Business Focused Evaluation Methods: A Case Study

Piew Datta[1]

[1]GTE Laboratories Incorporated, 40 Sylvan Rd.,
Waltham, Massachusetts USA 02451
pdatta@gte.com

Abstract. Classification accuracy or other similar metrics have long been the measures used by researchers in machine learning and data mining research to compare methods and show the usefulness of such methods. Although these metrics are essential to show the predictability of the methods, they are not sufficient. In a business setting other business processes must be taken into consideration. This paper describes additional evaluations we provided potential users of our churn prediction prototype, CHAMP, to better define the characteristics of its predictions.

1. Introduction

As data mining and machine learning techniques are moving from research algorithms to business applications, it is becoming obvious that the acceptance of data mining systems into practical business problems relies heavily on their integration in to business process. One critical aspect of building a practical and useful system is showing that the techniques can tackle the business problem. Traditionally, machine learning and data mining research areas have used classification accuracy in some form to show that the techniques can predict better than chance. The evaluation methods need to more closely resemble how the system will work if in place.

This paper focuses on the experimental evaluations we performed on a prototype called CHAMP, Churn Analysis Modeling and Prediction, developed for GTE Wireless (GTEW). CHAMP is a data mining tool used to predict which of GTEW customers will churn within the following two months. Although we were able to show that CHAMP was considerably more accurate at identifying churners than existing processes at GTEW, we needed to provide additional evaluations to persuade potential users of its benefits.

In the next section of this paper we provide some background about GTEW. Section 3 describes briefly each of CHAMP's components. Section 4 discusses several criteria we used to describe CHAMP's benefits to the GTEW marketing department. Many of these experiments are non-traditional methods used to evaluate CHAMP.

J.M. Zytkow and J. Rauch (Eds.): PKDD'99, LNAI 1704, pp. 316-322, 1999.

2. GTEW and Its Data Warehouse

GTEW provides cellular service to customers from various geographically diverse markets within the United States of America. GTEW currently has about 5 million customers in about 100 markets and is growing annually. As in all businesses, customers will sometimes terminate their service or switch providers for a variety of reasons. In the telecommunications industry this is referred to as churn. Although industry wide churn rates are only about 2% to 3% per month, this results in a considerable number of subscribers discontinuing service.

Currently GTEW accumulates information for its cellular customers in a relational data warehouse that collects data from many regional database sources. The warehouse contains over 200 fields consisting of billing and service data for each customer on a monthly basis and stores historical data going back for two years. Each month CHAMP analyzes this information to predict the possibility that any particular customer will churn based on historical data. Knowledge discovered by analyzing the characteristics of churners is used to guide marketing retention campaigns.

3. CHAMP: A Brief Overview

Members of the Knowledge Discovery in Databases project at GTE Laboratories developed CHAMP to help GTEW reduce customer churn. Since GTEW's data warehouse is updated monthly and since they have a diverse set of markets, we decided to build models monthly for each of GTEW's top 20 markets. This decision constrains CHAMP to be fully automated. Another goal was to ensure that CHAMP is scalable regarding customer records and fields. The last goal was to allow the models to be valid for a period of 60 days to allow the marketing department time to develop any desired campaigns. We developed CHAMP's overall design with these and other goals at the forefront. Readers interested in details should refer to Datta et. al. (1999)[1].

There are essentially two phases for applying data mining methods to identify churners: building models and applying models. For model building, initially the date and the market are provided to the prototype which retrieves relevant historical data from the remote data warehouse to create a local extract. The input to the model building component uses customer billing and usage information from three months previous and the dependent binary variable denoting whether the customer has churned in the previous 2 months. CHAMP's modeling method employs a hybrid of machine learning techniques. Initially we use a decision tree method (Quinlan, 1993 [2]) to rank fields according to their prediction capability and then use a cascade neural network (Fahlmann & Lebiere, 1988 [3]; Puskorius et al. 1991 [4]; Rumelhart, Hinton, & Williams, 1986 [5]) with the 30 highest ranked fields. The neural network uses a genetic algorithm (Koza, 1993 [6]) to find transformations and groupings of fields for increased model accuracy.

Once the model is built, the model is applied to current data. This data only contains customer billing and usage data and does not have customer churn information since we do not know if a customer will churn until the end of the month. The churn score generator uses the learned model and current data to produce a churn score for each customer, predicting if the customer will churn in the next 60 days.

The churn score ranging from 0 to 100 describes an individual customer's propensity to churn. Customers with a higher churn score have a higher propensity to churn.

4. Empirical Evaluations

In this section we describe the empirical evaluations we applied to better understand the characteristics of CHAMP on several differing markets. Some of these methods are applied traditionally, such as computing the lift and payoff of the learned models. Marketing professionals at GTEW suggested business oriented experiments aimed at taking some of the marketing processes currently in place and seeing how CHAMP will operate in regards to these constraints.

Generally, the data is prepared by randomly separating the entire dataset into two distinct sets: training and testing. The testing dataset is roughly 50% of the entire dataset. All experimental results are shown on the held aside testing set. We use five markets which vary considerably in size and geographic location, showing the generality of our results. We use these markets to demonstrate the performance of CHAMP across six different types of evaluation methods.

4.1 Traditional Evaluation Methods

We have validated models using both the lift[1] and payoff metrics (Datta et al., 1999 [1]; Masand et al., 1999 [7]; Masand & Piatetsky-Shapiro, 1996 [8]). An example of the lifts for different percentages of the sorted list is shown in Figure 1 for Markets 1, 2, and 3. The largest gain in lift for all three markets occurs for the first 5% to 10% as shown from the slope of the curve at these points. The first (top) decile is the first 10% of the sorted list. A lift of 1 means that the model predicts churn equal to chance and the lift eventually becomes 1 as the entire sorted list is used. These results show that CHAMP can predict churn behavior more accurately than chance.

Figure 2 shows the cumulative payoff[2] as incremental percentages of the sorted scores list are used for Markets 1, 2, and 3. The highest point in the curve where payoff is maximized varies dramatically for each market. If the customers falling to the left of the highest point are contacted this results in the highest payoff for the market. The highest payoff has a large range from $40,000 to $85,000 per month depending on the market.

4.2 Business Oriented Evaluation Experiments

In this section, we describe evaluations of CHAMP behavior of interest to marketing professionals. We typically run experiments on the first decile, top 10% of the sorted churn scores. As shown in Figure 2, this is where CHAMP has the largest lift.

[1] The prediction module produces a score for each customer and sorts customers according to score. The lift metric computes the gain in predictiveness for subsets of the sorted list over the base churn rate (i.e. churn as it is currently occurring in the market).

[2] We used a probability of 50% that a customer will continue service after being contacted, that contacting the customer costs $7 and that the customer will stay remain for 6 months. These numbers used to calculate the payoff are for illustrative purposes only and do not necessarily reflect actual numbers used in the business process.

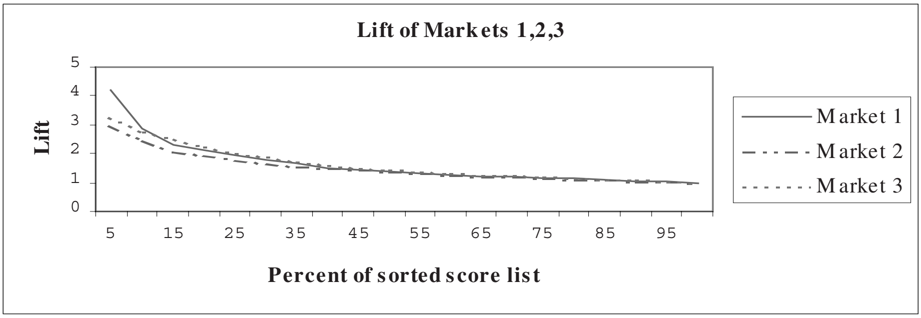

Fig. 1. Lift for Markets 1, 2, and 3. The highest gain in lift is between the top 1-10% of the sorted scores list.

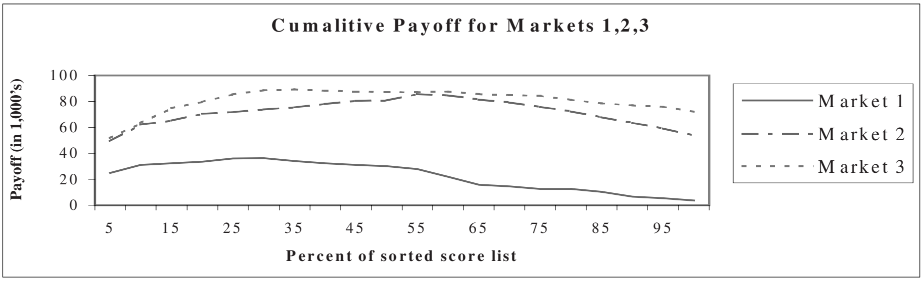

Fig. 2. Simulated cumulative payoff for Markets 1, 2, and 3. For these markets the highest payoff occurs at less than 50% of the sorted scores list.

Percentage of Churners Identified

From a marketing point of view it is important to know the percentage of the actual churners that are being captured in the highest decile, decile 1. In addition it is important to know how much lead time they will have before a customer with a higher propensity to churn will churn. We went back to our historical data and chose to look at the churners at a point in time, namely February 1998. We calculated the number of churners that appeared in decile 1 during the previous month, January 1998, and also looked at the number of customers that appeared at least once in decile 1 during the three months previous to February. Table 1 shows the results for 3 markets. CHAMP identified a fairly large percentage (28% to 36%) of churners (i.e. that is the customer appeared in the top decile at least once in the previous three months). These results also indicate that CHAMP can pick up clear signs of churning soon before customers actually churn. A smaller percentage of decile 1 customers churned in the following month (first column), although it is a larger percent than uniform for markets 2 and 3.

Table 1. Percent of churners identified by CHAMP scores for January 1998.

Market Name	Percent of churners in decile 1 in previous month	Percent of churners in decile 1 at least once in previous three months
Market 1	17%	28%
Market 2	19.5%	31%
Market 3	28%	36%

Estimated Overlap among High Propensity Churn Customers

Another aspect marketing professionals were interested in was the number of contacts they would have to make if they used CHAMP scores. There is some indication that some percentage of those that appeared in decile 1 from one month would also appear in the next month. GTEW has policies restricting the number of times a customer can be contacted for a specified period of time.

We conducted the following experiment. We identified the unique customers from decile 1 for two consecutive months, that is, if a customer appeared in decile 1 for both months, the customer was only counted once. We also conducted the same experiment for three consecutive months. Table 2 shows the results. The percentages were computed by dividing the number of unique customers for the period by the number of total customers in decile 1 for the period. For example, if we identify 50 unique customers for two months and decile 1 has a size of 30 customers a month then we would divide 50 by (30*2) and get 83%. These results show that a sizeable number of unique customers appear in decile 1 for consecutive months, showing the stability of the model. Depending on the marketing department's policy for contacting customers, they have some idea of the number of contacts they will need to make monthly.

Table 2. Percent of unique customers in decile 1 for consecutive months.

Market Name	Unique customers in 2 months	Unique customers in 3 months
Market 1	80%	69%
Market 2	75.5%	69%
Market 3	72%	59%

Aging Experiments

With dramatic changes in the cellular industry, an important issue related to modeling behavior is understanding the lifetime of the learned models and the decline in their predictive capability. In addition, it is also important to understand how long individual customer scores are valid. In this section we discuss two experiments focused on evaluating the lifetime of the models and scores.

We ran the first experiment, model aging, with models learned monthly, starting at March 1997. We created 4 different models, one for each successive month. We evaluated the models on customer data dated in June of 1997. Figure 3 shows the lift of the first decile for three markets. As can be seen on the graph, although the lift decreases slightly, there is no statistical difference when the older models are run on more recent customer data. Market 1 did have a significant drop in lift for the 3 and 4 month models. One possibility is that the data representing that time of the year did not include any major seasonal changes or new competitive offers for Markets 4 and 5 but some external factors such as new competition could have made the older models less accurate in Market 1. An experiment looking at longer delay periods between when the model is built and applied may better reflect seasonal trends.

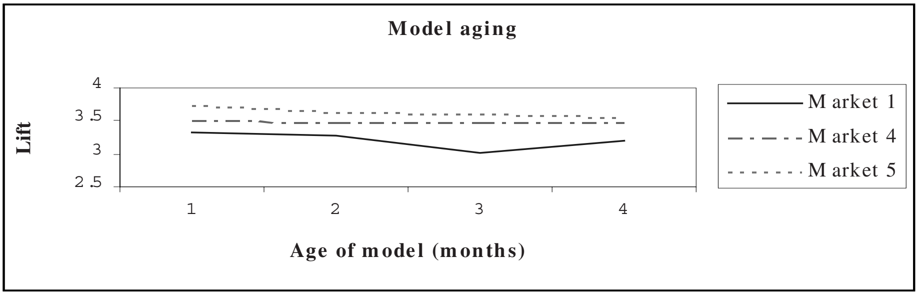

Fig. 3. Lift slowly decreases for Markets 1, 4, and 5 as the model ages.

In the second experiment we considered the lifetime of generated scores, that is when do customers in decile 1 with a high propensity to churn actually churn. To conduct this experiment, we followed a group of customers scored by CHAMP, from June 1997 until January 1998 to see whether they churned during the period and if so during which month. The results of Market 1 are illustrated in Figure 4. Decile 1 has a larger number of customers churning over the time period compared to the other deciles. In addition, in decile 1 the customers that churn tend to do so within the first few months. The lift for decile 1 in June 1997 is 4.07 which means churners concentrated in decile 1 are about four times as likely to churn when compared to the background churn rate. The lifts for July and August are 2.35 and 1.93 respectively. Decile 10 should contain the customers less likely to churn. The proof for this is shown in the lifts for decile 10 which are 0.22, 0.52, and 0.61 for June, July and August respectively. Although only about 40% of the customers in decile 1 have churned in a 6 month period, this prediction accuracy is still much higher than the percentage of background churn over the same period, about 16%-24% (assuming the industry average of about 2%-3% per month). The remaining markets have similar lift characteristics but are not shown for space considerations.

5. Summary and Discussion

The traditional lift experiments we conducted on CHAMP indicated that the learned models could predict churners in the upcoming months more effectively than current methods used by GTEW. We conducted additional experiments described in section 4.2 that focus on these questions. These experiments illustrated the benefits of using CHAMP that were not obvious to us initially. For example, in Figure 4 shows that the effectiveness of CHAMP customer scores extends over the time period that we initially built the models for, 60 days, and Figure 3 shows that models slowly decline in predictive capability over several months. These experiments not only helped explain CHAMP characteristics to users, but also the helped CHAMP developers and researchers. We expect end users of any data mining prototype or system to have a wide variety of questions regarding performance and applicability. This paper takes a first step in describing some of the questions not addressed by simple accuracy measurements.

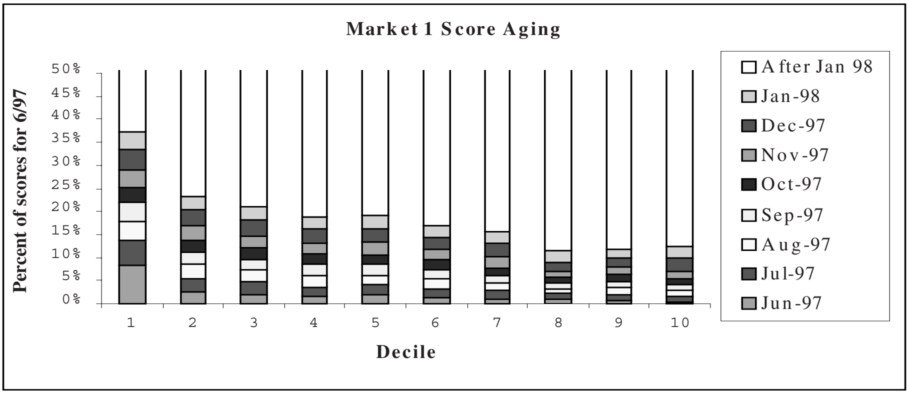

Fig. 4. Score aging results for Market 1. Those in decile 1 tend to churn at a higher rate not only for the next 2 months, but for the next 6 months. Note that the top of the decile bars have been cut off for space considerations. The bars reach 100%.

References

1. Datta, P., Masand, B., Mani, D. R. & Li, B.: Automated Cellular Modeling and Prediction on a Large Scale. Artificial Intelligence Review: Special Issue on Data Mining Applications. Kluwer Academic Publishers. To appear Oct. 1999.
2. Quinlan, J. R.: C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann (1993).
3. Fahlmann, S. E., & Lebiere, C.: The cascade-correlation learning architecture, Advances in Neural Information Processing Systems, volume 2. Morgan Kaufmann (1988).
4. Puskorius, Gint, Feldkamp, Lee: Decoupled Extended Kalman Filter Training of Feedforward Layered Networks. Proceedings of the International Joint Conference on Neural Networks, IEEE (1996).
5. Rumelhart, D. E., Hinton, G. E. & Williams, R. J.: Learning internal representations by error propagation. Parallel Distributed Processing: Explorations in the microstructure of cognition. Volume I: Foundations. Cambridge, MA: MIT Press/Bradford Books, (1986) pp 318 - 362.
6. Koza, J.: Genetic Programming. MIT Press (1993).
7. Masand, B., Datta, P., Mani, D. R. & Li, B. CHAMP: A Prototype for Automated Cellular Churn Prediction. Data Mining and Knowledge Discovery. Kluwer Academic Publishers (1999).
8. Masand, B. & Piatetsky-Shapiro, G.: A comparison of approaches for maximizing business payoff of prediction models. Proceedings of the Second International Conference on Knowledge Discovery & Data Mining. Seattle, WA. (1996). pp.195-201.