Journal of Comparative International Management

Applying and Evaluating Models to Predict Customer Attrition Using Data Mining Techniques

Tom Au, Guangqin Ma et Shaomin Li

Volume 6, numéro 1, june 2003

URI : https://id.erudit.org/iderudit/jcim6_1art02

Aller au sommaire du numéro

Éditeur(s)

Management Futures

ISSN

1481-0468 (imprimé)
1718-0864 (numérique)

Découvrir la revue

Citer cet article

Au, T., Ma, G. & Li, S. (2003). Applying and Evaluating Models to Predict Customer Attrition Using Data Mining Techniques. *Journal of Comparative International Management*, *6*(1), 10–22.

Résumé de l'article

As competition intensifies, retaining customers becomes one of the most serious challenges facing customer service providers. Customer attrition prediction models hold great promise as powerful tools for enhancing customer retention. Several statistical methods have been applied to develop models predicting customer attrition. Yet little research is done on the relative performance of models developed by different methods. The lack of knowledge about the performance of various prediction models is more pronounced due to the nonlinear nature of the combined causes of attrition (such as switching to another provider or canceling a service). The development of data mining techniques has made the comparison of prediction power of different models more efficient and easier. In this article we demonstrate how to use data mining techniques and software to fit and compare different customer attrition prediction models, using data from a major telecom service provider.

Journal of Comparative International Management
2003, Vol. 6, No. 1, 10-22

Applying and Evaluating Models to Predict Customer Attrition Using Data Mining Techniques

by

Tom Au
AT&T, USA

Shaomin Li*
Old Dominion University, USA

and

Guangqin Ma
AT&T, USA

As competition intensifies, retaining customers becomes one of the most serious challenges facing customer service providers. Customer attrition prediction models hold great promise as powerful tools for enhancing customer retention. Several statistical methods have been applied to develop models predicting customer attrition. Yet little research is done on the relative performance of models developed by different methods. The lack of knowledge about the performance of various prediction models is more pronounced due to the nonlinear nature of the combined causes of attrition (such as switching to another provider or canceling a service). The development of data mining techniques has made the comparison of prediction power of different models more efficient and easier. In this article we demonstrate how to use data mining techniques and software to fit and compare different customer attrition prediction models, using data from a major telecom service provider.

* Please address all correspondence to Shaomin Li, Department of Management, Old Dominion University, Norfolk, VA 23529, sli@odu-edu.

INTRODUCTION

Customer attrition refers to the phenomenon whereby a customer leaves a service provider.[1] As competition intensifies, preventing customers from leaving is a major challenge to many businesses such as telecom service providers (Ganesh, Arnold, and Reynolds, 2000). For example, in the telecom industry, the annual attrition rate is about 30 percent for wireless service; nearly half of all Internet subscribers leave their providers every year; 50 percent of heavy users ($50 or more per month) of long distance calls leave their carrier within a year (Institute for International Research, 1998). Research has shown that retaining existing customers is more profitable than acquiring new customers due primarily to savings on acquisition costs, the higher volume of service consumption, and customer referrals (Jacob, 1994; Walker, Boyd, and Larreche, 1999: 283). The importance of customer retention has been increasingly recognized by both marketing managers as well as research analysts (Jacob 1994; Li, 1994 and 1995; Keaveney, 1995; Walker, Boyd, and Larreche, 1999: 120-122 and 282-284; Ganesh, Arnold, and Reynolds, 2000). Keaveney (1995) examines customer switching behavior in service industries. She focuses on the quality of service and identifies eight main variables that may cause customer switching: price, inconvenience, core service failures, service encounter failures, failed employee responses to service failures, competitive issues, ethical problems, and involuntary factors. A limitation of Keaveny's study is that she does not examine the characteristics of customers who have switched. Ganesh, Arnold, and Reynolds (2000) examine the differences between switchers and stayers and conclude that customers who have switched services providers because of dissatisfaction are significantly different from stayers in their satisfaction and loyalty behaviors. These studies have contributed to our understanding of switching behavior. Understanding why customers leave is the first step in building an effective customer retention program. A second step is to identify the customers with high risk of leaving, which is the task of predicting customer attrition. Predicting customer attrition with high accuracy is vital for customer retention. In addition, a reliable prediction of changes in the customer population will improve business planning and resource allocation efficiency.

Predicting customer attrition is a challenging work due to the large quality of data and the difficulty of specifying the right statistical model. Customer leaving is not caused by a single reason; usually there are multiple reasons: the customer may no longer need a service, he/she may migrate to another type of service, or he/she may switch to a competitor for the same service. Each type of leaving indicates a unique situation. Furthermore, when a customer leaves, we often do not know which reason applies. Thus predicting customer leaving by any single cause is inappropriate and total customer attrition is not an additive sum of the attrition of each cause. From the perspective of customer behavior, switching to a competitor and canceling the service are different behaviors, so combining them together as customer attribution increases the heterogeneity

of the predicted variable. All these concerns give rise to the question of how to select the most efficient model to predict customer attrition. When the data set is large, the application of models and comparisons are cumbersome.

The development of techniques in data mining and knowledge discovery (Peacock, 1998a and 1998b) has greatly enhanced our ability to develop and compare models predicting customer attrition with nonlinear combinations of causes and large data sets. In this study, we develop a systematic way of dealing with the non-linearity issue associated with customer attrition and the issue of large data sets. We demonstrate the use of different data mining methods to develop predicting models of customer attrition and compare their predicting power, using data from a major telecom service company.

FITTING AND COMPARING PREDICTION MODELS

Fitting Models

The idea of using statistical models to predict customer attrition is not new (see, e.g., Ma and Li, 1993 and 1994). In general, when fitting customer attrition models, we study a data set that contains customer service duration and time of service status change, and customer/service characteristics. We then identify the association between customer attrition and customer- and service-related characteristics, such as duration and service arrangement:

Customer attrition = f (customer- and service-related characteristics).

Customer attrition is a combination of cancellation and switching to a competitor. When we cannot separate the two causes, we combine them into a single measure of attrition in our model. To evaluate the prediction power of a model, we build a model with two independent samples. We first develop a model using a "learning" (or "train") sample. We then validate the model by applying it to a "validation" (or "test") sample to determine the extent to which the model may be generalized beyond the original "learning" sample. This is a standard procedure for fitting scoring models (models that estimate the probability (score) of an event (such as attribution)) using data mining techniques.

Comparing Modeling Methods

There are several data mining methods that may be used to construct models to estimate customer attrition, such as the logistic method, the Cox regression method, the tree-based classification method, and, more recently, the artificial neural network method. Each may be more suitable for a particular application. Thus a critical issue is how to efficiently assess the performance of a model developed by one method relative to models developed by other competing methods. Such a comparison is especially important for attrition risk prediction, because, as mentioned earlier, customers may terminate a service for multiple reasons (e.g., service cancellation or switching to a competitor's service), and the combination of different reasons is not linear. Although compar-

isons of modeling methods are straightforward in theory, in practice they are very cumbersome, due to the magnitude of the observations and the variables.

Recent developments in data mining software, such as the SAS Enterprise Miner (SAS Institute, 1998), provide efficient tools to perform cross-validations of different prediction models. In this exercise, we set out to compare three different modeling methods that are most often used to predict customer attrition: (1) the regression-based method; (2) the tree-based method, and (3) the artificial neural network method.

The regression-based method. This is by far the most popular method to build models to predict customer leaving. It regresses the outcome of the variable of interest (such as attrition) on a number of variables that may co-vary with (predict) the outcome variable. It defines the probability of an outcome by the magnitude of the score obtained by adding or subtracting the coefficients assigned to the predicting variables.

The tree-based method. The tree-based method employs the technique of recursive partitioning of data with respect to the variable of interest. In the case of predicting customer attrition, it identifies subgroups of customers who are relatively homogeneous with respect to the risk of service termination. Unlike the regression-based method, recursive partitioning identifies subject subgroups based on Boolean combinations of variables. A branching, algorithm-like "tree" is created, with the "trunk" (the entire sample) or major branches split into two or more smaller branches based on the value of the single variable that minimizes a measure of within-group heterogeneity. The tree terminates in two or more "nodes," each of which defines a subgroup of relatively similar subjects with respect to the outcome of interest.

The artificial neural network method. This method automatically identifies patterns in data by a computer procedure. The artificial neural network method distinguishes two types of self-learning processes: *supervised* and *unsupervised* learning. Unsupervised learning is used to identify patterns in data, such as clustering customers based on certain criterion variables. In supervised learning, the goal is to predict one or more target variables from one or more input variables. Supervised learning is usually some form of nonlinear regression or discriminant analysis. The artificial neural network method has been increasingly used in business modeling as a powerful tool to examine large data sets with many variables. In our study, we use the *multilayer perceptrons procedures* (MLPs) in the SAS Enterprise Miner to develop the model. MLPs are general-purpose, flexible, nonlinear models (Sarle, 1994).

How to Compare Models: The Use of the ROC Curve

In predicting service termination, we assume that the subjects are in one of the two basic statuses: *at risk* or *not at risk*. Because of the diverse activity and speculation (i.e., there is no clear-cut distinction between the subjects at

risk and the subjects not at risk) and because of the limited information associated with the outcome, the subject at risk and the subject not at risk may not always be predicted correctly. The relative frequency of a possible incorrect prediction among the subjects not at risk (p= 1- specificity) and the relative frequency of a possible correct predictions among the subjects at risk (q=sensitivity) will depend on what decision threshold is adopted. The performance of a predicting model may vary depending on a specific threshold used. Thus, an objective evaluation of a risk-predicting model should examine the overall performance of the model under *all* possible decision thresholds, not only one particular decision threshold. To achieve this, we adopt a useful tool, the receiver-operating characteristic (ROC) curve, to evaluate the performance of a risk predictive model. This curve is the locus of the relative frequencies of p and q, which occupy different points on the curve corresponding to different decision thresholds. The area under this curve (on a unit square) is equal to the probability that the predictive model will correctly distinguish the "at risk" subjects from the "not at risk" subjects (Hanley and McNeil, 1982; Ma and Hall, 1993). The construction and analysis of ROC curves are built in the SAS Enterprise Miner for predictive model evaluations, which enable us to easily compare the predictive models we build using the three data mining methods.

COMPARISON DESIGN

Data Source

We choose to study a data set from a major telecom service company, which includes a segment of customers with a large number of service lines (in the order of millions) that are active at the beginning of the year. The unit of analysis is the individual subscriber line. For each active service line, information on service termination (yes/no) was collected over the next 3-month period (the first quarter of the chosen year). Customer usage and characteristics were collected over the 6-year period backward from the chosen year. This information includes type of service, service usage, marketing information, customer demographics, and service line transaction history.

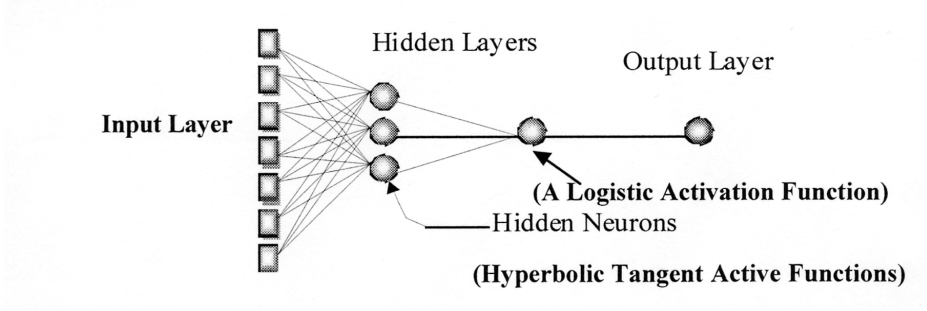
As in most event history analyses, the event (line termination) represents only a small percentage of the population (less than 5 percent in our case). The low rate of termination implies that if we draw a random sample to study the event, we need a very large sample to ensure the precision of the model estimation (or we would have to observe a long period of time to collect more cases of termination). To overcome this problem, we adopt a case-based sample (Prentice, 1986) in which we (1) include all the termination cases, and (2) mix these termination cases with a simple random sample of about 113,000 service lines. The sample size is approximately 182,000 with 38 percent termination cases. We then randomly split these observations into two samples of 50 percent each for learning (train) and validation (test).

Model Fitting and Testing

The SAS Enterprise Miner, a data-mining software, is used to develop and compare predicting models based on different estimating methods. We create and select 20 variables based on business experiences and our exploratory study. The SAS Enterprise Miner provides two variable selection methods: a regression-based method and a tree-based method. The regression-based method gauges the importance of input variables based on an R-square statistic. The tree-based method measures the importance of input variables based on a Chi-square statistic. We eliminate 10 variables where neither the R-square statistic nor the Chi-square statistic is significant at the 10 percent level.[2] We then use the three data mining methods (logistic regression, tree, and artificial neural network) to estimate models using the remaining variables.

The model based on logistic regression is estimated by using a backward stepwise procedure with a level of significance at 10 percent. The model that employs tree-based method is developed by using CART with a minimum number of observations in a node equal to 20 and a significant level for the Chi-square statistic equal to 10 percent. As for the model that uses artificial neural network method, after trying several MLP neural network estimations, (from one to seven hidden layers, and from three to ten hidden neurons), we select a model with one hidden layer and three hidden neurons (see Figure 1) because of its simplicity and because the other MLP models are not significantly better.

Figure 1. The Multilayer Perceptron (MLP) Model

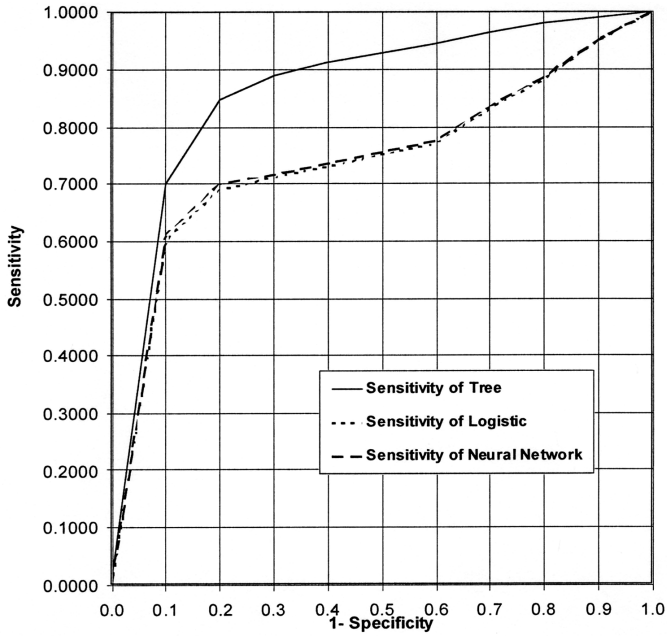


Given the large dataset we have (a total of 182,000 observations) and the random split of data into learning (train) and validation (test) samples, we expect the results of model estimation and validation to be highly consistent. However, in order to further insure that our results are consistent, we repeat our model estimate five times with new split of data each time. As can be seen in Table 2, the results are highly consistent.

RESULTS

The SAS Enterprise Miner automatically calculates the optimal decision threshold for each model based on the learning data set. The optimal decision threshold is the threshold that maximizes the number of correct classifications among the terminated service lines and minimizes the number of incorrect classifications among the active service lines. We calculate a total predictive accuracy (i.e., correct classification rate) for each model in the validation data set by using the model's optimal decision threshold. The total predictive accuracies for the logistic model, the tree model, and the MLP model (averaging from the results of five random data splits) are 79.9 percent, 83.1 percent, and 80.1 percent, respectively. The ROC curve for each model is generated on the test (validation) data set by varying the decision threshold. The ROC curve of the tree model is superior to the ROC curves of both the logistic model and the neural network model (see Figure 2).

Figure 2. ROC Curves of the Three Models



Variables Selected

Of the ten variables initially included in the models, the logistic regression method selects seven variables and rejects three; the tree method selects five variables and rejects the other five. All five variables selected by the tree method are a subset of the seven variables selected by the logistic method. The

MLP method uses all ten variables for its input/output calculations (see Table 1 for all the variables selected by each model).

Table 1. Predicting Variables Selected by Each Model

Variables Selected by all models (tree, logistic, and MLP):

Account Type (VIP/Premier/Special/Missing)
Customer Segment (Transnational/Core/Local/Other)
Location (East/Central/South/West/Other)
Last 6-Month Usage (0 – 3,000 [hours])
Indicator of being active since 1993 (Yes/No)

Additional Variables Selected by logistic and MLP models:

Current Tenure (0 – 72 [months])
Changes in Quarterly Usage (%)

Additiona Variables Selected by MLP model:

Advanced Features (Yes/No/Missing)
Type of Service (Customized/Standard/Other)
Number of previous terminations by the line

Variables Rejected by all models:

Customer sales volume
Customer size (number of employees)
Customer 4-digit SIC (standard industry classification) code
Line level usage distribution by weekday
Line level usage distribution by hour of day
Number of lines a customer has
Customer line internal management entity
Dual user (using two service providers)
Line level usage strata created based on monthly average
Last month's usage

Performance Characteristics

Table 2 shows various performance characteristics of the models, for both the learning (train) and the validation (test) samples, from estimations using five random splits of the data. The model developed by using the tree method consistently performs better than the other two models in both the learning and the validation samples. The prediction power of the model developed by logistic regression is similar to that of the model based on MLP neural network method.

Table 2. Performance Characteristics of the Models

Models	Learning			Validation		
	Logistic	Tree	MLP	Logistic	Tree	MLP
Samples						
A	79.5	84.3	80.0	80.9	83.4	80.9
B	79.4	83.6	79.6	79.5	82.8	79.5
C	79.4	84.0	79.7	79.5	83.5	80.1
D	79.2	83.1	79.5	79.7	82.3	80.2
E	79.1	83.4	79.4	79.8	83.4	79.9
Mean	79.3	83.7	79.6	79.9	83.1	80.1

As mentioned earlier, we construct ROC curves (generated by the SAS Enterprise Miner along the learning and validation processes) for the three models based on the validation sample (Figure 2). In terms of prediction accuracy, the tree model is uniformly (across all decision thresholds) superior to the logistic model and the MLP model. Areas under the ROC curves for the logistic model, the tree model, and the MLP model are 74.1 percent, 86.5 percent, and 74.6 percent, respectively. A nonparametric comparison of the areas under the three ROC curves shows that the tree model is significantly better than the logistic model and the MLP model (P-value < 0.0001). (For the test statistics, see DeLong, Delong, and Clarke-Pearson, 1988 and Hanley and McNeil, 1983.)

Predictions at the Level of Individual Lines

Table 3 compares the three models' agreement for predicting termination at the level of individual lines, that is, the chances that the three models produce the same outcome for an individual. The agreement of these models is measured by using a statistic called a κ-statistic (kappa) which measures the agreement between model predictions beyond expected chance (Fleiss, 1981). The values of the κ-statistic greater than 0.75 may be interpreted as having excellent agreement beyond chance. Thus the three models presented here are in excellent agreement at individual level predictions. The closest agreement is found between the logistic and the MLP models.

Table 3. Inter-Model Agreements on Individual Line Prediction

Models	κ-Statistic	Standard Error of κ
Tree vs. Logistic	0.8496	0.003312
Tree vs. MLP	0.8499	0.003312
Logistic vs. MLP	0.9251	0.003314

Evaluating Model Fit

The tree model performs best overall in both the learning and the validation samples. Comparing the tree model with the logistic model and the MLP model, we find that the difference is overwhelming, with about 1,800 to 3,600 more service lines (about 2 to 4 percent more of the learning or the validation sample) being correctly predicted by the tree model than those by the other two less-accurate models.

With regard to the individual level predictions, the logistic model and the MLP model provide comparable accuracy. These two models achieve similar sensitivity and specificity for all possible decision thresholds.

In this particular application, namely, the prediction of customer attrition, the tree-based method is the best among the three methods. The following features of the tree-based method may contribute to its superiority. First, it is a nonparametic method, which is less sensitive to data distribution. Second, in our application the tree-based method uses the fewest number of predicting variables, which may make the model more stable and less affected by the missing data problem, which refers to the situation where values of some variables are missing for a subset of observations (in our case the individual line subscribers). This is a common problem in marketing research that has propelled analysts to look for the most robust model that can overcome this impediment. The fewer the number of variables a model uses, the less the missing data problem is, and therefore the more stable and robust the model is.

A third feature relates to how the missing data are handled. Both the logistic regression method and the MLP method can only use cases with no missing data. This limitation causes a loss of information due to discarding observations with missing data and an inability to generate predictions for new observations with missing data. Unlike the logistic regression method or the MLP method, the tree method makes use of all the observations, even observations with missing data. This advantage may substantially contribute to its superiority over the other two methods.

However, the result whereby the model developed by the tree-based method is the best is specific to our task, namely, predicting customer attrition. In other tasks, another method may be better (Little and Rubin, 1987), which may be evaluated by using the steps we outlined in this article.

Lastly, as we mentioned earlier, the attritions in our data include multiple causes such as switching and cancellation. In this particular data set, the causes cannot be identified and we do not know the distribution of the combination of different causes. In this regard, the non-parametric nature of the tree model may lend it extra robustness to fit the data.

CONCLUDING REMARKS: MANAGERIAL IMPLICATIONS

In this study we demonstrate the general steps of applying data mining techniques to fit and evaluate customer attrition prediction models. The main goal of this exercise is to help marketing managers to achieve efficiency in business planning and resource allocation. Data mining techniques are becoming extremely useful due to the advances in computing power and the availability of large quantity of data. Managers who plan to use data mining to enhance their marketing effort should pay attention to the following points.

Theory-driven vs. data-driven statistical analyses. Social scientists, including business scholars, emphasize the importance of theory and in general disapprove data-driven statistical analyses. They tend to select predicting variables based on theoretical hypotheses about the relationship between causal concepts. On the other hand, marketing managers use statistical analyses primarily for the purpose of enhancing their business objectives such as profitability. In the case of retention, marketing managers are looking for variables that can effectively and efficiently distinguish stayers from non-stayers. Whether these variables make theoretical sense is a secondary concern at most. In the pre-data mining era, our ability to handle large quantity of data and test multiple models were limited. Using theory-driven statistical models would minimize the number of variables and observations we needed to handle. Theories would also help us to reduce the possibility of model misspecification, thus reducing the number of models we had to try.

However, the availability of data mining techniques has substantially reduced the above concerns. Data mining allows us to handle very large data sets with millions of observations and thousands of variables. We can use data mining to fit the data with many different models and evaluate their goodness of fit efficiently, making the concern for model misspecification virtually irrelevant. Please note here that we are not nullifying the importance of theories. For example, theoretical explorations on customer switching behaviors help us understand why customers switch (e.g., Keaveney, 1995), which serve as a foundation to develop marketing programs to retain customers and improve loyalty.

The roles of marketing research analysts and marketing managers. The introduction of data mining in marketing has redefined the roles of marketing research analysts and marketing managers. Data mining has substantially reduced the workload of marketing research analysts. Many tasks that previously performed by marketing analyst, such as data preparation and statistical analyses, are now done by using data-mining tools. The implication of this change is that first, companies may not need as many marketing research analysts as they used to; second, marketing managers with adequate data mining training may be able to perform many statistical analyses previously requiring

complicated statistical programming. The line between marketing research analysts and marketing managers is becoming more blurred due to the ease of using data mining tools. However, we would like to caution companies that using data mining without a good understanding of the statistical principles is dangerous. Companies should encourage managers to learn data mining techniques and at the same time should keep well-trained marketing research analysts on their staff, which may be fewer due to the reduced workload, to provide guidance on how to use the new data mining tools.

REFERENCES

DeLong, E.R., DeLong, D.M. and Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated ROC curves: A nonparametric approach. *Biometrics*, 44, 837-845.

Fleiss, J.L. (1981). *Statistical methods for rates and proportions*, 2nd ed. New York: John Wiley & Sons.

Ganesh, J., Arnold, M., and Reynolds, K. (2000). Understanding the customer base of service providers: An examination of the differences between switchers and stayers. *Journal of Marketin*g, 64 (July), 65-87.

Hanley, J.A. and McNeil, B.J. (1982). The meaning and use of the area under a ROC curve. *Radiology*, 143, 29-36.

——. (1983). A method of comparing the areas under ROC curves derived from the same cases. *Radiology*, 148, 839-843.

Institute for International Research (1998). Preventing churn in telecommunications. July 15-16, 1998, Washington, DC.

Jacob, R. (1994). Why some customers are more equal than others. *Fortune*, Sept. 19, 200.

Keaveney, S. (1995). Customer switching behavior in service industries: An exploratory study. *Journal of Marketing*, 59 (April), 71-82.

Li, S. (1994). "Applications of demographic techniques in modeling customer retention. In K. V. Rao and J. W. Wicks (eds.), *Applied demography*. Bowling Green, Ohio: Bowling Green State University, 183-197.

—— (1995). Survival analysis. *Marketing Researc*h, Fall, 17-23.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley & Sons.

Ma, G. and Hall, W.J. (1993). Confidence bands for ROC curves. *Medical Decision Making*, 13, 191-197.

Ma, G. and Li, S. (1993 and 1994). *Applications of the survival analysis techniques in modeling customer retention*. Workbook for the 4th and 5th Advanced Research Techniques Forums, American Marketing Association.

Peacock, P.R. (1998a). Data mining in marketing: Part 1. *Marketing Management*, Winter, 9-18.

—— (1998b). Data mining in marketing: Part 2. *Marketing Management*, Spring, 15-25.

Prentice, R.L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73(1), 1-11.

Sarle, W.S. (1994). Neural networks and statistical models. *Proceedings of the Nineteenth Annual SAS Users Group International Conference*, Cary, NC: SAS Institute, Inc.

SAS Institute (1988). SAS/STAT user's guide. Cary, NC: SAS Institute Inc.

SAS Institute (1998). *Getting started with Enterprise Miner*. Cary, NC: SAS Institute Inc.

Walker, O. C., Boyd, H. W., and Larreche, J. C. (1999). *Marketing Strategy: Planning and Implementation*, 3rd Ed. Boston, MA: Irwin.

NOTES

1. In service industries, customer attrition is covered by the broader term "churn," which can be either positive (gaining customers) or negative (losing customers). There are a number of terms to refer to negative churn: attrition, termination, defection, or leaving. In our study, we primarily use attrition, and only occasionally use the other terms, to describe negative churn.
2. We realize that if we want to guard against including any variables that do not contribute to the predicting power, we should lower the significance level. However, our objective is to increase the predicting power as much as we can. According to the SAS manual, "if you want to choose the model that provides the best prediction using the sample estimates, you need only guard against estimating more parameters than can be reliably estimated with the given sample size, so you should use a moderate significance level, perhaps in the range of 10 percent to 25 percent" (SAS, 1988: 820).