Scientific Research Publishing

Open Journal of Business and Management, 2020, 8, 866-880
https://www.scirp.org/journal/ojbm
ISSN Online: 2329-3292
ISSN Print: 2329-3284

P2P Borrower Default Identification and Prediction Based on RFE-Multiple Classification Models

Xianyan Hou

School of Management, Jinan University, Guangzhou, China
Email: houxianyan123@163.com

How to cite this paper: Hou, X.Y. (2020) P2P Borrower Default Identification and Prediction Based on RFE-Multiple Classification Models. *Open Journal of Business and Management*, **8**, 866-880.
https://doi.org/10.4236/ojbm.2020.82053

Received: February 28, 2020
Accepted: March 21, 2020
Published: March 24, 2020

Abstract

P2P network lending, as a new type of lending model for Internet finance, is favored by people because of its fast and low cost. However, borrower default has always been one of the core issues of platform concern. Because borrower characteristic data has the characteristics of high dimensionality and multicollinearity, how to select key features to judge borrowing default behavior has been a hot topic. To solve this problem, this paper uses the data of the lending club lending platform to introduce the recursive feature elimination method (RFE) to select key variables, and combines with the classification model to predict the borrower's default behavior. The research results show that the recursive feature elimination method can screen the key variables affecting the default of the borrower. After the recursive feature elimination method, the accuracy of the classification model is over 95%.

Keywords

P2P Networks Lending, Recursive Feature Elimination, Classification Model, Credit Default Risk

1. Introduction

With the development of Internet technology, private lending has developed from offline to online. P2P (peer-to-peer) online lending is a new model of Internet finance, in which lenders and borrowers borrow directly through the Internet platform and no longer need banks or other financial institutions as intermediaries. Its main characteristics are low threshold, convenient, both lenders and borrowers can complete transactions through the online, so it is favored by people, and constantly faced with risks and challenges. In particular, the "thun-

derstorms" of P2P platforms kept happening in 2018, which greatly affected the healthy development of P2P online lending industry. According to the "2018 Online Loan Industry Data Summary" released by Rong 360 Big Data Research Institute, as of the end of December 2018, there were 1082 online loan platforms operating normally in the country, 848 problem platforms, and 254 more problem platforms than 2017. In the current period, the number of active lenders and borrowers was 3.169 million and 6.0205 million. Compared with December 2017, the number of lenders borrowed was 6.865 million and the number of borrowers was 6.933 million. One of the reasons why there are so many problematic platforms is that a large number of borrowers have defaulted and lost contact, which has seriously damaged the interests of platforms and investors and hindered the sustainable development of P2P online lending. In P2P online lending industry, loan customers losing contact refers to loan customers who cannot repay the loans due and cannot be contacted by P2P platforms through various means. In recent years, there have been frequent incidents of P2P platforms running off the road and losing contact with each other. In fact, the malicious default of loan customers and the loss of contact have broken the capital chain of P2P platforms, which is one of the important reasons leading to the closure and loss of P2P platforms, bringing huge challenges to the sustainable and healthy development of P2P platforms. In February 2019, the Beijing Internet finance industry association published "a notice on the list of borrowers and institutions that evaded and abandoned their debts by online lending institutions in Beijing"[2], which published 300 borrowers who evaded and abandoned their debts. Among these 300 borrowers, more than 100 borrowers' overdue time occurred during the "storm" of P2P platforms in 2018. In addition, only 66 of the 300 borrowers have not lost contact, with a loss ratio of 78% and overdue amount of 164,100 yuan, which has brought huge losses to investors and platforms. Therefore, it is urgent for the current P2P industry to identify borrowers' default behaviors. This paper takes foreign Lending club as the research object. Lending club is a good P2P Lending platform in foreign countries. In this paper, the recursive feature elimination method is used to select the key information of the borrower, eliminate the multicollinearity of the data, and predict the default behavior of P2P borrowers by combining Logistic regression model, CART decision tree and BP neural network model.

2. Literature Review

At present, researches on P2P online lending mainly focus on the influencing factors of borrowers' default behaviors and how to choose the credit evaluation model to identify and predict borrowers' default situations. Lin *et al.* [1] believe that loan description is positively correlated with loan success rate, and nega-

[1]https://www.r360insights.com/insights/.

[2]https://mp.weixin.qq.com/s?src=11&timestamp=1583566563&ver=2201&signature=BcA1ZLKXaeIdEpHrzJFjyf374LfwkOtRIUYxBhf9vVnW62UYdmMjPvPN262X3TjvKgGHtd5z400fBdz1Fgu2PCHZuXSoc0bQNMW0e4utZ4SyaFkmbZTCOYlaF8lYW7*&new=1.

tively correlated with loan interest rate and borrower default rate. Kumar *et al.* [2] found that investors are more willing to invest in borrowers with high credit ratings and more loan descriptions, but the higher the loan amount, the higher the default rate. Emekter *et al.* [3] used the data published by Lengding Club to study the characteristics of P2P loans, and found that credit rating, debt-to-income ratio and FICO score were significantly correlated with loan defaults, among which the default rate of loans with low credit rating and long loan maturity was the highest. Lin *et al.* [4] found that gender, age, marital status, education level, working years, loan amount and historical times of default had significant effects on loan defaults based on data of P2P online lending platforms in China. Serrano-Cinca *et al.* [5] found that economic characteristics such as annual income, housing status and debt-to-income ratio were significantly correlated with delinquency among loan customers. Chen *et al.* [6] used data from Chinese platforms to study and found that the default rate of borrowers in P2P online lending was also related to gender. Generally speaking, the default rate of females was lower than that of males. Carlos *et al.* [7] found that the larger the loan amount and the higher the loan interest rate, the higher the default rate of the borrower. Puro *et al.* [8] pointed out that credit score, historical default times and total debt repayment ratio significantly affect the success rate of borrowing. Guo *et al.* [9] took into account the borrowers' historical credit score, loan amount, debt-to-income ratio and housing situation, etc. in terms of credit evaluation of P2P borrowers. Iyer *et al.* [10] used descriptive information (such as whether the borrower voluntarily disclosed the purpose of the loan) to predict the borrower's default behavior. He believed that the credit score not only relied on standard hard information, but soft information was more important than hard information in screening borrowers with low credit quality. Ge *et al.* [11] the social deterrence as a new potential mechanism to explain the default behavior of the borrowers, the borrower of loan data and popular social media sites combined state of social media data, the study found that the use of social media marketing activities can improve network platform's reputation and social media disclosed the borrower loan account information, can obviously reduce the loan default rates.

Gradually increased in recent years, the problem of P2P lending platform, the main cause of this problem is caused by the borrower overdue behavior caused by not getting paid on time, and P2P platform can't afford the risk of default, leaving investors interest is damaged, so to strengthen the credit of the borrower risk assessment is the key to the healthy development of the P2P lending platform. Early scholars mainly adopted traditional linear models, among which Logistic regression model was the most widely used. Wigintio [12] first applied Logistic model to credit score, believing that Logistic model had stronger explanatory power than linear discriminant model. Bekhet *et al.* [13] established Logistic regression model and radial basis function model to conduct credit score for Jordan commercial bank, and the results showed that the accuracy of Logistics

regression model was better than that of radial basis function, but the radial basis function model was better in identifying defaulting customers. In recent years, machine learning and artificial intelligence models have been gradually applied, which mainly include neural network [14], random forest [15] and support vector machine [16]. Malekipiribazari *et al.* [15] proposed a classification method based on random forest to predict the borrower status based on the data of Lending Club, and the results showed that the random forest method was better than FICO credit rating and LC grade in identifying excellent borrowers. Antonio *et al.* [17] used the sample data of 5500 borrowers from Peruvian microfinance institutions to construct the neural network credit score model based on the multi-layer perceptron, and compared the performance with the traditional linear discriminant analysis, quadratic discriminant analysis and Logistic regression model. The study found that the neural network credit score model was superior to the other three classical technologies. Xia *et al.* [18] first for data preprocessing, and then based on the characteristics of relative importance to eliminate redundant variables, and finally using Bayesian parameter optimization super parameter adaptive adjustment XGBoost, build sequence based on gradient enhanced machine integration credit scoring model, the accuracy of the confusion matrix, the error rate of the results show that the Bayesian parameter optimization model of performance is better than that of random search and web search and manual search. Neagoe *et al.* [19] used the German credit data set and the Australian credit data set to construct the credit score model (DCNN) based on the neural network classifier, and made a comparative analysis with the multi-layer perceptron (MLP) to evaluate the performance of the model through the confusion matrix.

Based on the above, we find that there are many factors influencing the borrower's default, but there are few researches on how to choose the variables that can significantly affect the borrower's default behavior scientifically and rationally. Therefore, this paper starts from variable selection and combines several classification models to study the default recognition and prediction of P2P borrowers. The marginal contributions of this paper are as follows: 1) at present, recursive feature elimination method is widely used in the fields of biology and medicine, and there are few researches on P2P network lending. This paper USES recursive feature elimination method to screen and rank high-dimensional variables. 2) The simulation effect and prediction accuracy of the classification model are compared, which provides references for the selection of P2P credit risk assessment model in the future.

3. Models and Methods

3.1. Classification Model

In this paper, logistic regression model, decision tree and BP neural network are used to identify and predict the default behavior of P2P borrowers. Logistic regression model [12] [13] is a classical classification method among statistical

learning methods, which is widely used in the prediction of default risk of P2P borrowers. Decision tree [18] and BP neural network [14] [17] are favored by more and more scholars due to their high prediction accuracy and high dimensional data processing.

1) Logistic regression model

Logistic regression model is a classical classification model, which uses the basic information of the borrower and the data of the loan information for analysis. The model is as follows:

$$g(z)=\frac{1}{1+\mathrm{e}^{-z}} \tag{1}$$

where Z can be estimated by the following multiple regression equation:

$$Z=\beta_0+\beta_1 X_1+\cdots+\beta_p X_p \tag{2}$$

If the probability of credit default of the borrower is $P(Y=1\mid X)=\pi(Y)$, the following formula can be obtained:

$$P(Y=1\mid X)=\frac{\exp\left(\beta_0+\beta_1 X_1+\cdots+\beta_p X_p\right)}{1+\exp\left(\beta_0+\beta_1 X_1+\cdots+\beta_p X_p\right)} \tag{3}$$

where Y is a binary variable, 1 represents default and 0 represents non-default. The maximum likelihood estimation method is adopted for parameter estimation, and the formula is:

$$l(\beta)=\prod_{i=1}^{p}\pi(Y_i)^{Z_i}\left[1-\pi(Y_i)\right]^{1-Z_i} \tag{4}$$

where $i=1,2,\cdots,p$, the formula of the de-log-likelihood function is as follows:

$$\log\left(l(\beta)\right)=\sum_{i=1}^{p}\left[\left(Z_i\log\left(\pi(Y_i)\right)\right)+(1-Z_i)\log\left(1-\pi(Y_i)\right)\right] \tag{5}$$

The goal of the maximum likelihood estimation method is what is the value of β when $l(\beta)$ takes the maximum value. In this paper, the gradient rise method is used to calculate the value of parameter β.

2) CART decision tree

At present, CART decision tree is one of the most widely used decision learning methods. The CART decision tree USES a gini index to partition attributes. The gini index represents the uncertainty of set D. The larger gini index is, the greater the uncertainty of set is. The formula of gini index is as follows:

$$Gini(D)=1-\sum_{k=1}^{k}\left(\frac{|C_k|}{|D|}\right)^2 \tag{6}$$

where C_k represents the sample subset of the k class, and k is the number of classes. If D is divided into D_1 and D_2 under feature A, then the gini index of set D is:

$$Gini(D,A)=\frac{|D_1|}{|D|}Gini(D_1)+\frac{|D_2|}{|D|}Gini(D_2) \tag{7}$$

3) BP neural network

The learning process of BP neural network includes information forward transmission and error reverse transmission BP neural network has the following structure shown in Figure 1, $X_1, X_2, \cdots, X_n$ is the input information of input layer, in P2P lending, the input information are the borrower's loan amount, loan interest rate, loan term and housing about ownership, etc., w_{ij} is the weight vector of input layer to hidden layer, w_{jk} is the hidden layer to output layer weights vector, $Y_1, Y_2, \cdots, Y_n$ is the BP neural network predictive value. A three-layer BP neural network is used in this paper. The output case is whether or not the borrower defaults, which is a dichotomy, 0 is non-default, 1 is default, so the output layer has only two neurons. For the determination of the number of hidden layer nodes, the following formula is adopted in this paper:

$$m = \sqrt{n+l} + \alpha \tag{8}$$

where m is the number of nodes in the hidden layer, n is the number of nodes in the input layer, and l is the number of nodes in the output layer, and α is a constant between 1 and 10.

3.2. Variable Selection Methods

1) Recursive feature elimination

Recursive feature elimination method (Recursive feature elimination, RFE) the main idea is repeated build model (SVM or the regression model) and choose the best (or worst) features, the selected feature selection, and then repeat the process on the characteristics of the residual, until all the characteristics of the traverse the select key characteristics. RFE adopts the feature sorting technology to select the feature subset. In this paper, the classification performance of SVM is taken as the evaluation function to select the feature. The flow chart of RFE is shown in Figure 2.

2) Pearson correlation coefficient

Pearson correlation coefficient is a linear correlation coefficient. Pearson correlation coefficient is a statistic used to reflect the degree of linear correlation between two variables. The correlation coefficient is represented by r, where n is the sample size, r describes the degree of strong linear correlation between the

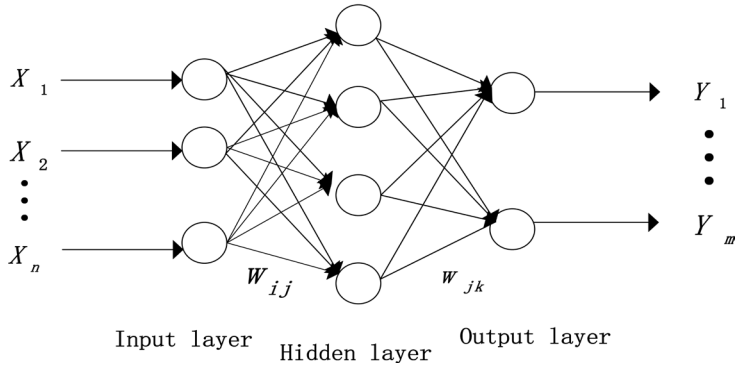

Figure 1. Structure diagram of BP neural network.

two variables, and the greater the absolute value of r is, the stronger the correlation is. The formula is as follows:

$$r_{kj} = \frac{\sum_{i=1}^{n}\left(x_{ik} - \overline{x}_k\right)\left(x_{ij} - \overline{x}_j\right)}{\sqrt{\sum_{i=1}^{n}\left(x_{ik} - \overline{x}_k\right)^2 \sum_{i=1}^{n}\left(x_{ij} - \overline{x}_j\right)^2}} \tag{9}$$

where r_{kj} represents the correlation coefficient between the k index and the j index, x_{ik} represents the i index value of the k index, $\overline{x}_k$ represents the average value of the k index, represents the average value of the third sample. x_{ij} represents the i index value of the j sample, and $\overline{x}_j$ represents the average value of the j index.

3.3. Model Performance Evaluation Method

1) Confusion matrix

In the two classification problems, there are four types of prediction results. Based on the confusion matrix, we can preliminarily judge the normal rate and error rate of the model, as shown in Table 1.

Where, TN represents the true category of samples as negative, and the number of samples judged as negative; FP represents the number of samples whose real category is negative, but which are judged to be positive. FN means the real category of the sample is positive, but the number of samples is judged as negative. TP represents the number of samples whose real category is positive and predicted to be positive. According to Table 1, the accuracy of the whole model

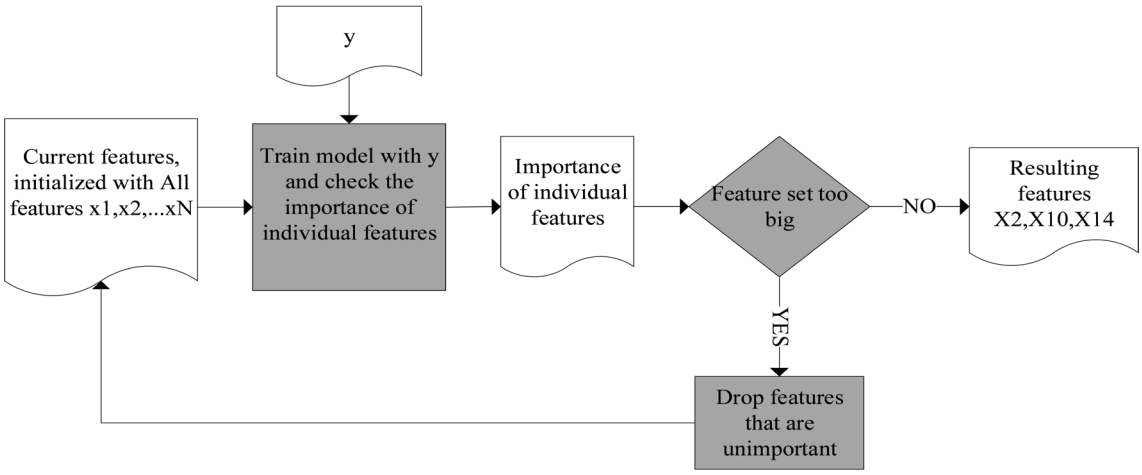

Figure 2. RFE flow chart.

Table 1. Confusion matrix.

		predicted	
		negative	positive
actual	negative	TN (True Negative)	FP (False Positive)
	positive	FN (False Negative)	TP (True Positive)

can be calculated, which can be expressed as:

$$\text{Accuracy} = \frac{TN + TP}{TN + FN + FP + TP} \tag{10}$$

2) ROC curve and AUC values

In binary classification problems, ROC curve and AUC value are often used to evaluate the merits of binary classifier. ROC curve is a comprehensive indicator of sensitivity and specificity of continuous variables. The horizontal axis of the ROU curve is the false positive rate (FPR), that is, the proportion of all negative cases in the partition instance, and the vertical axis is the true positive rate (TPR), that is, the proportion of all positive cases in the partition instance. For a binary classification problem, the instance is divided into a positive class (positive) or a negative class (negative). But in practice, there are four things that happen when you classify. a) If an instance is a positive class and is predicted to be a positive class, it is a True Positive class (TP); b) If an instance is a positive class, but is predicted to be a Negative class, it is False Negative (FN); c) If an instance is a negative class, but is predicted to be a positive class, it is False Positive (FP); d) If an instance is a Negative class, but is predicted to be a Negative class, it is True Negative (TN).

Among them, the True Positive Rate (TPR) represents the proportion of the actual positive instances in the positive classes predicted by the classifier to all positive instances. The formula can be expressed as:

$$TPR = \frac{TP}{TP + FN} \tag{11}$$

False Positive Rate (TPR) represents the proportion of the actual negative instances in the positive class predicted by the classifier to all negative instances:

$$FPR = \frac{FP}{FP + TN} \tag{12}$$

AUC (Area Under Curve) is defined as the Area Under ROC Curve. The closer the value of AUC is to 1, the better the stimulation effect of the model is.

4. Experimental Analysis

4.1. Data Source

In this study, Lending club provided data on borrowing targets in the third quarter of 2017 for experiments (data source URL: https://www.lendingclub.com/info/download-data.action). The platform of Lending Club asks customers to fill in the loan application form online or offline to collect the basic information of customers, including the applicant's age, gender, marital status, educational background, loan amount, and the applicant's property status, etc. Generally speaking, it also USES the information of third-party platforms such as credit investigation agencies or FICO. The original data in this paper contains 122,703 samples and 112 attributes. The classification label attribute "loan_stauts" has seven states In the original data: "Current",

"Fully Paid", "Charged Off", "Default", "In Grace Period", "Late (16 - 30 days)" and "Late (31 - 120 days)". Since "Current" belongs to the loan bid under repayment, and the repayment situation is not clear, the loan bid under "Current" status is deleted in this study. The final data of this paper includes 24,330 samples, including 16,770 non-default samples and 7560 default samples. The training set contains 70% data, and the test set contains 30% data. "Fully Paid" is regarded as non-default and the value is 0; "Charged Off", "Default", "In Grace Period", "Late (16 - 30 days)" and "Late (31 - 120 days)" shall be deemed as breach of contract, with the value of 1.

4.2. Variable Selection

Firstly, the attributes that have many missing values, most of the observed values are the same and have no significance to affect the borrower's default status were deleted. After preliminary screening, the remaining 87 attribute indexes were found. This paper adopts recursive feature elimination method to select key features. RFE uses the feature sorting technique to select the feature subset and takes the classification performance of SVM as the evaluation function. In this paper, the recursive feature elimination algorithm is used to screen out 15 features with the strongest correlation with the target variable. Then the Pearson correlation map is used to find out the redundancy features. From **Figure 3**, it can be found that the correlation coefficients of "installment", "grade" and "loan_amnt" are very high, respectively 0.95 and 0.98, which indicates that the two features of "installment" and "grade" have a strong correlation with the feature of "loan_amnt". Therefore, we need to delete the two feature attributes to avoid data multicollinearity, and the correlation Numbers of "total_rec_prncp", "last_pymnt" and "total_pymnt_amnt" are also very high.0.99 and 0.91, respectively, so two feature genera need to be removed from these three feature attributes as well.

Finally, the important features are sorted by the random forest algorithm, and the results are shown in **Figure 4**. The importance order of 15 variables is screened by recursive feature elimination method. The eigenvalues of "total_rec_prncp" and "last_pymnt" are greater than those of "total_rec_prncp" and "last_pymnt". According to the correlation of the three attributes mentioned above, two features need to be deleted to avoid data multicollinearity. Therefore, "total_rec_prncp" and "last_pymnt" are deleted from the three features. Similarly, it can be seen from **Figure 4** that the eigenvalue of "loan_amnt" is larger than that of "installment" and "grade", so the two features of "installment" and "grade" are deleted from the three features. For the remaining properties, "home_ownership_MORTGAGE," "home_ownership_OWN," and "home_ownership_RENT" belong to the "home_ownership" property. "Term_36 months" and "term_60 months" belong to the "term" attribute; "Verification_status_verification" belongs to the "verification_status" attribute.

Through the above, 9 attribute values were selected from 112 attributes in the final paper. The variable descriptions for the nine attributes are shown in **Table 2**.

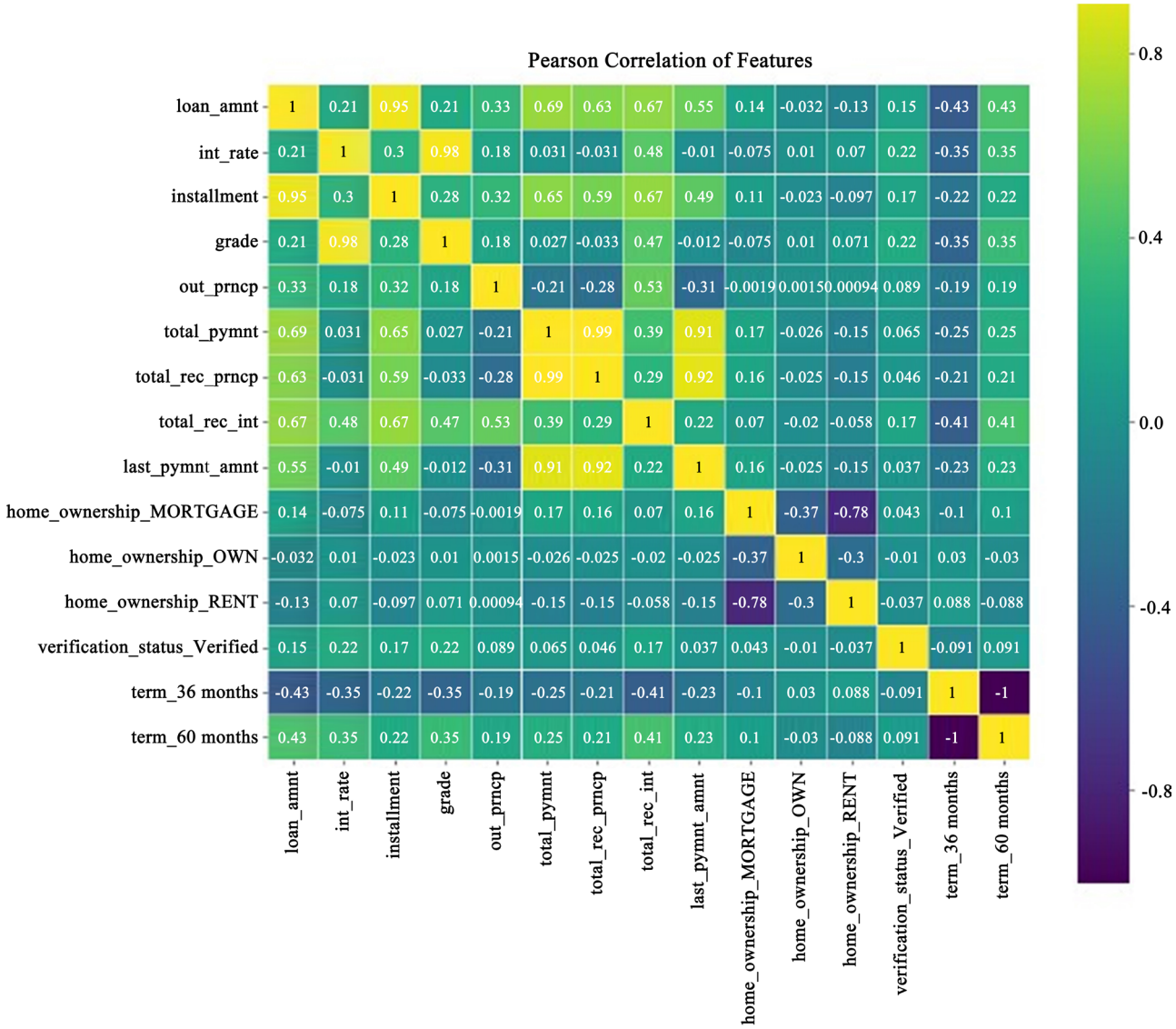

Figure 3. Pearson correlation coefficient diagram.

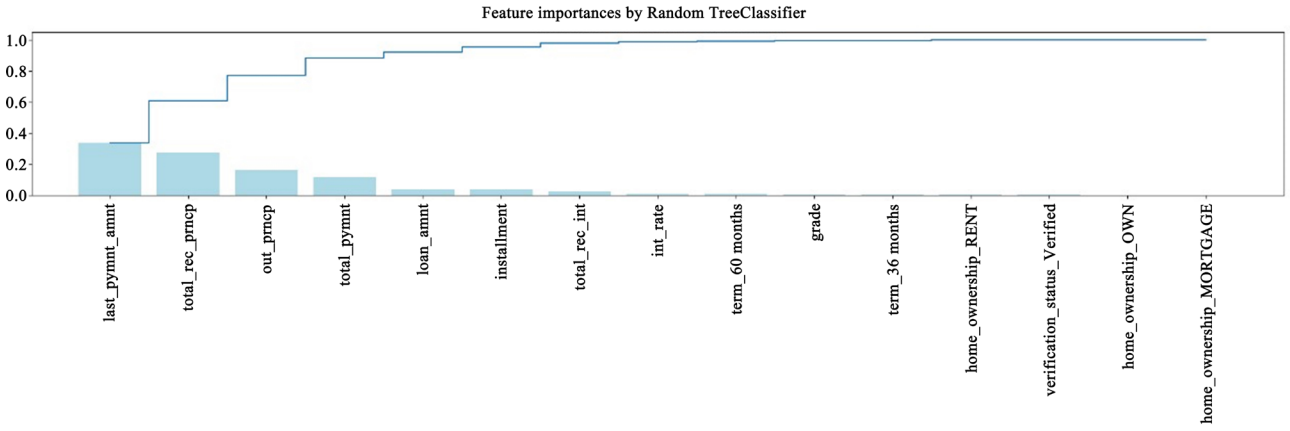

Figure 4. Important feature ordering diagram.

4.3. Analysis of Experimental Results

1) Analysis of ROU curve and ACU value

In this study, the ROU curve mentioned above is used to evaluate the advantages and disadvantages of the model. The detailed results of the ROC curves of the training samples and test samples of the logistic regression model, CART decision tree and BP neural network model in this study are shown in Table 3 below. As can be seen from Table 3, the ACU values of both the training samples and test samples of the Logistic regression model are 0.991, which is close to 1. This indicates that the Logistic regression model has a good simulation effect after screening important variables by recursive feature elimination method in this paper. The ACU values of the training samples and test samples of the CART decision tree were 0.957 and 0.965, respectively. Compared with the Logistic regression model, the ACU values of the CART decision tree were lower, but the AUC values of the training samples and test samples were both greater than 0.95,

Table 2. Description of variable properties.

Variable	Variable declaration	Variable types
loan_status	Current status of the loan	nominal variable
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.	continuous variable
int_rate	Interest Rate on the loan	continuous variable
out_prncp	Remaining outstanding principal for total amount funded	continuous variable
total_rec_int	Interest received to date	continuous variable
last_pymnyt_amnt	Last total payment amount received	continuous variable
home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER	nominal variable
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified	nominal variable
term	The number of payments on the loan. Values are in months and can be either 36 or 60.	nominal variable

Data source URL: https://www.lendingclub.com/info/download-data.action.

Table 3. Areas below the ROC curve.

Models	Samples	AUC value
Logistic	training samples	0.991
	test samples	0.991
CART	training samples	0.957
	test samples	0.965
BPNN	training samples	0.991
	test samples	0.992

indicating that the simulation effect was also good. The difference between the AUC values of the training samples and test samples of BP neural network algorithm and the ACU values of the training samples and test samples of Logistic regression model is not significant, which are 0.991 and 0.992 respectively. From the above analysis, it can be seen that Logistic regression model and BP neural network model have the best simulation effect among the four classification models after filtering out important variables by recursive feature elimination method, followed by CART decision tree. However, in general, the ACU values of the training samples and test samples of these three classification models are all greater than 0.9. From the definition of ACU, their simulation effects are very good.

2) Analysis of model results

In this paper, the above mentioned logistic regression model, CART decision tree and BP neural network model are respectively used to predict the credit default situation of borrowers. The accuracy of these three classification models in predicting borrowers' credit default in training samples and test samples is shown in Table 4.

As can be seen from the results in Table 4, the 8 variables finally screened by the recursive feature elimination method were input into the three classification models in this paper, and the prediction effect of the three classification models was very good. In a Logistic regression model, the training sample will not the borrower default judgment is a default on the accuracy of 95.5% of the borrowers, borrowers will default judgment for the default of the borrower's accuracy is 95.2%, the test samples of the time are divided into 95.9 and 95.1, and the training samples of the time difference is not big, it shows that in this article the classification of the Logistic regression model effect is very good, and its robustness is very good also. In the CART decision tree model, the accuracy rate of judging non-defaulting borrowers as non-defaulting borrowers in the training samples was 96.8%, 96.2%, 95.6% and 97.4% respectively in the test samples. Compared

Table 4. Model prediction accuracy.

models	classification	training samples			test samples		
		0	1	correct rate	0	1	correct rate
	0	11,241	529	95.5%	4794	206	95.9%
logistic	1	256	5081	95.2%	108	2115	95.1%
	correct rate			95.4%			95.7%
	0	11,390	380	96.8%	4782	218	95.6%
CART	1	201	5136	96.2%	57	2166	97.4%
	correct rate			96.6%			96.2%
	0	11,198	572	95.1%	4785	215	95.7%
BPNN	1	183	5154	96.6%	79	2144	96.4%
	correct rate			95.6%			95.9%

with the Logistic regression model, the classification effect of the decision tree was better. In BP neural network model, the training sample, not the borrower default judgment is not the default of the borrower's accuracy is 95.1%, the default of the borrower to default of the borrower's accuracy is 96.6%, the test sample of the correct rates were 95.7% and 96.4%, respectively, CART decision tree in the training sample and test sample, will not the borrower default judgment for the default of the borrower's accuracy and the effect of the Logistic regression model about the same, However, the accuracy rate of the BP neural network model in judging defaulted borrowers as defaulted borrowers was higher than that of the Logistic regression model, which was 1.4% higher in the test sample and 1.3% higher in the test sample. In P2P network lending, judging defaulted borrowers as non-defaulted borrowers will bring more losses than non-defaulted borrowers. Therefore, the classification effect of BP neural network is better than Logistic regression model, but worse than the classification effect of CART decision tree. Compared with Logistic regression model, CART decision tree and BP neural network model.

4.4. Summary

In this section, through the recursive feature method, Prosper correlation coefficient method and random forest feature selection, the paper finally screened out "loan_amnt", "int_rate", "out_prncp", "total_rec_int", "last_pymnyt-amnt", "home_ownership", "verification_status" and "term", which had a great impact on "loan_status". As can be seen from Figure 4, among them, "last_pymnyt-amnt" and "total_rec_int" have the largest characteristic values, indicating that these two indicators have the greatest impact on the borrower's default behavior. Through the confusion matrix, ROU curve and ACU value, the logistic regression model, the decision tree and BPNN ACU CART value is greater than 0.9, and the model of classification accuracy is greater than 0.95, so the performance of the model of the three classification and classification accuracy is very good, because the borrowers will default to the default of the borrower losses than to the default of the borrower to default borrowing National People's Congress, according to Table 4 shows that Logistics regression model of the test sample will default judgment for the default sample for 206. There are 218 CART decision trees and 215 BPNN, because the classification effect of Logistics regression model to judge the default samples as non-default samples is better than that of CART decision and BPNN.

5. Conclusions

Taking the user data published by lending club as the research object, this paper uses the method of recursive feature elimination combined with classification algorithm to identify and predict the credit default of borrowers. In this paper, it is found that 1) the recursive feature elimination method can screen the key variables affecting the borrower's default status; then, by sorting the key variables

from large to small, it is found that the borrower's latest repayment amount, loan amount and loan interest rate have a great impact on the borrower's default status. Pearson coefficient indicates that the borrower's credit rating and income have a strong correlation with the borrower's loan amount. 2) The experimental results show that the classification model with the recursive feature elimination method to select key variables has high accuracy. In this paper, CART decision tree has the highest accuracy, indicating that it has the best classification effect. The classification accuracy of Logistic regression model and BP neural network model is slightly lower than CART decision tree, but the ACU value is higher than CART decision tree, indicating that their simulation effect is better than CART decision tree.

Although the combination of the recursive feature elimination method and the classification algorithm used in this paper is effective in identifying and predicting the credit default situation of P2P online borrowers, it can be studied in the following two aspects: 1) try to apply oversampling or undersampling methods deal with data imbalances and observe the corresponding processing effects; 2) use other classification algorithms, such as random forest, support vector machine and other machine learning algorithms to incorporate experimental comparisons, and compare various classification algorithms for identification and the accuracy rate of predicting the default status of P2P online borrowers.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Lin, M.F., *et al.* (2009) Judging Borrowers by the Company They Keep: Social Networks and Adverse Selection in Online Peer-to-Peer Lending.

[2] Kumar, S. (2007) Bank of One: Empirical Analysis of Peer-to-Peer Financial Marketplaces. *AMCIS* 2007 *Proceedings*, 305. https://aisel.aisnet.org/amcis2007/305

[3] Emekter, T. and Jirasakuldech, L. (2015) Evaluating Credit Risk and Loan Performance in Online Peer-to-Peer (P2P) Lending. *Applied Economics*, **47**, 54-70. https://doi.org/10.1080/00036846.2014.962222

[4] Lin, X., *et al.* (2017) Evaluating Borrower's Default Risk in Peer-to-Peer Lending: Evidence from a Lending Platform in China. *Applied Economics*, **49**, 3538-3545. https://doi.org/10.1080/00036846.2016.1262526

[5] Carlos, S.C., Gutiérrez-Nieto, B., López-Palacios, L., *et al.* (2015) Determinants of Default in P2P Lending. *PLoS ONE*, **10**, e0139427. https://doi.org/10.1371/journal.pone.0139427

[6] Chen, D.Y., Hao, L. and Xu, H. (2013) Gender Discrimination towards Borrowers in Online P2PLending. *WHICEB* 2013 *Proceedings*, 55. https://aisel.aisnet.org/whiceb2013/55

[7] Serrano-Cinca, C. and Gutiérrez-Nieto, B. (2016) The Use of Profit Scoring as an Alternative to Credit Scoring Systems in Peer-to-Peer (P2P) Lending. *Decision Support Systems*, **89**, 113-122. https://doi.org/10.1016/j.dss.2016.06.014

[8] Puro, L., Teich, J.E., Wallenius, H., *et al.* (2010) Borrower Decision Aid for People-to-People Lending. *Decision Support Systems*, **49**, 52-60. https://doi.org/10.1016/j.dss.2009.12.009

[9] Guo, Y.H., Zhou, W.J., Luo, C.Y., Liu, C.R. and Xiong, H. (2016) Instance-Based Credit Risk Assessment for Investment Decisions in P2P Lending. *European Journal of Operational Research*, **249**, 417-426. https://doi.org/10.1016/j.ejor.2015.05.050

[10] Iyer, R., Khwaja, A., Luttmer, E.F. and Shue, K. (2016) Screening Peers Softly: Inferring the Quality of Small Borrowers. *Management Science*, **62**, 1554-1577. https://doi.org/10.1287/mnsc.2015.2181

[11] Ge, R., Feng, J., Gu, B., *et al.* (2017) Predicting and Deterring Default with Social Media Information in Peer-to-Peer Lending. *Journal of Management Information Systems*, **34**, 401-424. https://doi.org/10.1080/07421222.2017.1334472

[12] Wiginton, J. (1980) A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior. *The Journal of Financial and Quantitative Analysis*, **15**, 757-770. https://doi.org/10.2307/2330408

[13] Bekhet, H.A. and Eletter, S.F.K. (2014) Credit Risk Assessment Model for Jordanian Commercial Banks: Neural Scoring Approach. *Review of Development Finance*, **4**, 20-28. https://doi.org/10.1016/j.rdf.2014.03.002

[14] Hájek, P. (2011) Municipal Credit Rating Modelling by Neural Networks. *Decision Support Systems*, **51**, 108-118. https://doi.org/10.1016/j.dss.2010.11.033

[15] Malekipirbazari, M. and Aksakalli, V. (2015) Risk Assessment in Social Lending via Random Forests. *Expert Systems with Applications*, **42**, 4621-4631. https://doi.org/10.1016/j.eswa.2015.02.001

[16] Yao, X., Crook, J. and Andreeva, G. (2015) Support Vector Regression for Loss Given Default Modelling. *European Journal of Operational Research*, **240**, 528-538. https://doi.org/10.1016/j.ejor.2014.06.043

[17] Blanco, A., Pino-Mejías, R., Lara, J. and Rayo, S. (2013) Credit Scoring Models for the Microfinance Industry Using Neural Networks: Evidence from Peru. *Expert Systems with Applications*, **40**, 356-364. https://doi.org/10.1016/j.eswa.2012.07.051

[18] Xia, Y., Liu, C., Li, Y.Y., *et al.* (2017) A Boosted Decision Tree Approach Using Bayesian Hyper-Parameter Optimization for Credit Scoring. *Expert Systems with Applications*, **78**, 225-241. https://doi.org/10.1016/j.eswa.2017.02.017

[19] Neagoe, V.E., Ciotec, A.D. and Cucu, G.S. (2018) Deep Convolutional Neural Networks versus Multilayer Perceptron for Financial Prediction. *International Conference on Communications*, Bucharest, 14-16 June 2018, 201-206. https://doi.org/10.1109/ICComm.2018.8484751