CHAPTER 1 INTRODUCTION

1.1 THE DATA MINING PROCESS

The procedure used to perform data mining modeling and analysis has undergone a long transformation from the domain of academic research to a systematic industrial process performed by business and quantitative analysts. Several methodologies have been proposed to cast the steps of developing and deploying data mining models into a standardized process.

This chapter summarizes the main features of these methodologies and highlights the role of the data preparation steps. Furthermore, it presents in detail the definition and contents of the the mining view and the scoring view.

1.2 METHODOLOGIES OF DATA MINING

Different methodologies of data mining attempt to mold the activities the analyst performs in a typical data mining engagement into a set of logical steps or tasks. To date, two major methodologies dominate the practice of data mining: CRISP and SEMMA.

CRISP, which stands for Cross Industry Standard Process for data mining, is an initiative by a consortium of software vendors and industry users of data mining technology to standardize the data mining process. The original CRISP documents can be found on *http://www.crisp-dm.org/*. On the other hand, SEMMA, which stands for Sample, Explore, Modify, Model, Assess, has been championed by SAS Institute. SAS has launched a data mining software platform (SAS Enterprise Miner) that implements SEMMA. For a complete description of SEMMA and SAS Enterprise Miner, visit the SAS web site: *http://www.sas.com/*.

In addition to SEMMA and CRISP, numerous other methodologies attempt to do the same thing, that is, to break the data mining process into a sequence of steps to be followed by analysts for the purpose of promoting best practices and standardizing the steps and results.

This book does not delve into the philosophical arguments about the advantages of each methodology. It extracts from them the basic steps to be performed in any data mining engagement to lay out a roadmap for the remaining chapters of this book.

All methodologies contain the following set of main tasks in one form or another.

1. Relevant data elements are extracted from a database or a data warehouse into one table containing *all* the variables needed for modeling. This table is commonly known as the *mining view*, or *rollup file*. In the case when the size of the data cannot be handled efficiently by available data modeling tools, which is frequently the case, sampling is used.
2. A set of data exploration steps are performed to gain some insight about the relationships among the data and to create a summary of the properties. This is known as EDA (Exploratory Data Analysis).
3. Based on the results of EDA, some transformation procedures are invoked to highlight and take advantage of the relationships among the variables in the planned models.
4. A set of data mining models are then developed using different techniques, depending on the objective of the exercise and the types of variables involved. Not all the available variables are used in the modeling phase. Therefore, a data reduction procedure is often invoked to select the most useful set of variables.
5. The data mining models are evaluated and the best performing model is selected according to some performance criteria.
6. The population of data intended for the application of the model is prepared in an identical process to that used in the preparation of the mining view to create what is known as the *scoring view*. The selected optimal (best) model is used to score the scoring view and produce the *scores*. These scores are used by the different business units to achieve the required business objective, such as selecting the targeted customers for marketing campaigns or to receive a loan or a credit card.

Typically, these steps are performed iteratively, and not necessarily in the presented linear order. For example, one might extract the mining view, perform EDA, build a set of models, and then, based on the evaluation results of these models, decide to introduce a set of transformations and data reduction steps in an attempt to improve the model performance.

Of the six steps in the data mining process, the data extraction and preparation steps could occupy up to 80% of the project time. In addition, to avoid a "garbage in, garbage out" situation, we have to make sure that we have extracted the right and most useful data. Therefore, data extraction and preparation should have the priority in planning and executing data mining projects. Therefore, this book!

1.3 The Mining View

Most, if not all, data mining algorithms deal with the data in the form of a single matrix (a two-dimensional array). However, the raw data, which contains the information needed for modeling, is rarely stored in such form. Most data is stored in relational databases, where the data is scattered over several tables. Therefore, the first step in collecting the data is to roll up the different tables and aggregate the data to the required rectangular form in anticipation of using mining algorithms. This last table, with all the elements needed, or suspected to be needed, for the modeling work is known as the *mining view*, *rollup file*, or *modeling table*. The tools used to aggregate the data elements into the mining view are usually data management tools such as SQL queries, SAS procedures, or, in the case of legacy systems, custom programs (e.g., in C, C++, and Java).

The mining view is defined as the aggregated modeling data on the specified *entity* level. The data is assembled in the form of columns, with the entity being unique on the row level. The meaning of the *entity* in the preceding definition is related to the business objective of modeling. In most business applications, the entity level is the *customer level*. In this case, we assemble all the relevant data for each customer in the form of columns and ensure that each row represents a unique customer with all the data related to this customer included. Examples of customer-level mining views are customer acquisition, cross selling, customer retention, and customer lifetime value.

In other situations, the entity is defined as the *transaction*. For example, we try to create a fraud detection system, say for online credit card shopping. In this case, the entity level is the purchase transaction and not the customer. This is because we attempt to stop fraudulent *transactions*. Similarly, the entity level may be defined as the *product level*. This could be necessary, for example, in the case of segmentation modeling of products for a supermarket chain where hundreds, or even thousands, of products exist.

The mining view usually undergoes a series of data cleaning and transformation steps before it is ready for use by the modeling algorithm. These operations achieve two purposes:

1. Clean the data of errors, missing values, and outliers.
2. Attempt to create new variables through a set of transformations, which could lead to better models.

Data errors and missing values always occur as a result of data collection and transformation from one data system to another. There are many techniques for cleaning the data from such errors and substituting or imputing the missing values.

Typically, the required data transformations are discovered over several iterations of data preparation, exploration, and pilot modeling. In other words, not all the needed data preparation steps are, or could be, known in advance. This is the nature of knowledge discovery in data mining modeling. However, once specific

transformations have been established and tested for a particular dataset for a certain model, they must be recorded in order to be used again on the data to be scored by the model. This leads us to the next view: the *scoring view*.

1.4 The Scoring View

The scoring view is very similar to the mining view except that the dependent variable (variable to be predicted) is not included. The following are other differences between the mining view and the scoring view.

1. The scoring view is usually much larger than the mining view. The mining view is only a sample from the data population; the scoring view is the population itself. This has implications on the requirements of the hardware and software needed to manipulate the scoring view and perform the necessary transformations on it before using it in scoring.
2. The scoring view may contain only one record. This is the case of online scoring, in which one record is read at a time and its score is calculated. The mining view, for obvious reasons, must have many records to be useful in developing a model.
3. The variables needed to make the mining view are determined by attempting to collect all conceivable variables that may have association with the quantity being predicted or have a relationship to the problem being modeled. The scoring view, on the other hand, contains only the variables that were used to create the model. The model may contain derived and transformed variables. These variables must also be in the scoring view. It is expected, therefore, that the scoring view would have significantly fewer variables than the mining view.

The only special case in which the mining view becomes the scoring view as well is the development of time series models for forecasting. In this case, the mining view is used to fit the predictive model and simultaneously to predict future values, thus removing the distinction between the mining view and the scoring view. We do not deal with data preparation for time series modeling in this book.

The next chapter provides a more detailed description of both the mining view and the scoring view.

1.5 Notes on Data Mining Software

Many software packages are used to develop data mining models. The procedures developed in this text for data preparation are independent of the tool used for the actual model building. Some of these tools include data preparation capabilities, thus allowing analysts to perform functions similar to some of the procedures described

in this book. Most analysts prefer to separate the procedures of data preparation and modeling. We have adopted this attitude by developing the procedures described in this book as SAS macros to be implemented independently of the modeling software.

However, the techniques and procedures described in the book could also be applied using many of the data manipulation capabilities of these modeling tools.

This Page Intentionally Left Blank