Bagging and boosting classification trees to predict churn

Lemmens, A.; Croux, C.

Published in:
Journal of Marketing Research

Publication date:
2006

Citation for published version (APA):
Lemmens, A., & Croux, C. (2006). Bagging and boosting classification trees to predict churn. *Journal of Marketing Research*, *43*(2), 276-286.

AURÉLIE LEMMENS and CHRISTOPHE CROUX*

In this article, the authors explore the bagging and boosting classification techniques. They apply the two techniques to a customer database of an anonymous U.S. wireless telecommunications company, and both significantly improve accuracy in predicting churn. This higher predictive performance could ultimately lead to incremental profits for companies that use these methods. Furthermore, the results recommend the use of a balanced sampling scheme when predicting a rare event from large data sets, but this requires an appropriate bias correction.

Bagging and Boosting Classification Trees to Predict Churn

Classification issues are common in marketing literature. One of the most frequent topics envisioned as a classification task is consumer choice modeling (see, e.g., Chung and Rao 2003; Corstjens and Gautschi 1983; Currim, Meyer, and Le 1988; Guadagni and Little 1983; Kalwani, Meyer, and Morrison 1994). The current study considers a binary choice problem, namely, the prediction of customer churn behavior.

Several classification models exist, but one of the most popular is the (binary) logit model, which has been used extensively in marketing to solve binary or multiple choice problems (see, e.g., Andrews, Ainslie, and Currim 2002). More sophisticated models, which take into account the heterogeneity in consumer response, include finite mixture models (see, e.g., Andrews and Currim 2002; Wedel and Kamakura 2000) and hierarchical Bayes techniques (see, e.g., Arora, Allenby, and Ginter 1998; Yang and Allenby 2003). For binary choice problems, these approaches require the availability of panel data (i.e., data from several observations over time on multiple customers). However, in many applications (including the current one), a customer is observed only once over time, which makes it impossible to disentangle the individual effects from the random errors (Donkers et al. 2006).

In this article, we use the "bagging" and "boosting" classification models that originated in the statistical machine-learning literature. Bagging (Breiman 1996) consists of sequentially estimating a binary choice model—called a "base classifier" in machine learning—from resampled versions of a given calibration sample. The obtained classifiers form a committee from which a final choice model can be derived by simple aggregation. Although bagging is simple and easy to use, more sophisticated variants also exist. "Stochastic gradient boosting" (Friedman 2002) is one of the latest developments and includes weights in the resampling procedure.

Although bagging and boosting have received increasing attention in various fields (e.g., for the University of California, Irvine, machine-learning archive, see Friedman, Hastie, and Tibshirani 2000; for text categorization, see Nardiello, Sebastiani, and Sperduti 2003; for use in chemometrics, see Varmuza, He, and Fang 2003; for an application in fraud claim detection, see Viaene, Derrig, and Dedene 2002), to the best of our knowledge, marketing literature does not contain any reference to such models. Therefore, we attempt to fill this gap by empirically investigating whether bagging and stochastic gradient boosting can challenge more traditional choice models. In particular, we examine their performance in predicting customers' churn behavior for an anonymous U.S. wireless telecommunications company.[1] To evaluate the predictive accuracy of our churn model, we consider not only the misclassification rate, which may be misleading for rare events, such as churn, but also the Gini coefficient and the top-decile lift.

Churn is a marketing-related term that characterizes whether a current customer decides to take his or her busi-

*Aurélie Lemmens is a doctoral candidate (e-mail: aurelie.lemmens@econ.kuleuven.be), and Christophe Croux is a professor (e-mail: christophe.croux@econ.kuleuven.be), Department of Applied Economics, K.U. Leuven. The authors are grateful to Marnik Dekimpe for his valuable and helpful comments and to the Teradata Center for Customer Relationship Management at Duke University for the data and remarks. They also wish to thank the two anonymous *JMR* reviewers for their constructive comments. This research was funded by the Research Fund at K.U. Leuven and the Fonds voor Wetenschappelijk Onderzoek (Contract No. G.0385.03).

[1] This database was provided by the Teradata Center for Customer Relationship Management at Duke University during the Duke/NCR Churn Modeling Tournament.

ISSN: 0022-2437 (print), 1547-7193 (electronic)

ness elsewhere (in the current context, to defect from one mobile service provider to another). As with many other sectors (e.g., the newspaper business), churn is an important issue for both the U.S. and the European wireless telecommunications industry. Monthly churn rates amount to approximately 2.6% (Hawley 2003) as a result of increased competition, lack of differentiation, and saturation of the market. Because the cost of replacement of a lost wireless customer amounts to $300–$700 (depending on the source of information; see, e.g., Snel 2000) in terms of sales support, marketing, advertising, and commissions, churn may have damaging consequences for the financial wealth of companies. However, predicting churn enables the elaboration of targeted retention strategies to limit these losses (Bolton, Kannan, and Bramlett 2000; Ganesh, Arnold, and Reynolds 2000; Shaffer and Zhang 2002). For example, specific incentives may be offered to the most risky customer segments (i.e., the most inclined to leave the company) with the hope that they remain loyal. Other scientific studies also note the advantage of customer retention as a lower-cost operation than attracting new customers (Athanassopoulus 2000; Bhattacharya 1998; Colgate and Danaher 2000).

Despite the financial consequences that a 2% monthly churn rate may lead to, customer defection is still a statistically rare event. Consequently, when the churn predictive model is estimated on a random sample of the customer population, the vast majority of nonchurners in this proportional calibration sample (i.e., the number of churners in this randomly drawn sample is proportional to the real-life churn proportion) dominate the statistical analysis, which may hinder the detection of churn drivers and eventually decrease the predictive accuracy. To address this issue, the calibration sample size can be increased. However, this solution is usually not optimal (see the "Results" section; King and Zeng 2001a). A better solution to this issue is to apply a selective sampling scheme to increase the number of churners in the calibration sample. Such a sampling scheme is called "balanced sampling" (or "stratified sampling" in King and Zeng 2001a, b). Theoretically, a potentially better-performing classifier could be obtained from such a sample, especially for small sample sizes (see, e.g., Donkers, Franses, and Verhoef 2003; King and Zeng 2001a, b). We investigate whether these findings are still valid for large sample sizes.

The estimation of a classification model from a balanced sample usually overestimates the number of churners in real life. Several methods exist to correct this bias (see, e.g., Cosslett 1993; Donkers, Franses, and Verhoef 2003; Franses and Paap 2001, pp. 73–75; Imbens and Lancaster 1996; King and Zeng 2001a, b; Scott and Wild 1997). However, most of these corrections are dedicated to traditional classification methods, such as the binary logit model. Therefore, we subsequently discuss two easy correction methods for bagging and boosting, from which marketers can benefit to predict churn.

In summary, we investigate the following research questions: Do the recent developments in statistical machine learning outperform the traditional binary logit model in predicting churn? If so, what financial gains can be expected from this improvement, and what are the more relevant churn drivers, or "triggers," for which marketers can watch? Moreover, we propose two bias correction methods for balanced samples and compare their performance. Finally, using large sample configurations, we investigate whether a choice model estimated on a balanced sample, with the bias appropriately corrected for, outperforms a choice model estimated on a proportional sample.

We organize the remainder of this article as follows: The next section contains a description of the data. The three subsequent sections outline the bagging and boosting models, the bias correction methods for balanced sampling schemes, and the assessment criteria, respectively. We then empirically answer the aforementioned research questions and offer some conclusions.

DATA

We used a data set that the Teradata Center at Duke University provided. This database contains three data sets of mature subscribers (i.e., customers who had been with the company for at least six months) to a major U.S. wireless telecommunications carrier. The variable we attempt to predict is whether a subscriber churns during the period of 31–60 days after the sampling date (we know that the actual reported average monthly churn rate is approximately 1.8%). A delay of one month in measuring the churn variable is justified because the implementation of proactive customer retention incentives requires some time. In this case, marketers would have a one-month delay to target and retain customers before they churn. We coded the churn response as a dummy variable, where y = 1 if the customer churns and –1 if otherwise.

We used the first two data sets as calibration samples of 51,306 observations each.[2] The first data set is a "proportional calibration sample" (the proportion of churners in the sample is approximately 1.8%), and the second contains an "oversampled" number of churners such that the number of churners is perfectly balanced by the number of nonchurners. Selected at a future point in time, the third data set contains 100,462 customers, 1.8% of whom are churners. We used this third set as a validation (thus, we do not use it in our estimation) holdout sample to evaluate the performance of the prediction rules constructed from one of the aforementioned calibration samples. All samples contain a different set of customers.

To predict customers' churn potential, U.S. wireless operators usually take into account between 50 and 300 subscriber variables as explicative factors (Hawley 2003). From the high number of explicative variables contained in the initial database (171 variables), we retained 46 variables, including 31 continuous and 15 categorical variables. The retained predictors include behavioral (e.g., the average monthly minutes of use over the previous three months, the total revenue of a customer account, the base cost of a calling plan), company interaction (e.g., mean unrounded minutes of customer-care calls), and customer demographic (e.g., the number of adults in the household, the education level of the customer) variables (for an overview, see Table 1). We selected the variables by excluding all variables that contained more than 30% of missing values. Among the remaining variables, we selected those with the most poten-

[2] Originally, the second data set contained 100,000 observations, but we reduced its size (by taking a random subset from it) to ensure a fair comparison between both calibration samples.

Table 1
NONEXHAUSTIVE OVERVIEW OF THE CHURN PREDICTORS

Behavioral Predictors	*Company Interaction Predictors*	*Customer Demographics*
Billing adjusted total revenue over the life of the customer ("total revenue over life")	Having responded to an offer in the mail (yes/no)	Age of the first household member ("age")
Mean number of attempted calls placed ("mean attempted calls")	Mean minutes of use of customer care calls	Estimated income
Percentage change in monthly minutes of use versus previous three-month average ("change in monthly minutes of use")		Social group
Mean total monthly recurring charge ("base cost of the calling plan")		Marital status
Average monthly minutes of use over the previous six months ("average monthly minutes of use [six months]")		Geographic area
Mean number of completed calls ("mean completed calls")		Account spending limit
Mean number of peak calls ("mean peak calls")		Children in the household (yes/no)
Total number of months in service ("months in service")		Dwelling unit type
Mean number of inbound calls less than one minute ("mean inbound calls less one minute")		Number of days of current equipment ("Equipment days")
Mean of overage revenue ("mean overage revenue")		Refurbished or new handset
Mean number of monthly minutes of use ("mean monthly minutes of use")		Current handset price ("handset price")
Mean unrounded minutes of use of outbound wireless to wireless calls ("mean monthly minutes wireless to wireless")		

tial relevance, following the results of a principal components analysis.[3] Note that for an equal comparison, we consider the exact same set of variables for all investigated models.

The handling of missing values is operated differently for the continuous and the categorical predictors. For the continuous variables, we imputed the missing values by the mean of the nonmissing ones. Because not answering a question may be as informative as a specific response, for each observation, we added an extra predictor that indicated whether there was at least one imputation. For categorical predictors, we created an extra level for each of them that indicated whether the value was missing.

THE BAGGING AND BOOSTING MODELS

Both bagging and boosting originate from the machine-learning research community and are based on the principle of "classifier aggregation." This idea was inspired by Breiman (1996), who found gains in accuracy by combining several base classifiers, sequentially estimated from perturbed versions of the calibration sample. Among the several possible alternatives of base classifiers, classification trees (also known as CART; see Breiman et al. 1984) are a sensible choice (Breiman 1996). Their use is not widespread in marketing literature (for exceptions, see Baines et al. 2003; Currim, Meyer, and Le 1988; Haughton and Oulabi 1997), though they are powerful nonparametric methods. In recent years, statistical theory has been elaborated to provide a theoretical background for these techniques (e.g., for bagging, see Bühlmann and Yu 2002; for boosting, see Friedman, Hastie, and Tibshirani 2000; for a comprehensive review, see Hastie, Tibshirani, and Friedman 2001).

For the sake of conciseness, the following subsection contains a brief description of the bagging algorithm. In the next subsection, we provide further details about the main differences between bagging and stochastic gradient boosting, one of the most sophisticated versions of boosting to date (for an in-depth description of this method, see Friedman 2002).

Bagging

Bagging (i.e., a term derived from "bootstrap aggregating") is the simplest technique to upgrade, or to "boost," the performance of a given choice model. We denote the calibration sample as $Z = ([x_1, y_1], \ldots, [x_i, y_i], \ldots., [x_N, y_N])$, where N is the number of customers in the calibration sample. In this expression, $x_i = (x_{i1}, \ldots, x_{ik}, \ldots, x_{iK})$ represents a vector that contains the K predictors for customer i, and y_i (equal to 1 or –1) indicates whether customer i will churn. We estimate a base classifier $\hat{f}$ from this calibration sample, giving a score value of $\hat{f}(x)$ to each customer, where x is the characteristics of this subscriber. This score value indicates the risk to churn associated with each customer. For a specified cutoff value τ, we can predict customers as churners or nonchurners by computing

$$\hat{c}(x) = \text{sign}[\hat{f}(x) - \tau], \tag{1}$$

which takes values of +1 or –1. If $\hat{f}(x_i)$ is larger than τ, customer i is classified as a churner, but if $f(x_i)$ is smaller than

[3] Because the purpose of this article is to investigate the comparative performance of different models, we do not provide further details about variable selection, which mainly served to reduce computation time. Some experiments indicated that the performance of the classification rules barely changed, regardless of whether we implemented a variable selection procedure.

τ, customer i is classified as a nonchurner. When we use a classification tree as base classifier, the score is given by $\hat{f}(x) = 2\hat{p}(x) - 1$, where $\hat{p}(x)$ is the probability to churn as estimated by the tree. When working with a proportional calibration sample, we set $\tau = 0$. In the presence of a nonproportional calibration sample, the value of τ varies (see the "Correction for a Balanced Sampling Scheme" section).

From the original calibration set Z, we construct B bootstrap samples Z_b^*, b = 1, 2, …, B, by randomly drawing, with replacement, N observations from Z. Note that the size of the bootstrap samples equals the original calibration sample size. From each bootstrap sample Z_b^*, we estimate a base classifier, giving B score functions $\hat{f}_1^*(x)$, ..., $\hat{f}_b^*(x)$, ..., $\hat{f}_B^*(x)$. We aggregate these functions into the final score

$$(2) \qquad \hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}_b^*(x).$$

We can then carry out the classification using the following equation:

$$(3) \qquad \hat{c}_{bag}(x) = \text{sign}[\hat{f}_{bag}(x) - \tau_B], \text{ where } \hat{c}_{bag}(x) \in \{-1, 1\}.$$

Again, the cutoff value τ_B equals zero in the presence of a proportional calibration sample. To determine the optimal value of B (i.e., the number of bootstrap samples), a strategy consists of selecting B such that the apparent error rates (i.e., error rates on the calibration data) remain more or less constant for values larger than B. In our application, we set B = 100.[4]

As with traditional classification models, we can also obtain diagnostic measures for the estimated bagging model. These are important to give some face validity to the estimated model. For example, the estimated relative importance of each predictor in the construction of the classification rule can be investigated. For a single tree, the relative importance of a predictor can be computed, as Hastie, Tibshirani, and Friedman (2001) do.[5] For bagging (and, similarly, for boosting), the relative importance of an explicative variable is averaged over all B trees. In addition, the partial dependence of churn on a specified predictor variable can be investigated. This measure provides similar insight to the parameter estimates' values of a logit model, but it advantageously allows for nonlinear relationships between the predictors and the dependent variable. A partial dependence plot represents the impact of a predictor variable on the churn probability of a customer, conditional on all other predictors. In practice, the partial dependence of the dependent variable on a specified value of a predictor x_k is obtained by assigning the value of x_k to all observations of the calibration sample. The model is subsequently estimated, and the N resulting predicted probabilities are computed for the calibration data. The partial dependence on a specified value of x_k is eventually given by averaging over these N predicted probabilities. The partial dependence plot is obtained by letting the value assigned to x_k vary over a large range of values (for more details, see Friedman 2001).

Boosting and Stochastic Gradient Boosting

Several versions of boosting exist: the Real AdaBoost (Freund and Schapire 1996; Schapire and Singer 1999), LogitBoost (Friedman, Hastie, and Tibshirani 2000), and gradient boosting (Friedman 2001). Boosting is more complex than bagging and not as easy to put into practice. In this article, we focus on stochastic gradient boosting (Friedman 2002), one of the most recent boosting variants and the winning model of the Teradata Churn Modeling Tournament (Cardell, Golovnya, and Steinberg 2003).

The main difference between boosting and the previously described bagging procedure lies in the sampling scheme. Boosting consists of sequentially estimating a classifier to "adaptively reweighted" versions of the initial calibration sample Z_b^*, b = 1, 2, …, B. The adaptive reweighting scheme enables us to give previously misclassified customers an increased weight on the next iteration, whereas weights given to observations that were correctly classified previously are reduced. The idea is to force the classification procedure to concentrate on the customers that are difficult to classify.

Another main difference with bagging is that the initial choice model should preferably be "weak" (i.e., with a slightly lower associated error rate than random guessing). For stochastic gradient boosting, Friedman (2002) recommends the use of k-node trees as a base classifier, where k is approximately 6–9, depending on the issue. In addition, the number of required iterations is usually higher for stochastic gradient boosting than for bagging. In our application, we select B = 1000.

CORRECTION FOR A BALANCED SAMPLING SCHEME

Predictions made from a model estimated on a balanced calibration sample are known to be biased because they overestimate the proportion of churners in real life. Although appropriate bias correction methods already exist for some common classifiers (for the logit model, see, e.g., King and Zeng 2001b), to the best of our knowledge, no correction method for bagging and boosting currently exists. Hereinafter, we adapt to the bagging and boosting models two simple bias correction methods that King and Zeng (2001b) discuss.

The first correction consists of attaching a weight to each observation of the balanced sample. These weights are based on marketers' prior beliefs about the churn rate π_c (i.e., the proportion of churners) among their customers. For example, π_c can be taken as the empirical frequency of churners in a proportional sample; in the current context, this is 1.8%. Let $N_c^{balanced}$ be the number of churners in the balanced sample, where N is the total size of this sample. It is possible to weight the observations of a balanced calibration sample by attaching the weights

$$(4) \qquad w_i^c = \frac{\pi_c}{N_c^{balanced}} \quad \text{and} \quad w_i^{nc} = \frac{1-\pi_c}{N - N_c^{balanced}}$$

to the churners and the nonchurners, respectively. As such, the sum of the weights associated with the churners equals

[4] Other criteria could also be considered (e.g., the Gini coefficient, the top-decile lift).

[5] More precisely, a tree is composed of several nodes, from the root to the leaves (i.e., terminal nodes). Each nonterminal node is split into two child nodes on the basis of the value of the variable that provides the maximal reduction in the squared error rate. The relative importance of a variable x_k is then the sum of these improvements (reductions) over all nodes for which the predictor x_k was selected as a splitting variable.

the real-life proportion of churners. Note that the sum of the weights we defined in Equation 4 is always equal to one. When this weighting correction is applied to bagging and stochastic gradient boosting, a sequence of weighted decision trees is estimated, and the weights remain fixed through iterations. In a statistical context, assigning weights to customers is a valid approach to correct for stratified sampling. However, because the weights assigned to the churners are small, this correction might actually cancel the advantage of oversampling the churners and thus provide similar results to a proportional sample of the same size (see the "Results" section).

Rather than weighting the observations of a balanced sample, we could employ a more simple approach by taking a nonzero cutoff value τ_B in the bagging and boosting algorithms. The value of τ_B is such that the proportion of predicted churners in the calibration sample equals the actual a priori proportion of churners π_c. This correction is achieved for bagging (and, similarly, for boosting) by first sorting the values of $\hat{f}_{bag}(x)$ in the calibration sample from the largest to the smallest value, $\hat{f}_{bag}(x_{(1)}) \geq \hat{f}_{bag}(x_{(2)}) \geq \ldots \geq \hat{f}_{bag}(x_{(N)})$, and then taking

$$(5) \qquad \tau_B = \hat{f}_{bag}(x_{(j)}), \text{ where } j = N\pi_c.$$

This latter correction method can also be called "intercept correction" (or "prior correction" in King and Zeng 2001a, b), referring to a similar correction for the logit model (see, e.g., Franses and Paap 2001, pp. 73–75). Unlike the weighting correction, the intercept correction affects neither the estimated scores nor the ranking of the customers. We assess both corrections in the "Results" section.

ASSESSMENT CRITERIA

We assess the predictive performance of the investigated models using a holdout test sample (as described in the "Data" section). Because this sample has not been used for the estimation of the classification rules and is very large, it allows for a valid assessment of performance. We denote the validation or holdout test sample as ($[x_1, y_1], \ldots, [x_i, y_i], \ldots, [x_M, y_M]$) and the computed scores as $\hat{f}(x_i)$, for $i = 1, \ldots, M$, where M is the size of the validation sample.

Error Rate

The traditional performance criterion is the error rate, that is, the percentage of incorrectly classified observations in the validation set. For rare events, as Morrison (1969) notes, the error rate is often inappropriate. For example, a naive prediction rule stating that no customer of the validation set churns has an expected error rate of approximately 1.8%, from which the classification rule could be falsely considered good. Indeed, such a rule does not isolate any group of the potentially riskiest customers for a targeted retention strategy. Another drawback is that error rates do not take the numerical values of the scores $\hat{f}(x_i)$ into account, whereas these scores may contain relevant information for proactive marketing actions. The targeting of such incentives can indeed be based on the churn degree of risk (i.e., score) of each customer (e.g., targeting the 10% riskiest customers). In contrast, the top-decile lift and the Gini coefficient are based on these scores.

Top-Decile Lift

The top-decile lift focuses on the most critical group of customers and their churn risk. The top 10% riskiest customers (i.e., those who have score values among the 10% highest) represent a potentially ideal segment for targeting a retention marketing campaign. The top-decile lift equals the proportion of churners in this risky segment, $\pi_{10\%}$, divided by the proportion of churners in the whole validation set, π:

$$(6) \qquad \text{Top decile} = \frac{\hat{\pi}_{10\%}}{\hat{\pi}}.$$

The higher the top-decile lift, the better is the classifier. This measure enables us to control whether the targeted segment of risky customers indeed contains actual churners. As Neslin and colleagues (2006) extensively describe, top-decile lift is related directly to profitability. They define the incremental gain in financial profit from an increase in top-decile lift as

$$(7) \qquad \text{Gain} = N\alpha\hat{\pi}\ (\Delta\text{Top decile})[\gamma LVC - \delta(\gamma - \psi)],$$

where N is the customer base of the company, α is the percentage of targeted customers (in our context, 10%), ΔTop decile is the increase in top-decile lift, γ is the success rate of the incentive among the churners, LVC is the lifetime value of a customer (Gupta, Lehmann, and Stuart 2004), δ is the incentive cost per customer, and ψ is the success rate of the incentive among the nonchurners (for more details, see Neslin et al. 2006).

Gini Coefficient

Another interesting measure is the Gini coefficient (e.g., Hand 1997, p. 134). Instead of focusing only on the riskiest segment, this measure considers all scores, including the less risky customers. The top-decile lift and the Gini coefficient provide complementary information; a model can be good at identifying the riskiest segment but less effective at recognizing less risky customers. We first determine the fraction of all subscribers who have a predicted churn probability above a certain threshold. We consider a whole sequence of thresholds, each of which is given by a predicted score $\hat{f}(x_l)$, for $l = 1, 2, \ldots, M$, which results in M proportions:

$$(8) \qquad \pi_l = \frac{1}{M}\sum_{i=1}^{M} I[\hat{f}(x_i) > \hat{f}(x_l)].$$

For each threshold, we also compute the fraction of all churners who have a score value above this threshold:

$$(9) \qquad \pi'_l = \frac{1}{M_c}\sum_{i=1}^{M_c} I[\hat{f}(x_i) > \hat{f}(x_l) \text{ and } y_i = 1],$$

where M_c is the total number of actual churners in the validation set. We then define the Gini coefficient as

$$(10) \qquad \text{Gini coefficient} = \frac{2}{M}\sum_{l=1}^{M} (\pi'_l - \pi_l).$$

The larger the Gini coefficient, the better is the classification model.

RESULTS

This section addresses the research questions introduced in the beginning of the article. We show that (1) both bagging and boosting techniques significantly improve the classification performance of traditional classification models, (2) the correction methods for a balanced calibration sample reduce the classification error rate, and (3) the use of a balanced calibration sample improves the forecasting accuracy of the estimated choice models.

Do Bagging and Boosting Provide Better Results than Other Benchmarks?

We apply bagging and stochastic gradient boosting, with classification trees as base classifiers, to the balanced calibration sample.[6] As a benchmark, we estimate a binary logit choice model on the same sample. Other benchmark models, including the traditional discriminant analysis, a single classification tree, and a neural network, have also been investigated (see, e.g., Thieme, Song, and Calantone 2000; West, Brockett, and Golden 1997), but they appear to perform worse than the binary logit choice model in this empirical application. Neslin and colleagues (2006) recently compared the predictive performance of different methodological approaches for this particular database and found that the logit model and the decision tree were among the most competitive methodologies. To evaluate the relative performance of the different methods, we apply the estimated models to the holdout proportional test sample to obtain churn predictions for each of the customers who belong to this sample. From these predictions, we then compute the validated error rate, the Gini coefficient, and the top-decile lift that each of the three choice models reaches.

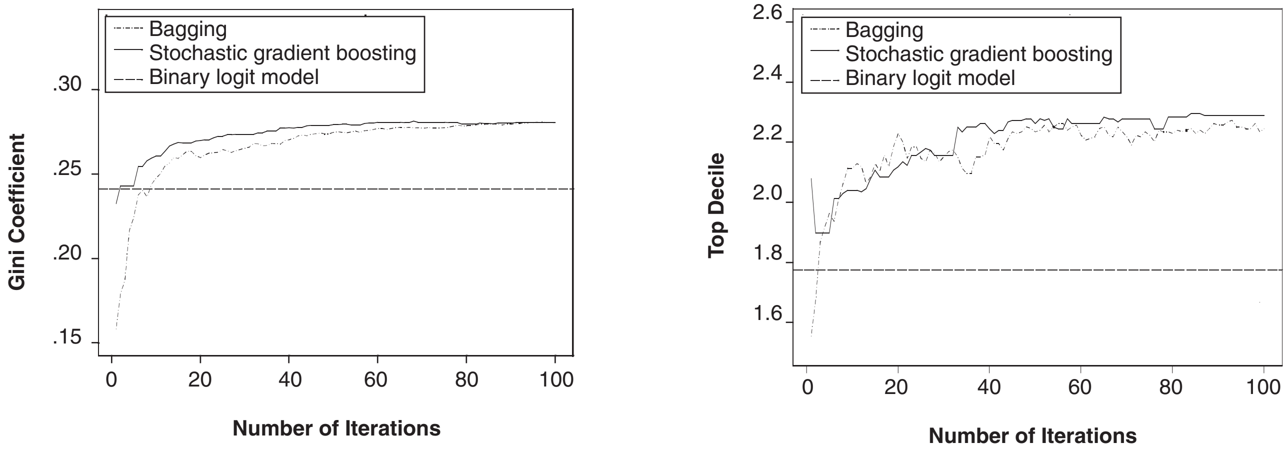
Figure 1 represents the Gini coefficient and the top-decile lift against the number of iterations for both bagging and stochastic gradient boosting.[7] The horizontal line in Figure 1 represents the performance of the binary logit model. The performance of bagging and boosting improves as B increases and stabilizes for large values of B. After the first few iterations, both models already outperform the logit benchmark, thus confirming many other examples (e.g., Hastie, Tibshirani, and Friedman 2001, pp. 246–49, 299–345).[8]

The relative gain in predictive performance is greater than 16% for the Gini coefficient and 26% for the top-decile lift. This improvement is statistically significant.[9] Stochastic gradient boosting performs similarly to bagging but is conceptually more complicated. Therefore, we consider bagging the most competitive approach, at least in this application. We can also evaluate the additional financial gains (Equation 7) expected from a retention marketing campaign that would be targeted using the scores predicted by the bagging rather than the logit model. If we consider N = 5,000,000 customers, a target group of α = 10%, γ = 30% success probability among the churners, LVC = $2,500 lifetime value, δ = $50 incentive cost, and ψ = 50% success probability among the nonchurners, the use of bagging as a scoring model (versus a logit model) for targeting a specific retention campaign is worth an additional $3,214,800.

Regarding the error rate, all three choice models perform poorly (see Table 2, Column 3), confirming that a balanced sampling scheme requires an appropriate bias correction, regardless of the choice model under consideration. In the next research question, we investigate whether a bias correction reduces these high error rates.

Figure 1

VALIDATED GINI COEFFICIENT AND TOP-DECILE LIFT FOR BAGGING, STOCHASTIC GRADIENT BOOSTING, AND A BINARY LOGIT MODEL AS A FUNCTION OF B



[6] We implemented bagging using the statistical software package Splus, whereas we computed stochastic gradient boosting using the MART software package for R that J.H. Friedman developed.

[7] Note that B is actually multiplied by ten for stochastic gradient boosting in Figure 1.

[8] The Gini coefficient and top-decile lift are –.06 and .49, respectively, for neural nets; .199 and 1.60, respectively, for discriminant analysis; and .091 and 1.37, respectively, for a single classification tree, compared with .24 and 1.77 for logit regression. These figures motivate our preference for the logit model as a benchmark.

[9] Standard errors, which we computed using a bootstrap procedure, are approximately .012 for the Gini coefficient and .09 for the top-decile lift.

Table 2

VALIDATED ERROR FOR PREDICTING CHURN FROM A BALANCED SAMPLE WITH INTERCEPT CORRECTION, WITH WEIGHTING CORRECTION, OR WITHOUT BIAS CORRECTION

Error Rate	*Intercept Correction*	*Weighting Correction*	*No Correction*
Binary logit model	.035	.018	.400
Bagging	.034	.025	.374
Stochastic gradient boosting	.034	.018	.460

Although the bagging and boosting models focus mainly on scoring customers for targeting purposes, we can also interpret the models. Figure 2 reports the 15 most important variables in explaining churn, using bagging.[10] Reported results offer some face validity. Among the particularly relevant churn triggers, we find the number of days of the current cellular phone ("equipment days"), the changes in minutes of consumption over the previous three months ("change in monthly minutes of use"), and the base cost of the calling plan the customer chose ("base cost of the calling plan"). Partial dependence plots provide additional insights into the way these variables affect churn.

It appears that the probability that a customer churns increases as his or her cellular phone becomes older (see Figure 3, "Equipment Days"). This rise is particularly important during the first year, which could be due to numerous operators proposing combined one-year-subscription and free cellular phone packages. After this delay, customers may be likely to defect from the company and buy a new package from a competitor. Figure 3 ("Change in Monthly Minutes of Use") indicates how the churn risk of a customer varies as his or her consumption habits change. When consumption decreases, the subscriber is more likely to churn. When his or her consumption is constant, the subscriber is less likely to defect. Finally, when consumption increases, the customer is slightly less (but still) likely to be loyal than when no change occurs.[11] Another interesting insight can be derived from Figure 4, which represents the partial dependence between churn and a combination of two churn drivers (i.e., the age of the customer ["age"] and the base cost of his or her calling plan). A customer is more likely to churn when his or her calling plan is cheaper. However, this relationship tends to be much stronger for younger customers than for older ones, indicating that some demographics are more likely to drop certain calling plans than others.

What Is the Best Bias Correction When Using a Balanced Calibration Sample?

We use two corrections to adapt the predicted probabilities obtained through the use of a balanced calibration sam-

[10] Boosting yields similar results, confirming the face validity of the results.

[11] Note that logit models cannot capture such nonmonotonic relations.

Figure 2

VARIABLES' RELATIVE IMPORTANCE FOR BAGGING

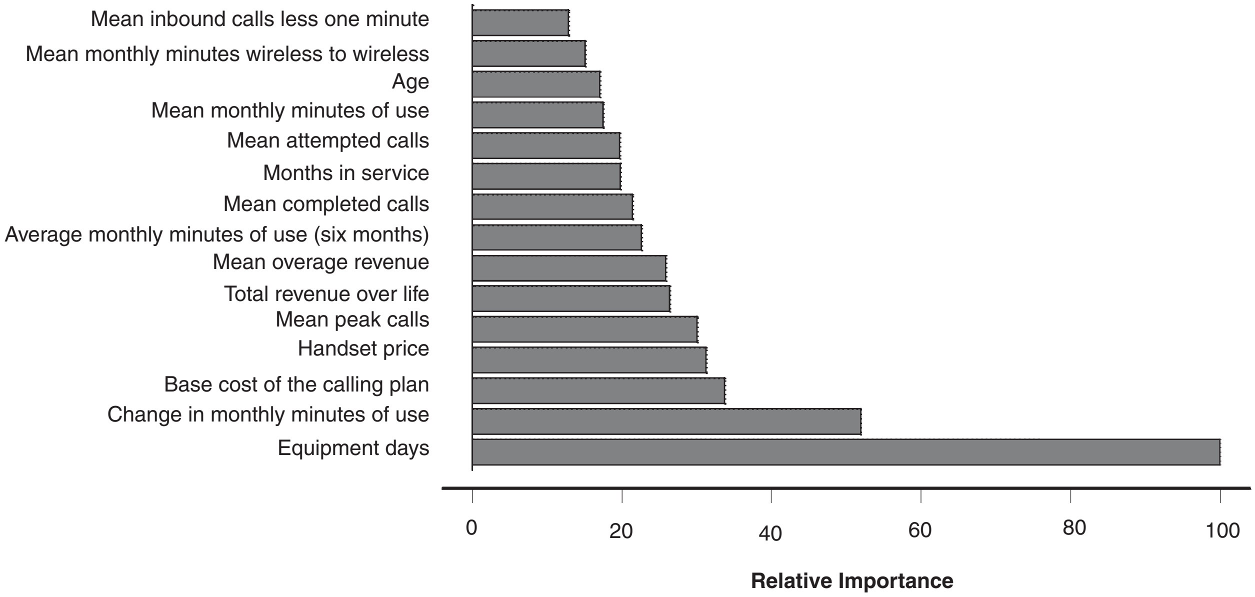


Figure 3

PARTIAL DEPENDENCE PLOTS FOR “CHANGE IN MONTHLY MINUTES OF USE” AND “EQUIPMENT DAYS” FOR BAGGING

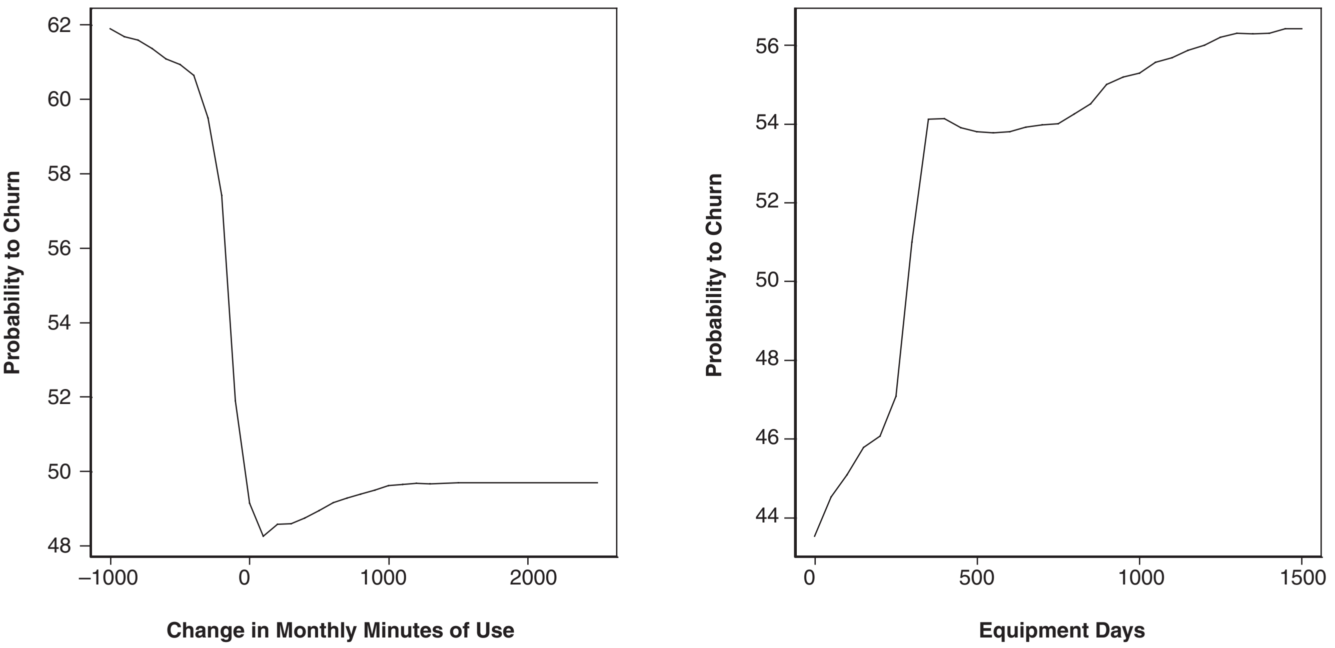
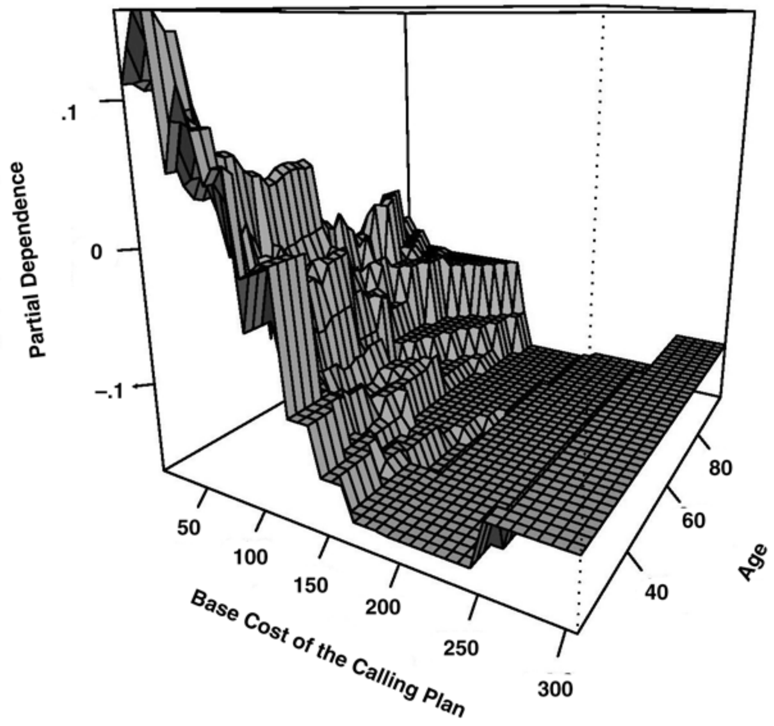


Figure 4

PARTIAL DEPENDENCE PLOT FOR THE “BASE COST OF THE CALLING PLAN” AND “AGE” FOR BAGGING



ple. Either of these two corrections reduces the error rate significantly (see Table 2).

The effectiveness of both corrections differs. For the error rate, the weighting correction seems to be the most appropriate bias correction method for all considered models. However, the weighting correction affects the estimated scores, their ranking, and, eventually, the Gini coefficient and the top-decile lift. This is not the case for the intercept correction method, which preserves the relative ranking of the attributed scores. Table 3 reports the Gini coefficient and the top-decile lift for bagging, stochastic gradient boosting, and the logit model (all estimated on the balanced sample) for both corrections. For all three models under consideration, the Gini coefficient and the top-decile lift obtained with the intercept correction are substantially better than those obtained with the weighting correction.

This confirms the prior assumption that weighting the observations of a balanced sample cancels the advantage of balanced sampling, even for large sample sizes. Because we consider the Gini coefficient and the top-decile lift more global measures of performance than the error rate, the intercept correction is the best compromise between no correction (i.e., a better Gini coefficient and top-decile lift but a worse error rate) and weighting correction (i.e., a worse Gini coefficient and top-decile lift but a better error rate), at least in this application.

Note that the intercept correction appears to perform well for stable markets (e.g., constant churn rate), but it is likely to be inefficient in dynamic markets (e.g., increasing churn rate). This constitutes a major limitation to the correction methods we propose in this study. Moreover, the lack of theory about the properties of these correction methods prevents us from generalizing our findings to any other setting.

Table 3

VALIDATED GINI COEFFICIENT AND TOP-DECILE LIFT FOR PREDICTING CHURN FROM A BALANCED SAMPLE WITH INTERCEPT CORRECTION AND WEIGHTING CORRECTION

	Intercept Correction[a]		*Weighting Correction*	
	Gini Coefficient	*Top Decile*	*Gini Coefficient*	*Top Decile*
Binary logit model	.241	1.775	.239	1.764
Bagging	.281	2.246	.161	1.549
Stochastic gradient boosting	.280	2.290	.187	1.632

[a]The Gini coefficients and top-decile lifts are the same for the "no-correction" method.

Does a Choice Model Estimated on a Balanced Sample, with Bias Appropriately Corrected for, Outperform a Choice Model Estimated on a Proportional Sample?

A balanced calibration sample is often advised when the variable to be predicted consists of a rare event, such as churn. However, our third research issue questions this advice. Indeed, given the high amount of observations in the proportional calibration sample, the absolute number of churners is still quite large, and a proportional sampling could still be efficient.

Table 4 compares the performance of bagging, stochastic gradient boosting, and the binary logit model, estimated from the proportional or the balanced sample (with intercept correction). The results of both the Gini coefficient and the top-decile lift indicate that the balanced sampling scheme is recommended for the three investigated classification models. For the error rate, the results are more in favor of proportional sampling. However, for the same reasons as in the preceding subsection, we consider the balanced sampling a better compromise than the proportional sampling, which performs poorly for the Gini coefficient and top-decile lift.

CONCLUSIONS

In this article, we discussed several new developments from the machine-learning and statistical classification literature in the context of marketing research. We presented one of the simplest versions of classifier aggregation (i.e., bagging) and one of the most sophisticated algorithms in this field (i.e., stochastic gradient boosting). We especially drew attention to the competitive performance of bagging, an easy-to-use procedure aimed at increasing the classification performance of an initial classification model, by repeatedly estimating a classifier to bootstrapped versions of the calibration sample. We summarize the main findings of this study in terms of three contributions: First, bagging and boosting provide substantially better classifiers than a binary logit model. In predicting churn, the gain in predictive performance has reached 16% for the Gini coefficient and 26% for the top-decile lift. Bagging and stochastic gradient boosting perform comparatively. The performance of the simple and easy-to-use bagging is especially noticeable. In addition to their higher predictive power, bagging and boosting provide good diagnostic measures, variable importance, and partial dependence plots, which offer face validity to the models and interesting insights into potential churn drivers.

Second, in the presence of a rare event, such as churn, we recommend a balanced sampling scheme over proportional sampling for all considered classification models (i.e., bag-

Table 4

VALIDATED GINI COEFFICIENT, TOP-DECILE LIFT, AND ERROR RATE WITH A BALANCED AND A PROPORTIONAL CALIBRATION SAMPLING

	Balanced Sample (Intercept Correction)			*Proportional Sample*		
	Gini Coefficient	*Top Decile*	*Error Rate*	*Gini Coefficient*	*Top Decile*	*Error Rate*
Binary logit model	.241	1.775	.035	.181	1.665	.018
Bagging	.281	2.246	.034	.237	1.886	.018
Stochastic gradient boosting	.280	2.290	.034	.113	1.560	.018

ging, boosting, and logit models), even for large data sets. However, to maintain the classification error rate at a reasonable level, it is necessary to correct the predictions obtained from a balanced sample. Third, intercept correction constitutes an appropriate bias correction for a balanced sampling scheme.

If companies take into account these recommendations, they might better identify the riskiest customer segments in terms of churn risk and thus ameliorate their retention strategy. Noteworthy losses could ultimately be avoided.

REFERENCES

Andrews, Rick L., Andrew Ainslie, and Imran S. Currim (2002), "An Empirical Comparison of Logit Choice Models with Discrete Versus Continuous Representations of Heterogeneity," *Journal of Marketing Research*, 39 (November), 479–87.

——— and Imran S. Currim (2002), "Identifying Segments with Identical Choice Behaviors Across Product Categories: An Intercategory Logit Mixture Model," *International Journal of Research in Marketing*, 19 (1), 65–79.

Arora, Neeraj, Greg M. Allenby, and James L. Ginter (1998), "A Hierarchical Bayes Model of Primary and Secondary Demand," *Marketing Science*, 17 (1), 29–44.

Athanassopoulos, Antreas D. (2000), "Customer Satisfaction Cues to Support Market Segmentation and Explain Switching Behavior," *Journal of Business Research*, 47 (3), 191–207.

Baines, Paul R., Robert M. Worcester, Jarrett David, and Roger Mortimore (2003), "Market Segmentation and Product Differentiation in Political Campaigns: A Technical Feature Perspective," *Journal of Marketing Management*, 19 (1–2), 225–49.

Bhattacharya, C.B. (1998), "When Customers Are Members: Customer Retention in Paid Membership Contexts," *Journal of the Academy of Marketing Science*, 26 (1), 31–44.

Bolton, Ruth N., P.K. Kannan, and Matthew D. Bramlett (2000), "Implications of Loyalty Program Membership and Service Experiences for Customer Retention and Value," *Journal of the Academy of Marketing Science*, 28 (1), 95–108.

Breiman, Leo (1996), "Bagging Predictors," *Machine Learning*, 24 (2), 123–40.

———, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone (1984), *Classification and Regression Trees*. New York: Chapman and Hall.

Bühlmann, Peter and Bin Yu (2002), "Analyzing Bagging," *Annals of Statistics*, 30 (4), 927–61.

Cardell, Scott N., Mikhail Golovnya, and Dan Steinberg (2003), "Churn Modeling for Mobile Telecommunications: Winning the NCR Teradata Center for CRM at Duke University: Salford Systems," paper presented at the 2003 INFORMS Marketing Science Conference, University of Maryland (June 12–15).

Chung, Jaihak and Vithala R. Rao (2003), "A General Choice Model for Bundles with Multiple-Category Products: Application to Market Segmentation and Optimal Pricing for Bundles," *Journal of Marketing Research*, 40 (May), 115–30.

Colgate, Mark R. and Peter J. Danaher (2000), "Implementing a Customer Relationship Strategy: The Asymmetric Impact of Poor Versus Excellent Execution," *Journal of the Academy of Marketing Science*, 28 (3), 375–87.

Corstjens, Marcel L. and David A. Gautschi (1983), "Formal Choice Models in Marketing," *Marketing Science*, 2 (1), 19–56.

Cosslett, S.R. (1993), "Estimation from Endogenously Stratified Samples," in *Handbook of Statistics*, Vol. 11, G.S. Maddala, C.R. Rao, and H.D. Vinod, eds. Amsterdam: Elsevier Science.

Currim, Imran S., Robert J. Meyer, and Nhan T. Le (1988), "Disaggregate Tree-Structure Modeling of Consumer Choice Data," *Journal of Marketing Research*, 25 (August), 253–65.

Donkers, Bas, Philip H.B.F. Franses, and Peter Verhoef (2003), "Selective Sampling for Binary Choice Models," *Journal of Marketing Research*, 40 (November), 492–97.

———, Richard Paap, Jedid-Jah Jonker, and Philip H.B.F. Franses (2006), "Deriving Target Selection Rules from Endogenously Selected Samples," *Journal of Applied Econometrics*, forthcoming.

Franses, Philip H. and Richard Paap (2001), *Quantitative Models for Marketing Research*. Cambridge, UK: Cambridge University Press.

Freund, Yoav and Robert E. Schapire (1996), "Experiments with a New Boosting Algorithm," in *Proceedings of the 13th International Conference on Machine Learning*, L. Saitta, ed. Bari, Italy: Morgan Kaufmann, 148–56.

Friedman, Jerome H. (2001), "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics*, 29 (5), 1189–1232.

——— (2002), "Stochastic Gradient Boosting," *Computational Statistics and Data Analysis*, 38 (4), 367–78.

———, Trevor Hastie, and Robert Tibshirani (2000), "Additive Logistic Regression: A Statistical View of Boosting," *Annals of Statistics*, 28 (2), 337–407.

Ganesh, Jaishankar, Mark J. Arnold, and Kristy E. Reynolds (2000), "Understanding the Customer Base of Service Providers: An Examination of the Differences Between Switchers and Stayers," *Journal of Marketing*, 64 (July), 65–87.

Guadagni, Peter M. and John D.C. Little (1983), "A Logit Model of Brand Choice Calibrated on Scanner Data," *Marketing Science*, 2 (3), 203–238.

Gupta, Sunil, Donald R. Lehmann, and Jennifer A. Stuart (2004), "Valuing Customers," *Journal of Marketing Research*, 41 (February), 7–18.

Hand, David J. (1997), *Construction and Assessment of Classification Rules*. Chichester, UK: John Wiley & Sons.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2001), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer-Verlag.

Haughton, Dominique and Samer Oulabi (1997), "Direct Marketing Modeling with CART and CHAID," *Journal of Direct Marketing*, 11 (4), 42–52.

Hawley, David (2003), "International Wireless Churn Management: Research and Recommendations," Yankee Group report, (June), (accessed January 2006), [available at http://www.ams.com/cme/pdfs/yankeechurnstudy.pdf].

Imbens, Guido W. and Tony Lancaster (1996), "Efficient Estimation and Stratified Sampling," *Journal of Econometrics*, 74 (2), 289–318.

Kalwani, Manohar U., Robert J. Meyer, and Donald G. Morrison (1994), "Benchmarks for Discrete Choice Models," *Journal of Marketing Research*, 31 (February), 65–75.

King, Gary and Langsche Zeng (2001a), "Explaining Rare Events in International Relations," *International Organization*, 55 (3), 693–715.

——— and ——— (2001b), "Logistic Regression in Rare Events Data," *Political Analysis*, 9 (2), 137–63.

Morrison, Donald G. (1969), "On the Interpretability of Discriminant Analysis," *Journal of Marketing Research*, 6 (May), 156–63.

Nardiello, Pio, Fabrizio Sebastiani, and Alessandro Sperduti (2003), "Discretizing Continuous Attributes in AdaBoost for Text Categorization," in *Proceedings of the 25th European Conference on Information Retrieval Research*, Fabrizio Sebastiani, ed. Heidelberg, Germany: Springer-Verlag, 320–34.

Neslin, Scott A., Sunil Gupta, Wagner Kamakura, Junxiang Lu, and Charlotte H. Mason (2006), "Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models," *Journal of Marketing Research*, 43 (May), 204–211.

Schapire, Robert E. and Yoram Singer (1999), "Improved Boosting Algorithms Using Confidence-Rated Predictions," *Machine Learning*, 37 (3), 297–336.

Scott, Alastair J. and Chris J. Wild (1997), "Fitting Regression Models to Case-Control Data by Maximum Likelihood," *Biometrika*, 84 (1), 57–71.

Shaffer, Greg and John Z. Zhang (2002), "Competitive One-to-One Promotions," *Management Science*, 48 (9), 1143–60.

Snel, Ross (2000), "Fighting the Fickle," *The Wall Street Journal (Europe)*, (September 18), (accessed January 2006), [available at http://interactive.wsj.com/public/current/articles/SB968956754675902638.htm].

Thieme, R. Jeffrey, Michael Song, and Roger J. Calantone (2000), "Artificial Neural Network Decision Support Systems for New Product Development Project Selection," *Journal of Marketing Research*, 37 (November), 499–507.

Varmuza, Kurt, Ping He, and Kai-Tai Fang (2003), "Boosting Applied to Classification of Mass Spectral Data," *Journal of Data Science*, 1 (4), 391–404.

Viaene, Stijn, Richard A. Derrig, and Guido Dedene (2002), "Boosting Naive Bayes for Claim Fraud Diagnosis," in *Lecture Notes in Computer Science 2454*, Y. Kambayashi, W. Winiwarter, and M. Arikawa, eds. Berlin: Springer, 202–211.

Wedel, Michel and Wagner A. Kamakura (2000), *Market Segmentation: Conceptual and Methodological Foundations*, 2d ed. Boston: Kluwer Academic Publishers.

West, Patricia M., Patrick L. Brockett, and Linda L. Golden (1997), "A Comparative Analysis of Neural Networks and Statistical Methods for Predicting Consumer Choice," *Marketing Science*, 16 (4), 370–91.

Yang, Sha and Greg M. Allenby (2003), "Modeling Interdependent Consumer Preferences," *Journal of Marketing Research*, 40 (August), 282–94.