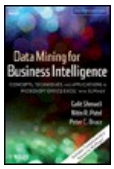

Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner, Second Edition

by Galit Shmueli, Nitin R. Patel and Peter C. Bruce
John Wiley & Sons (US). (c) 2010. Copying Prohibited.

Chapter 1: Introduction

1.1 What is Data Mining?

The field of data mining is still relatively new and in a state of evolution. The first International Conference on Knowledge Discovery and Data Mining (KDD) was held in 1995, and there are a variety of definitions of data mining. A concise definition that captures the essence of *data mining* is:

> Extracting useful information from large data sets.
>
> (Hand et al., 2001)
>
> A slightly longer version is:
>
> Data mining is the process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules.
>
> (Berry and Linoff, 1997, p. 5)
>
> Berry and Linoff later had cause to regret the 1997 reference to "automatic and semi-automatic means," feeling that it shortchanged the role of data exploration and analysis analysis (Berry and Linoff, 2000).
>
> Another definition comes from the Gartner Group, the information technology research firm:
>
> [Data Mining is] the process of discovering meaningful correlations, patterns and trends by sifting through large amounts of data stored in repositories. Data mining employs pattern recognition technologies, as well as statistical and mathematical techniques.
>
> (http://www.gartner.com/6_help/glossary, accessed May 14, 2010)

A summary of the variety of methods encompassed in the term *data mining* is given at the beginning of Chapter 2.

1.2 Where is Data Mining Used?

Data mining is used in a variety of fields and applications. The military use data mining to learn what roles various factors play in the accuracy of bombs. Intelligence agencies might use it to determine which of a huge quantity of intercepted communications are of interest. Security specialists might use these methods to determine whether a packet of network data constitutes a threat. Medical researchers might use it to predict the likelihood of a cancer relapse.

Although data mining methods and tools have general applicability, most examples in this book are chosen from the business world. Some common business questions that one might address through data mining methods include:

1. From a large list of prospective customers, which are most likely to respond? We can use classification techniques (logistic regression, classification trees, or other methods) to identify those individuals whose demographic and other data most closely matches that of our best existing customers. Similarly, we can use prediction techniques to forecast how much individual prospects will spend.

2. Which customers are most likely to commit, for example, fraud (or might already have committed it)? We can use classification methods to identify (say) medical reimbursement applications that have a higher probability of involving fraud and give them greater attention.

3. Which loan applicants are likely to default? We can use classification techniques to identify them (or logistic regression to assign a "probability of default" value).

4. Which customers are most likely to abandon a subscription service (telephone, magazine, etc.)? Again, we can use classification techniques to identify them (or logistic regression to assign a "probability of leaving" value). In this way, discounts or other enticements can be proffered selectively.

1.3 Origins of Data Mining

Data mining stands at the confluence of the fields of statistics and machine learning (also known as artificial intelligence). A variety of techniques for exploring data and building models have been around for a long time in the world of statistics:

linear regression, logistic regression, discriminant analysis, and principal components analysis, for example. But the core tenets of classical statistics— computing is difficult and data are scarce—do not apply in data mining applications where both data and computing power are plentiful.

This gives rise to Daryl Pregibon's description of data mining as "statistics at scale and speed" (Pregibon, 1999). A useful extension of this is "statistics at scale, speed, and simplicity." Simplicity in this case refers not to the simplicity of algorithms but, rather, to simplicity in the logic of inference. Due to the scarcity of data in the classical statistical setting, the same sample is used to make an estimate and also to determine how reliable that estimate might be. As a result, the logic of the confidence intervals and hypothesis tests used for inference may seem elusive for many, and their limitations are not well appreciated. By contrast, the data mining paradigm of fitting a model with one sample and assessing its performance with another sample is easily understood.

Computer science has brought us *machine learning techniques*, such as trees and neural networks, that rely on computational intensity and are less structured than classical statistical models. In addition, the growing field of database management is also part of the picture.

The emphasis that classical statistics places on inference (determining whether a pattern or interesting result might have happened by chance) is missing in data mining. In comparison to statistics, data mining deals with large datasets in open-ended fashion, making it impossible to put the strict limits around the question being addressed that inference would require.

As a result, the general approach to data mining is vulnerable to the danger of *overfitting*, where a model is fit so closely to the available sample of data that it describes not merely structural characteristics of the data but random peculiarities as well. In engineering terms, the model is fitting the noise, not just the signal.

1.4 Rapid Growth of Data Mining

Perhaps the most important factor propelling the growth of data mining is the growth of data. The mass retailer Wal-Mart in 2003 captured 20 million transactions per day in a 10-terabyte database (a terabyte is 1 million megabytes). In 1950, the largest companies had only enough data to occupy, in electronic form, several dozen megabytes. Lyman and Varian (2003) estimate that 5 exabytes of information were produced in 2002, double what was produced in 1999 (1 exabyte is 1 million terabytes); 40% of this was produced in the United States.

The growth of data is driven not simply by an expanding economy and knowledge base but by the decreasing cost and increasing availability of automatic data capture mechanisms. Not only are more events being recorded, but more information per event is captured. Scannable bar codes, point-of-sale (POS) devices, mouse click trails, and global positioning satellite (GPS) data are examples.

The growth of the Internet has created a vast new arena for information generation. Many of the same actions that people undertake in retail shopping, exploring a library, or catalog shopping have close analogs on the Internet, and all can now be measured in the most minute detail. In marketing, a shift in focus from products and services to a focus on the customer and his or her needs has created a demand for detailed data on customers.

The operational databases used to record individual transactions in support of routine business activity can handle simple queries but are not adequate for more complex and aggregate analysis. Data from these operational databases are therefore extracted, transformed, and exported to a *data warehouse*, a large integrated data storage facility that ties together the decision support systems of an enterprise. Smaller *data marts* devoted to a single subject may also be part of the system. They may include data from external sources (e.g., credit rating data).

Many of the exploratory and analytical techniques used in data mining would not be possible without today's computational power. The constantly declining cost of data storage and retrieval has made it possible to build the facilities required to store and make available vast amounts of data. In short, the rapid and continuing improvement in computing capacity is an essential enabler of the growth of data mining.

1.5 Why are There so Many Different Methods?

As can be seen in this book or any other resource on data mining, there are many different methods for prediction and classification. You might ask yourself why they coexist and whether some are better than others. The answer is that each method has advantages and disadvantages. The usefulness of a method can depend on factors such as the size of the dataset, the types of patterns that exist in the data, whether the data meet some underlying assumptions of the method, how noisy the data are, and the particular goal of the analysis. A small illustration is shown in Figure 1.1, where the goal is

to find a combination of *household income level* and *household lot size* that separate buyers (solid circles) from nonbuyers (hollow circles) of riding mowers. The first method (left panel) looks only for horizontal and vertical lines to separate buyers from nonbuyers, whereas the second method (right panel) looks for a single diagonal line.

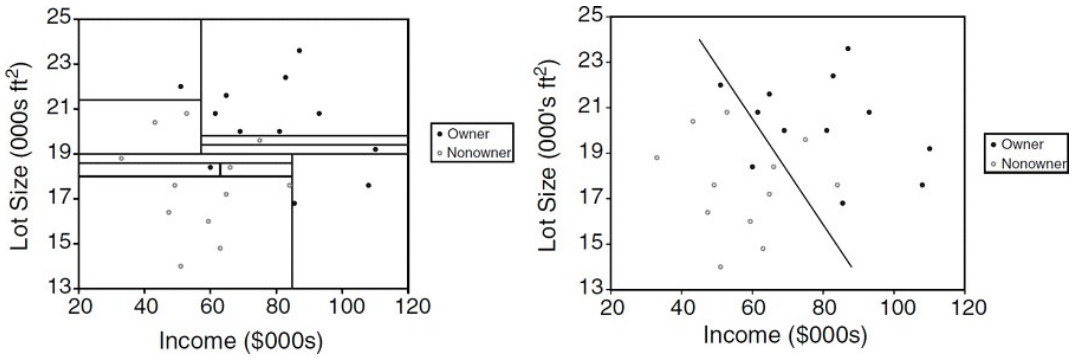

FIGURE 1.1: TWO METHODS FOR SEPARATING BUYERS FROM NONBUYERS

Different methods can lead to different results, and their performance can vary. It is therefore customary in data mining to apply several different methods and select the one that is most useful for the goal at hand.

1.6 Terminology and Notation

Because of the hybrid parentry of data mining, its practitioners often use multiple terms to refer to the same thing. For example, in the machine learning (artificial intelligence) field, the variable being predicted is the output variable or target variable. To a statistician, it is the dependent variable or the response. Here is a summary of terms used:

Algorithm Refers to a specific procedure used to implement a particular data mining technique: classification tree, discriminant analysis, and the like.

Attribute See **Predictor**.

Case See **Observation**.

Confidence Has a specific meaning in association rules of the type "IF *A* and *B* are purchased, *C* is also purchased." Confidence is the conditional probability that *C* will be purchased IF *A* and *B* are purchased.

Confidence Also has a broader meaning in statistics *(confidence interval)*, concerning the degree of error in an estimate that results from selecting one sample as opposed to another.

Dependent Variable See **Response**.

Estimation See **Prediction**.

Feature See **Predictor**.

Holdout Sample Is a sample of data not used in fitting a model, used to assess the performance of that model; this book uses the term *validation set* or, if one is used in the problem, *test set* instead of *holdout sample*.

Input Variable See **Predictor**.

Model Refers to an algorithm as applied to a dataset, complete with its settings (many of the algorithms have parameters that the user can adjust).

Observation Is the unit of analysis on which the measurements are taken (a customer, a transaction, etc.); also called *case, record, pattern, or row*. (Each row typically represents a record; each column, a variable.)

Outcome Variable See **Response**.

Output Variable See **Response**.

***P* (A | B)** Is the conditional probability of event *A* occurring given that event *B* has occurred. Read as "the probability that *A* will occur given that *B* has occurred."

Pattern Is a set of measurements on an observation (e.g., the height, weight, and age of a person).

Prediction The prediction of the value of a continuous output variable; also called *estimation*.

Predictor Usually denoted by *X*, is also called a *feature, input variable, independent variable*, or from a database perspective, a *field*.

Record See **Observation**.

Response usually denoted by *Y*, is the variable being predicted in supervised learning; also called *dependent variable, output variable, target variable, or outcome variable*.

Score Refers to a predicted value or class. *Scoring new data* means to use a model developed with training data to predict output values in new data.

Success Class Is the class of interest in a binary outcome (e.g., *purchasers* in the outcome *purchase/no purchase)*.

Supervised Learning Refers to the process of providing an algorithm (logistic regression, regression tree, etc.) with records in which an output variable of interest is known and the algorithm "learns" how to predict this value with new records where the output is unknown.

Test Data (or **test set)** Refers to that portion of the data used only at the end of the model building and selection process to assess how well the final model might perform on additional data.

Training Data (or **training set)** Refers to that portion of data used to fit a model.

Unsupervised Learning Refers to analysis in which one attempts to learn something about the data other than predicting an output value of interest (e.g., whether it falls into clusters).

Validation Data (or **validation set**) Refers to that portion of the data used to assess how well the model fits, to adjust some models, and to select the best model from among those that have been tried.

Variable Is any measurement on the records, including both the input *(X)* variables and the output *(Y)* variable.

1.7 Road Maps to This Book

The book covers many of the widely used predictive and classification methods as well as other data mining tools. Figure 1.2 outlines data mining from a process perspective and where the topics in this book fit in. Chapter numbers are indicated beside the topic. Table 1.1 provides a different perspective: It organizes data mining procedures according to the type and structure of the data.

Table 1.1: ORGANIZATION OF DATA MINING METHODS IN THIS BOOK, ACCORDING TO THE NATURE OF THE DATA[a]

	Continuous Response	Categorical Response	No Response
Continuous	Linear regression (6)	Logistic regression (10)	Principal components (4)
Predictors	Neural nets (11)	Neural nets (11)	Cluster analysis (14)
	k Nearest neighbors (7)	Discriminant analysis (12)	
		k Nearest neighbors (7)	
Categorical	Linear regression (6)	Neural nets (11)	Association rules (13)
Predictors	Neural nets (11)	Classification trees (9)	
	Regression trees (9)	Logistic regression (10)	
		Naive Bayes (8)	

[a]Numbers in parentheses indicate chapter number.

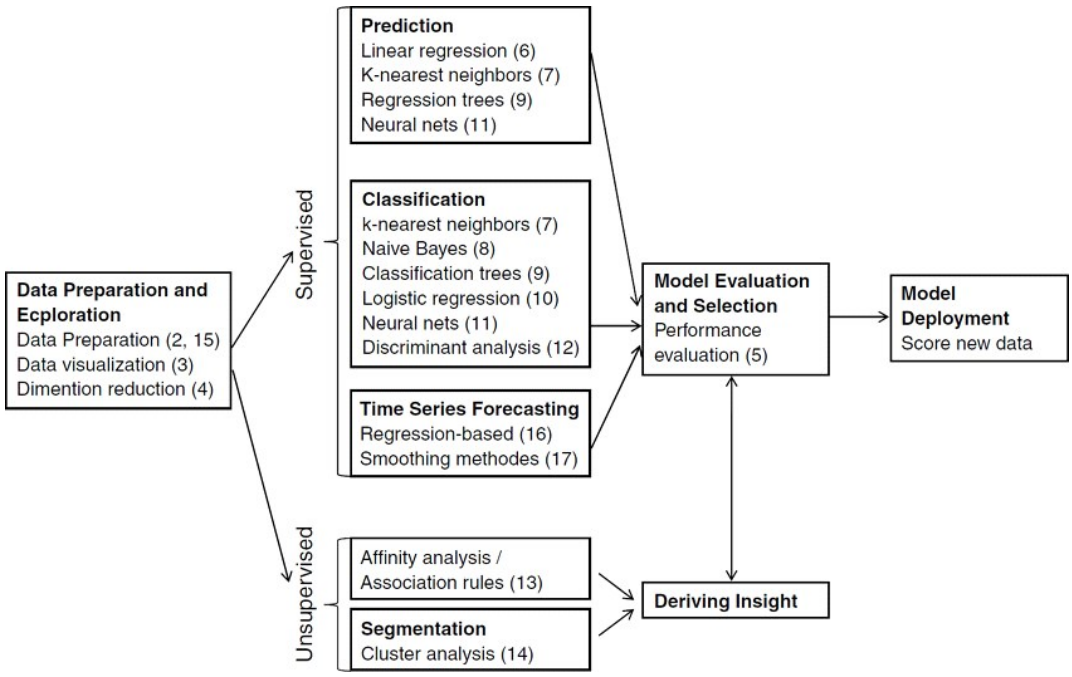

FIGURE 1.2: DATA MINING FROM A PROCESS PERSPECTIVE. NUMBERS IN PARENTHESES INDICATE CHAPTER NUMBERS

Order of Topics

The book is divided into five parts: Part I (Chapters 1–2) gives a general overview of data mining and its components. Part Ⅱ (Chapters 3–4) focuses on the early stage of data exploration and dimension reduction in which typically the most effort is expended.

Part III (Chapter 4) discusses performance evaluation. Although it contains a single chapter, we discuss a variety of topics, from predictive performance metrics to misclassification costs. The principles covered in this part are crucial for the proper evaluation and comparison of supervised learning methods.

Part IV includes eight chapters (Chapters 5–12), covering a variety of popular supervised learning methods (for classification and/or prediction). Within this part, the topics are generally organized according to the level of sophistication of the algorithms, their popularity, and ease of understanding.

Part V focuses on unsupervised learning, presenting association rules (Chapter 13) and cluster analysis (Chapter 14).

Part VI includes three chapters (Chapters 15–17), with the focus on forecasting time series. The first chapter covers general issues related to handling and understanding time series. The next two chapters present two popular forecasting approaches: regression-based forecasting and smoothing methods.

Finally, Part VII includes a set of cases.

Although the topics in the book can be covered in the order of the chapters, each chapter stands alone. It is advised, however, to read Parts I–Ⅲ before proceeding to the chapters in Parts IV–V, and similarly Chapter 15 should precede other chapters in Part VI.

USING XLMINER SOFTWARE

To facilitate hands-on data mining experience, this book comes with access to XLMiner, a comprehensive data miningadd-in for Excel. Forthose familiar with Excel, the use of an Excel add-in dramatically shortens the software learning curve. XLMiner will help you get started quickly on data mining and offers a variety of methods for analyzing data. The illustrations, exercises, and cases in this book are written in relation to this software. XLMiner has extensive coverage of statistical and data mining techniques for classification, prediction, affinity analysis, and data exploration and reduction. It offers a variety of data miningtools: neural nets, classification and regression trees, *k*-nearest neighbor classification, naive Bayes, logistic regression, multiple linear regression, and discriminant analysis, all for predictive modeling. It provides for automatic partitioning of data into training, validation, and test samples and for the deployment

of the model to new data. It also offers association rules, principal components analysis, *k*-means clustering, and hierarchical clustering, as well as visualization tools and data-handling utilities. With its short learning curve, affordable price, and reliance on the familiar Excel platform, it is an ideal companion to a book on data mining for the business student.

Installation Click on setup.exe and installation dialog boxes will guide you through the installation procedure. After installation is complete, the XLMiner program group appears under *Start* > *Programs* > *XLMiner*. You can either invoke XLM inerdirectlyor select the option to registerXLMinerasan Excel add-in.

Use Once opened, XLMiner appears as another menu in the top toolbar in Excel, as shown in Figure 1.3. By choosing the appropriate menu item, you can run any of XLMiner's procedures on the dataset that is open in the Excel worksheet.

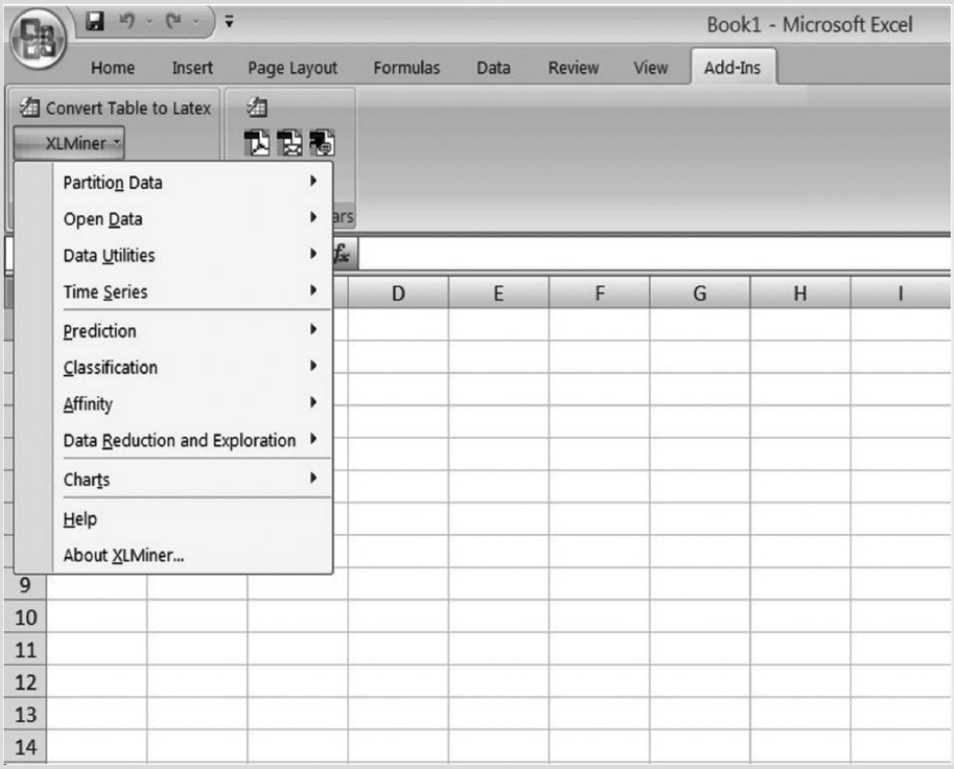

FIGURE 1.3: XLMINER SCREEN