Research on Prediction Model of Financial User Churn Based on Data Mining

Qiongqiong Zhu
e-mail: zhuqq501@126.com

School of Economics Shanghai University, Shanghai, China. Employer: Chint Electric Co., Ltd

Abstract-Today, with the rapid development of network technology, a large number of online financial services have shown a high degree of homogenization trend, and a large number of customers will face a large number of losses, which will also cause a great impact on the company. According to relevant investigations, the cost of attracting new customers is eight times more than that of retaining old customers. In addition, if a company's customer retention rate can increase by 5%, the company's profit will increase by 85%, so it is very critical for the company to reduce customer losses. This paper mainly uses data mining technology to make statistics of network financial data, and uses classification mode to predict it, so as to formulate retention strategies for enterprises and reduce the loss of users. Unbalanced data sets are processed by under-sampling, over-sampling and SMOTE methods respectively, and the prediction effects of three methods in decision tree model, Logistic regression model, SVM model under four kernel functions and BP neural network model are compared. The experimental results show that the model based on SMOTE method has the highest accuracy among the three sampling methods, and the BP neural network model has higher accuracy among the four models, so it is more suitable for the Internet financial user churn model.

Keywords-data mining; Finance; User churn

1. Preface

Network financing is a new financing method emerging in recent years, which flows funds through advanced technologies such as network technology, mobile phone and network. With the development of Internet technology, And the deepening of the financial industry. The domestic Internet finance business has developed rapidly, Online financial management with low price and high return is favored by consumers [1]. With the increasing number of network investment users, the cake of Internet finance is expanding, the traditional financial industry is transforming to Internet finance one after another, the number of Internet financial platforms is increasing, wealth management products are seriously homogenized, users are more mobile between platforms, and the possibility of losing users is increasing [2]. How to prevent the loss of customers is an important issue that network finance companies must take seriously at present. From the perspective of the operation mode of the company, the turnover of employees is the key factor affecting the sustainable development of the company.

Colgate (2019) conducted a questionnaire survey on the customer service of a commercial bank, analyzed its impact and predicted it[3]; Chih-Ping Wei (2020) adopts a method based on decision tree to analyze the possible user churn[4]. Breiman (2021) uses data mining (classification and regression) technology to study the statistical characteristics of users and predict the interference

BDEIM 2022, December 02-03, Zhengzhou, People's Republic of China

DOI 10.4108/eai.2-12-2022.2328766

of users[5]. He Benlan (2020), using random sampling method to analyze customer churn, and using SVM (SVM) to analyze customer churn[6]; Cao Guo, Xu Juan and Shen Lixiang (2021) used binary Logistic regression analysis method to build a multi-dimensional early warning model of customer flight risk of commercial banks, and used binary Logistic regression analysis to find that transaction frequency, recent transaction time and age have great influence on customer churn, among which transaction frequency and recent transaction time have the greatest influence on customer churn[7].

This paper combines the analysis of user churn characteristics and data mining technology in research methods, A set of methodology for predicting user churn is established, Has certain academic reference value, R language, which is commonly used in data mining and statistical analysis, And successfully completed the whole process from data analysis to data mining, combined the statistical theoretical basis with data mining related technologies, and established an Internet financial user churn prediction model, which provided certain academic value for the Internet financial industry user churn problem.

2. Selection of model variables and descriptive statistics of data sets

2.1 Selection of model variables

By tracking the user's behavior for a period of time, taking April 1, 2020 as an observation station, we observed and extracted the independent variable X data in the three months before that time point, and then extracted each factor Y in one month from April 1, 2019. Then, according to the wide table formed by the independent variable X corresponding to the factor Y, it was used to train the mode. After the model is established, August 30, 2020 is taken as the observation point to verify the accuracy of the model.

2.2 Descriptive statistics of data sets

Before building the model, we should first have a general concept of the whole data set, then describe and analyze the whole data, find out the abnormality and omission of the data, and then preprocess it. Using descriptive statistical method, we can master the variables in the data set as a whole: there are 22 variables, 7 category variables and 15 continuous variables in the data set; There were 2,912,000 observations, with 2.4400 (91.6%) users missing and 26.6800 (8.4%) retained. In addition, in the case of some data exceptions and omissions, Corresponding prejudgment must also be made. For example, the maximum age is 2017, and the highest proportion of funds returned in the next month is 3770%, which exceeds 1. This is totally illogica[10]l. Obviously, this is caused by the wrong data source. For example, there is a gap in the short-term investment ratio, which can be handled by deleting it. The final statistics were 290,300, of which 2.4200 were lost and 25.4100 were retained[11].

3. Establishment of loss prediction model

3.1 Training set and test set

Before data modeling, it is necessary to divide the data set into training set and test set. The training set is mainly used to train the model, while the test set is used to evaluate the accuracy

and prediction performance of the model. Therefore, 70% of the data set is used to train the model, and the remaining 30% is used to test the predicted results of the model.

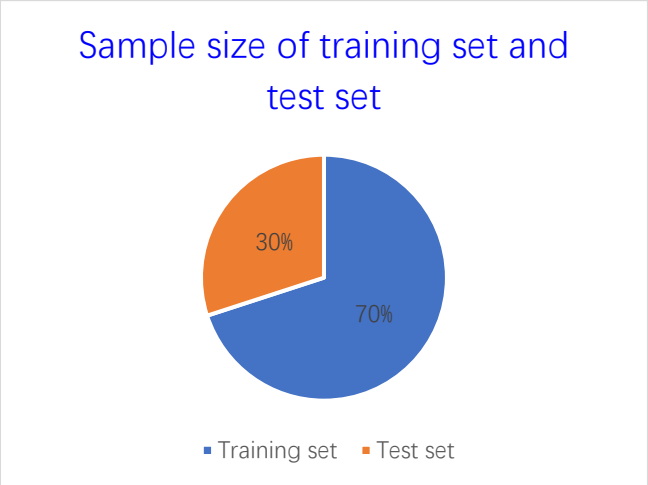

Figure 1 Sample size distribution of training set and test set

3.2 Unbalanced data processing

Before training the model, you need to check the balance of the training data set. By analyzing the binary distribution of dependent variables (churn variables), as shown in Figure1：

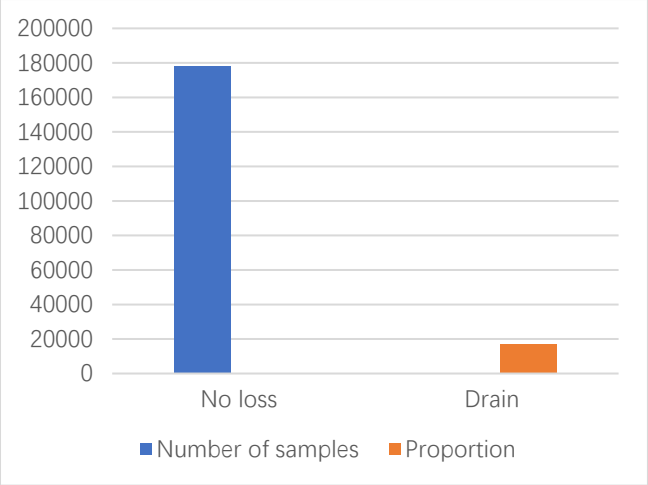

Figure 2 Distribution of dependent variables in training data set

The results show that the model is representative of non-equilibrium, which is 8.7% for normal sample and 91.3% for reverse sample, respectively.

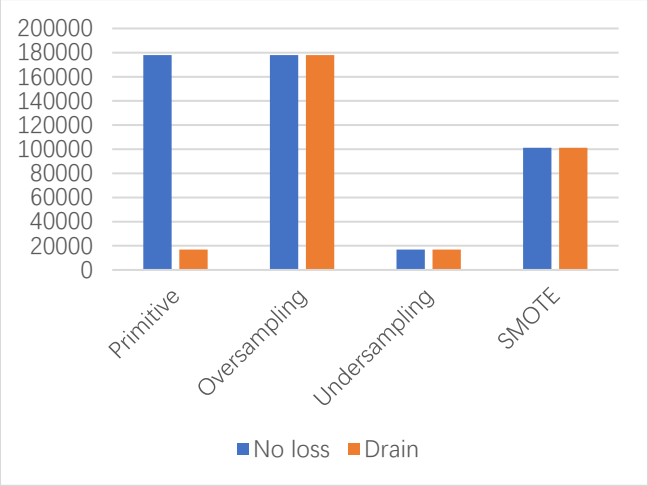

Figure 3 Sample number of three methods to deal with unbalanced data

Above three methods have been used to deal with the unbalanced data of the training set, forming three groups of training sets, and then training the model of these three groups of training sets respectively.

3.3 Decision tree model

As you can see from the three decision trees, The decision tree structure obtained by three different decision tree models is basically the same. They all use four factors as decision trees: last_cyc_amt_rate (capital turnover in the next month), m_max_dead_line (investment period in the last three months), m_invest_amt (monthly average investment in March) and short_prod_rate (short-term investment proportion), while ignoring other attribute variables that have little correlation with user turnover. It consists of five leaves[12]. According to this paper, The decision tree model constructed in this paper has good explanatory ability, Among them, the decision tree of SMOTE algorithm is: when the user's "ratio of payment in the next month to total assets" is classified, when the user's "ratio of payment in the next month to total assets" is less than 89%, the probability of 53% is judged as no loss, and when the user's "investment period in recent March" is greater than or equal to 3 months, the high probability is judged as no loss, which is in line with business logic. If the "average monthly investment in March" exceeds 65,000, a large number of users are likely to leave, which shows that users' investment and trust in them reduce losses. When the "short-term target investment ratio" reaches more than 82%, the users judge as "capital outflow", and there is a certain liquidity between them and the short-term target users. Finally, three different decision tree models are tested respectively, and the corresponding AUC values are obtained: AUC of decision tree model AUC based on oversampling; Under the condition of low sampling, the AUC value of decision tree is 0.921; The AUC of this algorithm in SMOTE algorithm is 0.936. The results show that SMOTE method is the three sampling methods, and the model based on SMOTE method has the highest accuracy, so it is more suitable for forecasting the flow of network financial personnel. Some of the following model algorithms use SMOTE to train uneven data sets.

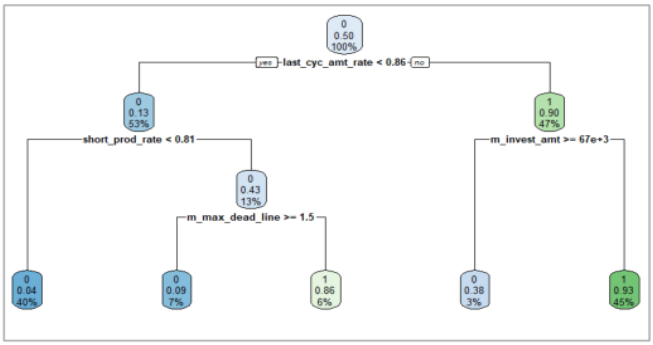

Figure 4 Decision tree model tree diagram (oversampling)

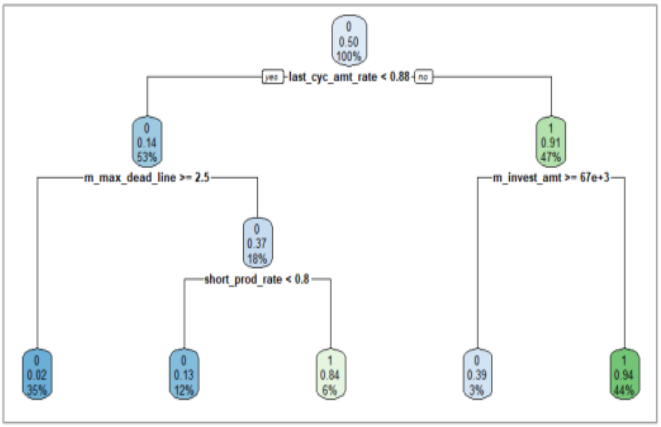

Figure 5 Decision tree model tree diagram (undersampled)

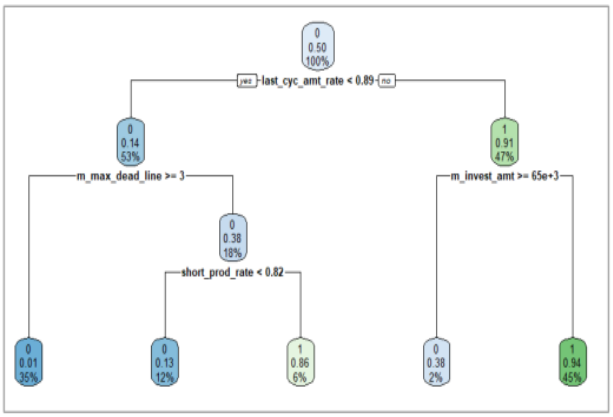

Figure 6 Decision tree model tree diagram (SMOTE method)

3.4 Logistic regression model

Then the logistic regression model is used to establish the data set, and then the SMOTE method is used to train the logistic regression model. After the model is constructed, the influence of each factor is tested. The results show that under different conditions, the P value of each factor is below 0.001, and under different conditions, the difference of each factor is very different. The results show that the above two factors have no significant effect on the prediction results.

After understanding the estimation of each parameter, the following mathematical formula can be obtained:

log (Loss Trends)

=-2.0970 +0.3511/Return Channel-. 0850 Is there an investment date?

-0.0092 years old +0.0003. Average registration time in the last three months +0.0083

-0.1833 Average monthly investment days in the last three months-0.3573

+-0.3167 Monthly Average Monthly Payment Date +3.9060 Monthly Cash Turnover Ratio

-2.4140 Investment ratio of short-term assets

In addition, we can also learn that five factors, such as rebate channel, investment time, average monthly login days, payment ratio in the next month and recent investment ratio, have a positive impact on user loss, that is, the greater the value of these five factors, the easier it is for users to lose, and the payment in the next month contributes the most to user loss. In addition, five factors such as the second investment date, age, average investment days in the last three months, investment cycle in the last three months and average recharge days in the last three months have an inverse effect on the loss of users, that is, the greater these five factors are, the less likely users are to lose.

3.5 SVM model

Finally, the model of support vector machine is established by using support vector machine method, and the simulation of SVM is carried out by using SMOTE method. The type of SVM is C-classification, which uses linear kernel, polynomial kernel, radial basis function kernel and Sigmoid kernel to model. This program takes place in R. On this basis, four different kernel functions are used to establish the model of support vector machine, which is compared with the actual situation, and its output is fuzzy matrix, and it is compared horizontally. From the four core functions, the classification accuracy of support vector machine (SVM) is investigated respectively.

Table1 Comparison of positive and negative examples of SVM model based on four kernel functions

Prediction algorithm	Real example	True negative example	False positive example	False negative example
Radial basis kernel function	6900	70189	6048	370
Polynomial kernel function	6857	69831	6406	413
Linear kernel function	6558	69079	7158	712
Sigmoid kernel function	6147	64815	11422	1123

As can be seen from Figure Table11, the correct classification number of normal and negative states of support vector machine is higher than that of SVM constructed by other three kernel functions, while the errors obtained by positive and negative kernel functions are smaller. Using the above fuzzy matrix, the corresponding accuracy, recall, accuracy, promotion coefficient and F value are obtained. Table2 shows the comparison of the prediction effects of various models.

Table2 Effect comparison of SVM models based on four kernel functions

Prediction algorithm	Precision rate	Recall rate	Accuracy	Lifting coefficient	F_1
Radial basis kernel function	0.533	0.949	0.923	6.121	0.683
Polynomial kernel function	0.517	0.943	0.918	5.939	0.668
Linear kernel function	0.478	0.902	0.906	5.492	0.625
Sigmoid kernel function	0.350	0.846	0.850	4.019	0.495

It can be seen from Table2 that the support vector machine of support vector machine is a better method. According to the data, the accuracy of SVM method using radial basis function kernel SVM method is 2%, 6% and 18% higher. Accuracy refers to the accuracy obtained during sampling when users are lost, while the support vector machine model based on RF has higher accuracy, which shows that SVM has higher accuracy in the sample database expected to be lost users; Similarly, the SVM based on radiative kernel has higher recall rate, accuracy rate, lifting coefficient and other three types of SVM. Therefore, four different kernel functions are used to build a support vector machine. Among them, the accuracy of RBF-based support vector machine is much higher. The following method is used to compare.

3.6 BP neural network model

Then the BP neural network method is used to establish the data set, and the SMOTE method is used to train the model. Firstly, the BP neural network is layered and calculated. BP neural network is divided into three layers: input layer, hidden layer and input layer. At the same time,

with the increase of the number of stealth layers, the error rate of the system is reduced, the learning cycle of the network is greatly prolonged, and the structure of the system becomes more complex. Therefore, the three-layer BP neural network used in this paper is a neural network including input layer, output layer and hidden layer. In order to reduce errors, we can increase the number of hidden layer units, but not too much, so as not to cause disharmony and lead to the normal operation of neuron system. The initial value is set to the initial value of the number of hidden layers. When other parameters are the same, the corresponding AUC value is obtained by training with the same experimental set. The larger the AUC value, the more the number of hidden layers is obtained. You can use the following formula to calculate the number interval between secret layers:

$$l = \sqrt{n + m + a}$$

Where l is the number of hidden layer units; n is the number of neurons in the input layer; m is the number of neurons in the output layer; A is the regulating factor, usually 1 to 10.

Table3 Influence of the number of hidden layer units on model results

Number of hidden layer units	AUC
5	0.934
6	0.932
7	0.926
8	0.943
9	0.93
10	0.934
11	0.929
12	0.941
13	0.94
14	0.935

It can be seen from Fig. 3 to Fig. 3 that the AUC is the largest when the number of hidden layers is 8, which indicates that the prediction performance of the network is better. Here, we set the number of hidden layers to 8 and the maximum number of exercises to 1000. Because the learning efficiency of neural network is also related to the learning efficiency of the whole network, through a large number of experiments and comparisons, the learning result of 0.01 is finally obtained. On this basis, the overall structure of BP neural network is established and adjusted accordingly, including the start-up function of hiding and output level, weighted update method and so on. Because the object of this experiment is classified variables, the more common activation function Sigmod function was selected. On the basis of weighting, an adaptive algorithm based on gradient is proposed. After making sufficient preparations, the BP neural network is modeled. This program is also carried out in R language. After modeling BP neural network, the ROC graph is drawn according to the training set, and the ROC graph is studied.

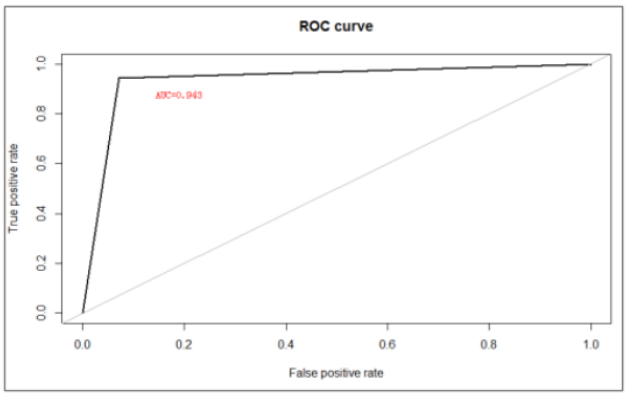

Figure7 ROC of BP Neural Network Training Set

It can be seen from Fig. 8 that the AUC value of the training set reaches 0.943, which shows that the established BP neural network has a strong fitting effect with the training data set.

4. Conclusion

In this paper, we use statistical principles and data mining technology to analyze and study the loss of Internet financial users quantitatively, and use data mining technology such as correlation analysis, decision tree, logistic regression, support vector machine and BP neural network to build a prediction model of Internet financial user loss. Finally, the model is tested by the experimental data, the accuracy of the model is tested, and the model is estimated, which provides great convenience for the company to retain the lost customers.

References

[1] Zhang Yusi. Research and application of mobile communication customer churn prediction model based on data mining [J]. Modern Information Technology, 2022, 6 (11): 41-45 +50. DOI: 10.19850/J.cnki.2096-4706.2022. 011. 011.

[2] Wang Xiangping. Application research of user churn analysis model based on embedding and model stacking fusion algorithm [J]. China New Communication, 2021, 23 (22): 98-100.

[3] Zhang Yili. Accurate marketing of new radio and television services based on heat map [J]. Radio and Television Network, 2021, 28 (07): 105-107. DOI: 10.16045/j.cnki.catvtec.2021.07.031.

[4] Deng Mingrong. Research on UDGM Marketing Strategy of Mobile Communication in 5G Environment [J]. China New Communication, 2021, 23 (13): 67-69.

[5] Yang Yingying. Early warning system of mobile phone customer churn based on data mining [J]. Modern Information Technology, 2020, 4 (13): 55-57. DOI: 10.19850/j.cnki.2096-4706.2020.13.018.

[6] Long Keshu, Deng Juan, Liu Xiaobin. Research on Predicting and Coping Strategies of Operator User Churn Based on Machine Learning Algorithm [J]. Information Record Materials, 2020, 21 (05): 202-205. DOI: 10.16009/j.cnki.cn13-1295/tq.2020. 05.135.

[7] Yuan Qingze, Wu Caixia. Application of data mining in early warning of user loss in communication enterprises [J]. Information and Communication, 2018 (08): 252-253.

[8] Wei Li, Zhao Chunjian. Research on mobile abnormal user model based on big data [J]. China High and New Technology, 2018 (09): 55-57. DOI: 10.13535/j.cnki.10-1507/n.2018. 09.04.
[9] Xu Naili. Research and Application of Telecom Customer Churn Prediction Model Based on Big Data Technology [J]. Information and Communication Technology, 2018, 12 (02): 66-71.
[10] Jia Qiong. On the application of data mining technology in enterprise customer relationship management [J]. Modern Marketing (late issue), 2018 (02): 129.
[11] Li Shuguo, Li Chujun. A new data mining method for user churn process in power supply service channels [J]. China Science and Technology Information, 2018 (02): 90-92.
[12] Dong Bo, Wang Xue. Predicting user churn based on hybrid data mining model [J]. Firepower and Command and Control, 2017, 42 (03): 156-160 +165.