Chapter #2

Initial activities oriented to reduce failure in information mining projects

Pablo Pytel
Program on Computer Science, National University of La Plata, La Plata; Information Systems Research Group, National University of Lanus, Remedios de Escalada; & GEMIS Group, Technological National University, Buenos Aires. Argentina

Paola Britos
Information Mining Research Group. National University of Rio Negro at El Bolson, Argentina

Ramón Garcia-Martinez
Information Systems Research Group, National University of Lanus, Remedios de Escalada, Argentina

1 INTRODUCTION

Most Traditional Software Engineering projects can be considered (at least) partial failures because few projects meet all their cost, schedule, quality, or requirements objectives (May 1998). From the challenged or canceled projects, the average project was 189 percent over budget, 222 percent behind schedule, and contained only 61 percent of the originally specified features. In 2005, it has been considered that from 5 to 15 percent of projects were abandoned before or shortly after delivery as hopelessly inadequate (Charette 2005). In other words, few projects truly succeed.

On the other hand, Information Mining projects are a special type of Software Engineering projects with the objective of extracting non-trivial knowledge which is located (implicitly) in the available data from different information sources (Schiefer *et al.* 2004). Commonly, instead of developing specific software, available software tools are used which already include the necessary techniques and algorithms (García-Martínez *et al.* 2011a). As a result, the features of Information Mining projects are different from Traditional Software Engineering projects and also from Knowledge Engineering projects (KE), even though the algorithms are based on artificial intelligence methods (García-Martínez *et al.* 2003). However, they share similar problems. Conducted studies about Information Mining projects have detected that not all projects are successfully completed (Edelstein *et al.* 1997, Strand 2000), ending most in failure. In 2000, it has estimated that 85% of the projects have failed to achieve its goals (Fayyad 2000). In other words, this means that from 100 developed projects only 15 have been successfully completed. After five years working, the community has been able to decrease this project failure rate to approximately 60% (Gondar 2005) and therefore it can be said that the community is working on the right lane but there are project elements that should be enhanced yet.

In this context, in this Chapter we have the objective of proposing two *ad-hoc* models to be used at the beginning of Information Mining project in order to increase the probability of successfully finishing it. First, Information Mining projects and its main characteristics are introduced (Section 2), then the problem is identified and the solution is proposed (Section 3) by defining the two models (Section 4). Finally, a conceptual proof is presented (Section 5) with the main conclusions (Section 6).

2 INFORMATION MINING PROJECTS

Information Mining is a sub-discipline of Information Systems which provides to Business Intelligence (Negash & Gray 2008) the necessary knowledge for the organizational decision making process. Information Mining involves more than the application of techniques to obtain this knowledge. As explained by García-Martínez *et al.* (2011a), two terms should be specified: "Data Mining" studies the technology (algorithms) to obtain knowledge from data repositories and "Information Mining" includes the application of processes and methodologies for successfully accomplishing the project goals. Consequently, Information Mining is closer to Software Engineering activities and Data Mining is closer to the developing tasks. Although Software Engineering provides several methods, techniques and tools, not all of them can be used because they are not focused on practical aspects. This means that specific methods, techniques, tools, and methodologies should be developed considering the main features of Information Mining projects.

The most used methodologies for Information Mining projects are CRISP-DM (Chapman *et al.* 2000), SEMMA (SAS 2008) and P3TQ (Pyle 2003). These methodologies are considered as proven by the community, but they exhibit problems when trying to define the phases related to project management (Vanrell *et al.* 2010). The elements of project management are mixed with project development process. On the other hand, tasks which should follow all the development process such as project monitoring, verification and measurement are not considered in the referenced methodologies.

Additionally, the features of Small and Medium-sized Enterprises (SMEs) should be considered for this type of

project, especially in Latin America. Normally, top-level managers (usually the company's owners) need non-trivial knowledge extracted from the available databases in order to solve a specific business problem with no special risks at stake. As the company's employees usually do not have the necessary experience, the project is performed by outsourcing consultants. The consultants need to elicit both the needs and expectations of the stakeholders, and also the features of the available data sources within the organization (i.e. existing data repositories). Although, the outsourcing consultants should have a minimum knowledge and experience in developing Information Mining projects, they might or not have experience in similar projects on the same business type which could facilitate the tasks of understanding the organization and its related data. As the data repositories are not so often properly documented, the organizational experts should be interviewed. However, experts are normally scarce and reluctant to get involved in the elicitation sessions. Thus, it is required the willingness of the personnel and the supervisors to identify the right features of the organization and the data repositories. As the project duration is quite short and the structure of the organization is centralized, it is considered that the elicited requirements will not change.

According to the Organization for Economic Cooperation and Development (OECD 2005): "SMEs (Small and Medium-sized Enterprises) constitute the dominant form of business organization in all countries world-wide, accounting for over 95 % and up to 99 % of the business population depending on country". However, although the importance of SMEs is well-known, there is no universal criterion to characterize them. Depending on the country and region, there are different quantitative and qualitative parameters used to recognize a company as SMEs. For instance, each country in Latin America has a different definition (Álvarez & Durán 2009): while Argentina considers as SME all independent companies with an annual turnover lower than USD 20,000 (USA dollars maximum amount that depends on the company's activities), Brazil includes all companies with 500 employees or less. On the other hand, the European Union defines as SMEs all companies with 250 employees or less, assets lower than USD 60,000 and gross sales lower than USD 70,000 per year. In that respect, International Organization for Standardization (ISO) has recognized the necessity to specify a software engineering standard for SMEs and thus it is working in the ISO/IEC 29110 standard "Lifecycle profiles for Very Little Entities" (ISO 2011). The term 'Very Little Entity' (VSE) was defined by the ISO/IEC JTC1/SC7 Working Group 24 (Laporte *et al.* 2008) as being "an entity (enterprise, organization, department or project) having up to 25 people."

On the other hand, the Information and Communication Technology (ICT) infra-structure of SMEs is analyzed. Ríos (2006) points out that more than 70% of Latin American SMEs have an ICT infrastructure, but only 37% have automated services and/or proprietary software. Normally commercial off-the-shelf software is used (such as spread-sheets managers and document editors) in order to register the management and operational information. The data repositories are not large (less than one million records) but implemented in several formats and technologies. Therefore, data formatting, data cleaning and data integration tasks will have a considerable effort if there is no available software tools to perform them because *ad-hoc* software should be developed for implementing such tasks.

3 ANALYSIS OF PROJECT FAIULRE

The most important reasons causing the failure of software development projects are, among others (Charette 2005):

- Unrealistic or unarticulated project goals.
- Poorly defined system requirements.
- Lack of communication among customers, developers, and users.
- Poor project management.
- Poor reporting of the project status.
- Inability to handle the project complexity.
- Stakeholder politics.
- Commercial pressures.
- Use of immature technology.
- Sloppy development practices.
- Unmanaged risks.
- Inaccurate estimates of needed resources.

The first three reasons are related to requirements handling and can be solved by applying methodologies and good practices (Wiegers 2003) considering the characteristics of the information mining projects (Britos *et al.* 2008). The next seven reasons are related to project manager activities that should be handled when executing the project (Vanrell *et al.* 2010).

The remaining two reasons are problems to be handled by the initial activities of the project. Before starting any traditional software project, the organization must decide whether it is appropriate doing it or not. Making such decisions is complex and depends on multiple factors; it is necessary to know both the impact of the software on the organization and its developing associated risks (Pressman 2004). This requires analyzing the project features by assessing the technical and economic feasibility of the project (commonly known as feasibility study). In expert system development projects, something similar happens. As the initial specifications for these systems are often uncertain, incomplete, and inconsistent, it is necessary to develop several prototypes for coherently define the system functionality, performance, and interfaces (García Martínez & Britos 2004). So, the Knowledge Engineering projects use more resources than traditional software development projects (Gómez *et al.* 1997). Then the feasibility study of these projects is highly important in order to identify the risks should be monitored and controlled during the project. Once the project is considered as feasible, it is necessary to predict the effort required to perform the project. With this information it is possible to estimate the necessary resources

and associated cost (Boehm *et al.* 2000). Although it is considered an activity required only for the project planning phase, the estimation of the project effort is also used as an indicator for the organization to decide if the project can be performed with the available resources. When the effort estimated for the project is too high, the management can decide to suspend the project or even cancel it. This is due, among other reasons, to problems undetected or wrongly handled.

The information mining project initial tasks are similar to a traditional software development project. By early detection of risks, its effects could be reduced during the project development. However, given that the features of information mining projects are different from traditional software and knowledge engineering projects, the models to study the feasibility and estimate the effort cannot be reused for this type of projects and it is necessary to propose specific ones.

4 PROPOSED MODELS

In this section two models are proposed to be used at the beginning of an information mining project. The first model aims to analyze the feasibility of the project (described in Section 4.1) and the second one allows estimating the resources and time required to perform the project (Section 4.2).

These models have been specified based on actual information mining projects collected by researchers from the following research groups: the Information Systems Research Group of the National University of Lanus (GISI-DDPyT-UNLa), the Information System Methodologies Research Group of the Technological National University at Buenos Aires (GEMIS-FRBA-UTN), and the Information Mining Research Group of the National University of Rio Negro at El Bolson (SAEB-UNRN).

Be advised that all these projects had been performed by applying the CRISP-DM methodology (Chapman *et al.* 2000). Therefore, the proposed models can be considered reliable only for Information Mining projects developed with this methodology.

4.1 *Feasibility model for information mining projects*

The proposal of the feasibility model for information mining projects requires the identification of the main conditions should be met to consider a project as feasible (section 4.1.1). Such a task is dependent on the project features. However, it is not usually easy to met these conditions by answering 'yes'/'no' questions (or by giving a numerical value). The proposed feasibility model should be able to handle a range of linguistic values to answer each condition. From such values, and by applying a pre-defined process, it would be possible to determine the overall project feasibility as detailed in Section 4.1.2.

4.1.1 *Conditions*

The main conditions are identified (Bolea *et al*. 2011, Davenport 2009, Fayyad 2000, Nemati & Barko 2003, Nie *et al.* 2009, Nothingli *et al.* 2011, Pipino *et al.* 2002, Sim 2003) and classified into three groups (or dimensions) based on the same criteria used for knowledge engineering (KE) projects feasibility test defined by García Martínez & Britos (2004) and Gómez *et al.* (1997):

• Conditions that determine the *plausibility* of the project include the factors that make it possible to perform the information mining project. A project can be performed if the following conditions are met: the available data repositories have current and representative data of the business problem to be solve, the business problem is understood, and the team has a minimum knowledge about the information mining process.

• Conditions that determine the *adequacy* of the project include the factors that determine whether information mining is the appropriate solution for the identified business problem (*i.e.*, it is the best solution for the problem). It is appropriate to apply information mining if the following conditions are met: the available data repositories have digital format (*i.e.*, they are not only available in paper), the business problem cannot be solved by using traditional statistical techniques, the business problem will not change during the project, and the data quality is good. The following metrics are used for assessing the data quality:

- Number of attributes and records (measuring the availability of enough data to apply the data mining process).
- Degree of credibility of the data (measures of how much you can trust on the data accuracy depending on the source and nature).

• Conditions that determine the *success* of the project, including the factors ensuring the project accomplishment. An information mining project will be successful if the following conditions are met: data repositories are implemented with technologies allowing easy data access and manipulation (*i.e.*, integration, cleaning, and formatting tasks), the project stakeholders (either high level managers, mid-level managers, or end-users) support the project, it is possible to perform the project planning considering best practices with necessary required time, and the team has experience in similar projects.

4.1.2 *Proposed procedure*

The five steps we propose to assess the project feasibility are:

Step 1: Determining the value of each project features.
Looking for characterizing an information mining project and evaluating its feasibility, the corresponding features should be identified from the interviews conducted in the organization. Such features (specified in Table 1) are based

on the conditions identified in Section 3.1. Each feature should be identified by using one of the following words: 'nothing', 'little', 'regular', 'much', and 'all'.

Table 1. Project features evaluated by the model.

Category	ID	Condition	Weight	Threshold
Data	P1	How much actual is considered the data from the repositories?	*8*	little
	P2	How representative is considered the data in the repositories in order to solve the business problem?	*9*	little
	A1	How much the data repositories have digital format?	*4*	little
	A2	How many attributes and records are available in the data repositories?	*7*	little
	A3	How much credibility has the available data?	*8*	little
	S1	In which degree the repository technology supports the manipulation of the data?	*6*	nothing
Business Problem	P3	How much the business problem is understood?	*7*	little
	A4	In which degree the business problem cannot be solved by traditional statistical techniques?	*10*	little
	A5	How stable is considered the business problem during the project?	*9*	little
Project	S2	How much the stakeholders support the project?	*8*	nothing
	S3	In which degree the project plan considers the required time to perform best practices during the project?	*7*	nothing
Project Team	P4	How much knowledge has the team about information mining?	*6*	little
	S4	How much experience has the team in similar projects?	*6*	nothing

For each feature of Table 1 the following attributes are defined:

• *Category*: used only to group the features according to what or who is concerned.

• *ID*: indicates a code to uniquely identify the property and the dimension to which it belongs (plausibility, adequacy, or success).

• *Condition*: describes the feature to be identified for characterizing the project.

• *Weight*: indicates the relative importance of each feature in the global model.

• *Threshold*: Indicates the value that the feature must be equal or bigger than. If that the feature does not exceed the threshold, it can be considered that the project is not feasible and is not necessary to continue with the next steps.

Step 2: Converting feature values into fuzzy intervals.

Once the linguistic values have been defined for each feature of Table 1, they should be translated into numeric values to calculate the project feasibility. The transformation process described in the feasibility test of KE projects (García Martínez & Britos 2004, Gómez *et al.* 1997) is used based on fuzzy expert systems (Jang 1997). For each word, the values of a fuzzy interval are defined and expressed by four numbers (ranging from zero to ten) that represents the breakpoints (or corner points) of the corresponding membership function. These intervals with the graphic representation of the membership function are shown in Figure 1.

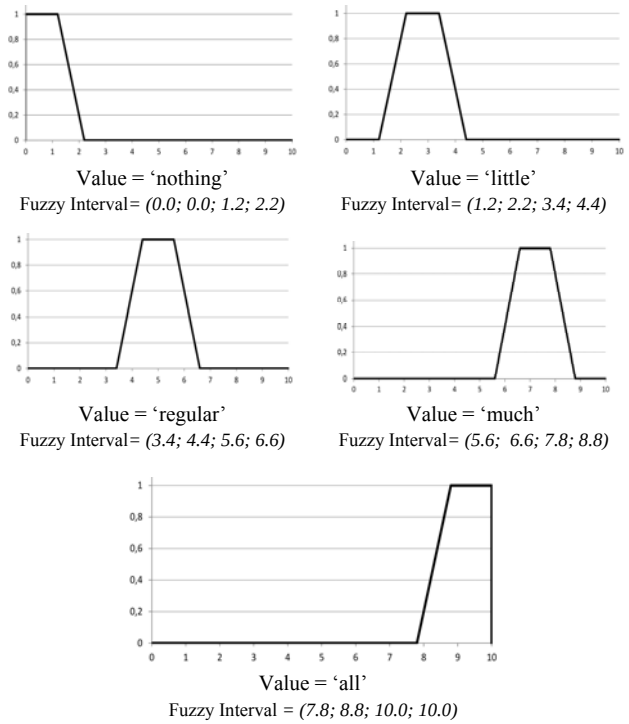


Figure 1. Membership function graphical and fuzzy interval assigned to each word.

Step 3: Calculating the value of each dimension.

In order to calculate the value of each project dimension, the fuzzy intervals (obtained in the previous step) are balanced considering its corresponding weight (as defined by Table 1). The interval representing the value of each dimension (Id) is calculated with the Formula # 1 of Table 2. This formula is formed by the combination of the harmonic mean and the arithmetic mean of the set of intervals. We aim to reduce the influence of low values when calculating the dimension value.

As the result of the formula, another fuzzy interval is achieved. In order to convert this interval into a single numeric value (Vd) the arithmetic average is used as shown in Formula # 2 of Table 2.

Step 4: Calculating the overall project feasibility.

Finally, the numerical values calculated in the previous step for each dimension (Vd) are combined by using a weighted

arithmetic mean (Formula # 3 of Table 2) obtaining the overall project feasibility value (OV).

Table 2. Formulas used by the model.

#	Formula
1	$$I_d = \left(\frac{1}{2} \cdot \frac{\sum_{i=1}^{n_d} W_{d_i}}{\sum_{i=1}^{n_d} \left(\frac{W_{d_i}}{F_{d_i}} \right)} \right) + \left(\frac{1}{2} \cdot \frac{\sum_{i=1}^{n_d} (W_{d_i} \cdot F_{d_i})}{\sum_{i=1}^{n_d} W_{d_i}} \right)$$ Where: I_d: represents the fuzzy interval calculated for the dimension *d* (using '*P*' for plausibility, '*A*' for adequacy, and '*S*' for success). W_{di}: represents the weight of the feature *i* for the dimension *d*. F_{di}: represents the fuzzy interval that has been assigned to the feature *i* for the dimension *d*. n_d: represents the quantity of features associated to the dimension *d*.
2	$$V_d = \frac{\sum_{i=1}^{4} I_{d_i}}{4}$$ Where: V_d: represents the numeric value calculated for the dimension *d*. I_{di}: represents the value of the position *i* of the fuzzy interval calculated for the dimension *d* .
3	$$OV = \frac{8 \cdot V_P + 8 \cdot V_A + 6 \cdot V_S}{22}$$ Where: *OV*: represents the overall project feasibility value. V_P: represents the value calculated for dimension plausibility. V_A: represents the value calculated for dimension adequacy. V_S: represents el value calculated for dimension success.

<u>Step 5:</u> *Interpreting the results.*
Once the numeric values for each dimension and the overall project feasibility value are calculated (steps 3 and 4 respectively), they should be analyzed. As a way to interpret the results of the feasibility of each dimension, it is recommended to plot the corresponding membership function of the obtained fuzzy interval (Id). The viability of the dimension can be considered as accepted if it exceeds the range of 'regular' value. Analyzing the numeric value of the dimension is another way to do it. If the dimension value (Vd) is greater than 5, the dimension can be considered as accepted.
On the other hand, for analyzing the feasibility of the project, the following criteria can be used: whether the three dimensions are accepted and the overall project feasibility (OV) is greater than 5, then the project is considered as feasible. Otherwise, it is not feasible.

In both cases, the engineer should also observe the weaknesses of the project to be strengthened (if project is not feasible).

4.2 *Effort estimation model for information mining projects oriented to SMEs*

The normal effort estimation method applied in traditional software development projects cannot be used at information mining projects because the considered features are different. For example COCOMO II (Boehm *et al.* 2000), one of the most used estimation method, uses the quantity of source code lines as a parameter. This is not useful for estimating an information mining project because the data mining algorithms are already available in commercial tools and then it is not necessary to develop software. Estimation methods in information mining projects should use more representative features, such as, the quantity of data sources, the level of integration within the data and the type of problem to be solved. In that respect, only one specific analytical estimation method for information mining projects has been found after a state-of-the-art review. Such a method, called Data Mining Cost Model (or DMCoMo), is defined by Marbán e*t al.* (2008). However, from a statistical analysis of DMCoMo performed by Pytel *et al.* (2011), it has been found that this method tends to overestimate the efforts mainly in little-sized projects that are usually required by SMEs.

Therefore, for specifying the effort estimation method oriented to SMEs, first, the cost drivers used to characterize a SMEs' project are defined (Section 4.2.1) and then the corresponding formula is presented (Section 4.2.2). This formula has been obtained by regression using real projects information.

4.2.1 *Cost Drivers*

Considering the features of information mining projects for SMEs indicated in Section 3, eight cost drivers are specified. Few cost drivers have been identified in this version because, as explained by Chen *et al.* (2005), when an effort estimation method is created, many of the non-significant data should be ignored. As a result, the model is pre-vented from being too complex (and therefore impractical), the irrelevant and co-dependent variables are removed, and the noise is also reduced.

The cost drivers have been selected based on the most critical tasks of CRISP-DM methodology: Domingos *et al.* (2006) indicate that building the data mining models and finding patterns is quite simple now, but 90% of the effort is included in the data pre-processing (*i.e.*, "Data Preparation" tasks performed at the phase III of CRISP-DM). From our experience, the other critical tasks are related to the "Business Understanding" phase (*i.e.*, "understanding of the business' background" and "identifying the project success" tasks). The proposed cost factors are grouped by three as follows:

Cost drivers related to the project:

- *Information mining objective type (OBTY)*

This cost driver analyses the objective of the information mining project and therefore the type of process to be applied based on the definition performed by García-Martínez *et al.* (2011b). The allowed values for this cost drivers are indicated in Table 3.

Table 3. Values of OBTY cost driver

Value	Description
1	It is desired to identify the rules that characterize the behavior or the description of an already known class.
2	It is desired to identify a partition of the available data without having a previously known classification.
3	It is desired to identify the rules that characterize the data partitions without a previous known classification.
4	It is desired to identify the attributes that have a greater frequency of incidence on the behavior or the description of an already known class.
5	It is desired to identify the attributes that have a greater frequency of incidence over a previously unknown class.

- *Level of collaboration of the organization (LECO)*

The level of collaboration from the members of the organization is analyzed by reviewing if the high-level management (*i.e.*, usually the SME's owners), the middle-level management (supervisors and department heads) and the operational personnel are willing to help the consultants to understand the business and the related data (especially in the first phases of the project). If the information mining project has been contracted, it is assumed that at least the high-level management should support it. The possible values for this cost factor are shown in Table 4.

Table 4. Values of LECO cost driver

Value	Description
1	Both managers and the organization's personnel are willing to collaborate on the project.
2	Only the managers are willing to collaborate on the project while the rest of the company personnel is not concerned with the project.
3	Only the high-level managers are willing to collaborate on the project while the middle-level manager and the rest of the company personnel is not concerned with the project.
4	Only the high-level managers are willing to collaborate on the project while the middle-level manager is not willing to collaborate.

Cost Drivers related to the available data:

- *Quantity and type of the available data repositories (AREP)*

The data repositories to be used by the information mining process are analyzed (including data base management systems, spread-sheets, and documents, among others). In this case, both the quantity of data repositories (public or private from the company) and the implementation technology are studied. In this stage, it is not necessary to know the quantity of tables in each repository because their integration within a repository is relatively simple as it can be performed with a query statement. However, depending on the technology, the complexity of the data integration tasks could vary. The following criteria can be used:

- If all the data repositories are implemented with the same technology, then the repositories are compatible for integration.
- If the data can be exported to a common format, then the repositories can be considered as compatible for integration because the data integration tasks will be performed by using the exported data.
- On the other hand, if there are non-digital repositories (*i.e.*, written paper), then the technology should not be considered compatible for the integration. But the estimation method is not able to predict the required time to perform the digitalization because it could vary on many factors (such as quantity of papers, length, format, and diversity, among others).

The possible values for this cost factor are shown in Table 5.

Table 5. Values of AREP cost driver

Value	Description
1	Only 1 available data repository.
2	Between 2 and 5 data repositories compatible technology.
3	Between 2 and 5 data repositories non-compatible technology.
4	More than 5 data repositories compatible technology.
5	More than 5 data repositories no-compatible technology.

- *Total quantity of available tuples in main table (QTUM)*

This variable ponders the approximate quantity of tuples (records) available in the main table to be used when applying data mining techniques. The possible values for this cost factor are shown in Table 6.

Table 6. Values of QTUM cost driver

Value	Description
1	Up to 100 tuples from main table.
2	Between 101 and 1,000 tuples from main table.
3	Between 1,001 and 20,000 tuples from main table.
4	Between 20,001 and 80,000 tuples from main table.
5	Between 80,001 and 5,000,000 tuples from main table.
6	More than 5,000,000 tuples from main table.

- *Total quantity of available tuples in auxiliaries tables (QTUA)*

This variable ponders the approximate quantity of tuples (records) available in the auxiliary tables (if any) used to add information to the main table (such as a table used for determining the product features associated with the product ID of the sales main table). Normally, these auxiliary tables include fewer records than the main table. The possible values for this cost factor are shown in Table 7.

- *Knowledge level about the data sources (KLDS)*

The knowledge level about the data sources studies if the data repositories and their tables are properly documented. In other words, if a document defining the technology in which

it is implemented, the features of the tables fields, and how the data is created, modified, and/or deleted.

Table 7. Values of QTUA cost driver

Value	Description
1	No auxiliary tables used.
2	Up to 1,000 tuples from auxiliary tables.
3	Between 1,001 and 50,000 tuples from auxiliary tables.
4	More than 50,000 tuples from auxiliary tables.

When this document is not available, it should be necessary to hold meetings with experts (usually in charge of the data administration and maintenance) for creating it. As a result, the project required effort should be increased depending on the collaboration of these experts to help the consultants.
The possible values for this cost factor are shown in Table 8.

Table 8. Values of KLDS cost driver

Value	Description
1	All the data tables and repositories are properly documented.
2	More than 50% of the data tables and repositories are documented and there are available experts to explain the data sources.
3	Less than 50% of the data tables and repositories are documented but there are available experts to explain the data sources.
4	The data tables and repositories are not documented but there are available experts to explain the data sources.
5	The data tables and repositories are not documented, and the available experts are not willing to explain the data sources.
6	The data tables and repositories are not documented and there are not available experts to explain the data sources.

Cost drivers related to the available resources:

• *Knowledge and experience level of the information mining team (KEXT)*
This cost driver studies the ability of the outsourcing consultants carrying out the project. Both the knowledge and experience of the team in similar previous projects are analyzed by considering the similarity of the business type, the data to be used and the expected goals. It is assumed that when there is greater similarity, the effort should be lower. Otherwise, the effort should be increased.
The possible values for this cost factor are shown in Table 9.

Table 9. Values of KEXT cost driver

Value	Description
1	The information mining team has worked with similar data in similar business types to obtain the same objectives.
2	The information mining team has worked with different data in similar business types to obtain the same objectives.
3	The information mining team has worked with similar data in other business types to obtain the same objectives.
4	The information mining team has worked with different data in other business types to obtain the same objectives.
5	The information mining team has worked with different data in other business types to obtain other objectives.

• *Functionality and usability of available tools (TOOL)*
This cost driver analyzes the features of the information mining tools to be utilized in the project and its implemented functionalities. Both the data preparation functions and the data mining techniques are reviewed.
The possible values of this cost factor are shown in Table 10.

Table 10. Values of TOOL cost driver

Value	Description
1	The tool includes functions for data formatting and integration (allowing the importation of more than one data table) and data mining techniques.
2	The tool includes functions for data formatting and data mining techniques, and it allows importing more than one data table independently.
3	The tool includes functions for data formatting and data mining techniques, and it allows importing only one data table at a time.
4	The tool includes only functions for data mining techniques, and it allows importing more than one data table independently.
5	The tool includes only functions for data mining techniques, and it allows importing only one data table at a time.

4.2.2 *Estimation Formula*

Once the values of the cost drivers have been specified, they were used to characterize 34 information mining projects with their actual effort collected by co-researchers as indicated before. A multivariate linear regression method (Weisberg 1985) has been applied to obtain a linear equation of the form used by COCOMO family methods (Boehm *et al.* 2000). As a result, the following formula is obtained:

$$PEM = 0.80\ OBTY + 1.10\ LECO - 1.20\ AREP - 0.30\ QTUM - 0.70\ QTUA + 1.80\ KLDS - 0.90\ KEXT + 1.86\ TOOL - 3.30$$

where PEM is the effort estimated by the proposed method for SMEs (in man-month), and the following cost drivers: information mining objective type (OBTY), level of collaboration from the organization (LECO), quantity and type of the available data repositories (AREP), total quantity of available tuples in the main table (QTUM) and in auxiliaries tables (QTUA), knowledge level about the data sources (KLDS), knowledge and experience level of the information mining team (KEXT), and functionality and usability of available tools (TOOL). The values for each cost driver are defined in Tables 3 to 10 respectively of Section 4.2.1.

5 CONCEPTUAL PROOF

As a way to test the proposed models, a small-sized project has been used. The project objective was the detection of evidence of causality between overall satisfaction and internet. The information from a survey conducted by the

organization to its customers has been used. This project has been completed successfully in 4 months with 3 people (*i.e.*, the real effort has been 12 man-month). First the feasibility model has been performed (Section 5.1) and then the estimation of the required effort has been calculated (Section 5.2).

5.1 *Analyzing the Viability of the conceptual project*

The steps proposed in Section 4.1 have been applied. The values of the project's features have been identified as indicated by Table 11 (as indicated by step 1).

Table 11. Values of the project features.

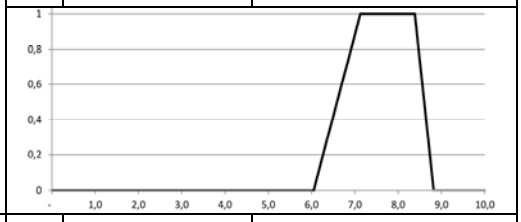
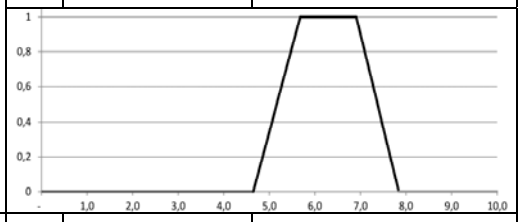
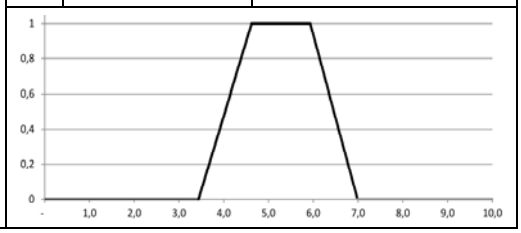
Category	ID	Condition	Assigned Value
Data	P1	How much actual is considered the data from the repositories?	*all*
	P2	How representative is considered the data in the repositories in order to solve the business problem?	*regular*
	A1	How much the data repositories are in digital format?	*all*
	A2	How many attributes and records are available in the data repositories?	*Much*
	A3	How much credibility has the available data?	*regular*
	S1	In which degree the repository technology aids the manipulation of the data?	*little*
Business Problem	P3	How much the business problem is understood?	*all*
	A4	In which degree the business problem cannot be solved by traditional statistical techniques?	*much*
	A5	How stable is considered the business problem during the project?	*regular*
Project	S2	How much the stakeholders support the project?	*much*
	S3	In which degree the project plan considers the required time to perform best practices during the project?	*regular*
Project Team	P4	How much knowledge has the team about information mining?	*all*
	S4	How much experience has the team in similar projects?	*much*

Then, these values are converted into fuzzy intervals (as indicated in step 2) to calculate the interval of each dimension (step 3) that is shown with its graphical representation in Table 12.

Later, the numerical value for each dimension and the overall project feasibility values are calculated (step 4) obtaining the following results:

- *Plausibility = 7.60*
- *Adequacy = 6.27*
- *Success = 5.25*
- *Overall project feasibility value =* ***6.47***

Table 12. Conversion and fuzzy intervals calculated for each dimension

Dimension	ID	Fuzzy interval of the feature	Fuzzy interval calculated for the Dimension (Vd)
Plausibility	P1	(7.8; 8.8; 10; 10)	*(6.05; 7.12; 8.39; 8.82)* The interval is greater than the 'much' value.
	P2	(3.4; 4.4; 5.6; 6.6)	
	P3	(7.8; 8.8; 10; 10)	
	P4	(7.8; 8.8; 10; 10)	
			
Adequacy	A1	(7.8; 8.8; 10; 10)	*(4.65; 5.68; 6.91; 7.84)* The interval is between 'regular' and 'much' values.
	A2	(5.6; 6.6; 7.8; 8.8)	
	A3	(3.4; 4.4; 5.6; 6.6)	
	A4	(5.6; 6.6; 7.8; 8.8)	
	A5	(3.4; 4.4; 5.6; 6.6)	
			
Success	S1	(1.2; 2.2; 3.4; 4.4)	*(3.44; 4.62; 5.93; 6.99)* The interval is greater than 'regular' value.
	S2	(5.6; 6.6; 7.8; 8.8)	
	S3	(3.4; 4.4; 5.6; 6.6)	
	S4	(5.6; 6.6; 7.8; 8.8)	
			

Finally, these values are interpreted (step 5) as follows: since the dimension values are above the required minimum, their feasibility is accepted. However, it should be noted that although the assessment of plausibility and adequacy is good, for the project success is very close to the minimum required value. This means that during the project the evaluated features should be monitored more closely. From the overall feasibility value it can be considered that the project is feasible to be performed.

5.2 *Estimating the Effort of the conceptual project*

As the project has been considered as feasible, the values of the cost drivers of the proposed estimation model (defined in Section 4.2) are identified in Table 13. With these values the estimated effort is calculated by applying the proposed formula. As it can be seen the proposal is very accurate to estimate the required error. The absolute error is lower than 6 man-day (*i.e.*, 0.18 man-month).

Table 13. Values of the cost drivers.

Category	ID	Value Description	Value
Project	OBTY	*It is desired to identify the attributes that have a greater frequency of incidence over a previously unknown class.*	*5*
	LECO	*Both managers and the organization's personnel are willing to collaborate on the project.*	*1*
Available Data	AREP	*Only 1 available data repository.*	*1*
	QTUM	*Between 1,001 and 20,000 tuples from main table.*	*3*
	QTUA	*Between 1,001 and 50,000 tuples from auxiliary tables.*	*3*
	KLDS	*The data tables and repositories are not documented and there are not available experts to explain the data sources.*	*6*
Available Resources	KEXT	*The information mining team has worked with different data in similar business types to obtain the same objectives.*	*2*
	TOOL	*The tool includes functions for data formatting and data mining techniques, and it allows importing only one data table at a time.*	*3*
Project Real Effort = *12,00 man-month*			
Effort Estimated by the Model = *12,18 man-month*			

6 CONCLUSIONS

Information mining is a sub-discipline of information systems which provides to business intelligence the needed non-trivial knowledge for making decision inside an organization. This knowledge is (implicitly) located in the available data from several information sources. Although such projects have different features, they share some of the problems of traditional software engineering and knowledge engineering projects. Most of the projects are not successfully completed, ending most in failure.

Among the reasons that produce project failure, two are highlighted: unmanaged risks and inaccurate estimates of needed resources. In order to handle these problems, two *ad-hoc* models are proposed to be used at the beginning of information mining projects. By early detection of risks, its effects could be reduced during development of the project. First one model has the objective of the analyzing the feasibility of the project. This means that based on the values of 13 features that characterize a project, the model allows "calculating" if the project can be performed (*i.e.*, its plausibility), if information mining is appropriate solution for the identified business problem (*i.e.*, adequacy) and if the project accomplishment can be achieved (*i.e.*, success). The model is able to manage five words for qualifying the features, because the engineers are not capable to answer the project features with yes/no or numerical values at the beginning of the project.

The second model allows estimating the resources and time required to perform the project based on the values of 8 project features (also known as cost drivers) and a formula. This model is oriented to estimate small projects which are normally needed by SMEs.

Finally, a conceptual proof is presented for applying both models to a real performed project.

AKNOWLEDGEMENT

The research reported in this Chapter has been partially granted by Research Projects 33A105 and 33B102 within National University of Lanus, and Research Project 40B133 within National University of Rio Negro.

REFERENCES

Álvarez, M. & Durán, J. 2009. *Manual de la Micro, Pequeña y Mediana Empresa. Una contribución a la mejora de los sistemas de información y el desarrollo de las políticas públicas*. San Salvador: CEPAL - Naciones Unidas. http://tinyurl.com/d5zarna.

Bolea, U., Jakličb, J., Papac, G., & Žabkard, J. 2011.*Critical Success Factors of Data Mining in Organizations*. Ljubljana.

Boehm, B., Abts, C., Brown, A., Chulani, S., Clark, B., Horowitz, E., Madachy, R., Reifer, D., Steece, B. 2000. *Software Cost Estimation with COCOMO II*. Prentice-Hall, Englewood Cliffs.

Britos, P., Dieste, O., García-Martínez, R. 2008. *Requirements Elicitation in Data Mining for Business Intelligence Projects*. EnIn Advances in Information Systems Research, Education and Practice. David Avison, George M. Kasper, Barbara Pernici, Isabel Ramos, Dewald Roode Eds. (Boston: Springer), IFIP Series, 274: 139–150.

Chapman P., Clinton J., Keber R., Khabaza T., Reinartz, T., Shearer, C., Wirth, R. 2000. *CRISP-DM 1.0 Step by step BI guide*. Edited by SPSS. http://tinyurl.com/crispdm

Charette, R. N. 2005. *Why software fails*. Spectrum, IEEE, 42(9), 42–49.

Chen, Z., Menzies, T., Port, D., et al. 2005 *Finding the right data for software cost modeling*. Software, IEEE, vol.22, no.6, pp. 38- 46, Nov.-Dec. 2005. Online (04/12).

Davenport, T. H. 2009. *Make Better Decisions*. Harvard Business Review, (November). Pp. 117-123

Domingos, P., Elkan, C., Gehrke, J., Han, J., Heckerman, D., Keim, D., et al.. 2006. 10 challeng-ing problems in data

mining research. *International Journal of Information Technology & Decision Making*, vol. 5, nº. 4, pp. 597–604.

Edelstein, H.A. & Edelstein, H.C.. Building, 1997. U*sing, and Managing the Data Warehouse*. Data Warehousing Institute. Prentice-Hall PTR, EnglewoodCliffs, NJ..

Fayyad, U.M. 2000. *Tutorial report*. Summer school of DM. Monash University, Australia.

García Martínez, R. & Britos, P. 2004. *Ingeniería de Sistemas Expertos*. 649 páginas. Editorial Nueva Librería. ISBN 987-1104-15-4.

García-Martínez, R., Britos, P., Pesado, P., Bertone, R., Pollo-Cattaneo, F., Rodríguez, D., Pytel, P., Vanrell. J. 2011a. *Towards an Information Mining Engineering.* En Software Engineering, Methods, Modeling and Teaching. Sello Editorial Universidad de Medellín. ISBN 978-958-8692-32-6. Páginas 83-99.

Garcia-Martinez, R., Britos, P., Pollo-Cattaneo, F., Rodriguez, D., Pytel, P. 2011b. *Information Mining Processes Based on Intelligent Systems*. Proceedings of II International Congress on Computer Science and Informatics (INFONOR-CHILE 2011), pp. 87-94. ISBN 978-956-7701-03-2.

García Martínez, R., Servente, M. y Pasquini, D. 2003. *Sistemas Inteligentes*. Editorial Nueva Librería. ISBN 987-1104-05-7

Gómez, A., Juristo, N., Montes, C. & Pazos, J. 1997. *Ingeniería del Conocimiento*. Centro de Estudios Ramón Areces. S.A., Madrid.

Gondar, J. E. 2005. *Metodología del Data Mining*. Number 84-96272-21-4. Data Mining Institute, S.L.

International Organization for Standardization (ISO). 2011 ISO/IEC DTR 29110-1 Software Engineering - Lifecycle Profiles for Very Little Entities (VSEs) - Part 1: Overview. International Organization for Standardization Geneva, Switzerland.

Jang, J. S. R. 1997. *Fuzzy inference systems*. Upper Saddle River, NJ: Prentice-Hall.

Laporte, C., Alexandre, S. & Renault, A. 2008. Developing International Standards for VSEs. IEEE Computer, vol. 41, nº 3, pp. 98—101.

Marbán, O., Menasalvas, E., Fernández-Baizán, C. 2008. *A cost model to estimate the effort of data mining projects (DMCoMo)*. Information Systems 33, pp. 133-150.

May, L. J. 1998. Major causes of software project failures. CrossTalk: *The Journal of Defense Software Engineering*, 11(6), 9-12.

Negash, S. & Gray, P. 2008. *Business Intelligence*. In Handbook on Decision Support Systems 2, eds. F. Burstein y C. Holsapple (Heidelberg, Springer), Pp. 175-193.

Nemati, H. R., & Barko, C. D. 2003. *Key factors for achieving organizational data-mining success*. Industrial Management & Data Systems, 103(4), pp. 282-292. doi:10.1108/02635570310470692. H

Nie, G., Zhang, L., Liu, Y., Zheng, X., & Shi, Y. 2009. *Decision analysis of data mining project based on Bayesian risk.* Expert Systems with Applications, 36(3), pp. 4589–4594.

Nothingli, A., Kakhky, E. N., & Nosratabadi, H. E.. *Evaluating the success level of data mining projects based on CRISP-DM methodology by a Fuzzy expert system*. Electronics Computer Technology (ICECT), 3rd International Conference on Kanyakumari, Vol. 6, pp. 161–165. IEEE. doi:10.1109/ICECTECH.2011.5942073

Organization for Economic Cooperation and Development: OECD SME and Entrepreneurship Outlook 2005. OECD Publishing. doi: 10.1787/9789264009257-en.

Pipino, L. L., Lee, Y. W. & Wang, R. Y. 2002. *Data quality assessment*. Communications of the ACM, 45(4), Pp. 211–218.

Pressman, R. 2004. *Software Engineering: A Practitioner's Approach*. Editorial Mc Graw Hill.

Pyle, D. 2003. *Business Modeling and Business intelligence*. Morgan Kaufmann

Pytel, P., Tomasello, M., Rodríguez, D., Pollo-Cattaneo, F., Britos, P., García-Martínez, R. 2011. *Estudio del Modelo Paramétrico DMCoMo de Estimación de Proyectos de Explotación de Información.* Proceedings XVII Congreso Argentino de Ciencias de la Computación, pp. 979--988. ISBN 978-950-34-0756-1.

Ríos, M. D. 2006. *El Pequeño Empresario en ALC, las TIC y el Comercio Electrónico*. Instituto para la Conectividad en las Américas. http://tinyurl.com/c97qkjd.

SAS .2008. *SAS Enterprise Miner: SEMMA* http://tinyurl.com/semmaSAS

Schiefer, J., Jeng, J., Kapoor, S. & Chowdhary, P. 2004. *Process Information Factory: A Data Management Approach for Enhancing Business Process Intelligence*. Proceedings IEEE International Conference on E-Commerce Technology. Pp. 162-169.

Sim, J. 2003. *Critical success factors in data mining projects*. Ph.D. Thesis, University of North Texas.

Strand, M. 2000. *The Business Value of Data Warehouses - Opportunities, Pitfalls and Future Directions*. Ph.D. Thesis, Department of Computer Science, University of Skovde.

Vanrell, J., Bertone, R., García-Martínez, R. 2010. *Modelo de Proceso de Operación para Proyectos de Explotación de Información*. Anales del XVI Congreso Argentino de Ciencias de la Computación. Pp. 674-682. ISBN 978-950-9474-49-9.

Wiegers, K. 2003. *Software Requirements*. Microsoft Press.

Weisberg, S. 1985. *Applied Linear Regression*. John Wiley & Sons, New York.