CHAPTER

3

The Data Mining Process

Chapter 1 describes the virtuous cycle of data mining as a business process that divides data mining into four stages:

1. Identifying the problem
2. Transforming data into information
3. Taking action
4. Measuring the outcome

This chapter shifts the emphasis to data mining as a technical process, moving from identifying business problems to translating business problems into data mining problems. The second stage, transforming data into information, is expanded into several topics including hypothesis testing, model building, and pattern discovery. The ideas and best practices introduced in this chapter are elaborated further in the rest of the book. The purpose of this chapter is to bring the different styles of data mining together in one place.

The best way to avoid breaking the virtuous cycle of data mining is to understand the ways it is likely to fail and take preventive measures. Over the years, the authors have encountered many ways for data mining projects to go wrong. This chapter begins with a discussion of some of these pitfalls. The rest of the chapter is about the data mining process. Later chapters cover the aspects of data mining methodology that are specific to the particular styles of data

mining — directed data mining and undirected data mining. This chapter focuses on what these approaches have in common.

The three main styles of data mining are introduced, beginning with the simplest approach — testing hypotheses typically by using ad hoc queries — and working up to more sophisticated activities such as building models that can be used for scoring, and finding patterns using undirected data mining techniques. The theme of the chapter is getting from a clear statement of the business goal to a clear understanding of the data mining tasks required to achieve the goal and the data mining techniques appropriate to the task.

What Can Go Wrong?

Data mining is a way of learning from the past in order to make better decisions in the future. The best practices described in this chapter are designed to avoid two undesirable outcomes of the learning process:

- Learning things that aren't true.
- Learning things that are true, but not useful.

Ancient mariners learned to avoid the rocks of Scylla and the whirlpool of Charybdis that protect the narrow straits between Sicily and the Italian mainland. Like the ancient sailors who learned to avoid these threats, data miners need to know how to avoid common dangers.

Learning Things That Aren't True

Learning things that aren't true is more dangerous than learning things that are useless because important business decisions may be made based on incorrect information. Data mining results often seem reliable because they are based on actual data processed in a seemingly scientific manner. This appearance of reliability can be deceiving. The data may be incorrect or not relevant to the question at hand. The patterns discovered may reflect past business decisions or nothing at all. Data transformations such as summarization may have destroyed or hidden important information. The following sections discuss some of the more common problems that can lead to false conclusions.

WARNING **The most careful and painstaking analysis, using the most sophisticated techniques, yields incorrect results when the data analyzed is incorrect or simply not relevant. In information technology circles, a popular aphorism is, "garbage in, garbage out."**

Patterns May Not Represent Any Underlying Rule

It is often said that figures don't lie, but liars do figure. When it comes to finding patterns in data, figures don't have to actually lie in order to suggest things that aren't true. So many ways to construct patterns exist that any random set of data points reveals one if examined long enough. Human beings depend so heavily on patterns in our lives that we tend to see them even when they are not there. We look up at the nighttime sky and see not a random arrangement of stars, but the Big Dipper, or the Southern Cross, or Orion's Belt. Some even see astrological patterns and portents that can be used to predict the future. The widespread acceptance of outlandish conspiracy theories is further evidence of the human need to find patterns.

Presumably, the reason that humans have evolved such an affinity for patterns is that patterns often do reflect some underlying truth about the way the world works. The phases of the moon, the progression of the seasons, the constant alternation of night and day, even the regular appearance of a favorite TV show at the same time on the same day of the week are useful because they are stable and therefore predictive. We can use these patterns to decide when it is safe to plant tomatoes, when to eat breakfast, and how to program the DVR. Other patterns clearly do not have any predictive power. If a fair coin comes up heads five times in a row, there is still a 50–50 chance that it will come up tails on the sixth toss.

The challenge for data miners is to figure out which patterns are useful and which are not. Consider the following patterns, all of which have been cited in articles in the popular press as if they had predictive value:

- The party that does not hold the presidency picks up seats in Congress during off-year elections.
- When the American League wins the World Series, Republicans take the White House.
- When the Washington Redskins win their last home game, the incumbent party keeps the White House.
- In U.S. presidential contests, the taller man usually wins.

The first pattern (the one involving off-year elections) is explainable in purely political terms. Every four years, just over half of the American voters get all excited and vote for their candidate for president. A few months later, the candidate takes over, and the disappointment begins — politicians simply cannot keep all the promises that their base expects. Two years later, in the Congressional elections, a backlash occurs, usually caused by disappointed supporters who

do not turn out to vote. Because this pattern has an underlying explanation, it seems likely to continue into the future, implying that it has predictive value.

The next two alleged predictors, the ones involving sporting events, seem just as clearly to have no predictive value. No matter how many times Republicans and the American League may have shared victories in the past (and the authors have not researched this point), there is no reason to expect the association to continue in the future.

What about candidates' heights? Since 1948 when Truman (who was short, but taller than Dewey) was elected, the election in which Carter beat Ford and the one in which Bush beat Kerry are the only ones where the shorter candidate won more popular votes. The 2000 election that pitted 6'1" Gore against the 6'0" Bush still fits the pattern, if one assumes that the pattern relates to winning the popular vote rather than the electoral vote. In 2008, the basketball-playing Obama out-polled the shorter McCain. Height does not seem to have anything to do with the job of being president. However, our language exhibits "heightism": we look *up* to people as a sign of respect, and look *down* on people to show disdain. Height is associated with better childhood nutrition, which in turn leads to increased intelligence and other indicators of social success. As this chapter explains, the right way to decide whether a rule is stable and predictive is to compare its performance on multiple samples selected at random from the same population. In the case of presidential height, the authors leave this as an exercise for the reader. As is often the case, the hardest part of the task is collecting the data — before the age of Google, determining the heights of unsuccessful presidential candidates from previous centuries was not easy!

The technical term for finding patterns that fail to generalize is *overfitting*. Overfitting leads to unstable models that work one day, but not the next, on one data set but not on another. Building stable models is the primary goal of the data mining methodology.

The Model Set May Not Reflect the Relevant Population

The *model set* is the data used to create a data mining model, and it necessarily describes what happened in the past. The model can only be as good as the data used to create it. For inferences to be valid, the model set must reflect the population that the model is meant to describe, classify, or score. A sample that does not properly reflect the population being scored or the overall population is *biased*.

A biased model set can lead to learning things that are not true. Unless the biases are taken into account, the resulting model is also biased. Biases can be hard to avoid. Consider:

- Customers are not like prospects.
- Survey responders are not like non-responders.
- People who read e-mail are not like people who do not read e-mail.

- People who register on a website are not like people who do not register.
- After an acquisition, customers from the acquired company are not necessarily like customers from the acquirer.
- Records with no missing values reflect a different population from records with missing values.

Consider the first point. Customers are not like prospects because they represent people who responded positively to whatever messages, offers, and promotions were made to attract customers in the past. A study of current customers is likely to suggest more of the same. If past campaigns have gone after wealthy, urban consumers, then any comparison of current customers with the general population would likely show that customers tend to be wealthy and urban. Such a model may miss opportunities in middle-income suburbs.

TIP Careful attention to selecting and sampling data for the model set is crucial to successful data mining.

The consequences of using a biased sample can be worse than simply a missed marketing opportunity. In the United States, there is a history of "redlining," the illegal practice of refusing to write loans or insurance policies in certain neighborhoods (usually low-income or minority neighborhoods). A search for patterns in the historical data from a company that had a history of redlining would reveal that people in certain neighborhoods are unlikely to be customers at all. If future marketing efforts were based on that finding, data mining would help perpetuate illegal and unethical practices.

Data May Be at the Wrong Level of Detail

In more than one industry, the authors have been told that usage often goes down in the month before a customer leaves. Upon closer examination, this may turn out to be an example of learning something that is not true. Figure 3-1 shows the monthly minutes of use for a group of cellular telephone subscribers who are recorded as stopping in month nine. For seven months, the subscribers use about 100 minutes per month. In the eighth month, their usage declines to about half that. And in the following month, there is no usage at all, because the subscribers have stopped. This suggests that a marketing effort triggered by a decline in usage might be able to save these customers.

These subscribers appear to fit a pattern where a month with decreased usage precedes abandonment of the service. Appearances are deceiving. These customers have no usage in month nine because the actual stop date is in month eight. On average, the stop date would be halfway through the month. These customers continue to use the service at a constant rate until they stop, presumably because on that day, the customers begin using a competing service. The putative period of declining usage does not actually exist and so certainly does

not provide a window of opportunity for retaining the customer. What appears to be a leading indicator is actually a trailing one.

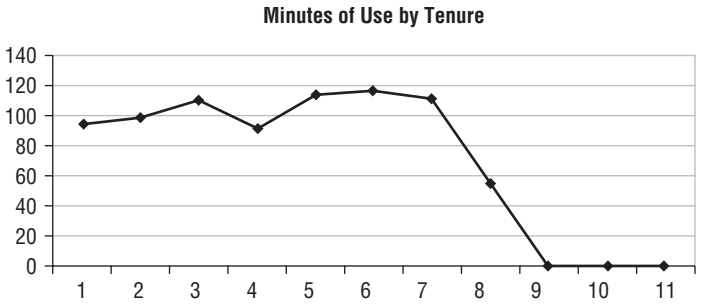


Figure 3-1: Does declining usage in month 8 predict attrition in month 9?

Figure 3-2 shows another example of confusion caused by aggregation. Sales appear to be down in October compared to August and September. The picture comes from a business that has sales activity only on days when the financial markets are open. Because of the way that weekends and holidays fell in 2003, October had fewer trading days than August and September. That fact alone accounts for the entire drop-off in sales.

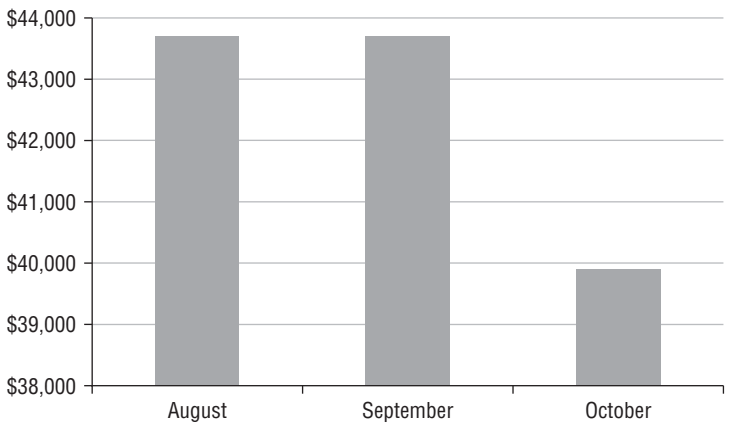


Figure 3-2: Did sales really drop off in October?

In the previous examples, aggregation leads to confusion. Failure to aggregate to the appropriate level can also lead to confusion. One member of a household might have a checking account with a low balance and little activity while another member of the same household has multiple large accounts. Treating the small account holder as a less-than-valuable customer might put the relationship with the entire household at risk. In this case, a total balance figure may be more important than the balance in any one account.

TIP When summarizing data, choose a level of aggregation that does not hide important patterns within a single period. A business with strong week-to-week changes should not report activity summarized monthly.

Learning Things That Are True, but Not Useful

Although not as dangerous as learning things that aren't true, learning things that aren't useful is more common. This can happen in several ways.

Learning Things That Are Already Known (or Should Be Known)

Data mining should provide new information. Many of the strongest patterns in data represent things that are already known. People over retirement age tend not to respond to offers for retirement savings plans. People living outside home delivery zones do not become newspaper subscribers. Even though they may respond to subscription offers, service never starts. People who do not own cars do not purchase car insurance.

Data mining can also discover patterns that should have been known to be true. In one interesting example, the authors were working on a project analyzing purchase patterns in grocery store data. When the first set of data arrived, we set out to find products that are purchased together. The first combinations were "eggs and meat," "eggs and milk," and "eggs and soda." The rules continued in the same vein — eggs were rushing off the shelves with just about every product in the store. At first, this seemed like a potential problem in the data. Then, one of our colleagues noted that the data came from the week before Easter. And, indeed, when people go grocery shopping before Easter, they often buy eggs to dye or hide for the Easter holiday.

The strongest patterns often reflect business rules. If data mining "discovers" that people who have anonymous call blocking also have caller ID, the reason is perhaps because anonymous call blocking is only sold as part of a bundle of services that also includes caller ID. If data mining "discovers" that maintenance agreements are sold with large appliances (as Sears once found), that is because maintenance agreements are almost always sold after the appliance. Not only are these patterns uninteresting, their strength may obscure less obvious but more actionable patterns.

Learning things that are already known does serve one useful purpose. It demonstrates that, on a technical level, the data mining techniques are working and the data is reasonably accurate. This can be comforting, even if not helpful. When data mining techniques are powerful enough to discover things that are known to be true, there is reason to believe that they can discover more useful patterns as well.

Learning Things That Can't Be Used

Data mining can uncover relationships that are both true and previously unknown, but still hard to make use of. Sometimes the problem is regulatory. A customer's wireless calling patterns may suggest an affinity for certain landline long-distance packages, but a company that provides both services may not be allowed to take advantage of the fact due to legal restrictions. Similarly, a customer's credit history may be predictive of future insurance claims, but regulators may prohibit making underwriting decisions based on such information. Or, in what is becoming a more prevalent example, a person's genetic material may suggest propensity for certain diseases — a characteristic that insurance companies in the United States and most European countries are barred from using.

Other times, data mining reveals that important outcomes are outside the company's control. A product may be more appropriate for some climates than others, but it is hard to change the weather. Mobile phone service may be worse in some regions for reasons of topography, but that is also hard to change.

TIP A study of customer attrition may show that a strong predictor of customers leaving is the way they were acquired. It is too late to go back and change that for existing customers, but that does not make the information useless. Future attrition can be reduced by changing the mix of acquisition channels to favor those that bring in longer-lasting customers.

Data miners must take care to steer clear of the Scylla of learning things that aren't true and the Charybdis of not learning anything useful. The methodologies laid out in Chapter 5 and Chapter 12 are designed to ensure that data mining efforts lead to stable models that successfully address business problems.

Data Mining Styles

Chapter 1 says that data mining involves the "exploration and analysis of large quantities of data to produce meaningful results." That is a broad enough definition to cover many different approaches. These come in three main styles:

- Hypothesis testing
- Directed data mining
- Undirected data mining

In hypothesis testing, the goal is to use data to answer questions or gain understanding. In directed data mining, the goal is to construct a model that explains or predicts one or more particular target variables. In undirected data mining, the goal is to find overall patterns that are not tied to a particular target. During

the course of a data mining project, you may spend time working in any or all of these styles depending on the nature of the problem and your familiarity with the data.

Although the three styles of data mining have some technical differences, they also have much in common. Many of the topics discussed in Chapter 5 in the context of directed data mining are also important for hypothesis testing and finding patterns. In fact, the first three steps of the directed data mining methodology — translating a business problem into a data mining problem, selecting appropriate data, and getting to know the data — could just as well be covered in this chapter.

Hypothesis Testing

Hypothesis testing is a part of almost all data mining endeavors. Data miners often bounce back and forth between approaches, first thinking up possible explanations for observed behavior (often with the help of business experts) and letting those hypotheses dictate the data to be analyzed, and then letting the data suggest new hypotheses to test.

A *hypothesis* is a proposed explanation whose validity can be tested by analyzing data. Such data may simply be collected by observation or generated through an experiment, such as a test marketing campaign. Hypothesis testing sometimes reveals that the assumptions that have been guiding a company's actions are incorrect. For example, a company's advertising is based on a number of hypotheses about the target market for a product or service and the nature of the responses. It is worth testing whether these hypotheses are borne out by actual responses.

Depending on the hypotheses, this may mean interpreting a single value returned from a simple query, plowing through a collection of association rules generated by market basket analysis, determining the significance of a correlation found by a regression model, or designing a controlled experiment. In all cases, careful critical thinking is necessary to be sure that the result is not biased in unexpected ways. Proper evaluation of data mining results requires both analytical and business knowledge. Where these are not present in the same person, making good use of the new information requires cross-functional cooperation.

By its nature, hypothesis testing is ad hoc, but the process has some identifiable steps, the first and most important of which is generating good hypotheses to test. Next comes finding or generating data to confirm or disprove the hypotheses.

Generating Hypotheses

The key to generating hypotheses is getting diverse input from throughout the organization and, where appropriate, outside it as well. Outsiders may question things that insiders take for granted — perhaps providing valuable

insight. Often, all that is needed to start the ideas flowing is a clear statement of the problem itself — especially if it is something that has not previously been recognized as a problem.

More often than one might suppose, problems go unrecognized because they are not captured by the metrics used to evaluate performance. If a company has always measured its sales force on the number of new sales made each month, the salespeople may never have given much thought to the question of how long new customers remain active or how much they spend over the course of their relationship. When asked the right questions, however, the sales force may have insights into customer behavior that marketing, with its greater distance from the customer, has missed.

The goal is to come up with ideas that are both testable and actionable. Consider the following hypotheses:

- Most customers who accept a retention offer would stay anyway.
- Families with high-school age children are more likely than others to respond to a home equity line offer.
- Customers who buy more distinct product types have higher overall spending.

All of these propositions might or might not be true, and in each case, knowing the answer suggests some concrete action. If the first hypothesis is true, stop spending money to retain customers who are not at risk of leaving or find a better way of targeting retention offers to customers who really are going to leave. If the second hypothesis is true, continue the current marketing focus on this group. If the third hypothesis is correct, encourage salespeople to do more cross selling.

Testing Hypotheses Using Existing Data

It is often possible to test a new hypothesis by looking for evidence in existing historical data. For example, a manufacturer of medical devices sold to hospitals had the hypothesis that customers who bought products in multiple categories tended to spend more. As a first step, they looked at average sales by number of distinct products and produced the chart shown in Figure 3-3.

The chart clearly shows that customers who buy many kinds of product generate substantially more revenue per customer, but it does not show to what extent cross-selling drives additional revenue. Larger institutions naturally spend more, and perhaps they are also more likely to need products from multiple categories. Perhaps high revenue and multiple product categories are both driven by customer size — something not in the company's control. That, too, is a testable hypothesis: Group customers by size and type and look for a relationship between distinct products and revenue within each group.

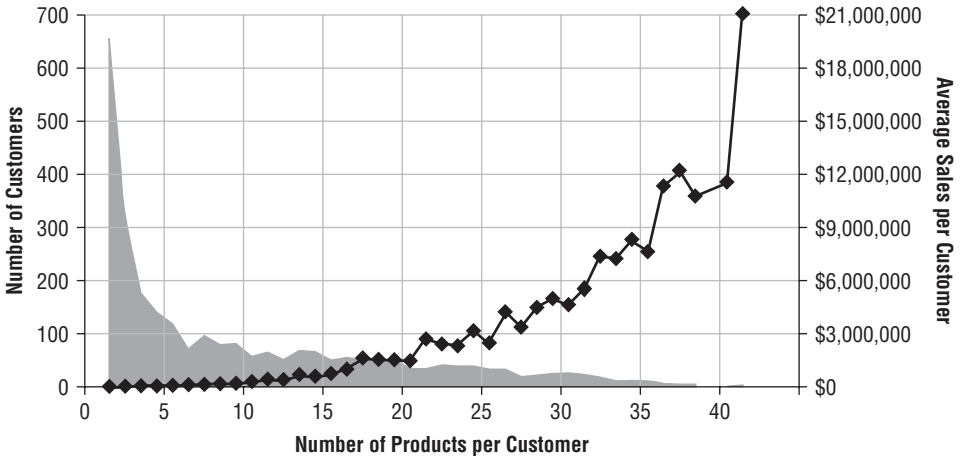


Figure 3-3: Customers who buy more product types spend more money.

Testing long-held beliefs can be harder because historical data reflects whatever assumptions have been made in the past. If families with high-school children have always been targeted for a particular product, this fact will be reflected in higher adoption rates for those families. This does not prove that they are the most responsive segment; some other group, such as small business owners, might have responded even more. In such cases, conducting a controlled experiment is advisable.

Small changes in what and how data is collected can greatly increase its value for analysis. For example, use different web addresses or call-in numbers in different ads and keep track of how each response comes in.

TIP **Each time a company solicits a response from its customers, whether through advertising or a more direct form of communication, it has an opportunity to gather information. Slight changes in the design of the communication, such as including a way to identify the channel when a prospect responds, can greatly increase the value of the data collected.**

Hypothesis Testing and Experimentation

Although many hypotheses can be tested against historical data, many cannot. Take the hypothesis that the people who accepted a retention offer would have stayed with or without the added enticement. Historical data describes who received the offer, who accepted the offer, and who ended up staying, but unless the campaign was set up as a proper experiment with a control group, it does not answer the question about what would have happened if the offer had not been made. This question cannot be answered by comparing retention for those who received the offer and those who did not because the two groups almost certainly differ in systematic ways.

If the offer went to customers considered to be at high risk for attrition, then people who did not get the offer may have better retention even if the offer really saved a large number of customers. On the other hand, if the offer went to customers considered particularly valuable, they may have better retention than non-recipients for reasons that have nothing to do with the offer. A valid test of the program's effectiveness requires comparing two groups of customers that are similar in every way except for the thing being tested. Data like that may not occur naturally, so you have to design an experiment to generate it. Experimental design and analysis is a broad field in statistics. This section covers some key points about specifics common to marketing tests.

Test and Control

The most basic experimental design involves creating two groups. One, known as the *test group* or *treatment group*, receives some sort of treatment such as an e-mail or phone call. The other group, known as the *control group*, does not receive the treatment. The two groups are picked to be as similar as possible — the same average age, the same average income, the same distribution of men and women, the same distribution of customer tenure, and so on. That may sound painstaking, but it is not. Basically, choose an overall group, and then randomly divide it into the test and control group. As long as the test and control groups are large enough, the laws of probability ensure that the groups are similar to each other (and to the whole population). If you want ensure that the groups are representative for certain key traits (say sex and tenure), then sort the population by these fields and take every *n*th record for the control group.

After the experiment, any significant difference between the groups can confidently be attributed to the treatment. Chapter 4 explains the concept of statistical significance and how to test for it.

A/B Tests

An A/B test compares two (or possibly more) treatments. Customers are randomly assigned to group A or group B. The two groups receive different treatments such as different advertising messages, web page layouts, prices, or payment options. Analytically oriented companies routinely run A/B tests to determine the effect of even seemingly minor changes because small changes can have large and unanticipated effects.

One online retailing company found that adding a box where customers could enter a discount coupon code reduced the proportion of customers who made purchases by a significant 6.5 percent. Most shoppers did not have coupons and apparently the invitation to supply a discount code caused people without one to think they were getting a bad deal. Perhaps such shoppers were encouraged to search for a coupon on Google, possibly finding a better price in the process.

A/B testing is usually associated with direct marketing and web-based retailing because in these environments controlling which customers get

which messages is relatively simple. A/B testing is also useful for less directed kinds of advertising such as billboards, radio, and television. The trick is to run different campaigns in similar markets. Such tests are called *paired tests*, because they depend on pairs of different markets (or store locations or whatever) to be as similar as possible for testing purposes. One half of the pair gets the treatment and the other half is the control. Chapter 9 discusses paired tests in more detail.

Champion/Challenger Tests

A common form of A/B testing compares a new treatment, the *challenger*, with the existing treatment, the *champion*. This idea is often applied to data mining models used to score customers. The new model should not be adopted until it is shown to be better than the old one.

Amazon.com is particularly adept at this form of A/B testing. Everything on its website — from the placement of product reviews and product descriptions to the number of user comments and keywords — has been tested against the "champion" best layout. In Amazon's live environment, visitors to the website are chosen randomly for the test group to see a modified layout. After a few hours or days, enough data has been gathered to suggest whether the tested modifications to the layout produce higher or lower sales than the champion. If improvements are significant, the test becomes the new champion.

Case Study in Hypothesis Testing: Measuring the Wrong Thing

This is a story about a company that makes recommendation software for retailing websites. Its clients, the retailers, leave some blank areas on particular web pages, such as the product pages, the shopping cart, and check-out pages. The recommendation software provides product recommendations to fill in the blanks when customers are shopping at the site. When a customer purchases the recommended item, the software company makes a commission. The goal, of course, is to increase the overall sales on the site, which benefits the retailers and encourages them to keep using the recommendation software.

The software company had a conundrum: According to all its metrics, its recommendations were improving year after year. More customers were clicking on and purchasing the recommended items. However, retailers complained that revenues were not rising as much as expected. In some head-to-head tests, the sophisticated recommendation software was not doing as well as simple general rules developed by clients.

This is not a well-formed problem for directed data mining. What is the target variable? It is also not a good candidate for undirected pattern finding; the pattern is all too clear. It is a perfect fit for hypothesis testing. The data miner's job was to brainstorm about what might be going wrong and then test the resulting hypotheses.

The software company approached Data Miners (the consulting company founded by the authors) to help make sense of this conundrum. We received data from an A/B test that had yielded disappointing results. In an A/B test, half the shoppers were randomly selected to receive recommendations from the company while the other half received competing recommendations from the retailer. This data included an order line table with details about each item such as its price, product category, and, in cases where the shopper had clicked on a recommendation for the product, a click ID. For each click, a clicks table showed which of several recommendation algorithms had generated the recommendation, and what item the shopper had been looking at when the recommendation was made.

Using simple SQL queries, we found that customers on our client's side of the test indeed clicked on more recommendations and, on both sides of the test, customers who clicked were more likely to make a purchase. More purchases should mean more money. And more money should mean the retailers are happy.

How could the A side — our client's side — lose given these metrics? The first clue was that the average price of items clicked was lower on the A side than the B side. Our first hypothesis was that A was recommending a different mix of products than B, but that was easily disproven. We kept trying out other hypotheses until we found two that, together, explained what was happening:

- The A side's recommendations yielded more substitutions and fewer cross-sells.
- Many of the A side's recommendations were down-sells.

Cross-sells are when consumers buy recommended products *in addition* to products they are already considering, resulting in a larger total purchase. A substitution is when consumers buy recommended products *instead* of the original ones. A cross-sell is more valuable to the retailer because it increases the amount the customer spends. However, our client's commission was only based on whether or not the end consumer purchased its recommendation. The retailer designed its recommendations to generate cross-sells. Where it did recommend substitutions, the recommended product was nearly always something pricier — up-selling. By comparison, our client's recommendations were down-sells on average.

Our conclusion was that our client had been measuring the wrong thing. Its recommendations "improved" over time in the sense of attracting more clicks, but clicks are not useful by themselves. The easiest way to attract clicks is to show shoppers cheaper substitutes for the items they are looking at. This behavior generated commissions for our client, but (inadvertently) at the expense of the retailer who ended up selling a cheaper item and paying a commission for the privilege! We recommended that the software company change its commission structure so it would be rewarded for incremental revenue rather than clicks: a valuable result from data mining using hypothesis testing.

Directed Data Mining

Directed data mining is another style of data mining. Directed data mining focuses on one or more variables that are targets, and the historical data contains examples of all the target values. In other words, directed data mining does not look for just any pattern in the data, but for patterns that explain the target values. A very typical example is retention modeling. The historical data contains examples of customers who are active and others who have stopped. The goal of directed data mining is to find patterns that differentiate between factors that cause customers to leave and customers to stay.

In statistics, the term *predictive modeling* is often used for directed data mining. In the authors' opinion, this is a bit of a misnomer, because although predictive modeling is definitely one aspect of directed data mining, it has other aspects, as well. Chapter 5 differentiates between predictive modeling and profile modeling, based on the temporal relationship between the target variable and the inputs. Predictive modeling is specifically when the target comes from a timeframe later than the inputs; profile modeling is specifically when the target and inputs come from the same timeframe.

Undirected Data Mining

Undirected data mining is a style of data mining that does not use a target variable, at least not explicitly. In directed data mining, different variables play different roles. Target variables are the objects of study; the rest of the variables are used to explain or predict the values of the targets. In undirected data mining, there are no special roles. The goal is to find overall patterns. After patterns have been detected, it is the responsibility of a person to interpret them and decide whether they are useful.

The term *undirected* may actually be a bit misleading. Although no target variable is used, business goals must still be addressed. The business goals addressed by undirected data mining may sound just as directed as any other goals; "Find examples of fraud," is an example of a business goal that might call for either directed or undirected data mining depending on whether training data contains identified fraudulent transactions. A directed approach would search for new records that are similar to cases known to be fraudulent. An undirected approach would look for new records that are unusual.

Increasing average order size is another example of a business goal that could be addressed using undirected data mining. Association rules, an undirected data mining technique, reveal patterns about which items are frequently sold together. This information could be used to increase order sizes through improved cross-selling.

Sometimes, the business goals themselves may be a bit vague and the data mining effort is a way to refine them. For example, a company might have a goal

of developing specialized services for different customer segments without having a clear idea of how customers should be segmented. Clustering, an undirected data mining technique, could be used to discover customer segments. Studying the segments might yield insight into what segment members have in common, which in turn might suggest common needs that a new product could address.

Goals, Tasks, and Techniques

A data mining consultant the authors know says that he lives in fear of clients reading a magazine article that mentions some particular data mining technique by name. When a vice president of marketing starts asking about neural networks versus support vector machines, it is probably time to reset the conversation. Data mining always starts with a business goal, and the first job of the data miner is to get a good understanding of that goal. This step requires good communication between people in upper management who set the goals and the analysts responsible for translating those goals into data mining tasks. The next job is to restate the business goal in terms of data mining tasks, and only then are particular data mining techniques selected.

Data Mining Business Goals

The data mining applications in the previous chapter provide several good examples of business goals:

- Choose the best places to advertise.
- Find the best locations for branches or stores.
- Acquire more profitable customers.
- Decrease exposure to risk of default.
- Improve customer retention.
- Detect fraudulent claims.

The rest of this book also contains many examples of data mining being used in the real world to solve real problems. Not all business goals lend themselves to data mining directly; sometimes they need to be turned into data mining business goals. For data mining to be successful, the business goal should be well-defined and directed towards particular efforts that are amenable to analysis using available data. A data mining business goal can usually be expressed in terms of something measurable such as incremental revenue, response rate, order size, or wait time.

Achieving any of these goals requires more than just data mining, of course, but data mining has an important role to play. The first step is to design a high-level

approach to the problem. To acquire more profitable customers, you might start by learning what drives profitability for existing customers and then recruit new customers with the right characteristics. Decreasing exposure to credit risk might mean predicting which of the customers currently in good standing are likely to go bad and preemptively decrease their credit lines. Improving customer retention might focus on improving the experience of existing customers or on recruiting new customers with longer expected tenures. The high-level approach suggests particular modeling tasks.

Data Mining Tasks

Data mining tasks are technical activities that can be described independently of any particular business goal. If a business goal is well-suited to data mining, it can usually be phrased in terms of the following tasks:

- Preparing data for mining
- Exploratory data analysis
- Binary response modeling (also called binary classification)
- Classification of discrete values and predictions
- Estimation of numeric values
- Finding clusters and associations
- Applying a model to new data

Data mining projects typically involve several of these tasks. Take the example of deciding which customers to include in a direct marketing campaign. Exploratory data analysis suggests which variables are important for characterizing customer response. These variables could then be used to find clusters of similar customers. A customer's cluster assignment could be an important explanatory variable in a binary response model. And, of course, the whole point of creating the model is to apply it to new data representing prospective customers to score them for propensity to respond to the campaign.

Preparing Data for Mining

Preparing data for mining is the subject of Chapters 18 through 20. The amount of effort required depends on the nature of the data sources and the requirements of particular data mining techniques. Some data preparation is nearly always required and it is not unusual for data preparation to be the most time-consuming part of a data mining project. Some data preparation is required to fix problems with the source data, but much of it is designed to enhance the information content of the data. Better data means better models.

Typically, data from a variety of sources must be combined to form a customer signature with one record per customer and a large number of fields to capture everything of interest about them. Because the source data is usually not at the customer level, building the customer signature requires many transformations. Transactions must be summarized in useful ways. Trends in time series might be captured as slopes or differences. For data mining techniques that work only on numbers, categorical data must somehow be represented numerically. Some data mining techniques cannot handle missing values, so missing values must somehow be dealt with; the same goes for outliers. When some outcomes are rare, using stratified sampling to balance the data may be necessary. When variables are measured on different scales, standardizing them may also be necessary.

Data preparation may involve creating new variables by combining existing variables in creative ways. It may also involve reducing the number of variables using principal components and other techniques.

Exploratory Data Analysis

Exploratory data analysis is not a major focus of this book, but that is not because we think it is unimportant. In fact, one of the authors (Gordon) has written a book that is largely devoted to this data mining task: *Data Analysis Using SQL and Excel*. The product of exploratory data analysis may be a report or a collection of graphs that describe something of interest. Exploratory data analysis can also be used for adding new measures and variables in the data.

Profiling is a familiar approach to many problems, and it need not involve any sophisticated data mining algorithms at all. Profiles are often based on demographic variables, such as geographic location, sex, and age. Because advertising is sold according to these same variables, demographic profiles can turn directly into media strategies. Simple profiles are used to set insurance premiums. A 17-year-old male pays more for car insurance than a 60-year-old female. Similarly, the application form for a simple term life insurance policy asks about age, sex, and smoking — and not much more.

Powerful though it is, profiling has serious limitations. One is the inability to distinguish cause and effect. As long as the profiling is based on familiar demographic variables, this is not noticeable. If men buy more beer than women, we do not have to wonder whether beer drinking might be the cause of maleness. We can safely assume that the link is from men to beer and not vice versa.

With behavioral data, the direction of causality is not always so clear. Consider a couple of examples from real data mining projects:

- People who have purchased certificates of deposit (CDs) have little or no money in their savings accounts.
- Customers who use voice mail make a lot of short calls to their own number.

Not keeping money in a savings account is a common behavior of CD holders, just as being male is a common feature of beer drinkers. Beer companies seek out males to market their product, so should banks seek out people with no money in savings in order to sell them certificates of deposit? Probably not! Presumably, the CD holders have no money in their savings accounts because they used that money to buy CDs. A more common reason for not having money in a savings account is not having any money, and people with no money are not good prospects for investment accounts. Similarly, the voice mail users call their own number so much because in this particular system that is one way to check voice mail. The pattern is useless for finding prospective users.

Binary Response Modeling (Binary Classification)

Many business goals boil down to separating two categories from each other — the good from the bad, the sheep from the goats, or (at the risk of being sexist and ageist) the men from the boys. In a direct marketing campaign, the good respond and the bad do not. When credit is extended, the good pay what is owed and the bad default. When claims are submitted, the good are valid and the bad are fraudulent. There are techniques, such as logistic regression, that are specialized for these sorts of yes or no models.

Depending on the application, a response model score can be the class label itself or an estimate of the probability of being in the class of interest. A credit card company wanting to sell advertising space in its billing envelopes to a ski boot manufacturer might build a classification model that put all of its cardholders into one of two classes, skier or non-skier. More typically, it would assign each cardholder a propensity-to-ski score. Anyone with a score greater than or equal to some threshold is classified as a skier, and anyone with a lower score is considered not to be a skier.

The estimation approach has the great advantage that the individual records can be rank ordered according to the estimate. To see the importance of this, imagine that the ski boot company has budgeted for a mailing of 500,000 pieces. If the classification approach is used and 1.5 million skiers are identified, then it might simply place the ad in the bills of 500,000 people selected at random from that pool. If, on the other hand, each cardholder has a propensity-to-ski score, it can contact the 500,000 most likely candidates.

Classification

Classification, one of the most common data mining tasks, seems to be a human imperative. To understand and communicate about the world, we are constantly classifying, categorizing, and grading. We divide living things into phyla, species, and genera; matter into elements; dogs into breeds; people into races; steaks and maple syrup into USDA grades.

Classification consists of assigning a newly presented object to one of a set of predefined classes. The classification task is characterized by a well-defined definition of the classes, and a model set consisting of preclassified examples. The task is to build a model of some kind that can be applied to unclassified data in order to classify it.

Examples of classification tasks that have been addressed using the techniques described in this book include:

- Classifying credit applicants as low, medium, or high risk
- Choosing content to be displayed on a Web page
- Determining which phone numbers correspond to fax machines, which to voice lines, and which are shared
- Spotting fraudulent insurance claims
- Assigning industry codes and job designations on the basis of free-text job descriptions

In all of these examples, there are a limited number of classes, and the task is to assign any record into one or another of them.

Estimation

Classification deals with discrete outcomes: yes or no; measles, rubella, or chicken pox. Estimation deals with continuously valued outcomes. Given some input data, estimation comes up with a value for some unknown continuous variable such as income, order size, or credit card balance.

Examples of estimation tasks include:

- Estimating a family's total household income
- Estimating the lifetime value of a customer
- Estimating the value at risk if a customer defaults
- Estimating the probability that someone will respond to a balance transfer solicitation
- Estimating the size of the balance to be transferred

The product of the estimates created in the last two bullet points is the expected value of the balance transfer offer. If the expected value is less than the cost of making the offer, the solicitation should not be made.

Finding Clusters, Associations, and Affinity Groups

Determining what things go together in a shopping cart at the supermarket, and finding groups of shoppers with similar buying habits are both examples of undirected data mining. Products that tend to sell together are called

affinity groups and customers with similar behaviors comprise *market segments*. Retailers can use affinity grouping to plan the arrangement of items on store shelves or in a catalog so that items often purchased together will be seen at the same time. Marketing people can design products and services to appeal to particular segments.

Affinity grouping is one simple approach to generating rules from data. If two items, say cat food and kitty litter, occur together frequently enough, you can think of how to use this information in marketing campaigns. It also brings up another issue: What are customers not buying that they should? A customer who buys lots of kitty litter should also be buying cat food — where are they getting it?

Clustering is the task of segmenting a heterogeneous population into a number of more homogeneous subgroups or clusters. What distinguishes clustering from classification is that clustering does not rely on predefined classes. In classification, each record is assigned a predefined class on the basis of a model developed through training on preclassified examples.

In clustering, there are no predefined classes and no examples. The records are grouped together on the basis of self-similarity. It is up to the user to determine what meaning, if any, to attach to the resulting clusters. Clusters of symptoms might indicate different diseases. Clusters of customer attributes might indicate different market segments.

Clustering is often a prelude to some other form of data mining or modeling. For example, clustering might be the first step in a market segmentation effort: Instead of trying to come up with a one-size-fits-all rule for "what kind of promotion do customers respond to best," first divide the customer base into clusters or people with similar buying habits, and then ask what kind of promotion works best for each cluster. Chapters 13 and 14 cover techniques for cluster detection in detail.

Applying a Model to New Data

Many of the tasks listed earlier usually involve applying a model to new data. This is not true of exploratory data analysis, and it may or may not be true of clustering, but for binary response modeling, classification, and estimation, the data used to create the model contains known values of the target variable. One reason for applying a model to data where the target value is already known is to evaluate the model. After the model has been deployed, its purpose is to score new data where the probability of response, class, or value to be estimated is unknown.

Applying a model to new data is called *scoring*. The data to be scored must contain all the input variables required by the model along with a unique identifier for each row. The result of scoring is a new table with at least two columns — the identifier and the score.

Data Mining Techniques

The title of this book starts with "Data Mining Techniques," and most of the chapters describe individual techniques.

In many cases, data mining is accomplished by building models. In one sense of the word, a model is an explanation or description of how something works that reflects reality well enough that it can be used to make inferences about the real world. Without realizing it, human beings use models all the time. When you see two restaurants and decide that the one with white tablecloths and real flowers on each table is more expensive than the one with Formica tables and plastic flowers, you are making an inference based on a model you carry in your head based on your past experience. When you set out to walk to one of the restaurants, you again consult a mental model of the town.

In a more technical sense of the word, a model is something that uses data to classify things, make predictions, estimate values, or to produce some other useful result. As shown in Figure 3-4, pretty much anything that can be applied to data to produce a score of some kind fits the definition of a model.

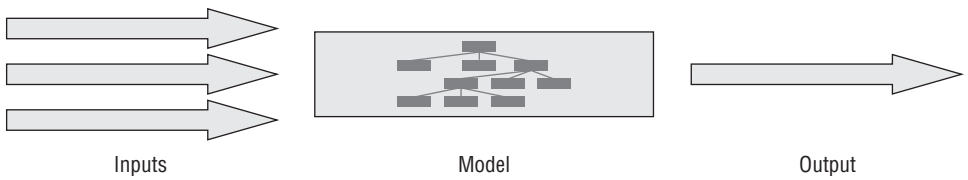


Figure 3-4: Models take an input and produce an output.

A data mining model serves two purposes. The first purpose is to produce scores that you can use to guide decisions. The second is to provide insight into the relationship between the explanatory variables used to build the model and the target. Depending on the application, one or the other of these purposes may be more important than the other.

Data mining techniques fall into two categories: They can be either directed or undirected, which means respectively whether techniques themselves require or do not require target variables. Directed and undirected techniques should not be confused with directed and undirected data mining, because both types of techniques can be used for both types of data mining.

Formulating Data Mining Problems: From Goals to Tasks to Techniques

Business goals, data mining tasks, and data mining techniques form a kind of staircase from the general to specific and from non-technical to technical. Formulating a data mining problem involves descending this staircase one step

at a time; going first from business goals to data mining tasks and then from data mining tasks to data mining techniques. Typically, each step requires the involvement of different staff with different skill sets. Setting and prioritizing goals is the responsibility of upper management. Translating these goals into data mining tasks and using data mining techniques to accomplish them is the role of data miners. Gathering the requisite data and transforming it into a suitable form for mining often requires cooperation with database administrators and other members of the information technology group.

Choosing the Best Places to Advertise

A company is trying to reach new profitable customers. Where should it look? Google AdWords? A reality TV show about cooking? A magazine? If so, which one? *Architectural Digest*? *People en Español*? *Rolling Stone*?

Many factors affect the decision, including overall cost, cost per impression, and cost per conversion. Data mining can provide input to the decision by matching the demographics of the advertising vehicle to the demographics of the best customers. Behavioral data for the profitable customers does not help, because advertising is based only on demographic data.

One possible approach is:

1. Profile existing profitable customers using demographic and geographic characteristics such as age, sex, occupation, marital status, and neighborhood characteristics. Use this profile to define the prototypical profitable customer.
2. Define the audience of each potential advertising vehicle using the same variables used to profile profitable customers.
3. Estimate the distance from each advertising channel to the prototypical profitable customer. This distance is the advertising channel's similarity score; as in golf, smaller is better.
4. Advertise in the venues with the lowest scores.

This is an example of a similarity model, which is covered in Chapter 6.

Determining the Best Product to Offer a Customer

What is the best next offer to make to a customer? This question is an example of cross-selling that occurs in many industries.

There are several possible approaches to this problem, depending, among other things, on the number of products to choose from. If the number of products is manageably small, a good approach is to build a separate model for each product so every customer can be given as many scores as there are products, as shown in Figure 3-5. A customer's best offer is the product for which he or she has the highest score (possibly excluding products the customer already has).

1. For each product, build a binary response model to estimate the propensity of customers for the product.
2. Set the propensity score to 0 for customers who already have a product.
3. Using the propensity scores, design a decision procedure that assigns the best product to each customer, based on something like the highest propensity or the highest expected profit.

Natural choices for Step 1 include decision trees, neural networks, and logistic regression.

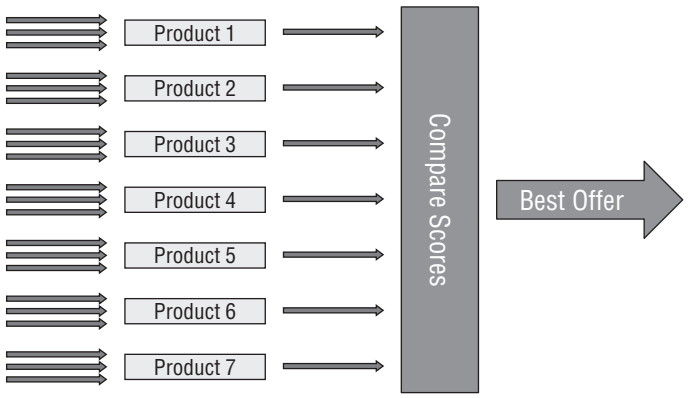


Figure 3-5: Individual propensity scores for each product are compared to determine the best offer.

A binary response model is not the only method for developing propensity scores. Another method is to cluster the data using input variables and see which products predominate in each cluster. The proportion of the cluster with a given product can be assigned as the propensity score. This method would use k-means clustering or another clustering technique.

Finding the Best Locations for Branches or Stores

What are the best locations for new stores? In this scenario, performance data for existing stores is available along with data about the catchment area — the natural market area from which each store draws its customers. The idea is to find the combination of explanatory variables that predicts good performance for a store.

The following modeling tasks are one approach to addressing this problem:

1. Build a model to estimate some store performance metric based on the available explanatory variables for the catchment area.
2. Apply the model to candidate locations so the highest scoring locations can be selected.

This is basically an estimation model, which can use a variety of techniques, such as neural networks, regression, or MBR.

An alternative approach is to classify the stores as good or bad, and then build a model that predicts these groups. Often, a good way to approach this is using the *excluded middle* approach: The profitability of each store is divided into thirds — high, medium, and low. Remove the "medium" stores and build a model to separate the high from the low (a case study in Chapter 15 takes this approach for finding the factors that distinguish stores in Hispanic areas from those in non-Hispanic areas):

1. Classify existing stores as good or bad and build a model that can distinguish between the two classes.
2. Apply the model to candidate locations so the good one can be selected.

Likely explanatory variables include the population within driving distance, the number of competitors within driving distance, and demographic factors. This is a profiling model because the goal is to link current performance with current conditions. The modeling techniques are those used for classification, such as logistic regression, decision trees, and MBR.

Segmenting Customers on Future Profitability

A method for defining profitability has been established, such as the total revenue or net revenue generated by customers over the course of one year. The goal is to segment customers today based on their anticipated profitability over the next year.

There are many ways to approach profitability calculations. This approach removes some of the more difficult areas, such as predicting how long a customer will remain a customer (and hence deciding on future discount rates), and how to attribute network effects to the customers.

For this approach, turn the clock back one year and take a snapshot of each customer who was active on that date. Then, measure the total revenue during the following year. This is the model:

1. Prepare the data for modeling by turning the clock back one year and taking a snapshot of each customer who was active on that date. Then, measure the total revenue during the following year. This creates a prediction model set.
2. Use this model set to estimate how much someone will be worth in the next year.
3. Segment the anticipated revenue into thirds, to get high, medium, and low anticipated revenue.

Step 2 requires building an estimation model, using a technique such as neural networks, MBR, or regression.

A slight variation on this approach would be to classify the customers in the model set as high, medium, or low generators of revenue in the upcoming year. This would use a classification model, which might use decision trees (with a three-way target) or three logistic regression models (one for each of the three groups).

Decreasing Exposure to Risk of Default

The goal of this business problem is to detect warning signs for default while there is still time to take steps to decrease exposure. One detection method uses a binary response model, with a target of "default." The model set is a snapshot of all customers at a given point in time (for example, the first of the year) and a flag that indicates whether or not they default in the three months after the snapshot date. New customers can then be scored with the binary response model to predict their probability of default. Perhaps customers with high levels of default should have their credit lines lowered.

Such a binary response model could be built using a variety of techniques, such as logistic regression, decision trees, or neural networks. Undirected techniques, such as clustering, could even be used. Build clusters on the input variables, and then measure the ability of the clusters to separate the target values. This is an example of using an undirected technique for a directed model.

Another approach combines the probability of default with the amount of default. This two-stage model estimates how much a customer would owe after defaulting. The model set for this consists only of customers who have defaulted, with the target being the amount owed. This model would be used to calculate the expected value of the loss, which is the probability of default multiplied by the estimated amount owed. The estimate of the amount owed could be built using MBR, neural networks, regression, or possibly decision trees.

Yet another approach would be to treat this as a time-to-event problem, estimating when a customer is likely to default. In this case, the model set consists of all customers, with their start date, end date, and whether or not the customer defaulted. The model would estimate the amount of time until a customer defaults. When scoring new customers, if the estimated time to default is in the near future, then actions would be taken to mitigate the default. This type of model would typically be built using survival analysis.

Improving Customer Retention

There are many different ways to improve customer retention:

- Find customers most at risk of leaving and encourage them to stay.
- Quantify the value of improving operations, so customers will stick around.
- Determine which methods of acquiring customers bring in better customers.
- Determine which customers are unprofitable, and let them leave.

This section only discusses the first of these.

The task list for determining who will stay is similar to the task list for any binary response model. Build a model set that consists of customers who stay and go, and let the model find the patterns that distinguish between them. This provides a model score that you can then use for a retention effort.

This type of binary response model can be built using many techniques, such as decision trees, neural networks, logistic regression, and MBR. An alternative approach would be to estimate the remaining customer tenure using survival analysis, and apply the retention message to those customers most likely to leave in the near future.

Sometimes the most important output from a model is not the scores it produces, but the understanding that comes from examining the model itself. The model may be able to explain whether customers are primarily leaving due to service disruptions, price sensitivity, or other causes. However, this requires using a technique that can explain its results. Decision trees and logistic regression are the best of the bunch for explicability.

Detecting Fraudulent Claims

The translation of this goal into modeling tasks depends on whether examples of known fraud are available. If so, this is a directed data mining task:

1. Build a profiling model that is capable of distinguishing fraudulent claims from legitimate ones.
2. Use the model to score all claims that come in. Mark claims that score higher than some threshold for additional scrutiny before approval.

Decision trees and logistic regression are likely techniques for building the profiling model in Step 1.

Sometimes, fraud is suspected, but it is not clear which transactions are fraudulent. This situation calls for undirected data mining:

1. Form clusters of similar claims. Most claims will probably fall into a few large clusters representing different types of legitimate claim.
2. Examine the smaller clusters to see what makes them special.

The claims in the smaller clusters may also be perfectly legitimate. All that the clustering exercise shows is that they are unusual. Some unusual claims turn out to be fraudulent, so all are worth further scrutiny.

ONE GOAL, TWO TASKS: WINNING A DATA MINING CONTEST

Every year, contestants from academia and industry test their data mining skills in a contest held in conjunction with the annual KDD (Knowledge Discovery and Data Mining) conference. One year, it was clear that what separated winners from losers was not the algorithms they used or the software they employed, but how they translated the business problem into data mining tasks.

The business problem was to maximize donations to a non-profit charity. The data was a historical database of contributions.

Exploring the data revealed the first insight: the more often someone contributed, the less money they contributed each time. Expecting the best donors to be those who respond most frequently is quite reasonable.

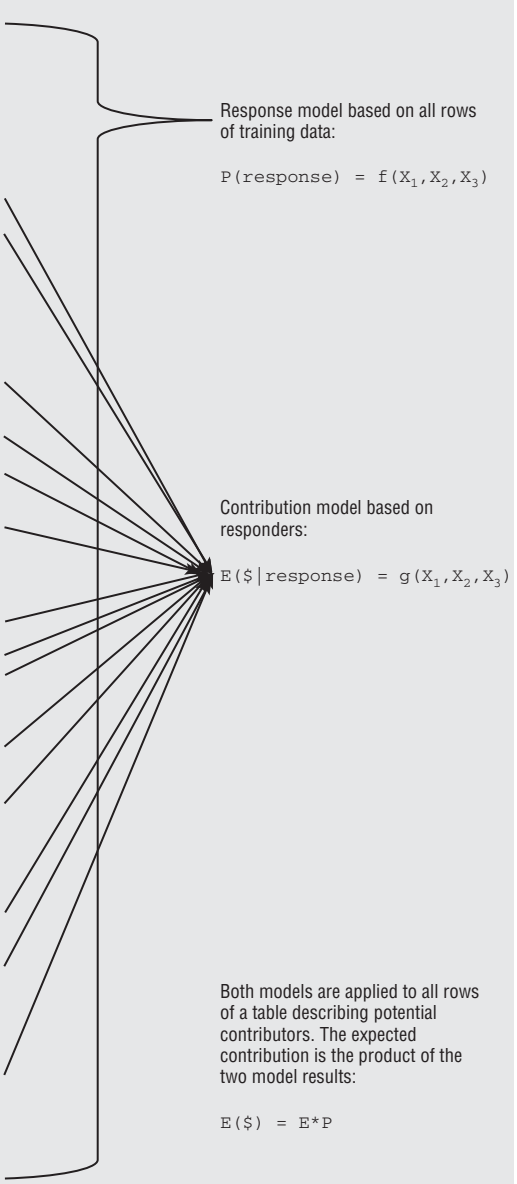
In this case, though, people seem to plan their charitable giving on a yearly basis. They might donate a lump sum all at once, or space their contributions over time. More checks does not always mean more money. This suggests that the decision to make a donation is separate from the decision of how large a donation to make. The two decisions are quite likely influenced by different factors. Perhaps people of all income levels are more likely to donate to a veterans' organization if they themselves have served in the military. After they have decided to contribute, income level may have an influence on the sizes of the donations.

These insights led to the winning approach, which was to model response and contribution size separately. The response model is built on a training set that contains both contributors and non-contributors. This is a binary outcome classification task.

The contribution size model is built on a training set consisting only of contributors. This is an estimation task. The following figure shows the two models and how their results are combined to produce an expected response value for each prospect.

The three winning entries all took this approach of combining models. The majority of contestants, on the other hand, built a single model with *amount contributed* as the target. These models treated the entire problem as an estimation task with a lack of response represented as a contribution of zero dollars.

CustomerID	Response	Contribution	X1	X2	X3
292129	0		A	39,220	1
292130	0		A	39,749	1
292134	0		C	40,052	1
197549	0		A	39,485	1
292137	0		A	39,749	1
291800	0		A	39,610	1
292138	0		A	39,749	0
332806	0		A	39,860	0
292140	0		A	39,686	1
347807	1	$40	C	40,139	0
292141	0		A	39,749	1
292143	1	$30	C	40,027	0
409542	0		A	40,050	0
292848	0		C	40,012	1
292850	0		C	40,151	1
292851	0		A	39,750	0
292852	0		C	39,997	1
292853	0		A	39,750	1
292857	0		A	39,750	1
292859	1	$30	A	39,994	1
292860	0		A	39,750	0
292861	0		A	39,750	0
292862	1	$30	C	39,859	0
292863	0		C	39,877	1
292864	1	$40	C	40,071	1
292868	0		A	39,750	0
403246	0		A	40,035	0
292869	1	$30	D	40,132	0
292870	0		C	39,788	0
292871	0		A	39,750	1
292872	0		A	39,750	1
292873	0		C	39,997	1
292874	1	$40	C	40,150	1
292878	0		A	39,750	1
292879	1	$40	C	40,132	0
292880	1	$30	C	39,859	1
292881	0		C	39,879	0
24583	0		A	38,966	0
292884	0		A	39,750	1
126612	1	$40	A	40,016	0
292886	0		A	39,288	1
292887	0		A	39,750	1
292888	1	$40	A	40,113	0
292889	0		C	39,795	0
390095	0		A	40,000	1
292893	0		A	39,462	1
292894	0		A	40,118	1
292964	0		D	40,138	0
292897	1	$30	C	39,859	1
292900	0		A	39,750	1
292901	0		C	39,808	1
292902	1	$30	C	39,859	0
292905	0		A	39,750	1
292908	0		A	39,750	0
292909	0		A	39,750	1
292911	0		A	39,750	1
292913	0		C	39,798	1
292914	1	$30	D	40,132	0
292915	0		A	39,750	0
292916	0		C	39,812	0
292917	0		A	39,750	0
292919	0		A	39,750	1
292920	0		D	40,114	0



A two-stage model for the expected value of a contribution

What Techniques for Which Tasks?

You can use all the data mining techniques described in this book in creative ways for applications outside the ones with which they are most often associated. Each major family of techniques has a chapter (or even more than one

chapter). The individual technique chapters include examples of how to apply the techniques for various purposes. Still, some techniques are better suited to some tasks. When choosing a technique, ask yourself these questions:

- Is there a target or targets?
- What is the target data like?
- What is the input data like?
- How important is ease of use?
- How important is explicability?

The answers to these questions narrow the choice of techniques.

Is There a Target or Targets?

All directed data mining techniques, including regression, decision trees, and neural networks, require training with known values for the target variables. When the data does not contain such a target, one of the undirected techniques such as clustering or exploratory data analysis is needed.

What Is the Target Data Like?

When the target is numeric and can take on a wide range of values, a technique that produces continuous values is appropriate. Linear regression models can produce any value from negative infinity to infinity, as can neural networks. When the task is to estimate the value of a continuous target, these are natural choices. Regression trees and table lookup models can all be used to estimate numeric values also, but they produce a relatively small number of discrete values. Memory-based reasoning is another choice for numeric targets that can produce a wide range of values, but never outside the range of the original data.

When the target is a binary response or categorical variable, techniques that produce a probability of being in each class are called for. Decision trees are a very natural fit for these kinds of problems, as are logistic regression and neural networks. Depending on other aspects of the problem, and on the nature of the inputs, other techniques such as similarity models, memory-based reasoning, and naïve Bayesian models may be good choices.

What Is the Input Data Like?

Regression models, neural networks, and many other techniques perform mathematical operations on the input values and so cannot process categorical data or missing values. It is of course possible to recode categorical data or replace categorical fields with numeric fields that capture important features of the categories. It is also possible to input missing values. These operations can be

time-consuming and inaccurate, however. As the number of categorical fields and fields with missing values goes up, so does the appeal of decision trees, table lookup models, and naïve Bayesian models, all of which can easily handle categorical fields and missing values. When the inputs are numeric and do not contain missing values, regression models and neural networks may be able to make use of more of the information in the data.

How Important Is Ease of Use?

Some techniques require much more data preparation than others. For example, neural networks require all inputs to be numeric and within a small range of values. They are also sensitive to outliers and unable to process missing values. Others, such as decision trees, are much more forgiving and require less data preparation, but may not do as good a job. There is often a trade-off between power, accuracy, and ease of use. As an extreme example, genetic algorithms require so much work on the part of the miner that they are rarely used if an alternative approach is available.

Since the first edition of this book appeared back in the 1990s, data mining software tools have made great strides in the ease-of-use area. The best ones provide user interfaces that support best practices and make even complex techniques such as neural networks relatively user-friendly.

How Important Is Model Explicability?

For some problems, getting the right answer fast is paramount. A modern, no-envelope-required automatic teller machine must be able to recognize handwritten amounts accurately in order to accept checks for deposit. Although it would certainly be fascinating to learn how the algorithm differentiates American "7s" from European "1s", there is no urgent need to do so. In the brief interval between when a credit card is swiped and the approval code is transmitted, the transaction is scored for likelihood to be fraudulent. Getting this decision right is important. Approving a fraudulent transaction has an immediate and obvious cost; rejecting a legitimate transaction annoys a valuable customer. In both these examples, getting the right answer is clearly more important than having a clear explanation of how the decision was made.

At the other extreme, some decisions — whether to grant or deny credit, for example — may be subject to regulatory review. Explaining that credit was denied because the applicant had too many open lines and too great a ratio of debt to income is fine. Saying, "The model identified the applicant as high risk, but we have no idea why," is unacceptable.

Different techniques offer different trade-offs between accuracy and explicability. Decision trees arguably offer the best explanations because each leaf has a precise description in the form of a rule. Although this means that the score for

any given record can be explained, it does not mean that a large, complex tree is easy to understand as a whole. The trade-off is that decision trees may not make use of as much of a variable's inherent information as other techniques that make use of the value directly instead of simply comparing it to a splitting value.

With a bit of attention to data preparation, regression models also shed a lot of light on what contributes to a score. When explanatory variables have been standardized, the relative magnitude of the model coefficients show how much each one contributes to the score. In a regression, every small change in the value of an explanatory variable has an effect on the score. In that sense, the regression model makes more use of the information provided by the explanatory variables than do decision trees.

Neural networks are quite flexible and are capable of modeling quite complex functions very accurately, but are essentially inexplicable. Each of these techniques provides a different trade-off between best scores and best explanations. Knowing the strengths and weaknesses, you must decide on the techniques that are most appropriate for your application.

Table 3-1 shows which techniques are typically used for which tasks. As the table makes clear, pretty much any of the directed techniques can be used for classification, prediction, and estimation problems. The final choice is driven by the extent to which the model should be able to tell a story in addition to producing scores, and by characteristics of the data to be mined.

Table 3-1: What Techniques for Which Tasks?

TASK	BEST FIT	ALSO CONSIDER
Classification and prediction	Decision trees, logistic regression, neural networks	Similarity models, table look-up models, nearest neighbor models, naïve Bayesian models
Estimation	Linear regression, neural networks	Regression trees, nearest neighbor models
Binary response	Logistic regression, decision trees	Similarity models, table look-up models, nearest neighbor models, naïve Bayesian models
Finding clusters and patterns	Any of the clustering algorithms	Association rules

Lessons Learned

The data mining process can fail in many ways. Failure can take several forms, including simply failing to answer the questions you set out to answer, as well as "discovering" things you already know. An especially pernicious form of

failure is learning things that aren't true. This can happen in many ways: when the data used for mining is not representative; or when it contains accidental patterns that fail to generalize; or when it has been summarized in a way that destroys information; or when it mixes information from time periods that should be kept separate.

There are three styles of data mining. Exploratory data mining produces insights or answers questions rather than producing models used for scoring. Exploratory data mining often involves coming up with hypotheses that can be proven or disproven using data. Exploratory data mining is very important; however, it is not the subject of the advanced techniques in this book.

Directed data mining is used when the historical data contains examples of what is being looked for. For an attrition model, this assumes that the historical data contains examples of customers who have and have not stopped. For a customer value model, this assumes that it is possible to estimate customer value using the historical data. The target (or targets) of the model are these variables. The "explanatory" variables in the model are the inputs.

Undirected data mining does not use a target variable. It is like throwing the data at the computer and seeing where it lands. Making sense of undirected data mining requires understanding and interpreting the results. Without a target, there is no way for the computer to judge whether or not the results are good.

You can use all three data mining styles separately or in combination to accomplish a wide range of business goals. The data mining process starts with a business goal. The data mining process involves translating the business goal into one or more data mining tasks. After the tasks have been defined, the nature of the task, the type of data available, the way that results will be delivered, and the trade-off between model accuracy and model explicability all influence the choice of data mining technique.

Whichever technique you choose, and regardless of the data mining style, using data mining effectively requires some knowledge of statistics, the subject of the next chapter.