Defection Detection: Improving Predictive Accuracy of Customer Churn Models

Scott A. Neslin
Dartmouth College

Sunil Gupta
Columbia University

Wagner Kamakura
Duke University

Junxiang Lu
Comerica Bank

Charlotte Mason
University of North Carolina

February 7, 2004

The authors express their appreciation 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 to an anonymous wireless telephone carrier that provided the data for this study. We also thank participants in the Tuck School of Business, Dartmouth College, Marketing Workshop for comments.

Defection Detection: Improving Predictive Accuracy of Customer Churn Models

Abstract

This paper investigates how methodological factors contribute to the accuracy of customer churn predictive models. The study is based on a tournament in which researchers from business and academia downloaded data from a publicly accessible website, estimated a churn prediction model on that data, and made predictions on two validation databases. These predictions were merged with the actual churn records for the validation data and scored on two criteria, top-decile lift and the Gini coefficient. A total of 33 participants submitted 45 entries.

The results suggest several important findings. First, methods do matter: The differences observed in predictive accuracy across submissions could change profit contribution by $100,000's. Second, models have staying power: The churn models in our data typically have very little decrease in performance if used to predict churn for a database compiled three months after the calibration data. Third, researchers use a variety of "approaches" to develop churn models, described by a combination of estimation technique, variable selection procedure, time allocations to various steps in the model-building process, and number of variables included in the model. These approaches can be labeled "Logit," "Trees," "Novice," "Discriminant," and "Explain." We find that the Logit and Tree approaches are associated with relatively higher predictive performance, the Novice approach is associated with middle-of-the road predictive performance, while the Discriminant and Explain approaches are associated with lower predictive performance.

We discuss implications of these results for both researchers and practitioners.

Defection Detection: Improving Predictive Accuracy of Customer Churn Models

1. INTRODUCTION

Customer churn – the propensity of customers to cease doing business with a company in a given time period – has become a significant problem for many firms. These include publishing, investment services, insurance, electric utilities, health care providers, credit card providers, banking, Internet service providers, telephone service providers, online services, and cable services operators. For example, in the Internet service industry, reported annual churn rates range from 21% (Forrester 2002) to 63.2% (Network World 2001). In the wireless telephone industry, annual churn rates have been reported to range from 23.4% (Wireless Review 2000) to 46% (Telephony Online 2002). Figure 1 provides more detail for the wireless industry. The rates in Figure 1 vary by company but all companies lose at least a quarter of their customers from one year to the next. It can be shown that if no new customers are acquired then the average lifetime of an existing customer is equal to 1 / c, where c is the annual churn rate. For the company with 25% churn, that means an average lifetime of 4 years, whereas with a churn rate of 50%, the average lifetime of a customer is 2 years. Obviously, customer churn figures directly in how long a customer stays with a company, and in turn the customer's lifetime *value* to that company.

[Figure 1 Goes About Here]

There are two basic approaches to managing customer churn. Untargeted approaches rely on superior product and mass advertising to increase brand loyalty and retain customers. A good example of this is AOL's recent efforts to decrease churn

through better software and content (Business Week 2003). Targeted approaches rely on identifying customers who are likely to churn, and then either providing them with a direct incentive or customizing a service plan to stay. An example of this is to provide customers with market competitive service plans by segmenting their telecommunications calling behavior. There are two types of targeted approaches: reactive and proactive. With the reactive approach, the firm waits until the customer contacts the firm to cancel his or her account. The firm then offers the customer an incentive, e.g., a rebate, to stay. With the proactive approach, the firm tries to identify *in advance* customers who are likely to churn at some later date. The firm then targets these customers with special programs or incentives to forestall the customer from churning.

Targeted proactive programs have potential advantages of lower incentive costs (since the incentive may not have to be as high as when the customer has to be "bribed" at the last minute not to leave) and not training customers to negotiate for better deals under the threat of churning. However, these systems can be very wasteful if churn predictions are inaccurate, because then firms are wasting incentive money on customers who would have stayed anyway. The goal then is to predict customer churn as accurately as possible.

There are numerous predictive modeling techniques for predicting customer churn. These vary in terms of statistical technique (e.g., neural nets versus logistic regression), variable selection method (e.g., theory versus stepwise selection), number of variables included in the model, and time spent in total on the modeling exercise as well as how a given time budget is allocated across various tasks in the model-building

process. There are a multitude of combinations of approaches available, and it is not clear *a priori* which methods should be best, or in fact, if it even matters.

The purpose of this paper 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 be used to predict customer churn at time $t+x$, where x is some later time period, on the order say of a few months? Or do churn models have to be re-estimated on a monthly basis?
- 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 most likely to be successful; which are less likely to be successful?

We answer these questions through an analysis of model predictions on real data, collected via a "churn modeling tournament." The tournament was administered through the Teradata Center for Customer Relationship Management, located at the Fuqua School of Business, Duke University. Data were publicly provided to all model-builders interested in participating. Each participant estimated a churn prediction model based on the data, and then used the model to generate predictions for a validation data set. The tournament administrators (the co-authors of this article) then assessed the predictive accuracy of these submissions. We also collected data from each submission on the

methodologies used to construct the model used for that submission. We were then able to conduct a *post hoc* "meta-analysis" of the results to answer our research questions.

The benefits to the tournament approach are scale, generalizability, and insight. The effort encompassed in the total number of submissions far exceeds what we as co-authors could have done with available time and knowledge resources. Therefore, the scale of the study is elevated by pooling the resources of many participants. The generalizability of our results to the real world is enhanced by the diversity of participants – from both academia and industry – and the variety of educational backgrounds and methodological approaches they represent. These by definition represent the approaches researchers use in the real world. Finally, the tournament allowed us to study certain factors that would have been impossible to study if we had created all the predictive models ourselves. A good example is how one selects variables to be in the model. As model-builders, we have our own "bias" on how to accomplish this crucial step. By drawing on a variety of researchers, we have some who emphasize theory, some who emphasize common sense, some who emphasize customer behavior, some who use automatic stepwise selection, and some who emphasize variable simplification, and several combinations of the above. Given the competitive nature of the tournament it is reasonable to assume that each entry to the tournament is the best out of a variety of models tried by the participant. It would be very difficult to achieve the insights this diversity and effort provide if we had done the analysis ourselves.

The disadvantage of the tournament approach is the lack of control (e.g., see Kumar, Rao, and Soni 1995). We could not assign certain researchers to analyze the data in particular ways. We do not have a factorial design, and this will make it difficult to

tease out certain factors. However, the scale, generalizability, and insight generated by the tournament provide a compelling set of advantages that we believe counter-balance the lack of control. We hope to demonstrate this with our results, and will discuss how this study can help others define issues to be studied on a controlled basis.

The paper proceeds as follows. We first present a profitability analysis that shows the potential dollar benefits of increased accuracy. The numbers we compute here provide a benchmark for answering our first question, as to whether the differences in predictive accuracy we see in our data are economically meaningful. Second, we discuss the tournament structure. Then we report our results. Finally we summarize and discuss the implications of our findings both for future research efforts and for practice.

2. THE IMPORTANCE OF PREDICTIVE ACCURACY

One way to quantify the accuracy of a predictive model is by calculating its lift, e.g., among say the 10% of customers predicted as most likely to churn, what percentage of them actually do, relative to the percentage of all customers who churn. The higher the lift, the more accurate the model, and intuitively, the more profitable a targeted proactive churn management program will be. In this section, we formulate profitability as a function of lift, and perform calculations to assess the potential value of increased lift in practice. We first define the following terms:

N = Total number of customers.

α = Fraction of customers who are targeted for the churn management program.

β = Fraction of targeted customers who are in fact would-be churners.

δ = Cost of the customer incentive to the firm. E.g., if the company offers customers a $50 rebate, the cost is $50.

γ = Fraction of targeted would-be churners who are wooed back to the company by the incentive, i.e., the success rate of the incentive.

c = Cost of contacting a customer to offer him or her the incentive.

LVC = Lifetime value of the customer, i.e., the value of the customer to the firm if the customer is retained.

Π = Profit contributed by the churn management program.

Given these definitions, the profit contributed by the churn management program is:

$$\Pi = N\alpha\{ \beta\gamma(LVC-c-\delta) + \beta(1-\gamma)(-c) + (1-\beta)(-c-\delta) \} \quad (1)$$

The first term within the brackets reflects profit contribution among the $\beta\gamma$ fraction of contacted customers who are would-be churners and decide based on the incentive to stay with the company. The firm retrieves their lifetime value at a cost of $c+\delta$. The second term within the brackets reflects profit contribution among the $\beta(1-\gamma)$ fraction of contacted would-be churners who do not accept the offer and leave the firm. The loss from these customers is *c*, since they do not accept the offer. The third term within the brackets reflects profit contribution among the $(1-\beta)$ fraction of contacted customers who are not would-be churners. We assume these customers accept the offer and cost the company $c+\delta$. These customers represent the wasted money for the firm. They were not going to churn yet the firm spent incentive money on them.[1]

[1]There are two ways to complicate the calculation to reflect ways in which firms might avoid this waste. First, one could assume that only a fraction of these customers will accept the offer. In the case of a rebate, it is safe to say that fraction is 100%, which is what we assume. But other types of incentive, e.g., a discounted cell phone, may not be accepted by all customers. Second, the customer may be "delighted" by

Note this formulation takes as given the percentage of the customer base to be contacted (α). Implicitly, we are only concerned with the "false positives" represented by 1-β. In general, the firm might also be concerned with the percentage of customers not contacted (1-α) who are churners and might have been retained (the "false negatives"). There is undoubtedly an inverse relationship between α and β, i.e., in contacting more customers, we are reaching into the lower deciles where a smaller percentage of customers are would-be churners. To find an optimal α, one would have to recognize that larger α will decrease the false negatives (1-α) but increase the false positives (1-β would get larger). In our example, we will assume α is specified at 10% - the company has decided, perhaps due to budget constraints, to contact 10% of its customer base – its top decile of predicted churners.

The term β reflects the accuracy of the model, and is related to the concept of lift as follows. Let:

β_0	=	The fraction of all the firm's customers who will churn.
λ	=	"Lift" from the predictive model, i.e., how much more likely the contacted group of customers is to churn relative to all the firm's customers. λ=1 would mean that the model provides essentially no predictive power, since the targeted customers are no more likely to churn than the population as a whole. λ=2 means that the targeted customers are twice as likely to churn as the population as a whole. In our numerical example, we will use α=.10, so λ is "top-decile lift".

Then, we can express β as:

---

the surprising offer, surprising in that they were quite satisfied with the firm, were not going to churn, yet were offered this unexpected bonus (Rust and Oliver 2000). One could add a "delight" factor to the model, which would represent the gains due to increased lifetime value caused by customer delight.

$$\beta = \lambda\beta_0 \quad (2)$$

Substituting equation (2) into equation (1) and re-arranging terms, we get:

$$\Pi = N\alpha\{ (\gamma LVC+\delta(1-\gamma))\beta_0\lambda - \delta - c \} \quad (3)$$

The incremental gain in profit from a unit increase in predictive accuracy "λ" is the slope of equation (3), namely:

$$GAIN = N\alpha\{ (\gamma LVC+\delta(1-\gamma))\ \beta_0 \} \quad (4)$$

Equation (4) tells us that the gains in profit arising from improved predictive accuracy of the churn predictive model depend on the following factors:

- Size of the campaign (Nα): To the extent that company has a large customer base and will contact many of them in the churn management campaign, the potential gain due to increased predictive accuracy is larger.
- Higher customer lifetime value (LVC): The purpose of the campaign is to recapture the lifetime value of would-be churners. To the extent that this value is larger, predictive accuracy is more important.
- Higher incentive cost (δ): High incentive cost means there will be a lot of money wasted if predictive accuracy is poor. Therefore, high incentive cost implies that the gains to predictive accuracy will be increased.

- Higher success rate (γ): Since we presume LVC > δ, higher success rate, i.e., contacted would-be churners are likely to be retained by the incentive, means that predictive accuracy is more important. This is because the expected benefit from finding a would-be churner (γLVC) is higher.
- Higher base churn rate (β_0): To the extent that customer churn is a big problem (higher β_0), higher predictive accuracy offers more of a gain in profits.

The above results provide interesting insights. Our purpose, however, is to attach some plausible dollar values to the gains from higher predictive accuracy. We therefore assume reasonable values for the parameters in equation (4) and calculate the gains in profit from improved lift. In doing so, we consider the "typical" wireless telecommunications company, because an anonymous firm in this industry provided the customer data for the tournament. Accordingly, we assume:

N	=	5,000,000. We assume the wireless company has 5,000,000 subscribers.
α	=	.10. We assume the company will contact 10% of its customers in the churn management campaign.
δ	=	\$50. We assume the company will offer a \$50 rebate to all targeted customers.
γ	=	.1, .3, or .5. There are no published statistics on how many would-be churners accept a \$50 offer and stay with the company. We use three values, 10%, 30%, and 50%, to provide a reasonable range.
c	=	\$0.50. We assume it costs 50 cents to contact customers to offer them the \$50 rebate. This could be done via a separate mailing piece.

β_0 = .018. The average monthly churn rate for the customers in our data is 1.80%, so we use this as the baseline churn level. Note that 1.8% monthly churn translates to $1-.982^{12} = 20\%$ annual churn.

LVC = \$1500, \$2500, \$4000. Monthly revenues per customer for the typical telecommunications company are in the \$40-\$60 range. If we use \$40, a monthly churn rate of 1.8%, and a 5% annual discount rate, the lifetime value of the customer is \$1812. A heavy user customer might spend \$100 per month, which would result in an LVC of \$4530. We use the values \$1500, \$2500, and \$4000 as reasonable numbers for our illustrative calculations.[2]

Using the above assumptions, Table 1 displays the gains in profit from improving lift λ by a tenth of a unit, i.e., from 2 to 2.1, etc. The table shows that even in the least favorable scenario, customer lifetime value of \$1500 and only a 10% success rate, a gain in just a tenth of a point in lift results in \$175,500 additional profit contribution. At the other extreme, at an LVC of \$4000 and a 50% success rate, a gain of a tenth of a point in lift generates \$1,822,500 additional profits. Of course the numbers depend importantly on the lifetime value of the recovered customer and the success rate in preventing him or her for churning. However, in all cases, a seemingly small increase in lift, say from 2 to 2.1, would be worth \$100,000's to the company. A gain of half a point, say from 2 to 2.5, would be worth five times that, easily in the range of \$1,000,000.

[Table 1 Goes About Here]

The calculations demonstrate that with reasonable assumptions for the company that provided the data for this tournament, it is easy to see how relatively small gains in predictive accuracy can be worth \$100,000's in increased profits. Therefore, if the entries to the tournament generate roughly the same lift, say all are close to λ=2, the differences in lift we observe, while possibly statistically significant, are not managerially

[2] Note this calculation does not include the variable cost of providing service. Telecommunications companies are saddled with huge fixed cost due to the investment they have made with infrastructure, so we assume the goal of the company is revenue growth to cover these costs.

significant. However, if the differences we observe are on the order of half a point or more, the differences are managerially very important.

3. TOURNAMENT STRUCTURE

3.1 Overview

The format of the tournament was relatively simple. The Teradata Center for Customer Relationship Management at the Fuqua School of Business, Duke University, provided data freely available at its website. The tournament was publicized through a number of vehicles targeting academics and practitioners potentially interested in predictive modeling. These vehicles included the ELMAR electronic mail network, the INFORMS Society for Marketing Science membership list, the INFORMS newsletter, FuquaNet, the Peppers and Rogers Group newsletter, and publicity provided by the Marketing Science Institute (MSI) and the Direct Marketing Educational Foundation (DMEF). Participants were asked to register for the tournament at the website and then could download the data. Also provided at the website were a description of the tournament, which provided an industry overview, an overview of the data, prediction criteria, and procedures, as well as a detailed description of the data. There were a total of four prediction criteria, and cash prizes of $2000 were awarded to the winners on each criterion.

When participants finished developing their models and had calculated predictions for the validation databases, they uploaded the predictions to the website. Participants could submit more than one set of predictions, representing different models, if they wanted. At that time, they also completed a brief survey that asked them about the

methodologies they had used in developing their model. Research assistants merged the uploaded validation data with the actual churn results for the validation data, and scored each entry in terms of the four criteria. Roughly half the entries were cross-validated by separate scoring accomplished by the co-authors of this article.

The tournament was launched on September 1, 2002, and the final day for submitted entries was January 13, 2003.

3.2 Data

The data bases are summarized in Figure 2 and Table 2. The calibration data consisted of 100,000 customers for whom there were 171 potential predictor variables. The data were compiled for a three month period and then whether or not the customers churned in the fifth month was recorded. The purpose of the one-month lag between the last observed predictors and the churn period was to consider the logistics of a telecommunications company implementing the results from a churn predictive model in a Customer Relationship Management (CRM) environment. In such an implementation it would take about one month to set up a campaign, target customers in the system, substitute the customer's predictors in the model, generate a prediction, prioritize customers, and take an appropriate action (e.g. sending them an incentive). So from a campaign effectiveness perspective, the time period one month after the last observed predictor is the crucial churn period.

[Figure 2 Goes About Here]

[Table 2 Goes About Here]

The calibration data included a churn indicator. Fifty per cent of the customers in the calibration data were churners. This is much larger than the base rate for the company, which was about 1.8%. However, the key to this exercise was to create an ordered list of customers, prioritized by how likely they are to churn. In low-incidence statistical problems, it is often preferred to over-sample "responders," or churners in this case, so as to provide the statistical techniques with enough information to profile churners versus non-churners (King and Zeng 2001).

There were two validation databases. The first, or "Current Score Data," was compiled at the same time as the calibration data. There were 51,306 customers in this database, 1.80% of whom were churners. Participants had access to the same predictor variables in this database as were available in the calibration data, but they did not have access to churn. The point was for the participants to predict churn, upload the predictions, and then we merged their predictions with the actual churn records and scored their entry.

The second validation database, the "Future Score Data," was structured the same as the Current Score Data, except it was compiled roughly three months after that database. This is to replicate the situation of a model estimated at a given point in time and then applied months later. If predictive accuracy falls off markedly for these data, it means that customer churn models have to be re-estimated frequently. However, if predictive accuracy does not fall off significantly, it means that churn models have "staying power," and can at least be used for three months before being re-estimated. The Future Score Data consisted of 100,462 customers with the same predictors available as for the calibration data. A churn indicator was of course not available, but 1.80% of

these customers were churners, in accordance with the firm's monthly baseline churn rate.

The 171 variables in the database fell into three main categories: customer behavior such as minutes of use, revenue, handset equipment, and trends in usage; company interaction data such as customer calls to the customer service center, and customer household demographics, including age, income, geographic location, and home ownership. The predictors were described in detail in a spreadsheet downloadable from the tournament website. Appendix 1 contains a full listing.

3.3 Prediction Criteria

Two prediction criteria were used for each of the two validation databases, resulting in four "contests" in all. The two prediction criteria were top decile lift and the Gini coefficient. The top-decile lift (λ) was defined as the percentage among the 10% of customers predicted to be most likely to churn who actually churned (β) divided by the baseline churn rate (β_0=.018) (see equation 2). Lift is probably the most commonly used prediction criterion in predictive modeling, and its relevance is demonstrated by the profit calculations in the previous section. Lift relates directly to profitability.

The second prediction coefficient, the Gini coefficient, has not been applied as often in database marketing, although can be calculated from the cumulative lift curve, which is commonly used. Figure 3 shows a cumulative lift curve. This tells us what cumulative percentage of churners are accounted for by the x% of customers predicted to be most likely to churn. Obviously one would want cumulative lift to be as high as possible. For example, one would like the 25% of customers predicted most likely to

churn to account for 100% of churners if possible. Note the diagonally increasing line with unit slope in Figure 3 is the random lift curve. If predictions are random, the top 25% of customers will account for 25% of churners.

[Figure 3 Goes About Here]

The goal is to generate a lift curve as separated as possible from the random lift curve. This is measured by the Gini coefficient – the area between an entry's cumulative lift curve and the random lift curve. To specify the Gini coefficient, define:

n = number of customers.

v_i = % of churners who have predicted probability of churn equal to or higher than customer *i*.

$\hat{v}_i$ = % of customers who have predicted probability of churn equal to or higher than customer *i*.

The Gini coefficient is then defined as (Alker 1965; Statistics.Com 2002):

$$GINI = \left(\frac{2}{n}\right)\sum_{i=1}^{n}(v_i - \hat{v}_i) \qquad (5)$$

The Gini coefficient can range from zero to one. Higher Gini reflects better separation between the achieved lift curve and random lift. Note that top-decile lift λ focuses on lift among the top 10% of customers, whereas Gini measures lift along the full continuum. It is plausible that a given method might be very good at identifying the top candidates for churn, but does not do as well on identifying the mid-range candidates. If that is the case, Gini and λ will be relatively uncorrelated.

4. RESULTS

4.1 Submissions

The tournament generated 44 entries from 33 participants. Half of these were practitioners and half were academics. Table 3 lists the participating organizations. They include universities, major companies with obvious interest in managing customer churn, and consulting companies.

[Table 3 Goes About Here]

Figure 4 displays various submission statistics documenting the methodologies used for model development. These data were gathered from the survey the participants completed when they uploaded their data. Figure 4a shows a variety of estimation methods, with logistic regression and decision trees being most common, but neural nets, discriminant analysis, and cluster analysis were also used. Figure 4b shows the methods participants used for the challenging task of selecting variables to be included in their models. This is measured on a 1-7 scale, where higher numbers mean that the participant relied on the method more extensively. We see that exploratory data analysis, common sense, and stepwise procedures were used most often to select variables for the model. These are very hands-on, practical approaches. Participants also used "theory," factor analysis, and cluster analysis to help select variables. A major aspect of the data is the 171 potential predictors. Factor and cluster analyses were presumably used to combine the variables into a more parsimonious set of predictors. Figure 4c shows a histogram of the number of variables included in the final model. There is an interesting bimodality here. Most entries used fewer than 80 predictors, usually less than 40. However, a few entries used more than 140 predictors. Note that while 171 predictors were provided, one

could create many more variables by including interactions and coding category variables in various ways. Figure 4d shows that more than half of participants divided the calibration data into estimation and holdout samples in the course of developing their models.

[Figure 4 Goes About Here]

Figure 4e shows the average number of hours participants spent on five steps required to develop a model: downloading, data cleaning, creating variables, estimation, and preparing prediction files. First we note that the average entry required 60 hours of work in total. Across our 44 entries, that means 2440 hours in total. So assuming a an 8-hour workday and a 5-day workweek, the authors would have had to work 61 weeks to generate the entries, and these entries would not have encompassed the rich diversity of approaches reflected in the data and illustrated in Figure 4.

4.2 Overall Performance

Table 4 displays overall performance statistics. These include descriptive statistics for each of the two prediction criteria on each of the two validation databases, and the correlations among the measures.

[Table 4 Goes About Here]

The table suggests three important conclusions:

- Entries vary significantly in predictive performance: The range in lift performance is roughly 2 units, from 1.07 to 3.01. The standard deviation of lift is roughly one half point. Given the results of our profitability calculations, these differences are managerially significant, potentially meaning additional

$100,000's in profit for the firm that uses a well-performing method, compared to the firm that uses a low-performing method.

- There is little significant fall-off in prediction between Current and Future Score Data: The average lift goes from 2.14 to 2.13, and the average Gini goes from .269 to .265 as we move from Current to Future Score Data. This means the predictive accuracy a model achieves now holds up for at least three months.
- The predictive criteria are highly correlated with each other: All the correlations are above 0.9. The correlations are higher for two different measures on the same data base as opposed to for the same measure used on different databases. For example, current lift and future lift are correlated .939, while current lift and current Gini are correlated .982. This suggests that only one criterion is needed.

4.3 Factors Determining Performance

4.3.1 Correlations

Figure 5 shows the correlation between various methodological characteristics and prediction performance. For each characteristic, there are four correlations, corresponding to the four prediction measures (current lift, future lift, current Gini, future Gini). These correlations tend to be the same, for a given methodological characteristic, mirroring our observations for Table 4.

[Figure 5 Goes About Here]

Figure 5 shows some very interesting results. Use of stepwise variable selection techniques, using more variables, spending time on estimation, and use of logistic regression are relatively highly correlated with performance. Total hours spent, sub-

dividing data into calibration and holdout, and using EDA, theory, and common sense for variable selection are somewhat correlated with performance. On the negative side, spending more time on preparing prediction files is highly negatively correlated with performance. This could reflect the entrant's inexperience with this type of exercise or with the particular software he or she used for the task.

It is tempting to spend a lot of time interpreting these results, but the various factors are highly correlated with each other. For example, use of stepwise procedures to select variables was correlated .600 with use of exploratory data analysis and .610 with use of logistic regression. What this says is that participants who used logistic regression tended to use stepwise procedures and exploratory data analysis to select variables. Therefore, it really is not appropriate to ask how each of these methodological factors contributed separately to predictive performance – they are part of one "approach" commonly used by our participants. To try to sort out say the contribution of stepwise versus the contribution of exploratory data analysis would be fruitless because of multicollinearity in the data.

Accordingly, we undertook a factor analysis of the methodological characteristics shown in Figure 5. The factors in this analysis would reveal the overall approaches used by our participants. We then relate the factor scores from this analysis to overall performance, to assess the performance of the various approaches.

4.3.2 Factor Analysis

Table 5 shows the factor loadings matrix from factor analyzing the methodological characteristics[3]. Variable definitions are provided in Appendix 2.

[Table 5 Goes About Here]

The loadings are relatively easy to interpret. The first factor we label the "Logit" approach. This is characterized by use of logistic regression for estimation, exploratory data analysis and stepwise for variable selection, and allocating relatively less time to preparing prediction files. The second factor we label the "Tree" approach. This factor is characterized by heavy reliance on decision trees for estimation, little use of any of the variable selection techniques we specified, especially exploratory data analysis and stepwise. Participants who scored high on this factor allocated a lot of their time to estimation, and less of their time to creating variables. They tended to spend relatively more time in total on the exercise, and divided the data into calibration and holdout samples. This interpretation of the "Tree" approach makes sense. The time in developing tree models is spent on estimating the model, i.e., growing and pruning the trees. As part of this process, the procedure identifies and defines predictor variables. This is a relatively time consuming process.

The third factor we call the "Novice" approach. Participants who score high on this factor do not have a particular estimation preference. They rely heavily on common sense in selecting variables for the model. They allocated a lot more time than average to downloading the data, although did not spend as much time in total on the exercise, and did not sub-divide data into calibration and holdout samples. The use of the word

[3] We used principal components analysis and Varimax rotation of the loadings. Seven Eigenvalues were greater than 1, but we selected the 5-factor solution as most interpretable and there was a noticeable fall-off in the Eigenvalue plot after five factors.

"Novice" is a bit presumptuous, but is motivated by the time spent on downloading and not sub-dividing the data.

The fourth factor we label the "Disciminant" approach. It relies heavily on discriminant analysis for model estimation and cluster analysis for selecting variables. Those who scored high on this factor tended to allocate less time on data cleaning and more time on estimation, and used a large number of variables. This is somewhat paradoxical because presumably the cluster analysis would have created a smaller group of variables, although the use of cluster analysis means simply that the participants used cluster analysis in the process of creating variables. The final variable selection may have ignored this analysis.

The fifth factor we label the "Explain" approach. This factor 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 those who score high on this factor were interested in understanding churn as much as predicting churn. This again is a subjective interpretation, but the self-reported reliance on theory suggests at least the instincts of these participants was to understand what was driving performance. Consistent with the use of factor analysis and cluster analysis, these participants tended to use fewer variables in their models.

In summary, the factor analysis suggested five general approaches to estimating customer churn: Logistic, Trees, Novice, Discriminant, and Explain. The next task is to see how higher scores on these factors relate to performance.

4.3.3 Regression Results of Performance

Table 6 displays the regressions of the each of the four performance measures against the five factor scores derived in the previous section. The R^2's are in the range of .35-.47, acceptable for cross-sectional "meta analysis" regressions of this type (Farley and Lehmann 1986), and the F-statistics are all statistically significant at conventional levels.

[Table 6 Goes About Here]

The results are consistent across the four criteria and suggest the following:

- Logit and Tree approaches are positively associated with predictive performance. Participants who scored high on these factors tended to perform significantly better than those who scored lower. Since the factors are orthogonal, this means these are two independent approaches, either of which tends to do well.
- The Novice approach tends to do acceptably well. 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 better than the Discriminant and Explain modelers.
- The Discriminant and Explain approaches tend not to do as well. Often the coefficients here are not significantly different than zero, but the signs are always negative, consistently across criteria.

5. SUMMARY AND DISCUSSION

5.1 Summary

This research has used a churn modeling tournament to investigate the accuracy of statistical models for predicting whether a customer will leave the company in a specified time period. The tournament attracted 44 entries contributed by 33 individual participants, representing both academics and practitioners. Participants spent an average of 60 hours working on each entry. The principal findings of the study are as follows:

1. *Method matters*: The differences in predictive accuracy among the tournament entries are managerially meaningful, representing $100,000's in additional profits for the firm that could achieve accuracy similar to the higher-performing entries relative to the lower-performing entries. This calculation is based on a profit calculation for a proactive customer churn management program administered by a wireless telecommunications company (a wireless carrier donated the data for the tournament).

2. *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. This conclusion assumes a data structure as in Figure 3, where predictor variables are compiled over a three month period, then we skip a month for readying the prediction and providing the incentive, and then predicting churn for the next month. The model calibrated on these data could be used roughly three months later without any loss of predictive accuracy.

3. *Model builders utilize distinct methodological "approaches" to developing churn models.* There are several elements that go into developing a predictive model. These include the estimation technique, the variable selection technique, and the time

allocated to various tasks. We found that model builders utilize methodological approaches that are combinations of these elements. We identified five such approaches: *Logistic*, where the model builder uses logistic regression for estimation and exploratory analysis and stepwise procedures for selecting variables; *Tree*, where the model builder uses decision trees for estimation, allocates a lot of time to the estimation task, and uses several variables in the final model; *Novice*, where the model builder has no estimation technique preference, relies on common sense to select variables, does not spend much time on the exercise, and does not divide data into calibration and holdout samples; *Discriminant*, where the model builder uses discriminant analysis for estimation and includes many predictors in the model; and *Explain*, where the analyst emphasizes theory and data simplification tools such as factor and cluster analyses to select variables.

4. *Logistic and Tree approaches perform relatively well, and equally so, the Novice approach has middle-of-the-road performance, and Discriminant and Explain approaches have the lowest performance.* These conclusions hold across two criteria – top-decile lift and Gini coefficient – and across two types of data – data collected simultaneously to the calibration data and data collected roughly three months later. The key conclusion is that different approaches yield different average levels of performance. It matters what methodological approach one uses to model customer churn.

5.2 Implications

The results of this study have several implications for practice and for future research. For practice, perhaps the most basic yet important conclusion is that firms should constantly be on the lookout for better prediction techniques. This is based on our

finding that method matters. It is quite possible that a new method can improve lift a few tenths of point over the models the firm is currently using. In turn, our profitability calculations suggest this can mean significantly increased levels of profit. Stated differently, it is worth spending $100,000's of dollars more to achieve nominally small (tenths of a point) but managerially meaningful increases in predictive lift.

A second implication for practice is for the firm wishing to initiate a modeling effort in predicting customer churn. Our results suggest that using a Logistic or Tree *approach* will result in that firm achieving a relatively good level of predictive ability. We italicize "approach" to emphasize that these approaches are more than the use of a particular statistical model – logistic regression or decision tree. They encompass distinct patterns on other methodological elements as well. For example, what we find to be part of the logistic approach is a strong emphasis on exploratory data analysis and stepwise. So users of this approach should make sure they emphasize calculating preliminary frequency counts and correlations so they understand the basic relationships in the data, and then making sure they have the computing power to run what may amount to fairly large stepwise analyses. Users of decision trees need to realize they must spend a lot of time on estimation, pruning and experimenting with various tree structures. Table 5 describes these approaches in more detail.

A third implication is that churn models do not have to be estimated every month. The results in this study suggest that every few months is at least sufficient. This means that rather than spending time re-estimating models using the same approach, more time can be spent exploring and testing new approaches on previously collected data.

Finally, top-decile lift appears to be a good prediction criterion for churn models, and can be assessed on validation databases collected at the same time as the calibration databases. Top-decile lift is highly correlated with the Gini coefficient, an arguably broader measure of predictive accuracy, but more cumbersome to calculate. In addition, top-decile lift can be directly related to churn program profitability, as illustrated by equations (1)-(4) above.

For researchers, one of the perhaps "eye-opening" implications of this work is the importance of modeling approach or "style". The academic literature pays much attention to statistical technique, e.g., various types of neural net or logistic models (cf. Kumar et al. 1995). Our research supports these efforts, but also points toward the need to understand methods of variable selection and the model building process. A statistical technique such as decision trees can be very successful, but we find that success occurs *if* the statistical technique is coupled with a lot of time allocated to estimating the model. A user of decision trees would presumably meet with less success if he or she allocated more time to data cleaning and creating variables rather than estimation.

Another interesting implication is that churn "explanation" is not the same as churn prediction. The Explain approach, which relies on theory and insight-generating factor and cluster analyses, did not achieve the as high predictive performance as other approaches. One immediate caveat is that we as co-authors subjectively labeled this approach as "Explain," and we believe that is consistent with our factor analysis loadings, but we did not explicitly ask a question about whether users of this approach emphasized explanation at the cost of prediction. However, the potential conflict between explanation and prediction is a perennial issue for model-builders, especially in a time-

constrained modeling environment. They have to ask themselves, "Should I spend my time trying to understand what variables are driving the analysis, or spend my time trying another model or set of variables to see if I can tweak predictive performance just a little." We should also add, on this point, that while we have emphasized predictive performance as the key criterion for this tournament, understanding is also an important criterion, as it may help in the design of the incentive or ensure that the customers selected for the incentive are consistent with the overall market segmentation strategy of the firm.

A final set of implications for researchers involves the direction for future research. First would be to research on a controlled basis the relative contributions of variable selection and time allocations to model accuracy. We find stepwise procedures are associated with good performance, but that could be because it was frequently coupled with logistic regression and exploratory data analysis. We need to tease out the separate contributions of these elements. This would provide insight not only on *which* approaches do better, but *why*. A second avenue of future research is to investigate how well the approaches identified in this study generalize to other venues. For example, is the logistic/stepwise/exploratory data analysis approach the universal way that model builders use logistic regression? Or are there other logistic *approaches*, perhaps even more powerful? A third avenue for future research is indeed to continue developing and testing additional statistical techniques. For example, we were unable to provide specific evaluations of genetic algorithms and various machine learning approaches due to their lack of representation in our data.

A fourth area for future research is the accuracy of dynamic models. The approach here would be to observe customers over a time series and predict when they will churn. This could be accomplished using a hazard model (Lu 2002). Note this requires a fundamentally different data set-up than the one used in this research, where we observed customers over a 3-month period during which they did not churn, and then tried to predict whether they would churn one month later. This is a very pragmatic approach, linked nicely to the implementation of a proactive churn management offer. However, it might be worthwhile to compile the type of database that would support dynamic models.

Future research could also tunnel in on the types of data that are important for churn prediction models. For example, while our data included behavioral, customer interaction, and demographic variables, it contained very little in the way of marketing efforts. This could be inferred through geographic variables and is reflected somewhat in the customer interaction data, but we did not have complete data on marketing efforts.

A final avenue for future research is in designing optimal targeted proactive programs. For example, does it pay to intervene earlier or later? What is the optimal contact percentage (α)? What is the impact of an accepted offer on future churn rates? Does this simply train the customer to hold out for better offers, or does it delight the customer and increase usage? An especially important area is the design of optimal "tiered" rebate structures, linked to customers' monthly revenue or profit

While tournaments are used in other areas of statistical forecasting, this is to our knowledge the first use of tournaments to study churn management, a crucial problem facing many companies and just recently attracting the interest of researchers. Of course

our efforts are limited by the particular data we used (provided by a telecommunications company) and by the particular set of model builders who contributed to our "meta" database. However, we have 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.

Appendix 1
Variable Descriptions for Tournament Database

Continuous Variables	Explanation	% Missing
ADJMOU	Billing adjusted total minutes of use over the life of the customer	0.000%
ADJQTY	Billing adjusted total number of calls over the life of the customer	0.000%
ADJREV	Billing adjusted total revenue over the life of the customer	0.000%
ATTEMPT_MEAN	Mean number of attempted calls	0.000%
ATTEMPT_RANGE	Range of number of attempted calls	0.000%
AVG3MOU	Average monthly minutes of use over the previous three months	0.000%
AVG3QTY	Average monthly number of calls over the previous three months	0.000%
AVG3REV	Average monthly revenue over the previous three months	0.000%
AVG6MOU	Average monthly minutes of use over the previous six months	2.839%
AVG6QTY	Average monthly number of calls over the previous six months	2.839%
AVG6REV	Average monthly revenue over the previous six months	2.839%
AVGMOU	Average monthly minutes of use over the life of the customer	0.000%
AVGQTY	Average monthly number of calls over the life of the customer	0.000%
AVGREV	Average monthly revenue over the life of the customer	0.000%
BLCK_DAT_MEAN	Mean number of blocked (failed) data calls	0.000%
BLCK_DAT_RANGE	Range of number of blocked (failed) data calls	0.000%
BLCK_VCE_MEAN	Mean number of blocked (failed) voice calls	0.000%
BLCK_VCE_RANGE	Range of number of blocked (failed) voice calls	0.000%
CALLFWDV_MEAN	Mean number of call forwarding calls	0.000%
CALLFWDV_RANGE	Range of number of call forwarding calls	0.000%
CALLWAIT_MEAN	Mean number of call waiting calls	0.000%
CALLWAIT_RANGE	Range of number of call waiting calls	0.000%
CC_MOU_MEAN	Mean unrounded minutes of use of customer care (see CUSTCARE_MEAN) calls	0.000%
CC_MOU_RANGE	Range of unrounded minutes of use of customer care calls	0.000%
CCRNDMOU_MEAN	Mean rounded minutes of use of customer care calls	0.000%
CCRNDMOU_RANGE	Range of rounded minutes of use of customer care calls	0.000%
CHANGE_MOU	Percentage change in monthly minutes of use vs previous three month average	0.891%
CHANGE_REV	Percentage change in monthly revenue vs previous three month average	0.891%
COMP_DAT_MEAN	Mean number of completed data calls	0.000%
COMP_DAT_RANGE	Range of number of completed data calls	0.000%
COMP_VCE_MEAN	Mean number of completed voice calls	0.000%
COMP_VCE_RANGE	Range of number of completed voice calls	0.000%
COMPLETE_MEAN	Mean number of completed calls	0.000%
COMPLETE_RANGE	Range of number of completed calls	0.000%
CUSTCARE_MEAN	Mean number of customer care calls	0.000%
CUSTCARE_RANGE	Range of number of customer care calls	0.000%
DA_MEAN	Mean number of directory assisted calls	0.357%
DA_RANGE	Range of number of directory assisted calls	0.357%
DATOVR_MEAN	Mean revenue of data overage	0.357%
DATOVR_RANGE	Range of revenue of data overage	0.357%
DROP_BLK_MEAN	Mean number of dropped or blocked calls	0.000%
DROP_BLK_RANGE	Range of number of dropped or blocked calls	0.000%
DROP_DAT_MEAN	Mean number of dropped (failed) data calls	0.000%
DROP_DAT_RANGE	Range of number of dropped (failed) data calls	0.000%
DROP_VCE_MEAN	Mean number of dropped (failed) voice calls	0.000%

DROP_VCE_RANGE	Range of number of dropped (failed) voice calls	0.000%
EQPDAYS	Number of days (age) of current equipment	0.001%
INONEMIN_MEAN	Mean number of inbound calls less than one minute	0.000%
INONEMIN_RANGE	Range of number of inbound calls less than one minute	0.000%
IWYUS_VCE_MEAN	Mean number of inbound wireless to wireless voice calls	0.000%
IWYLIS_VCE_RANGE	Range of number of inbound wireless to wireless voice calls	0.000%
MONTHS	Total number of months in service	0.000%
MOU_CDAT_MEAN	Mean unrounded minutes of use of completed data calls	0.000%
MOU_CDAT_RANGE	Range of unrounded minutes of use of completed data calls	0.000%
MOU_CVCE_MEAN	Mean unrounded minutes of use of completed voice calls	0.000%
MOU_CVCE_RANGE	Range of unrounded minutes of use of completed voice calls	0.000%
MOU_MEAN	Mean number of monthly minutes of use	0.357%
MOU_OPKD_MEAN	Mean unrounded minutes of use of off-peak data calls	0.000%
MOU_OPKD_RANGE	Range of unrounded minutes of use of off-peak data calls	0.000%
MOU_OPKV_MEAN	Mean unrounded minutes of use of off-peak voice calls	0.000%
MOU_OPKV_RANGE	Range of unrounded minutes of use of off-peak voice calls	0.000%
MOU_PEAD_MEAN	Mean unrounded minutes of use of peak data calls	0.000%
MOU_PEAD_RANGE	Range of unrounded minutes of use of peak data calls	0.000%
MOU_PEAV_MEAN	Mean unrounded minutes of use of peak voice calls	0.000%
MOU_PEAV_RANGE	Range of unrounded minutes of use of peak voice calls	0.000%
MOU_RANGE	Range of number of minutes of use	0.357%
MOU_RVCE_MEAN	Mean unrounded minutes of use of received voice calls	0.000%
MOU_RVCE_RANGE	Range of unrounded minutes of use of received voice calls	0.000%
MOUIWYLISV_MEAN	Mean unrounded minutes of use of inbound wireless to wireless voice calls	0.000%
MOUIWYLISV_RANGE	Range of unrounded minutes of use of inbound wireless to wireless voice calls	0.000%
MOUOWYLISV_MEAN	Mean unrounded minutes of use of outbound wireless to wireless voice calls	0.000%
MOUOWYLISV_RANGE	Range of unrounded minutes of use of outbound wireless to wireless voice calls	0.000%
OWYLIS_VCE_MEAN	Mean number of outbound wireless to wireless voice calls	0.000%
OWYLIS_VCE_RANGE	Range of number of outbound wireless to wireless voice calls	0.000%
OPK_DAT_MEAN	Mean number of off-peak data calls	0.000%
OPK_DAT_RANGE	Range of number of off-peak data calls	0.000%
OPK_VCE_MEAN	Mean number of off-peak voice calls	0.000%
OPK_VCE_RANGE	Range of number of off-peak voice calls	0.000%
OVRMOU_MEAN	Mean overage minutes of use	0.357%
OVRMOU_RANGE	Range of overage minutes of use	0.357%
OVRREV_MEAN	Mean overage revenue	0.357%
OVRREV_RANGE	Range of overage revenue	0.357%
PEAK_DAT_MEAN	Mean number of peak data calls	0.000%
PEAK_DAT_RANGE	Range of number of peak data calls	0.000%
PEAK_VCE_MEAN	Mean number of inbound and outbound peak voice calls	0.000%
PEAK_VCE_RANGE	Range of number of inbound and outbound peak voice calls	0.000%
PLCD_DAT_MEAN	Mean number of attempted data calls placed	0.000%
PLCD_DAT_RANGE	Range of number of attempted data calls placed	0.000%
PLCD_VCE_MEAN	Mean number of attempted voice calls placed	0.000%
PLCD_VCE_RANGE	Range of number of attempted voice calls placed	0.000%
RECY_SMS_MEAN	Mean number of received SMS calls	0.000%
RECV_SMS_RANGE	Range of number of received SMS calls	0.000%
RECV_VCE_MEAN	Mean number of received voice calls	0.000%
RECV_VCE_RANGE	Range of number of received voice calls	0.000%
RETDAYS	Number of days since last retention call	96.017%

REV_MEAN	Mean monthly revenue (charge amount)	0.357%
REV_RANGE	Range of revenue (charge amount)	0.357%
RMCALLS	Total number of roaming calls	85.777%
RMMOU	Total minutes of use of roaming calls	85.777%
RMREV	Total revenue of roaming calls	85.777%
ROAM_MEAN	Mean number of roaming calls	0.357%
ROAM_RANGE	Range of number of roaming calls	0.357%
THREEWAY_MEAN	Mean number of three way calls	0.000%
THREEWAY_RANGE	Range of number of three way calls	0.000%
TOTCALLS	Total number of calls over the life of the customer	0.000%
TOTMOU	Total minutes of use over the life of the customer	0.000%
TOTMRC_MEAN	Mean total monthly recurring charge	0.357%
TOTMRC_RANGE	Range of total monthly recurring charge	0.357%
TOTREV	Total revenue	0.000%
UNAN_DAT_MEAN	Mean number of unanswered data calls	0.000%
UNAN_DAT_RANGE	Range of number of unanswered data calls	0.000%
UNAN_VCE_MEAN	Mean number of unanswered voice calls	0.000%
UNAN_VCE_RANGE	Range of number of unanswered voice calls	0.000%
VCEOVR_MEAN	Mean revenue of voice overage	0.357%
VCEOVR_RANGE	Range of revenue of voice overage	0.357%
Category Variables	**Explanation**	**% Missing**
ACTVSUBS	Number of active subscribers in household	0.000%
ADULTS	Number of adults in household	23.019%
AGE1	Age of first household member	1.732%
AGE2	Age of second household member	1.732%
AREA	Geographic area	0.040%
ASL_FLAG	Account spending limit	0.000%
CAR_BUY	New or used car buyer	1.732%
CARTYPE	Dominant vehicle lifestyle	68.412%
CHILDREN	Children present in household	65.928%
CHURN	Instance of churn between 31-60 days after observation date	0.000%
CRCLSCOD	Credit class code	0.000%
CREDITCD	Credit card indicator	1.732%
CRTCOUNT	Adjustments made to credit rating of individual	96.500%
CSA	Communications local service area	0.000%
DIV_TYPE	Division type code	81.459%
DUALBAND	Dualband	0.001%
DWLLSIZE	Dwelling size	38.308%
DWLLTYPE	Dwelling unit type	31.909%
EDUC1	Education of first household member	86.478%
ETHNIC	Ethnicity roll-up code	1.732%
FORGNTVL	Foreign travel dummy variable	1.732%
HND_PRICE	Current handset price	0.847%
HHSTATIN	Premier household status indicator	37.923%
HND_WEBCAP	Handset web capability	0.001%
INCOME	Estimated income	25.436%
INFOBASE	InfoBase match	22.079%
KID0_2	Child 0 - 2 years of age in household	1.732%
KID3_5	Child 3 - 5 years of age in household	1.732%
KID6_10	Child 6 - 10 years of age in household	1.732%

KID11_15	Child 11 - 15 years of age in household	1.732%
KID16_17	Child 16 - 17 years of age in household	1.732%
LAST_SWAP	Date of last phone swap	58.000%
LOR	Length of residence	30.190%
MAILFLAG	DMA: Do not mail flag	98.523%
MAILORDR	Mail order buyer	64.363%
MAILRESP	Mail responder	62.889%
MARITAL	Marital status	1.732%
MODELS	Number of models issued	0.001%
MTRCYCLE	Motorcycle indicator	1.732%
NEW_CELL	New cell phone user	0.000%
NUMBCARS	Known number of vehicles	49.366%
OCCU1	Occupation of first household member	73.353%
OWNRENT	Home owner/renter status	33.706%
PCOWNER	PC owner dummy variable	81.534%
PHONES	Number of handsets issued	0.001%
PRE_HND_PRICE	Previous handset price	57.515%
PRIZM_SOCIAL_ONE	Social group letter only	7.388%
PROPTYPE	Property type detail	71.788%
REF_QTY	Total number of referrals	95.545%
REFURB_NEW	Handset: refurbished or new	0.001%
RV	RV indicator	1.732%
SOLFLAG	Infobase no phone solicitation flag	98.039%
TOT_ACPT	Total offers accepted from retention team	96.017%
TOT_RET	Total calls into retention team	96.017%
TRUCK	Truck indicator	1.732%
UNIQSUBS	Number of unique subscribers in the household	0.000%
WRKWOMAN	Working woman in household	87.491%

Appendix 2

Definition of Variables for Factor Analysis (Table 5)

Variable Type	Variable Name	Definition
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.
	Discrim	0-1 Indicator: 1 => used discriminant analysis in estimation.
Variable Selection	EDA	Extent to which used exploratory data analysis in variable selection (1-7 scale)
	Theory	Extent to which used theory in variable selection (1-7 scale)
	Sense	Extent to which used common sense in variable selection (1-7 scale).
	Stepwise	Extent to which used stepwise procedure in variable selection (1-7 scale).
	Factor	Extent to which used factor analysis in variable selection (1-7 scale).
	Cluster	Extent to which used cluster analysis 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 Files	Fraction of total time spent on exercise allocated to preparing prediction files.
Total Time	Total	Total time in hours spent on exercise.
Sub-Division	Sub-Divide	0-1 Indicator: 1 => divided calibration data into estimation and holdout samples.
Number of Variables	Number of Variables	Number of variables included in final model.

References

1. Alker, Hayward R., Jr. (1965) *Mathematics and Politics*, New York: The Macmillan Company.
2. *Business Week* (2003) "AOL: Scrambling to Halt the Exodus," Catherine Yang (Author), August 4, 2003, Issue 3844, p. 62.
3. Farley, John U. and Don R. Lehmann (1986), *Generalizing about Market Response Models: Meta-Analysis in Marketing*. Lexington Books, Lexington, MA.
4. Forrester Research (2002) "AOL Isn't Hemorrhaging Subscribers," *Consumer Technographics North America Brief*, by Jed Kolko with Jennifer Gordon, June 13, 2002.
5. King, Gary and Langche Zeng (2001) "Logistic Regression in Rare Events Data," *Political Analysis* 9(2), 137-63.
6. Kumar, Akhil, Vithala R. Rao, and Harsh Soni (1995) "An Empirical Comparison of Neural Network and Logistic Regression Models," *Marketing Letters*, 6 (4), 251-264.
7. 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 114-27, http://www.sas.com/usergroups/sugi/proceeedings/.
8. Network World (2001) "What the Cost of Customer Churn Means to You," November 12, 18 (46), 43.
9. Rust, Roland T., and Richard L. Oliver (2000) "Should We Delight the Customer," *Journal of the Academy of Marketing Science*, 28 (1), 86-94.
10. Statistics.Com (2002) http://www.statistics.com/content/glossary/g/gini.html.
11. Telephony Online (2002) "Standing by Your Carrier," March 18, http://currentissue.telophonyonline.com/.
12. Wireless Review (2000) "They Love Me, They Love Me Not," November 1, 17 (21), 38-42.

Table 1

Profit Gains from Increasing Predictive Performance (Lift) by One Tenth of A Point (see Equations 1-4)

		Success Rate (γ)		
		10%	30%	50%
Lifetime Value of Customer (LVC)	$1,500	$175,500	$436,500	$697,500
	$2500	$265,500	$706,500	$1,147,500
	$4000	$400,500	$1,111,500	$1,822,500

Table 2

Descriptive Statistics of Customer Data

	Calibration Data	Current Score Data	Future Score Data
Sample Size	100,000	51,306	100,462
# of Predictor Variables	171	171	171
Churn Indicator	Yes	No	No
Churn %	50%	1.80%	1.80%

Table 3

Participating Organizations

- Teradata
- Singapore Management University
- HEC / Montreal
- Analytex
- Seoul National University
- Columbia University
- Penn State University
- Utah State University
- UNIBANCO
- Stanford University
- Keleuven University
- Lingnan University
- Northwestern University
- Tel Aviv University
- AOL Deutschland
- Hudson's Bay Company
- Allant Group
- Data Reward
- University of Southern California
- UCLA
- Bank of America
- Sobeys
- Verizon Wireless

Table 4

Overall Performance

Descriptive Statistics

Criterion	Mean	Std Dev	Min	Max
Lift Current	2.14	0.53	1.07	2.9
Lift Future	2.13	0.53	1.19	3.01
Gini Current	0.269	0.1	0.06	0.41
Gini Future	0.265	0.09	0.05	0.4

Correlation Matrix

	Lift Current	Lift Future	Gini Current	Gini Future
Lift Current	1			
Lift Future	0.939	1		
Gini Current	0.982	0.929	1	
Gini Future	0.939	0.969	0.949	1

Table 5

Factor Loadings

		"Logit"	"Trees"	"Novice"	"Discrim"	"Explain"
Estimation	Logit	**0.723**	-0.307	0.240	-0.383	-0.001
	Neural	-0.665	-0.208	-0.185	-0.135	-0.151
	Tree	-0.091	**0.722**	-0.116	-0.042	-0.096
	Discrim	-0.060	-0.229	-0.041	**0.843**	0.109
Variable Sel.	EDA	**0.433**	**-0.580**	0.132	-0.041	0.014
	Theory	0.192	-0.023	0.258	0.004	**0.818**
	Sense	-0.061	-0.129	**0.604**	0.272	0.213
	Stepwise	**0.703**	-0.409	0.043	-0.101	0.160
	Factor	-0.273	-0.024	0.039	-0.209	**0.780**
	Cluster	0.284	-0.135	-0.065	**0.500**	**0.694**
Relative Time	Downloading	0.005	-0.054	**0.840**	-0.037	0.061
	Data Cleaning	0.213	-0.284	-0.404	**-0.546**	0.016
	Creating Variables	-0.124	**-0.672**	-0.021	0.056	-0.045
	Estimation	0.131	**0.754**	0.190	**0.378**	-0.064
	Preparing Prediction Files	**-0.724**	-0.296	0.204	0.056	0.244
Total Time	Total	-0.211	**0.493**	**-0.684**	-0.120	0.050
Sub-divde	Sub-Divide	-0.059	-0.233	**-0.486**	0.122	-0.046
Vars	Number of Variables	-0.068	**0.415**	0.035	**0.717**	**-0.320**

Table 6

Regression Results

		Current Lift		**Current Gini**		**Future Lift**		**Future Gini**	
		Stndrd. Coef.	P-Value	Stndrd. Coef,	P-Value	Stndrd. Coef	P-Value	Stndrd. Coef.	P-Value
Approach Factor Score	**Logit**	0.432	0.007	0.455	0.000	0.411	0.007	0.473	0.002
	Tree	0.334	0.031	0.401	0.009	0.373	0.013	0.398	0.007
	Novice	0.182	0.227	0.147	0.309	0.224	0.125	0.192	0.168
	Discriminant	-0.188	0.213	-0.141	0.331	-0.191	0.188	-0.177	0.202
	Explain	-0.064	0.668	-0.059	0.679	-0.154	0.287	-0.113	0.413
Statistics	**R-Squared**	0.37		0.41		0.42		0.46	
	F p-value	0.015		0.006		0.006		0.002	
	Sample Size	35		35		35		35	

Figure 1

Churn Rates at Wireless Telecommunications Companies

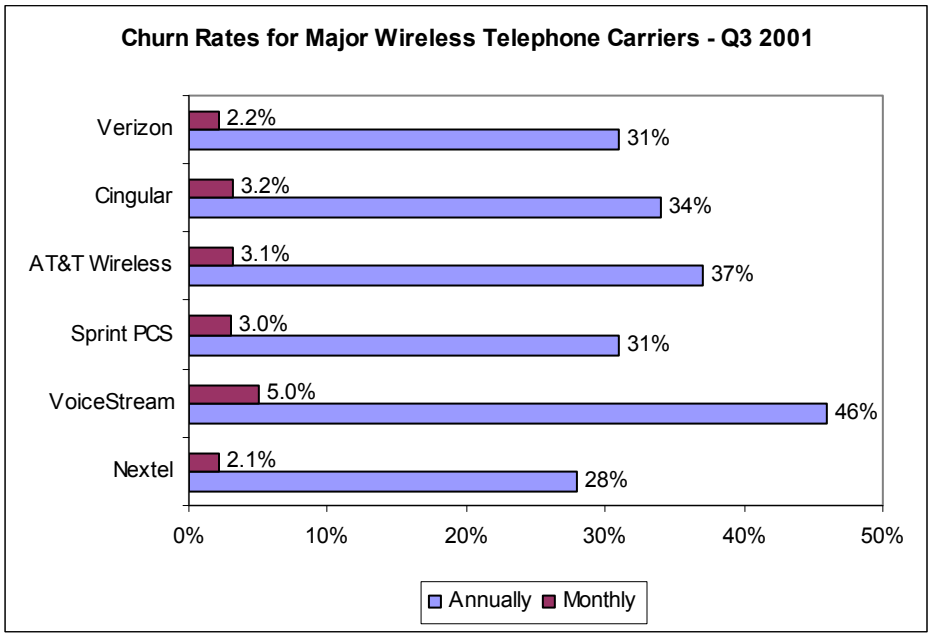

Figure 2

Database Structure

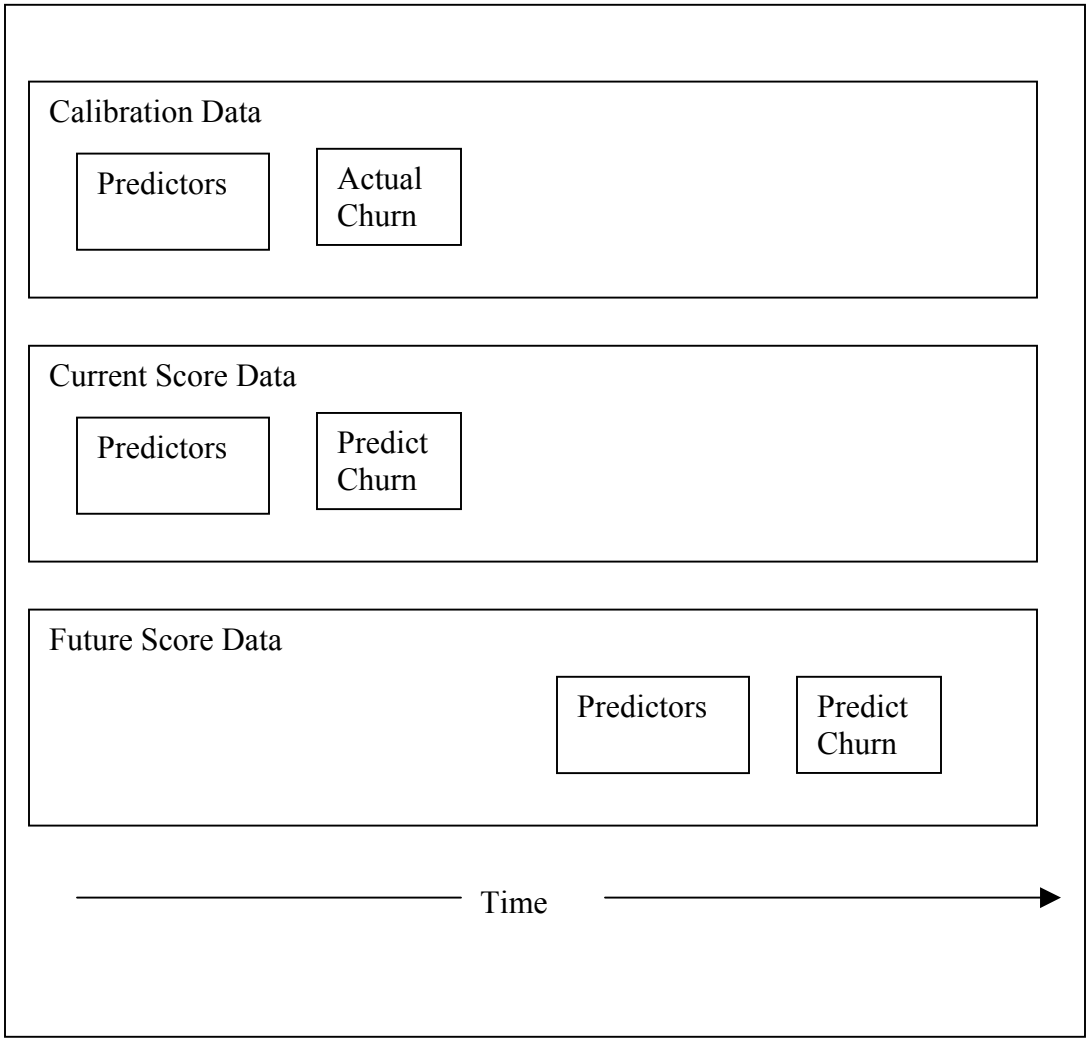

Figure 3

Cumulative Lift Chart for Calculating Gini Coefficient

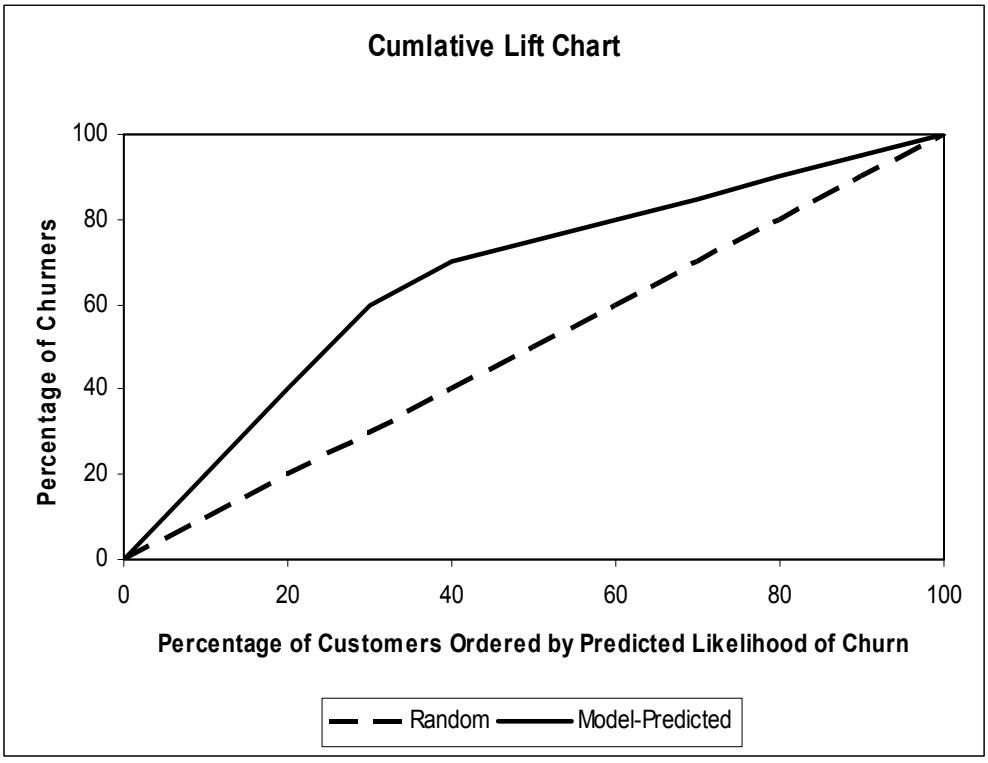

Figure 4

Submission Statistics

Figure 4a: Estimation Methods

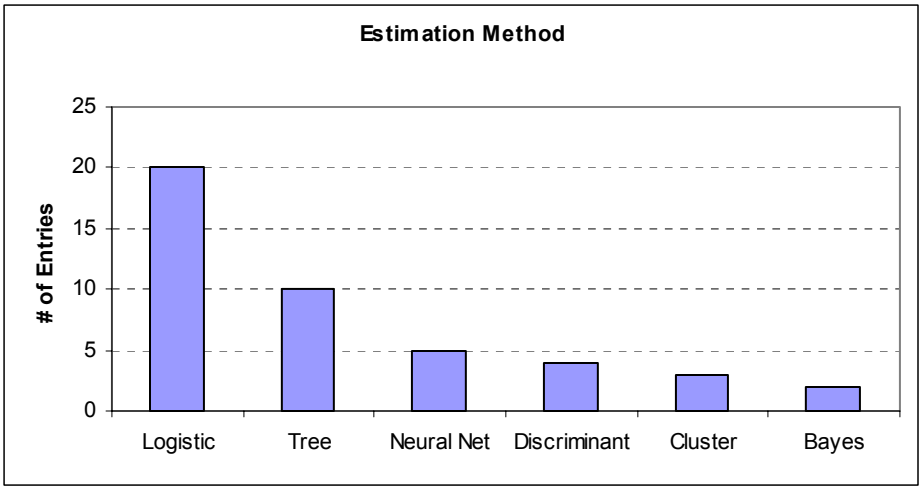

Figure 4b: Variable Selection Methods

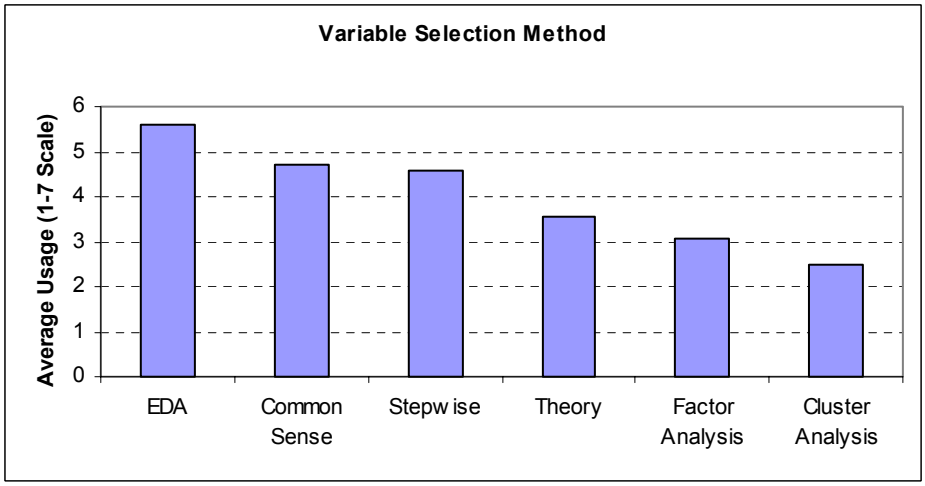

Figure 4c: Number of Variables Used in Final Model

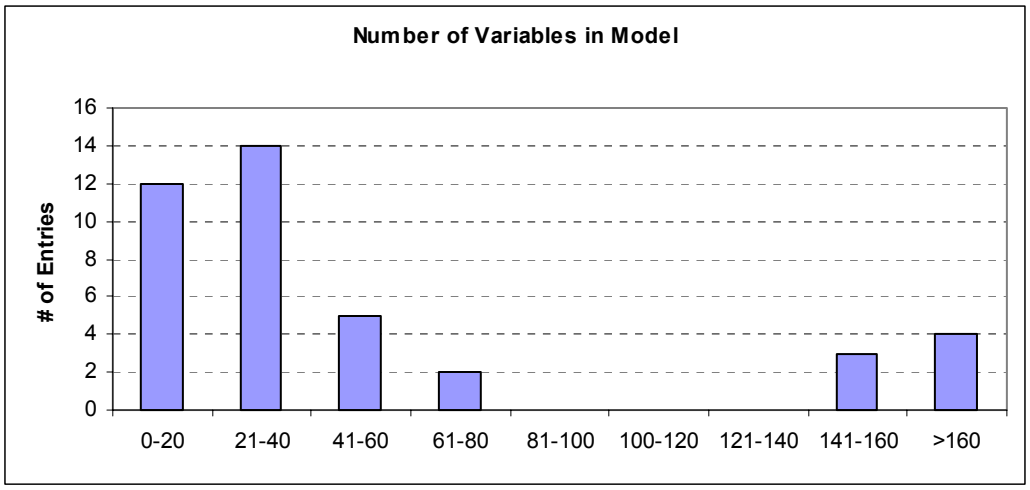

Figure 4d: Dividing Calibration Sample into Estimation and Holdout Samples

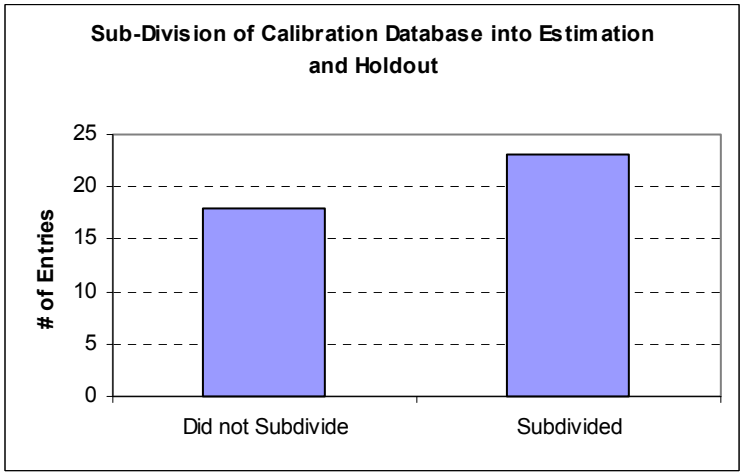

Figure 4e: Time Allocations

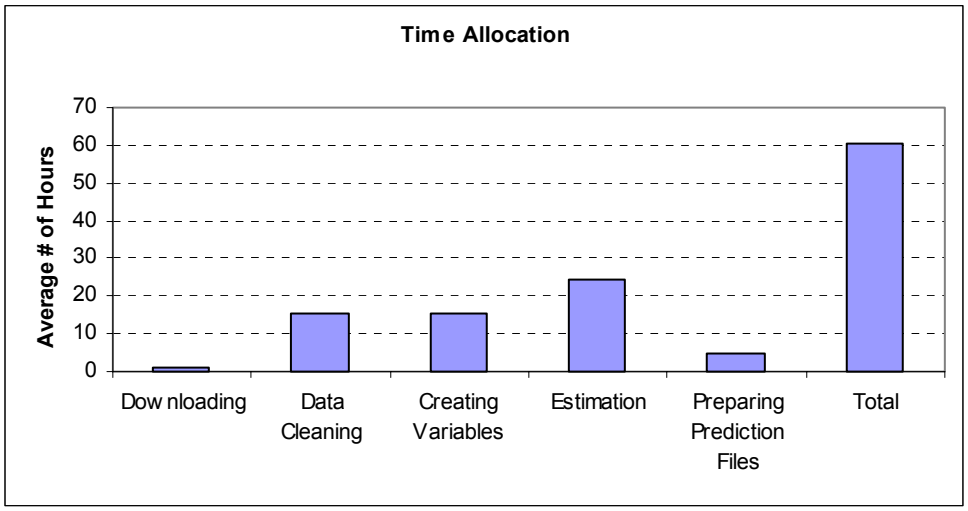

Figure 5

Correlations between Methodological Element and Prediction Performance