Churn Modeling Tournament

TERADATA CENTER FOR
CUSTOMER RELATIONSHIP MANAGEMENT
AT DUKE UNIVERSITY

DUKE
THE FUQUA SCHOOL OF BUSINESS

Scott Neslin, Director
Sanyin Siang, Managing Director

Sunil Gupta, Advisory Board
Wagner Kamakura, Advisory Board
Junxiang Lu, Advisory Board
Charlotte Mason, Advisory Board

Overview

Predictive modeling – the statistical process of "scoring" and targeting customers for a marketing campaign – is a significant database marketing tool and an important component of a firm's customer relationship management (CRM) effort. The promise of predictive modeling is the ability to predict what actions customers will take, thereby allowing firms to target their marketing efforts more effectively. One area of particular importance is customer "churn," in this case, customer voluntary churn, when current customers decide to take their business elsewhere or voluntarily terminate their service. Annual churn rates have been reported to be in the 20% - 40% range for telecommunication and other technology industries. This puts a premium on developing models that accurately predict which customers are most likely to churn, so proactive steps (e.g. appropriate communication and treatment programs) can be taken to prevent customers from churning. The purpose of the Churn Modeling Tournament is to learn which methods work best for predicting churn, thereby enhancing our overall understanding of predictive modeling.

The Teradata Center for Customer Relationship Management at Duke University (the Center) will provide data to those interested in participating in the tournament. The data consist of calibration and validation samples of customers from a major wireless telecommunications company. The calibration sample includes observed churn and a set of potential predictor variables. The two validation samples include the same predictor variables, but no churn variable. Participants will submit their predictions of likelihood to churn. We will merge those predictions with the actual churn records to evaluate predictive accuracy. The entries with the best prediction records will be the "winners." Cash prizes will be awarded.

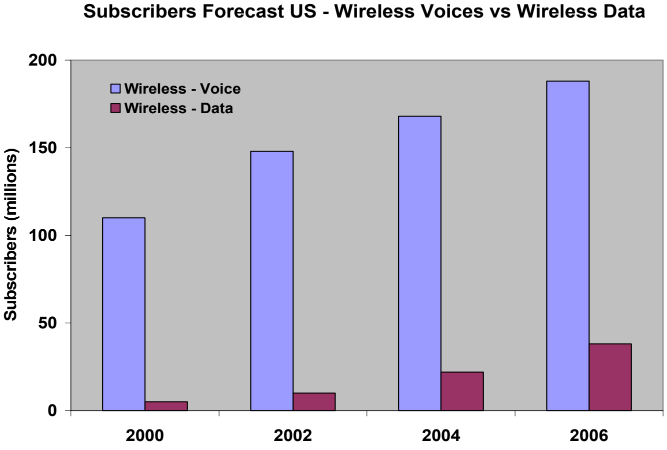
After the tournament is completed, we will conduct a "meta-analysis" of the results. We will determine which particular methodologies tend to work best and to what degree. We will make these results available to the general public and attempt to publish them in an academic, as well as a practitioner, journal. While the authors of any resulting paper will be the judges/organizers of the tournament, all entrants will be listed and acknowledged.

In summary, the tournament provides modelers with (1) the opportunity to gain recognition and win cash prizes, (2) the opportunity to participate in a collective effort that will enhance academic and practitioner knowledge of predictive modeling, and (3) the opportunity to learn first-hand about the challenges of predictive modeling in a particularly relevant context - churn.

Industry Background

The Wireless Industry*:* During the last five years, the wireless sector has been one of the fastest-growing businesses in the economy. With a unique value proposition – freedom and connectivity – the number of subscribers doubled every two years during the 90's. Wireless stocks grew as fast as those of many dot-coms, start-ups emerged everywhere, and IPO's raised record amounts of money. These events shaped the new telecommunications landscape as we know it today. And there is promise for more developments to come (*Business Week* 2002; *Wireless News Factor* 2002):

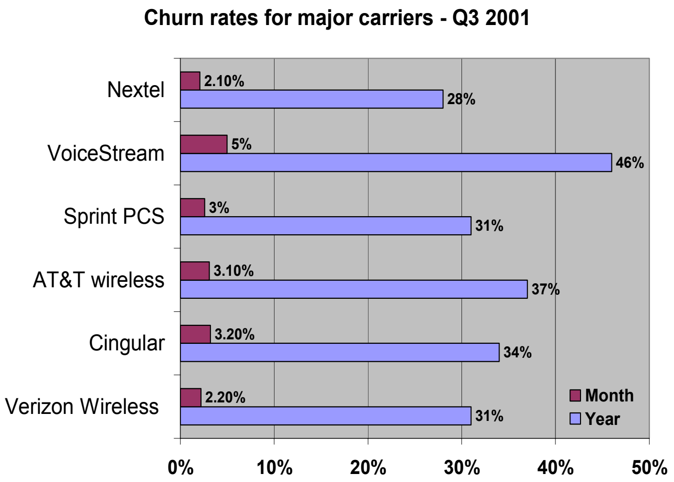
- By 2003, 25% of all telephone minutes will be accounted for by wireless services.
- By 2006, the US penetration in the wireless-voice market is expected to hit 189 million subscribers, while that of the wireless-data market is expected to jump to 38 million subscribers.
- Of all wireless customers, 70% are using digital networks that allow carriers to efficiently offer more appealing services.
- Investment in network infrastructure has increased by 17% and the number of cell sites increased by 22.3%, indicating a clear upward trend in US coverage and quality.



Source: InfoTech (2002)

***Industry Turmoil**:* Despite the vertiginous levels of growth and promise, serious charges to industry profitability have recently emerged: (a) Consolidation: From the nearly 60 cellular companies, virtually all of them are now bankrupt, bought out, or struggling with heavy debts. Only six big players now account for 80% of the wireless pie. (b) Growth: Subscriber growth rates went from 50% yearly to 15% - 20% in 2002 and analysts predict a meager 10% growth rate in 2003 (*Business Week Online, 2002*). (c) Competition: As an obvious result (and to the consumer's delight), firms engaged in a devastating price war that not only eroded revenue growth but also endangered their ability to meet their titanic debts. (d) Customer Strategy: The industry paradigm has arguably changed from one of "make big networks, get customers" to "make new services, please customers." In short, the industry has moved from an acquisition orientation to a retention orientation.

***The Elusive Customer**:* Until now, firms have been able to acquire customers without much effort. Demand for wireless services has been such that if a customer decided to drop his service and switch to another carrier, another new customer was right behind him. The priority was to maintain the customer acquisition rate high, often at the expense of customer retention. But this situation has changed. As the well of wireless subscribers has begun to run dry, churn – the customer's decision to end the relationship and switch to another company – has become a major concern. Last year the industry average churn rate was 20% - 25% annually, which translates to approximately 2% churn per month. This means that companies lose 2% of their customers every month. Third quarter, 2001, statistics show annual churn rates in an even higher range, 28% - 46% annual churn.



Source: *Telephony Online*, 2002

The reasons for the high level of churn are: (a) variety of companies, (b) the similarity of their offerings, and (c) the cheap prices of handsets. In fact, the biggest current barrier to churn – the lack of phone number portability – is likely to change in the short term. Companies are now beginning to realize just how important customer retention is. In fact, one study finds that "the top six US wireless carriers would have saved $207 million if they had retained an additional 5% of customers open to incentives but who switched plans in the past year" (Reuters 2002). Over the next five years, the industry's biggest marketing challenge will be to control churn rates by identifying those customers who are most likely to leave and taking appropriate steps to retain them. The first step therefore is predicting churn likelihood at the customer level.

Data Description

The data provided have generously been provided to the Center by a major wireless carrier. The data are organized into three data files: Calibration, Current Score Data, and Future Score Data.

	Calibration	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
Customer ID	1,000,001 – 1,100,000	2,000,001 – 2,051,306	3,000,001 – 3,100,462

The Calibration Data contain the "dependent variable" – churn – as well as several potential predictors. The Current and Future Score Data contain the predictors but not churn. Participants in the tournament will therefore estimate models on the calibration data and use these models to predict for the Current and Future Score Data.

The "Data Documentation" spreadsheet provides detailed descriptions of all the variables. The predictors include three types of variables: *behavioral data* such as minutes of use, revenue, handset equipment; *company interaction data* such as customer calls into the customer service center, and *customer household demographics*.

Customers were selected as follows: mature customers, customers who were with the company for at least six months, were sampled during July, September, November, and December of 2001. For each customer, predictor variables were calculated based on the previous four months. Churn was then calculated based on whether the customer left the company during the period 31-60 days after the customer was originally sampled. The one-month treatment lag between sampling and observed churn was for the practical concern that in any application, a few weeks would be needed to score the customer and implement any proactive actions.

The actual percentage of customers who churn in a given month is approximately 1.8%. However, churners were over sampled when creating the Calibration sample to create a roughly 50-50 split between churners and non-churners (the exact number is 49,562 churners and 50,438 non-churners). Over sampling was not undertaken in creating the Current Score and Future Score validation samples. This is to provide a more realistic predictive test. The Current Score data contain a different set of customers from the Calibration data, but selected at the same point in time. The Future Score data contain a different set of customers selected at a future point in time. One interesting aspect of the tournament will be to investigate the accuracy for same-period versus future predictive accuracy.

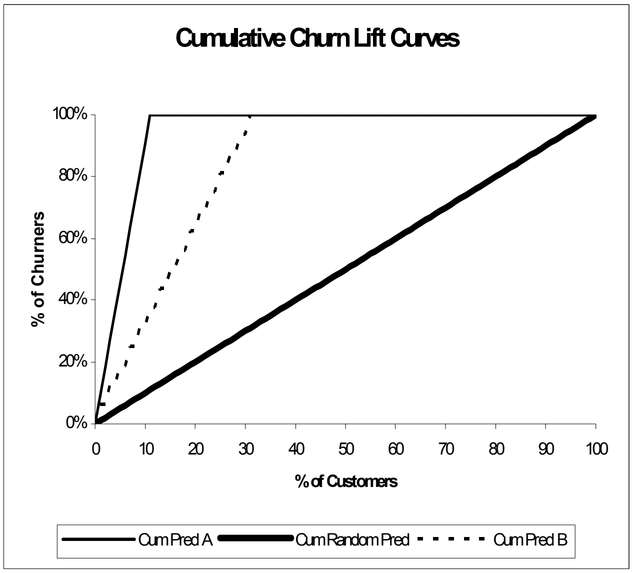
Prediction Criteria

We will calculate two measures of predictive accuracy for each submitted data file – Top Decile Lift and Gini Coefficient. Top Decile Lift measures whether the 10% of customers predicted most likely to churn actually churn. The Gini Coefficient measures predictive accuracy across the entire set of customers, not just the top 10%.

***Top Decile Lift**:* We will sort the customers in the submitted file from predicted most likely to predicted least likely to churn. We then will take the top 10% and calculate the exact percentage that did in fact churn. To create an index, we will divide by the average churn rate across all customers. So for the Current Score Data, we will find the churn rate among the 10% x 51,306 = 5131 customers predicted most likely to churn[1]. Assume for example this turns out to be 3.9%, and that the churn rate among all customers in the Current Score Data turns out to be 1.80%. Therefore, the Top Decile Lift would be 3.9/1.80 = 2.17, or "2.17 to 1" lift.

***Gini Coefficient**:* The Gini Coefficient is used in economics to measure phenomena such as income inequality (Wolff 2002; Sydsaeter and Hammond 1995). In database marketing, the Gini Coefficient works off the "Cumulative Lift Curves," shown in the figure to the right.

A Cumulative Lift Curve plots the top x% predicted customers versus the percentage of churners accounted for by these customers. For example, in the figure, the 10% of customers predicted most likely to churn by Method B account for 31.3% of all churners. The top 20% predicted customers account for 62.5% of all churners. That is better than random prediction (shown by the Cum Random line), where the top 10% would account for 10% of churners, and the top 20% would account for 20% of churners.



Generally, the higher the lift curve, the better. That is, a method predicts more accurately to the extent that the area between its cumulative lift curve and the lift curve for random prediction is large. In the figure, Method A is obviously better than Method B.

The Gini Coefficient is the area between a method's cumulative lift curve and the random lift curve. Technically, it should be calculated as an integral (Sydsaeter and Hammond 1995) but we will approximate it by a numerical measure (Alker 1965; Statistics.Com, 2002) since we have a finite number of customers and no closed form formula for the cumulative lift curve for a given method.

[1] Note the rounding. 10% of 51,306 is 5,130.6, which we round to 5131.

The formula we use is:

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

where:

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*.

v_i is the height of the method's cumulative lift curve at the i^{th} most likely predicted-to-churn customer, and $\hat{v}_i$ is the height of the random cumulative lift curve. The difference provides the "length" for calculating the area between the random and method prediction curves. The term *1/n* approximates the "width" on the x-axis. The Gini Coefficient sums these lengths-times-widths across customers, providing an approximation to the area between the method's lift curve and the random lift curve. The calculation is multiplied by "2" to ensure that the maximum possible Gini Coefficient is 1[2]. The Gini Coefficient for Method A in Figure 3 is .84; the Gini for Method B is .69. Random prediction will achieve a Gini of 0 (as seen in the formula above since for random prediction, $\hat{v}_i = v_i$) and higher Gini will correspond to more separation between the method's lift curve and random, which means better prediction.

Note that a method with a high first decile lift obviously has a head start toward achieving a high Gini Coefficient. But one method may do quite well on first decile lift but not well thereafter, and another method that doesn't do quite as well in first decile lift may do better overall.

[2] Note while this is the theoretical maximum, a Gini Coefficient exactly equal to 1 is possible with a finite data set only if there is only one churner and that churner is ranked first.

Procedures

***Eligibility**:* Any interested party is welcome to participate. We anticipate receiving entries from four main groups: academic faculty, students, model builders working in industry, and software providers. Persons affiliated with the Center are not eligible to win.

***Request for Data**:* Data are available for downloading from our website at http://faculty.fuqua.duke.edu/teradatacenter/ after the participant has registered. The data are in SAS dataset (v. 8.0) or CSV format. The data can also be placed on a CD and sent to the participant.

***Requirements for Participation**:* The following are required for participating in the tournament:

1. The participant will submit predictions for both the Current Score Data and the Future Score Data, as well as for the Calibration Data.
2. When submitting the predictions, the participant will complete a brief questionnaire that asks questions regarding the methodology.
3. The participant agrees to be interviewed if follow-up clarifications on their method are needed[3].
4. The participant gives the Center permission to use the materials submitted to the tournament in resulting publications.

***Submitting Predictions**:* The participant needs to create three files, one each for the Calibration, Current Score, and Future Score data. Each file will have two columns:

1. Customer ID: As stipulated in the original file.
2. Prediction: This can either be a rank order or a numeric score. If a rank order, we will assume that a lower rank order means more likely to churn (i.e., the customer ranked "1" is more likely to churn than the one ranked "2", etc.). If a numeric score, we will assume that a higher score means more likely to churn. In calculating top decile lift, if there are ties in the scores, i.e. two customers have the same score; the customer appearing first will be counted ahead of the subsequent customer.

In summary, the submitted Calibration prediction file will have 100,000 rows and 2 columns, the Current Score prediction file will have 51,306 rows and 2 columns, and the Future Score data will have 100,462 rows and 2 columns. Winners will be determined based on the predictive accuracy for the Current and Future Score Data, but we will use the Calibration predictions to measure out-of-sample prediction degradation.

[3] We understand that some entrants may be reluctant to discuss certain details of their methodology due to confidentiality. We will make every effort to ask reasonable questions in any follow-up.

Submissions can be uploaded onto the website under the subsection "Submit Results" in the modeling tournament section under the Current News & Events heading. The Center will also accept submissions on CDs sent to the Center address and addressed to "Churn Modeling Tournament". Please remember to include your username on all materials.

***Performance Feedback**:* Each entrant will receive his or her prediction evaluation scores for feedback purposes and so the entrant can compare with the statistics we will calculate across all entrants. This feedback will be made available after the tournament officially closes on January 1, 2003 and will be in the format of a scale from 1 – 10.

***Time Frame**:* The time frame for the tournament is August 1, 2002 through January 1, 2003. We will accept requests for data up to December 15, 2002, and will accept predictions that are received by 5:00 PM EDT, January 1, 2003.

***Governance**:* A governing board will monitor the process of the tournament to assure its integrity and to award prizes. Research assistants employed by the Teradata Center for Customer Relationship Management at Duke will calculate the predictive accuracy statistics as described above. Following are the members of the governing board:

Prof. Scott Neslin (Director), *Dartmouth College*
Sanyin Siang, *Teradata Center for Customer Relationship Management at Duke*
Prof. Sunil Gupta, *Columbia University*
Prof. Wagner Kamakura, *Duke University*
Dr. Junxiang Lu, *Industry Consultant*
Prof. Charlotte Mason, *University of North Carolina*

***Frequently Asked Questions (FAQs)**:* This description, the data, and the variable descriptions are intended to provide all information necessary for participating in the tournament. However, questions are allowed regarding the procedures, databases, and objectives. All questions and responses will be posted as FAQs on the Teradata Center website and any entrant can examine them.

***Prizes**:* For each of the four predictions (Current Score Data lift and Gini; Future Score Data lift and Gini), we will award $2000 to the entrant who receives the highest prediction score. It is possible for one entrant to win in each of the four categories. In case of a tie, the prize will be sub-divided. Prizes will be based on maximal scores. There will be no statistical testing for significant differences, etc.

Teradata Center for Customer Relationship Management at Duke University

The Teradata Center for Customer Relationship Management at Duke University advances the field of CRM through research and learning. Although various organizations exist internationally for studying CRM, the Center leverages the intellectual resources of a leading academic institution and business partnership to merge theory and practical business experience. The Center's overarching goal is to prioritize, facilitate and disseminate academic research and curriculum design aimed at advancing the field of CRM. It currently funds global CRM research, produces case studies and papers, develops curricula and provides data sets for use by other institutions and industry. The findings and offerings may influence the way corporations, students and academics view marketing.

The Center gratefully acknowledges the contributions of Emilio del Rio and Michael Kurima, who provided invaluable research assistance and data analysis in the development and organization of the tournament.

References

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

Business Week (2002) “What Ails Wireless,” April 1, 2002.

Business Week Online (2002) “Who’ll Survive the Cellular Crisis?”, February 15, 2002, http://www.businessweek.com/technology/content/feb2002/tc20020215_8884.htm.

InfoTech (2002) “Wireless Voice and Data Services Penetration,” June 18, 2002.

Reuters (2002) http://news.com.com/2100-1033-276151.html?legacy=cnet.

Statistics.Com (2002) http://www.statistics.com/content/glossary/g/gini.html.

Sydsaeter, Knut, and Peter J. Hammond (1995) *Mathematics for Economic Analysis*, Englewood Cliffs, NJ: Prentice Hall.

Telephony Online (2002) “Standing by Your Carrier”, March 19, 2002.

Wireless News Factor (2002) “Report: US Wireless Industry Flexes Muscles,” May 21, 2002.

Wolff, Edward N. (2002) “The Impact of IT Investment on Income and Wealth Inequality in the Postwar US Economy,” *Income Economics and Policy,* 14, 233-251.