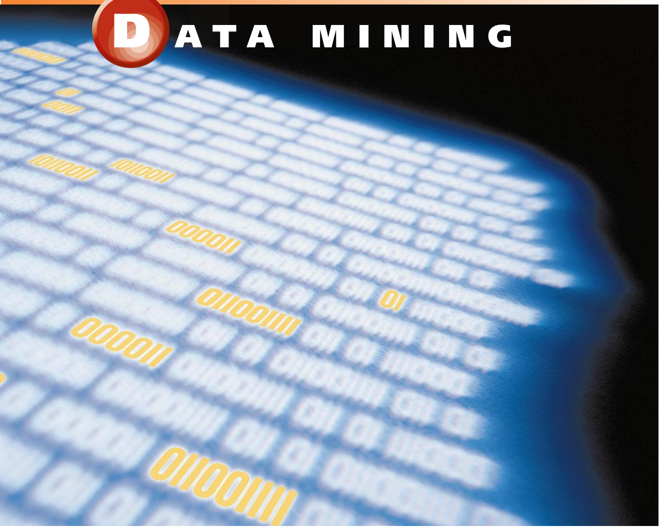


Assessing Loan Risks: A Data Mining Case Study

Rob Gerritsen

Imagine what it would mean to your marketing clients if you could predict how their customers would respond to a promotion, or if your financial clients could predict which applicants would repay their loans. Data mining has come out of the research lab and into the real world to do just such tasks.

Defined as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (*Advances in Knowledge Discovery and Data Mining*, U.M. Fayyad et al., eds., MIT Press, Cambridge, Mass., 1996), data mining frequently uncovers patterns that predict future behavior. It is proving useful in diverse industries like banking, telecommunications, retail, marketing, and insurance. Basic data mining techniques and models also proved useful in a project for the US Department of Agriculture.

Basic data mining techniques helped the Rural Housing Service better understand and classify problem loans.

TRACKING 600,000 LOANS

The USDA's Rural Housing Service administers a loan program that lends or guarantees mortgage loans to people living in rural areas. To administer these nearly 600,000 loans, the department maintains extensive information about each one in its data warehouse. As with most lending programs, some USDA loans perform better than others.

Last-resort lender

The USDA chose data mining to help it better understand these loans, improve the management of its lending program, and reduce the incidence of problem loans. The department wants data mining to find patterns that distinguish borrowers who repay promptly from those who don't. The hope is that such patterns could predict when a borrower is heading for trouble.

Primarily, it's the USDA's role as lender of last resort that drives the difference between how it and commercial lenders use data mining. Commercial lenders use the technology to predict loan-default or poor-repayment behaviors at the time they decide to make a loan. The USDA's principal interest, on the other hand, lies in predicting problems for loans already in place. Isolating problem loans lets the USDA devote more attention and assistance to such borrowers, thereby reducing the likelihood that their loans will become problems.

Training exercise

The USDA retained my company, Exclusive Ore, to provide it with data mining training. As part of that training exercise, my colleagues and I performed a preliminary study with data extracted from the USDA data warehouse. The data, a small sample of current mortgages for single-family homes, contains about 12,000 records, roughly 2 percent of the USDA's current database. The sample data includes information about

- the loan, such as amount, payment size, lending date, and purpose;

Inside

1520-9202/99/$10.00 © 1999 IEEE

- the asset, such as dwelling type and property type;
- the borrower, such as age, race, marital status, and income category; and
- the region where the loan was made, including the state and the presence of minorities in that state.

Table 1. Data mining taxonomy.

Predictive Models		Descriptive Models	
Classification	Regression	Clustering	Association

Using data mining techniques, we planned to sift through this information and extract the patterns and characteristics common in problem loans.

BUILDING DATA MODELS

Data mining builds models from data, using tools that vary both by the type of model built and, within each model domain, by the type of algorithm used. As Table 1 shows, at the highest level this taxonomy of data mining separates models into two classes: predictive and descriptive.

Predictive models

As its name suggests, a predictive model predicts the value of a particular attribute. Such models can predict, for example,

- a long-distance customer's likelihood of switching to a competitor,
- an insurance claim's likelihood of being fraudulent,
- a patient's susceptibility to a certain disease,
- the likelihood someone will place a catalog order, and
- the revenue a customer will generate during the next year.

When, as in the first four examples, a prediction relates to class membership, the model is called a *classification model*, or simply a *classifier*. The classes in our first four examples might be, respectively, loyal versus disloyal, legitimate versus fraudulent, susceptible versus indeterminate versus unsusceptible, and buyer versus nonbuyer. In each case, the classes typically contain few values.

When, as in the final example, the model predicts a number from a wide range of possible values, the model is called a *regression model* or a *regressor*.

Descriptive models

The class of descriptive models encompasses two important model types: clustering and association. *Clustering* (also referred to as segmentation) lumps together similar people, things, or events into groups called clusters. Clusters help reduce data complexity. For example, it's probably easier to design a different marketing plan for each of six targeted customer clusters than to design a specific marketing plan for each of 15 million individual customers.

Association models involve determinations of affinity—how frequently two or more things occur together. Association is frequently used in retail, where it is called *market basket analysis*. Such an analysis will generate rules like "when people purchase Halloween costumes they also purchase flashlights 35 percent of the time." Retailers use these rules to plan shelf placement and promotional discounts.

Although a descriptive model is not predictive, the converse does not hold: Predictive models often *are* descriptive. Actually, a predictive model's descriptive aspect is sometimes more important than its ability to predict. For example, suppose a researcher builds a model that predicts the likelihood of a particular cancer. The researcher might be more interested in examining the factors associated with that cancer—or its absence—than with using the model to predict if a new patient has the disease. Almost all predictive models can be used descriptively.

Algorithmic implementations

Several algorithms exist that implement the models I've described. Classifiers are most commonly implemented with neural network, decision tree, Naïve Bayes, or *k*-nearest-neighbor algorithms. Regressors can be implemented with neural networks or decision trees. Clustering and association models each also have several well-known algorithms.

WORKING WITH CLASSIFIERS

How do you choose from among these algorithms? In our case, the USDA needed to use a specific type of predictive model, the classifier, to catalog which loan types would likely go into default. When choosing an algorithm for a predictive model, you must weigh three important criteria: accuracy, interpretability, and speed.

Accuracy

You measure accuracy by generating predictions for cases with known outcomes and then compare the predicted value to the actual value. For classifiers a prediction is either right or wrong, so we can state the accuracy as percentage correct, or as an error rate (percentage wrong). Despite the claims you may encounter from various software vendors, no "most accurate" algorithm exists. In some cases, Naïve Bayes will produce the most accurate classifier; in others, a model built with a decision tree, neural network, or *k*-nearest-neighbor algorithm will be more accurate.

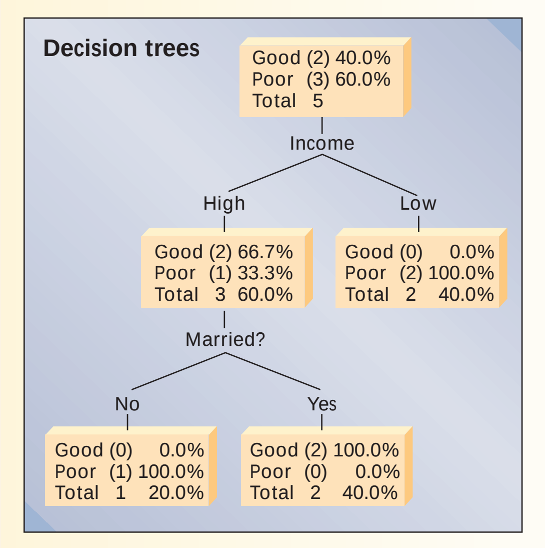
Worse, you cannot determine in advance which algorithm will produce the most accurate model for a particular data set. Thus, we usually try to apply at least two algorithms to

Predictive Modeling Techniques

Most commercial classification and regression tools use one or more of the following techniques.

Decision tree. **As its name implies, this algorithm generates a tree-like graphical representation of the model it produces. Usually accompanied by rules of the form "if condition then outcome," which constitute the text version of the model, decision trees have become popular because of their easily understandable results. Some commonly implemented decision tree algorithms include Chi-squared automatic interaction detection (CHAID), and classification and regression trees (CART). Although all these algorithms do classification extremely well, some can also be adapted to regression models.**

Figure A. Sample decision tree model for classifying loan prospects by income and marital status.



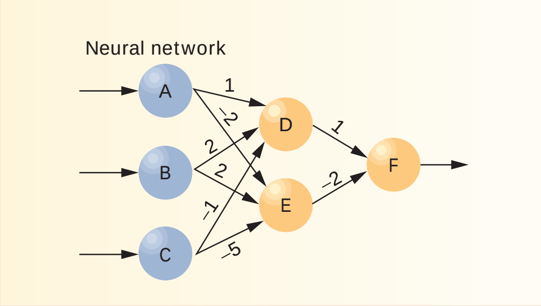
Neural networks. **Based on an early model of human brain function, neural networks do classification and regression equally well. More complex than other techniques, neural networks have often been described as a "black box" technology. They require setting numerous training parameters and, unlike decision trees, provide no easily understandable output.**

Naïve Bayes. **This technique limits its inputs to categorical data and applies only to classification. Named after Bayes's theorem, the technique acquired the modifier "naïve" because the algorithm assumes that variables are independent when they may not be. Simplicity and speed make Naïve Bayes an ideal exploratory tool. The technique operates by deriving conditional probabilities from observed frequencies in the training data.**

K-nearest neighbor. **Also known as K-NN, this algorithm differs from other techniques in that it has no distinct training phase—the data itself becomes the model. To make predictions for a new case using this algorithm, you find the group with most similar cases ("k" refers to the number of items in this group) and use their predominant outcome for the predicted value.**

To read more about predictive modeling techniques, check the Data Mining section, Technology subsection of Exclusive Ore's home page at http://www.xore.com.

Figure B. Sample neural network.



a data set to see which has the best accuracy. Our experience shows that neural networks and decision trees frequently have somewhat higher accuracy than Naïve Bayes, but not always. For regression, we have found that neural networks sometimes provide the highest accuracy.

Interpretability

How easy is it to understand the patterns found by the model? The *k*-nearest neighbor algorithm, an exception because it actually does not produce a model, has no interpretability value and thus scores worst here. Neural network models also produce little useful information, although you can write software that analyzes the neural-network model for information about the patterns and relationships in the model. Naïve Bayes and decision tree models produce the most extensive interpretive information. A Naïve Bayes model will tell you which variables are most important with respect to a particular outcome: "The use of voice mail is the strongest indicator of loyalty," and "customers who use voice mail more than 20 times a month are 15 times less likely to close their accounts in the next month," for example. Decision trees can find and report interactions—for example, "customers who never use voice mail, who live in Connecticut, whose accounts have been open between six and 14 months, and whose usage in the last three months has declined from the preceding three months, have a 23 percent probability of closing their accounts next month."

Speed

Two processes associated with predictive modeling emphasize speed: the time it takes to

- train a model and
- make predictions about new cases.

The *k*-nearest-neighbor algorithm's zero training time makes it the fastest trainer, but the model makes predictions extremely slowly. The three other major algorithms make predictions just as quickly as each other (and much more quickly than *k*-nearest-neighbor), but vary significantly in their training time, which lengthens in proportion to the number of passes each must make through the training data. Naïve Bayes trains fastest of the three because it takes only one pass through the data. Decision trees vary, but typically require 20 to 40 data passes. Neural networks may need to pass over the data 100 to 1,000 times or more.

PUTTING ALGORITHMS TO WORK

At the USDA, our goal was to build a model that would predict the loan classification based on information about the loan, borrower, and property. Often, to maximize our processing and results-generating efficiency, we use several algorithms together. Because of Naïve Bayes' speed and interpretability, we use it for initial explorations, then follow up with decision tree or neural network models.

Self-taught tools

To build a predictive model, a data mining tool needs examples: data that contains known outcomes. The tool will use these examples in a process—variously named learning, induction, or training—to teach itself how to predict the outcome of a given process or transaction. The column of data that contains the known outcomes—the value we eventually hope to predict—also has various names: the dependent, target, label, or output variable. Finally, all other variables are variously called features, attributes, or the independent or input variables. Data mining's eclectic nature fostered this inconsistency in naming—the field encompasses contributions from statistics, artificial intelligence, and database management; each field has chosen different names for the same concept.

The dependent or output variable we used for the USDA loan classification model has five values: problemless, substandard, loss, unclassified, and not available. Approximately 80 percent of loans fell into the problemless category. For each of the 12,000 mortgages in the sample, we knew in advance and included the correct loan classification.

Data mining consists of a cycle of generating, testing, and evaluating many models. The data mining cycle for our USDA project illustrates this process and highlights some common problems that modelers face in mining data.

Descriptive Modeling Techniques

Most descriptive modeling tools use one or more of the following techniques.

Clustering. **A descriptive technique that groups similar entities and allocates dissimilar entities to different groups, clustering can find customer-affinity groups, patients with similar profiles, and so on. Clustering techniques include a special type of neural net called a Kohonen net, as well as k-means and demographic algorithms. Highly subjective, clustering requires using a distance measure, like the nearest neighbor technique. Because clusters depend completely on the distance measure used, the number of ways you can cluster the data can be as high as the number of data miners doing the clustering. Thus, clustering always requires significant involvement from a business or domain expert who must judge whether the resulting clusters are useful.**

Association and sequencing. **Using these techniques can help you uncover customer buying patterns that you can use to structure promotions, increase cross-selling or anticipate demand. Association helps you understand what products or services customers tend to purchase at the same time, while sequencing reveals which products customers buy later as follow-up purchases. Often called market basket analysis, these techniques generate descriptive models that discover rules for drawing relationships between the purchase of one product and the purchase of one or more others.**

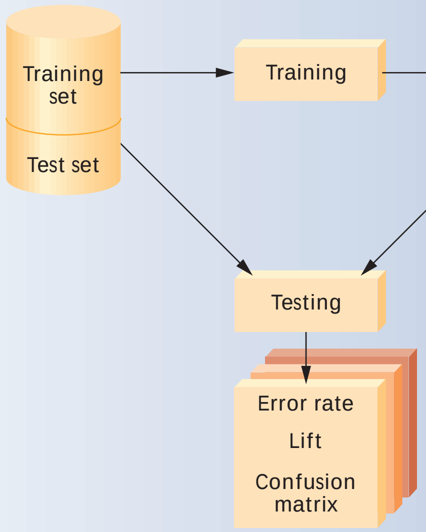
To read more about descriptive modeling techniques, check the Data Mining section, Technology subsection of Exclusive Ore's home page at http://www.xore.com.

Building models and a test database

We built the models using two thirds of the data—8,000 rows—and set aside the remaining data as an independent data set for testing the models. Testing reveals how well a model predicts the target variable—in this case, loan classification. During testing, we apply the model to the test data and predict the loan classification for each borrower. Because we also know the actual loan classification, we can compare the predicted value to the actual value for all 4,000 cases. From this data we can easily compute an accuracy score, the *predictive accuracy*.

The first model we built performed poorly, giving a pre-

Figure 1. Training a model. Analysts used two-thirds of the sample data to train the model, then used the remaining third as a test set to check various accuracy measures.



dictive accuracy of around 50 percent. This result prompted us to look closer at some of the noncategorical variables, like loan and payment amounts. We found that the skewed distribution of these values negatively affected the model. Payment amount is a good example of this effect: Although a few loans required large monthly payments of up to $60,000, most required payments smaller than $400.

DRILLING DEEPER

Commercial implementations of the Naïve Bayes algorithm requires the "binning" of numeric values. The algorithm we used in our case study automatically binned all numeric values into five bins. Since payment amounts range from $0 to $60,000, it divided the bins into five ranges of 12,000 each, starting with $0 to $11,999 and ending with $48,000 to $60,000. It then assigned each borrower's payment amount to a payment-amount bin, which the data mining algorithms use in place of the payment amount itself. Although binning itself is not a significant requirement, we found that the exact binning method used can significantly influence results.

Adjusting bin ranges

Because the loan's payment amount has a non-normal and nonuniform distribution, the equal-range bins we used by default proved poor predictors. Since 80 percent of the loans required payments of $400 or less, almost 99 percent of the loans landed in the first bin, which ranged between 0 to 12,000. As a result, after binning we could no longer distinguish between a loan with a truly small payment, like $100, and one with a fairly significant payment of $1,000 or even $10,000.

Yet someone with a small payment might be much less likely to have trouble than would a borrower with a large payment. Thus the default binning eradicated any relationship between loan amount and repayment behavior. Although we use data mining to look for patterns, in this case our binning may have actually removed one.

Indeed, redesigning the bins so that each contained approximately one-fifth of the total population improved the model's accuracy to 67 percent, with accuracy climbing up to 76 percent in predicting the problemless and loss categories. These results showed clearly that our default binning had removed important patterns.

Pruning irrelevant values

Our revised accuracy ratings proved too good to last, however. We now found something we'd overlooked: The sample data includes a "total loan amount due" field. When a lender stops paying, the value in this field grows continually larger as more payments become past due. At some point during default, depending on the loan's conditions, the entire principal falls due. The initial classification model therefore relied on this field as an excellent predictor for substandard and loss loans. This model is not very useful because it relies on after-the-fact information. However, when we removed this field from the data available for mining, overall accuracy dropped to 46 percent, and accuracy for predicting the loss category fell to 37 percent.

Having swung from too-good-to-be-true results back to abominable ones, we reviewed the data again, focusing this time on the loan class itself. Two loan class values—unclassified and not available—each occurred in less than 1 percent of cases. Having no interest in predicting membership in these class values, we decided to discard rows that contained either of them. This left three different values for loan class: problemless, substandard, and loss. Because we sought to predict those loans that might require attention, we combined the substandard and loss classes into a single not OK class. For consistency, we changed the problemless class's name to OK.

At first glance, the model we now generated—with an overall predictive accuracy of 82 percent—appeared pretty good. However, closer examination showed that, because it predicted only 20 percent of all problem loans, the Not OK class's accuracy fell disappointingly short. Being the most important class relative to taking actions on potential problem loans, Not OK's performance showed that our models required further refinement.

Refining with decision trees

After initially exploring the data with the Naïve Bayes classification algorithm, we also trained a decision tree model. As is often the case, the decision tree algorithm exhibited improved accuracy, generating a predictive accuracy of almost 85 percent. Accuracy in predicting the Not OK class also improved slightly, to 23 percent.

Yet accuracy, in and of itself, was not our only goal. When you account for costs or savings, you may find that even a seemingly low accuracy can yield significant benefits. The numbers that follow result from pure speculation on our part, and do not reflect actual costs or problem frequencies at the USDA.

First, assume that the average problem loan costs $5,000 and that that the USDA encounters 50,000 problem loans annually. If early intervention can prevent 30 percent of such cases, and each intervention costs $500, the USDA can still save approximately $11.5 million annually even if our data mining anticipates only 23 percent of all Not OK loans.

This figure does not, however, account for the cost of intervening with accounts that actually would not have become a problem. On the decision tree, about 29 percent of the accounts predicted as being Not OK will actually be OK—another statistic produced by testing the model. If we assume that such nonrequired interventions also cost $500, the net annual savings drops to $9.1 million. Yet even this adjusted figure shows that a low accuracy rate could still significantly reduce costs.

Data mining increases understanding by showing which factors most affect specific outcomes. For the USDA, the initial models revealed that the important factors to loan outcome included loan type, such as regular or construction; type of security, such as first mortgage or junior mortgage; marital status; and monthly payment size. We based our models on a small data sample and plan additional validation to determine the true effect of these factors.

The USDA's preliminary data mining study sought to demonstrate the technology's potential as a predictor and learning tool. In the near future, the department plans to expand the limited number of attributes available for data mining. In particular, it plans to include payment histories in the warehouse, and we hope that this data will help further improve the model's accuracy. Eventually, the USDA will use these models to identify loans for added attention and support, with the goal of reducing late payments and defaults. ■

***Rob Gerritsen** is a founder and president of Exclusive Ore Inc., a data mining and database management consultancy. Contact him at rob@xore.com.*

See our new cyber home!

Visit *IT Pro online* for extras like

- hot links to IT sites
- video clips of *IT Profiles*
- career resources

IEEE Computer Society

computer.org/itpro/

Profile

Steve Deering: Multicast Waits for Its Killer App

Most Internet routers already support multicasting, so adoption could be quick.

Steve Deering

Profile

Andy Grove Takes a Close Look at Networking

Intel has become one of the biggest networking companies and its cofounder has plenty to say about the state of the network.

Andy Grove, Intel

To View Clips...

Video clips require a Real Player to view. If you do not have a player, you may downloa