1

Why Data Mining?

Somerville, Massachusetts, home to one of the authors of this book, is also home to a woman from Cameroon who braids hair. Every few weeks, she arrives at my house with a bag of combs, brushes, and oils, and sets to work transforming my daughter's long, bushy mane into neat cornrows. Some of her clients prefer to come to her, but this particular client finds the four-hour procedure more pleasant if she can set up her chair directly in front of the VCR with her cordless phone, and her Indigo Girls and Ani Difranco tapes, close at hand.

Because she is familiar with the client's likes and dislikes, the hair artist brings no extensions or beads with her to our house. She remembers how each client likes her hair: the diameter of the braids; the distance between braids; the orientation of the cornrows; whether the braids should be formed into fanciful loops or left to hang down straight; whether to weave in ribbons or beads. In short, the customer and the service provider have an established, learning relationship. The longer that relationship lasts—the more the provider learns about the customer—the harder it becomes for the customer to switch to a new provider.

Small retail businesses rely on their knowledge of the customer to inspire loyalty. For nearly 20 years I have done my wine shopping at a small, local store called The Wine Cask. The owner and staff know my tastes and my price range. When I ask them for advice, I know that it will be based on their knowledge of me and my preferences as well as on their knowledge of their stock.

The people at The Wine Cask know a lot about wine. But, while that knowledge does give them an advantage over a big discount liquor store, and even over many specialty wine stores, it is their knowledge of me, the customer, that keeps me coming back. Another shop could open across the street and hire a staff of expert oenophiles, but it would take them many months to get to know me the way those at The Wine Cask do.

LARGE FIRMS NEED LEARNING RELATIONSHIPS TOO

It is widely recognized that firms of all sizes need to learn to emulate what small, service-oriented businesses have always done well—creating one-to-one relationships with their customers. Managers of companies large and small have sent books like *The One to One Future: Building Relationships One Customer at a Time* by Don Peppers and Martha Rogers and *The Loyalty Effect: The Hidden Force behind Growth, Profits, and Lasting Value* by Frederick Reichheld to the top of the business best-seller charts.

In every industry, forward-looking companies are trying to move towards the one-to-one ideal of understanding each customer individually and to use that understanding to make it easier for the customer to do business with them rather than with a competitor. These same firms are learning to look at the lifetime value of each customer so they know which ones are worth investing money and effort to hold on to and which ones to let drop. This change in focus from broad market segments to individual customers requires changes throughout the enterprise, but nowhere more than in marketing, sales, and customer support.

Data Warehousing Provides the Enterprise with a Memory

A small business builds relationships with its customers by *noticing* their needs, *remembering* their preferences, and *learning* from past interactions how to serve them better in the future. How can a large enterprise accomplish something similar when most customers may never interact personally with company employees? Even where there is customer interaction, it is likely to be with a different sales clerk or anonymous call-center employee each time, so how can the enterprise notice, remember, and learn from these interactions? What can replace the creative intuition of the sole proprietor who recognizes customers by name, face, and voice, and remembers their habits and preferences?

In a word, nothing. But through the clever application of information technology, even the largest enterprise can come surprisingly close. In large commercial enterprises, the first step—noticing what the customer does—has already largely been automated. On-line transaction processing (OLTP) systems are everywhere, collecting data on seemingly everything.

These days, we all go through life generating a constant stream of transaction records. When you pick up the phone to order a canoe paddle from L.L. Bean or a satin bra from Victoria's Secret, a transaction record is generated at the local phone company showing the time of your call, the number you dialed, and the long distance company to which you have been connected. At the long distance company, more records are generated recording the duration of your call and the exact routing it takes through the switching system. This data will be combined with other records that store your billing plan, name, and address in order to generate a bill. At the catalog company, your call is logged again along with information about the particular catalog from which you ordered and any special promotions you are responding to. When the customer service representative that answered your call asks for your credit card number and expiration date, the information is immediately relayed to a credit card verification system to approve the transaction. All too soon, the transaction reaches the bank that issued your credit card where it will appear on your next monthly statement. When your order, with its item number, size, and color, goes into the cataloger's order entry system, it will spawn still more records in the billing system and the inventory control system. Within hours, your order is also generating transaction records in a computer system at UPS or FedEx where it may be scanned many times between the warehouse and your home, allowing you to call an 800 number or check the shipper's Web site to track its progress.

The customer-focused enterprise regards every record of an interaction with a client or prospect—each call to customer support, each point-of-sale transaction, each catalog order, each visit to a company Web site—as a learning opportunity. But, learning requires more than simply gathering data. In fact, many companies gather hundreds of gigabytes of data from and about their customers without learning anything! Data is gathered because it is needed for some operational purpose, such as inventory control or billing. And, once it has served that purpose, it languishes on tape or gets discarded.

For learning to take place, data from many sources—billing records, scanner data, registration forms, applications, call records, coupon redemptions, surveys—must first be gathered together and organized in a consistent and useful way. This is called *data warehousing*. Data warehousing allows the enterprise to remember what it has noticed about its customers. Next, the data must be analyzed, understood, and turned into actionable information. That is where *data mining* comes in.

A Note on Terminology

Data mining brings together ideas and techniques from a variety of fields that have very different vocabularies. Statisticians, artificial intelligence (AI) researchers, database administrators, and marketing people use different words to mean the same thing and the same words to mean different things.

For statisticians, economists, and other quantitative researchers, "data mining" is a pejorative term. It refers to the practice of selectively trying to find data that will support a particular hypothesis. Just as one can find support for any opinion by searching long enough in the Bible, it is usually possible to find data to support any theory.

To a lawyer, to say that something is "actionable" means that it is grounds for a law suit. Yet this is a word that business people use frequently to describe information in response to which some positive action can be taken.

For this book, we have adopted the terminology of the marketing profession, and of database marketing in particular. Where a statistician sees dependent and independent variables, and an AI researcher sees features and attributes, we see records and fields. And when we say that we use data mining to find actionable market information, we certainly don't mean that we are cooking the books and expect to be sued!

Data Mining Provides the Enterprise with Intelligence

The data warehouse provides the enterprise with a memory. But, memory is of little use without intelligence. Intelligence allows us to comb through our memories noticing patterns, devising rules, coming up with new ideas to try, and making predictions about the future. This book describes the tools and techniques that add intelligence to the data warehouse. Using these techniques, you will be able to exploit the vast mountains of data generated by interactions with your customers and prospects in order to get to know them better.

Who is likely to remain a loyal customer and who is likely to jump ship? What products should be marketed to which prospects? What determines whether a person will respond to a certain offer? Which telemarketing script is best for this call? Where should the next branch be located? What is the next product or service this customer will want? Answers to questions like these lie buried in your corporate data, but it takes powerful data mining tools to get at them.

In this book you will learn how these tools work, and the strengths and weaknesses of each. We describe data mining techniques in sufficient detail that even if you have no previous experience with analytical modeling or data mining, you will understand what you can and cannot expect to accomplish with each approach, and how to pick the right tool or tools for a given data mining task. You will also learn a general methodology for data mining, how to select and prepare data sources so as to get the best out of the data mining techniques, and how to evaluate the results you obtain through data mining.

WHAT IS DATA MINING?

Data mining, as we use the term, is the exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules. For the purposes of this book, we assume that the *goal* of data mining is to allow a corporation to improve its marketing, sales, and customer support operations through better understanding of its customers. Keep in mind, however, that the data mining techniques and tools described here are equally applicable in fields ranging from law enforcement to radio astronomy, medicine, and industrial process control.

In fact, hardly any of the data mining algorithms we describe were first invented with commercial applications in mind. The commercial data miner employs a grab bag of techniques borrowed from statistics, computer science, and artificial intelligence research. The choice of a particular combination of techniques to apply in a particular situation depends on both the nature of the data mining task to be accomplished and the nature of the available data.

Later in the book, we define the tasks well-suited for data mining: *classification*, *estimation*, *prediction*, *affinity grouping*, *clustering*, and *description*. Some of these tasks are best approached in a top-down manner called *hypothesis testing*. In hypothesis testing, a database recording past behavior is used to verify or disprove preconceived notions, ideas, and hunches concerning relationships in the data.

Other tasks are best approached in a bottom-up manner called *knowledge discovery*. In knowledge discovery, no prior assumptions are made; the data is allowed to speak for itself. Knowledge discovery comes in two flavors—directed and undirected. Directed knowledge discovery attempts to explain or categorize some particular data field such as income or response. Undirected knowledge discovery attempts to find patterns or similarities among groups of records without the use of a particular target field or collection of predefined classes. All of these activities fall under our definition of data mining.

WHY NOW?

Although most of the data mining techniques described in this book have existed, at least as academic algorithms, for years or decades, it is only in the last several years that commercial data mining has caught on in a big way. This is due to the convergence in the 1990s of a number of factors:

- The data is being produced.
- The data is being warehoused.
- The computing power is affordable.
- The competitive pressure is strong.
- Commercial data mining software products have become available.

Let's look at each factor in turn.

Data Is Being Produced

Data mining only makes sense when there are large volumes of data. In fact, most data mining algorithms *require* large amounts of data in order to build and train the models that will then be used to perform classification, prediction, estimation, or other data mining tasks.

A few industries, including telecommunications and credit cards, have long had an automated, interactive relationship with customers that generated many transaction records, but it is only relatively recently that the automation of everyday life has become so pervasive. Today, the rise of supermarket point-of-sale scanners, automatic teller machines, credit and debit cards, pay-per-view television, home shopping, electronic funds transfer, automated order processing, electronic ticketing, and the like means that data is being produced and collected at unprecedented rates.

Data Is Being Warehoused

Not only is a large amount of data being produced, but also, more and more often, it is being extracted from the operational billing, reservations, claims processing, and order entry systems where it is generated and being fed into a data warehouse to become part of the corporate memory.

Data warehousing brings together data from many different sources in a common format with consistent definitions for keys and fields. It is generally not possible (and certainly not advisable) to perform computer- and input/output (i/o)-intensive data mining operations on an operational system that the business depends on just to function. In any case, operational systems store data in a format designed to optimize performance of the operational task. This format is generally not well-suited to decision-support activities like data mining. The data warehouse, on the other hand, is designed exclusively for decision support. This simplifies the job of the data miner.

Computing Power Is Affordable

Data mining algorithms typically require multiple passes over huge quantities of data. Many are computationally intensive as well. The continuing dramatic decrease in prices for disk, memory, processing power, and i/o bandwidth has brought once-costly techniques that were used only in a few government-funded laboratories into the reach of ordinary businesses.

The successful introduction of parallel relational database management software by major suppliers such as Oracle, Informix, Red Brick, Sybase, Tandem, and IBM, has brought the power of parallel processing into many corporate data centers for the first time. These parallel database server platforms provide an excellent environment for large-scale data mining.

Competitive Pressure Is Strong

Some of the most information-rich industries such as telecommunications, insurance, and financial services are experiencing greatly increased competition. Companies in these information-rich sectors of the economy have long had the data and resources to perform data mining. Now, for the first time, they have a strong business incentive to do so. Industries that have not traditionally been information-rich are striving to become so. Several trends are increasing the competitive importance of information:

- An increasingly service-based economy
- The advent of mass customization
- The increasing importance of information as a product in its own right

Every Business Is a Service Business

For companies in the service sector, information confers competitive advantage. That is why hotel chains record your preference for a non-smoking room and car rental companies record your preferred type of car. In addition, companies that have not traditionally thought of themselves as service providers are beginning to think differently. Does an automobile dealer sell cars or transportation? If the latter, it makes sense for the dealership to offer you a loaner car whenever your own is in the shop, as many are now doing.

Even commodity products can be enhanced with service. A home heating oil company that monitors your usage and delivers oil when you need more, sells a better product than a company that expects you to remember to call to arrange a delivery before your tank runs dry and the pipes freeze! Credit card companies, long distance providers, airlines, and even computer retailers often compete as much on service as they do on price.

Mass Customization

Mass customization means producing individually tailored products by combining selections from a large set of standard components. In the process, mass customizers necessarily gather data on sizes or preferences of each customer. In addition to allowing the vendor to understand more about individual customers, this data can be mined for insights on the market as a whole.

Levi-Strauss offers a line of custom-fit jeans. The company has equipped some of its stores with laser devices that record a customer's measurements. This data, along with the customer's stated preferences on fit (tight, baggy, etc.) is sent to the factory where the correct set of standard components will be assembled to order. The Custom

Foot does the same thing with shoes. Andersen Windows uses a mass customization system to allow customers to design their own windows.

Individual, Inc. goes farther by offering a mass customized product that gets smarter over time. Individual has an electronic clipping service. Customers provide a profile of their interests, then rate the articles they are sent as "very relevant," "somewhat relevant," or "not relevant." These ratings are used to automatically adjust the parameters of the information retrieval engine so that after a few weeks, the customer is receiving a customized newsletter containing only relevant articles.

Information as Product

Custom Clothing Technology Corporation, the company that designed the measuring technology used to fit custom Levis is currently working to develop a similar system for fitting brassieres. Eventually, it hopes to supply custom measurements for any type of clothing. When that day comes, the size database itself will be a valuable product.

Many catalog shoppers would be happy to register their individual measurements with an information broker in return for a personal size code. Then, when calling up Lands End or J. Crew they could simply supply their size code once and order everything from hats to shoes. Your grandchildren could put their size codes on their gift wish lists and you could carry your spouse's code in your wallet. Of course, for the system to work, the retailers would have to subscribe to a service provided by the information broker in order to check for the most up-to-date measurements associated with those codes. The information broker could charge a small fee for each inquiry in addition to the subscription charge.

It hasn't happened in body sizes yet, but information brokers are already doing a healthy business in many areas. IMS collects data on prescriptions from pharmacies and sells it back to drug companies who use it to track who is prescribing what to whom and to determine commissions for their sales forces. AC Nielsen does the same with television viewing habits. Market research firms like the NPD Group collect data on purchasing habits and resell it as reports.

Any company that collects valuable data is in a position to become an information broker. The *Cedar Rapids Gazette* takes advantage of its dominant position in a 22-county area of Eastern Iowa to offer direct marketing services to local businesses. The paper uses its own obituary pages and wedding announcements to keep its marketing database current.

Commercial Data Mining Software Products Have Become Available

There is always a lag between the time when new algorithms first appear in academic journals and excite discussion at conferences, and the time when commercial software incorporating those algorithms becomes available. There is another lag between the initial availability of the first products and the time that they achieve wide acceptance. For data mining, the period of widespread availability and acceptance is only just beginning.

Many of the techniques we discuss in this book started out in the artificial intelligence field. After a few years in universities and government labs, a new technique will start to be used by a few early adopters in the commercial sector. At this point in the evolution of a new technique, the software is typically available in source code to the intrepid user willing to retrieve it via FTP, compile it, and figure out how to use it by reading the author's Ph.D. thesis. Only after a few pioneers become successful with a new technique, does it start to appear in real products that come with user's manuals and help lines.

All of the techniques discussed in this book are available in commercial software products, but the techniques are at different points along the road from laboratory to data center. Neural networks and market basket analysis are almost old-hat by now. Decision tree methods and automatic cluster detection are close behind. Genetic algorithms and memory-based reasoning are only just beginning to make the transition.

HOW DATA MINING IS BEING USED TODAY

This whirlwind tour of a few interesting applications of data mining is intended to demonstrate the wide applicability of the data mining techniques discussed in this book. We hope that these vignettes will convey something of the excitement of the field and possibly suggest ways that data mining could be profitably employed in your own work.

Naming Names: Companies Have Privacy Concerns Too

We realize that it is tiresome to read case studies about "a major retailer" or a "regional Bell operating company." Unfortunately, much of the authors' experience in data mining has been inside companies that prefer not to publicize their data mining activities. Many of our clients believe that they gain significant competitive advantage through their data mining activities and would just as soon keep that knowledge to themselves.

Out of respect for these concerns we generally do not name any of the companies with which we have worked and some of the details of case studies drawn from our own experience, or from unpublished conversations with other practitioners, have been changed in ways that do not alter the moral of the story, but do protect the identity of the companies involved.

Wherever the actual names of companies appear in this book, the source of the information is published material from the companies themselves or from the press.

The Feds Use Data Mining to Track Down Criminals

The federal government of the United States was an early adopter of data mining technology. As part of the investigation of the Oklahoma City bombing case, the Unabomber case, and many lower-profile crimes, the FBI used automatic link analysis to sift through thousands of reports submitted by agents in the field looking for connections and possible leads.

The Treasury Department uses data mining to hunt for suspicious patterns in international funds transfer records; patterns that may indicate money laundering or fraud. Data mining tool vendors report that the Internal Revenue Service has expressed a lot of interest as well—so, watch out!

A Supermarket Becomes an Information Broker

Supermarkets are in a position to notice a lot about their customers these days, but most of them do not yet have the technical capability to link the purchase data captured by the check-out scanners with individual purchasers and households. One that does is the Safeway Corporation.

Safeway, like several other large chains, has turned itself into an information broker. The supermarket purchases address and demographic data directly from its customers by offering them discounts in return for using a Safeway Savings Club card when they make their purchases. In order to obtain the card, shoppers voluntarily divulge personal information of the sort that makes good input for predictive computer models.

From then on, each time the shopper presents the discount card, his or her transaction history is updated in a data warehouse somewhere. With every trip to the store, the shoppers teach the retailer a little more about themselves. The supermarket itself is probably more interested in aggregate patterns (what sells well together, what should be shelved together) than in the behavior of individual customers, but the information gathered on individuals is of great interest to the *manufacturers* of the products that line the store's aisles.

Of course, the store assures the customers that the information thus collected will be kept private and it is. Rather than selling Coca-Cola a list of frequent Pepsi buyers and vice-versa, the chain sells *access* to customers who, based on their known buying habits and the data they have supplied, are likely prospects for a particular supplier's product. Safeway charges 5.5 cents per name to suppliers who want their coupon or special promotional offer to reach just the right people. Since the coupon redemption also becomes an entry in the shopper's transaction history file, the precise response rate of the targeted group is a matter of record. Furthermore, a particular customer's response or lack thereof to the offer becomes input data for future predictive models.

American Express and other charge card suppliers do much the same thing, selling advertising space in and on their billing envelopes. The price they can charge for space in the envelope is directly tied to their ability to correctly identify people likely to respond to the ad. That is where data mining comes in.

A Business Based on Community Knowledge

One of the reasons for learning to understand the behavior of individual customers is to be able to generalize so as to make predictions about the behavior of other, similar people. This is what Peppers and Rogers call "developing community knowledge." In their book, *Enterprise One to One: Tools for Competing in the Information Age*, they cite Firefly (http://www.ffly.com) as an example of a business based on community knowledge.

It is also an example of a business entirely dependent on data mining. Firefly asks its members to rate music and movies. Based on the likes and dislikes revealed by these ratings, subscribers are automatically clustered into groups of like-minded people. Once a new subscriber has been assigned to a cluster, the system can make recommendations of movies and music he or she will probably enjoy because other members of the cluster have rated them highly.

The beauty of the system is that the more you use it, the better it understands which other subscribers tastes are similar to your own and the better its advice becomes.

Cross Selling

USAA is an insurance company that markets to active duty and retired military personnel and their families. The company credits information-based marketing, including data mining, with a doubling of the number of products held by the average customer to 4.5. USAA keeps detailed records on its customers and uses data mining to predict where they are in their life cycles and what products they are likely to need.

Another company that has used data mining to improve its cross-selling ability is Fidelity Investments. Fidelity maintains a data warehouse filled with information on all of its retail customers. This information is used to build computer models that predict what other Fidelity products are likely to interest each customer. When an existing customer calls Fidelity, the phone representative's screen shows exactly where to lead the conversation.

In addition to improving the company's ability to cross sell, Fidelity's retail marketing data warehouse has allowed the financial services powerhouse to build models of what makes a loyal customer and thereby increase customer retention. These models caused Fidelity to retain a marginally profitable bill-paying service that would otherwise have been cut. It turned out that people who used the service were far less likely than the average customer to take their business to a competitor. Cutting the service would have encouraged a profitable group of loyal customers to shop around.

A central tenet of the one-to-one marketing philosophy is that it is more profitable to focus on "wallet share" or "customer share," the amount of business you can do with each customer, than on market share. From financial services to heavy manufacturing, innovative companies are using data mining to increase the value of each customer.

Warranty Claims Routing

A diesel engine manufacturer receives a constant stream of warranty claims from independent dealers who have performed maintenance on engines covered by the manufacturer's warranty. Each of the claims must be examined by an expert to determine if the labor and parts used seem reasonable and appropriate. There has long been a set of rules used to exempt certain claims that are considered so routine that they can be paid without any detailed examination.

The company is now investigating the use of data mining to expand the number of claims paid automatically by discovering an expanded set of rules to describe classes of claims always approved by the professional adjusters. This automated claims routing has the potential to save the company millions of dollars.

Holding on to Good Customers

Data mining is being used to promote customer retention in any industry where customers are free to change suppliers at little cost and competitors are eager to lure them away. Banks call it attrition. Cellular phone companies call it churn. By any name, it is a big problem. By gaining an understanding of *who* is likely to leave and *why*, a retention plan can be developed that addresses the right issues and targets the right customers.

Southern California Gas is just one example of a former regulated monopoly that now has to compete for customers. Before deregulation, the company didn't even have a marketing department. Now it has a database marketing program that integrates customer usage and billing data with credit information and U.S. Census data. By applying data mining techniques to this data the utility was able to figure out who would benefit most from a level payment plan. Direct mail based on the model yielded response rates between 7 and 11 percent, a phenomenal result for direct mail. The utility has also learned that some groups of customers are much more price-sensitive than others, with small, commercial heating customers being the quickest to put on a sweater and turn off the gas.

It costs more to bring in a new customer than it does to hold on to an existing one, but often the incentive offered to retain a customer is quite expensive. Data mining is the key to figuring out which customers should get the incentive, which customers will stay without the incentive, and which customers should be allowed to walk.

Weeding Out Bad Customers

In many industries, some customers cost more than they are worth. These might be people who consume a lot of customer support resources without buying much. Or, they might be those annoying folks who carry a credit card they rarely use, are sure to pay off the full balance when they do, but must still be mailed a statement every month. Even worse, they might be people who owe you a lot of money when they declare bankruptcy.

The same data mining techniques that are used to spot the most valuable customers can also be used to pick out those that should be turned down for a loan; those who should be allowed to wait on hold the longest time; and those who should always be assigned a middle seat near the engine (or is that just our paranoia showing?).

Revolutionizing an Industry

In 1988, the idea that a credit card issuer's most valuable asset is the information it has about its customers was pretty revolutionary. It was an idea that Richard Fairbank and Nigel Morris shopped around to 25 banks before Signet Banking Corporation decided to give it a try.

Signet acquired behavioral data from many sources and used it to build predictive models. Using these models it launched the highly successful balance transfer card product that changed the way the credit card industry works. In 1994, Signet spun off the card operation as Capital One which is now one of the top 10 credit card issuers with over $11 billion in loans outstanding. The same aggressive use of data mining technology that fueled such rapid growth is also responsible for keeping Capital One's loan loss rates among the lowest in the industry.

Data mining is at the heart of the marketing strategy of all the so-called monoline credit card banks, First USA, MBNA, Advanta, and Capital One. At least one of these, First USA, having been aquired by Bank One, is now positioned to offer a full range of banking services. Data mining will allow the combined company to mine its extensive credit card portfolio for cross-selling opportunities for car loans, mortgages, and other general banking services.

Capital One is making plans to diversify into other information businesses, both financial and nonfinancial. Meanwhile, Signet, Capital's former parent company, is putting information-based marketing and information-based decision-making at the center of its approach to all its banking services.

Credit card divisions may have led the charge of banks into data mining, but other divisions are not far behind. At First Union, a large North Carolina-based bank, data mining techniques are used to predict which customers are likely to be moving soon. For most people, moving to a new home in another town means closing the old bank account and opening a new one, often with a different company. First Union set out to improve retention by identifying customers who are about to move and making it easy for them to transfer their business to another First Union branch in the new location. Not only has retention improved markedly, but also a profitable relocation business has developed. In addition to setting up a bank account, First Union now arranges for gas, electricity, and other services at the new location.

And Just About Anything Else

These applications should give you a feel for what is possible using data mining, but they do not come close to covering the full range of applications. The data mining techniques described in this book have been used to find quasars, design army uniforms, detect second-press olive oil masquerading as "extra virgin," teach machines to read aloud, and recognize handwritten letters. They will, no doubt, be used to do many of the things your business will require to grow and prosper in the coming century.