CHAPTER 3

# Data Mining Methodology: The Virtuous Cycle Revisited

The great American photographer Ansel Adams did not just step outside and "snap" pictures of sunsets in the West. Even though the images were captured by just pressing a button, he had to plan the photos and wait until just the right moment. One great irony of photography is that the most natural and unposed shots are often the most work, requiring a great deal of planning and preparation. The same is true in data mining. Being successful in data mining requires planning and understanding the business problem.

To continue the analogy, photography has a whole range of technical options for developing prints: one-hour photo labs, amateur darkrooms, professional darkrooms, and digital photography. Mastering photography requires understanding the development process as well as composition. The ultimate result is a combination of the aesthetic and the technical.

Similarly, mastering data mining requires combining the business and the technical. Data mining connects business needs to data; it is about understanding customers and prospects, understanding products and markets, understanding suppliers and partners, understanding processes—all by leveraging the data collected about them. A basic understanding of the technical side of data mining is critical for success, especially the process of transforming data into information.

Failure to follow the rules of photography can lead to fuzzy, poorly exposed pictures—even when using the best equipment. The same is true of data

mining—failure to follow the rules of good model building will lead to inaccurate information and poor decisions.

The data mining methodology presented in this chapter is designed for those companies that want to build a core competency in data mining, although anyone who relies on data analysis can benefit from it. We originally introduced this methodology in *Data Mining Techniques* (John Wiley & Sons, 1997); here we have elaborated and condensed it to focus on its most important parts. Following the methodology leads to better models that support more informed decisions. This methodology is built around the virtuous cycle of data mining, which highlights the business aspects of data mining while recognizing the interplay between the business and technical. The discussion on the methodology begins, appropriately, with the two different styles of data mining.

### PEOPLE ARE NECESSARY!

**There is a fear that as computers become more and more powerful, they will eventually replace people in many different fields. With respect to data mining, the day is very far off! At the technical level, data mining is a set of tools and techniques that make people more productive. Automated algorithms can spot patterns. People will always be needed to know when the patterns are relevant, what problems need to be addressed, when the results are meaningful, and so on.**

## Two Styles of Data Mining

There are two styles of data mining. *Directed data mining* is a top-down approach, used when we know what we are looking for. This often takes the form of predictive modeling, where we know exactly what we want to predict. *Undirected data mining* is a bottom-up approach that lets the data speak for itself. Undirected data mining finds patterns in the data and leaves it up to the user to determine whether or not these patterns are important.

These two approaches are not mutually exclusive. Data mining efforts often include a combination of both. Even when building a predictive model, it is often useful to search for patterns in the data using undirected techniques. These can suggest new customer segments and new insights that can improve the directed modeling results.

### Directed Data Mining

The top part of Figure 3.1 shows a model as a black box. What this means is that we do not care what the model is doing; we just want the most accurate result possible. This is the approach used when we know what we are looking for, when we can direct the data mining effort toward a particular goal.

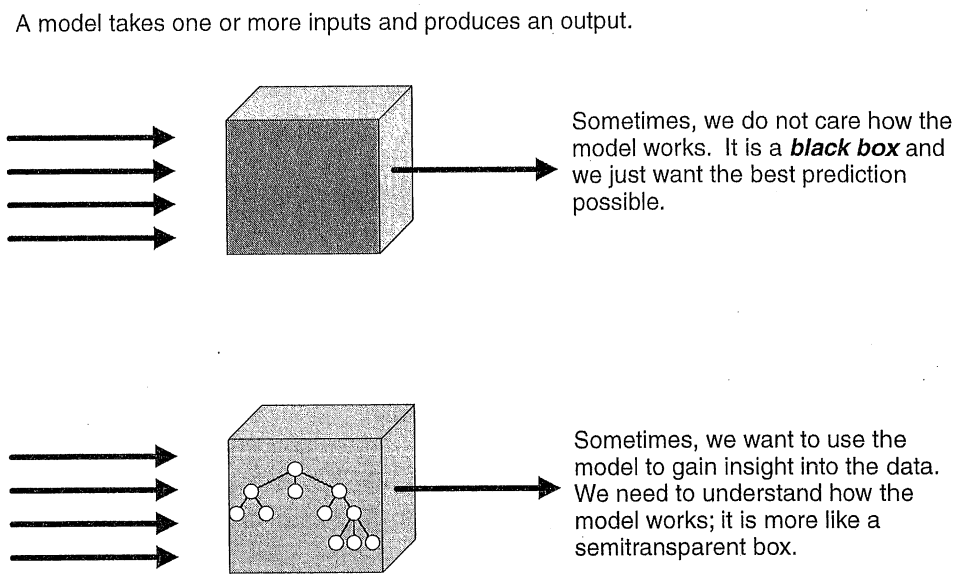



**Figure 3.1** Data mining uses both black-box models and semitransparent models.

Typically, we are using already known examples, such as prospects who already received an offer (and either did or did not respond), and we are applying information gleaned from them to unknown examples, such as prospects who have not yet been contacted. Such a model is called a *predictive model*, because it is making predictions about unknown examples. Predictive models answer questions such as:

- Who is likely to respond to our next offer, based on the history of previous marketing campaigns?
- What is the right medical treatment, based on past experience?
- Which machine is most likely to be the next one to fail?
- Which customers are likely to leave in the next six months?
- What transactions are likely to be fraudulent, based on known examples of fraud?

The predictive models use experience to assign scores (and confidence levels) to some relevant outcome in the future. One of the keys to success is having enough data *with the outcome already known* to train the model. Predictive models are never 100 percent accurate. They are helpful because making informed decisions in business should lead to better results.

In the past, predictive modeling has often been a very specialized function—the work of actuaries, corporate forecasters, and statisticians. Although their

work is quite important, they have been relatively far removed from daily business concerns; they provide little assistance on the front line, for the typical brand manager, account manager, or product manager. These specialists speak their own language. Consider an example from the risk management group in one credit card company. They have a label for customers who earn high incomes, are unlikely to go bankrupt, and use the card infrequently. For most of the company, this segment represents the highest potential value. For the risk management group, they are simply "possible future write-off." Quite a difference in perception!

Nowadays, even amateurs can attempt to build predictive models on their desktops. Unfortunately, it is easier to build misleading models than predictive models. The goal in making predictions is to learn from the past, and to learn in such a way that the knowledge can be applied to the future. Perhaps the most important insight in this chapter is this: the best model is not the one with the highest lift when it is being built. It is the model that performs the best on unseen, future data. Although there is no 1-2-3 recipe for building perfect models, this chapter covers the fundamentals needed to build effective models, and we will see these lessons applied in the case studies.

## Undirected Data Mining

Sometimes, though, predictive accuracy is not the only or even the primary goal. Undirected data mining is about discovering new patterns inside the data. These patterns provide insight, and this insight might even prove very informative.

We represent this form of data mining with *semitransparent* boxes. Unlike directed data mining, we want to know what is going on, we want to know how the model is coming up with an answer. In the case study on cellular churn, we see how to use a decision tree to discover an important segment of customers.

The case study is a good example, showing that undirected data mining and directed data mining are both valuable in many data mining efforts. Undirected data mining is often used during the data exploration steps. What is in the data? What does it look like? Are there any unusual patterns? What does the data suggest for customer segmentation? These types of questions are answered using tools that support clustering, visualization, and market basket analysis.

At the same time, some predictive modeling techniques, notably decision trees, explain the models they produce. These techniques sometimes provide important insights in addition to the predictions they make. Two things are happening. An example of directed data mining is that a decision tree makes predictions. An example of undirected data mining is that a person looks at a decision tree and possibly notices an interesting pattern.

Undirected data mining is necessarily interactive. Advanced algorithms can find patterns in the data, but only people can determine whether the patterns have any significance and what the patterns might mean.

## The Virtuous Cycle of Data Mining

The virtuous cycle of data mining, depicted in Figure 3.2, highlights the fact that data mining does not exist in a vacuum. Data mining has a purpose. In the case studies, we will see data mining applied to many problems in many industries. Its most common applications are in marketing, specifically for customer relationship management; we see it applied to prospecting for new customers, retaining existing ones, and increasing customer value. We also see it applied to understanding customer behavior and optimizing manufacturing processes.

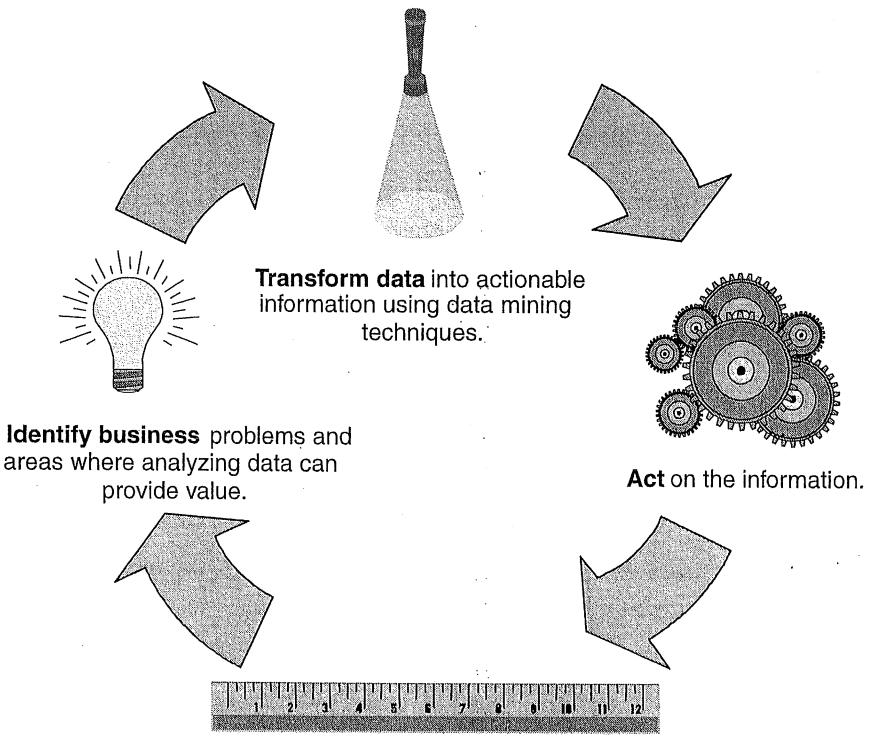



**Figure 3.2** The virtuous cycle of data mining leads to a learning organization.

Clearly, data mining has many applications. And, although they may have much in common, every application has its own unique characteristics. Industries differ from each other; within a single industry, different companies have different strategic plans and different approaches. All of this affects the approach to data mining.

The previous chapter outlined three ways that companies can incorporate data mining solutions into their business. Here, we focus primarily on the companies that want to build core competencies in data mining, because data analysis supports critical business processes.

The virtuous cycle is a high-level process, consisting of four major business processes:

- Identifying the business problem
- Transforming data into actionable results
- Acting on the results
- Measuring the results

There are no shortcuts—success in data mining requires all four processes. Results have to be communicated and, over time, we hope that expertise in data mining will grow. Expertise grows as organizations focus on the right business problems, learn about data and modeling techniques, and improve data mining processes based on the results of previous efforts. In short, successful data mining is an example of *organizational learning*.

## Identifying the Right Business Problem

Defining the business problem is the trickiest part of successful data mining because it is exclusively a communication problem. The technical people analyzing data need to understand what the business really needs. Even the most advanced algorithms cannot figure out what is most important.

*A necessary part of every data mining project is talking to the people who understand the business.* These people are often referred to as *domain experts*. Sometimes there is a tendency to want to treat data analysis as a strictly technical exercise. Resist this tendency! Only domain experts fully understand what really needs to be done; and ultimately, they are likely to receive the credit or blame for bottom-line results.

While taking into account what the domain experts have to say, it is also important not to be constrained by their expertise. Important results often

come from "thinking outside the box"—ignoring supposed wisdom by understanding what is really happening.

Working with the business people allows you to answer questions such as the following:

- Is the data mining effort really necessary?
- Is there a particular segment or subgroup that is most interesting?
- What are the relevant business rules?
- What do they know about the data? Are some data sources known to be invalid? Where should certain data come from?
- What do their intuition and experience say is important?

Answering these questions requires a combination of skills that includes being able to work with the domain experts as well as understanding the technology and data. There is no specific right or wrong way, because this is an interactive process between people. The next few sections begin with business scenarios (shown in italics) that should help shed light on the process.

**Domain experts have a very good idea of what is going on. They can focus data mining efforts by answering questions such as:**

**What is really important to the business?**

**What information, that you cannot now get, would you be able to act on immediately?**

**Is data mining really necessary to solve this problem?**

**What important business rules are relevant in this case?**

**What does your experience and intuition add to the equation?**

**It is also important to keep the business people in the loop, so they are aware of new insights gleaned along the way and so the data mining remains focused on areas of value to the business.**

## Is the Data Mining Effort Necessary?

*A Senior Vice President in the credit card group of a large bank has spent tens of thousands of dollars developing a response model. This predictive model is designed to identify the prospects who are most likely to respond to the bank's next offering. The VP is told that by using the model, she can save money: using only 20 percent of the prospect list will yield 70 percent of the responders. However, despite these findings, she replies*

*that she wants every single responder—not just some of them. Getting every responder requires using the entire prospect list, since no model is perfect. In this case, data mining is not necessary.*

*Moral: She could have saved tens of thousands of dollars by not building predictive models in the first place.*

Data mining is not always necessary. As this example shows, not every marketing campaign requires a response model. Sometimes, the purpose of the campaign is to communicate to everyone. Or, the communication might be particularly inexpensive, such as a billing insert or e-mail, so there may be less reason to focus only on responders.

## Is There a Particular Segment or Subgroup That Is Most Interesting?

*On the other side of the globe, another marketing group at a cellular telephone provider has specified that it wants a propensity-to-churn score for all existing customers. A propensity-to-churn score is a number ranging from 0 to 1, where 0 means that the model finds no indications of churn and 1 means that the model has the highest confidence of churn. The marketing group wants to do an intervention (marketing-speak for communicating to the customers in some way) on the 10,000 customers most likely to churn. However, the results are very disappointing. It turns out that they really want to retain their elite customers—and of the 10,000 on their list, fewer than a quarter were elites. The campaign was disappointing.*

*Moral: The campaign would have been more successful by focusing on the right customer segment when building the models.*

Matching expectations is a key to successful data mining efforts. In this case, the need was for a list with only elite members, and the business people did not realize that modeling elites separately would produce better results. Mismatches of this type occur in many ways. For instance, a marketing campaign wants to focus on consumers, but the models return a mix of consumer and commercial customers; a model is built in one geography or on one customer segment and then applied somewhere else. And so on.

## What Are the Relevant Business Rules?

*In the cross-selling example in Chapter 10, we will see that a bank was putting together a marketing campaign to sell brokerage accounts to existing customers. The predictive models did a very good job of determining who would want a brokerage account; one of the key determinants was whether the customer had a private banking account. Alas, bank rules forbid including private bank customers in marketing campaigns.*

*Moral: The modeling effort needed to understand business rules to avoid a particular segment of customers.*

## What about the Data?

*A large, converging telecommunications provider wants to use data mining to figure out who has fax machines. This requires starting with a list of known fax machines and determining who calls them and who does not call them. Where can they obtain this list?*

*A large retailer wants to analyze returns—merchandise returned by customers. Where can it get the data on returns and link them to the original purchaser and market basket? How are returns represented in the data?*

*Moral: Domain experts know where data resides and how it is stored. Using this information can save a lot of time in understanding the data.*

The business people understand the business environment and business processes. Sometimes, they also have a very good understanding of what should be in the data—or what cannot possibly be. In some other cases, they have some important data that never makes it into data warehouses. This often includes information residing on their desktops, which may have important attributes about competitors, suppliers, particular products, and so on.

Of course, the IT group responsible for databases and applications also understands the data and may be the only group that recognizes obscure values in the data.

## Verifying the Opinion of Domain Experts

Domain experts provide experience and intuition, but this can be a double-edged sword. Their intuition can be a source of valuable insight, allowing the data mining effort to focus on particular sources of data or suggesting ways to segment data for building multiple models (such as building a separate churn model for high-value customers and low-value customers).

Their experience can also be a source of gotchas. For instance, after extensive analysis, the data may reveal that call forwarding is always sold with call waiting. The business people already know that, because call forwarding is always bundled with call waiting. They also know what has worked in the past and what has not.

However, as far as possible, their intuition and experience should be verified by looking for patterns that support them. The purpose of data mining is to let the data speak; and data does not lie (although the data may be dirty). If they believe that the best customer is married with an age between 35 and 50, is this true in the data? If they believe that customers who purchased something in the last six months are most likely to make another purchase, is this true? If they believe that home equity customers have highly variable incomes (such as sales people), is this true? And so on.

## Transforming Data into Actionable Results

The heart of data mining is transforming data into actionable results, and there will be much more to say about this topic throughout the book. Here, our goal is to give an outline of the major steps taken to build data mining models and to emphasize that this is an iterative process.

Figure 3.3 illustrates the basic steps taken to transform data into actionable results. Throughout this book, we will be delving into these topics in greater detail.

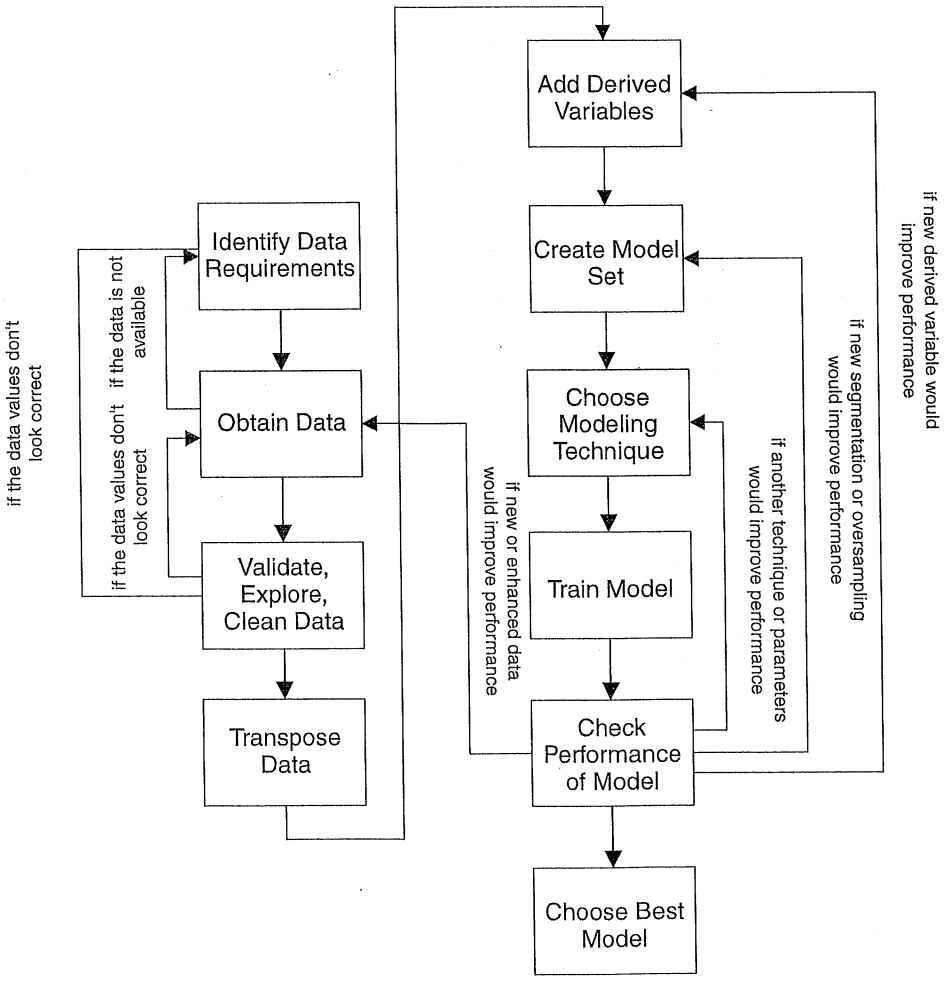



**Figure 3.3** Building data mining models is an iterative process.

## Identify and Obtain Data

The first step in the modeling process is identifying and obtaining the right data. Often, the right data is simply whatever data is available, reasonably clean, and accessible. In general, more data is better.

It is important to verify that the data meets the requirements for solving the business problem. For instance, if the business problem is to identify particular customers, then the data must contain information about each individual customer. There may be additional detail data—such as transaction-level information—but it must also be possible to tie this data back to individual customers.

In addition, we want the data to be as complete as possible when modeling. This can make it impractical to use survey data or other data available only for survey respondents. If the data from the survey proves valuable, then how will nonrespondents be scored? (In other situations, using survey data for data mining may prove very fruitful; just don't expect to apply the resulting model to nonrespondents without some extra work.)

The purpose of the data mining effort may be to identify customer segments, perhaps for the purpose of directing advertising or purchasing lists of prospective customers. In this case, the data needs to contain fields that are appropriate for purchasing advertising space and lists. This often includes fields supplied by outside list providers, location information, demographics, and so on.

When doing predictive modeling the data also needs to contain the desired outcome. One brick-and-mortar retailer was trying to set up a catalog for their identified customers (members). They were building a response model, based on three earlier catalog mailings; this model would determine who was likely to respond to the catalog based on previous responses to test catalogs in the past. They had the following data, as shown in Figure 3.4:

- Marketing data about all members
- Responses data to previous catalogs
- Tons of transaction-level detail about what members had purchased

The one thing they were missing was who had been sent the earlier catalogs! Knowing who has responded without knowing who had been contacted is almost useless. Without this information, the attempts at predicting response were doomed.

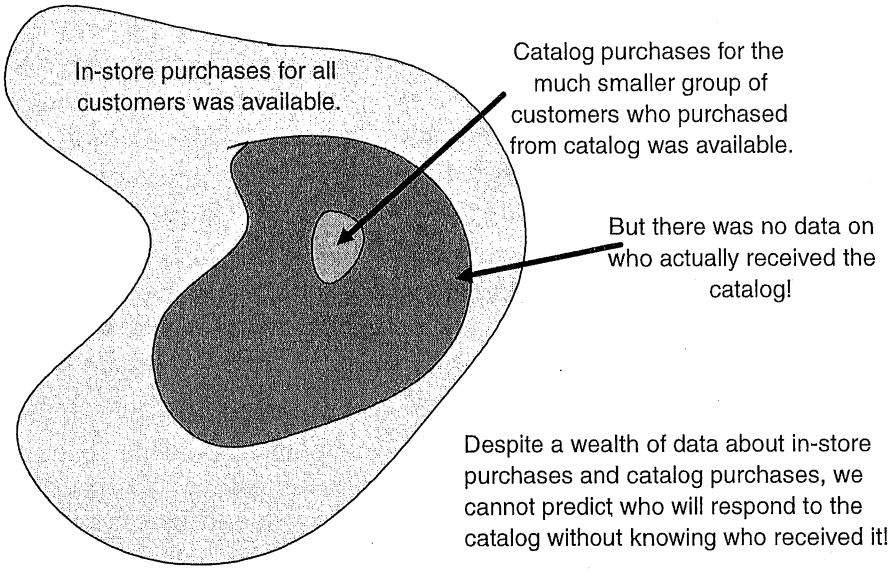



**Figure 3.4** Without the right historical data, data mining will yield disappointing results.

**WARNING**

**Data from OLAP (online analytic processing) systems is not usually sufficient for data mining purposes. This data has already been summarized along important business dimensions, such as time, geography, and product lines. In doing so, the customer dimension is often not one of the underlying dimensions—the number of customers is often so large that a customer dimension would require a much larger amount of storage. So, data mining efforts centered on customers require additional sources of data.**

## Validate, Explore, and Clean the Data

The next step is to validate the data and to cleanse it. The outcome of the data mining effort depends critically on the data.

- Are the fields populated? Will missing data be a big problem?
- Are the field values legal? That is, are numerical fields within proper bounds and are code fields all valid?
- Are the field values reasonable?
- Are the distributions of individual fields explainable?

Data inaccuracies creep in from many different places. Usually, fields that are critical to the business are quite accurate. So, the amount of money billed each month is usually accurate as is the billing address. However, data that is not used (which is the vast majority of the fields) is often inaccurate. The business collects the data but no one ever really looks at it. Data mining, because it is hungry for lots and lots of data, is often the first business process that really uses most of the data fields.

### OUR FAVORITE DATA QUALITY QUOTE

**"The data is clean because it is automatically generated—no human ever touches it." The data in question contained information about file transfers on a distributed computer system. Data packets were often sent between systems. When we looked at the data, though, 20 percent of the transactions described files that arrived before they had even been sent! Evidence of spectacular network performance? Not really. Not only did people never touch the data, but they didn't set the clocks on the computers either.**

## Transpose the Data to the Right Granularity

Granularity refers to the level of the data that is being modeled. Data mining algorithms work on individual rows of data. So, all the data describing a customer (or whatever we are interested in) must be in a single row. The data needs to be summarized to the right level of granularity.

For instance, an automobile insurance company may keep track of every vehicle covered for every year of coverage. For each year, there might be information on the type of vehicle, the total number of claims, the total cost of claims, the estimated value of the car, and so on. However, a marketing application is unlikely to be interested in this data at the vehicle level. People buy insurance policies, and these policies often cover more than one car. So the data has to be transformed from the year-vehicle format to a summarized format by policy, as shown in Figure 3.5.

This is an insurance example, but the problem is omnipresent. Billing data is recorded by month, and all months of data need to be in the same record. Data on different products has to be combined for individual customers. Market basket data needs to be summarized to describe household behavior. Web clicks need to be summarized to describe a single visit. And so on and so on and so on.

This summarization is often a tricky process. Because of the complex data formats, the full power of a programming language is often needed for the summarization. For this reason, tools, such as SAS, SPSS, Ab Initio, and PERL, are often used.

| Policy | Year | Car | Premium | Claims | Number of Claims |
|---|---|---|---|---|---|
| 000 000 0001 | 1997 | CAR01 | | | |
| 000 000 0001 | 1998 | CAR01 | | | |
| 000 000 0001 | 1998 | CAR02 | | | |
| 000 000 0001 | 1999 | CAR01 | | | |
| 000 000 0001 | 1999 | CAR02 | | | |
| 000 000 0002 | 1997 | CAR01 | | | |
| 000 000 0002 | 1997 | CAR02 | | | |
| 000 000 0002 | 1998 | CAR01 | | | |
| 000 000 0002 | 1999 | CAR01 | | | |
| 000 000 0002 | 1999 | CAR02 | | | |
| 000 000 0003 | 1998 | CAR01 | | | |
| 000 000 0003 | 1999 | CAR01 | | | |
| 000 000 0004 | 1999 | CAR01 | | | |

This is an example of car insurance data. Every year the insurance company keeps track of claims by policy number, year, and car on the policy.

Data mining algorithms typically require the data in the format of one policy per row when we want to make predictions about policies. All rows need to have the same number of columns.

This requires transposing the data and calculating new values for the columns.

| Policy | Number of Years | Number of Car-Years | Number of Claims | Value of Claims | Total Premiums |
|---|---|---|---|---|---|
| 000 000 0001 | 3 | 5 | | | |
| 000 000 0002 | 3 | 5 | | | |
| 000 000 0003 | 2 | 2 | | | |
| 000 000 0004 | 1 | 1 | | | |

**Figure 3.5** An example of summarizing automobile insurance data.

## Add Derived Variables

The next step in the process is adding derived variables. Derived variables have values based on combinations of other values inside the data. Some simple examples of derived variables are

- Total number of transactions and sum of dollar amount
- Number of months when new charges were equal to 0
- Growth in usage from beginning of the period to the end
- Ratio of usage attributed to international, long-distance, and local calls
- Ratio of weight to height squared (the obesity index)

These are all examples of derived variables coming from the data within a single row. Another type of derived variable gives information about that row relative to all the others. For instance:

- **The revenue decile for each customer.** This is determined by taking the total amount spent in a given period for all customers and assigning 1 for the customers in the top decile, 2 for the customers in the second decile, and so on.
- **The churn rate by type of wireless phone.** This is determined by taking the most recently available churn information and determining the rate for each type of handset.
- **Profitability by demographics.** This is determined by taking historical profitability information by age and gender.
- **Fraud by amount of transaction.** This is determined by determining the amount of fraud that has historically been identified for transactions of different sizes.

These variables are powerful because past behavior is often a strong predictor of future behavior. In the wireless industry, for instance, handset churn rates are an important part of every churn model. However, these historical variables are not enough. Obtaining better predictive results requires combining the historical information with other types of data.

This type of derived variable is often available through an OLAP system. In fact, these types of variables show that there are many synergies between OLAP and data mining. Sometimes you can get the data from an OLAP system; sometimes you want to calculate these historical variables directly from the data used for data mining.

## Prepare the Model Set

The model set is the data that is used to actually build the data mining models. Once the data has been cleaned, transposed, and derived variables added, what more needs to be done?

There are a few things that we still have to take into account. If we are building a predictive model from historical data, then what is the frequency of the rarer outcomes in the model set? A good rule of thumb is that we want between 15 and 30 percent density of the rarer outcomes.

Consider fraud. The data may contain fewer than 1 percent cases of known fraud. Almost any model that we build on such a model set will be 99 percent accurate—by simply predicting no fraud. Very accurate and entirely useless.

There are several ways of handling such rare data. The most common is oversampling, which we discuss in Chapter 7.

At this point, we also need to divide the model set into the training, test, and evaluation sets. Only some of the data is used to create the model initially; other data is held back to refine the model and to predict how well it works.

We may also decide to build different models on different segments of data. For instance, when building a cross-sell model, we may start by building a model for the propensity to buy each different product. In this case, we might create a separate model set for each product as a prelude to making a prediction about that product.

## Choose the Modeling Technique and Train the Model

Once upon a time, training a neural network or building a decision tree was difficult, often requiring custom coding. Fortunately, data mining tools have eliminated most of the cumbersome details in building models, as well as providing much friendlier graphical environments in which to work. So, in a way, the actual building of models is the least time-consuming part of data mining, because it is now as simple as point and click.

At this point, we have entered the technical realm of building models. There are a variety of different data mining techniques to choose from; each technique has advantages and disadvantages, which are discussed in Chapter 5.

When time is available, multiple different models are often built on the same model set. The best algorithm and set of parameters is generally unknown in advance, and experimentation can determine what best fits the data in the model set.

The specifics of training a model depend, of course, on the algorithm chosen and on the tool being used. Some tools can generate lots of different models and will choose the best one automatically. Others are more interactive, requiring the user to determine which is the best.

## Check Performance of the Models

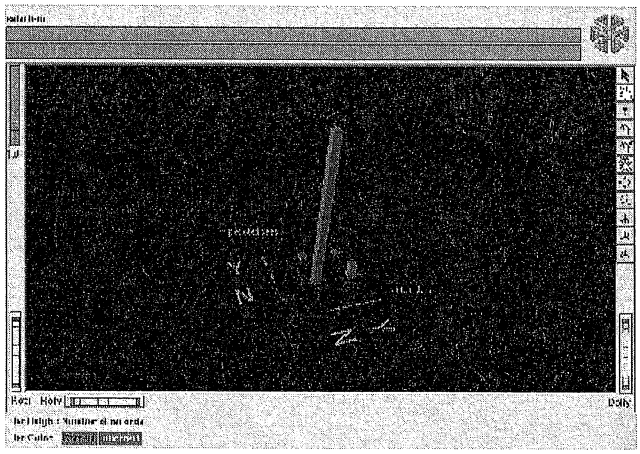
The final step is to check the performance of different models on the data. Different data mining techniques have different ways of measuring results. To compare the results between different models, though, we want to see how well the model performs on unseen data. This hold-out set is the evaluation set (which is part of the model set).

Figure 3.6 shows a confusion matrix, both graphically and as a table. This tells us how many predictions made by a predictive model are correct and how many are incorrect. Which is the best model depends on the business problem.

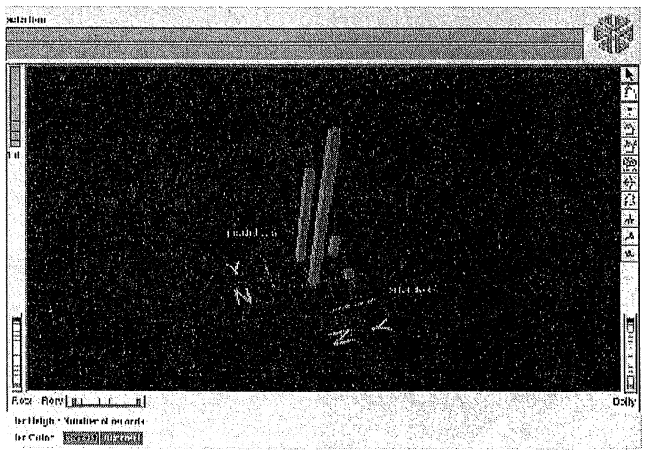
## Anticonfusion Matrix?

**The name *confusion matrix* puts off a lot of people. In fact, on hearing the word confusion, the concept suddenly becomes difficult to understand. Perhaps a better name would be a prediction accuracy chart. Alas, the name has stuck. A confusion matrix is measuring whether a model is confused or not; that is, whether the model is making mistakes in its predictions. Far from leading to confusion, they help clarify the performance of models—eliminating a bit of confusion in understanding the models.**

For a marketing response problem, we want to get as many potential responders as possible and we are not too concerned if a bunch of nonresponders are also in the data. That is, we do not care about false positives.



| | | Actual | |
|---|---|---|---|
| | | Y | N |
| Predicted | Y | 2% | 4% |
| | N | 12% | 82% |



| | | Actual | |
|---|---|---|---|
| | | Y | N |
| Predicted | Y | 7% | 40% |
| | N | 3% | 50% |

**Figure 3.6** Using confusion matrixes to see how well a model performs.

For a medical diagnostic test for cancer, the situation is quite different. We might use such a model as an initial screen. When there is an indication of cancer, the doctors would then run more elaborate tests. For this type of problem, we want to be sure that when the test says "no" it really means "no." In this case, we care a lot about false negatives—and we want as few as possible.

The lift chart or cumulative gains chart is another way of comparing the performance of different models. Figure 3.7 shows a cumulative gains chart for several different models. The greater the area between the line for the model and the diagonal line, the better the model.

What is the cumulative gains chart? The horizontal axes represent the ranking of the data according to the model. So, the 10 percent mark represents the 10 percent of the data that scores the highest with the model. The vertical axis shows the actual density of the outcome. If the top 20 percent of the data has 70 percent of the desired outcomes, then the model is doing well. It is far from the diagonal baseline. In this case, we would say that the model has a lift of 3.5; it is doing 3.5 times as well as we would expect.

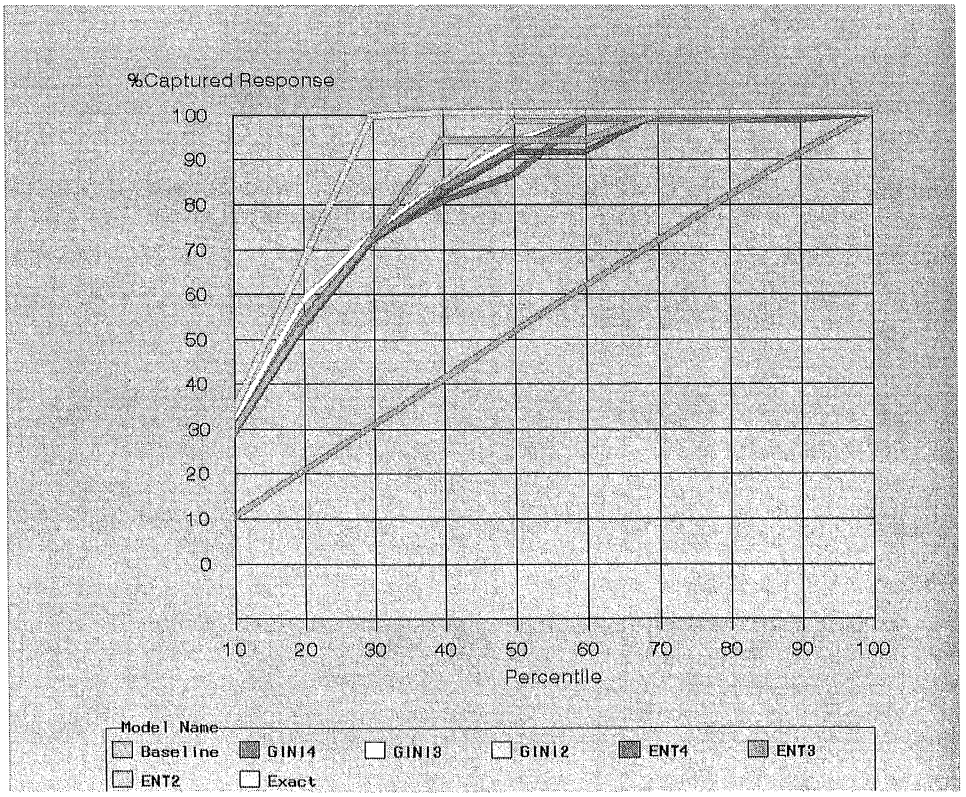



**Figure 3.7** A comparison of different models using a cumulative gains chart.

## Acting on the Results

Data mining serves no purpose if we never act on the results of the model. Acting on the results can take several different forms:

**Insights.** During the course of modeling, we may have learned new facts from the data. These may lead to insights about the business and about the customers. The insights need to be communicated!

**One-time results.** The results may be focused on a particular activity, such as a marketing campaign. In this case, the marketing campaign should be carried out, based on the propensities determined by the model.

**Remembered results.** The results may provide interesting information about customers, and this information should be accessible through a data mart or a data warehouse. Predicted customer profitability and best next offer are two examples that are often worthy of "publication" in a database.

**Periodic predictions.** The model may be used to score customers periodically, to determine the best next offer to make them or to determine whom to target for retention efforts.

**Real-time scoring.** The model itself may be incorporated into another system to provide real-time predictions. For instance, the profitability of a customer may be updated based on the results of online transactions, or a Web site may be customized based on a customer's predicted needs.

**Fixing data.** Sometimes the data mining effort uncovers data problems that significantly affect the performance of models. In these cases, sometimes the only actions are those that will result in cleaner, more complete data for future efforts.

The type of action can have an effect on the modeling. For instance, when using the scores that the model produces, there is always a time lag needed to deploy the results. This time lag is due more to the availability of recent data and to deployment scheduling than it is to the process of actually scoring the data. The time lag needs to be taken into account when building the models.

Deploying a model in real-time imposes requirements on how the results are scored and, perhaps, on the complexity of the data transformations permitted in the model. In these cases, we may want to export the model as code, such as C, C++, SAS code, C/SQL, or using database access languages.

Sometimes it is also valuable to incorporate a bit of experimental design into the process. For example, if we are predicting customer response to a product, we might actually have three different groups:

- A group of customers chosen based on the results of the data mining model, who get the marketing message

- A group of customers chosen at random, who get the marketing message
- A group of customers, chosen at random, who do not get the marketing message

What we hope is that the first group will have a high response rate, the second group will have a mediocre response rate, and the third will have a negligible response rate. In any case, we can test both the strength of the marketing message (the difference in response between the second and third groups) and the strength of the data mining (the difference between the first and second groups).

Even when data mining is used in a production environment, it is still valuable to have an additional, random group exposed to the marketing message. This helps future data mining efforts by bringing in an unbiased sample. For instance, if we predict that young couples between 25 and 35 with moderate incomes are most likely to purchase a home, then we might direct all marketing messages to them. Well, if we ever look at the data, we will find—lo and behold—that this group has a high response rate. This phenomenon is known as sample bias. By including a random group of people, not predicted to respond, we prepare ourselves for finding new patterns that may arise in the future.

## Measuring the Model's Effectiveness

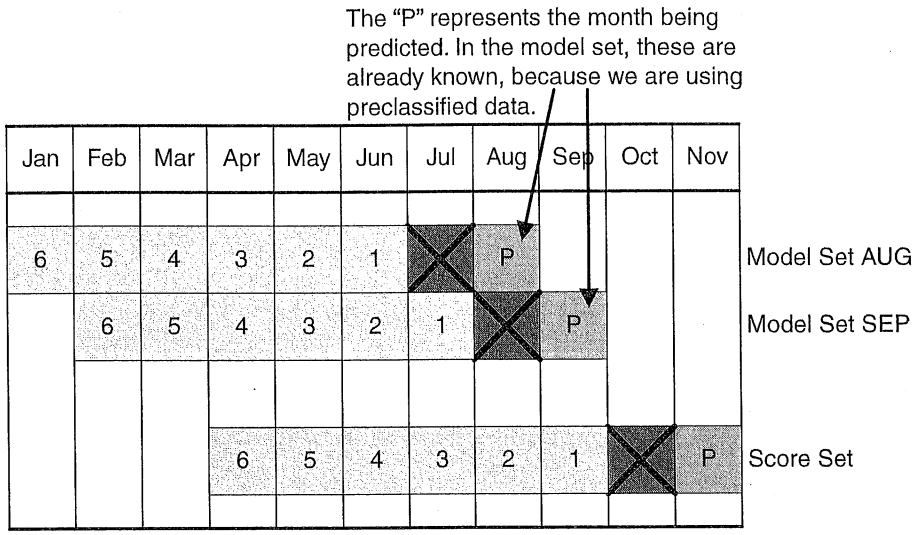
We have already discussed measuring the effectiveness of models on a hold-out set. This is for the purpose of evaluating the model. We also need to compare the results to what actually happened in the real world. This is particularly true when we are making predictions about future behavior.

Did the predicted behavior actually happen? That is, did the prospects accept the offer, did the customers purchase the new product or service, did they churn? The only way to really know is by comparing what actually happened to the prediction.

We have already discussed lift charts and confusion matrixes as methods of comparing the results of models. These are readily adapted to compare actual results to predicted results.

For a predictive model, actual results will usually be worse than predicted. The discrepancy arises because models perform less well the farther they get from the model set—the set of data used to create the model. Typically, the data set we are acting on is more recent than the data set used to create the model. The model captures patterns from the past and, over time, the patterns become less relevant. Figure 3.8 illustrates what typically happens on a monthly basis. Because the model set is older than the score set, we expect the results to degrade over time.

This model time chart shows that six months of historical data is being used to predict one month into the future.




Model performance usually degrades over time. We expect the model's predictions to be a bit less accurate on the score set than on the model set.

**Figure 3.8** A model time chart shows that the score set is usually more recent than the model set.

## What Makes Predictive Modeling Successful?

"Consistency is the hobgoblin of little minds," is a frequent misquote of Ralph Waldo Emerson (he actually wrote "A foolish consistency . . ."). Quite the opposite is true for predictive modeling. Predictions are only useful because they are consistent—and especially, because they are consistent over time. Otherwise, they would have no predictive value. With hard work and a bit of luck, our predictive models will not produce foolish consistencies.

### Time Frames of Predictive Modeling

Although the inner workings of the data mining techniques are interesting, it is possible to approach predictive models without considering the details of the techniques. Models simply transform inputs into predictions, whether using statistical regressions, neural networks, decision trees, nearest neighbor approaches, or even some technique waiting to be invented in the future. There are really two things to do with predictive models, as shown in Figure 3.9:

- The models are created using data from the past to make predictions. This process is called *training* the model.
- The models are then run on another set of data to assign outcomes. This process is known as *scoring*, and often predicts future outcomes based on the most recent data.

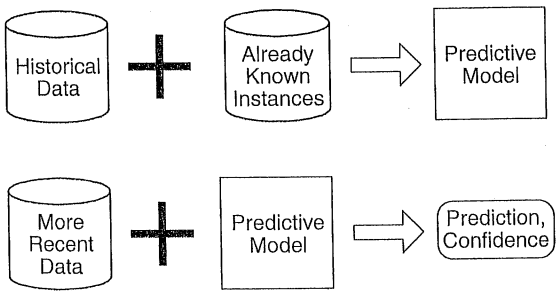
There are two time frames associated with developing models. The first is when they are being trained. At this point in time, the data is historical *and the outcomes are known*. These are the records used for training. The second time frame is when the models are being scored. At this point in time, the input data is available, but the outcomes are not known. The role of modeling is to assign probable outcomes.

When predictive models are being created, their performance can be measured only on the past data—because that is the only data that is available. Often, it is possible to achieve good results in the past that do not generalize well in the future, resulting in a model that looks very good on the model set and fails miserably when applied in practice. The methods given in this chapter help to reduce the likelihood of developing poor models.

Fortunately, a good predictor of the future is the past, so predictive modeling has proven to be a good approach for many types of problems. However, the past is by no means perfect. To make effective use of predictive models, we need to understand not only how to build them but also when they work well and when they don't.

## Modeling Shelf-Life

Looking at time frames also brings up two critical questions about models and their predictions:




**Training a Predictive Model** is the process of creating a model using historical data and already known instances of what you are trying to predict.

**Scoring a Predictive Model** is the process of applying the model to unseen data to make new predictions.

**Figure 3.9** Predictive models must be trained (created) before they are used (scored).

**What is the shelf-life of a model?** The things being modeled change over time, such as the business environment, technology, and customer base changes. This means that a model created five years ago, or last year, or last month, may no longer be valid. When this happens, you need to train a new model on more recent data.

**What is the shelf-life of a prediction?** Predictions also have a shelf-life. They are valid during a particular time frame. The classic example is predicting what will happen during a particular month (such as churn or making a purchase). And then using the prediction during a different month.

The whole process of predictive modeling is based on some key assumptions. These assumptions shed light on the process of building models.

## Assumption 1: The Past Is a Good Predictor of the Future

Using predictive models assumes that the past is a good predictor of the future. If we know how patients reacted to a drug in the past, we can be confident that similar patients will have roughly similar reactions in the future. Or if certain customers who are going bankrupt have behaved in a certain way, then similar customers will behave in similar ways in the future. Or, customers who bought widgets last month are similar to customers who will buy widgets next month. And so on.

It is important to recognize that this is an assumption about the problem being addressed and about the business environment. It is usually a pretty reasonable assumption, too. However, there are some cases where the past may not be a particularly good predictor of the future, because it can be hard to capture significant external events in the data. For instance, retail sales decrease during cold weather and blizzards. Lumber sales go up after a hurricane. Mortgage lending increases when interest rates go down. Trying to predict retail sales, lumber, or response to mortgage campaigns may produce anomolous predictions. One year, South Florida may look like a strong market for lumber sales; another year South Carolina, and another Louisiana. This is a case where past sales may not directly predict future ones.

Such seasonal patterns appear in many places. Electricity usage during the summer is driven by heat waves. The Christmas season and back-to-school season drive many retail sales. A strike at a competitor or currency devaluation may totally change the market. All of these can significantly change the assumption that the past is a good predictor of the future. For retail sales, for instance, it is usually a good idea to have at least one year of data available for data mining.

More subtle reasons may prevent the past from being a good predictor of the future. Definitely last year's Christmas-time toy fad is unlikely to sell as well

this year. Perhaps a more insidious danger is that data collected about customers during a booming economy may not have much predictive power in a less stable economy. During the very volatile world markets in 1998, several financial institutions went under for precisely this reason—the models they developed during years of relatively stable financial markets were not applicable in the more volatile markets.

External factors, to a greater or lesser degree, will always have an influence on the models being built. How do we know when the past—and especially the data we have about the past—is a good predictor of the future? Well, we can never know for sure. This is one major reason why it is critical to include domain experts in the modeling process. They understand the business and the market, and they often have insight about important factors that affect predictive models. It is also critical to include enough of the right data to make good decisions—especially when seasonal factors play a major role.

## Assumption 2: The Data Is Available

The data is available. Such a simple statement, and yet there are so many challenges in practice. Data may not be available for any number of different reasons:

- The data may not be collected by the operational systems. This is rather common: the telephone switch does not record the fact that a customer turned off call waiting; the cash register does not record the employee who rang the sale; the customer service system is unable to capture the customer ID, and so on.
- Another department may own the data, so it is difficult to get.
- The data resides in a data warehouse or other database, but the database is too busy most of the time to prepare extracts.
- The data is owned by an outside vendor, who manages the operational or decision support systems.
- The data is in the wrong format or the keys do not match.
- And so on.

It is important to remember that the data used to build the model must also be available to apply the model. For example, it is tempting to use survey responses as inputs into predictive models. There are several issues with this, but the most basic is that the same inputs must be available for scoring as well. So, if the responses are useful, it might mean having to survey all customers in order to apply the model. Such a survey is prohibitively expensive. An alternative is to try to find an outside vendor who can supply the same information—but if you can buy the information, why do a survey? (Of course,

analyzing survey data using data mining techniques can be very enlightening and valuable.)

A similar situation arises when a sample of the customer base is used to build a prototype data warehouse. This is one approach to building a data warehouse, and it has a nice, iterative feel. And, predictive models built using all the relatively clean and complete data from the prototype warehouse look really good—but they can't be scored for the larger customer base until the same data is available about all the customers. In this situation, it is often more valuable to build a less effective model using fewer attributes—but a model that can be scored for everyone.

Themes about data quality recur throughout the book, because data quality is the biggest issue facing data miners and other analysts. And ensuring that the right data is available is critical to building successful predictive models.

## Assumption 3: The Data Contains What We Want to Predict

To apply the lessons of the past to the future, we need to be comparing apples to apples and oranges to oranges. Predictive modeling works best when the goal is to predict the same outcome over and over and over again. In areas such as predicting risk for life insurance or predicting credit worthiness for mortgages, modeling has proven quite effective.

Often, the business people phrase their needs very ambiguously. They say, "we are interested in people who do not pay their bills." Does that mean missing a payment for one month? For three months? For six months? For four of the last six months? Does nonpayment mean not making the minimum payment or making no payment at all? These decisions can have a big effect on the resulting models.

Sometimes, predictive modeling can be hard to apply at all. For instance, it may or may not help in trying to predict which customers are likely candidates for an entirely new product. In this case, the new product may be targeted at a new customer base—and all information collected about existing customers is unlikely to help. Or, the new product may be targeted at a specific segment, such as high-end customers.

Sometimes business users have unreasonable expectations from their data. For instance, when building a response model, it is important to know two things: who responded to the campaign and who received the campaign. For advertising campaigns, the second group is not known—we would have to know who saw the advertisement. Without this information, it is not possible to build an effective response model. We can, however, compare the responders to a random sample of the general population.

## Lessons Learned

The virtuous cycle of data mining focuses on using data mining to derive business benefit. It consists of the following stages:

**Identifying the right business problem.** This stage uses the domain experts to identify important business problems and the data needed to resolve them. However, it is important to verify the assumptions made by the domain experts.

**Transforming the data into actionable results.** Building models is an iterative process that needs to focus on how the results will be used.

**Acting on the results.** The purpose of data mining is to act on the results, and this can happen in different ways. Sometimes, the results are insights into the business; sometimes results are used only once. At other times, they may be remembered and put into a data warehouse. Sometimes, the goal is to provide the capability for real-time scoring (particularly true in e-commerce). Disappointing results often show a need for richer and cleaner data.

**Measuring results.** The final stage is to measure the results of actions. These measurements feed the virtuous cycle by providing more questions and more data for additional data mining efforts.

Predictive modeling is the most common application of data mining. Its success rests on three assumptions. The first is that the past is a good predictor of the future. The second is that the data is available. And the third is that the data contains what we want to predict. Its success also requires taking into account the time frame of the model, and acting on the results before the model and predictions expire.