Research on personal credit scoring model based on multi-source data

Haichao Zhang, Ruishuang Zeng, Linling Chen* and Shangfeng Zhang

College of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou 310018, China

*Corresponding author’s e-mail: 392061026@qq.com

Abstract. In the Internet financial personal credit loan business, it is necessary to construct a credit scoring model for users, and the problems of unbalanced user categories, high data dimensions and sparse features make it difficult to model the credit situation of users. This paper adopts the idea of grouping modeling. It proposes an improved BIV value feature screening method and a weighted average model based on Logistic Regression, Random Forest and Catboost, which provides a set of solutions for user modeling in this scenario. The grouping modeling idea pre-groups the customers and reduces the feature sparsity problem. The improved BIV value shows the influence of each feature on the results and points out the mutation threshold. The oversampling method alleviates the category imbalance problem. AUC is used as the model result evaluation index, and the results show that the classification effect of the model is good. The results show that customers with a long history of credit history and a history of good credit behavior have lower credit risk.

1. Introduction

Internet finance has developed rapidly and has become an indispensable channel for individuals and small and micro enterprises to finance. Taking P2P network loans as an example, in January 2019, the P2P online loan industry reached a volume of 103.707 billion RMB in China. However, because the traditional financial credit system relies too much on the central bank's credit information system, and the amount of data stored by the central bank's credit information system is not enough to support the personal credit assessment in the context of the Internet, Internet finance has encountered difficulties in its rapid development. In 2018, a number of well-known P2P companies such as Union finance have closed down, and Internet finance companies such as Dig Fortune and 51 Credit Cards have been laid off. The development of the Internet finance industry has been questioned by all sectors of society. In order to solve this problem, many institutions have made full use of the advantages of high mobile payment penetration rate, and tried to apply big data to credit evaluation, including Alibaba's “Sesame Credit” and Tencent's “Tencent Credit”, etc., has been widely recognized and applied in the business community (Zhang, 2016). By mining the rich information behind big data, using the theory of machine learning, constructing a personal credit scoring model, and establishing a strong and robust credit risk assessment system to improve the management level of the Internet financial platform and ensure the healthy and sustainable development of Internet finance have a very important theoretical basis (Song, 2015; Liu and Qu, 2016).

Credit scoring has become the primary tool for credit risk assessment. The research of credit scoring system has gradually become a research hot-spot, and research methods are also advancing. Based on statistical theory methods, some scholars have begun to compare the application effects of
different credit scoring methods (Xiong Wei, 2009). The existing literature has done a good job in comparing the credit scoring ability of various models. Common methods are: Linear Discriminant Analysis (Libarati et al., 2017), Logistic Regression (Shi and Chang, 2015), Decision Tree (Gao and Wang, 2017), Support Vector Machine (Shen, 2005; Minghui, 2008), etc. (Benyacoub et al., 2014; Chen et al., 2018). However, scholars have not reached a consensus on the merits and demerits of various scoring methods (Shi and Jin, 2003, 2004; Li et al., 2007). Sometimes the conclusions are even completely opposite. The majority of this phenomenon is because there are differences in the collected credit data sets and the principles followed in the data cleansing process and the setting of the collective model parameters.

Wiginton (1980) showed that the credit score of Logistic Regression is superior to discriminant analysis. West (2000) also compared the credit scoring ability of various models, in addition to the common linear discriminant analysis, decision tree, K-nearest neighbor, Logistic Regression and other methods, it also includes five kinds of neural network models. The results show that Logistic Regression has achieved the highest prediction accuracy. Kan (2010) proposed to calculate the WOE and IV values, select all the indicators whose IV value is greater than a certain threshold, and then select certain indicators according to the size of the WOE complex correlation coefficient, and establish a Logistics Stepwise Regression model and a Logistic Regression model respectively. Li (2015) proposed that the credit risk assessment process faces complexity, non-linearity and uncertainty. The traditional evaluation methods adopted by commercial banks are difficult to apply. To this end, the Random Forest algorithm in the field of combined classification frontier research is applied to the evaluation process. The comparison with the traditional model evaluation results shows that the evaluation model achieves faster speed and higher evaluation accuracy, which effectively improves the evaluation efficiency. Dorogush et al. (2018) believe that the traditional model can not fully extract the effective information in the typed variables. For example, the Cart decision tree cannot process the discrete features of the string type. Therefore, it is necessary to convert the categorical variables and assign them to another model. And because the gradient lifting tree model is easy to over-fitting, it should be limited by complexity to enhance the generalization ability of the model. At this point, Catboost can effectively solve the above two problems.

Stacking&Blending Models is an effective technique to increase the accuracy of the model by increasing the diversity of the algorithm and reducing the generalization error. The most common Stacking&Blending method in practice is voting weighted, and only the established model is predicted on the test set, no retraining is required. In summary, this paper has obtained a series of data related to personal credit assessment, such as basic information, loan information and credit card information, from the Internet finance company. The Stacking&Blending Models method of weighted voting is used to fuse the prediction results of Logistic, Random Forest and Catboost models in order to obtain a more accurate credit scoring method.

The rest of this paper is organized as follows. Section 2 is data processing and index selection. Section 3 is introduction of Logistic Regression, Random Forest, Catboost and Weighted Average Regression Model. Section 4 is model solving and testing. Section 5 is model result analysis. Finally, the conclusions are presented in section 6.

2. Data and Indicators

2.1. Data Processing
This paper is based on credit evaluation related and desensitized data authorized by an internet finance company, and is obtained after confirming with the data provider that the data can be provided to the researcher for research purposes only and encrypted.

There are 24,660 samples of raw data, of which 26.4% are rated as risk customers and 73.6% are non-risk customers. Data content can be roughly divided into personal basic information and personal historical financial data. Not all customers have the same type of financial history behavior (for example, some customers have no loan experience, some customers have not done credit cards, etc.).
Personal historical financial data includes loan data, credit card data, overdue history, outstanding liabilities, specific business, and institutional inquiry summary tables of the target sample. Personal basic information includes household registration information and the first 6 digits of the ID number.

2.1.1. **Personal basic information processing.** According to the first 6 digits of the ID card, the user's birth address can be obtained and visually analyzed (the more red, the more the local number), the result shown in Figure 1 below. According to the work address zip code, user's birth address can be got, which is distributed as shown in Figure 2 below.

Figure 1 shows that the birthplaces of applicants for loan applications are mainly concentrated in the southeast coastal cities, mainly including Nanjing and southeastern Zhejiang and southeastern cities of Fujian. These cities are often more developed economies, and users have more opportunities for Internet finance. Figure 2 shows that the distribution of job sites for loan applicants is more uniform, but it is also densely distributed in economically developed cities (Beijing, Shanghai, Guangzhou, Shenzhen, etc.).

Due to the scattered distribution of users' cities, this paper evaluates the birthplace of users according to the economic development of the city. This paper evaluates the user's birthplace according to the economic development of the city, and uses the first- and second-tier cities as the criteria to create the following characteristics: birth place score (0 points in first-tier cities, 1 point in new first-tier cities, 2 points in second-tier cities, and 3 points in third-tier cities and below). According to the characteristics of the work city zip code, create characteristics: work site scores (0 points in the first-tier cities, 1 point in the new first-tier cities, 2 points in the second-tier cities, 3 points in the third-tier cities and below).

2.1.2. **Credit Information Processing.** The distribution of credit information customer numbers is shown in Table 1 below.

| Form name              | Proportion of customer numbers included (base 24660) | Original feature number | Overview                                          |
|------------------------|------------------------------------------------------|-------------------------|---------------------------------------------------|
| Loan information       | 84.70%                                               | 18                      | Historical loan status of target customer loans   |
| Specific business      | 38.90%                                               | 6                       | Target customers' early repayment and deferred behavior |
| Credit card information| 99.80%                                               | 13                      | Target customer's information on each credit card |
| Overdue history        | 79.40%                                               | 6                       | Target customer overdue repayment information     |
| Unliquidated liabilities| 99.70%                                               | 11                      | Target customer outstanding debt information      |

![Figure 1. Distribution of birthplaces of users applying for loans](image1.png)

![Figure 2. Distribution of users' work places for applying for loans](image2.png)
Institutional query summary 100% 11 Total number of queries by different agencies

From the data distribution in Table 1, we find that a considerable part of the table, only some customers appear, which means that there may be obvious user group stratification, and the difference between different groups is large, then It is unreasonable to put all users in the same frame (Bravo et al., 2015). Group modeling considers the sparseness of data features, and divides the sample into multiple sets of data according to the sparseness of the data. If the data is not divided into multiple groups, a large number of sample sparse cases will occur. For example, the presence of customers in a particular business table is only 38.9% of the total, i.e., only 38.9% of the samples have specific business behavior. If grouping is not modeled, missing values (usually using constant padding) need to be filled in order to make the model work. It will result in 61.1% of the samples showing zero or all other identical values in the specific business-related features, which seriously reduces the utility of all features in the part. Another advantage of grouping modeling used in this paper is that the samples can be pre-grouped to make the results more reasonable. That is, people who have had specific trading behaviors only compare with those who have had specific trading behaviors, making the model more refined.

2.1.3. Structural related features. This paper refers to the factors of interest in the US FICO scoring system and creates the following types of features based on raw data.

1) Repayment history: According to the information in the credit information table, the average loan repayment time, the total historical loan amount, the credit card repayment speed, the credit card history overdue total days, the credit card overdue amount, the forced early repayment status, the active early repayment status, the overdue repayment, the number of times and the overdue duration are created.

2) Number of credit accounts: According to the loan information table and the credit card information table, the number of loans, the number of institutions, the number of credit cards held, etc. are created.

3) Credit time limit class: According to the loan information table and the credit card information table, the time from the first loan and first credit card to the current time are created.

4) Credit type class being used: According to the loan information table and the credit card information table, the loan and credit card account types and the number of credit accounts of each type are created.

5) Other relevant information categories: average consumption and maximum consumption of credit card in the past 6 months, amount and time of deferred repayment of specific business, amount and time of prepayment of specific business, time of inquiry by the organization and corresponding reasons.

2.2. Indicator selection

According to the problem that the data is unbalanced and the dimension is high, in order to prevent the result of the inefficient feature interference model, this paper intends to use the BIV value (Binary Information Value) based on the IV value (Information Value) optimization for feature screening.

The IV value is a measure of the degree of influence of a single feature on the result label. The core is that the sample can be grouped by one feature, and the label distribution of the sample after the group is compared with the original label distribution. The larger the change, the larger the IV value of the feature. However, the IV value can only process the categorical variable and cannot process the continuous value. And as the categories in the variable increase, its IV value will only increase and not decrease. What’s more, if the frequency of a tag in the sample is 0, then ln(0) or ln(∞) will occur.

Therefore, this paper proposes the following solutions:

1) Refer to the Decision Tree algorithm for the processing of variables. After sorting the continuous value variables from large to small, each variable is segmented as a node to generate a
binary tree. The two ends of the binary tree are respectively classified as one type, which is converted into two categories of categorical variables.

(2) Since the number of categories in the variable increases, the IV value of the variable only increases and does not decrease. Therefore, when the variable is pre-processed, the variable is directly restricted. First, the category variable is converted into a dummy variable, so that the categories of categorical variables are only two. Secondly, for continuous variables, because of the Decision Tree theory, each continuity variable is only divided into two categories (greater than the threshold and less than the threshold). The advantage of this processing is that both the category variable and the continuous variable are divided into two categories, the number of categories is the same, and the variables can be directly compared.

(3) Based on the actual data, a threshold to prevent numerical overflow is added when calculating the probability. That is to say, if 0 occurs, this paper uses a super small number (0.0001) instead, so that the calculation can continue. The adjusted IV value is called BIV value (Binary Information Value).

3. Models and Methods

3.1. Logistic Regression

The model form of Logistic Regression appears as follows:

\[ z_i = \sum_{j=1}^{n} w_j x_{ij} + b, \ j \text{ is the eigenvalue} \]

\[ \sigma(z_i) = \frac{1}{1 + e^{-z_i}} \]

(1)

(2)

The cross entropy is expressed as a loss function as follows:

\[ L = \sum_{i=1}^{n} y_i \ln \sigma(z_i) + (1 - y_i)(1 - \ln(\sigma(z_i))), i \text{ is the number of samples} \]

Aim \[ \arg \min_{w} L \]

The above is the problem to be optimized by Logistic Regression. The gradient descent method and the chain derivation law are used to find the optimal solution.

\[ \frac{\partial L}{\partial w_{y_j}} = \frac{\partial L}{\partial z_i} \frac{\partial z_i}{\partial w_{y_j}} \]

Among them, \[ \frac{\partial L}{\partial w_{y_j}} = (y_i - \sigma(z_i)) \cdot x_{ij}, \frac{\partial L}{\partial z_i} = y_i - \sigma(z_i), \frac{\partial z_i}{\partial w_{y_j}} = x_{ij} \]

Under the condition of step \( \gamma \), the update strategy of parameter \( w \) is finally obtained as follows:

\[ w_{y_j} = w_{y_j} - \gamma(y_i - \sigma(z_i)) \cdot x_{ij} \]

(4)

(5)

3.2. Random Forest and Catboost algorithm

Ensemble learning refers to the combination of several weak classifiers to produce a strong classifier. Weak learners are classifiers whose classification accuracy is only slightly better than random guessing (error rate < 50%). The key to the success of the ensemble learning algorithm is to ensure the diversity of the weak classifier, and the more stable performance improvement can be obtained by integrating the unstable learning algorithm. Boosting and bagging are the basic algorithms in the field of ensemble learning in machine learning.

3.2.1. Bagging introduction. Bagging is the bagging method. The algorithm process is as follows:

(1) Extract the training set from the original sample set. Each round draws \( n \) training samples from the original sample set using a method of putting back random samples (in the training set, some samples
may be extracted multiple times, and some samples may not be drawn at once). A total of \( k \) rounds are taken to obtain \( k \) training set (The training sets are independent of each other).

(2) Each time a training set is used to obtain a model. A total of \( k \) models are obtained for \( k \) training sets (There is no specific classification algorithm or regression method, we can use different classification or regression methods according to specific problems, such as Decision Trees, Perceptrons, etc.)

(3) For the classification problem: the \( k \) models obtained in the previous step are classified by voting method. For the regression problem, the mean of the above models is calculated as the final result (all models have the same importance).

3.2.2. Random Forest Process. The process of the Random Forest algorithm used in this paper is as follows:

(1) The input is sample set \( D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\} \), and the weak classifier iteration number \( T \). The output is the final strong classifier \( f(x) \).

For \( t = 1, 2, \ldots, T \): the \( t \)th random sampling of the training set, a total of \( m \) acquisitions, to obtain a sample set \( D_m \) containing \( m \) samples, training the \( m \)th decision tree model \( G_m(x) \) with the sample set \( D_m \), when training the nodes of the decision tree model Select a part of the sample features among all the sample features on the node, and select an optimal feature among the randomly selected partial sample features to make the left and right subtree partitions of the decision tree.

(2) In the classification algorithm prediction, one of the categories or categories in which the \( T \) weak classifiers cast the most votes is the final category.

3.2.3. Boosting Introduction. The main idea of Boosting is to assemble weak classifiers into a strong classifier. Under the PAC (Probability Approximation Correct) learning framework, the weak classifiers must be assembled into a strong classifier.

Two core questions about Boosting:

(1) The weight or probability distribution of the training data will be changed after each round of change. By improving the weights of the samples classified by the weak classifier in the previous round and reducing the weight of the previous round of correct classification examples, the classifier has a better effect on the mis-classified data.

(2) Weigh the weak classifier by the error rate of the weak classifier. The weak classifiers are linearly combined by the addition model. For example, AdaBoost adopts a weighted majority voting method, that is, increases the weight of the classifier with a small error rate, and reduces the weight of the classifier with a large error rate.

3.2.4. Catboost Process. (1) Converting classification features into numerical features

The sub-type variable is connected with the dependent variable. According to the distribution of the dependent variable in each category, the sub-type variable is assigned and converted into a numerical variable. The calculation formula is as follows:

\[
\frac{\sum_{i=1}^{p} [x_{\sigma_j,k} = x_{\sigma_j,k}] Y_{\sigma_j} + a \cdot P}{\sum_{i=1}^{p} [x_{\sigma_j,k} = x_{\sigma_j,k}] Y_{\sigma_j} + a}
\]

(6)

\( p \) is the total number of samples, \( k \) is the feature number, \( Y_{\sigma_j} \) is the dependent variable of the \( j \)th sample, \( a \) is the weight of the a priori value, and \( P \) is the a priori value of the dependent variable.

(2) Model iterative algorithm

In order to alleviate the problem that the gradient lifting tree is easy to over-fitting, Catboost limits the model of each iteration, that is, only establishes a complete symmetric tree model, which has the advantages of increasing the prosperous ability of the model and speeding up the calculation of the model.
3.3. Weighted average model

In this paper, the model Stacking&Blending method of weighted voting is used to fuse the prediction results of the above three models, and the generalization error is reduced to reduce the accuracy of the model. The Stacking&Blending method is the voting method, that is, each test set sample is placed in Logistic Regression, Random Forest, and Catboost three algorithms, the results obtained by each algorithm are $n_j$, and different weights $w_j$ are given to different models. The $w_j$ used in this paper is determined by the accuracy of multiple trainings of a single model. The higher the accuracy of a single model, the greater the weight of the model. Sum it out to get the final prediction result $O_j$ of each test set sample, and the formula is as follows:

$$O_j = \sum_{i=1}^{n} w_j n_{ij}$$  \hspace{1cm} (7)

The model weights are shown in Table 2 below.

| Model name         | Weights |
|--------------------|---------|
| Logistic Regression| 0.2     |
| Random Forest      | 0.4     |
| Catboost           | 0.4     |

4. Model solving and testing

This paper calls sklearn, statsmodels and Catboost in Python to build a model for the five sets of data. The test results of the model are as follows.

4.1. Specific business groups with loans, overdue behaviors

The analysis results are shown in Table 3 and Table 4 below.

## Table 2. Weight table of the model

| Model name         | Weights |
|--------------------|---------|
| Logistic Regression| 0.2     |
| Random Forest      | 0.4     |
| Catboost           | 0.4     |

## Table 3. Model test results of Specific business groups with loans, overdue behaviors

| Dep. Variable: Y | No. Observations: 6808 |
|------------------|-------------------------|
| Model: Logit     | Df: Residuals: 6799     |
| Method: MLE      | Df Model: 8             |
| converged: True  | Pseudo R-squ.: 0.044444 |
|                  | Log-Likelihood: -4228.9 |
|                  | LL-Null: -4425.6        |

## Table 4 The significant test results of specific business groups with loans, overdue behaviors

| coef     | std err | z    | P>|z| | P =0.025 | P=0.975 |
|----------|---------|------|------|----------|---------|
| Number of credit cards settled | -0.0080 | 0.0040 | -2.1760 | 0.0300 | -0.0150 | -0.0010 |
| Number of institutions that have been loaned | -0.0160 | 0.0110 | -1.4600 | 0.1440 | -0.0380 | 0.0060 |
| Number of personal consumption loans | 0.0020 | 0.0040 | 0.5530 | 0.5800 | 0.0050 | 0.0090 |
| Specific business times | 0.0049 | 0.0050 | 1.0600 | 0.2890 | 0.0040 | 0.0140 |
| Number of credit cards held | -0.0300 | 0.0050 | -5.7760 | 0.0000 | -0.0410 | -0.0200 |
| The time from the first loan to the current time | 0.0086 | 0.0010 | 9.8300 | 0.0000 | 0.0070 | 0.0100 |
| The time from the first credit card to the current time | -0.0010 | 0.0010 | -1.1600 | 0.2460 | -0.0020 | 0.0010 |
| Other types of loans queried by | -0.0090 | 0.0030 | -3.1320 | 0.0020 | -0.0140 | -0.0030 |
The results show that the number of institutions that have been loaned, the number of credit cards held, the time from the first loan to the current time, the time from the first credit card to the current time, and the number of other types of loans that the agency has inquired, a total of five variable coefficients passed the significance test with a significance level of 0.05, that is, the above five independent variables have significant effects on the dependent variables.

4.2. No specific business group with loans and overdue behavior

The results of the analysis are shown in Table 5 and Table 6 below.

Table 5 Model test results of no specific business group with loans and overdue behavior

| Dep. Variable: Y | No. Observations: 6808 |
|------------------|------------------------|
| Model: Logit     | Df Residuals: 6799     |
| Method: MLE      | Df Model: 8            |
| converged: True  | Pseudo R-squ.: 0.04444 |
| Log-Likelihood:  | -4228.9               |
| LL-Null:         | -4425.6               |

|                     | Coef   | Std err | z     | P>|z| | P =0.025 | P=0.975 |
|---------------------|--------|---------|-------|------|----------|---------|
| Total historical loan | 0.0000 | 0.0000  | -0.6500 | 0.5160 | 0.0000  | 0.0000  |
| Total number of historical loans | -0.0310 | 0.0150  | -2.0110 | 0.0440 | -0.0610 | -0.0010 |
| Personal consumption loans | 0.0093 | 0.0160  | 0.5890  | 0.5560 | -0.0220 | 0.0400  |
| Number of personal business loans | 0.0263 | 0.0160  | 1.6500  | 0.0990 | -0.0050 | 0.0580  |
| Other loan times | 0.0145 | 0.0190  | 0.7840  | 0.4330 | -0.0220 | 0.0510  |
| Overdue loans as a percentage of total loans | 0.9712 | 0.0520  | 18.7300 | 0.0000 | 0.8700  | 1.0730  |
| Number of credit cards held | -0.0148 | 0.0030  | -4.4700 | 0.0000 | -0.0210 | -0.0080 |
| The time from the first loan to the current time | 0.0047 | 0.0010  | 9.1790  | 0.0000 | 0.0040  | 0.0060  |
| The time from the first credit card to the current time | -0.0019 | 0.0000  | -3.8270 | 0.0000 | -0.0030 | -0.0010 |
| Other types of loans queried by the organization | -0.0069 | 0.0020  | -4.2820 | 0.0000 | -0.0100 | -0.0040 |
| Total loan amount | 0.0000 | 0.0000  | -1.3290 | 0.1840 | 0.0000  | 0.0000  |
| Number of outstanding loan application agencies | -0.1252 | 0.0160  | -7.8410 | 0.0000 | -0.1570 | -0.0940 |
| Total loans for the last 6 months | 0.0000 | 0.0000  | -0.0720 | 0.9430 | 0.0000  | 0.0000  |

The results show that the total number of historical loans, the number of personal consumption loans, the number of personal business loans, the overdue loans accounted for the total loan ratio, the number of credit cards held, the time from the first loan to the current time, other types that the agency has inquired, the number of loans and the number of outstanding loan application institutions, a total of eight variables passed the significance test with a significance level of 0.05, that is, eight independent variables have significant influence on the dependent variable.
4.3. **Overdue behavior group with loans**

The results of the analysis are shown in Table 7 and Table 8 below.

| Table 7 Model test results of overdue behavior group with loans |
|---------------------------------------------------------------|
| Dep. Variable: Y | No. Observations: 3814 |
| Model: Logit | Df Residuals: 3808 |
| Method: MLE | Df Model: 5 |
| converged: True | Pseudo R-squ.: 0.04335 |
| Log-Likelihood: -2379.1 | LL-Null: -2486.9 |

| Table 8 The significant test results of overdue behavior group with loans |
|--------------------------------------------------------------------------|
| coef | std err | z | P>|z| | P =0.025 | P=0.975 |
|-----------------------------------|-----------|---|-------|----------|---------|
| Total number of historical loans | -0.0092 | 0.0030 | -2.6200 | 0.0090 | -0.0160 | -0.0020 |
| Specific business times | 0.0121 | 0.0070 | 1.6810 | 0.0930 | 0.0020 | 0.2600 |
| The time from the first loan to the current time | 0.0109 | 0.0010 | 8.9140 | 0.0000 | 0.0090 | 0.0130 |
| The time from the first credit card to the current time | 0.0006 | 0.0010 | 0.6010 | 0.5480 | -0.0010 | 0.0020 |
| Other types of loans queried by the organization | -0.0026 | 0.0030 | -0.7950 | 0.4270 | -0.0090 | 0.0040 |
| Number of credit cards held | -0.0225 | 0.0070 | -3.3130 | 0.0010 | -0.0360 | -0.0090 |

The results show that the total number of historical loans, the number of credit cards held, the time from the first loan to the current time, and the time from the first credit card to the current time, a total of four variables passed the significance test with a significance level of 0.05, that is, all four independent variables have significant effects on the dependent variable.

4.4. **No loan group**

The results of the analysis are shown in Table 9 and Table 10 below.

| Table 9 Model test results of No loan group |
|--------------------------------------------|
| Dep. Variable: Y | No. Observations: 3981 |
| Model: Logit | Df Residuals: 3975 |
| Method: MLE | Df Model: 5 |
| converged: True | Pseudo R-squ.: 0.07025 |
| Log-Likelihood: -2252.8 | LL-Null: -2423.1 |

| Table 10 The significant test results of no loan group |
|-------------------------------------------------------|
| coef | std err | z | P>|z| | P =0.025 | P=0.975 |
|-----------------------------------------------|-----------|---|-------|----------|---------|
| Total number of historical loans | -0.0187 | 0.0070 | -2.5420 | 0.0100 | -0.0330 | -0.0040 |
| The time from the first loan to the current time | 0.0108 | 0.0010 | 7.4730 | 0.0000 | 0.0080 | 0.0140 |
| The time from the first credit card to the current time | 0.0003 | 0.0010 | 0.3170 | 0.7500 | -0.0020 | 0.0020 |
| Number of mortgages inquiries | -0.2507 | 0.1370 | -1.8310 | 0.0600 | -0.5190 | 0.0180 |
| Other types of loans queried by the organization | -0.0395 | 0.0060 | -6.8310 | 0.0000 | -0.0510 | -0.0280 |
The number of times the organization was inquired for the qualification examination

|                          | Coefficient | Standard Error | t-value | P-value |
|--------------------------|-------------|----------------|---------|---------|
| The number of historical loans | -0.2197     | 0.0690         | -3.1970 | 0.0000  |
| Time from the first loan to the current time | -0.3540     | -0.0850        |         |         |

The results show that the total number of historical loans, the time from the first loan to the current time, the number of other types of loans that the institution has inquired, and the number of times the institution has been inquired for the qualification examination, a total of five variables passed the significance test with a significance level of 0.05, that is, all five independent variables have significant effects on the dependent variable.

4.5. Model test results analysis

Based on the results of the model, the following conclusions can be drawn:

1. The coefficient of the time from the first credit card to the current time is significant, and the symbol is positive. According to the processing method of the time series feature, the larger the value, the closer the time is to the present, that is, the earlier the customer uses the loan for the first time, the smaller the corresponding credit risk.

2. The coefficient of historical loan times is significant and the symbol is negative. It can be concluded that the more the number of historical loans, the lower the credit risk of customers. Based on the current lending situation in China, customers who are able to obtain multiple loans often rely on this customer to be highly trusted by some lending platforms, which can indicate that their credit risk is low.

3. The coefficient of other types of loan times that the institution has inquired is significant and the symbol is negative. Other types of loans are loans other than mortgages, car loans, and consumer loans. The reason for the impact is similar to the historical loan number coefficient. However, the correlation coefficient between the two is small, and there is no multi-collinearity, that is, the credit risk of customers who can apply for more different channels of loans should be considered lower.

4. In most models, the number of credit card variables held is significant and the sign is negative, but the proportion of customers holding one or more credit cards is almost 100%, so not having a credit card means that the customer's credit risk is low. In combination with the optimal BIV value and corresponding threshold in the third chapter, we should consider that when the customer holds more than or equal to three, the credit risk is small, and the risk decreases as the number of credit cards held increases.

5. The correlation coefficient between each model variable is small, that is, the model does not have multiple co-linearity problems.

5. Model results analysis

5.1. Model effect evaluation method

Since the problem to be dealt with is a two-category problem with unbalanced categories, the indicator used to evaluate the model effect is AUC (Area Under Curve). The AUC is the area under the curve. This curve is the Receiver Operating Characteristic Curve. The origin of the ROC curve is based on a series of different two-category methods (demarcation value or decision threshold), with the true positive rate (sensitivity) as the ordinate and the false positive rate (1-specificity) as the abscissa. AUC is a model evaluation index that can only be used for the evaluation of the two-category model. For the two-category model, there are many other evaluation indicators, such as logloss, accuracy and precision. Many machine learning models predict the classification problem as probability. If you want to calculate accuracy, you need to convert the probability into a category first. This requires manually setting a threshold. If the prediction probability for a sample is higher than this threshold, This sample is placed in a category below this threshold and placed in another category. So this threshold greatly affects the accuracy.

Because of the different losses caused by the two types of mistakes of “Take true as false” and “Take false as true” under different realistic conditions, for example, the potential bad debt losses
brought by overdue customers are far from being compensated by the profits of a normal customer. It is reasonable to use the AUC value to evaluate the two-category model without knowing the difference between the two errors. The specific classification threshold can be set by the model user according to different goals.

The value of AUC ranges from 0.5 to 1. It is generally considered that the model is more reasonable when the AUC of the binary model is greater than 0.7.

5.2. Evaluation of model results

10% of the samples containing the tags were randomly selected as the test set, and the remaining samples were used as the training set training results for model training and prediction results comparison.

5.2.1. Specific business groups with loans, overdue behaviors. The number of samples included in the model is 6225 (26%), and the ROC curve and AUC values are shown in Figure 3.

5.2.2. No specific business group with loans and overdue behavior. The number of samples included in the model is 11189 (45%), and the ROC curve and AUC values are shown in Figure 4.

5.2.3. Overdue behavior group with loans. The number of samples included in the model is 3480 (14%), and the ROC curve and AUC values are shown in Figure 5.

5.2.4. No loan group. The number of samples included in the model is 3766 (15%), and the ROC curve and AUC values are shown in Figure 6.

![Figure 3. ROC curve and AUC value of Specific business groups with loans, overdue behaviors](image1)

![Figure 4. ROC curve and AUC value of No specific business group with loans and overdue behavior](image2)
The above four models are multiplied by their corresponding sample proportions according to their corresponding AUC values. The final model has an AUC of 0.73, that is, the overall model results are reasonable and have strong discriminating ability.

5.3. Model evaluation
From the above results, it can be considered that the modeling prediction effect is better and the features after feature selection are more interpretable. The following are some conclusions based on data characteristics, feature representation and model results.

(1) Historical credit information has a strong reference value for judging the credit risk of customers. That is, the individual credit history and personal basic characteristic data can be put into the model to obtain a more accurate credit risk evaluation score.

(2) Since the overall model of this paper works well, it is reasonable to group the samples. When judging whether a customer has a high credit risk, they should be placed in similar groups for comparison, instead of compare it directly to all customer information.

(3) The BIV value modified by this paper can calculate the optimal partition threshold and partitioning effect corresponding to each feature under the condition of unbalanced categories, which can provide reference for the dimension reduction of screening features.

(4) The lower the probability that a customer with a variety of credit history (for example, the first loan time, the first credit card time, etc.) is rated as a risk customer. The longer the credit history usually means the older a person, the longer the working hours, the richer the capital accumulation, and the lower the credit risk. According to the BIV value performance, users with a credit history of more than one and a half years have a lower credit risk.

(5) The more number of loans processed and the number of credit cards held has a lower credit risk. Loans and credit cards need to be approved. If a customer can make a loan or credit card in a financial institution, they must at least pass the credit screening of these financial institutions. And a customer who is recognized by many organizations for their credit often pays more attention to their own credit situation. According to the performance of BIV, customers who have made three or more loans or have more than two credit cards have lower credit risk.

6. Research conclusions
In this paper, the grouping modeling idea is used to pre-group customers and reduce the feature sparsity problem. A BIV value feature screening method and a weighted average model based on Logistic Regression, Random Forest and Catboost are proposed to solve the user credit score. In order to solve the difficulty of user credit scoring, a set of solutions is provided. The results show that the classification effect of the model is good and has the following practical significance.
(1) Established a personal credit rating standard for the company
This paper uses big data and machine learning methods to establish a personal credit scoring method for enterprises, which has certain reference value for preventing credit risk.

(2) Speed up customer review
This paper proposes a personal credit scoring modeling technique based on data provided by Internet finance companies, and calculates a credit risk score between 0 and 1 for each sample. In the case of ensuring that customers can provide real and effective information, each new sample data entry system can automatically score it. If the score is higher than a certain threshold, it will be automatically approved. If the customer is below a certain threshold, the customer will be automatically rejected. Customers with a certain range of values will have a little manual intervention. In most cases, the user only needs to confirm that the model is running correctly, and does not need to personally observe a large amount of customer data, which is convenient for implementing the retail credit factory model, greatly improving work efficiency and reducing a large amount of labor costs.

(3) Meeting the changing needs of decision makers by adjusting thresholds
The model proposed in this paper has an AUC of 0.73 on the actual data, and the overall effect of the model is better, and the user can modify the threshold of the model decision by actually major demand (maintaining credit risk or increasing traffic). When credit risk is the main goal, the threshold can be appropriately reduced to reduce the potential credit risk customers. When the main goal is to increase the business volume, the threshold can be appropriately increased to allow more customers to obtain loans, and the corresponding higher tolerance credit risk.

(4) Provide a single variable analysis method
The BIV value screening feature method proposed in this paper calculates the maximum utility and corresponding threshold of each feature, so that the enterprise can briefly analyze the potential credit risk value corresponding to the information when the new customer data is less.

Acknowledgments
This paper is supported by the First Class Discipline of Zhejiang-A (Zhejiang Gongshang University-Statistics), Zhejiang province advantage subject (Zhejiang Gongshang University-Statistics). The authors would like to thank Professor Shangfeng Zhang for excellent theoretical support and critically reviewing the manuscript.

References
[1] Benyacoub, B., Bernoussi, S. E., & Zoglat, A. (2014). Building classification models for customer credit scoring. International Conference on Logistics & Operations Management. IEEE.
[2] Bravo, Cristián, Thomas, L. C., Weber, R. (2015). Improving credit scoring by differentiating default behaviour. Journal of the Operational Research Society, 66(5), 771-781.
[3] Chen Q.W., Wang W., Ma D., Mao W. (2018). Class-imbalance credit scoring using Ext-GBDT ensemble. Application Research of Computers, 35(02):421-427.
[4] Dorogush, A. V., Ershov, V., Gulin, A. (2018). Catboost: gradient boosting with categorical features support.
[5] Gao S., Wang C.B. (2017). Personal Credit Scoring Based on Decision Tree C5.0 Algorithm. 2017 7th International Conference on Education, Management, Computer and Society (EMCS 2017).
[6] Kan S.X. (2010). Comparison and Verification of the Effect of Commercial Bank Credit Rating Screening Financial Indicators. (Doctoral dissertation, Shandong University).
[7] Li J. (2015). A Green-credit Risk Assessment Research based on Random Forest Algorithm. Financial Theory & Practice (11), 14-18.
[8] Li Y.H., Li X.S., Jiang Y.X. (2007). Study on Information Entropy as a Solution to Measure Risk in Financial Market. Operations Research and Management Science, 16(5):111-116.
[9] Liu X.H., Qu D.Y. (2016). Substitute Credit Scoring Based on Big Data Credit Reference.
Credit Reference, 34(3), 33-36.

[10] Shen C.H. (2005). Research on Personal Credit Evaluation Method in Consumer Credit Based on Support Vector Machine. (Doctoral dissertation, China Agricultural University).

[11] Shi Q.Y., Jin Y.H. (2003). Consumer credit scoring model: a survey. Statistical Research, 20(8):36-39.

[12] Shi Q.Y., Jin Y.H. (2004). The Comparative Analysis of the Application of Several Scoring Models of Consumer Credit in China. Statistical Research, 21(6), 43-5.

[13] Shi X.K., Chang Z.Y. (2015). Application of Two Kinds of Biased Logistic Distributions in Credit Score Model. Statistics & Decision (14):19-23.

[14] Song L.P., Zhang L.K., Xu W. (2015). P2P network lending personal credit risk assessment. Finance and Accounting Monthly (35), 94-96.

[15] West, D. (2000). Neural network credit scoring models. Computers & Operations Research, 27(11-12), 1131-1152.

[16] Wiginton, John, C. (1980). A note on the comparison of logit and discriminant models of consumer credit behavior. The Journal of Financial and Quantitative Analysis, 15(3), 757.

[17] Xiong W. (2009). Research on Prevention and Evaluation of Personal Credit Risk. Money China (7), 4-6.

[18] Zhang W.Y. Exploring the Improved Personal Credit Scoring Model of Ant Financial Services in its Disruptive Innovation Process[P]. 3d International Conference on Applied Social Science Research (ICASSR 2015), 2016.

[19] Liberati, C., Camillo, F., Saporta, G. (2017). Advances in credit scoring: combining performance and interpretation in kernel discriminant analysis. Advances in Data Analysis and Classification, 11(1), 121-138.

[20] Minghui, J. (2008). Construction and application of pso-svm model for personal credit scoring. Chinese Journal of Management, 5(4), 158-161.