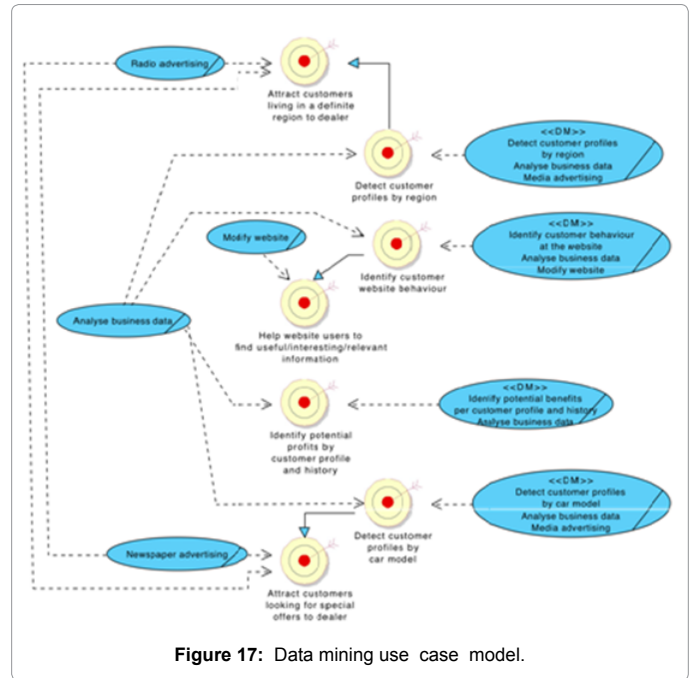


Figure 17: Data mining use case model.

also another context condition: the profiles to be analysed in the above regions should be confined to the set of profiles describing the audience of the radio station or newspaper or magazine that are going to be used in the campaign.

Then again the variables describing or defining these profiles (age, sex, income, occupation, education, household status, life style) delimit what type of variables should be used to perform the data mining tasks in databases. For all these reasons, both the Radio Advertising and Newspaper Advertising business use cases and their associated goal have been related to this data mining use case through the business goals hierarchy.

On the other hand, in order to rate the success of the “Detect Customer Profiles by Region: Analyse Business Data Media Advertising” data mining use case, we have to look at how the data mining use case realizes its real business value. As shown in Figure 17, the marketing campaign success criterion is to manage to attract the potential customers to the dealer (an X% of the regional population, a percentage that is usually accurately measured by the advertising agencies in response to their usual campaigns). Tracing the data mining use case goals, we find that the business value of the data mining use case is to increase this percentage (sale of vehicles due to this data mining use case). Therefore, if this does not happen, and without going into what could be the root of this problem or what type of corrective actions could be taken, the truth is that this is the criterion that company will use to measure the success of this particular data mining use case and that will determine its survival.

Other data mining use cases can be evaluated in other ways without being traced back to the core business goals, whose success can be hard to directly correlate to the data mining use case. For example, “<<DM>> Identify Customer Website Behaviour: Analyse Business Data” can be hard to correlate to “Attract Potential Customers that are Internet Users to Dealer” because, one might argue, its success or failure is also linked to the success or failure of the marketing campaigns run in other media. In this case, it could be evaluated, thanks to traceability, at the level of a support business goal like “Help Website Users to Find

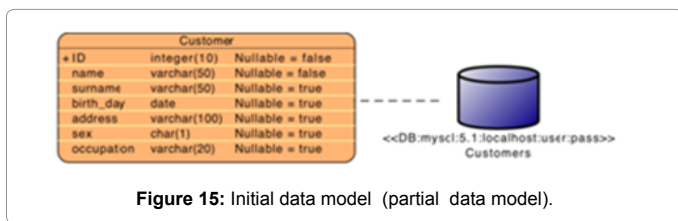


Figure 15: Initial data model (partial data model).

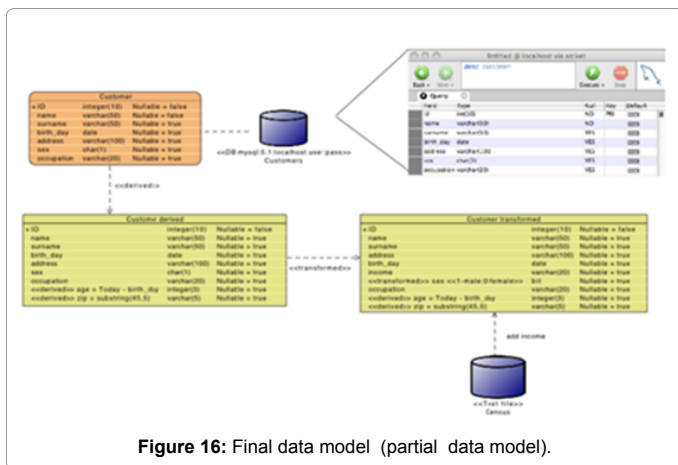


Figure 16: Final data model (partial data model).

⁶This figure shows part of the traceability model between the business and data mining use cases. It does not model the entire project.

Useful/Interesting/Relevant Information” (by means of a usability test [44]).

Finally, other data mining use cases can be evaluated directly without being traced to other goals. This would be the case of “Identify Potential Profits by Customer Profile and History: Analyse Business Data” that can be evaluated directly and fairly accurately by means of a statistical analysis of historical data. In conclusion, the traceability provided by UML from data mining use cases to business goals can detect the goal against which the use cases are to be evaluated, as well as providing the contextual information required for development.

Data mining goal model: Together with the data mining use cases, we must establish the data mining objectives for each use case. The data mining goals are established in terms of the business goals as they should be and are a *translation* of the business problem to problems expressed in data mining terms [10]. Typical data mining goals are: find typologies, cluster data, create a predictive model, etc., which will later be used to somehow achieve the business goal. In the case study, the goals are: create vehicle typologies, identify customer profiles or typologies, associate vehicle typology with customer profile, and find age-related and previous vehicle-related buying patterns.

Figure 18⁷ shows the data mining goal model diagram for our example. It shows that there are four use cases with their associated data mining goals forming a dependency, where a change in the data mining use case will lead to changes in the data mining goal. The relationship between the data mining goals and the business goals, i.e. their *translation*, is established by means of the hierarchy taken from the data mining use case model (Figure 17).

To build the data mining use cases diagram, we also have to include the data mining actors. A data mining use case has actors. These actors are the external agents that will ultimately use the knowledge gathered from the respective data mining use cases. In this case study, the external actor is the business analyst.

The data mining goals are used separately or can be combined for each business goal. For example, the “Detect Customer Profiles by Vehicle Model: Analyse Business Data – Media Advertising” data mining use case, whose final business goal is to “Attract potential customers looking for special offers to dealer” (Figure 17) can be tackled by combining the identification of customer typologies (identify customer-d profile in Figure 18) with the identification of vehicle types on offer (identify vehicle typologies-d in Figure 18), and launching an advert specially targeting that customer typology and vehicle type through the “Newspaper Advertising” or “Radio Advertising” business use case.

One and the same data mining goal can be achieved differently depending on the data mining use case. For example, Identify Customer Typologies can be executed using just basic data, like age and sex, or more sophisticated data, like occupation and income, depending on whether the associated business use case requires the customer profile to be associated with these data. In both cases, they are concerned with identifying customer typologies, but they each use different data and output different results, meaning that they must also be denoted differently. Figure 18 shows the Identify vehicle typologies-a and Identify vehicle typologies-d cases.

There are eight data mining goals in the example. When these goals are the same, they have been tagged with letters (a, b, c, d) for each data mining use case (Figure 18). For example, the “Identify Customer

Website Behaviour: Analyse Business Data – Modify Website” data mining use case has the following associated data mining goals:

- Identify customer-a profile: This goal intends to find out what brand customers captured through the website are like
- Identify vehicle typologies-a: This goal intends to describe major groups of vehicle types that the brand has clustered by features (engine, fuel, size, driver associated life style, price, etc.).
- Find vehicle-customer association-a: This goal aims to associate customer typology with vehicle typology.

In a real project, although the same sort of goals are stated differently, they may end up converging. For example, the “Identify customer-a profile” and “Identify customer-d profile” business goals could end up being developed using the same data and techniques, thereby cutting the number of data mining goals.

Data mining technique model: The data mining techniques models are obtained from the data mining use cases model containing data mining goals. They show which data mining technique will be used to achieve each data mining goal. For simplicity’s sake, Figure 19 shows just one of the data mining technique models. Specifically, we show the data mining techniques diagram for the Identify customer-a profile data mining goal. This data mining use goal is the one that will be developed in detail from now on. There will be similar models for the other data mining goals.

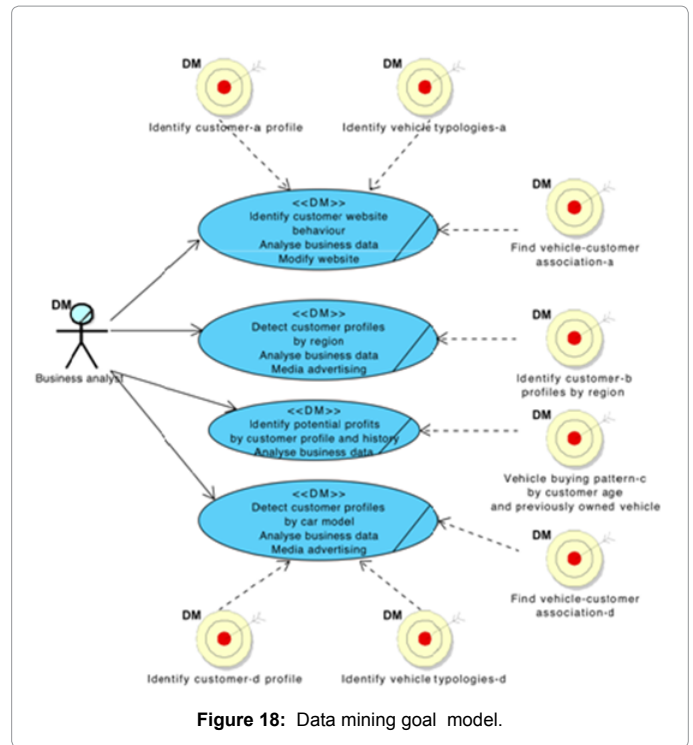


Figure 18: Data mining goal model.

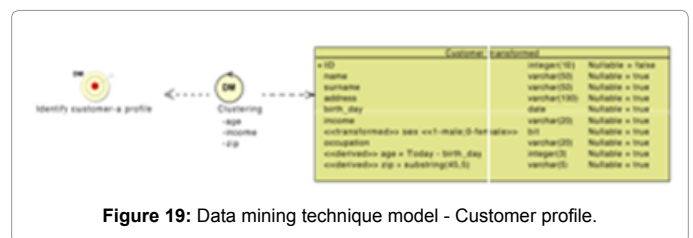


Figure 19: Data mining technique model - Customer profile.

⁷The figure shows only part of the data mining goal model

Figure 19 shows that clustering is the chosen data mining technique to be applied to achieve the Identify customer-a profile data mining goal. The data for applying this technique are to be found in the Customer Transformed data element, and, of all the available data, this technique will use age and zip code, which are derived attributes, and the income attribute.

Data mining algorithm model: The developer will decide which of the data mining algorithms is to be used from the data mining techniques model, and will create the model with the respective parameters. Figure 20 shows the data mining algorithms model derived from the data mining techniques model shown in Figure 19. In this case, k-means was the selected data mining algorithm for the clustering technique. As the algorithms for use depend on the tool, each algorithm to be modelled in the diagrams will also include different parameters depending on the data mining tool to be used. Figure 20 also shows the parameters for the k-means algorithm in the SPSS Clementine tool that has been used (the SPSS Clementine window depicted is not part of DM-UML). Using predefined UML Profiles. There are cases in the literature where UML profiles have been predefined to represent some specific data mining problems, for example, for clustering [30], for association [ZT07] or for classification [33]. If for the problem in question there is a specific profile, it can be used to model the data mining techniques and algorithms. Figure 21 shows the result of modelling the diagrams in Figure 19 and Figure 20 [16]. for clustering.

Data mining model: Finally, we build the data mining models diagram, which represents the data mining elements within the data mining tool to be used. Depending on the tool, we will be able to save the work space, generated model or both. Figure 22 shows the data mining model diagram (k-means: SPSS Clementine) for the algorithm diagram shown in Figure 20. Additionally, we add separate elements of Clementine, k-means.gm, which is the file that Clementine uses to save a data mining model, and Typologies.str, which is the file where Clementine stores the work space (stream). Additionally, Figure 22 shows the correspondence between the DM-UML elements and the respective elements in SPSS Clementine, although this correspondence is not part of DM-UML. Together with the diagrams in the example, textual descriptions of each of the elements could be added to clarify the element contents.

Porting DM-UML to a commercial data mining tool

In sum, Figure 23 shows the correspondence of the DM-UML elements of the data mining use case examined in this case study in a commercial data mining tool. It indicates how the use of the DM-UML models proposed in this article are portable to the implementation of a data mining project with a data mining tool, which, in this case, is SPSS Clementine. As we can see, DM-UML thoroughly documents

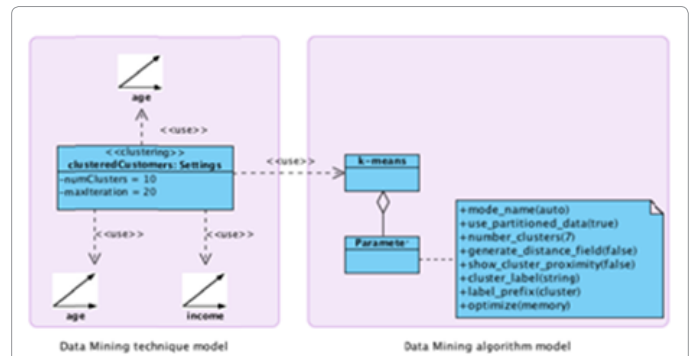


Figure 21: Specific UML profile for clustering technique [30].

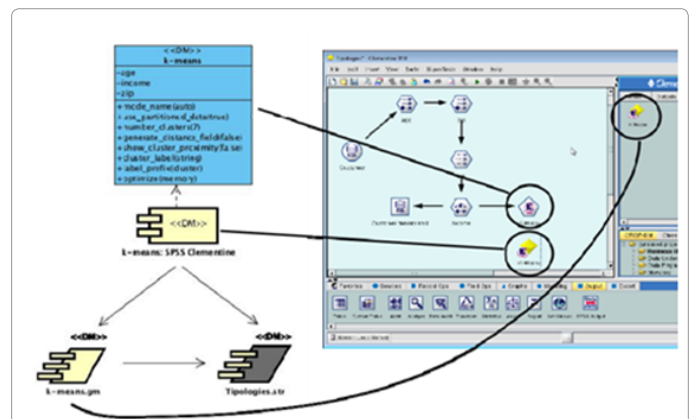


Figure 22: Data mining model model – Customer topology.

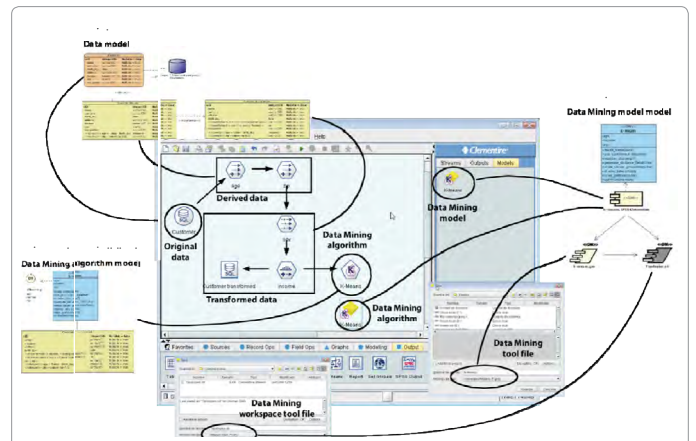


Figure 23: DM-UML artefacts vs. SPSS Clementine.

the project implementation in the selected data mining tool. Business modelling is not part of what is ported to the data mining tool, as it is designed to enable the creation of data mining models.

Conclusions

Project modelling and documentation is a well-known and growing problem as projects become more complex and the generated phase-to-phase information flow increases. So far data mining has made little progress in this direction. On the one hand, different data mining process standards have stated the need and criticality of project modelling and documentation, identifying the information to be output

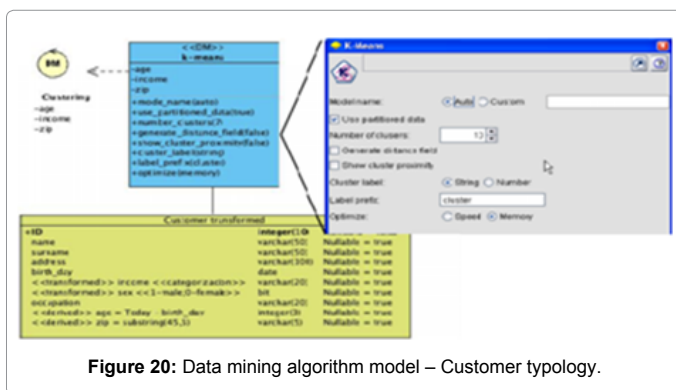


Figure 20: Data mining algorithm model – Customer topology.

at each stage as input for the next phase. However, they failed to specify or detail how that information should be structured. At the other end of the scale, though, we have found literature proposing a type of UML-based modelling for very specific process tasks and techniques. In this paper we have extended these seminal works to other non-technical parts of the data mining process, proposing DM-UML. DMUML is based on the UML modelling language. UML represents a collection of best engineering practices that have proven successful for modelling large and complex systems.

DM-UML covers all the phases of a data mining project, all the models are connected and depend on each other, and the way they are modelled assures that any change in the description of a key element is properly transferred to all its dependencies. It is also a very useful and transparent tool for modelling and connecting the business understanding or modelling phase with the remainder of the project right through to deployment, as well as a of communicating with the non-technical stakeholders involved in the project, which has always been an open question in data mining [11,37]. In this paper we have shown an example of a real application of DM-UML modelling. Its full development has been omitted for reasons of space. However, our case study does give a good idea of what the final result of DM-UML application is. Project documentation increases developers' workload, but this additional effort is an investment that helps to improve customer understanding, prevent misunderstandings between team members doing the analysis separately, and keep a record of all the techniques employed and their success, ruling out unnecessary repetitions. Also it saves project learning time when new members join the team or new goals are set.

References

1. Becker K, Ghedini C (2005) A documentation infrastructure for the management of data mining projects. *Inform Software Tech* 47: 95–111.
2. Berry MJA, Lino G (2004) *Data mining techniques: for marketing, sales, and customer support*. 2nd edn. Wiley Computer Publishing.
3. Fayyad U, Piatetsky SG, Smith P, Uthurusamy R (1996) *Advances in Knowledge Discovery and Data Mining*. American Association for Artificial Intelligence, California, USA.
4. Wirth R, Hipp J (2000) Crisp-dm: Towards a standard process model for data mining. Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining.
5. Wirth R, Shearer C, Grimmer U, Reinartz T, Schloßer J et al. (1997) Towards process oriented tool support for knowledge Discovery in database. *Data Min Knowl Disc* 1263: 243-253.
6. Hipp J, Lindner G (1999) Analysing warranty claims of automobiles; an application description following the crisp-dm data mining process 1749: 31–40.
7. Ghedini C, Becker K (2001) A documentation model for kdd application management support. International Conference of the Chilean Computer Science Society.
8. Brachma RJ, Anand T (1996) Advances in knowledge discovery and data mining, chapter. *The Process of Knowledge Discovery in Databases*, American Association for Artificial Intelligence, Menlo Park, CA, USA 37–57.
9. Zantout H, Marir F (1999) Document management systems from current capabilities towards intelligent information retrieval: an overview. *Int J Inform Manage* 19: 471-484.
10. Chapman P, Clinton J, Kerber R, Khazana T, Reinartz T et al. (2000) Crisp-dm 1.0 step by step data mining guide. CRISP-DM Consortium.
11. Marban O, Segovia J, Menasalvas E, Fernandez BC (2009) Toward data mining engineering: A software engineering approach. *Inform Syst* 34: 87–107.
12. Haramundanis K (1995) Documentation project management: some problems and solutions. In SIGDOC 95: Proceedings of the 13th annual international conference on Systems documentation: emerging from chaos: solutions for the growing complexity of our jobs, New York, USA.
13. Haramundanis K (1991) *The Art of Technical Documentation*. Digital Press.
14. Engels G, Heckel R, Sauer S (2000) UML-A Universal Modeling Language, Application and Theory of Petri Nets 2000 1825: 24-38.
15. Booch G, Rumbaugh J, Jacobson I (1999) *The Unified Modeling Language User Guide*. Addison-Wesley Professional.
16. Zubco J, Trujillo J (2007) A UML 2.0 profile to design association rule mining models in the multidimensional conceptual modeling of data warehouses. *Data Knowl Eng* 63: 44-62.
17. OMG (2007) *Omg Unified Modeling Language (omg UML) Superstructure version 2.1.2*.
18. Koch N (2007) Classification of model transformation techniques used in UML based web engineering. *IET Software* 1: 98-111.
19. Wu X, Chen J, Li R, Sun W, Zhang G, et al. (2006) Modeling a web-based remote monitoring and fault diagnosis system with UML and component technology. *J Intell Inf Syst* 27: 5-19.
20. Lee HK, Lee WJ, Chae HS, Kwon YR (2007) Specification and analysis of timing requirements for real-time systems in the cbd approach. *Real-Time Syst* 36: 135–158.
21. Felfernig A (2007) Standardized configuration knowledge representations as technological foundation for mass customization. *IEEE T Eng Manage* 54: 41-56.
22. Willard B (2007) UML for systems engineering. *Computer Stand Inter* 29: 69-81.
23. Secchi C, Bonfe M, Fantuzzi C (2007) On the use of UML for modeling mechatronic systems. *IEEE Transactions on Automation Science and Engineering* 4: 105-113.
24. Luján-Mora S, Trujillo J, Song I (2006) A UML profile for multidimensional modeling in data warehouses. *Data Knowl Eng* 59: 725-769.
25. Prat N, Akoka J, Comynwattiau I (2006) A UML-based data warehouse design method. *Decis Support Syst* 42: 1449-1473.
26. Lin J (2004) An object-oriented analysis method for customer relationship management information systems. *Inform Software Tech* 46: 433-443.
27. Lin J (2007) An object oriented development method for customer knowledge management information systems. *Knowl-Based Syst* 20: 17-36.
28. Frank E, Witten IH (2005) *Data Mining: Practical Machine Learning Tools with Java Implementations*. Morgan Kaufmann Publishers, New Jersey, USA.
29. Data Mining Group (2009) *Data mining group pmml 4.0 general structure of a pmml document*.
30. Zubco J, Pardillo J, Trujillo J (2007) Integrating clustering data mining into the multidimensional modeling of data warehouses with UML profiles. *Data Warehousing and Knowledge Discovery* 4654: 199-208.
31. Xu WL, Kuhnert L, Foster K, Bronlund J, Potgieter J, et al. (2007) Object-oriented knowledge representation and discovery of human chewing behaviours. *Engineering Applications of Artificial Intelligence* 20: 1000-1012.
32. Rizzi S (2004) UML-based conceptual modeling of pattern-bases. *Intl. Workshop on Pattern Representation and Management, 9th Int. Conference on Extending Database Technology (EDB)*.
33. Zubco J, Trujillo J (2006) Conceptual modeling for classification mining in data warehouses. *Data Warehousing and Knowledge Discovery* 4081: 566-575.
34. Maksimchuk RA, Naiburg EJ (2004) *UML for Mere Mortals*. Addison-Wesley Professional.
35. Fuentes L, Vallecillo A (2004) An introduction to UML profiles. *European Journal for the Informatics Professional*.
36. Marban O, Segovia J (2009) UML profile for data mining projects.
37. CRISP-DM Consortium. *Crisp-2.0: Updating the methodology*.
38. Stoyan H, Hognl O, Müller M (2000) The knowledge discovery assistant: Making data mining available for business users. *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*.
39. Rennolls K (2006) Visualization and bayesian nets to link business aims through kdd to deployment. *17th International Conference on Database and Expert Systems Applications (DEXA'06)*, Krakow, Poland.

-
40. Fogelman F (2006) Data mining in the real world. What do we need and what do we have. Workshop on Data Mining for Business Applications.
41. Dasu T, Koutsofios E, Wright J (2006) Zen and the art of data mining. Workshop on Data Mining for Business Applications. The 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
42. Fayyad UM, Piatetsky SG, Smyth P (1996) Advances in knowledge discovery and data mining, chapter From Data Mining to Knowledge Discovery: An Overview. American Association for Artificial Intelligence, Menlo Park, CA, USA.
43. Frutos S, Menasalvas E, Montes C, Segovia J (2003) Calculating economic indexes per household and censal section from official Spanish Intelligent Data Analysis.
44. J. Nielsen (1993) Usability engineering. Morgan Kaufmann Publications, USA.