Research on the Installment Risk of P2P Network Loan

Bo LI, Du-yu LIU*

Southwest Minzu University, Key Laboratory of Electronic and Information Engineering, State Ethnic Affairs Commission, Chengdu Sichuan, 610041 China
Email: bobo_8898@126.com

Abstract. The problems about some borrowers default in the rapid development of P2P network loan causes economic losses to online lending platforms and investors. Based on the situation of borrowers repaying loans in installments, this paper uses automatic binning to select features, builds a risk monitoring model, and predicts whether the borrower will perform next month. The model can discover the signs of borrower's default in advance, so that the platform can take preventive measures earlier and prevent the problem of platform fund circulation caused by insufficient repayment. In addition, it can provide reference for the platform to estimate the monthly payment amount. In this paper, the borrower data of 2016-2018 on Lending Club is used. The risk monitoring models of borrowers are based on CART algorithm, random forest algorithm and XGBoost algorithm respectively. The precision accuracy of the algorithms above is above 95%. The repayment amount of borrowers at last month, the borrower's occupation, the total amount of borrowing, the borrower's monthly repayment amount, and whether the borrower is working with the debt settlement company which are very effective in analyzing the willingness of the borrower to perform on time next month. Therefore, they could be used as the main basis for analysis.

Keywords: Peer to peer lending, CART decision tree, random forest, XGBoost

1 Introduction

P2P network loan is a business operation mode in a small amount of idle funds in the society is gathered which lend them to individuals or enterprise them with capital needs. It is also called unsecured micro-credit loans [1]. In recent years, the defaults and fraudulent behaviors of borrowers seriously damage the interests of investors and platforms and hinder the development of the P2P network loan [2]. This issue has attracted widespread attention in the industry and academia. At present, The research on P2P network loan mainly focuses on four aspects: (1) Research on Operation Mode and Related Mechanism of P2P [3][4]; (2) Research on Influencing factors of borrower default behavior [5]; (3) Research on the investment behavior bias of Investor [6]; (4) Build the default prediction model before loan[7]. Most researchers focus on the prediction model before lending, restrict risky applications, or reject higher-risk applications, but the risks already in borrowing are rarely studied. Therefore, considering the situation of borrowers repaying the loan in installments, this paper intends to conduct risk assessment analysis and forecasting of credits that have occurred, and provide reference for the platform. Platform can take preventive measures earlier and prevent the problem of platform fund circulation caused by insufficient repayment.

The structure of this paper as follows: In the second part, the principle of the algorithm are introduced, including: data binning, SMOTE, CART, Random Forest, XGBoost; in Section 3, data preprocessing and feature selection are described; in Part 4, The loan risk monitoring model is constructed. The prediction performance of each model and the weight distribution of each feature each model have been compared.

2 Related Theory

2.1 Data Binning

Data binning [8] is a process of discretization of continuous variables and combination of multi-state discrete variables into few-state discrete variables. Data binning has many advantages [9], such as: (1) The iterative speed of model is fast. (2) Robustness of outlier data is better.(3) The fitting ability of the model is stronger. (4) Model is easier to explain. (5) All independent variables are transformed to similar scales. This article uses python's woe package to automatical be binned the data, the principle is similar to the binary tree. For a single variable data set (denoted as D0), the data are sorted from small to large and divide it into 10 parts and each part dividing point is recorded. Traversing the IV value of each dividing points on the dichotomy of D0, if the maximum IV value that has been divided which is greater than the pre-divided IV value +0.01 (Default value, the user can set it by himself), then the data of D0 is divided with this point (D0 can be divided into DL and DR, D0 is replaced by DL and DR in turn. the data set can be divide by iterating to achieve the effect of binning), otherwise it would not been divided with this point.

WOE (Weight Of Evidence) indicates the logarithm of the ratio that the proportion of positive classes in D0 to the proportion of negative classes in D0 in the i-th box data after binning. The formula is as follows:

$$\mathrm{WOE(i)} = \ln\frac{\mathrm{good(i)}}{\mathrm{bad(i)}} \tag{1}$$

where i is the box number. The bad (i) is the ratio of the number of defaults in i box to the number of defaults in D0. The good (i) is the ratio of the number of positive classes in i box to the number of positive classes in D0.

IV (Information Value) is used to measure the amount of information of variable. The larger the IV value is, the more information the variable contains. The formula for calculating the IV value is as follows:

$$\mathrm{IV} = \sum_{\mathrm{i=1}}^{\mathrm{N}} \big(\mathrm{good(i)} - \mathrm{bad(i)}\big) * \mathrm{WOE(i)} \tag{2}$$

where N is the number of boxes.

The number of the IV value represents the predictive power of a variable. Usually, if the IV of a variable is less than 0.03, the variable is considered to have no predictive ability; if the IV of a variable is Greater than or equal to 0.5, the variable is considered to have a great predictive power.

2.2 SMOTE Algorithm

Category imbalance exists in many scenarios, such as fraud detection, risk identification, data concentration fraud, and high-risk samples. It is much smaller than other types of samples. Some machine learning algorithms tend to lean toward classes with large numbers when building models if the data set which is Unbalanced categories is used directly. The learning is not enough for the classes with small number, the effect of learning will not achieve the goal and the model will lose efficacy.

The full name of SMOTE is Synthetic Minority Oversampling Technique, which is an improved scheme for random oversampling algorithms [10]. The SMOTE algorithm analyses a small number of samples and according to the samples add some new samples by artificial synthesis to a balance data set [11]. The basic flows of the SMOTE algorithm are as follows:

1. For each sample X in a few classes, the distance from all samples in the sample set will be calculated by Euclidean distance to obtain the K nearest neighbor of X (the nearest K samples from X).
2. According to the imbalance ratio of the sample class, SMOTE algorithm will set a sampling rat N. for each minority sample X, N neighbor samples will be set randomly from its K nearest neighbors.
3. For each randomly selected neighbor Xn, a synthetic sample X_new is constructed with the X sample according to the following formula.

$$\mathrm{X\_new} = \mathrm{X} + \mathrm{rand(0,1)} * |\mathrm{X} - \mathrm{Xn}| \tag{3}$$

This indicates that a random point is taken as a synthetic sample X_new on the line connecting the sample X and the neighbor Xn.

2.3 CART Decision Tree

The decision tree classification [12] algorithm is an instance-based induction learning method. The algorithm extracts the tree structure from the training sample set: the tree is composed of nodes and leaves. The node records the classification judgment of a feature and leaf represents the last category. A classification path rule is formed between the nodes and the leaves. When testing a new sample, the root node of the tree begins. According to the judgment basis of the node, proceed to the next node along the corresponding path rule to continue testing until reaches a certain leaf. The category of the leaf is the category of the new sample.

The decision tree generated by CART is a simple binary tree. The binary recursively divides the data set. One node only divides the data set into left and right sub-parts, and enters the left sub-tree and the right sub-tree respectively. The CART classification decision tree divides dataset based on the Gini coefficient. The Gini value is defined to be the dataset. The smaller the Gini value is, the higher the purity of the sample is. If the sample set is D, the Gini value of D is calculated as the follows:

$$\mathrm{Gini(D)} = \sum_{k=1}^{|y|} \sum_{k' \neq k} p_k p_{k'} = 1 - \sum_{k=1}^{|y|} {p_k}^2 \tag{4}$$

where D represents the data set. k represents the kth category in D. $|\mathrm{y}|$ are the number of categories in D, and $\mathrm{p_k}$ represents the proportion of the kth category in D.

The Gini coefficient is defined to be the feature value, which reflects that feature A is used to divide the data set and improve the purity of the data. The smaller the Gini coefficient is, the better the data set is segmented according to this feature, and the better the purity of the data set is. The data set D is segmented according to feature A. It's Gini coefficient is calculated as follows:

$$Gini(\mathrm{D}, \mathrm{A}) = \frac{|D_1|}{D} Gini(D_1) + \frac{|D_1|}{|D_2|} Gini(D_2) \tag{5}$$

A represents the segmentation judgment of feature A. $\mathrm{D_1}$ and $\mathrm{D_2}$ represent left subtree data set and the right subtree data set when the data set has been divided two parts, respectively.

2.4 Random Forest

Random Forest (RF) is an algorithm that integrates results of multiple trees by Ensemble Learning. Its basic unit is the decision tree [1]. The algorithm use bagging to extract a certain number of samples from the original data set as a training set of a small tree. On the original data set, a number of small trees is constructed in the above way. Finally, by voting, the majority votes are taken as the final classification result.

There are two factors that affect the performance of random forest classification. (1) The number of small trees (2) the number of features selected for building small trees. In a certain range, as the number of small trees increases, the classification ability of random forests will increase. When the number of small tree features increases, the correlation between small trees increases and the classification performance of random forests will also improve. However, increasing the number of small trees and the number of features will increase the computational load of the algorithm. Moreover, when the number of small trees and the number of features m increase beyond a certain range, it is limited to improve model performance by increasing the number of small trees. Therefore, the optimization of random forests is usually to find smaller number of small trees and smaller number of features m [13].

2.5 XGBoost Algorithm

XGBoost, also called extreme Gradient Boosting, is a gradient-enhanced integration algorithm based on CART decision tree [14]. The basic idea is to combine multiple decision tree models. A value is randomly set as the initial prediction value of all data. The distance between the predicted value and the true value of the data is used to be as a reference for the spanning tree to construct a decision tree that shortens these distances. The constructed tree aims to reduce the distance between the predicted value of the previous tree and the true value of the data. The new tree is constructed through iteration continuously by using the idea of gradient descent, so that the predicted value is closer to the real value and achieves the purpose of learning.

XGBoost uses the gain as a criterion when segmenting data sets. The formula is as follows:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \tag{6}$$

where $\frac{G_L^2}{H_L+\lambda}$ represents the score of the left leaf after division. $\frac{G_R^2}{H_R+\lambda}$ represents the score of the right leaf after division, $\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}$ indicates the score before the node partitioned. $G_j = \sum_{i\in I_j} g_i$, $H_j = \sum_{i\in I_j} h_i$, $g_i = l'\left(y_i, \hat{y}_i{}^{T-1}\right), h_i = l''\left(y_i, \hat{y}_i{}^{T-1}\right)$, λ is the variable introduced when defining the complexity of the decision tree. It represents the smoothness degree of leaf weight and the range of values is [0, ∞]. γ indicated in the complexity cost of dividing the node into leaves. The data of the segmentation data is higher than γ to divide the data. The larger the γ is, the more conservative the model is.

3 Data Preprocessing and Variable Selection

3.1 Data Sources

The Lending Club was launched in 2007 and was successfully listed on the New York Stock Exchange in 2014. Since its establishment, it has accumulated a large amount of loan information. It is currently one of the largest P2P platforms in the US and accounts for 65% of P2P market of US. . The data of this article has based on the Lending Club Loan Data from kaggle's public dataset, which was updated in March 2019 and contains borrowing data from 2007 to 2018 in the last, with more than 2 million instances. The data set contains 144 fields and a category label: loan_status. The distribution of the original category of loan_status in the past 3 years is shown in the following table.

Table 1. Distribution table of raw data on Loan_status

Category	Amount
Fully Paid (No default record, full payment of arrears)	1041952
Current (The current state is normal)	919695
Charged off (The repayment task has not been executed for 120 days, and the arrears may not be repaid.)	261655
Late (31-120 days) (Reimbursement has not been performed within 31 to 120 days)	24897
In Grace Period (Overdue but repaid all arrears)	8952
Late (16-30 days) (Reimbursement has not been completed within 16 to 30 days)	3737
Does not meet the credit policy. Status: Fully Paid (Borrowing is not in compliance with the law, the borrower has never breached the contract, and the debt is paid in full.)	1988
Does not meet the credit policy. Status: Charged Off (Borrowing is not in line with the law, and borrowers are likely to not repay)	761
Default (Amount owed will not be repaid)	31

The object of this paper is the amortized loan and the non-defaulting borrower in the repayment state. Based on the fixed information of the borrower and the repayment record of last month (Excluding cumulative record characteristics, separate analysis of information contained in a single month), this paper predicts whether the borrower can fulfill the contract on time in the next month and excavates the factors affecting the performance of the borrower in the next month. Therefore, in the Lending Club dataset, the default borrower with the loan_status marked Charged Off, Late, and Default is the negative class of the whole study. The normal performance borrower labeled loan_status as Fully Paid is the positive class of the whole study. According to the value of the issue_d field in the data, the borrower's repayment record from 2016 to 2018 is used to construct the borrower risk monitoring model and predict whether the monthly loaner will be able to perform repayment in the next month.

3.2 Data Preprocessing

The data selected in this paper contains more than 500,000 instances and 144 fields, which are divided into two categories. The positive class (denoted as 0) contains 370793 instances, the negative class (denoted as 1) contains 133401 instances. The fields whose default rate greater than 20% are removed and the remaining 86 fields of data are analyzed by the missingno drawing in python. The defaults of some fields, such as: term, grade, sub_grade... are as follows:

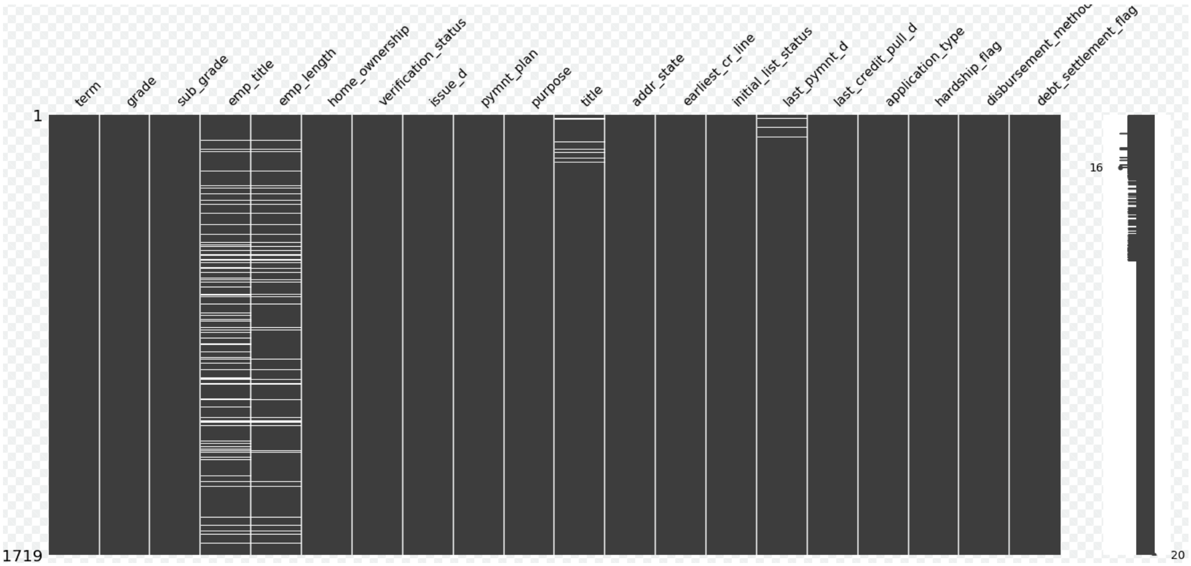


Figure 1. Partial field default distribution map

In the figure, black indicates a value, and white indicates a default. The abscissa corresponds to the field name. The ordinate starts from 1 and increases from top to bottom. It represents the index number of a single record in the data set. The rightmost thumbnail is a brief description of the default on the left side. The figure above shows the default of 20 fields in the first 1719 data. There are 16 fields without default values. The black columns corresponding to the emp_title field and emp_length field have more white, which means these two fields have more default values.

The data contains discrete fields and continuous fields. When the default value processing is performed, "UNKNOW" is inserted in the data blank for the discrete field, and the default position is inserted into the continuous field by the mode of the field. The python woe package is used to performing binning processing on the data processed by the default value, and the data_woe after the binning and the corresponding IV value of each variable are obtained. According to the IV value of each variable, the 10 variables with the most predictive ability were selected. The elastic network would be used to filter the variables. Finally, the following 9 variables were retained for building the model:

1) debt_settlement_flag
2) dti
3) emp_title
4) term
5) grade
6) total_bc_limit
7) last_pymnt_amnt
8) installment
9) loan_amnt

After above features automatically binned, the binning results of some fields (debt_settlement_flag, grade, installment) are visualized as follows:

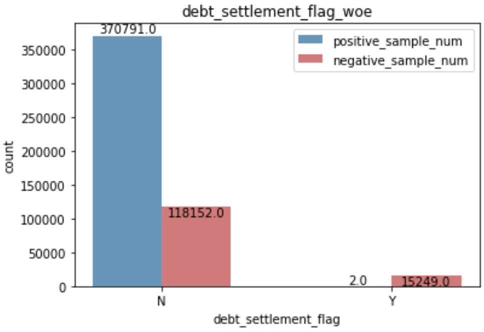


Figure 2. debt_settlement_flag binning result

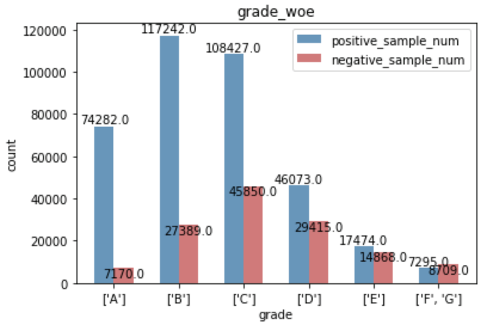


Figure 3. grade binning results

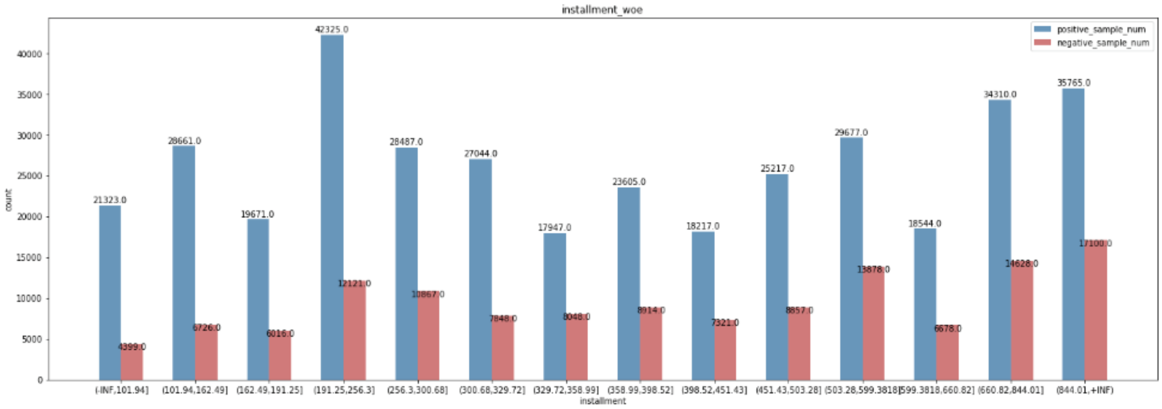


Figure 4. installment binning results

After the binning is completed, nine features re-selected in data_woe of binning data set. New_data_woe is built by nine features and prediction result loan_status. New_data_woe is used to build a risk monitoring model

3.3 Data Balance Processing

There are 370,000 positive samples in the new_data_woe data set, accounting for 73% of the total sample. Because the ratio of positive and negative class samples is unbalanced, SMOTE oversampling is used to insert synthetic positive class samples into data sets. Finally, the ratio of positive and negative samples on the new_data_woe_smo data set is 1:1. The data set new_data_woe_smo is split according to a ratio of 7:3, 70% of the data is used to build the model, and 30% of the data is used to test the performance of the model.

4 Case Analysis

In this paper, loan_status is used as a predictive classification category. category 1 indicates that the default has not be repaid by the borrower in recent months, and category 0 indicates that the borrower has repaid on time in recent months. CART-lender risk monitoring, random forest-lender risk monitoring, and XGBoost-borrower risk monitoring is constructed by the debt_settlement_flag, dti, emp_title, grade, term, total_bc_limit, last_pymnt_amnt, installment, and loan_amnt.

4.1 Related Parameter Selection

In the three models, CART-debtor risk monitoring does not require parameter selection; random forest-borrower risk monitoring mainly selects the appropriate number of trees and the value of the feature number m. max_features = 7, n_estimators = 300 are set in random forests by GridSearcherCV() of python; XGBoost-borrower risk monitoring has a large number of parameters. Usually, the default parameters are selected to obtain good prediction results. Therefore, several common parameters would be selected, and others adopt default values. The XGBoost parameters in this paper are selected as follows: max_depth=6; learning_rate=0.01; n_estimators=100; objective="binary:logistic";Gamma=0.1.

4.2 Model Evaluation

To evaluate the performance of each model, the classification performances of three models are compared from four aspects: precision, recall, f1-score and AUC (area under ROC curve). The results are as follows:

Table 2. Comparison of performance classification of each model

	Class	Precision	Recall	F1-score	AUC
CART- debtor risk monitoring	0	0.95	0.95	0.95	0.96
	1	0.95	0.95	0.95	
	avg/total	0.95	0.95	0.95	
RF- debtor risk monitoring	0	0.97	0.95	0.96	0.99
	1	0.95	0.97	0.96	
	avg/total	0.96	0.96	0.96	
XGBoost- debtor risk monitoring	0	0.99	0.92	0.96	0.99
	1	0.93	0.99	0.96	
	avg/total	0.96	0.96	0.96	

Compare the ROC curves of each model as follows:

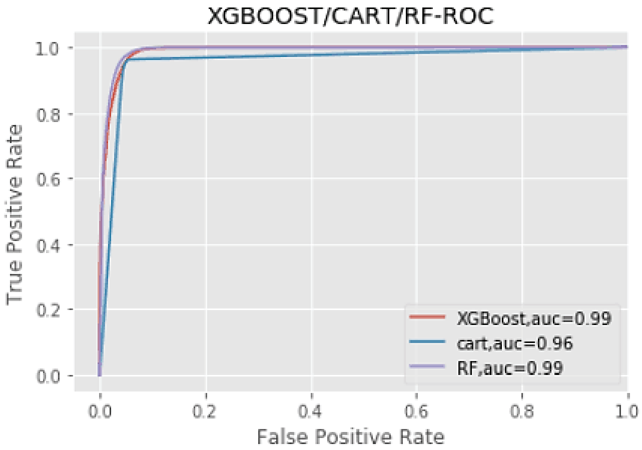
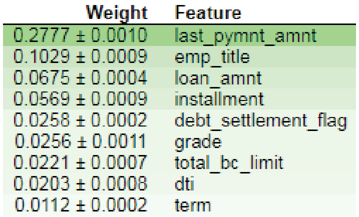


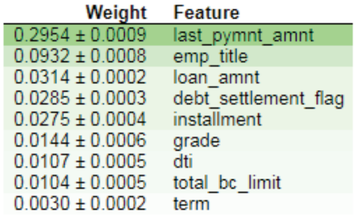
Figure 5. Comparison of ROC curves based on XGBoost/CART/RF model

Comparing the feature weights of each model from high to low is as follows:



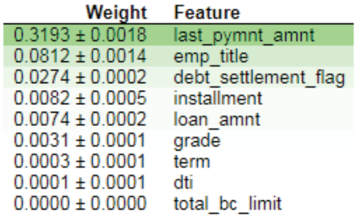
Weight	Feature
0.2777 ± 0.0010	last_pymnt_amnt
0.1029 ± 0.0009	emp_title
0.0675 ± 0.0004	loan_amnt
0.0569 ± 0.0009	installment
0.0258 ± 0.0002	debt_settlement_flag
0.0256 ± 0.0011	grade
0.0221 ± 0.0007	total_bc_limit
0.0203 ± 0.0008	dti
0.0112 ± 0.0002	term

Figure 6. CART-feature weight ranking



Weight	Feature
0.2954 ± 0.0009	last_pymnt_amnt
0.0932 ± 0.0008	emp_title
0.0314 ± 0.0002	loan_amnt
0.0285 ± 0.0003	debt_settlement_flag
0.0275 ± 0.0004	installment
0.0144 ± 0.0006	grade
0.0107 ± 0.0005	dti
0.0104 ± 0.0005	total_bc_limit
0.0030 ± 0.0002	term

Figure 7. RF-feature weight ranking



Weight	Feature
0.3193 ± 0.0018	last_pymnt_amnt
0.0812 ± 0.0014	emp_title
0.0274 ± 0.0002	debt_settlement_flag
0.0082 ± 0.0005	installment
0.0074 ± 0.0002	loan_amnt
0.0031 ± 0.0001	grade
0.0003 ± 0.0001	term
0.0001 ± 0.0001	dti
0.0000 ± 0.0000	total_bc_limit

Figure 8. XGBOOST-feature weight ranking

The precision of CART-lender risk monitoring, random forest-lender risk monitoring, and XGBoost-borrower risk monitoring are 95%, 96%, and 96%, respectively. Comparing with the ROC curves of each model, it can be concluded that XGBoost-borrower risk monitoring is close to RF-lender risk monitoring in predictive performance. The model based on integrated algorithm is better than CART-borrower risk in predicting performance. Comparing the distribution of feature weights among the three models, the top five features of feature weights are last_pymnt_amnt, emp_title, loan_amnt, installment, debt_settlement_flag, and the order of position is slightly different. The remaining four features have lower feature weights in the three models. In the XGBoost-borrower risk monitoring model, the feature weights of term and dti are close to 0, and the feature weight of total_bc_limit is 0.

5 Conclusion

Based on the performance record of borrower by Lending Club in 2016-2018, this paper constructs a borrower risk monitoring model and predict borrower whether performs the next month. The data is preprocessed before the model builds. The Woe package of Python is used to binning is used to each field of the record by the Woe package of Python. By calculating the IV value of each field, 10 initial features with the most predictive ability are selected. Nine features are selected from the elastic network: debt_settlement_flag, dti, emp_title, grade, term, total_bc_limit, last_pymnt_amnt, installment, loan_amnt. The CART-lender risk monitoring model, RF-lender risk monitoring model and XGBoost-borrower risk monitoring model are constructed with the above nine characteristics, and predicted on the test set. Finally, the prediction results of each loaner risk monitoring model are compared from the aspects of precision, recall, f1-score, AUC, ROC curve and ranking of feature weight.

The precision of all models are 95% accurate. RF-debtor risk monitoring and XGBoost-borrower risk monitoring based on integrated algorithms are better than CART-borrower risk monitoring. By comparing the rankings of each characteristic weight in the three models. it is concluded that the lender's repayment amount last month, the lender's occupation, the total amount of the loan, the monthly planned repayment amount of the lender and whether the lender cooperates with the debt settlement company to analyze whether the willingness of lender to perform on time next month is outstanding. Therefore, these can be used as the main reference factor. The credit rating, loan cycle, income and debt ratio, and bank loan total limit of borrower could be used as the secondary reference factors. Compared with many pre-lending risk prediction models of P2P networks, this paper proposes the interfering factors of lenders' staging risk, which has important reference value for the forecast of borrowers' willingness to perform in the next month.

References

1. QuYanting.Random Forest Prediction Model of P2P Network Lending Default[D].Chongqing University,2018.
2. Chaohui Wang. Credit Risk Assessment Model based on Feature Generation and Historical Records[D].Zhejiang University,2018.
3. Freedman S M, Jin G Z. Learning by Doing with Asymmetric Information: Evidence from Prosper.com[J]. NBER Working Papers, 2011:203--212.
4. XINGYU XHOU. Research on the operation mode of P2P network Lending industry—Taking Lufax as an example[D].Zhejiang University,2018.
5. CHEN Xiao DING Xiao-yu WANG Bei-fen. A Study of the Overdue Behaviors in Private Borrowing——Empirical Analysis Based on P2P Network Borrowing and Lending [J].Finance Forum,2013,18(11):65-72.
6. Shen D , Krumme C , Lippman A . Follow the profit or the herd? Exploring social effects in peer-to-peer lending[C]// IEEE Second International Conference on Social Computing. IEEE, 2010.
7. ZHANG Ning CHEN Qin. P2P loan default prediction model based on TF-IDF algorithm [J].Journal of Computer Applications,2018,38(10):3042-3047.
8. Zhang Mingjin Wang Mingwei. Use of Binning-based CARS method for feature selection from gene expression data [J]. Computers and Applied Chemistry, 2015, 32(8):001004-1006.
9. Yangqiujie. The Research on Random Forest Based on IV Feature Selection[D]. HeFei University of Technology, 2010.
10. Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic Minority Over-sampling Technique[J]. Journal of Artificial Intelligence Research, 2002, 16(1):321-357.
11. LIU Xiang-dong LI Fen. The Evaluation of the Borrower's Credit Risk in Peer-to-Peer Lending under the Background of Big Data:Evidence from RenRen Dai[J].Statistics&Information Forum,2016,31(05):41-48.
12. Breiman L , Firedman J H ,Olshen R A , et al. Classification and Regression Trees. Wadsworth , Inc.1984.
13. N. Schnitzler,P.-S. Ross,E. Gloaguen. Using machine learning to estimate a key missing geochemical variable in mining exploration: Application of the Random Forest algorithm to multi-sensor core logging data[J]. Journal of Geochemical Exploration,2019,205.
14. Chen T , Guestrin C . XGBoost: A Scalable Tree Boosting System[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016.