Forecasting of Leaving Probability of Customer by Using Data Mining

Lin, Lung Shu[1] and Chen, Wen Chin[2]
[1]Marketing Department of Chunghwa Telecom Co., Ltd.
21-3 Hsinyi Road, Sec. 1, Taipei, Taiwan 100 R. O. C.
shuh@cht.com.tw
[2]The Laboratory of Chunghwa Telecom Co., Ltd.
9, Lane 74, Hsinyi Road, Sec. 4, Taipei, Taiwan 100 R. O. C.
wenchin@cht.com.tw

ABSTRACT

To predict the leaving customers in telecommunication service, in this study, a forecasting procedure is proposed based on data mining. In the procedure, significant variables are first investigated by testing collected data of customers' behavior in this procedure. Then these variables are employed to fit appropriate forecasting models by adopting data mining tools as Neural network, Decision tree and Logistic curve. The method of "Lift value" [1] is adopted to evaluate these models and select an optimal one. Finally, the proposed optimal model is applied to forecast the leaving probability of ADSL (Asymmetric Digital Subscriber Line) subscriber in telecommunication. The obtained results show our proposed model is five times in accuracy than those not use any forecasting model.

1. INTRODUCTIONS

In a competitive market, customer may move between different operators due to their attractive price, promotion or special service. Therefore, operators often need to pay much cost to retain their customers. If an operator can forecast their potential leaving customers or know these customers' probabilities of moving to other operators, they can focus on retaining these potential leaving customers, and thus can improve the retention rate and reduce the retention cost.

In the earlier time, the regression models in statistics were often employed to forecast leaving customer [5]. Since these models are well defined and have strict assumptions for adopted data, they may produce large forecasting error for unstructured data. Recently, methods in data mining as Neural network, Decision tree etc. [1, 2, 3] with no well defined assumptions have used to solve this problem effectively. By using customers' trading records in data warehouse, they forecasted the leaving probability of individual customer by establishing pattern of these data.

In this study, a studying procedure of forecasting leaving customer is proposed based on the tools in data mining. Besides, to improve the accuracy of forecasting model in this procedure, significant variables that can fully explain why customers move to other operators should be investigated first [7]. The more the significant factors are investigated, the more accurate are the forecasting results. Then these variables are employed to fit appropriate forecasting models by adopting data mining tools as Neural network, Decision tree and Logistic curve. The method of "Lift value" [1] is adopted to evaluate these models and select an optimal one. Finally, the proposed optimal model is applied to forecast the leaving probability of ADSL subscriber in telecommunication. The obtained results show our proposed model is five times in accuracy than those not use any forecasting model.

Figure 1 illustrates the studying procedure in this study. Section 2 describes and validates the collected data, and then significant variables are investigated in Section 3. We establish forecasting models and select an optimal one in Section 4. Finally, a real case is implemented and conclusions are drawn in Section 5.

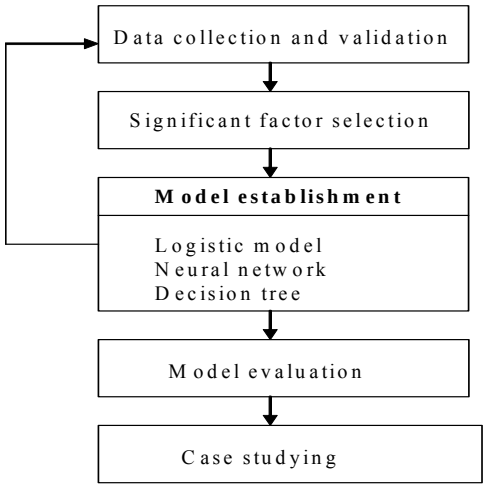

Figure 1. The studying procedure

2. DATA COLLECTION AND VALIDATION

2.1 Data collection

When the tools in data mining are used to establish a forecasting model, large amount of leaving and non-leaving data with customers' trading records should be collected to test the models. In this study, 32 variables with ADSL customers' trading records are drawn [6] from data warehouse and classified into four types: the basic attribute of customer, maintenance and service, billing and communication data. To select an optimal model, these data are also divided into three groups: training sample, validation sample and execution sample. Training sample containing 70 thousands samples is used to establish forecasting models. Validation sample containing 50 thousands samples is used to test and select an optimal model. The final group, execution sample containing 3.5 thousands samples is used to calculate the leaving probability of each customer.

2.2 Data validation

Before implemented to test the forecasting models, the data should be validated to improve the accuracy of the forecasting models. In this study, the collected data are validated as following:

- Data aggregation : Linking the related data and table to make integrated application.
- Data clarification : Detecting the error, unreasonable or incomplete data.
- Dummy variable : Producing new variables by combining two or more variables.
- Data transformation : Transferring (or normalizing) data with large scale into normal scale.

3. SIGNIFICANT FACTOR INVESTIGATION

Since it is too complicate in computation when 32 variables are included in a model, and besides, due to not all variables have capability to explain customer's behavior, only signification variables (factors) are selected to simple the proposed forecasting model in this study. There are several methods [4, 7] can be used to test the significant variables from different view points. In this study, we adopt the t-test in statistics to select significant variables since the amount of training data are very large, up to 70 thousands. The obtained results of t-test are shown in table 1. The larger is the T* value in Table 1, the more significant is the variable.

Table1. Testing results of part of variables.

var.	leaving or not	sample	mean	derivative	testing result T*	model transf.
v1	leaving	42000	10.89	3.07	7.49	Ln transf.
	not leaving	18000	10.48	3.72		
…			…			
v32	leaving	42000	295.85	288.34	15.30*	no
	not leaving	18000	143.81	316.98		

Note: v1 is the first variable, …, v32 is the 32nd variable.

4. FORECASTING MODEL ESTABLISHMENT AND EVALUATION

The data warehouse provides the enterprise with a memory, but the memory is of little use without intelligence. Data mining is the exploration by automatic means using data in data warehouse to provide intelligence for enterprise. It allows corporation to improve its marketing, sales and customer support operations through better understanding of its customers by data classification, estimation, forecasting, clustering and description etc.

Considering the techniques and tools of forecasting in data mining, the Neural network, Decision tree and Logistic curve methods are appropriate to predict the target in various fields presented in many documents [1, 2, 3], and hence are selected to fit the training data in this study. The results are evaluated by a method called "Lift evaluating chart" and an optimal model is selected to forecast the leaving probabilities of ADSL customers.

4.1 Neural network

To establish a Neural network model for the collected data, the relative weight of importance of each significant variable should be first computed in this model by iteration. Part of the obtained results are sorted and shown in Table 2 when 70 thousands training samples are implemented. In Table 2, we know the variable 21 (v21 is the average amount of download during six months), with relative weight of importance 0.306, is the most important variable to describe whether a customer choose ADSL service or not.

Table 2. The relative weights of importance of variables

Variable	Relative weight of importance
V21	30.6%
V13	26.3%
V7	25.9%
:	

Results in Table 3 are calculated by Neural network model when the 50 thousands validation samples are applied. In Table 3, row represents the actual leaving situation and column is forecasting results. The original leaving rate of 50 thousands samples is 1.12% ((384+176)/50000) which means one can catch 1.12 leaving customer after he surveys 100 customers randomly. However, we can obtain 4.0% (176/4324) forecasting rate when the Neural network model is adopted for the same samples.

Table 3. Results obtained by Neural network model.

	forecasting results		
actual situation	not leaving	leaving	forecasting rate
not leaving	45116	4324	91.3%
leaving	384	176	31.4%
overall percentage	91.0%	9.0%	90.6%

4.2 Decision tree

The Decision tree, using simple rules to classify the training data, is an useful tool to forecast an objective and is easy understanding in its computation.

Figure 2. The results of Decision tree

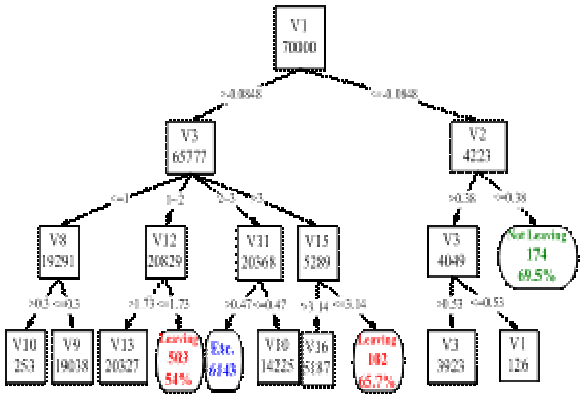

Figure 2 shows part of the results after 70 thousands training data are solved by Decision tree algorithm. From Figure 2, we see the leaving probability of a customer is 0.657 if he has v1>0.0848, v3>3 and v15<3.14

The results are shown in Table 4 when 50 thousands validation samples are implemented, and thus the forecasting rate is computed as 3.7% (233 / 6251).

Table 4. Results obtained by Decision tree model.

	forecasting results		
actual situation	not leaving	leaving	forecasting rate
not leaving	43189	6251	87.4%
leaving	327	233	41.6%
overall percentage	87.0%	13.0%	86.8%

4.3 Logistic curve

The Logistic model is effective used to forecast the objective with two categories (ex, leaving or not leaving). In this study, we use following Logistic model to forecast the leaving probability of ADSL customer:

$$Lscore = e^{ax+c} / 1 + e^{ax+c},$$

and the data transformation model is

new data = (original data – minimum data) / (maximum data – minimum data).

After 70 thousands training samples are implemented, we have

$$aX + c = 0.0001x_1 + 19.827x_6 - 8.314x_{15} - 0.415x_7 + 17.691$$

The finally results are shown in Table 5 after 50 thousands validation samples are implemented. In Table 5, we obtain the forecasting rate 3.2% (148 / 4615).

Table 5. Results obtained by Logistic curve model.

	forecasting results		
actual situation	not leaving	leaving	forecasting rate
not leaving	44825	4615	90.7%
leaving	412	148	26.4%
overall percentage	90.5%	9.5%	89.9%

4.4 Model evaluation

In this study, we adopt the method of "Lift evaluating chart" to evaluate forecasting rate and select the optimal model with the largest Lift value. The value of"Lift"is defined as following:

Lift value at score x % = forecasting rate at score x % / population churn rate

According to the empirical rule, if the value of Lift is greater than 3 at score 10%, the forecasting model is acceptable. In Table 3, the value of Lift is 3.57(4.0% / 1.12%) at score 9%, which shows Neural network model is effective. Also, we have the Lift values of Decision tree and Logistic curve 3.8 and 2.9 at score 9%.

Figure 3 illustrates Lift values of the three forecasting models at various score percentage. At score 5% in Figure 3 for instance, the Lift value of Neural network is better than those of other models'. Due to under the constraint of company's budget, no more than 5% ADSL customers are selected for retention. Therefore, according to the results shown in Figure 3, Neural network model is the optimal selection to forecast the leaving probability of ADSL customer at score 5% in this study.

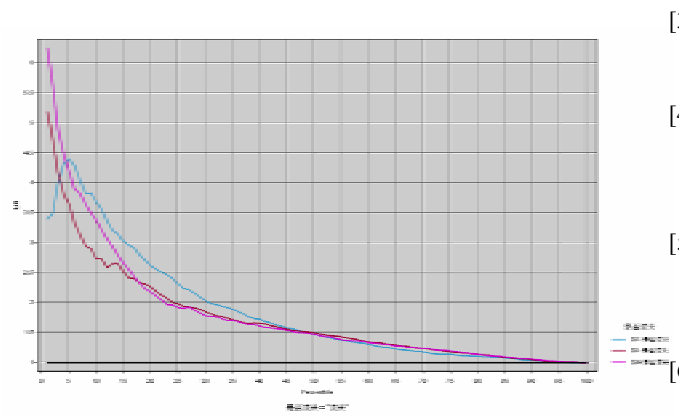

Figure3. The Left value of three forecasting models.

5. A CASE STUDY AND CONCLUSIONS

In the section, we adopt the proposed Neural network model to forecast the leaving probabilities of 3.5 thousands ADSL customers for the next two months. 1052 customers with larger leaving probabilities are selected from 3.5 thousands execution samples at score 3%. After two months observation, 56 out 1052 customers really leave, and thus the forecasting rate is 5.3% (56 / 1052) which is much greater than the population's leaving probability 0.98%. Besides, the Lift value 5.4 at score 3% shows our proposed model is accurate and effective.

Keywords: data mining, forecasting model, Neural network, Decision tree, Logistic curve

8. REFERENCES

[1] M. J. A. Berry and G. Linoff, Data Mining Techniques For Marketing, Sales, and Customer Support, New York: John Wiley & Sons, Inc., 1997.

[2] C. Glymour, D. Madigan, D. Pregibon and P. Smith, Statistical Inference and Data Mining, Association for Computer of the ACM, new York, Nov. 1996.

[3] B. C. Hsieh and R. L. Ye, "Applications of Statistics in Data Mining", Statistic Report, Taiwan, 2000.

[4]. W. Mendenhall, D. D. Wackerly and R. L. Scheaffer, Mathematical Statistics with Applications. Boston: PWS-Kent Pub. Co., 1990.

[5]. J. Neter, W. Wasserman and M. H. Kuther, Applied Linear Statistical Models. Boston: Richard D. IRWIN, INC., pp. 295-, 1989.

[6] R. L. Scheaffer, W. Mendenhall, and L. Y. Ott, Elementary Survey Sampling. Boston: PWS-KENT, 1990.

[7] H. F. Wang and L. S. Lin, "α-complete information in factor space". IEEE Transactions on Fuzzy Systems, 1998.