

# Research and Application of Internet Financial User Churn Prediction Model Based on Data Mining

Bohui Wan

Tongji University, Yangpu, Shanghai, 200092, China

## Abstract

With the rapid development of network technology, there are more and more Internet financial platforms, and network financial products have become more homogenized. The possibility of user loss is increasing, and user loss will have a great impact on enterprises. Based on the data of lost users in Internet finance, 21 variables consisting of user attributes, asset factors, and behavioral factors were selected, and the imbalanced data set was processed through the SMOTE method. The experimental results prove that the model based on the SMOTE method has high accuracy, and the BP neural network model has a higher accuracy rate, so it is more suitable for the Internet financial user churn model.

## Keywords

Data mining; Internet finance; BP neural network; Churn prediction.

## 1. Introduction

With the development of the Internet and its continuous penetration into the financial industry, the Internet financial model has developed rapidly in my country, and the low-cost and high-yield financial products of online financial management are increasingly popular among users. The number of online investment users continues to grow, and the cake of Internet finance is increasing. Traditional financial industries are transforming to Internet finance. The number of Internet financial platforms continues to increase sharply. Wealth management products are becoming more homogeneous, and users have greater liquidity between platforms. The possibility of user loss is also increasing [1]. Therefore, preventing the loss of users has become a major problem that Internet financial companies have to pay attention to. From the perspective of the company's own business model, user loss management is related to the company's survival and development.

## 2. Related Work

Research on user churn based on data mining technology is mainly applied to user relationship management, establishing targeted and personalized operation methods, promoting new users, increasing user viscosity, modeling user churn data, and analyzing the reasons for user churn through analysis. Adopt corresponding policies to prevent user loss. The foreign companies that have applied this research are mainly the telecommunications industry and commercial banks. Some scholars have elaborated on the application of data mining technology in bank user relationship management. They believe that user relationship management is a bridge between users and the bank, and data mining has sublimated the service capabilities of user relationship management [2]. Through the analysis of user behavior characteristics, we can understand and maintain users well, discover their interests and preferences, so as to quickly respond to user needs, create more considerate and satisfactory services for users, and greatly enhance the user viscosity of the enterprise. Operational auxiliary tools developed through data mining

technology can provide efficient pattern recognition and prediction capabilities, so that companies can effectively carry out marketing activities.

Wei used the decision tree algorithm to predict potential user churn [3]. Breiman used data mining algorithms (classification and regression algorithms) to analyze user demographic characteristics and perform user disturbances in user service logs. The forecast [4] experts and scholars such as Gardell analyzed the factors affecting user churn in the mobile communication industry, and predicted the user churn through a tree model (Tree Net) [5]. Ahmad recorded various data Collected, researched the maintenance of telephone banking users of a certain bank [6].

The above researches are all based on the research on the churn model algorithm, and there are few practical applications of the user churn reason analysis and churn model, and the use of questionnaire surveys to obtain user churn characteristics will waste the cost of user data acquisition. The collected user-related data also has great limitations, which will seriously affect the accuracy of user churn prediction. Based on a comprehensive user information database, this paper will capture and store user login and access data, investment behavior data, user attributes, user asset status and other multi-dimensional indicators in advance, and analyze the factors that affect user churn in as many dimensions as possible, and establish churn Model, and apply the model to practice for verification and effective operation.

### **3. The Establishment of A Prediction Model for the Loss of Internet Financial Users**

If data mining technology can determine in advance whether a user will be lost and analyze user preferences, user operations staff can implement personalized operating strategies for this user in advance, allowing users to increase investment or extend the investment period, and save it. To the effect of retaining users in advance. Through data mining technology, it is possible to discover the laws between data from a large amount of user data, extract valuable information, facilitate data analysis, understand user characteristics, build user churn prediction models, predict whether users will churn, and retain valuable users in advance. Realize the double growth of value users and business, and improve the competitiveness of enterprises.

#### **3.1. Selection of Model Variables**

Combining the characteristics and business characteristics of Internet financial companies, the final definition of user churn is: the user's net assets after the last withdrawal is less than 100 yuan (from the platform withdrawing 100 yuan) and no recharge within 30 days, the user is defined as a churn user. The time of the last withdrawal is the lost time. As long as the churn users who have this characteristic within a period of time are taken as the target users, the basic attributes of these users, the purchase of investment products, the frequency of withdrawal, the frequency of login and other behavioral characteristics of these users are analyzed, and a user churn prediction model is established, and then the model is applied. Go to the current users, determine whether the users have a tendency to churn and the probability of churn, and let the enterprise adopt the corresponding retention policy in advance to prevent the loss of this part of users.

Through analysis and research, it is found that the main factors affecting user churn are the purchase of investment products in the past three months, the changes in user assets in the past three months, user payment status and user attributes. This article will measure the impact on user churn from three aspects of user attributes, asset factors, and behavioral factors, and finally determine the following 21 main variables that affect Internet financial user churn, as shown in Table 1.

**Table 1. Variables of the Internet Finance Loss Model**

| Variable type          | Variable name     | Variable description                                                        | Types   | Variable interpretation                                                                   |
|------------------------|-------------------|-----------------------------------------------------------------------------|---------|-------------------------------------------------------------------------------------------|
| Dependent variable Y   | is_churn          | Whether to lose                                                             | int     | Net assets after the last withdrawal <100 and No recharge within 30 days<br>1: yes, 0: no |
| Independent variable X | sex               | gender                                                                      | varchar | gender                                                                                    |
|                        | is_rebate         | Whether to rebate channel                                                   | varchar | Whether to rebate channel                                                                 |
|                        | level             | User value level                                                            | varchar | User value level                                                                          |
|                        | chnl_level        | Channel level                                                               | varchar | Channel level                                                                             |
|                        | is_fis_to_inv     | Is the investment on the first investment day equal to the total investment | varchar | Is the investment on the first investment day equal to the total investment               |
|                        | is_second_date    | Is there a second investment day                                            | varchar | Is there a second investment day                                                          |
|                        | age               | age                                                                         | int     | age                                                                                       |
|                        | invest_time       | Investment duration                                                         | int     | Investment duration                                                                       |
|                        | m_sign_day        | Average sign-in days in the past 3 months                                   | int     | Average sign-in days in the past 3 months                                                 |
|                        | m_login_day       | Average number of login days per month in the past 3 months                 | int     | Average number of login days per month in the past 3 months                               |
|                        | d_netassets       | Average daily net assets in the past 3 months                               | money   | Average daily net assets in the past 3 months                                             |
|                        | d_dueinpandi      | Nearly 3 months to be received every day                                    | money   | Nearly 3 months to be received every day                                                  |
|                        | m_invest_count    | Average monthly investment days in recent 3 months                          | int     | Average monthly investment days in recent 3 months                                        |
| Variable type          | Variable name     | Variable description                                                        | Types   | Variable interpretation                                                                   |
| Independent variable X | m_invest_amt      | Average monthly investment amount in recent 3 months                        | money   | Average monthly investment amount in recent 3 months                                      |
|                        | m_max_dead_line   | Investment period in the last 3 months                                      | float   | Investment period in the last 3 months                                                    |
|                        | m_recharge_count  | Average monthly recharge days in the past 3 months                          | int     | Average monthly recharge days in the past 3 months                                        |
|                        | m_recharge_amt    | Average monthly recharge amount in recent 3 months                          | money   | Average monthly recharge amount in recent 3 months                                        |
|                        | m_max_invest_amt  | Maximum investment amount in recent 3 months                                | money   | Maximum investment amount in recent 3 months                                              |
|                        | m_max_netassets   | The largest net assets in the past 3 months                                 | money   | The largest net assets in the past 3 months                                               |
|                        | last_cyc_amt_rate | Proportion of payment in the next month                                     | float   | Proportion of payment in the next month                                                   |
|                        | short_prod_rate   | Proportion of short-term standard investment                                | float   | Proportion of short-term standard investment                                              |

### 3.2. Model

Decision tree algorithms are often used to solve classification and regression problems. It is the process of classifying data through a series of rules. It discovers laws and relationships by systematically decomposing and separating the information contained in the data set. The most widely used decision tree algorithm is the selection criteria of various attributes based on information entropy. For example, the ID3 algorithm uses information gain as the basis for the selected attributes to describe the quality of the selected attributes; the C4.5 algorithm uses the information gain rate as a criterion for selecting attributes, and can handle continuous numerical features.

#### (1) Information entropy

Information entropy is used to describe the impurity of selected attributes, and is used to measure the uncertainty of random variables. Entropy is the expected value of information, the formula is:

$$\begin{aligned} \text{Entropy} &= -(P_1 \log_2 P_1 + P_2 \log_2 P_2 + \dots + P_k \log_2 P_k) \\ &= -\sum_{i=1}^k P_i \log_2 P_i \end{aligned}$$

Among them is the sample set  $S$ ,  $n$  is the number of target attributes, and  $P_i$  is the proportion of  $I$  belonging to the category in  $S$ . The greater the entropy, the greater the impurity; the smaller the entropy, the smaller the impurity.

#### (2) Information gain

In the ID3 algorithm, information gain is used to describe the criteria for the quality of the selected attribute, and the formula is:

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{i=1}^k \frac{m_i}{n} \text{Entropy}(S_i)$$

Among them, entropy ( $S$ ) is the overall information entropy of the sample set  $S$ , the number of samples is  $n$ , the lower node is obtained according to the attribute  $A$ , assuming that it has  $k$  values, the  $i$ -th value is  $m_i$  samples, and the information entropy is denoted as  $\text{Entropy}(S_i)$  the greater the information gain, the better the selected attribute, and the more conducive to classification, the smaller the information gain, the worse the selected attribute, and the less conducive to classification. Therefore, the maximum value is selected as the basis for the selected attribute.

#### (3) Information gain rate

The C4.5 algorithm uses the information gain rate as the criterion for the quality of the selected attribute, and its formula is:

$$\text{GainRatio}(S, A) = \frac{\text{Gain}(S, A)}{\text{Entropy}(S)}$$

The information gain rate is the ratio of the information gain  $\text{Gain}(S, A)$  to the overall information  $\text{Entropy}(S)$ . The C4.5 algorithm can process continuous numerical features. The idea is to divide the continuous numerical features by two (multiple) divisions. Discretization, for each discrete scheme, calculate the information gain rate, and choose the scheme with the largest information gain rate to discretize the continuous numerical features.

### 3.3. BP Neural Network Algorithm

Neural network is a neural network formed by many neurons, which can simulate the interaction between the real world and objects of the biological nervous system. Artificial neural network processing data information is to train the neural network through data samples, so that it has the memory ability and discrimination ability similar to the human brain, and realizes various information processing functions [7].

BP neural network is a multi-layer feedforward neural network. Its main characteristics are: the signal propagates forward, and the error propagates backward [8]. The algorithm process of BP neural network mainly has two stages [9]. The first stage is the forward propagation of information. Information is input from the input layer, passes through the hidden layer, and finally reaches the output layer; the second stage is the inverse of the error. To propagate, from the output layer to the hidden layer, and finally to the input layer. In this propagation process, the weights and biases from the hidden layer to the output layer, and the weights and biases from the input layer to the hidden layer are sequentially adjusted. The structure diagram of BP neural network is shown in Figure 1.

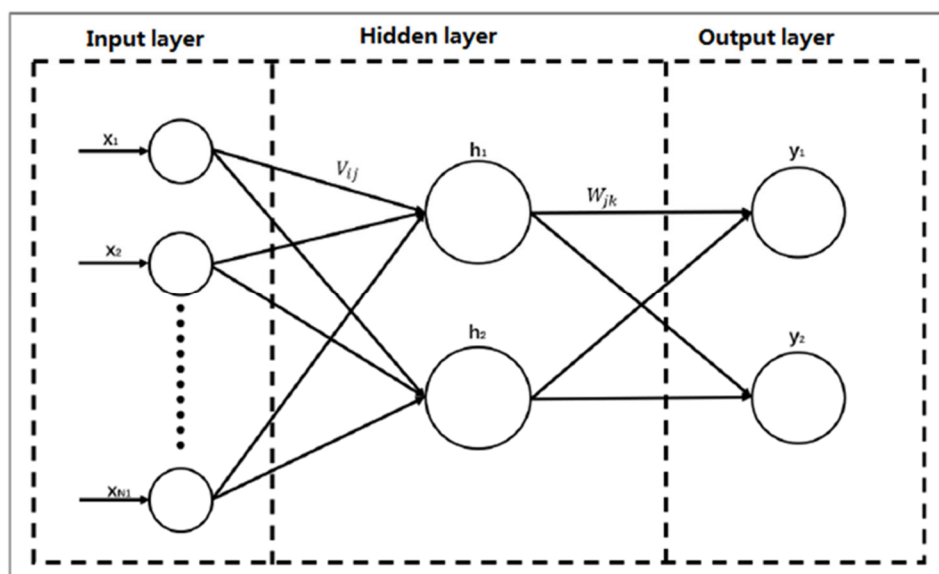


Figure 1. BP neural network structure diagram

## 4. Experiment

### 4.1. Training Set and Test Set

Before data modeling, the data set needs to be divided into a training set and a test set. The training set is mainly used to train the model, and the test set is used to evaluate the accuracy of the model and evaluate the predictive performance of the model. Therefore, 70% of the data in the data set is used to train the model, and the remaining 30% is used to test the results of the model's prediction.

Training set and test set sample size distribution

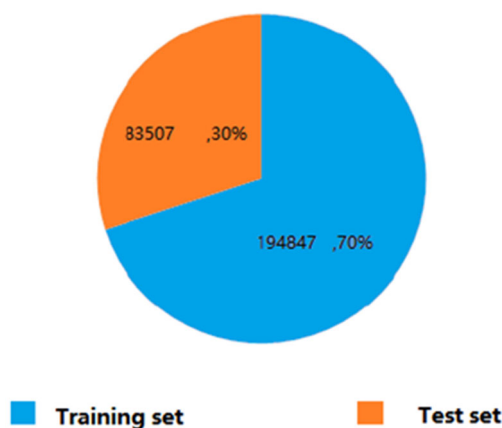


Figure 2. Training set and test set sample size distribution

## 4.2. Decision Tree Model

Use the rpart function of the R language to establish the decision tree model of the above three balanced training sets Z1, Z2, and Z3. At the same time, through the printcp function, the complexity parameter matrix of the decision tree can be calculated, and the structure and error of the decision tree can be analyzed.

**Table 2.** Decision tree

| CP       | nsplit | rel error | xerror  | xstd    |
|----------|--------|-----------|---------|---------|
| 0.771382 | 0      | 1         | 1.00569 | 0.00222 |
| 0.044132 | 1      | 0.22862   | 0.2287  | 0.00141 |
| 0.010948 | 3      | 0.14035   | 0.14042 | 0.00114 |
| 0.01     | 4      | 0.12941   | 0.1295  | 0.00109 |

CP is the complexity parameter of the decision tree as a penalty factor to control the size of the tree. In short, the larger the CP value, the smaller the tree split size (nsplit). The output parameter (rel error) indicates the average deviation ratio between the current classification model tree and the empty tree. xerror is the cross-validation error, and xstd is the standard deviation of the cross-validation error.

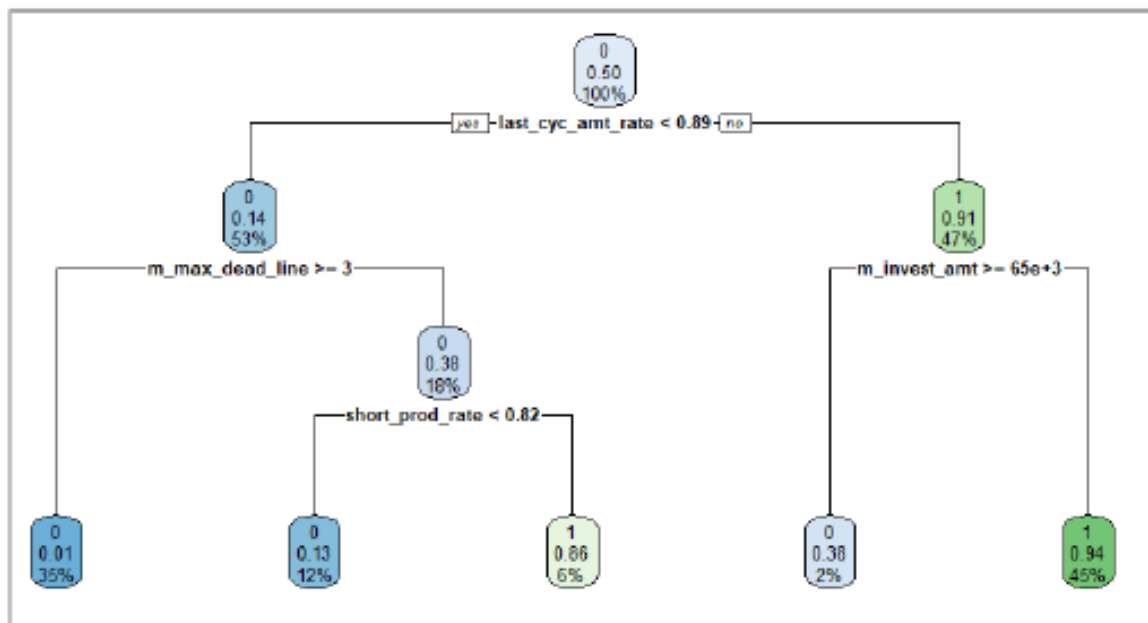


**Figure 2.** Tree diagram of decision tree model (SMOTE method)

The decision tree first categorizes the user's "repayment in the next month as a proportion of total assets". When the user's "repayment in the next month as a proportion of total assets" is less than 89%, there is a 53% probability of being judged as non-loss. When the user's "investment period in the last 3 months" is greater than or equal to 3 months, it is highly likely that the user will not lose, which is in line with business logic. AUC value of decision tree model based on SMOTE method: 0.936. The data shows that the model established by the SMOTE method has the highest accuracy, so it is more suitable for establishing an Internet financial user churn prediction model.

### 4.3. BP Neural Network Model

The BP neural network algorithm is used to model the data set, and the data set processed by the SMOTE method is also used for model training.

The first is to determine the number of layers and units of the BP neural network. It is already known that the BP neural network is composed of an input layer, a hidden layer and an output layer. Here, what needs to be determined is obviously the number of hidden layers. Since the increase in the number of hidden layers is conducive to the reduction of errors, the training time of the network will be greatly increased and the structure will be more complicated [8]. Therefore, this article adopts the conventional three-layer BP neural network model, which is a neural network composed of an input layer, an output layer and a hidden layer.

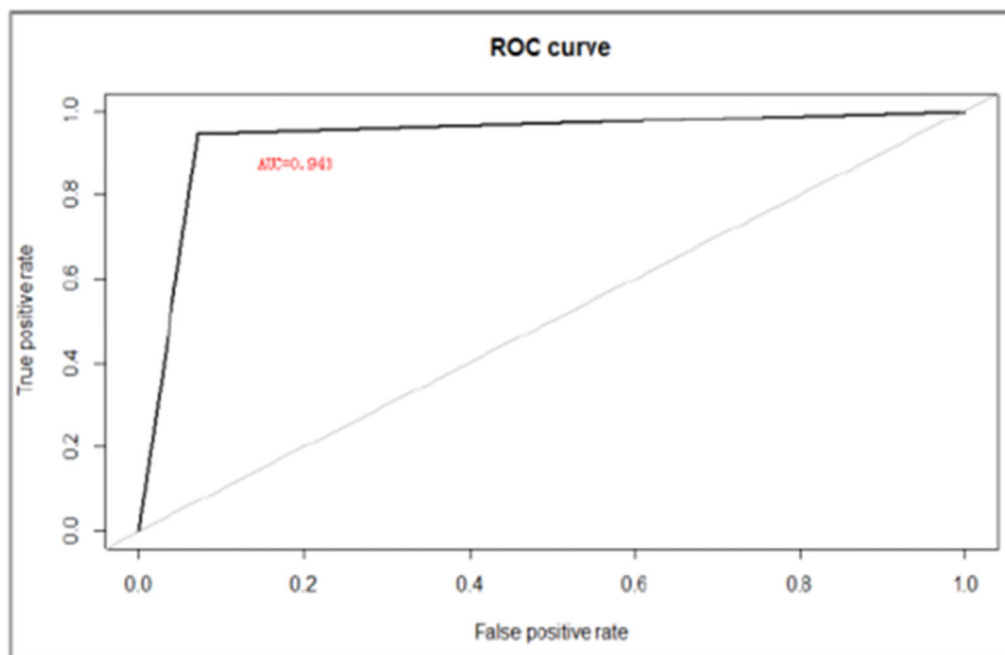


**Figure 3.** ROC of BP neural network training set

It can be seen from Figure 3 that the AUC value of the training set is 0.943, indicating that the established BP neural network has a strong fitting effect with the training data set.

## 5. Conclusions

Based on statistical theory and data mining tools, this paper establishes a user churn prediction model for Internet financial users, and conducts quantitative analysis and research on the problem of Internet financial user churn, using correlation analysis, decision tree algorithms, and BP neural network data Mining technology, established an Internet financial user churn prediction model, output the possibility of each user's churn through the model, and combined statistical methods, data mining algorithms, business logic, and model evaluation and optimization methods to achieve statistics and business applications Combination of qualitative and quantitative analysis methods. Finally, the accuracy of the model was verified through the test set, and the lost user group was predicted, which helped enterprises retain lost users in advance.

## References

- [1] Guo Zhongyuan. Research on the Development of Internet Finance in China[J].Global market, 2017, 3:1-2.

- [2] Qin H. The Application of Data Mining in Telecommunication churn customer[J].Research Journal of Applied Sciences Engineering&Technology,2012,4(11):1504-1507.
- [3] Wei C P, Chiu I T. Turning telecommunications call details to churn prediction: a data mining approach [J].Expert Systems with Applications,2002,23:103-112.
- [4] Breiman L, Friedman J, Olshen R, Stone C. Classification and regression trees[J]. Pacific Grove: Wadsworth, 1984.
- [5] Levin N, Zahavi J. Predictive modeling using segmentation[J]. Journal of Interactive Marketing, 2001, 15(2):2-23.
- [6] Ahmad R, Buttle F. Retaining telephone-banking customers at Frontier Bank [J]. The international Journal of Bank Marketing, 2002, 20(1):5-16.
- [7] Hecht-Nielsen R. Theory of the backpropagation neural network[C]. International joint conference on neural networks 1989, IJCNN, IEEE: 593-605.
- [8] An Xuanmiao. Research on mobile subscriber churn prediction algorithm based on BP neural network[D]. Shenyang: Northeastern University, 2016.
- [9] Kandel E, R, Schwartz, J. Principles of Neural Science[M], Amsterdam: Elsevier, 1989: 1-56.
- [10] Steinhaus H. Sur la division des corp materiels en parties[J]. Bull Acad, Pol Sci 1:801-804.