[1] Churn prediction with limited information in fixed-line telecommunication

Jiayin Qi ,Yingying Zhang, Huaying Shu , Yuanquan Li, Lei Ge
Economics and Management School,
Beijing University of Posts and Telecommunications
10 Xi Tu Cheng Road, Haidian District
Beijing，China , 100876
Office：8610-62282890；Mobile: 8610-13311326990；
Fax : 8610-62281232 ; Email：ssfqjy@263.net

***Abstract*- Based on the Comparison of three churn-prediction models, the predictive and explanatory effect of decision tree model was proved. By using the decision tree model, we found that duration of service use was the most predictive variable for churn analysis in the fixed-line service provider. At the same time, payment type, the average of several months' total calling fee, the proportion of rent fee and international IP call fee of different month, especially the latest 3 months, were also effectual predictors. With the reduction of time periods of history data, the index "churner captured proportion in top ranks" declined very little. However, the processing data size decreased dramatically. So, using relatively fewer, latest months' data to predict subscribers' churn trend would be an effective way.**

Keywords: Churn Prediction; Decision Tree Model; Neural Network; Lift Value; Captured Proportion

I. INTRODUCTION

Customer churn phenomenon becomes more and more serious in fixed-line telecommunication service industry (Shin-Yuan Hung, Hsiu-Yu Wang, 2004; Chih-Ping Wei, I-Tang Chiu, 2002; Michelle L. Hankins, 2003). The churn of subscribers causes a huge loss of fixed-line service providers (Shin-Yuan Hung, Hsiu-Yu Wang, 2004; Mattersion, 2001). In order to build customer loyalty and survive in the ever growing competitive market, a churn prediction method becomes necessary for the fixed-line service providers. However, today's researches on churn prediction in telecommunication industry mostly concentrate on mobile service field, rarely on fixed-line service field. One of the main reasons is the less amount of qualified information in the fixed-line service providers for churn prediction. Researches on churn prediction in the mobile telecom industry mainly use user demographics, contractual data, customer service logs and call patterns aggregated from call details (Chih-Ping Wei, I-Tang Chiu, 2002). However, some of the data sets above for churn prediction especially the call details are unavailable or imperfect in many fixed-line service providers. Thus, a churn predicted model with limited information is needed.

In response to the limitation of qualified information, especially the lack of call details of subscribers and the unreliability and unrepresentative of subscribers' demographic data in the fixed-line service provider, we proposed a churn-prediction model that derives limited predictors(i.e. input variables) from only subscribers' contractual information and service usage bill details in this study. An experiment result proved that limited but appropriately designed predictors still can predict the churn probability of fixed-line subscriber effectively.

II. CHURN PREDICTION

Churn denotes that a customer moves from one provider to another (Berson et al, 2000). Churn prediction is a prediction of subscribers' churn probability. Classification technologies are widely used in churn prediction, including decision tree, logit regression, and neural network and so on. The decision tree approach induces a tree-based classification model to describe relationships between variables and decision outcomes and uses the resulting decision tree for subsequent prediction purposes (Chih-Ping Wei, I-Tang Chiu, 2002); The logit regression approach induces a regression equation to reflect the relationships between the independent variables (i.e. the predictors)and the dependent logit variable (i.e. the natural logarithm of the odds of the dependent occurring or not) ; The neural network approach uses a pre-determined network topology to produce a set of adequately weighted links, which jointly differentiate individual decision outcomes, based on the respective variable values.

When a classification technique is used in churn prediction, the input variables are always the predictors

This research sponsored by National Natural Science Foundation of China (Customer value decision support theory and telecommunication industry application，project No. : 70371056); Meanwhile, sponsored by Information Management and Information Economy Key Lab of Ministry of Education of the People's Republic of China, project No,: 0607-41)
Jiayin Qi is an assistant Professor of Economics and Management School, Beijing University of Posts and Telecommunications, Beijing China, 100876 (Post Code) and her research fields are CRM and DSS etc. Email：ssfqjy@263.net.

derived from subscribes' demographic information, service contractual information, transaction record and billing details and so on. And the decision outcomes are always the values implying the status of the subscriber in the prediction period—"churner" or "non churner".

III DATA PREPARATION

In general, the services providers in China can only save the customer billing details for the latest seven months, the previous history data have to be saved offline, and it is hard to get. So the initial date is from February 2005 to August 2005 .Since the actual churn rate is considerably low and small percentage of churners' data will imperil the resulting learning effectiveness and might result in a "null" prediction system that simply predicts all subscribers as non-churners, we employed an over sampling procedure (BERRY, Michael J. A., LINOFF Gordon, 2000), resulting in a final data set of 17,223 observations, including 5,167 churners, which means the proportion of churners is about 30%.

As the demographic information of subscribers is unreliable and call details information is incomplete, we decided to abandon them. Thus only contractual data and billing details data are used in the next procedures. After cleaning the noises in the selected data sets, we then derived four categories of predictors:

I. Duration of service use

II. Payment type

III. Amount and structure of monthly service fees

We aggregated different fees into total fee, average fees over different amount months and five other fee categories: local call fee, domestic toll call fee, domestic IP call fee, international toll call fee, international IP call fee and fee of calling mobile subscribers. The structure of monthly service fees were measured by two kinds of variables, one is proportion variable, the other is consumption level rate variable. The formulas to derive these variables are as follows:

1） Proportions variables. Take proportion of local call fee as examples:
Proportion of local call fee=local call fee / total fee

2） Consumption level rate variable
Consumption level Rate=total fee/the average total fee of the same month over all observations

IV. Changes of the monthly service fees

Variables of Changes of the monthly service fees include:

1） **The growth rate of the second three months of total fee (i.e. May, June, July, 2005)** is gotten by ***(Average total fee over the second three months- Average total fee over the first three months) / Average total fee over the first three months***

2） The growth rate of the second three months of consumption level rate(i.e. May, June, July, 2005)
This variable is measured similarly to the above one, replacing all "total fee" with "consumption level rate".

3） The growth rate of the latest months of total fee (i.e. August, 2005) is measured by:
(Total fee of August, 2005- Average total fee over the second three months) / Average total fee over the first six months (**i.e. Form February, 2005 to July, 2005*****)***

4） The growth rate of the latest months of consumption level rate (i.e. August, 2005)
This variable is measured similarly to the above one, replacing all "total fee" with "consumption level rate".

5） The quantity of abnormal fluctuation of total fee is measured by:
|Total fee of August, 2005- Maximum total fee over the first six months|+|Total fee of August, 2005- Minimum total fee over the first six months|-| Maximum total fee over the first six months - Minimum total fee over the first six months|

6） The quantity of abnormal fluctuation of consumption level rate
This variable is measured similarly to the above one, replacing all "total fee" with "consumption level rate".

7） The quantity rate of abnormal fluctuation of total fee is measured by:
(|Total fee of August, 2005- Maximum total fee over the first six months|+|Total fee of August, 2005- Minimum total fee over the first six months|-| Maximum total fee over the first six months - Minimum total fee over the first six months|) / Average total fee over the first six months (**i.e. Form February, 2005 to July, 2005*****)***

8） The quantity rate of abnormal fluctuation of consumption level rate
This variable is measured similarly to the above one, replacing all "total fee" with "consumption level rate".

After derived the variables above, we then integrated them into one table, with adding a new variable "class", which with value 1 standing for churner and 0 standing for non-churner.

IV ALTERNATIVE/COMPETING MODELS SELECTION

In this chapter, we built three alternative /competing models by using the integrated data sets , then we compared these three models and get the most optical one as the churn prediction model.

A. Build three kinds of models with the technologies of Decision Tree, Neural Network and Regression

We used the Enterprise Miner of SAS System For Window V8e as our modeling tool. The details of initial

settings and structures of these three kinds of models were showed in the Appendix.

B. Select the most optimal technology for the churn prediction problem we are dealing with

In this phase, the predictive and explanatory power of three churn prediction technologies was compared: decision tree, neural network and regression. TABLE 1 below summarized the results of three churn prediction methods.

TABLE 1. THE COMPARISONS OF DECISION TREE, NEURAL NETWORK AND REGRESSION

Technology used	Captured rate in top ranks (10%)	Lift value in top ranks (10%)
Decision Tree	98.77%	9.8773
Neural Network	92.00%	9.2003
Regression	90.42%	9.0424

TABLE 1 above shows that Decision Tree model has the highest Captured rate (10%) 98.77% and LIFT value (10%) 9.8773, which means that the Decision Tree model is the most predictive model, And what's more important, the results of Decision Tree models are always easier to be understood. Thus, in this study, the Decision Tree model is selected as the optimal model for churn prediction.

V KEY PREDICTORS AND MODEL EVOLUTION

We selected decision tree technology to build the churn prediction model in the paper. The result of decision tree model shows that 6 variables can be selected as effective predictors (TABLE 2):

TABLE 2. THE EFFECTS OF THE SIX TYPES OF VARIABLES

Variable name	Importance value
Duration of service use	1.0000
Proportion of rent fee in April, 2005	0.6156
Payment type	0.5973
Average total fee over the first six months (i.e. from February to July, 2005)	0.5569
Proportion of rent fee in August, 2005	0.2088
Proportion of international IP call fee in June, 2005	0.1469

As shown in the TABLE2 above, "duration of service use" is the most predictive variable, "payment type" and several variables of amount and structure of monthly service fees are also effective predictors for churn prediction of the investigated provider's subscribers. And what's more, most of the selected variables are generated in the latest 3 months. This led us to raise such a question, "How will the prediction performance evolve when we reduce early monthly data sets for churn prediction"

Subsequently, we used Decision Tree technology to construct and run models using different number of latest monthly data as inputs. The results are summarized in TABLE3 .

TABLE 3. THE COMPARISON OF PREDICTION EFFECT WITH DIFFERENT INPUTS

Number of monthly data used	Captured rate in top ranks (10%)	Lift value in top ranks (10%)
6	98.92%	9.8923
5	98.84%	9.8842
4	98.84%	9.8842
3	98.78%	9.8779

TABLE 4. THE COMPARISON OF SIZE OF DATA SETS OF DIFFERENT NUMBER OF LATEST MONTHLY DATA

The Number of Monthly Data	Size	Decreasing Rate
6 Months	22400KB	2.18%
5 Months	18600KB	18.78%
4 Months	15300KB	33.19%
3 Months	11700KB	48.91%

As shown in TABLE 3 and TABLE 4, we can find that with the reduction of month's data for prediction, the model performance index "churner captured proportion in top ranks" declined very little. However, with the reduction of data, the data to be processed decreased dramatically. And furthermore , when the number of monthly data decreased from 5 to 4, the Captured rate in top ranks (10%) and Lift value in top ranks (10%) kept almost unchanged ,but the data size decreased sharply from 18600KB to 15300KB, so the 4 months data is enough and effective.

VI CONCLUSION AND FUTURE RESEARCH

This study focuses on the churn prediction of fixed-line service providers and tells what variables may be effectual, which can decrease the run time and the amount of data to process breathtakingly. There are several factors that affect in a negative way the quality of the prediction models. Firstly, input variables. Not all the variables have significant effect in predicting the churn, so the variable selection is necessary. Secondly, data quality. Unreliable and incomplete data will play negative role. And finally, technology for modeling. Sometimes, the combination of two kinds of basic data mining technologies can bring better results.

REFERENCES

[1]. Chih-Ping Wei , I-Tang Chiu. Turning telecommunication call details to churn prediction: a data mining approach. Expert System with Application 23, 2002, pp103-112

[2]. Shin-Yuan Hung, Hsiu-Yu Wang. Applying Data Mining to Telecom Churn Management. 8th Pacific Asia Conference on Information System, Shanghai, 2004

[3]. Mattersion, Rob. Telecom Churn Management. APDG Publishing, NC, 2001.

[4]. SAS Institute "Best Price In Churn Prediction", a SAS institute white paper, 2000

[5]. Berson, Alex, Smith, Stephen and Thearling, K. "Building Data Mining Applications for CRM", McGraw-Hill, New York, NY, 2000.

[6]. BERRY, Michael J. A., LINOFF Gordon: Mastering Data Mining: The Art and Science of Customer Relationship Management. John Wiley & Sons, 2000.
[7]. Michelle L. Hankins. Carriers Struggle to Control Churn. Billing World & OSS Today, January, 2003.

APPENDIX

TABLE 5. THE PRIME SETTINGS AND STRUCTURE OF DECISION TREE METHOD

Splitting criterion	chi-square test(significance level is 0.200)
Model assessment measure	proportion of event in top 10%
Sub-tree creation strategy	best assessment value
number of leaves nodes	15

TABLE 6. THE PRIME SETTING OF NEURAL NETWORK METHOD

Network architecture		multilayer perceptron
Input nodes		interval ,nominal and ordinal input nodes
Hidden nodes	number of neuron	15
	activation function	hyperbolic tangent
	combination function	linear-general
	bias	bias
target nodes	combination function	linear-general
	error function	multiple Bernoulli
	bias	bias

TABLE 7. THE PRIME SETTING OF REGRESSION METHOD

Model type	Logistic
Link function	Logit
Select method	Stepwise
Criteria	SBC
Optimized method	Quasi-Newton

EQUATION 1 , THE REGRESSION MODEL:

$$\text{Log}[p/(1-p)]=7.02\text{Intercept}(22.44)+0.479\text{Paytype1}(3.60)+2.56\ \text{Paytype2}\ (12.08)-1.59\text{localrate509}\ (-7.83)-0.004\text{duration1}\ (-26.20)$$

(p is the churn rate, and the data in the brackets following the variables is effect t-score)