SCOTT A. NESLIN, SUNIL GUPTA, WAGNER KAMAKURA, JUNXIANG LU, and CHARLOTTE H. MASON*

This article provides a descriptive analysis of how methodological factors contribute to the accuracy of customer churn predictive models. The study is based on a tournament in which both academics and practitioners downloaded data from a publicly available Web site, estimated a model, and made predictions on two validation databases. The results suggest several important findings. First, methods do matter. The differences observed in predictive accuracy across submissions could change the profitability of a churn management campaign by hundreds of thousands of dollars. Second, models have staying power. They suffer very little decrease in performance if they are used to predict churn for a database compiled three months after the calibration data. Third, researchers use a variety of modeling "approaches," characterized by variables such as estimation technique, variable selection procedure, number of variables included, and time allocated to steps in the model-building process. The authors find important differences in performance among these approaches and discuss implications for both researchers and practitioners.

Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models

Customer churn has become a significant problem for firms in publishing, financial services, insurance, electric utilities, health care, banking, Internet, telephone, and cable service industries. In the cellular phone industry, annual churn rates range from 23.4% (*Wireless Review* 2000) to 46% (Fitchard 2002). Customer churn figures directly in how long a customer stays with a company and, in turn, the customer's lifetime value (CLV) to that company.

A way to manage customer churn is to predict which customers are most likely to churn and then target incentives to those customers to induce them to stay. This approach enables the firm to focus its efforts on customers who are truly at risk to churn, and it potentially saves money that would be wasted in providing incentives to customers who do not need them. However, the approach assumes that customer churn can be predicted with acceptable accuracy.

The purpose of this article is to identify which methodological approaches work best for predicting customer churn. We focus on three research questions:

- Does method make a difference? Are differences in predictive accuracy across various techniques managerially meaningful?
- Do models have staying power? Can a model estimated at time t predict customer churn at time t + x, where x is some later time period?
- Which methods work best? How do the various statistical techniques, variable selection approaches, and time allocation strategies contribute to predictive accuracy? What overall approaches are likely to be successful?

*Scott A. Neslin is Albert Wesley Frey Professor of Marketing, Tuck School of Business, Dartmouth College (e-mail: scott.neslin@dartmouth.edu). Sunil Gupta is Meyer Feldberg Professor of Business, Graduate School of Business, Columbia University (e-mail: sg37@columbia.edu). Wagner Kamakura is Ford Motor Company Professor of Global Marketing, Fuqua School of Business, Duke University (e-mail: kamakura@duke.edu). Junxiang Lu is Vice President of Comerica Bank (e-mail: jlu@comerica.com). Charlotte H. Mason is Associate Professor of Marketing, Kenan-Flagler Business School, University of North Carolina (e-mail: charlotte_mason@unc.edu). The authors express their gratitude to Sanyin Siang (Managing Director, Teradata Center for Customer Relationship Management at the Fuqua School of Business, Duke University); research assistants Sarwat Husain, Michael Kurima, and Emilio del Rio; and an anonymous wireless telephone carrier that provided the data for this study. The authors also thank participants in the Tuck School of Business, Dartmouth College, Marketing Workshop, for comments and the two anonymous *JMR* reviewers for their constructive suggestions. Finally, the authors express their appreciation to former editor Dick Wittink (posthumously) for his invaluable insights and guidance.

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

Journal of Marketing Research
Vol. XLIII (May 2006), 204–211

We answer these questions by providing a descriptive analysis of data collected in a tournament administered by the Teradata Center for Customer Relationship Management (CRM) at the Fuqua School of Business at Duke University. Data were provided to all model builders who were interested in participating. Each participant estimated a churn prediction model and then used the model to generate predictions for validation data. We also surveyed participants on the methodologies they used. We were then able to conduct a post hoc "meta-analysis" of the results to answer our research questions.

Tournaments are used in fields other than marketing, notably the knowledge discovery and data-mining tournaments (see, e.g., http://kdd05.lac.uic.edu/kddcup.html). Benefits include scale (we have 33 participants who submitted 44 entries), generalizability (participants included academics and practitioners involved in modeling churn), and insight (the diversity of approaches created a rich database for distilling the methodological factors that influence predictive accuracy). The disadvantage of the tournament is the lack of a factorial design for various combinations of methods (see Kumar, Rao, and Soni 1995). However, the scale, generalizability, and insight generated by the tournament provide a compelling set of advantages.

This article aims to contribute to a research stream on the prediction of key marketing phenomena, such as market share (Ghosh, Neslin, and Shoemaker 1984), interpurchase times (Helsen and Schmittlein 1993), and new product acceptance (Kumar, Rao, and Soni 1995). The field of CRM has renewed the emphasis on predicting phenomena, such as direct mail response (Levin and Zahavi 1998, 2001) and next product to buy (Knott, Hayes, and Neslin 2002). Our research is distinctive in its emphasis on churn, its use of a tournament, and its relating of predictive accuracy to profitability of a churn management campaign.

PREDICTIVE ACCURACY AND CHURN MANAGEMENT PROFITABILITY

We formulate profitability of a single churn management campaign as a function of the ability of the predictive model to identify would-be churners:

- N = the total number of customers,
- α = the fraction of customers who are targeted for the churn management program,
- β = the fraction of targeted customers who are would-be churners,
- δ = the cost of the customer incentive to the firm,
- γ = the fraction of targeted would-be churners who decide to remain because of the incentive (i.e., the success rate of the incentive),
- c = the cost of contacting a customer to offer him or her the incentive,
- CLV = the customer lifetime value (i.e., the value to the firm if the customer is retained), and
- A = the fixed administrative costs of running the churn management program.

Given these definitions, the profit that a single churn management campaign contributes is as follows:

$$(1) \qquad \Pi = N\alpha[\beta\gamma(CLV - c - \delta) + \beta(1 - \gamma)(-c) + (1 - \beta)(-c - \delta)] - A.$$

The first term within the brackets reflects profit contribution among the $\beta\gamma$ fraction of contacted customers who are would-be churners and decide to stay on the basis of the incentive. The second term reflects the cost of contacting the $\beta(1 - \gamma)$ fraction of would-be churners who do not accept the offer and leave the firm. The third term reflects the cost among the $(1 - \beta)$ fraction of contacted customers who are not would-be churners but accept the offer.

The term β reflects the model's accuracy:

- β_0 = the fraction of all the firm's customers who will churn, and
- λ = "lift" (i.e., how much more likely the contacted group of customers is to churn than all the firm's customers). Thus, $\lambda = 1$ means that the model provides essentially no predictive power because the targeted customers are no more likely to churn than the population as a whole. As such, λ should be greater than one.

We can then express β as

$$(2) \qquad \beta = \lambda\beta_0.$$

Substituting Equation 2 into Equation 1 and rearranging terms, we obtain

$$(3) \qquad \Pi = N\alpha\{[\gamma CLV + \delta(1 - \gamma)]\beta_0\lambda - \delta - c\} - A.$$

The incremental gain in profit from a unit increase in predictive accuracy λ is the slope of Equation 3, namely,

$$(4) \qquad GAIN = N\alpha\{[\gamma CLV + \delta(1 - \gamma)]\ \beta_0\}.$$

The gain in profit from improved accuracy increases when (1) the size of the campaign is larger ($N\alpha$), (2) the potential recaptured CLV is higher, (3) the campaign's success rate (γ) is higher, (4) the incentive cost (δ) is higher (because more incentive money will be wasted if accuracy is poor), and (5) the base churn rate (β_0) is higher. We use Equation 4 to assess the profitability impact of using different methods to predict churn.

TOURNAMENT STRUCTURE

Overview

The Teradata Center for CRM provided data freely on its Web site. The tournament was publicized through several vehicles that target academics and practitioners. Participants downloaded the data along with full descriptions of the tournament and the data. After participants developed their models and calculated their predictions, they uploaded the predictions to the Web site and completed a survey about their methodological approach. The predictions were merged with the actual churn results and scored in terms of four criteria. Cash prizes were awarded to the winners.

Data

The data consisted of one calibration and two validation databases. The calibration data contained 171 potential predictor variables for 100,000 customers. The predictors were calculated over a three-month period, and churn was measured in the fifth month. The one-month lag between the predictors and the churn month reflects the logistics of implementing a churn management campaign with the goal of reaching customers *before* they churn. Although the monthly churn rate for the company was 1.8%, 50% of the customers in the calibration data were churners. This is because with low-incidence data, oversampling on inci-

dence provides more information, in this case, to profile churners versus nonchurners (King and Zeng 2001).

The validation databases included the same predictors as the calibration data but no churn indicator. The current score data were compiled at the same time as the calibration data and included 51,306 customers, 1.80% of whom were churners. The future score data were compiled roughly three months later to provide guidance on model "shelf life." Although shelf life is a practical concern, there are few guidelines to determine it. It can depend on the business environment, technology, and customer base (Berry and Linoff 2000; Rud 2001). In wireless telecom, frequent changes in market conditions, including new service plans, new equipment, and features (from the firm or its competitors), make model shelf life particularly relevant. The future score data included 100,462 customers, 1.80% of whom were churners.

The 171 predictors included customer behavior, such as minutes of use, revenue, handset equipment, and trends in usage; company interaction data, such as calls to the customer service center; and customer household demographics, including age, income, geographic location, and home ownership. The data did not include any previous targeted marketing efforts.

Two prediction criteria were used for each validation database, resulting in four "contests" in all. They were top-decile lift (λ in Equation 3 when $\alpha = .10$) and the Gini coefficient (Alker 1965; Statistics.com 2002). Lift is probably the most commonly used prediction criterion in predictive modeling, and its relevance is demonstrated by the direct link between lift and profitability we demonstrated previously. The Gini coefficient represents the area between the cumulative lift curve and random prediction, so it is a broader measure than top-decile lift.

RESULTS

Submissions

The tournament attracted 44 entries from 33 participants. Half were academics, and half were consultants and practitioners from companies with an interest in managing customer churn. The survey indicated that logistic regression (used by 45%) and decision trees (23%) were the most common estimation techniques, but neural nets (11%), discriminant analysis (9%), cluster analysis (7%), and Bayes (5%) were used as well. Most participants (88%) explored more than one estimation technique (average = 3.27). For variable selection, entrants relied on exploratory data analysis (EDA; average rating = 5.58 on a 1–7 scale), common sense (4.72), and stepwise procedures (4.59). Participants also used theory (average rating = 3.56), factor analysis (3.05), and cluster analysis (2.52). Most entries (82%) used fewer than 80 predictors (usually fewer than 40). However, a few entries (18%) used more than 140 predictors. More than half (56%) of the participants divided the calibration data into estimation and holdout samples. The average entry required 60 hours of work in total, which broke down as follows: downloading (1.03 hours), cleaning data (15.22 hours), creating variables (15.18 hours), estimating (24.33 hours), and preparing prediction files (4.71 hours). Across the 44 entries, this amounts to 2440 effort hours, or 61 40-hour workweeks in total.

Overall Performance

The performance statistics that we summarize in Table 1 suggest three important conclusions:

- Entries vary significantly in predictive performance: The range in top-decile lift is approximately 2 units, from 1.07 to 3.01, with a standard deviation of approximately one-half unit.
- There is little falloff in prediction between current score data and future score data: The average lift decreases from 2.14 to 2.13, and the average Gini decreases from .269 to .265.
- The predictive criteria are highly correlated: All the correlations are greater than .9, though the correlations are higher for two different measures on the same database than for the same measure on different databases.

The range in top-decile-lift results suggests that the best-performing models identify 10% of the customers who are three times more likely to churn than average (see the previous definition of lift and Equation 2), whereas the worst-performing models perform barely better than random. That appears to be a wide spread in performance. The current score data versus the future score data results suggest that there is little falloff in predictive ability at least three months after initial model calibration. The finding that top-decile-lift and Gini-coefficient criteria are highly correlated is notable. The Gini coefficient takes into account not only top-decile lift but performance in the subsequent deciles as well. The more commonly used top-decile lift correlates strongly with Gini, suggesting that predictive performance in the top decile is key, and top-decile lift can be used by itself as a measure of predictive performance.

Factors Determining Performance

Because the measures describing the approaches that participants used for variable selection, statistical technique, time allocation, and so forth, were highly correlated, we conducted a factor analysis of these measures. We then related the factor scores to overall performance. Table 2 shows the loadings matrix from our factor analysis of these measures.[1]

[1]We used principal components analysis and Varimax rotation. Seven eigenvalues were greater than 1, but we selected the five-factor solution as the most interpretable. There are two notable technical aspects of the factor analysis: First, it includes dichotomous and continuous variables. We used Pearson correlation coefficients for all pairs of items (dichotomous with dichotomous, dichotomous with continuous, and continuous with continuous) because though the phi coefficient is often recommended to measure correlation between two dichotomous variables and though the point biserial correlation is often recommended to measure correlation between dichotomous and continuous variables, the Pearson correlation equals these measures for the dichotomous case (see http://v8doc.sas.com/sashtml/ [search for "biserial correlation"]). Second, the 5 relative-time items (Table 2) sum to one, making one of them redundant. However, we included all 5 items in the factor analysis for interpretability. This makes the correlation matrix among all 20 items singular, so it would be impossible to extract 20 factors. However, we extracted only 5 factors, so SPSS, which we used for the factor analysis, was able to do this. Note that we included practitioner/academic in the analysis to capture participant-specific effects. This variable was correlated with other variables, but we included it because it added explanatory power to the regressions we report in the next section. The insights from these regressions are unchanged if we omit the practitioner variable from the factor analysis. If we omit practitioner from the factor analysis but include it in the regressions, the results are similar, but the statistical significance and magnitudes of the factors are reduced, as would be expected because our sample size is not large and because practitioner is correlated with the methodological factor scores.

Table 1
OVERALL PERFORMANCE

A: Descriptive Statistics

Criteria	*M*	*SD*	*Minimum*	*Maximum*
Lift current	2.14	.53	1.07	2.9
Lift future	2.13	.53	1.19	3.01
Gini current	.269	.1	.06	.41
Gini future	.265	.09	.05	.4

B: Correlation Matrix

	Lift Current	*Lift Future*	*Gini Current*	*Gini Future*
Lift current	1			
Lift future	.939	1		
Gini current	.982	.929	1	
Gini future	.939	.969	.949	1

We label the first factor the "logit" approach.[2] It entails the use of logistic regression, EDA, and stepwise procedures for variable selection and the allocation of relatively less time in the preparation of prediction files. Practitioners are associated with this factor. The second factor ("tree") is characterized by a heavy reliance on decision trees and a particularly low reliance on EDA and stepwise procedures for variable selection. Participants who scored high on this factor allocated a lot of their time to estimation. They spent relatively more time in total and subdivided the data into calibration and holdout samples. This makes sense because in the development of tree models, time is spent on estimating the model (i.e., "growing" and "pruning" the trees).

Participants who scored high on the third factor ("practical") did not have a particular estimation preference but instead relied heavily on common sense in their variable selection. They allocated more time than average to download data but less time in total on the exercise, and they did not subdivide the data. Practitioners were fairly strongly associated with this factor.

The fourth factor ("discriminant") relies heavily on discriminant analysis and cluster analysis for selecting variables. Those who scored high on this factor allocated less time on data cleaning and more time on estimation, and they used many variables. This is somewhat paradoxical because presumably, the cluster analysis would have created a smaller group of variables.

The fifth factor ("explain") was associated with no particular estimation technique but was strongly associated with self-reported use of theory, factor analysis, and cluster analysis for variable selection. This suggests that "explainers" were equally interested in both understanding and predicting churn. Consistent with the use of factor analysis and cluster analysis, these participants tended to use fewer variables, and consistent with a desire for deeper understanding, they explored several estimation techniques before selecting their final one.

[2]Note that we use the common practice of subjectively labeling the factors. This is subject to the usual caveat that a brief label cannot capture the full richness of the analysis. For example, the label "logit" does capture that factor, but we remind the reader that these factors are much more than the statistical technique used for estimation.

In summary, the factor analysis suggested five general approaches to estimating customer churn: logit, trees, practical, discriminant, and explain. The next task is to examine how higher scores on these factors are related to performance.

Regression Results of Performance

Table 3 reports the regressions of each of the four performance measures against the five factor scores. The R-square values range from .45 to .50, which is acceptable for cross-sectional "meta-analysis" regressions of this type (Farley and Lehmann 1986), and are all statistically significant.[3]

The results are consistent across the four performance measures and suggest the following:

- Logit and tree approaches are positively associated with predictive performance. Because the factors are orthogonal, they are two independent approaches, and both tend to do well.
- The practical approach is the "middle-of-the-road." The coefficient for this variable is often not significantly different from zero at conventional levels, but the trend is clear: Participants who score high on this factor tend not to do as well as the logit and tree modelers but do better than the discriminant and explain modelers.
- The discriminant and explain approaches do not do as well. Often, the coefficients are not significantly different from zero, but the signs are consistently negative.

Profit Impact of Results

We use Equation 4 to assess the profit impact of our results. We express this on a per-customer basis because firms differ in size. To calculate Equation 4, we use $\delta = \$50$, representing, for example, a month's service rebate as an incentive, and $\beta_0 = .018$, which is the base churn rate for our data. In addition, CLV varies by customer, and we investigate a range of values ($500, $1,500, and $2,500).[4] No published data are available for acceptance rate (γ), so we investigated 10%, 30%, and 50% rates.

First, note that even if we assume the most conservative CLV and γ, the differences in results shown in Table 1 are meaningful from a profit standpoint. Substituting these values into Equation 4 (and setting $N\alpha = 1$ to put it on a customer basis) results in a profit gain of $(.10 \times 500 + 50 \times [1 - .1]) \times .018 = \$1.71$ increase in per-customer profit per unit change in lift. A half-unit change (the standard deviation of lift in Table 1 is approximately .5) yields $.85 per customer. For a company with 5 million customers, contacting 10% of them for a churn management campaign yields an impact of $427,500 ($.855 \times 500,000). This means that the variation

[3]Note that the sample size after we account for missing data is n = 35. Six of these observations represent multiple submissions from the same applicant (four respondents made two submissions, and one made three). Given our sample size, it would be difficult to assess potential correlations in these observations. We reran the analysis, deleting the multiple submissions; this left us with n = 29, and the results were very similar.

[4]On one extreme, if we use an average monthly churn of 1.8% (equivalent to 80% annual retention), a revenue per month of $58 (the average for our data), and an annual discount rate of 5% and if we assume that the company is interested in cash flow and not gross profits, CLV is $2,973. At another extreme, if we use an average monthly churn of 5.4% (three times the average), a revenue per month of $40, an annual discount of 5%, and a gross profit margin of 65% (e.g., U.S. Cellular's gross profit margin is 63%; see http://money.cnn.com/news/companies/research/research.html?pg=fi&osymb=USM&sid=), CLV is $610.

we observe in the results in Table 1 can easily amount to changes in profit in the hundreds of thousands of dollars by using one method rather than another.

Second, we use the regression results to calculate the impact of methodological approach on profits. In particular, we calculate the impact of a standard deviation change in

Table 2
FACTOR LOADINGS

		Logit	*Trees*	*Practical*	*Discriminant*	*Explain*
Estimation	Logit	.695	–.409	.207	–.378	.089
	Neural	–.642	–.072	–.103	–.181	–.196
	Tree	–.020	.698	–.164	–.045	–.061
	Discriminant	–.116	–.186	–.19	.872	.072
Variable selection	EDA	.375	–.610	.153	–.409	.028
	Theory	.086	–.113	.178	.050	.797
	Sense	–.015	–.132	.633	.255	.152
	Stepwise	.930	–.506	–.003	–.065	.560
	Factor	–.292	.057	.017	–.214	.787
	Cluster	.214	–.156	–.068	.519	.658
Relative time	Downloading	.059	–.068	.811	–.060	.145
	Data cleaning	.127	–.387	–.436	–.477	–.087
	Creating variables	–.221	–.679	–.061	.119	–.006
	Estimation	.268	.786	.233	.288	.004
	Preparing prediction files	–.748	–.183	.239	.047	.180
Total time	Total	–.162	.513	–.675	–.122	.017
Subdivde	Subdivide	–.036	–.136	–.423	.085	–.028
Vars	Number of variables	.012	.460	.085	.689	–.336
Exploration	Number of techniques explored	.198	.037	.194	.002	.836
Practitioner	Practitioner respondent	.657	.344	.355	–.184	–.049

Notes: The five factors accounted for 66.0% of the variance in the above 20 items.

Notes: Variable Definitions

Estimation	Logit	0–1 indicator: 1 = used logistic regression in estimation.
	Neural	0–1 indicator: 1 = used neural nets in estimation.
	Tree	0–1 indicator: 1 = used decision tree in estimation.
	Discriminant	0–1 indicator: 1 = used discriminant analysis in estimation.
Variable selection	EDA	Extent to which EDA was used in variable selection (1–7 scale).
	Theory	Extent to which theory was used in variable selection (1–7 scale).
	Sense	Extent to which common sense was used in variable selection (1–7 scale).
	Stepwise	Extent to which stepwise procedure was used in variable selection (1–7 scale).
	Factor	Extent to which factor analysis was used in variable selection (1–7 scale).
	Cluster	Extent to which cluster analysis was used in variable selection (1–7 scale).
Relative time	Downloading	Fraction of total time spent on exercise allocated to data downloading.
	Data cleaning	Fraction of total time spent on exercise allocated to data cleaning.
	Creating variables	Fraction of total time spent on exercise allocated to creating variables.
	Estimation	Fraction of total time spent on exercise allocated to estimation.
	Preparing prediction file	Fraction of total time spent on exercise allocated to preparing prediction files.
Total time	Total	Total time in hours spent on exercise.
Subdivision	Subdivide	0–1 indicator: 1 = divided data into estimation and holdout samples.
Number of variables	Number of variables	Number of variables included in final model.
Exploration	Techniques explored	Number of estimation techniques explored (ranging from 1 to 7).
Participant	Practitioner	0–1 indicator: 1 = practitioner, 0 = academic.

Table 3
REGRESSION RESULTS

		Current Lift		*Current Gini*		*Future Lift*		*Future Gini*	
		Standard Coefficient	*p Value*	*Standard Coefficient*	*p Value*	*Standard Coefficient*	*p Value*	*Standard Coefficient*	*p Value*
Approach factor score	Logit	.527	.001	.548	.000	.512	.001	.567	.000
	Tree	.293	.042	.348	.015	.332	.017	.342	.012
	Practical	.187	.185	.144	.294	.242	.075	.197	.133
	Discriminant	–.232	.102	–.178	.196	–.248	.068	–.223	.089
	Explain	.000	.998	–.006	.964	–.105	.429	–.060	.640
Statistics	R-square	.452		.474		.503		.46	
	F *p* value	.003		.002		.001		.002	
	Sample size	35		35		35		35	

each factor score on profits. The results appear in Table 4. They show ample variation across methods and that using the less accurate methodological approach (discriminant) can lose almost as much money as the more accurate approaches (logit and tree) can gain. Again, for a company with 5 million subscribers, contacting 10% of them for a churn management campaign can result in hundreds of thousands of dollars. For example, a standard deviation increase in the logit approach generates $.46 × 500,000 = $230,000 in the most conservative case.

SUMMARY AND DISCUSSION

Conclusions

This research used a churn-modeling tournament to investigate the accuracy of statistical models for the prediction of customer churn. The principal findings of the study are as follows:

- *Method matters.* The differences in predictive accuracy among the tournament entries are managerially meaningful, representing hundreds of thousands of dollars in additional profits.
- *Models have staying power.* Our results suggest that the predictive ability of churn prediction models does not diminish appreciably after a period of approximately three months.
- *Model builders use distinct methodological approaches to develop churn models.* Multiple elements go into the development of a predictive model, including the estimation technique, the variable selection technique, and the allocation of time to various tasks. We identified five distinct methodological approaches that are combinations of these elements.
- *Logistic and tree approaches perform relatively well, the practical approach has average performance, and discriminant and explain approaches have the lowest performance.* These conclusions hold across two criteria (i.e., top-decile lift and Gini coefficient) and across two types of data (i.e., data collected simultaneously with the calibration data and data collected approximately three months later).

Implications

Our results have implications for both future researchers and practitioners. For researchers, we conclude the following:

- *The entire modeling approach should be considered when developing or evaluating prediction methodologies.* We find that the statistical technique is just one part of the overall approach, and the overall approach is strongly related to predictive accuracy.
- *Explanation does not mean prediction.* This is a common adage among model builders, but we illustrated it vividly in the context of churn prediction. The explain approach worked relatively poorly as a predictive tool.
- *Exploring several estimation techniques to develop one model may not pay off.* Exploring several statistical methodologies was associated with the explain approach, which did not perform well. These results might be idiosyncratic to our sample, but further analysis revealed that there was a group of entrants that explored logistic regression, rejected it in favor of discriminant analysis, and ended up with relatively poor predictive results.

Implications for practitioners include the following:

- *They should continue to search for better techniques.* This follows from our finding that method matters. If we had not found this, we could simply tell practitioners to stick with their current technique and perhaps try to make it more efficient, but method does matter, so companies should constantly test new procedures.
- *For companies starting up a predictive modeling function, logit and tree approaches are good techniques with which to*

Table 4

PROFIT IMPACT PER CUSTOMER OF CHANGES IN METHODOLOGICAL APPROACH

Methodological Approach	*CLV ($)*	*Success Rate (γ)*		
		10%	*30%*	*50%*
Logit	500	$ .46	$ .90	$1.34
	1,500	.95	2.37	3.79
	2,500	1.44	3.83	6.23
Tree	500	.30	.59	.87
	1,500	.62	1.54	2.45
	2,500	.93	2.49	4.04
Practical	500	.22	.43	.63
	1,500	.45	1.12	1.79
	2,500	.68	1.81	2.94
Discriminant	500	–.22	–.44	–.65
	1,500	–.46	–1.15	–1.83
	2,500	–.70	–1.86	–3.02
Explain	500	–.10	–.19	–.28
	1,500	–.20	–.49	–.78
	2,500	–.30	–.79	–1.28

Notes: Numbers represent the impact of a standard deviation increase in the factor score for each method on per-customer profits of a single churn management campaign. We calculate this as follows: The standardized coefficients in Table 3 represent the standard deviation change in lift per standard deviation change in the factor score. We multiplied these coefficients by .53 (the standard deviation in lift from Table 1) to yield the unit change in lift per standard deviation change in factor score. We then multiplied the unit change in lift by the profit impact per unit lift change that Equation 4 depicts (assuming that $N\alpha = 1$, to put the results on a per-customer basis). We used $\delta = \$50$ for the incentive amount, $\beta_0 = .018$ as the base churn rate (because this is the churn rate for our data), and CLV and γ as specified for each of the cells in Table 4.

begin. We admonish practitioners that the approaches mean more than just the statistical technique. For example, those investing in the tree approach should expect to spend a lot of time on model development (mostly on estimation) and should expect to use many variables in the final model.

- *Models last at least three months*. Practitioners do not need to develop a new model every month. Our results suggest at least a three-month shelf life for churn prediction models, and we recommend testing longer horizons.

Further Avenues for Investigation

A key message is that there is more to prediction than just the estimation technique. Two approaches that have received much recent attention are variable selection methods and missing-values techniques. The variable selection problem can be approached with methods of data reduction that minimize the loss of information on the dependent variable (Cook and Lee 1999; Li 1991). As for missing values, though we had limited data from our participants, there were indications that use of mean substitution along with a dummy variable to indicate missing data led to greater predictive accuracy. In general, if missing values seem random, recent advances in imputation of missing data (Schafer 1997) can be applied to create multiple, complete data sets.

In terms of the statistical technique itself, machine learning and nonparametric statistics have generated a plethora of approaches that emphasize predictive ability. Two popular approaches to overcome the curse of dimensionality are generalized additive models (Hastie and Tibshirani 1990) and multivariate adaptive regression splines (Friedman 1991). Support vector machines (Vapnik 1996) transform the raw data into a "featured space" that can classify objects using linear planes (Friedman 2003; Kecman 2001). Predictions can also be improved by combining models. The machine-learning literature on bagging, the econometric literature on the combination of forecasts, and the statistical literature on model averaging suggest that weighting the predictions from many different models can improve predictive ability. Our winning entry used the power of combining several trees to improve prediction, with each tree typically no larger than two to eight terminal nodes, through a gradient tree-boosting procedure (Friedman 2001).

Another avenue for future work is to unravel the relative contributions of variable selection, time allocations, and other elements of model approach. For example, we find that stepwise variable selection is associated with good performance, but this could be because it was often coupled with logistic regression. A related area for further research would be to disentangle the impacts of researcher and methodology. Our work suggests that these issues are intertwined (e.g., practitioners were associated with the logit approach), but again, it would be valuable to tease out their separate effects and test them in a field setting (see, e.g., Knott, Hayes, and Neslin 2002).

Another area is dynamic procedures, such as hazard models (Lu 2002). This requires a fundamentally different data setup than the one we used; we observed customers over three months during which they did not churn and then tried to predict whether they would churn one month later. However, it might be worthwhile to compile the type of database that would support dynamic models.

Further research could also zero in on the types of data that are important for churn prediction models. Although our data included behavioral, customer interaction, and demographic variables, they conained little in the way of marketing efforts. In particular, data on previous targeted offers to reduce churn would allow targeting based on *response* to churn reduction efforts rather than targeting based on who is likely to churn. The two need not be the same.

Although tournaments are used in other areas of statistical forecasting, to our knowledge, this is the first use of tournaments to study churn management, a crucial problem facing many companies that has just recently attracted the interest of researchers. Our efforts are limited by the particular data we used and by the particular set of model builders who contributed to our "meta" database. However, we learned a great deal from this first study. The managerial problem of controlling customer churn and the challenging statistical problems of predicting churn accurately should generate a fruitful line of research in this area.

REFERENCES

Alker, Hayward R., Jr. (1965), *Mathematics and Politics*. New York: The Macmillan Company.

Berry, Michael J.A. and Gordon S. Linoff (2000), *Mastering Data Mining: The Art and Science of Customer Relationship Management*. New York: John Wiley & Sons.

Cook, R. Dennis and Hakbae Lee (1999), "Dimension Reduction in Binary Response Regression," *Journal of the American Statistical Association*, 94 (December), 1187–1200.

Farley, John U. and Don R. Lehmann (1986), *Generalizing About Market Response Models: Meta-Analysis in Marketing*. Lexington, MA: Lexington Books.

Fitchard, Kevin (2002), "Standing by Your Carrier," *Telephony Online*, (March 18), (accessed December 16, 2005), [available at http://telephonyonline.com/mag/telecom_standing_carrier/index.html].

Friedman, J.H. (1991), "Multivariate Adaptive Regression Splines," *Annals of Statistics*, 19 (1), 1–67 (see also Discussion and Rejoinder, 67–141).

——— (2001), "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics*, 29 (4), 1189–1232.

——— (2003), "Recent Advances in Predictive (Machine) Learning," working paper, Department of Statistics, Stanford University.

Ghosh, Avijit, Scott A. Neslin, and Robert W. Shoemaker (1984), "A Comparison of Market Share Models and Estimation Procedures," *Journal of Marketing Research*, 21 (May), 202–210.

Hastie, T.J. and R.J. Tibshirani (1990), *Generalized Additive Models*. New York: Chapman and Hall.

Helsen, Kristiaan and David C. Schmittlein (1993), "Analyzing Duration Times in Marketing: Evidence for the Effectiveness of Hazard Rate Models," *Marketing Science*, 11 (4), 395–409.

Kecman, Vojislav (2001), *Learning and Soft Computing: Support Vector Machines, Neural Networks and Fuzzy Logic Models*. Cambridge, MA: MIT Press.

King, Gary and Langche Zeng (2001), "Logistic Regression in Rare Events Data," *Political Analysis*, 9 (2), 137–63.

Knott, Aaron I., Andrew Hayes, and Scott A. Neslin (2002), "Next-Product-to-Buy Models for Cross-Selling Applications," *Journal of Interactive Marketing*, 16 (3), 59–75.

Kumar, Akhil, Vithala R. Rao, and Harsh Soni (1995), "An Empirical Comparison of Neural Network and Logistic Regression Models," *Marketing Letters*, 6 (4), 251–64.

Levin, Nissan and Jacob Zahavi (1998), "Continuous Predictive Modeling: A Comparative Analysis," *Journal of Interactive Marketing*, 12 (2), 5–22.

——— and ——— (2001), "Predictive Modeling Using Segmentation," *Journal of Interactive Marketing*, 15 (2), 2–22.

Li, Ker-Chau (1991), "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association*, 86 (June), 316–42.

Lu, Junxiang (2002), "Predicting Customer Churn in the Telecommunications Industry: An Application of Survival Analysis Modeling Using SAS," *SAS User Group International (SUGI27) Online Proceedings*, Paper No. 114-27, (accessed December 21, 2005), [available at http://www2.sas.com/proceedings/sugi27/p114-27.pdf].

Rud, Olivia Parr (2001), *Data Mining Cookbook: Modeling Data for Marketing, Risk, and Customer Relationship Management*. New York: John Wiley & Sons.

Schafer, J.L. (1997), *Analysis of Incomplete Multivariate Data*. New York: Chapman and Hall.

Statistics.com (2002), Definition of "Gini Coefficient," (accessed December 16, 2005), [available at http://www.statistics.com/content/glossary/g/gini.php].

Vapnik, V. (1996), *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.

Wireless Review (2000), "They Love Me, They Love Me Not," (November 1), 38–42.

Yang, Catherine (2003), "AOL: Scrambling to Halt the Exodus," *BusinessWeek*, (August 4), 62.

Zahavi, Jacob and Nissan Levin (1997), "Applying Neural Computing to Target Marketing," *Journal of Direct Marketing,* 11 (Fall), 76–93.