CHAPTER 1

INTRODUCTION TO DATA MINING

About 13 million customers per month contact the West Coast customer service call center of the Bank of America, as reported by *CIO Magazine*'s cover story on data mining in May 1998 [1]. In the past, each caller would have listened to the same marketing advertisement, whether or not it was relevant to the caller's interests. However, "rather than pitch the product of the week, we want to be as relevant as possible to each customer," states Chris Kelly, vice president and director of database marketing at Bank of America in San Francisco. Thus, Bank of America's customer service representatives have access to individual customer profiles, so that the customer can be informed of new products or services that may be of greatest

Discovering Knowledge in Data: An Introduction to Data Mining, By Daniel T. Larose
ISBN 0-471-66657-2 Copyright © 2005 John Wiley & Sons, Inc.

interest to him or her. Data mining helps to identify the type of marketing approach for a particular customer, based on the customer's individual profile.

Former President Bill Clinton, in his November 6, 2002 address to the Democratic Leadership Council [2], mentioned that not long after the events of September 11, 2001, FBI agents examined great amounts of consumer data and found that five of the terrorist perpetrators were in the database. One of the terrorists possessed 30 credit cards with a combined balance totaling $250,000 and had been in the country for less than two years. The terrorist ringleader, Mohammed Atta, had 12 different addresses, two real homes, and 10 safe houses. Clinton concluded that we should proactively search through this type of data and that "if somebody has been here a couple years or less and they have 12 homes, they're either really rich or up to no good. It shouldn't be that hard to figure out which."

Brain tumors represent the most deadly cancer among children, with nearly 3000 cases diagnosed per year in the United States, nearly half of which are fatal. Eric Bremer [3], director of brain tumor research at Children's Memorial Hospital in Chicago, has set the goal of building a gene expression database for pediatric brain tumors, in an effort to develop more effective treatment. As one of the first steps in tumor identification, Bremer uses the Clementine data mining software suite, published by SPSS, Inc., to classify the tumor into one of 12 or so salient types. As we shall learn in Chapter 5 classification, is one of the most important data mining tasks.

These stories are examples of *data mining*.

WHAT IS DATA MINING?

According to the Gartner Group [4], "Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques." There are other definitions:

- "Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner" (Hand et al. [5]).
- "Data mining is an interdisciplinary field bringing togther techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large data bases" (Evangelos Simoudis in Cabena et al. [6]).

Data mining is predicted to be "one of the most revolutionary developments of the next decade," according to the online technology magazine *ZDNET News* [7]. In fact, the *MIT Technology Review* [8] chose data mining as one of 10 emerging technologies that will change the world. "Data mining expertise is the most sought after . . ." among information technology professionals, according to the 1999 *Information Week* National Salary Survey [9]. The survey reports: "Data mining skills are in high demand this year, as organizations increasingly put data repositories online. Effectively analyzing information from customers, partners, and suppliers has become important to more companies. ‘Many companies have implemented a data warehouse strategy and are now starting to look at what they can do with all that data,’ says Dudley Brown, managing partner of BridgeGate LLC, a recruiting firm in Irvine, Calif.”

How widespread is data mining? Which industries are moving into this area? Actually, the use of data mining is pervasive, extending into some surprising areas. Consider the following employment advertisement [10]:

STATISTICS INTERN: SEPTEMBER–DECEMBER 2003

Work with Basketball Operations

Resposibilities include:

- Compiling and converting data into format for use in statistical models
- Developing statistical forecasting models using regression, logistic regression, **data mining**, etc.
- Using statistical packages such as Minitab, SPSS, XLMiner

Experience in developing statistical models a differentiator, but not required.

Candidates who have completed advanced statistics coursework with a strong knowledge of basketball and the love of the game should forward your résumé and cover letter to:

Boston Celtics
Director of Human Resources
151 Merrimac Street
Boston, MA 02114

Yes, the Boston Celtics are looking for a data miner. Perhaps the Celtics’ data miner is needed to keep up with the New York Knicks, who are using IBM’s Advanced Scout data mining software [11]. Advanced Scout, developed by a team led by Inderpal Bhandari, is designed to detect patterns in data. A big basketball fan, Bhandari approached the New York Knicks, who agreed to try it out. The software depends on the data kept by the National Basketball Association, in the form of “events” in every game, such as baskets, shots, passes, rebounds, double-teaming, and so on. As it turns out, the data mining uncovered a pattern that the coaching staff had evidently missed. When the Chicago Bulls double-teamed Knicks’ center Patrick Ewing, the Knicks’ shooting percentage was extremely low, even though double-teaming should open up an opportunity for a teammate to shoot. Based on this information, the coaching staff was able to develop strategies for dealing with the double-teaming situation. Later, 16 of the 29 NBA teams also turned to Advanced Scout to mine the play-by-play data.

WHY DATA MINING?

While waiting in line at a large supermarket, have you ever just closed your eyes and listened? What do you hear, apart from the kids pleading for candy bars? You might hear the beep, beep, beep of the supermarket scanners, reading the bar codes on the grocery items, ringing up on the register, and storing the data on servers located at the supermarket headquarters. Each beep indicates a new row in the database, a new "observation" in the information being collected about the shopping habits of your family and the other families who are checking out.

Clearly, a lot of data is being collected. However, what is being learned from all this data? What knowledge are we gaining from all this information? Probably, depending on the supermarket, not much. As early as 1984, in his book *Megatrends* [12], John Naisbitt observed that "we are drowning in information but starved for knowledge." The problem today is not that there is not enough data and information streaming in. We are, in fact, inundated with data in most fields. Rather, the problem is that there are not enough trained *human* analysts available who are skilled at translating all of this data into knowledge, and thence up the taxonomy tree into wisdom.

The ongoing remarkable growth in the field of data mining and knowledge discovery has been fueled by a fortunate confluence of a variety of factors:

- The explosive growth in data collection, as exemplified by the supermarket scanners above
- The storing of the data in data warehouses, so that the entire enterprise has access to a reliable current database
- The availability of increased access to data from Web navigation and intranets
- The competitive pressure to increase market share in a globalized economy
- The development of off-the-shelf commercial data mining software suites
- The tremendous growth in computing power and storage capacity

NEED FOR HUMAN DIRECTION OF DATA MINING

Many software vendors market their analytical software as being plug-and-play out-of-the-box applications that will provide solutions to otherwise intractable problems without the need for human supervision or interaction. Some early definitions of data mining followed this focus on automation. For example, Berry and Linoff, in their book *Data Mining Techniques for Marketing, Sales and Customer Support* [13], gave the following definition for data mining: "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" (emphasis added). Three years later, in their sequel, *Mastering Data Mining* [14], the authors revisit their definition of data mining and state: "If there is anything we regret, it is the phrase 'by automatic or semi-automatic means' . . . because we feel there has come to be too much focus on the automatic techniques and not enough on the exploration and analysis. This has misled many people into believing that data mining is a product that can be bought rather than a discipline that must be mastered."

Very well stated! Automation is no substitute for human input. As we shall learn shortly, humans need to be actively involved at every phase of the data mining process. Georges Grinstein of the University of Massachusetts at Lowell and AnVil, Inc., stated it like this [15]:

> Imagine a black box capable of answering any question it is asked. Any question. Will this eliminate our need for human participation as many suggest? Quite the opposite. The fundamental problem still comes down to a human interface issue. How do I phrase the question correctly? How do I set up the parameters to get a solution that is applicable in the particular case I am interested in? How do I get the results in reasonable time and in a form that I can understand? Note that all the questions connect the discovery process to me, for my human consumption.

Rather than asking where humans fit into data mining, we should instead inquire about how we may design data mining into the very human process of problem solving.

Further, the very power of the formidable data mining algorithms embedded in the black-box software currently available makes their misuse proportionally more dangerous. Just as with any new information technology, *data mining is easy to do badly*. Researchers may apply inappropriate analysis to data sets that call for a completely different approach, for example, or models may be derived that are built upon wholly specious assumptions. Therefore, an understanding of the statistical and mathematical model structures underlying the software is required.

CROSS-INDUSTRY STANDARD PROCESS: CRISP–DM

There is a temptation in some companies, due to departmental inertia and compartmentalization, to approach data mining haphazardly, to reinvent the wheel and duplicate effort. A cross-industry standard was clearly required that is industry-neutral, tool-neutral, and application-neutral. The Cross-Industry Standard Process for Data Mining (CRISP–DM) [16] was developed in 1996 by analysts representing DaimlerChrysler, SPSS, and NCR. CRISP provides a nonproprietary and freely available standard process for fitting data mining into the general problem-solving strategy of a business or research unit.

According to CRISP–DM, a given data mining project has a life cycle consisting of six phases, as illustrated in Figure 1.1. Note that the phase sequence is *adaptive*. That is, the next phase in the sequence often depends on the outcomes associated with the preceding phase. The most significant dependencies between phases are indicated by the arrows. For example, suppose that we are in the modeling phase. Depending on the behavior and characteristics of the model, we may have to return to the data preparation phase for further refinement before moving forward to the model evaluation phase.

The iterative nature of CRISP is symbolized by the outer circle in Figure 1.1. Often, the solution to a particular business or research problem leads to further questions of interest, which may then be attacked using the same general process as before.

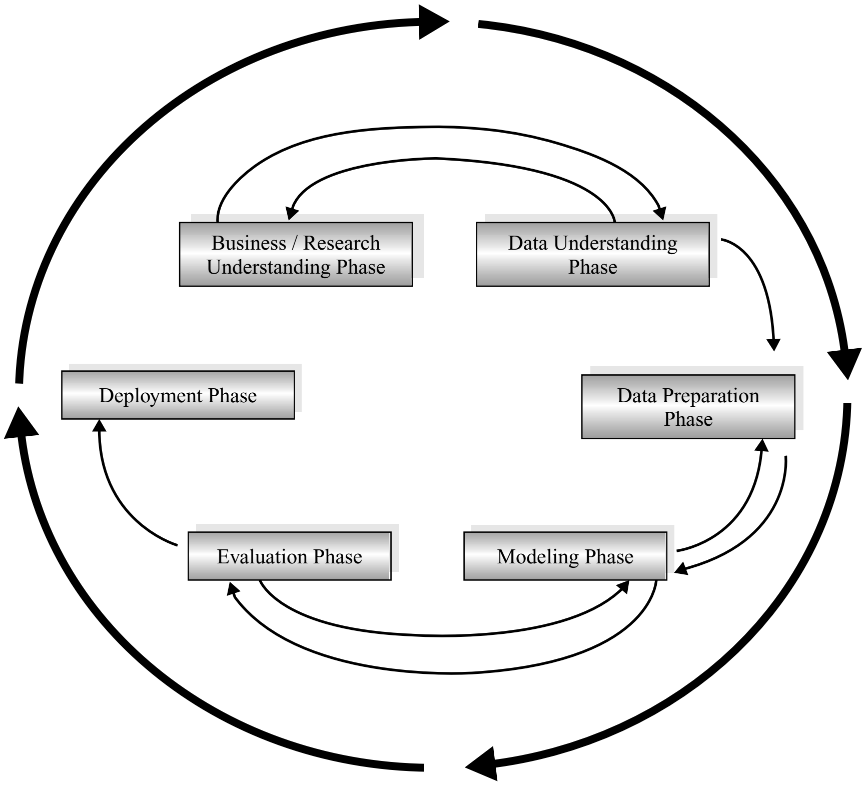

Figure 1.1 CRISP–DM is an iterative, adaptive process.

Lessons learned from past projects should always be brought to bear as input into new projects. Following is an outline of each phase. Although conceivably, issues encountered during the evaluation phase can send the analyst back to any of the previous phases for amelioration, for simplicity we show only the most common loop, back to the modeling phase.

CRISP–DM: The Six Phases

1. *Business understanding phase.* The first phase in the CRISP–DM standard process may also be termed the research understanding phase.
   a. Enunciate the project objectives and requirements clearly in terms of the business or research unit as a whole.
   b. Translate these goals and restrictions into the formulation of a data mining problem definition.
   c. Prepare a preliminary strategy for achieving these objectives.
2. *Data understanding phase*
   a. Collect the data.

   b. Use exploratory data analysis to familiarize yourself with the data and discover initial insights.

   c. Evaluate the quality of the data.

   d. If desired, select interesting subsets that may contain actionable patterns.

3. *Data preparation phase*

   a. Prepare from the initial raw data the final data set that is to be used for all subsequent phases. This phase is very labor intensive.

   b. Select the cases and variables you want to analyze and that are appropriate for your analysis.

   c. Perform transformations on certain variables, if needed.

   d. Clean the raw data so that it is ready for the modeling tools.

4. *Modeling phase*

   a. Select and apply appropriate modeling techniques.

   b. Calibrate model settings to optimize results.

   c. Remember that often, several different techniques may be used for the same data mining problem.

   d. If necessary, loop back to the data preparation phase to bring the form of the data into line with the specific requirements of a particular data mining technique.

5. *Evaluation phase*

   a. Evaluate the one or more models delivered in the modeling phase for quality and effectiveness before deploying them for use in the field.

   b. Determine whether the model in fact achieves the objectives set for it in the first phase.

   c. Establish whether some important facet of the business or research problem has not been accounted for sufficiently.

   d. Come to a decision regarding use of the data mining results.

6. *Deployment phase*

   a. Make use of the models created: Model creation does not signify the completion of a project.

   b. Example of a simple deployment: Generate a report.

   c. Example of a more complex deployment: Implement a parallel data mining process in another department.

   d. For businesses, the customer often carries out the deployment based on your model.

You can find out much more information about the CRISP–DM standard process at `www.crisp-dm.org`. Next, we turn to an example of a company applying CRISP–DM to a business problem.

CASE STUDY 1

ANALYZING AUTOMOBILE WARRANTY CLAIMS: EXAMPLE OF THE CRISP–DM INDUSTRY STANDARD PROCESS IN ACTION [17]

Quality assurance continues to be a priority for automobile manufacturers, including Daimler Chrysler. Jochen Hipp of the University of Tubingen, Germany, and Guido Lindner of DaimlerChrysler AG, Germany, investigated patterns in the warranty claims for DaimlerChrysler automobiles.

1. Business Understanding Phase

DaimlerChrysler's objectives are to reduce costs associated with warranty claims and improve customer satisfaction. Through conversations with plant engineers, who are the technical experts in vehicle manufacturing, the researchers are able to formulate specific business problems, such as the following:

- Are there interdependencies among warranty claims?
- Are past warranty claims associated with similar claims in the future?
- Is there an association between a certain type of claim and a particular garage?

The plan is to apply appropriate data mining techniques to try to uncover these and other possible associations.

2. Data Understanding Phase

The researchers make use of DaimlerChrysler's Quality Information System (QUIS), which contains information on over 7 million vehicles and is about 40 gigabytes in size. QUIS contains production details about how and where a particular vehicle was constructed, including an average of 30 or more sales codes for each vehicle. QUIS also includes warranty claim information, which the garage supplies, in the form of one of more than 5000 possible potential causes.

The researchers stressed the fact that the database was entirely unintelligible to domain nonexperts: "So experts from different departments had to be located and consulted; in brief a task that turned out to be rather costly." They emphasize that analysts should not underestimate the importance, difficulty, and potential cost of this early phase of the data mining process, and that shortcuts here may lead to expensive reiterations of the process downstream.

3. Data Preparation Phase

The researchers found that although relational, the QUIS database had limited SQL access. They needed to select the cases and variables of interest manually, and then manually derive new variables that could be used for the modeling phase. For example, the variable *number of days from selling date until first claim* had to be derived from the appropriate date attributes.

They then turned to proprietary data mining software, which had been used at DaimlerChrysler on earlier projects. Here they ran into a common roadblock—that the data format requirements varied from algorithm to algorithm. The result was further exhaustive preprocessing of the data, to transform the attributes into a form usable for model algorithms. The researchers mention that the data preparation phase took much longer than they had planned.

4. Modeling Phase

Since the overall business problem from phase 1 was to investigate dependence among the warranty claims, the researchers chose to apply the following techniques: (1) Bayesian networks and (2) association rules. Bayesian networks model uncertainty by explicitly representing the conditional dependencies among various components, thus providing a graphical visualization of the dependency relationships among the components. As such, Bayesian networks represent a natural choice for modeling dependence among warranty claims. The mining of association rules is covered in Chapter 10. Association rules are also a natural way to investigate dependence among warranty claims since the confidence measure represents a type of conditional probability, similar to Bayesian networks.

The details of the results are confidential, but we can get a general idea of the type of dependencies uncovered by the models. One insight the researchers uncovered was that a particular combination of construction specifications doubles the probability of encountering an automobile electrical cable problem. DaimlerChrysler engineers have begun to investigate how this combination of factors can result in an increase in cable problems.

The researchers investigated whether certain garages had more warranty claims of a certain type than did other garages. Their association rule results showed that, indeed, the confidence levels for the rule “If garage *X*, then cable problem,” varied considerably from garage to garage. They state that further investigation is warranted to reveal the reasons for the disparity.

5. Evaluation Phase

The researchers were disappointed that the support for sequential-type association rules was relatively small, thus precluding generalization of the results, in their opinion. Overall, in fact, the researchers state: “In fact, we did not find any rule that our domain experts would judge as interesting, at least at first sight.” According to this criterion, then, the models were found to be lacking in effectiveness and to fall short of the objectives set for them in the business understanding phase. To account for this, the researchers point to the “legacy” structure of the database, for which automobile parts were categorized by garages and factories for historic or technical reasons and not designed for data mining. They suggest adapting and redesigning the database to make it more amenable to knowledge discovery.

6. Deployment Phase

The researchers have identified the foregoing project as a pilot project, and as such, do not intend to deploy any large-scale models from this first iteration. After the pilot project, however, they have applied the lessons learned from this project, with the goal of integrating their methods with the existing information technology environment at DaimlerChrysler. To further support the original goal of lowering claims costs, they intend to develop an intranet offering mining capability of QUIS for all corporate employees.

What lessons can we draw from this case study? First, the general impression one draws is that uncovering hidden nuggets of knowledge in databases is a rocky road. In nearly every phase, the researchers ran into unexpected roadblocks and difficulties. This tells us that actually applying data mining for the first time in a company requires asking people to do something new and different, which is not always welcome. Therefore, if they expect results, corporate management must be 100% supportive of new data mining initiatives.

Another lesson to draw is that intense human participation and supervision is required at every stage of the data mining process. For example, the algorithms require specific data formats, which may require substantial preprocessing (see Chapter 2). Regardless of what some software vendor advertisements may claim, you can't just purchase some data mining software, install it, sit back, and watch it solve all your problems. Data mining is not magic. Without skilled human supervision, blind use of data mining software will only provide you with the wrong answer to the wrong question applied to the wrong type of data. The wrong analysis is worse than no analysis, since it leads to policy recommendations that will probably turn out to be expensive failures.

Finally, from this case study we can draw the lesson that there is no guarantee of positive results when mining data for actionable knowledge, any more than when one is mining for gold. Data mining is not a panacea for solving business problems. But used properly, by people who understand the models involved, the data requirements, and the overall project objectives, data mining can indeed provide actionable and highly profitable results.

FALLACIES OF DATA MINING

Speaking before the U.S. House of Representatives Subcommittee on Technology, Information Policy, Intergovernmental Relations, and Census, Jen Que Louie, president of Nautilus Systems, Inc., described four fallacies of data mining [18]. Two of these fallacies parallel the warnings we described above.

- ***Fallacy 1.*** There are data mining tools that we can turn loose on our data repositories and use to find answers to our problems.
  - ◦ *Reality.* There are no automatic data mining tools that will solve your problems mechanically "while you wait." Rather, data mining is a process, as we have seen above. CRISP–DM is one method for fitting the data mining process into the overall business or research plan of action.
- ***Fallacy 2.*** The data mining process is autonomous, requiring little or no human oversight.
  - ◦ *Reality.* As we saw above, the data mining process requires significant human interactivity at each stage. Even after the model is deployed, the introduction of new data often requires an updating of the model. Continuous quality monitoring and other evaluative measures must be assessed by human analysts.
- ***Fallacy 3.*** Data mining pays for itself quite quickly.
  - ◦ *Reality.* The return rates vary, depending on the startup costs, analysis personnel costs, data warehousing preparation costs, and so on.
- ***Fallacy 4.*** Data mining software packages are intuitive and easy to use.
  - ◦ *Reality.* Again, ease of use varies. However, data analysts must combine subject matter knowledge with an analytical mind and a familiarity with the overall business or research model.

To the list above, we add two additional common fallacies:

- ***Fallacy 5.*** Data mining will identify the causes of our business or research problems.
  - *Reality.* The knowledge discovery process will help you to uncover patterns of behavior. Again, it is up to humans to identify the causes.
- ***Fallacy 6.*** Data mining will clean up a messy database automatically.
  - *Reality.* Well, not automatically. As a preliminary phase in the data mining process, data preparation often deals with data that has not been examined or used in years. Therefore, organizations beginning a new data mining operation will often be confronted with the problem of data that has been lying around for years, is stale, and needs considerable updating.

The discussion above may have been termed *what data mining cannot or should not do*. Next we turn to a discussion of what data mining can do.

WHAT TASKS CAN DATA MINING ACCOMPLISH?

Next, we investigate the main tasks that data mining is usually called upon to accomplish. The following list shows the most common data mining tasks.

- Description
- Estimation
- Prediction
- Classification
- Clustering
- Association

Description

Sometimes, researchers and analysts are simply trying to find ways to *describe* patterns and trends lying within data. For example, a pollster may uncover evidence that those who have been laid off are less likely to support the present incumbent in the presidential election. Descriptions of patterns and trends often suggest possible explanations for such patterns and trends. For example, those who are laid off are now less well off financially than before the incumbent was elected, and so would tend to prefer an alternative.

Data mining models should be as *transparent* as possible. That is, the results of the data mining model should describe clear patterns that are amenable to intuitive interpretation and explanation. Some data mining methods are more suited than others to transparent interpretation. For example, decision trees provide an intuitive and human-friendly explanation of their results. On the other hand, neural networks are comparatively opaque to nonspecialists, due to the nonlinearity and complexity of the model.

High-quality description can often be accomplished by *exploratory data analysis*, a graphical method of exploring data in search of patterns and trends. We look at exploratory data analysis in Chapter 3.

Estimation

Estimation is similar to classification except that the target variable is numerical rather than categorical. Models are built using "complete" records, which provide the value of the target variable as well as the predictors. Then, for new observations, estimates of the value of the target variable are made, based on the values of the predictors. For example, we might be interested in estimating the systolic blood pressure reading of a hospital patient, based on the patient's age, gender, body-mass index, and blood sodium levels. The relationship between systolic blood pressure and the predictor variables in the training set would provide us with an estimation model. We can then apply that model to new cases.

Examples of estimation tasks in business and research include:

- Estimating the amount of money a randomly chosen family of four will spend for back-to-school shopping this fall.
- Estimating the percentage decrease in rotary-movement sustained by a National Football League running back with a knee injury.
- Estimating the number of points per game that Patrick Ewing will score when double-teamed in the playoffs.
- Estimating the grade-point average (GPA) of a graduate student, based on that student's undergraduate GPA.

Consider Figure 1.2, where we have a scatter plot of the graduate grade-point averages (GPAs) against the undergraduate GPAs for 1000 students. Simple linear regression allows us to find the line that best approximates the relationship between these two variables, according to the least-squares criterion. The regression line, indicated in blue in Figure 1.2, may then be used to estimate the graduate GPA of a student given that student's undergraduate GPA. Here, the equation of the regression line (as produced by the statistical package *Minitab*, which also produced the graph) is $\hat{y} = 1.24 + 0.67x$. This tells us that the estimated graduate GPA $\hat{y}$ equals 1.24 plus

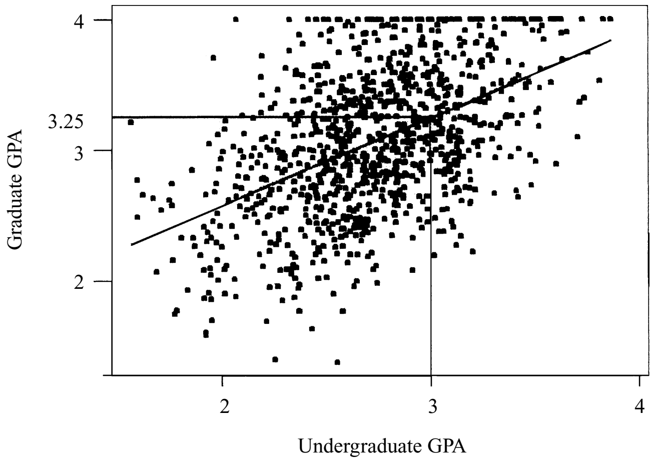

Figure 1.2 Regression estimates lie on the regression line.

0.67 times the student's undergraduate GPA. For example, if your undergrad GPA is 3.0, your estimated graduate GPA is $\hat{y} = 1.24 + 0.67(3) = 3.25$. Note that this point $(x = 3.0,\ \hat{y} = 3.25)$ lies precisely on the regression line, as do all linear regression predictions.

The field of statistical analysis supplies several venerable and widely used estimation methods. These include point estimation and confidence interval estimations, simple linear regression and correlation, and multiple regression. We examine these methods in Chapter 4. Neural networks (Chapter 7) may also be used for estimation.

Prediction

Prediction is similar to classification and estimation, except that for prediction, the results lie in the future. Examples of prediction tasks in business and research include:

- Predicting the price of a stock three months into the future (Figure 1.3)
- Predicting the percentage increase in traffic deaths next year if the speed limit is increased
- Predicting the winner of this fall's baseball World Series, based on a comparison of team statistics
- Predicting whether a particular molecule in drug discovery will lead to a profitable new drug for a pharmaceutical company

Any of the methods and techniques used for classification and estimation may also be used, under appropriate circumstances, for prediction. These include the traditional statistical methods of point estimation and confidence interval estimations, simple linear regression and correlation, and multiple regression, investigated in Chapter 4, as well as data mining and knowledge discovery methods such as neural network (Chapter 7), decision tree (Chapter 6), and *k*-nearest neighbor (Chapter 5) methods. An application of prediction using neural networks is examined later in the chapter in Case Study 2.

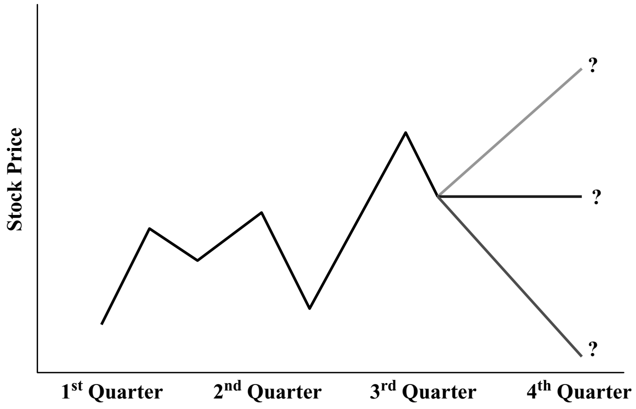

Figure 1.3 Predicting the price of a stock three months in the future.

Classification

In classification, there is a target categorical variable, such as *income bracket*, which, for example, could be partitioned into three classes or categories: high income, middle income, and low income. The data mining model examines a large set of records, each record containing information on the target variable as well as a set of input or predictor variables. For example, consider the excerpt from a data set shown in Table 1.1. Suppose that the researcher would like to be able to *classify* the income brackets of persons not currently in the database, based on other characteristics associated with that person, such as age, gender, and occupation. This task is a classification task, very nicely suited to data mining methods and techniques. The algorithm would proceed roughly as follows. First, examine the data set containing both the predictor variables and the (already classified) target variable, *income bracket*. In this way, the algorithm (software) "learns about" which combinations of variables are associated with which income brackets. For example, older females may be associated with the high-income bracket. This data set is called the *training set*. Then the algorithm would look at new records, for which no information about income bracket is available. Based on the classifications in the training set, the algorithm would assign classifications to the new records. For example, a 63-year-old female professor might be classified in the high-income bracket.

Examples of classification tasks in business and research include:

- Determining whether a particular credit card transaction is fraudulent
- Placing a new student into a particular track with regard to special needs
- Assessing whether a mortgage application is a good or bad credit risk
- Diagnosing whether a particular disease is present
- Determining whether a will was written by the actual deceased, or fraudulently by someone else
- Identifying whether or not certain financial or personal behavior indicates a possible terrorist threat

For example, in the medical field, suppose that we are interested in classifying the type of drug a patient should be prescribed, based on certain patient characteristics, such as the age of the patient and the patient's sodium/potassium ratio. Figure 1.4 is a scatter plot of patients' sodium/potassium ratio against patients' ages for a sample of 200 patients. The particular drug prescribed is symbolized by the shade of the points. Light gray points indicate drug Y; medium gray points indicate drug A or X;

TABLE 1.1 Excerpt from Data Set for Classifying Income

Subject	Age	Gender	Occupation	Income Bracket
001	47	F	Software engineer	High
002	28	M	Marketing consultant	Middle
003	35	M	Unemployed	Low
⋮				

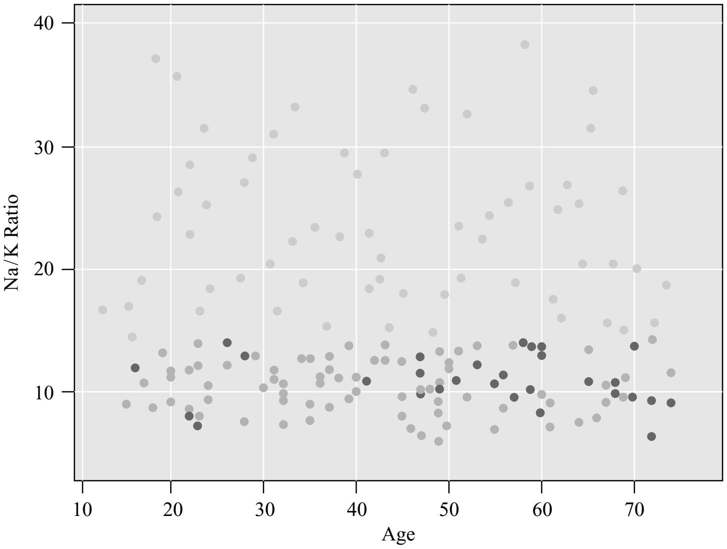

Figure 1.4 Which drug should be prescribed for which type of patient?

dark gray points indicate drug B or C. This plot was generated using the Clementine data mining software suite, published by SPSS.

In this scatter plot, Na/K (sodium/potassium ratio) is plotted on the *Y* (vertical) axis and age is plotted on the *X* (horizontal) axis. Suppose that we base our prescription recommendation on this data set.

1. Which drug should be prescribed for a young patient with a high sodium/potassium ratio?
   ◦ Young patients are on the left in the graph, and high sodium/potassium ratios are in the upper half, which indicates that previous young patients with high sodium/potassium ratios were prescribed drug Y (light gray points). The recommended prediction classification for such patients is drug Y.
2. Which drug should be prescribed for older patients with low sodium/potassium ratios?
   ◦ Patients in the lower right of the graph have been taking different prescriptions, indicated by either dark gray (drugs B and C) or medium gray (drugs A and X). Without more specific information, a definitive classification cannot be made here. For example, perhaps these drugs have varying interactions with beta-blockers, estrogens, or other medications, or are contraindicated for conditions such as asthma or heart disease.

Graphs and plots are helpful for understanding two- and three-dimensional relationships in data. But sometimes classifications need to be based on many different predictors, requiring a many-dimensional plot. Therefore, we need to turn to more sophisticated models to perform our classification tasks. Common data mining methods used for classification are *k*-nearest neighbor (Chapter 5), decision tree (Chapter 6), and neural network (Chapter 7). An application of classification using decision trees is examined in Case Study 4.

Clustering

Clustering refers to the grouping of records, observations, or cases into classes of similar objects. A *cluster* is a collection of records that are similar to one another, and dissimilar to records in other clusters. Clustering differs from classification in that there is no target variable for clustering. The clustering task does not try to classify, estimate, or predict the value of a target variable. Instead, clustering algorithms seek to segment the entire data set into relatively homogeneous subgroups or clusters, where the similarity of the records within the cluster is maximized and the similarity to records outside the cluster is minimized.

Claritas, Inc. [19] is in the clustering business. Among the services they provide is a demographic profile of each of the geographic areas in the country, as defined by zip code. One of the clustering mechanisms they use is the PRIZM segmentation system, which describes every U.S. zip code area in terms of distinct lifestyle types (Table 1.2). Just go to the company's Web site [19], enter a particular zip code, and you are shown the most common PRIZM clusters for that zip code.

What do these clusters mean? For illustration, let's look up the clusters for zip code 90210, Beverly Hills, California. The resulting clusters for zip code 90210 are:

- *Cluster 01:* Blue Blood Estates
- *Cluster 10:* Bohemian Mix
- *Cluster 02:* Winner's Circle
- *Cluster 07:* Money and Brains
- *Cluster 08:* Young Literati

TABLE 1.2 The 62 Clusters Used by the PRIZM Segmentation System

01 Blue Blood Estates	02 Winner's Circle	03 Executive Suites	04 Pools & Patios
05 Kids & Cul-de-Sacs	06 Urban Gold Coast	07 Money & Brains	08 Young Literati
09 American Dreams	10 Bohemian Mix	11 Second City Elite	12 Upward Bound
13 Gray Power	14 Country Squires	15 God's Country	16 Big Fish, Small Pond
17 Greenbelt Families	18 Young Influentials	19 New Empty Nests	20 Boomers & Babies
21 Suburban Sprawl	22 Blue-Chip Blues	23 Upstarts & Seniors	24 New Beginnings
25 Mobility Blues	26 Gray Collars	27 Urban Achievers	28 Big City Blend
29 Old Yankee Rows	30 Mid-City Mix	31 Latino America	32 Middleburg Managers
33 Boomtown Singles	34 Starter Families	35 Sunset City Blues	36 Towns & Gowns
37 New Homesteaders	38 Middle America	39 Red, White & Blues	40 Military Quarters
41 Big Sky Families	42 New Eco-topia	43 River City, USA	44 Shotguns & Pickups
45 Single City Blues	46 Hispanic Mix	47 Inner Cities	48 Smalltown Downtown
49 Hometown Retired	50 Family Scramble	51 Southside City	52 Golden Ponds
53 Rural Industria	54 Norma Rae-Ville	55 Mines & Mills	56 Agri-Business
57 Grain Belt	58 Blue Highways	59 Rustic Elders	60 Back Country Folks
61 Scrub Pine Flats	62 Hard Scrabble		

Source: Claritas, Inc.

The description for cluster 01, Blue Blood Estates, is: “Established executives, professionals, and ‘old money’ heirs that live in America’s wealthiest suburbs. They are accustomed to privilege and live luxuriously—one-tenth of this group’s members are multimillionaires. The next affluence level is a sharp drop from this pinnacle.”

Examples of clustering tasks in business and research include:

- Target marketing of a niche product for a small-capitalization business that does not have a large marketing budget
- For accounting auditing purposes, to segmentize financial behavior into benign and suspicious categories
- As a dimension-reduction tool when the data set has hundreds of attributes
- For gene expression clustering, where very large quantities of genes may exhibit similar behavior

Clustering is often performed as a preliminary step in a data mining process, with the resulting clusters being used as further inputs into a different technique downstream, such as neural networks. We discuss hierarchical and k-means clustering in Chapter 8 and Kohonen networks in Chapter 9. An application of clustering is examined in Case Study 5.

Association

The *association* task for data mining is the job of finding which attributes “go together.” Most prevalent in the business world, where it is known as *affinity analysis* or *market basket analysis*, the task of association seeks to uncover rules for quantifying the relationship between two or more attributes. Association rules are of the form “If *antecedent*, then *consequent*,” together with a measure of the support and confidence associated with the rule. For example, a particular supermarket may find that of the 1000 customers shopping on a Thursday night, 200 bought diapers, and of those 200 who bought diapers, 50 bought beer. Thus, the association rule would be “If buy diapers, then buy beer” with a support of $200/1000 = 20\%$ and a confidence of $50/200 = 25\%$.

Examples of association tasks in business and research include:

- Investigating the proportion of subscribers to a company’s cell phone plan that respond positively to an offer of a service upgrade
- Examining the proportion of children whose parents read to them who are themselves good readers
- Predicting degradation in telecommunications networks
- Finding out which items in a supermarket are purchased together and which items are never purchased together
- Determining the proportion of cases in which a new drug will exhibit dangerous side effects

We discuss two algorithms for generating association rules, the a priori algorithm and the GRI algorithm, in Chapter 10. Association rules were utilized in Case Study 1. We examine another application of association rules in Case Study 3.

Next we examine four case studies, each of which demonstrates a particular data mining task in the context of the CRISP–DM data mining standard process.

CASE STUDY 2

PREDICTING ABNORMAL STOCK MARKET RETURNS USING NEURAL NETWORKS [20]

1. Business/Research Understanding Phase

Alan M. Safer, of California State University–Long Beach, reports that stock market trades made by insiders usually have abnormal returns. Increased profits can be made by outsiders using legal insider trading information, especially by focusing on attributes such as company size and the time frame for prediction. Safer is interested in using data mining methodology to increase the ability to predict abnormal stock price returns arising from legal insider trading.

2. Data Understanding Phase

Safer collected data from 343 companies, extending from January 1993 to June 1997 (the source of the data being the Securities and Exchange Commission). The stocks used in the study were all of the stocks that had insider records for the entire period and were in the S&P 600, S&P 400, or S&P 500 (small, medium, and large capitalization, respectively) as of June 1997. Of the 946 resulting stocks that met this description, Safer chose only those stocks that underwent at least two purchase orders per year, to assure a sufficient amount of transaction data for the data mining analyses. This resulted in 343 stocks being used for the study. The variables in the original data set include the company, name and rank of the insider, transaction date, stock price, number of shares traded, type of transaction (buy or sell), and number of shares held after the trade. To assess an insider's prior trading patterns, the study examined the previous 9 and 18 weeks of trading history. The prediction time frames for predicting abnormal returns were established as 3, 6, 9, and 12 months.

3. Data Preparation Phase

Safer decided that the company rank of the insider would not be used as a study attribute, since other research had shown it to be of mixed predictive value for predicting abnormal stock price returns. Similarly, he omitted insiders who were uninvolved with company decisions. (Note that the present author does not necessarily agree with omitting variables prior to the modeling phase, because of earlier findings of mixed predictive value. If they are indeed of no predictive value, the models will so indicate, presumably. But if there is a chance of something interesting going on, the model should perhaps be given an opportunity to look at it. However, Safer is the domain expert in this area.)

4. Modeling Phase

The data were split into a training set (80% of the data) and a validation set (20%). A neural network model was applied, which uncovered the following results:

a. Certain industries had the most predictable abnormal stock returns, including:
   - *Industry group 36:* electronic equipment, excluding computer equipment
   - *Industry Group 28:* chemical products
   - *Industry Group 37:* transportation equipment
   - *Industry Group 73:* business services

b. Predictions that looked further into the future (9 to 12 months) had increased ability to identify unusual insider trading variations than did predictions that had a shorter time frame (3 to 6 months).

c. It was easier to predict abnormal stock returns from insider trading for small companies than for large companies.

5. Evaluation Phase

Safer concurrently applied a multivariate adaptive regression spline (MARS, not covered here) model to the same data set. The MARS model uncovered many of the same findings as the neural network model, including results (a) and (b) from the modeling phase. Such a *confluence of results* is a powerful and elegant method for evaluating the quality and effectiveness of the model, analogous to getting two independent judges to concur on a decision. Data miners should strive to produce such a confluence of results whenever the opportunity arises. This is possible because often more than one data mining method may be applied appropriately to the problem at hand. If both models concur as to the results, this strengthens our confidence in the findings. If the models disagree, we should probably investigate further. Sometimes, one type of model is simply better suited to uncovering a certain type of result, but sometimes, disagreement indicates deeper problems, requiring cycling back to earlier phases.

6. Deployment Phase

The publication of Safer's findings in *Intelligent Data Analysis* [20] constitutes one method of model deployment. Now, analysts from around the world can take advantage of his methods to track the abnormal stock price returns of insider trading and thereby help to protect the small investor.

CASE STUDY 3

MINING ASSOCIATION RULES FROM LEGAL DATABASES [21]

1. Business/Research Understanding Phase

The researchers, Sasha Ivkovic and John Yearwood of the University of Ballarat, and Andrew Stranieri of La Trobe University, Australia, are interested in whether interesting and actionable association rules can be uncovered in a large data set containing information on applicants for government-funded legal aid in Australia. Because most legal data is not structured in a manner easily suited to most data mining techniques, application of knowledge discovery methods to legal data has not developed as quickly as in other areas. The researchers' goal is to improve the delivery of legal services and just outcomes in law, through improved use of available legal data.

2. Data Understanding Phase

The data are provided by Victoria Legal Aid (VLA), a semigovernmental organization that aims to provide more effective legal aid for underprivileged people in Australia. Over 380,000 applications for legal aid were collected from the 11 regional offices of VLA, spanning 1997–1999, including information on more than 300 variables. In an effort to reduce the number of variables, the researchers turned to domain experts for assistance. These experts selected seven of the most important variables for inclusion in the data set: gender, age, occupation, reason for refusal of aid, law type (e.g., civil law), decision (i.e., aid granted or not granted), and dealing type (e.g., court appearance).

3. Data Preparation Phase

The VLA data set turned out to be relatively clean, containing very few records with missing or incorrectly coded attribute values. This is in part due to the database management system used by the VLA, which performs quality checks on input data. The age variable was partitioned into discrete intervals such as “under 18,” “over 50,” and so on.

4. Modeling Phase

Rules were restricted to having only a single antecedent and a single consequent. Many interesting association rules were uncovered, along with many uninteresting rules, which is the typical scenario for association rule mining. One such interesting rule was: *If place of birth = Vietnam, then law type = criminal law*, with 90% confidence.

The researchers proceeded on the accurate premise that association rules are interesting if they spawn interesting hypotheses. A discussion among the researchers and experts for the reasons underlying the association rule above considered the following hypotheses:

- *Hypothesis A:* Vietnamese applicants applied for support only for criminal law and not for other types, such as family and civil law.
- *Hypothesis B:* Vietnamese applicants committed more crime than other groups.
- *Hypothesis C:* There is a lurking variable. Perhaps Vietnamese males are more likely than females to apply for aid, and males are more associated with criminal law.
- *Hypothesis D:* The Vietnamese did not have ready access to VLA promotional material.

The panel of researchers and experts concluded informally that hypothesis A was most likely, although further investigation is perhaps warranted, and no causal link can be assumed. Note, however, the intense human interactivity throughout the data mining process. Without the domain experts’ knowledge and experience, the data mining results in this case would not have been fruitful.

5. Evaluation Phase

The researchers adopted a unique evaluative methodology for their project. They brought in three domain experts and elicited from them their estimates of the confidence levels for each of 144 association rules. These estimated confidence levels were then compared with the actual confidence levels of the association rules uncovered in the data set.

6. Deployment Phase

A useful Web-based application, WebAssociator, was developed, so that nonspecialists could take advantage of the rule-building engine. Users select the single antecedent and single consequent using a Web-based form. The researchers suggest that WebAssociator could be deployed as part of a judicial support system, especially for identifying unjust processes.

CASE STUDY 4

PREDICTING CORPORATE BANKRUPTCIES USING DECISION TREES [22]

1. Business/Research Understanding Phase

The recent economic crisis in East Asia has spawned an unprecedented level of corporate bankruptcies in that region and around the world. The goal of the researchers, Tae Kyung Sung from Kyonggi University, Namsik Chang from the University of Seoul, and Gunhee Lee of Sogang University, Korea, is to develop models for predicting corporate bankruptcies that maximize the interpretability of the results. They felt that interpretability was important because a negative bankruptcy prediction can itself have a devastating impact on a financial institution, so that firms that are predicted to go bankrupt demand strong and logical reasoning.

If one's company is in danger of going under, and a prediction of bankruptcy could itself contribute to the final failure, that prediction better be supported by solid "trace-able" evidence, not by a simple up/down decision delivered by a black box. Therefore, the researchers chose decision trees as their analysis method, because of the transparency of the algorithm and the interpretability of results.

2. Data Understanding Phase

The data included two groups, Korean firms that went bankrupt in the relatively stable growth period of 1991–1995, and Korean firms that went bankrupt in the economic crisis conditions of 1997–1998. After various screening procedures, 29 firms were identified, mostly in the manufacturing sector. The financial data was collected directly from the Korean Stock Exchange, and verified by the Bank of Korea and the Korea Industrial Bank.

3. Data Preparation Phase

Fifty-six financial ratios were identified by the researchers through a search of the literature on bankruptcy prediction, 16 of which were then dropped due to duplication. There remained 40 financial ratios in the data set, including measures of growth, profitability, safety/leverage, activity/efficiency, and productivity.

4. Modeling Phase

Separate decision tree models were applied to the "normal-conditions" data and the "crisis-conditions" data. As we shall learn in Chapter 6, decision tree models can easily generate rule sets. Some of the rules uncovered for the normal-conditions data were as follows:

- If the productivity of capital is greater than 19.65, predict *nonbankrupt* with 86% confidence.
- If the ratio of cash flow to total assets is greater than −5.65, predict *nonbankrupt* with 95% confidence.
- If the productivity of capital is at or below 19.65 *and* the ratio of cash flow to total assets is at or below −5.65, predict *bankrupt* with 84% confidence.

Some of the rules uncovered for the crisis-conditions data were as follows:

- If the productivity of capital is greater than 20.61, predict *nonbankrupt* with 91% confidence.
- If the ratio of cash flow to liabilities is greater than 2.64, predict *nonbankrupt* with 85% confidence.
- If the ratio of fixed assets to stockholders' equity and long-term liabilities is greater than 87.23, predict *nonbankrupt* with 86% confidence.
- If the productivity of capital is at or below 20.61, *and* the ratio of cash flow to liabilities is at or below 2.64, *and* the ratio of fixed assets to stockholders' equity and long-term liabilities is at or below 87.23, predict *bankrupt* with 84% confidence.

Cash flow and *productivity of capital* were found to be important regardless of the economic conditions. While *cash flow* is well known in the bankruptcy prediction literature, the identification of *productivity of capital* was relatively rare, which therefore demanded further verification.

5. Evaluation Phase

The researchers convened an expert panel of financial specialists, which unanimously selected *productivity of capital* as the most important attribute for differentiating firms in danger of bankruptcy from other firms. Thus, the unexpected results discovered by the decision tree model were verified by the experts.

To ensure that the model was generalizable to the population of all Korean manufacturing firms, a control sample of nonbankrupt firms was selected, and the attributes of the control sample were compared to those of the companies in the data set. It was found that the control sample's average assets and average number of employees were within 20% of the data sample.

Finally, the researchers applied multiple discriminant analysis as a performance benchmark. Many of the 40 financial ratios were found to be significant predictors of bankruptcy, and the final discriminant function included variables identified by the decision tree model.

6. Deployment Phase

There was no deployment identified per se. As mentioned earlier, deployment is often at the discretion of users. However, because of this research, financial institutions in Korea are now better aware of the predictors for bankruptcy for crisis conditions, as opposed to normal conditions.

CASE STUDY 5

PROFILING THE TOURISM MARKET USING *k*-MEANS CLUSTERING ANALYSIS [23]

1. Business/Research Understanding Phase

The researchers, Simon Hudson and Brent Ritchie, of the University of Calgary, Alberta, Canada, are interested in studying intraprovince tourist behavior in Alberta. They would like to create profiles of domestic Albertan tourists based on the decision behavior of the tourists. The overall goal of the study was to form a quantitative basis for the development of an intraprovince marketing campaign, sponsored by Travel Alberta. Toward this goal, the main objectives were to determine which factors were important in choosing destinations in Alberta, to evaluate the domestic perceptions of the "Alberta vacation product," and to attempt to comprehend the travel decision-making process.

2. Data Understanding Phase

The data were collected in late 1999 using a phone survey of 13,445 Albertans. The respondents were screened according to those who were over 18 and had traveled for leisure at least 80 kilometers for at least one night within Alberta in the past year. Only 3071 of these 13,445 completed the survey and were eligible for inclusion in the study.

3. Data Preparation Phase

One of the survey questions asked the respondents to indicate to what extent each of the factors from a list of 13 factors most influence their travel decisions. These were then considered to be variables upon which the cluster analysis was performed, and included such factors as the quality of accommodations, school holidays, and weather conditions.

4. Modeling Phase

Clustering is a natural method for generating segment profiles. The researchers chose *k*-means clustering, since that algorithm is quick and efficient as long as you know the number of clusters you expect to find. They explored between two and six cluster models before settling on a five-cluster solution as best reflecting reality. Brief profiles of the clusters are as follows:

- *Cluster 1: the young urban outdoor market*. Youngest of all clusters, equally balanced genderwise, with school schedules and budgets looming large in their travel decisions.
- *Cluster 2: the indoor leisure traveler market.* Next youngest and very female, mostly married with children, with visiting family and friends a major factor in travel plans.
- *Cluster 3: the children-first market*. More married and more children than any other cluster, with children's sports and competition schedules having great weight in deciding where to travel in Alberta.
- *Cluster 4: the fair-weather-friends market*. Second-oldest, slightly more male group, with weather conditions influencing travel decisions.
- *Cluster 5: the older, cost-conscious traveler market*. The oldest of the clusters, most influenced by cost/value considerations and a secure environment when making Alberta travel decisions.

5. Evaluation Phase

Discriminant analysis was used to verify the “reality” of the cluster categorizations, correctly classifying about 93% of subjects into the right clusters. The discriminant analysis also showed that the differences between clusters were statistically significant.

6. Deployment Phase

These study findings resulted in the launching of a new marketing campaign, “Alberta, Made to Order,” based on customizing the marketing to the cluster types uncovered in the data mining. More than 80 projects were launched, through a cooperative arrangement between government and business. “Alberta, Made to Order,” television commercials have now been viewed about 20 times by over 90% of adults under 55. Travel Alberta later found an increase of over 20% in the number of Albertans who indicated Alberta as a “top-of-the-mind” travel destination.

REFERENCES

1. Peter Fabris, Advanced navigation, *CIO Magazine*, May 15, 1998, `http://www.cio.com/archive/051598_mining.html`.
2. Bill Clinton, New York University speech, *Salon.com*, December 6, 2002, `http://www.salon.com/politics/feature/2002/12/06/clinton/print.html`.
3. *Mining Data to Save Children with Brain Tumors*, SPSS, Inc., `http://spss.com/success/`.
4. The Gartner Group, `www.gartner.com`.
5. David Hand, Heikki Mannila, and Padhraic Smyth, *Principles of Data Mining*, MIT Press, Cambridge, MA, 2001.
6. Peter Cabena, Pablo Hadjinian, Rolf Stadler, Jaap Verhees, and Alessandro Zanasi, *Discovering Data Mining: From Concept to Implementation*, Prentice Hall, Upper Saddle River, NJ, 1998.
7. Rachel Konrad, Data mining: Digging user info for gold, *ZDNET News*, February 7, 2001, `http://zdnet.com.com/2100-11-528032.html?legacy=zdnn`.
8. The Technology Review Ten, *MIT Technology Review*, January/February 2001.
9. Jennifer Mateyaschuk, The 1999 National IT Salary Survey: Pay up, *Information Week*, `http://www.informationweek.com/731/salsurvey.htm`.
10. The Boston Celtics, `http://www.nba.com/celtics/`.
11. Peter Gwynne, Digging for data, *Think Research*, `domino.watson.ibm.com/comm/wwwr_thinkresearch.nsf/pages/datamine296.html`.
12. John Naisbitt, *Megatrends*, 6th ed., Warner Books, New York, 1986.
13. Michael Berry and Gordon Linoff, *Data Mining Techniques for Marketing, Sales and Customer Support*, Wiley, Hoboken, NJ, 1997.
14. Michael Berry and Gordon Linoff, *Mastering Data Mining*, Wiley, Hoboken, NJ, 2000.
15. Quoted in: Mihael Ankerst, The perfect data mining tool: Interactive or automated? Report on the SIGKDD-2002 Panel, *SIGKDD Explorations*, Vol. 5, No. 1, July 2003.
16. Peter Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinart, Colin Shearer, and Rudiger Wirth, *CRISP–DM Step-by-Step Data Mining Guide*, 2000, `http://www.crisp-dm.org/`.
17. Jochen Hipp and Guido Lindner, Analyzing warranty claims of automobiles: an application description following the CRISP–DM data mining process, in *Proceedings of the*

5th International Computer Science Conference (ICSC '99), pp. 31–40, Hong Kong, December 13–15, 1999, © Springer.

18. Jen Que Louie, President of Nautilus Systems, Inc. (`www.nautilus-systems.com`), testimony before the U.S. House of Representatives Subcommittee on Technology, Information Policy, Intergovernmental Relations, and Census, *Congressional Testimony*, March 25, 2003.
19. `www.Claritas.com`.
20. Alan M. Safer, A comparison of two data mining techniques to predict abnormal stock market returns, *Intelligent Data Analysis*, Vol. 7, pp. 3–13, 2003.
21. Sasha Ivkovic, John Yearwood, and Andrew Stranieri, Discovering interesting association rules from legal databases, *Information and Communication Technology Law*, Vol. 11, No. 1, 2002.
22. Tae Kyung Sung, Namsik Chang, and Gunhee Lee, Dynamics of modeling in data mining: interpretive approach to bankruptcy prediction, *Journal of Management Information Systems,* Vol. 16, No. 1, pp. 63–85, 1999.
23. Simon Hudson and Brent Richie, Understanding the domestic market using cluster analysis: a case study of the marketing efforts of Travel Alberta, *Journal of Vacation Marketing*, Vol. 8, No. 3, pp. 263–276, 2002.

EXERCISES

1. Refer to the Bank of America example early in the chapter. Which data mining task or tasks are implied in identifying "the type of marketing approach for a particular customer, based on the customer's individual profile"? Which tasks are not explicitly relevant?

2. For each of the following, identify the relevant data mining task(s):

a. The Boston Celtics would like to approximate how many points their next opponent will score against them.

b. A military intelligence officer is interested in learning about the respective proportions of Sunnis and Shias in a particular strategic region.

c. A NORAD defense computer must decide immediately whether a blip on the radar is a flock of geese or an incoming nuclear missile.

d. A political strategist is seeking the best groups to canvass for donations in a particular county.

e. A homeland security official would like to determine whether a certain sequence of financial and residence moves implies a tendency to terrorist acts.

f. A Wall Street analyst has been asked to find out the expected change in stock price for a set of companies with similar price/earnings ratios.

3. For each of the following meetings, explain which phase in the CRISP–DM process is represented:

a. Managers want to know by next week whether deployment will take place. Therefore, analysts meet to discuss how useful and accurate their model is.

b. The data mining project manager meets with the data warehousing manager to discuss how the data will be collected.

c. The data mining consultant meets with the vice president for marketing, who says that he would like to move forward with customer relationship management.

- **d.** The data mining project manager meets with the production line supervisor to discuss implementation of changes and improvements.
- **e.** The analysts meet to discuss whether the neural network or decision tree models should be applied.

4. Discuss the need for human direction of data mining. Describe the possible consequences of relying on completely automatic data analysis tools.

5. CRISP–DM is not the only standard process for data mining. Research an alternative methodology. (*Hint:* SEMMA, from the SAS Institute.) Discuss the similarities and differences with CRISP–DM.

6. Discuss the lessons drawn from Case Study 1. Why do you think the author chose a case study where the road was rocky and the results less than overwhelming?

7. Consider the business understanding phase of Case Study 2.

- **a.** Restate the research question in your own words.
- **b.** Describe the possible consequences for any given data mining scenario of the data analyst not completely understanding the business or research problem.

8. Discuss the evaluation method used for Case Study 3 in light of Exercise 4.

9. Examine the association rules uncovered in Case Study 4.

- **a.** Which association rule do you think is most useful under normal conditions? Under crisis conditions?
- **b.** Describe how these association rules could be used to help decrease the rate of company failures in Korea.

10. Examine the clusters found in Case Study 5.

- **a.** Which cluster do you find yourself or your relatives in?
- **b.** Describe how you would use the information from the clusters to increase tourism in Alberta.