Advanced Materials Research Vols. 989-994 (2014) pp 1517-1521
© (2014) Trans Tech Publications, Switzerland
doi:10.4028/www.scientific.net/AMR.989-994.1517

Submitted: 2014-05-25
Accepted: 2014-05-30
Online: 2014-07-16

Research on Customers Churn Prediction Model Based on Logistic

Jiang Min[1], Chu Na[2], Bi Xiaoming[3]

Department of Computer, Dalian Neusoft University of Information, China

[1]jiangmin@neusoft.edu.cn, [2]chuna@neusoft.edu.cn, [3]bixiaoming@neusoft.edu.cn

Keywords: Loss of customers; Logistic regression; Prediction model; Model validation

Abstract. At present, the competition is increasingly fierce between the securities company, whether can effectively prevent the loss of users, reducing loss rate is a difficult problem at present each securities company urgently needs to solve. The model based on the principle of data mining, proposes a prediction method based on Logistic regression algorithm. Prediction model is built based on Logistic regression algorithm and the validity and accuracy of the model is verified by experiment, provides a new method and thinking for the securities company customer churn prediction.

Introduction

With the continuous development of the securities market in China, there has been a shift in securities companies' business operations to a more centralized and tightly-controlled model. As a result of the increasingly fierce competitions and the emphasis on value-added services, customer loyalty has become a common problem faced by the securities industry. Therefore, a customer-oriented service model and more targeted marketing strategies are believed to be the key to building and retaining the competitiveness of the securities companies.

Companies that have no clearly-defined contractual agreements with their clients quite often find themselves in constant risks of loosing customers and profits. Therefore, how to find the prospective customers and successfully retain them has become the focus of many securities companies in their short-term and long-term business strategies. Trading behaviors over a period of time, causes for customer loss and other factors are analyzed to determine the appropriate strategy or means to develop and maintain customers and remain profitable.

Using the Logistic regression analysis method, this research has established an effective customer churn prediction model by analyzing historical transaction data up to six months. The aim of this research is to develop a tool that the securities companies can use to analyze trends and forecast customers that are in a higher risk of losing so that timely measures can be deployed to retain customers and ensure they stay loyal and increase profitability.

Logistic regression prediction principle

Logistic regression is an important method for the analysis of categorical data statistics. Beginning in 1980s, it has been used for event prediction, the study of non-linear relationships between the probability of an event and the influencing factors of the event, and the identification of risk factors of the incident [1].Logistic regression analysis considerably overcomes the shortcomings of the linear hypothesis method, by reducing the restrictions of overall distribution. There are no assumptions about the distribution of the independent variables,and the argument can be continuous variables, discrete variables and dummy variables. Logistic regression determines the strain response based on the model fitting and the model approach between the variables and the dependent variable. Its main purpose is to classify, and estimates the probability of events.

For dichotomous questions, assuming the individual options $y=1$ indicates that the client properly and $y=0$ indicates that the customers churn and that $x=(x_1,x_2,\cdots,x_k)$ indicates that the associated descriptive variables can be obtained in the database, and then the established mathematical expressions for the Logistic regression model is:

$$\text{logit}\left(p(y=1)\right)=\ln\left(\frac{p(y=1)}{1-p(y=1)}\right)=\beta_0+\beta_1x_1+\cdots+\beta_kx_k \quad (1)$$
$$=\beta_0+\sum_{i=1}^{k}\beta_ix_i$$

$$P\left(y=1\mid x\right)=\frac{\exp\{\beta_0+\sum_{i=1}^{k}\beta_ix_i\}}{1+\exp\{\beta_0+\sum_{i=1}^{k}\beta_ix_i\}} \quad (2)$$

$$P\left(y=0\mid x\right)=\frac{1}{1+\exp\{\beta_0+\sum_{i=1}^{k}\beta_ix_i\}}$$

Among them $\beta_i\left(i=1,\cdots,k\right)$ is coefficient of each variable after logistic regression,Its meaning is: when x_i change an unit of measurement each caused by the change in the natural logarithm; β_0 behalf Intercept. Formula（2） represents the probability of loosing certain customers. Among them $P\left(y=1\mid x\right)$ expresses the probability of a normal customer, $P\left(y=0\mid x\right)$) indicates the probability of churn.

Table 1 Users monthly data table

Month	Total commission	Trading clients	Average commission	Quality customers	Premium customers of commission
200805	4275736.27	12033	355.33	2398	1308278.04（30.60%）
200807	16469120	40214	409.54	7859	6473401.56（39.06%）
200808	8836430.38	33689	262.29	6357	4206377.16（47.60%）
200809	8043580.58	31018	259.32	5740	3938777.11（48.97%）
...	...	...	...	...	...
200810	8161183.24	30319	269.18	5392	4107470.57（50.33%）

Table 2 Partial correlation coefficient table

English name	Variable name	Correlation coefficient
LJMRYJ_LJMCYJ_H_ZSZ_4M_BL	Nearly four months total commissions and buying and selling half of the total market capitalization ratio	0.502226533
LJMC_ZSZ_4M_BL	Cumulative proportion of the total market capitalization and sold nearly 4 months	0.460734314
LJMCYJ_ZSZ_4M_BL	The cumulative proportion of the total market value and selling commission nearly four months	0.445815741
LJMC_ZSZ_2M_BL	Cumulative proportion of the total market capitalization and sold nearly 2 months	0.443599602

Logistic regressions customer churn prediction

Data preprocessing

In the data preprocessing, we will group data into two categories: customer information data and customer transactions data.

Customer information data. Customer information data refers to the customer information, such as the customer's gender, age, marital status, educational status, occupation and capital scale. Due to historical reasons, our customers' electronic documentation is incomplete, missing most of the data; only the size of the customer's financial data can be exported via the customer's transaction data. Therefore the customer information data model structure we use includes only the size of their capitals.

Customer transaction behavior data. Customer transaction behavior data includes: Changes in the position data of customers, the number of monthly transactions, turnovers, monthly commission fees, etc. Users with the monthly data are shown in Table 1.

Logistic regression model structure

Data Preparation. The prediction model established in this article uses data from a securities company over a year, in the four months from September to December statistical data was collected. The time period used to predict customer behaviors are likely to trigger future marketing response. January to March next year is forecast to month, using the model created to predict customer marketing response.

Define variables and variables derived. In order to ensure the accuracy of the model prediction, we use 91 factors securities company's customer data to make predictions. These 91 factors include asset data, transaction data, basic customer information, customer service and product information such as customer subscriptions, five aspects. Through the chain, the result was worse than the method by the 91 factors derived 275 variables.

Data preprocessing and correlation analysis. According to this study, the correlation of explanatory variables and target variable, select the 150 most relevant variables.

Logistic regression modeling

Screening AR variables on the basis of Table 2, we do logistic regression to 150 variables use SAS. Getting lost probability between the univariate regression and univariate of each customer. After the loss probability Sort, Getting the value m% is the proportion of the cumulative loss of customers top n% (n=1…100). We obtain a curve between m% and n% (as shown in Figure 1), then obtain the AR values for each variable. Choose the high value of the variable AR from them. Also, delete the experience value of less than 0.73 of the variable AR. After this step, we obtain the 45 variables.

Select the maximum likelihood estimates of large and significant variables use Logistic regression forward stepwise regression method, stepwise regression, definition the loss of value of the target variable as 0, then

$$P_{\text{outflow}}=1\text{-P}=\frac{1}{1+\exp\{\beta_0+\sum_{i=1}^{k}\beta_i x_i\}}$$

When the regression coefficient β_i is positive, variable x is the larger, P_{outflow} is smaller; Contrary, When β_i is negative,variable x is smaller,P_{outflow} is larger. Therefore, the size of the sign and the loss probability of regression coefficients showed a negative relationship. After this step, we obtained 21 variables.

Empirical Analysis

Model evaluation phase is to test the model has been established. Model checking is the important step for the data mining, there are two principles in determining the goals and data mining preparation stage, the first is, the prediction model can be assessed and must be realized, and the second is, the resulting model can be used in a real-world environment.

Select the observation period and object. Before proceeding with data collection, we define the time interval churn models examined below. On observing the object, we screened for individual customers conditions include: Customer average monthly assets between the 1,000 yuan to 10 million yuan in the observation period; Opening time from the observation period over four months; There are transactions within the observation period 4 months.

Determine the variables and regression coefficients. Performed logistic regression calculation again for the 21 variables, confirm the regression coefficients and constant coefficient of these variables. The constant coefficient of the churn model is 0.2708204106.

Estimation Results. KS (Kolmogorov-Smirnov) is mainly used to validate the ability of the model, that distinguish the lost objects, verify the ability of the model accurately distinguish normal customers and the churn. This model predicts KS value is 44.678, as shown in Table 3, indicating that the model has a strong ability to identify for the churn, so this model prediction is credible.

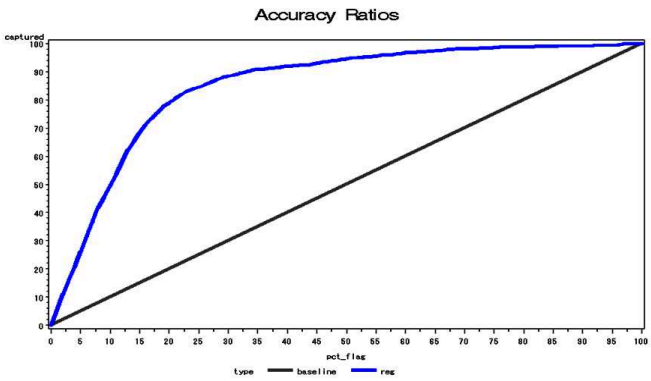

Fig. 1 AR value is 0.73

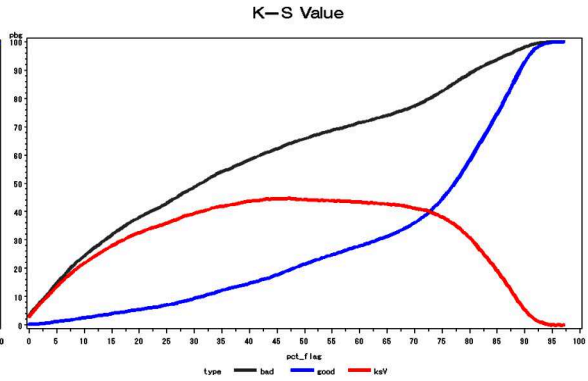

Fig. 2 Model Checking KS value

Table3 KS score sheet

Score	Cumulative target customers(CC)	Cumulative non-target customers(CNC)	Cumulative total client(CTC)	CC/ CTC	CNC/CTC	KS values
47	5812	218488	224300	63.87515112	19.19683415	44.678

Verify the correct rate and the feasibility to the model. This article randomly selected 10 customer data in the sample to test and verify substituting classification function, and then we obtain the results whether the customer has been lost, as shown in Table 4.

Table 4 Customer churn prediction analysis table

Customer ID	Var1 Assets	Var 2 Market cap	Var 3 Funds	Var 4 Positions	Var 5 Growth rate	Var 6 Than	Var Worse than	Outflow (Y)
100739135	1	0.2232	0.1713	0	8	80	6.25	1
100916315	3	0.1247	0.7521	2	1	30	-0.025	0
...	...	...	...	...	...	...	...	...
2760963	2	0.3369	0.2227	0	5	90	5.02	1

From the table, we can get the following conclusions: Customer ID100916315 variable difference ratio less than 0, that means the ratio of the loss of customers and normal customers less than 0, value of (Y) is equal to 0(loss or not), means that the customer has been lost. Contrary, the difference ratio of other nine customers is greater than 0, then Y=1, means that customers is not loss, so the conclusion of customer churn analysis by Logistic regression model are consistent with the actual situation.

Conclusion

This paper studies how to use data mining technology to establish customer churn prediction model to solve the problem of customer loss experienced by a securities firm. Real data was

collected and analyzed to design the model of the customer churn prediction using the Logistic regression method. The findings of the study provide countermeasures and recommendations for effectively improving customer turnover.

Acknowledgement

This research is supported by the Doctoral Program of Dalian Neusoft University of Information.

References

[1] Liu Bin, Qiu Huayong. Design and implementation of a comprehensive securities company customer analysis system. Computer Systems & Applications, 2010, 19(10):126-130.
[2] Wang Weijun. Based on data mining securities business customer churn analysis. University of Electronic Science and Technology（SOCIAL SCIENCES）, 2009, 11(1):18-22.
[3] Frederick F. Reichheld and W. Earl Sasser, Jr., Zero Defections: Quality Comes to Service. Harvard Business Review, Sept.-Oct. 1990, 105-111.
[4] BUCKINX W, DIRK Van den Poel .Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. European Journal of Operational Research 2003, 164: 252-268.
[5] Wu Bin, Ying Li. Churn Prediction Model and Application Based Logistics regression algorithm securities customers. Financial electronic, 2013, 65-67.
[6] David Heckerman, Dan Geiger, and David M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. Machine Learning, 1995, 20(3):241-243.
[7] Jia Lin, Li Ming. The establishment and implementation of the model based on data mining customer churn. Computer Engineering and Applications, 2004, 4:185-187.

Materials Science, Computer and Information Technology
10.4028/www.scientific.net/AMR.989-994

Research on Customers Churn Prediction Model Based on Logistic
10.4028/www.scientific.net/AMR.989-994.1517

DOI References
[6] David Heckerman, Dan Geiger, and David M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. Machine Learning, 1995, 20(3): 241-243.
http://dx.doi.org/10.1007/BF00994016