A Comparison Study on the Era of Internet Finance China Construction of Credit Scoring System Model

Hongjun Zeng
Business School
Guangxi University
Nanning, China
E-mail: hongjunzeng@protonmail.com

Abstract
At present, China's Internet finance has flourished, showing a variety of business models and operating mechanisms. Through Internet technology, financial institutions can speed up business processing and bring users a better service experience. However, there are also problems such as credit risk and user fraud, and it is urgent to improve the level of risk control through credit scoring models. Because of this, this article uses the borrower data of a Chinese financial institution from January 2017 to June 2017 as the original data, and then uses the Spearman rank correlation test to screen out the variables with reliable explanatory power from the many variables of the sample data, and then Based on the variables selected, R 3.4.3 and SPSS 23.0 were used to construct a random forest model, discriminant analysis model, and logistic regression model. In general, different models perform differently under different sample characteristics, but the discriminant analysis has been better applicable. This paper compares the judgment accuracy of these three types of models and tries to establish a more effective financial credit scoring method, to solve the problem of constructing China's credit scoring system model under the current Internet financial background.

Keywords: Credit Scoring System, Random Forest, Discriminate Analysis, Logistic Regression, Comparison Study.

I. Introduction
1.1 Research Background
The objective, comprehensive, and accurate individual credit rating model is an essential component of the personal credit rating system (Hand & Henley, 1997). The existing personal credit scoring system through Internet technology, speed up business processing, bring users a better service experience (Yu et al., 2009). However, there are some problems, such as credit risk and customer fraud. Therefore, it is urgent to improve the level of risk control through the credit score model. The credit investigation institution shall use the rich information collected to make comprehensive credit evaluation on individuals (Dhillon & Torkzadeh, 2006). Based on abundant personal credit history and credit behavior data, the credit behavior pattern obtained by adopting the data mining method can more accurately predict the future credit performance of individuals, improve the efficiency of operation, reduce the cost of credit granting, and accurately estimate the risk of consumer credit, which is an essential tool for the internal scoring of financial institutions (Hsieh & Hung, 2010). Therefore, the establishment of an accurate credit scoring system is of considerable significance to enterprises. The model of individual credit rating is to use statistical analysis method and data mining technology to analyze the primary personal information data and transform the current personal information data into a specific credit risk value with high recognition (West, 2000; Huang et al., 2007).

In the past, China mainly relied on the experience of credit officers to judge the credit status of customers. There has been severe information asymmetry between credit institutions and customers (Stiglitz, 1993) which makes credit institutions unable to accurately measure the credit status and risk of lenders, which may lead to credit errors and directly threaten the interests of credit institutions and the healthy development of credit market (Hoff & Stiglitz, 1990). Although other countries have a very mature experience in credit scoring and have used the combination of traditional statistics and machine learning to evaluate customer credit quantitatively (Thomas, 2000) but because there is no unified data source and credit evaluation system in China at present, foreign experience is not applicable, so it is necessary to form a set of personal credit reporting system in line with Chinese characteristics and find a suitable credit scoring method (Allen et al., 2007).

Based on the above conditions, this paper bases on the underlying theory and practice apply the appropriate methods of data mining and statistics and uses the historical business data of a loan institution as the original data Based on relevant Study Experience. In order to construct the evaluation system of Chinese personal consumption credit, we will provide some reference to the financial institutions and government.
1.2 Literature Review
Durand (1941) applied Discriminant Analysis to credit scores of commercial banks. Discriminant analysis is based on the original classification, when a new analytical sample is encountered, i.e., pass. This classification method is used to select specific evaluation criteria as the basis for judging the group in which the new sample is located (Eisenbeis, 1977; Wind, 1978; Day et al., 1978). On this basis, new discriminant samples can be classified into known taxonomic groups. Commonly used discriminants Distance discrimination, Bayesian discrimination and Fisher discrimination are the methods of analysis (Lachenbruch & Goldstein, 1979; Ripley, 1994). Discriminant analysis was also used to develop the credit model (Desai et al., 1996; Dorronsoro et al., 1997). FICO scores constructed with discriminant analysis as the core are widely used in the field of credit scoring by Chen & Chen (2010) used the latest semi-supervised nonparametric discriminant analysis (SND), sparse tensor discriminant analysis (STDA), semi-supervised discriminant analysis (SDA), sparse discriminant analysis (Sparse DA), Fisher discriminant analysis (FDA), and multivariate discriminant analysis (MDA) to construct credit score models, respectively, and the results showed that SND, STDA, and SDA performed better than other discriminant analyses.

Wiginton (1980) used discriminant analysis and logistic regression to construct a credit score model from 1967 to 1968. The results indicated that logistic regression was superior to discriminant analysis. Shi & He (2015) introduced the idea of asymmetric function in credit rating, took the distribution function of biased logistic distribution as the inverse function of connection, and conducted a comparative empirical analysis using personal credit data of a financial institution. The results indicate that the effect of the biased logistic regression model was better than that of the ordinary logistic regression model, and the effect of the biased logistic regression model was better than that of the decision tree, neural network and support vector machine in 10% default data set. Sohn et al., (2016) applies a fuzzy logistic regression model that was established by using the data of 4446 loan applicants and loan default results and compared with traditional logistic regression. It was found that fuzzy logistic regression could improve prediction performance. Compared with discriminant analysis, logistic regression is easy to calculate and requires more relaxed data distribution. So far, logistic regression is the most commonly used credit score model.

Since individual credit scoring models have their advantages, scholars have begun to study combination models, which are divided into heterogeneous integration models and homogeneous integration models. According to the definition of random forest, Random forest is a homogeneous integration of decision trees. Su (2018) proposed a personal credit scoring model based on the accompanying forest combination. Using the data of a commercial bank in Germany for empirical analysis, compared with KNN, radial basis based neural network, decision tree, gradient boosting decision tree and support vector machine, the random forest model not only has high accuracy but also has the characteristics of being able to handle noisy data and good generalization ability. According to the German credit data, Li (2017) respectively established the Logistic credit score model and random forest credit score model, and the results showed that the accuracy of the random forest was superior to that of logistic regression. As long as the coefficients of the combined model are set well, the combined model may be superior to the single model inaccuracy or other aspects. The two-stage scoring model proposed by Shi (2005) a logistic regression model based on the neural network, is validated with credit card customer data of a commercial bank. It is found that the accuracy of the new model is higher than logistic regression, and the robustness is also greater than neural network model, indicating that the new model combines the advantages of a single model and avoids the disadvantages of a single model. Yang (2018) used the results of a linear discriminant analysis model as one of the input variables of the BP neural network. The results of empirical analysis show that the combined model has better prediction accuracy than the single model, and overcomes the problem of single model robustness. A heterogeneous integration model based on bagging algorithm and stacking algorithm is proposed by Xia et al., (2018). Empirical analysis shows that the performance of this heterogeneous integration model is better than that of the logistic regression model, support vector machine, decision tree and random forest model.

1.3 Practical Application of Personal Credit Score Model
At present, the FICO score is the most commonly used in the US credit information market. Fair Isaac Company issues the FICO score. There are three forms of FICO score, which are respectively applied to the three significant US credit administrations (Berger & Udell, 2002) See Table 1.

| Reimbursement History | 35% | Repayment records of various credit accounts |
|-----------------------|-----|------------------------------------------|
|                       |     | Public record                            |
|                       |     | Overdue reimbursement                     |
|                       |     | Number of credit accounts to be reimbursed|
|                       |     | Credit Account Balance                    |

Table 1. American FICO Personal Credit Score Model
The credit scores derived from the model for the FICO score ranged between 300 and 850 points. The higher the score, the smaller the credit risk of the customer. Nevertheless, the score itself does not tell whether a customer is good or bad, and lenders often use the score as a reference for their loan decisions (Allen et al., 2004). Each lender will have its lending strategy and standards, and each product will have its risk level, which determines the acceptable credit score level.

Generally speaking, if the borrower’s credit score reaches 680 points or above, the lender can consider the borrower’s credit outstanding and can agree to the payment without hesitation. If the borrower’s credit score is below 620, the lender either asks the borrower to add collateral or looks for various reasons to reject the loan. If the borrower’s credit score is between 620 and 680 points, the lender will conduct further investigation and verification and use other credit analysis tools to handle the case.

The sesame credit is a subsidiary of China Alibaba Group Ant Finance. It belongs to an Independent third-party credit reporting institution; see Table 2, and gold garments objectively present their credit status through techniques such as cloud computing and machine learning. The sesame credit is different from the traditional credit reporting agency (Yip & McKern, 2016). Alibaba Cloud has a vast database as a backdrop, with the unique advantages of Internet technology and data. On this basis, sesame credit evaluates the credit rating of users through the credit model algorithm (Lin et al., 2015).

Nevertheless, it also suffers from the applicability of the credibility model. The problem of credit information sharing not only affects the comprehensiveness of data dimension but also affects the accuracy of the model measurement (Ennew & Binks, 1999; Wu, 2008). Therefore, the actual credit status of the client information subject cannot get a very accurate response in the sesame credit score. The applicability of the credit model also requires time for slow collection and validation; the primary data source for sesame credit depends on industry data, and the dimensions of data collection are not complete (Nwana, 1996). While sesame credit already collects a tremendous amount of information, Alibaba’s social system is slightly lacking, so it has less control over data on social behavior; it also lacks credit data on financial institutions. At present, Sesame Credit has not been able to intervene in the Central Bank’s credit system, and major banks and financial institutions have not been able to obtain their credit data, which also leads to the lack of personal use of bank credit information data in calculating Sesame Credit scores (Kostka, 2019; Creemers, 2018). Sesame Credit has no personal credit data from banks, and it is difficult for Sesame Credit to master the more accurate personal income of users, as well as essential assessment data such as debt information and related assets.

Table 2. Ali sesame credit score

| Identity Characteristics | 15% |
|--------------------------|-----|
| Credit history           | 35% |
| Compliance               | 20% |
| Personal relationships    | 5%  |
| Behavioral Preference    | 25% |

2. Research Preparation

2.1 Basic Data

This paper uses issued after archived by a financial institution in China of the historical business data in the first half of 2017 is the original data, which includes the report number, ID number, loan date, agent, local nationality, working province, education level, marital status, salary, and fund. There are 30000 raw data.
2.2 Selection of Research Methods

In this paper, data mining and statistical correlation methods are used, i.e., The software R 3.4.3 extension package and SPSS 23.0 were used to construct random forest models and to apply discriminant analysis methods, then establish Logistic Model, and the effect of each model after the actual operation of the comparative analysis.

2.3 Statistical Approach

This chapter selects the status of overdue repayment as the explanatory variable and selects the agent, local nationality, working province, education level, marital status, salary, presence or absence of funds and gender (as known from the information in the ID card data archive), as well as the provincial gross product, per capita disposable income, per capita consumption expenditure, regional fixed asset investment, regional fixed-asset investment index and unemployment rate that can be found by the working province (Anonymous, 2017).

The defined and explained variables are shown below:

| Table 3. Explained and Explanatory variables |
|--------------------------------------------|

| Variable Type | Variable | Variable Name | Grade | Comments |
|---------------|----------|---------------|-------|----------|
| Explained Variable | Y | Presence or absence of late repayment | 0: No overdue; 1: Overdue |
| XI | Loan grade | 1-12 | Every half month is the first grade, the earlier the loan grade is higher |
| X2 | Agent | 1-11 | The higher the frequency of use, the higher the grade |
| X3 | Whether there is local nationality | 0: Not local; 1: Local |
| X4 | Working provinces | Derive the macroeconomic variables of X10-X15, the model does not use this variable |
| X5 | Educational level | 1-3 | 1: Specialty or below; 2: Undergraduate; 3: Master's degree or above |
| X6 | Marital status | 1-4 | 1: Not married; 2: Married; 3: Divorced; 4: Widowed |
| X7 | Compensation | 1-7 | Higher pay, higher rank |
| X8 | Whether there is fund | 0: No fund; 1: Fund |
| X9 | Gender | 0: Female; 1: Male |
| X10 | Gross product of the province | 1-29 | The greater the GDP in the province, the higher the rank |
| X11 | Per capita disposable income | 1-29 | The higher the per capita disposable income, the higher the rank |
| X12 | Per capita consumption expenditure | 1-29 | The higher the per capita consumption expenditure, the higher the rank |
| X13 | Regional Fixed Assets Investment | 1-28 | The higher the regional fixed asset investment, the higher the grade |
| X14 | Regional fixed asset investment index | 1-28 | The higher the regional fixed asset investment index, the higher the grade |
| X15 | Unemployment | 1-16 | The higher the unemployment rate, the higher the grade |
2.4 Data Preprocessing
In summary, the borrowers with or without deferred repayment of a financial institution in China from January 2017 to June 2017 shall be taken as the total sample of data processing.

- First, the data were processed, and we found that the amount of data for the vacancy values of the samples that did not contain agents was huge, and the user, no salary levels of the agents, were not included, for which the data were divided into two sample sets by whether or not the agents were included and analysed separately. Then, analysis of the data found that the sample data are microscopic, the lack of macroscopic data support, the conclusions may not be accurate and complete. Therefore, the working province containing the agent sample and the province of origin without the agent sample (known from the first two digits of the ID card) were converted into six indicator representatives related to economic development, namely, the province's gross product, per capita disposable income, per capita consumption expenditure, regional fixed asset investment, regional fixed-asset investment index, and unemployment rate (all data resources given by China Statistical Yearbook 2017 obtained).

- Selection of samples.
  Select whether to include the full sample remaining from the agent.
- Added blank and missing values
  Since the data has been split into two data sets for analysis, the samples with vacancy values in the two data sets were filtered out, respectively, and then approximately 95% of the sample size remained in each data set, and the data integrity of these samples was functional.
  Therefore, in combination with the above analysis, a small number of samples with vacant values are directly sieved out to obtain the final sample with or without agents.

3. Analysis and Finding
3.1 Descriptive Statistical Analysis
The collected data samples were first subjected to descriptive statistical analysis using SPSS 23.0.

Table 4. Descriptive statistics including samples of agents

| Variable | N    | Minimum value | Maximum value | Mean   | St. Dev. | Skewness | Kurtosis |
|----------|------|---------------|---------------|--------|----------|----------|----------|
| Y        | 5775 | 0             | 1             | 0.121  | 0.326    | 2.2327   | 3.415    |
| X1       | 5775 | 1             | 12            | 4.486  | 2.381    | 0.544    | 0.238    |
| X2       | 5775 | 1             | 11            | 10.573 | 1.017    | -3.742   | 18.456   |
| X3       | 5775 | 0             | 1             | 0.689  | 0.463    | -0.815   | -1.336   |
| X5       | 5775 | 1             | 3             | 1.264  | 0.454    | 1.253    | 0.062    |
| X6       | 5775 | 1             | 4             | 1.742  | 0.533    | 0.022    | 0.513    |
| X7       | 5775 | 1             | 7             | 3.588  | 1.414    | 0.876    | 0.216    |
| X8       | 5775 | 0             | 1             | 0.363  | 0.481    | 0.571    | -1.675   |
| X9       | 5775 | 0             | 1             | 0.684  | 0.465    | -0.793   | -1.371   |
| X10      | 5775 | 1             | 29            | 19.953 | 6.878    | -0.709   | -0.558   |
| X11      | 5775 | 1             | 29            | 18.329 | 7.348    | -0.362   | -1.252   |
| X12      | 5775 | 1             | 28            | 16.963 | 7.381    | -0.386   | -1.276   |
| X13      | 5775 | 1             | 28            | 18.920 | 7.619    | -0.683   | -0.883   |
| X14      | 5775 | 1             | 16            | 8.344  | 3.953    | 0.257    | -1.225   |
| X15      | 5775 | 1             | 16            | 9.283  | 4.310    | -0.053   | -1.168   |

Table 5. Descriptive Statistics for Non-Agent Sample Variables

| Y | N    | Minimum value | Maximum value | Mean   | Standard deviation | Skewness | Kurtosis |
|---|------|---------------|---------------|--------|--------------------|----------|----------|
| X1| 20134| 0             | 1             | 0.044  | 0.206              | 4.421    | 17.547   |
| X3| 20134| 0             | 1             | 0.569  | 0.495              | -0.279   | -1.922   |
From descriptive statistics, it can be seen that the degree of steepness or smoothness varies significantly among different variables, as does the degree of skew.

### 3.2 Basic Analysis of Variables

First, it can be seen that in the sample containing agents, the number of deferred repayments is 698, accounting for about 12%. The number of performance articles was 5077, or about 88 percent. As shown in the figure below:

![Figure 1. Percentage of samples containing agents with or without deferred repayment](image1.png)

Based on the above analysis, it can be initially seen that the relative Contains The user of the agent has a high probability of deferred repayment Users without agents. Furthermore, overall, nearly 90% of people have not extended their repayment terms. An analysis that did not include a sample of agents was then performed. It can be seen that in the samples without agents, the number of deferred repayments is 895, accounting for about 4.5%; the number of performances is 19,239, accounting for about 95.5%. As shown in the figure below:

![Figure 2. Percentage of samples not containing agent with or without deferred repayment](image2.png)
4. Screening of Explanatory Variables for Samples with Agents

4.1 Correlation of Explanatory Variables with Explanatory Variables

In this chapter, 14 variables are selected to research the influencing factors of the borrower's deferred repayment, which are loan grade \( (x_1) \), agent \( (x_2) \), local nationality \( (x_3) \), education level \( (x_5) \), marital status \( (x_6) \), salary \( (x_7) \), fund availability \( (x_8) \), gender \( (x_9) \), provincial gross product \( (x_{10}) \), per capita disposable income \( (x_{11}) \), per capita consumption expenditure \( (x_{12}) \), regional fixed investment \( (x_{13}) \), regional fixed investment index \( (x_{14}) \), unemployment rate \( (x_{15}) \). The sample containing the agent does not select the working province because the working province itself has no substantial meaning. For this reason, we added six macro data variables corresponding to the provinces.

SPSS 23.0 was first used in this paper, followed by Pearson correlation test preliminarily explored the relationship between the explanatory variables and the explained variables. Explanatory variables were screened by the size and significance requirement of the correlation coefficient between the dependent and independent variables.

Table 6. Person Correlation Test Results

| Y | X1  | X2  | X3  | X4  | X5  | X6  | X7  | X8  | X9  | X10 | X11 | X12 | X13 | X14 | X15 |
|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Y | -0.179 | -0.293 | 0.020 | 0.015 | -0.031 | -0.045 | 0.001 | X9 | -0.001 | X10 | -0.142 | X11 | -0.036 | X12 | -0.052 |
|   | X1  | X2  | X3  | X5  | X6  | X7  | X8  | X9  | X10 | X11 | X12 | X13 | X14 | X15 |

According to the correlation test, except \( X_8 \) and \( X_9 \) failed the significance test, and all other variables passed the significance test. Inquiry Considering the remaining variables as 12 There is only one, so it is not screened according to the correlation of variables, finally selected \( X_1, X_2, X_3, X_5, X_6, X_7, X_9, X_{10}, X_{11}, X_{12}, X_{13}, X_{14}, X_{15} \) as explanatory variables.

4.2 Multicollinearity Analysis

In order to ensure the accuracy of the model results, it is necessary to test whether there is multicollinearity between the variables, and the results are shown in the following table, The inflation factor of 12 variables can be seen VIF Between 0-10 Between, can judge There was no severe multicollinearity among the 12 variables.

Table 7. Test for multicollinearity

| Model | Unnormalized coefficient | Normalization factor | T | Significance | Collinearity statistics |
|-------|--------------------------|----------------------|---|-------------|------------------------|
|       | B                        | Standard error       | Beta |             | Franchise | VIF   |
| I     | (constant) (quantity)    | 1.524                | 0.051 | 29.927 | 0         |              |
| X1    | -0.021                   | 0.002                | -0.156 | -12.824 | 0         | 0.954 | 1.048 |
| X2    | -0.092                   | 0.004                | -0.288 | -23.993 | 0         | 0.979 | 1.021 |
| X3    | 0.007                    | 0.009                | 0.01   | 0.83    | 0.407 | 0.924 | 1.082 |
| X5    | -0.013                   | 0.009                | -0.018 | -1.488 | 0.137 | 0.952 | 1.05  |
| X6    | -0.012                   | 0.007                | -0.02  | -1.644 | 0.1    | 0.941 | 1.063 |
| X7    | -0.008                   | 0.003                | -0.034 | -2.755 | 0.006 | 0.93  | 1.075 |
| X10   | -0.007                   | 0.001                | -0.157 | -5.002 | 0         | 0.143 | 7.004 |
| X11   | 0.017                    | 0.002                | 0.389   | 7.077  | 0     | 0.127 | 9.331 |
| X12   | -0.016                   | 0.003                | -0.364 | -6.123 | 0     | 0.114 | 9.033 |
| X13   | -0.002                   | 0.001                | -0.046 | -1.66  | 0.097 | 0.187 | 5.341 |
| X14   | -0.019                   | 0.001                | -0.227 | -17.504 | 0     | 0.843 | 1.187 |
| X15   | 0.003                    | 0.001                | 0.043   | 3.059  | 0.002 | 0.705 | 1.418 |
5. Modeling the Credit Method Evaluation Model with an Agent Sample

5.1 Modeling of Various Types of Models

Random Forest Models: Random forest refers to a classifier that uses multiple trees to train and predict samples, which overcomes overfitting. Produces, has the excellent anti-jamming ability, can estimate the sample characteristic in the classification importance degree accurately and the algorithm is easy to understand (Mao et al., 2018).

Randomly selected in all samples 80% of data as training data, where 564 records for deferred repayment, 4059 records for on-time repayment, with each variable as the characteristics of training, by making with R 3.4.3 Randomize the original software package to implement the modeling process.

Selection of variables: Selecting appropriate variables not only improves accuracy but also reduces the complexity of the model calculation process, thereby improving the model Run Efficiency. First, the variables are initially selected based on their correlation, from the perspective of significance, excluding x8 (with or without funds) and x9 (sex), and then introduce x10 (intra-provincial GDP), x11 (per capita disposable income), x12 (per capita consumption expenditure), x13 (regional fixed-asset investment), x14 (regional fixed-asset investment index), x15 (provincial unemployment rate) to replace x4 (working province). When there are 12 screening variables, the on-time repayment (0) is wrongly judged as delayed repayment (1), and the error rate is 2.7%. In contrast, the delayed repayment (1) is wrongly judged as on-time repayment (0). The error rate is 40.6%, and the overall error rate is 7.3%.

Table 8. Preliminary training results of random forest model

| Original Result | Training results | Prediction accuracy |
|-----------------|------------------|---------------------|
| 0               | 4059             | 3448                | II                  | 97.3%                          |
| 1               | 564              | 229                 | 335                 | 59.4%                          |

Overall accuracy: 92.7%

Considering that there are too many variables, and the variables with less correlation may affect the training effect of the model, resulting in a decrease in the accuracy of prediction. So, we should eliminate some irrelevant variables step by step to make the model achieve the best prediction effect. According to the importance of variable features in the random forest model from small to large in order of deletion. For example, the figure below shows the importance degree of each variable at the first elimination, as shown in Figure 3, the one with the lowest elimination importance (level of education).

Figure 3. Importance of each variable

Repeat the above steps according to the change of status of each variable during each elimination. The variables removed in turn are x6 (marital status), x7 (salary), x14 (unemployment rate), x3 (whether local), x13 (regional fixed asset investment), x11 (per capita disposable income), x12 (per capita consumption expenditure), x14 (regional fixed-asset investment index), x2 (agent), x1 (loan grade), the corresponding accuracy rate is shown in the table 10 below:
Table 9. Distribution of each variable

| Number of variables | 12 | 11 | 10 | 9  | 8  | 7  | 6  | 5  | 4  | 3  | 2  | 1  |
|---------------------|----|----|----|----|----|----|----|----|----|----|----|----|
| Wrong repayment on time | 2.73% | 1.87% | 2.59% | 1.36% | 1.92% | 2.14% | 2.19% | 2.32% | 2.39% | 2.59% | 3.10% | 0.00% |
| Error rate Delay repayment error | 40.6% | 45.9% | 38.1% | 44.8% | 40.0% | 34.7% | 33.6% | 32.0% | 32.4% | 33.5% | 62.7% | 100.0% |
| Error rate Total Error Rate | 7.35% | 7.25% | 6.92% | 6.66% | 6.58% | 6.12% | 6.04% | 5.95% | 6.06% | 6.36% | 10.3% | 12.20% |

From the figure 4, it can be seen that when the variable is 4, 5, 6, when the error rate is low. When the variables are selected as 4, 5, 6 by predicting the training samples and comparing the correct rate, we can see that the correct rate is equal and the highest when four or five variables are selected, indicating that the probability of making the above two types of errors has decreased at this time, which indicates a significant improvement in the accuracy rate. Considering the original accuracy and training accuracy, the final selected variable in this chapter is 5. So, in this case, the solution chosen for this model is expected to be optimal.

Selection the trees for test: The choice of the number of trees directly affects the accuracy of the random forest training results. If there are too few tree choices, the predicted results will be unsatisfactory; if there are too many tree choices, the results will be more accurate, and it has no significant effect and will directly affect the Speed of model operation. In this paper, 200 trees are selected to explore the influence of the number of trees on the accuracy of judgment. The results are as follows:
As shown in the figure 5 above, the graph abscissa represents the number of trees, and the ordinate represents the judgment error rate of the model, where green represents the error rate of the model in judging deferred repayment, red represents the error rate of the model in judging on-time repayment, and black represents the total error rate. It is evident from the figure that when the tree of the tree is at 50 trees, the error rates of on-time repayment and deferred repayment have reached the lowest point. Based on this, the judgment standard of the model can be inferred. The rate of confirmation is approximately 94%.

Random forest model predicts the final result: According to the selection of variables and the setting of model parameters, the final variables selected in this paper are per capita consumption expenditure, x14 (regional fixed-asset investment index), x2 (agent), x1 (loan grade), x10 (Province GDP). Parameter tree the choice is 200 trees.

Table 10. Random Forest Model Final Training Results

| Original Result | Training Results | Prediction accuracy |
|-----------------|------------------|---------------------|
| 0               | 4059             | 3965                | 94                  | 97.7%               |
| 1               | 564              | 181                 | 383                 | 67.9%               |

Overall accuracy: 94.1%

From the result, the overall accuracy of the model is 94.1%, of which the judgment of the people who repay on time is more accurate, and its accuracy up to 97.7%; while the judgment of the people who delayed repayment was slightly unsatisfactory, with an accuracy of approx. 67.9%. The reason for this may be related to the selection of sample size.

5.2 Discriminate Analysis

According to the previous Person According to the results of correlation coefficient analysis, ten variables that have a large to small correlation with the explained variable (whether deferred repayment or not) are selected as the observed variables. These are X1, X2, X6, X7, X10, X11, X12, X13, X14, X15, respectively, and there will be full samples of agents as training data. The final results obtained by the calculation method of discriminant analysis are as follows:

Table 11. Discriminant analysis results

| Original Result | Training Results | Prediction accuracy |
|-----------------|------------------|---------------------|
| 0               | 5077             | 4106                | 971                 | 80.9%               |
| 1               | 698              | 259                 | 439                 | 78.1%               |

Overall accuracy: 78.7%

5.3 Logistic Regression

Introduction to the entropy weight method: Entropy weight is a method based on actual weights, the amount of information contained in each index, and A. The smaller the entropy, the higher the variability of the exponent. The greater its role, the higher the weight of comprehensive evaluation. Computational program entropy weight method is simple and straightforward; the index data is effectively used, excluding the influence of subjective factors (Bikker & Haaf, 2002).

Data Normalization: Standardize the data of each indicator. It is assumed that m indicators are given, x1, x2, x3, ..., where xi = {x1, x2, ..., xm}, assuming that the values normalized to the respective indicator data are y1, y2, ..., ym, then

\[ Y_i = \frac{X_i - \min (X_i)}{\max (X_i) - \min (X_i)} \]

Seeking the entropy of information of Each Index: According to the definition of entropy of information, the entropy of information of a group of data

\[ E_t = \frac{1}{\ln (n)} \sum_{i=1}^{n} P_{ij} \ln P_{ij} \]

and. Among them If \( P_{ij} = 0 \) Then define \( \lim P_{ij} \neq 0 \)

Determine the weight of each indicator: According to the calculation formula of information entropy, the information entropy of each index is calculated as E1, E2,... Em. The weights of each indicator were calculated by information entropy:
\[ W_i = \frac{1 - E_i}{k + \sum E_i} \quad (i = 1, 2, \ldots, k) \]

Table 12. Results of entropy weight method

| Variable | Variable name                  | Weight  |
|----------|--------------------------------|---------|
| X1       | LOAN DATE                      | 0.079153|
| X2       | AGENT                          | 0.001991|
| X3       | IS_LOCAL                       | 0.108171|
| X5       | EDU_LEVEL                      | 0.394842|
| X6       | MARRY_STATUS                   | 0.110361|
| X7       | SALARY                         | 0.042161|
| X10      | Gross Domestic Product         | 0.022438|
| X11      | disposable income per capita   | 0.02952 |
| X12      | Per capita consumption and expenditure | 0.035831 |
| X13      | Regional Fixed Assets and Investment | 0.031807 |
| X14      | Regional Fixed Asset Investment Index | 0.044339 |
| X15      | Unemployment                   | 0.045766|

Next, according to the definition of the correlation matrix between the credit index and each factor, if the correlation coefficient is positive, then the factor entropy weight is also positive, if the correlation coefficient is negative, then the factor entropy weight is also negative.

Table 13. Weighted value of user credit index

| Variable | Variable name                  | Weight  |
|----------|--------------------------------|---------|
| X1       | LOAN DATE                      | -0.079153|
| X2       | AGENT                          | -0.001991|
| X3       | IS_LOCAL                       | 0.108171|
| X5       | EDU_LEVEL                      | 0.394842|
| X6       | MARRY_STATUS                   | -0.110361|
| X7       | SALARY                         | -0.042161|
| X10      | Gross Domestic Product         | -0.022438|
| X11      | disposable income per capita   | -0.02952 |
| X12      | Per capita consumption and expenditure | -0.035831 |
| X13      | Regional Fixed Assets and Investment | -0.031807 |
| X14      | regional fixed asset investment index | -0.044339 |
| X15      | Unemployment                   | 0.045766 |
5.4 Logistic Regression Based on Person Correlation Coefficient

Significance test: Now, use the software calculated Logistic regression by SPSS 23.0. The non-missing value data of all the sample data, including the agent shall be included in the logistic binary regression equation, and the logistic regression shall be performed again after the removal of variables by using the conditional forward method. The test results are shown in the table 14.

Table 14. Significance test

| Variable     | Variable name                  | B     | S.E  | Wald     | Sig | Exp (B) |
|--------------|--------------------------------|-------|------|----------|-----|---------|
| X1           | LOAN DATE                      | 5.516 | 417  | 16162.631| 1   | .000    |
| X2           | AGENT                          | 512.998| 28.092| 13333.482| 1   | .000    |
| X5           | EDU LEVEL                      | -.967 | 374  | 6.699    | 1   | .010    |
| X7           | SALARY                         | 2.2491| 1.136| 4.810    | 1   | .028    |
| X10          | Gross Domestic Product         | 4.405 | .947 | 21.639   | 1   | .000    |
| X11          | disposable income per capita   | -18.979| 1.557| 148.512  | 1   | .000    |
| X12          | Per consumption expenditure    | 15.616| 1.325| 138.841  | 1   | .000    |
| X13          | Regional Fixed Assets Investment| 1.802| .563 | 10.255   | 1   | .001    |
| X14          | regional fixed asset investment index| 10.652| .633 | 283.180  | 1   | .000    |
| X15          | Unemployment                   | 2.470 | .519 | 22.665   | 1   | .000    |
| Y            | Constant                       | 10.470| .601 | 303.975  | 1   | 10,000  |

After gradually removing the non-significant variables, the regression results are obtained. See Table 14. It can be seen that these explanatory variables in the table have a robust explanatory effect on the explained variables so that they can be retained in the model. It can also be judged from the previous multicollinearity test results that these variables do not have multicollinearity, and the tolerance between the variables is relatively high, which will not have a significant impact on the accuracy of the parameter estimation results of the regression model.

Table 15. Likelihood Ratio Test

| Step | 2 Logarithmic likelihood value | Cox Snell R Square | Nagelkerke R Square |
|------|--------------------------------|--------------------|---------------------|
| 10   | 3140.094 b                     | .176               | .337                |

According to the estimation results, in Table 15, and The smaller the - 2log-likelihood, the higher the value of Cox Snell R square and Nagelkerke R square, and thus the better fit of this model.

Table 16. Hosmer-Lemeshow Inspection

| Step | Chi-square | Df | SIG.  |
|------|------------|----|-------|
| 10   | 89.960     | 8  | .000  |
The overall situation of the significance test of the regression equation is shown in Table 16. For logistic analysis, the chi-square of the Hosmer-Leme show goodness-of-fit test was 89.960, and the probability P-value significance level was less than 0.05, so the goodness-of-fit between explanatory variables and logit (P) was significant, hence the model was reasonable.

5.5 Establish a Personal Risk Assessment Model

The logistic regression model can be expressed as:

$$\log \frac{P}{1-P} = \alpha + \beta_1X_1 + \ldots + \beta_mX_m$$

According to this logistic regression test, the results of SPSS 23.0

Table 17. Logistic Regression Test Results

| Original Result | Training results | Training accuracy |
|-----------------|------------------|-------------------|
| 0               | 5077             | 5001              | 0                  | 1                 |
|                 | 5001             | 76                | 98.5%              |
| 1               | 698              | 535               | 23.4%              |
|                 | 163              |                   | Overall accuracy: 89.4% |

As can be seen from Table 17, In this paper, the average forecast accuracy is 89.4%, of which the forecast accuracy is 23.4% for deferred repayments and 98.5% for on-time repayments. The model has high accuracy in predicting customers' non-deferred repayment, while the accuracy of judging customers' deferred repayment is very low. Therefore, further tests are needed to determine the accuracy of deferred repayment and on-time repayment.

5.6 Summary and Prediction

Impact of variable screening on the model: In this chapter, the variables are first screened by the Person correlation coefficient test, and then the judgment accuracy of each model is analyzed.

Finding of Comparison of Models: by the detection of the three models described above. A summary of the predictive accuracy of each model was obtained, see Table 18.

Table 18. Prediction accuracy of each model

| Model name          | Performance accuracy rate: 0 | Total Accuracy |
|---------------------|------------------------------|----------------|
|                     | Breach Correct               |                |
| Random forest       | 97.7%                        | 94.1%          |
| Discriminate Analysis | 80.9%                        | 78.7%          |
| Logistic            | 98.5%                        | 89.4%          |
|                     | 23.4%                        |                |

No overdue payments Record documented Dataset of samples in this paper, and this chapter tend to use the random forest to distinguish the data set of samples, i.e., If for a sample data set with an overdue repayment record, use Discriminate Analysis is more appropriate in a way that enhances the probability of judging overdue payments.
6. Screening of Explanatory Variables for Sample Data without Agents

6.1 Correlation of Explanatory Variables with Explained Variables

In this chapter, 12 variables are selected to study the influence factors of the borrower's deferred repayment. They are the loan grade (x1), whether local nationality (x3), education level (x5), marital status (x6), fund or not (x8), gender (x9), provincial gross product (x10), per capita disposable income (x11), per capita consumption expenditure (x12), regional fixed investment (x13), regional fixed investment index (x14), unemployment rate (x15). The data samples not including agents generally lack working provinces and salaries, so we added six macro data variables corresponding to the province of origin. So, this chapter first uses SPSS 23.0 to pass Person correlation test was used to explore the correlation between explanatory variables and explained variables. The explanatory variables were screened by the magnitude of the correlation coefficient between the dependent and independent variables and the requirement of significance between the two.

Table 19. Person Correlation Test

|     | Y     |     |
|-----|-------|-----|
| X1  | -0.179| X10 | -0.001|
| X3  | 0.020 | X11 | -0.036|
| X5  | 0.015 | X12 | -0.052|
| X6  | -0.031| X13 | -0.131|
| X8  | -0.045| X14 | -0.150|
| X9  | 0.001 | X15 | 0.067 |

According to the correlation test results, except for Fig. X6 failed the significance test, and all other variables passed the significance test. Considering that the remaining variables are 11 One, so it is not filtered according to the correlation magnitude of variables and finally selected the explanatory variables were X1, X3, X5, X8, X9, X10, X11, X12, X13, X14, X15.

6.2 Multicollinearity Analysis

In order to ensure the accuracy of the prediction results of the constructed model, the first step was to use SPSS 23.0 to test for the presence of multicollinearity between variables. The results are shown in Table 21. It can be seen that the inflation factor VIF (Variance inflation factor) of 11 variables is between 0 and 10, from which it is judged that there is no severe multicollinearity between the 11 variables.

Table 20. Result of multicollinearity

| Model | Unnormalized coefficient | Standardization coefficient | Significance | Collinearity statistics |
|-------|--------------------------|-----------------------------|--------------|------------------------|
|       | B                        | Standard error              | Beta         | T                      | Tolerance | VIF  |
| 1     | (Constant)               | 0.139                       | 0.011        | 13.188                 | 0         | 1.028|
|       | X1                       | 0.003                       | 0.001        | 0.034                  | 4.853     | 0.973| 1.028|
|       | X3                       | 0.019                       | 0.003        | 0.046                  | 6.401     | 0.941| 1.062|
|       | X5                       | -0.009                      | 0.003        | -0.025                 | -3.555    | 0.972| 1.029|
|       | X6                       | 0.018                       | 0.003        | 0.043                  | 6.095     | 0.959| 1.042|
|       | X9                       | 0.025                       | 0.003        | 0.054                  | 7.79      | 0.982| 1.018|
|       | X10                      | -0.002                      | 0.001        | -0.067                 | -3.72     | 0.146| 6.856|
|       | X11                      | 0.004                       | 0.001        | 0.119                  | 4.473     | 0.128| 9.726|
|       | X12                      | -0.004                      | 0.001        | -0.144                 | -5.118    | 0.108| 9.871|
|       | X13                      | -0.002                      | 0           | -0.086                 | -5.318    | 0.181| 5.511|
7. Modeling the Evaluation Model of Sample Data Credit Method without Agents

7.1 Modeling of Various Models

This chapter randomly selects data samples that do not contain agents from the 80% data as training data, where 708 for deferred repayment records, 15449. The bar is the on-time repayment record, and each variable is used as the training feature. R software is a random package to realize the modeling process.

Selection of variables: The variables were initially censored first. According to the correlation of each variable, the significance of each variable was judged and eliminated first x4 (marital status), again Introduced according to practical significance. Both the X7 (Fig. Province GDP), x8 (per capita disposable income), x9 (per capita consumption expenditure), x10 (regional fixed asset investment), x11 (regional fixed-asset investment index), x12 (provincial unemployment rate) to replace the working province variable and the native place variable and. When the screening variable is In 11 cases, repayment on time (0) is judged by the model as deferred repayment (1) The error rate is 0%, while deferred repayment (1) is judged by the model as on-time repayment (0) The error rate is 100%, overall The error rate is 4.4%.

Table 21. Preliminary training results of random forest model

| Original Result | Training results | Prediction accuracy |
|-----------------|------------------|---------------------|
|                 |                  |                     |
| 0               | 15449            | 0                   | 100%                |
| 1               | 708              | 0                   | 0%                  |

Overall accuracy: 95.6%

Considering the plethora of variables, among which the less relevant variables may affect the training effect of the model, resulting in the quasi-prediction Decreased certainty. Therefore, we consider eliminating some irrelevant variables step by step to make the model achieve the best prediction effect. According to Sen, the particular importance of forest variables is deleted from small to large. The screening rule is to eliminate the variables with the lowest degree of correlation based on the importance of each variable, as shown in the figure, and to eliminate X2 (whether local or not) with the lowest degree of correlation.
Through analysis, it is found that no matter several variables are eliminated; the training results are shown in the figure 6.

Selection trees for test. This chapter selects in order to investigate the influence of the number of trees selected on the accuracy of judgment, 200 trees are selected. The results are shown as follows:

![data.rf](image)

Figure 7. Verification of accuracy

As Figure 7 shown above, the graph abscissa represents the number of trees, and the ordinate represents the judgment error rate of the model, where green represents the error rate of the model-predicted deferred repayment, red represents the error rate of the model-predicted on-time repayment, and black represents the overall error rate. According to the figure above, the error rate is the same regardless of the number of trees.

Table 2. Random Forest Model Training Final Results

| Original results | Training results | Prediction accuracy |
|------------------|------------------|---------------------|
|                  | 0                | 1                   |                     |
| 0                | 15449            | 15449               | 100%                |
| 1                | 708              | 708                 | 0%                  |

Overall accuracy: 95.6%

From the prediction results in Table 2, it can be seen that the overall prediction of the model the accuracy rate is 95.6%, of which the judgment of the people who repay on time is more accurate, and its prediction accuracy up to 100%; and the prediction accuracy for the deferred payoff population is very low, i.e., is 0%. The reason for this result may also be related to the selection of sample size.

7.2 Discriminate Analysis

Currently, according to the previous analysis, the test results of the Person correlation coefficient test and analysis are selected according to the explained variables (whether to postpone repayment). Significant correlative relationship 1 one variable served as its observed indicator. These variables are, respectively, x1, x 3, x 5, x8, x9, x10, x11, x12, x1 3, x1 4, x1 5, the data from the samples containing agents without missing data were used as the training set data. Run through discriminate analysis was performed by SPSS 23.0 software, after which the results of the discriminate analysis were output.

Table 23. Prediction Results of Discriminate Analysis

| Original Results | Prediction results | Prediction accuracy |
|------------------|--------------------|---------------------|
|                  | 0                  | 1                   |                     |
| 0                | 19239              | 14531               | 75.5%               |
| 1                | 895                | 216                 | 679                 | 75.5%               |

Overall accuracy: 75.5%
7.3 Logistic Regression Analysis of Entropy Weight Method: According to the research method mentioned above, the result of entropy weight method is:

Table 2. Entropy Weighting Results

| Variables | Name of variable                  | Weight number |
|-----------|-----------------------------------|---------------|
| X1        | LOAD DATE Sort                    | 0.095096      |
| X2        | IS_LOCAL                          | 0.097815      |
| X3        | EDU_LEVEL                         | 0.131457      |
| X5        | HAS_FUND                          | 0.102543      |
| X6        | Gender                            | 0.088183      |
| X7        | Gross Domestic Product            | 0.089162      |
| X8        | disposable income per capita      | 0.07671       |
| X9        | Per capita consumption expenditure| 0.084632      |
| X10       | Regional Fixed Assets             | 0.088587      |
| X11       | Regional Fixed Asset Number       | 0.076631      |
| X12       | Unemployment                      | 0.069183      |

Next, according to the correlation coefficient matrix between the credit index and each factor, if this correlation coefficient is positive, then this factor entropy weight is also positive; if this correlation coefficient is negative, the factor entropy weight is also negative.

Table 25. User Credit Index Weighting

| Variable | Variable name                           | Weight     |
|----------|-----------------------------------------|------------|
| X1       | LOAD DATE Sort                          | + 0.095096 |
| X2       | IS_LOCAL                                | + 0.097815 |
| X3       | EDU_LEVEL                               | - 0.131457 |
| X5       | HAS_FUND                                | + 0.102543 |
| X6       | Gender                                  | + 0.088183 |
| X7       | Gross Domestic Product                  | - 0.089162 |
| X8       | disposable income per capita            | + 0.07671  |
| X9       | Per capita consumption expenditure      | - 0.084632 |
| X10      | Regional Fixed Assets                   | - 0.088587 |
| X11      | Regional Fixed Asset Number             | - 0.076631 |
| X12      | Unemployment rate                       | 0.069183   |
7.4 Logistic Regression

Significance test: The model was performed using SPSS 23.0 software Logistic regression, including all samples containing agent data in Data with non-missing values were included Logistic binary regression equation, and use conditional forward method to eliminate the variables, and then conduct a new round of Logistic regression. The test results are as shown in the table 26.

Table 26. Logistic Regression Results

| Variables | Variable name            | B     | S.E  | Wald   | Sig   | Exp (B) |
|-----------|--------------------------|-------|------|--------|-------|---------|
| X1        | LOAD DATE Sort           | 662   | .130 | 26.013 | .000  | 1.1939  |
| X2        | IS_LOCAL                 | 3.3194| .573 | 331.062| .000  | 24.381  |
| X3        | EDU_LEVEL                | -2.533| .688 | 13.548 | .000  | .079    |
| X5        | HAS_FUND                 | 4.830 | .828 | 33.995 | .000  | 125.245 |
| X6        | Gender                   | 7.209 | .960 | 56.327 | .000  | 1350.972|
| X7        | Gross Domestic Product   | - ...323 | .137 | 5.596  | .018  | .724    |
| X8        | disposable income per capita | 2.393 | .270 | 78.810 | .000  | 10.949  |
| X9        | Per consumption expenditure | -2.696 | .269 | 100.502| .000  | .067    |
| X10       | Regional Fixed-Assets    | -1.081| .114 | 90.492 | .000  | .339    |
| X11       | regional fixed-asset investment index | -3.118 | .224 | 194.432| .000  | .044    |
| Constant  | Constant                 | - ...165 | .230 | .518   | .472  | 848     |

After removing the variables with low significance step by step, it can be easily observed that these explanatory variables have a strong explanatory effect on the explained variables, so they should be kept in the model. It can also be learned from the multicollinearity test performed previously that the absence of multicollinearity in these several variables and the high tolerance between variables do not significantly affect the precision of the results of parameter estimation by the regression model.

Table 27.Likelihood Ratio Test

| Step | -2 value | Logarithmic likelihood | Cox Snell R Square | Nagelkerke |
|------|----------|------------------------|--------------------|------------|
| 10   | 6524.276b| .039                   | .127               |

According to the estimation results, in Table 28, - 2The smaller the log-likelihood, the higher the value of the Cox Snell R square and Nagelkerke R square, and the higher the fit of the model, so that the model can be considered to have a better fit.

Table 28.Hosmer-Lemeshow Fit test

| Step | Chi-Square | Df | SIG. |
|------|------------|----|------|
| 10   | 47.721     | 8  | 10,000|

The overall situation of the Hosmer-Lemeshow goodness-of-fit test of the regression equation is shown in the table, and it can be observed that the chi-square is 47.721 and the probability P-value significance level is less than 0.05. Hence, the correlation between the explanatory variables and logit (P) is significant, which can justify the model.
Establish a personal risk assessment model: In summary, logistic regression model can be expressed as:

$$\log it(p) = \ln \frac{p}{1-p} = \alpha + \beta_1 X_1 + \ldots + \beta_m X_m$$

The results of this logistic regression test were obtained by running SPSS 23.0 software:

Table 29. Logistic Regression Test Results

| Original Result | Training results | Training accuracy |
|-----------------|------------------|-------------------|
|                 | 0                | 1                 |
| 0               | 19239            | 19239             | 0 | 100.0% |
| 1               | 895              | 895               | 0 | 0.0%   |

Overall accuracy: 95.6%

In this chapter, the average forecast accuracy is 95.6%, of which the forecast accuracy is 0.00% for customers with deferred repayment and 100.0% for customers with timely repayment. The model has high accuracy in predicting customers' non-deferred repayment, while the accuracy of judging customers' deferred repayment is very low. Consequently, it is necessary to improve the test model further to improve the judgment of deferred repayment and on-time repayment accuracy.

7.5 Summary and Prediction

Impact of screening of variables on the model: In this chapter, the average forecast accuracy is 95.6%, of which the forecast accuracy is 0.00% for customers with deferred repayment and 100.0% for customers with timely repayment. The model has high efficiency in predicting customers' non-deferred compensation, while the accuracy of judging customers' deferred repayment is very low. Hence, it is necessary to improve the test model further to improve the judgment of partial compensation and on-time repayment accuracy.

Comparison of Models: The results were predicted by aggregating the three models described above. Obtain the prediction accuracy of each model, see Table 30.

Table 30. Prediction accuracy of each model

| Model            | 0 Correct rate of performance | 1 Correct rate of breach of contract | Total accuracy rate |
|------------------|-------------------------------|-------------------------------------|---------------------|
| Random Forest    | 100%                          | 0%                                  | 95.6%               |
| Discriminate Analysis | 75.5%                  | 75.9%                               | 75.5%               |
| Logistic         | 100%                          | 0%                                  | 95.6%               |

This chapter considers that there is no agent borrowing. Although the random forest model and the logistic regression model were both accurate at 95.6%, they were valid at 0% for the overdue population and did not work well for real-world applications.
8. Conclusion and Recommendation

8.1 Conclusion
In the sample with agents, the overall correct prediction rate of the random forest was 94.1%, discriminate analysis was 78.7%, and logistic regression was 89.4%. The prediction probability of random forest for overdue and non-overdue repayment was balanced, so the random forest model was more accurate and reliable for the general population. Nevertheless, the accuracy of discriminant analysis for overdue repayment prediction was higher than that of random forest. Discriminate analysis is suitable for the detection of the population with incomplete records.

In the non-agents sample, both random forest and logistic regression predicted 95.6% correctly, while discriminant analysis was only 75.5%. Nevertheless, discriminate analysis is more appropriate as a financial credit scoring model because the probability of predicting correctly using discriminant analysis is more than 75% for both overdue and non-overdue people. Although the prediction accuracy of the random forest model and the logistic regression model is 95.6%, the prediction accuracy for overdue repayment is 0, which is not practical for practical application.

8.2 Policy Recommendation
First, it is time to build a personal credit information system in line with China's national conditions. Compared with other countries, the construction of the personal credit information system in China started relatively late (Han et al., 2013; Cheng & Suyang, 2014). At present, a perfect and reasonable personal credit information system has not been formed, and personal credit information is lacking severely. Especially with the rapid development of China’s consumer credit market in recent years, a complete personal credit information system is urgently needed to guide the healthy development of the market (Huang et al., 2016). At present, most of the personal credit scores of the traditional credit agencies in China are still in the stage of subjective judgment and have high randomness (Hu & Ge, 2018). Although the personal credit scoring methods in some foreign countries are relatively mature and have been quantified by a large number of artificial intelligence and statistical methods, there are significant controversies on the performance and stability of each method, and China does not have the functional conditions to apply these methods (Sachs et al., 2007). Thus, China needs to build a personal credit system with Chinese characteristics.

Second, the construction of a suitable personal credit evaluation index system. The following two problems should be considered when establishing the evaluation index system. On the one hand, the evaluation system constructed should be able to make full use of all the data. On the other hand, the evaluation system constructed should be able to evaluate individual credit from multiple perspectives.

Finally, establish a suitable personal credit scoring model. In this paper, we tried other credit scoring models before determining the objective evaluation model, but the model discrimination ability and robustness are not as good as the random forest model. Discrete Analysis and logistic regression selected in this paper.

9. Research Prospects
Although this paper discusses a variety of personal consumption credit evaluation model, respectively, the random forest model, discriminant analysis, logistic regression model for empirical analysis and comparison, proved that the combination of model optimization role, but there are still some shortcomings in practical applications, mainly in:

First, as many variables as possible should be introduced. Variables used in this paper involve fewer types due to data type limitations. If the data can reflect the customer credit behavior, the effect of the model will be significantly improved. Due to the limited sample size in this research, further tests are needed to determine the accuracy of deferred repayment and on-time repayment.

Finally, efforts should be made to produce multilevel classifications. In this paper, the sample according to whether overdue agents and no agent sample, but in reality, is far from that simple.

References
Allen, F., Qian, J., & Qian, M. (2007). China’s Financial System: Past, Present, and Future, Available at SSRN: https://ssrn.com/abstract=978485.
Allen, L., DeLong, G., & Saunders, A. (2004). Issues in the credit risk modeling of retail markets, Journal of Banking & Finance, 28(4), 727-752.
Anonymous. (2017). Chinese Statistical Yearbook. Retrieved from http://www.stats.gov.cn/tjsj/ndsj/2017/indexeh.htm
Berger, A.N., & Udell, G.F. (2002). Small business credit availability and relationship lending: The importance of bank organisational structure, The economic journal, 112(477), 32-53.
Bikker, J.A., & Haaf, K. (2000). Measures of competition and concentration in the banking industry: a review of the literature, Economic & Financial Modelling, 1-46.
Chen, H.C., & Chen, Y.C. (2010). A comparative study of discrimination methods for credit scoring, The 40th International Conference on Computers & Industrial Engineering.

Creemers, R. (2018). China's Social Credit System: An Evolving Practice of Control, Available at SSRN: https://ssrn.com/abstract=3175792 or http://dx.doi.org/10.2139/ssrn.3175792.

Cheng, C., & Shuyang, O. (2014). The status quo and problems of the building of china's social credit system and suggestions, International Business & Management, 8(2), 169-173.

Dhillon, G., & Torkzadeh, G. (2006). Value-focused assessment of information system security in organizations, Information Systems Journal, 6(3), 293-314.

Durand, D. (1941). Appendix B: Application of the Method of Discriminant Functions to the Good-and Bad-Loan Samples, NBER Chapters, in: Risk Elements in Consumer Installment Financing, Technical Edition, 125-142, National Bureau of Economic Research, Inc.

Day, G.S., Shocker, A.D., & Srivastava, R.K. (1978). Customer-oriented approaches to identifying product-markets. Journal of Marketing, 43(4), 8-19.

Desai, V.S., Crook, J.N., & Overstreet, G.A. Jr. (1996). A comparison of neural networks and linear scoring models in the credit union environment, European Journal of Operational Research, 95(1), 24-37.

Dorronsoro, J.R., Giné, F., Sgn, & Cruz, C.S. (1997). Neural fraud detection in credit card operations, IEEE Transactions on Neural Networks, 8(4), 827-834.

Ennew, C.T., & Binks, M.R. (1999). Impact of participative service relationships on quality, satisfaction and retention: an exploratory study, Journal of Business Research, 46(2), 121-132.

Eisenbeis, R.A. (1977). Pitfalls in the application of discriminant analysis in business, finance, and economics, The Journal of Finance, 32(3), 875-900.

Hsieh, N.C., & Hung, L.P. (2010). A data driven ensemble classifier for credit scoring analysis, Expert systems with Applications, 37(1), 534-545.

Huang, C.L., Chen, M.C., & Wang, C.J. (2007). Credit scoring with a data mining approach based on support vector machines, Expert systems with Applications, 34(4), 847-856.

Hoff, K., & Stiglitz, J.E. (1990). Introduction: Imperfect information and rural credit markets: Puzzles and policy perspectives, The World Bank Economic Review, 4(3), 235-250.

Hand, D.J., & Henley, W.E. (1997). Statistical classification methods in consumer credit scoring: a review, Journal of the Royal Statistical Society, 160(3), 523-547.

Han, K., Lee, Y., & Park, C. (2013). Legal frameworks and credit information systems in China, Korea, and Singapore, Asian-Pacific Economic Literature, 27(1), 147-155.

Huang, Z., Lei, Y., & Shen, S. (2016). China's personal credit reporting system in the internet finance era: challenges and opportunities, China Economic Journal, 9(3), 288-303.

Hu, Y., & Ge, Z. (2018). The Development Dilemma and Countermeasures of China's Personal Credit Industry in the Era of Large Data, ATCI 2018: International Conference on Applications and Techniques in Cyber Security and Intelligence, 1023-1030.

Kostka, G. (2019). China's social credit systems and public opinion: Explaining high levels of approval, New Media & Society, 21(7), 1565-1593.

Lin, Z., Whinston, A.B., & Fan, S. (2015). Harnessing Internet finance with innovative cyber credit management, Financial Innovation, 5(DOI:10.1186/s40854-015-0004-7).

Lachenbruch, P.A., & Goldstein, M. (1979). Discriminant analysis, Biometrics, 35(1), 69-85.

Li, M. (2017). Comparative data mining analysis of personal credit scoring models, Times Finance, 23(6), 295-298.

Mao, Q., Hu, F., & Hao, Q. (2018). Deep learning for intelligent wireless networks: A comprehensive survey, IEEE Communications Surveys & Tutorials, 20(4), 2595-2621.

Nwana, H.S. (1996). Software agents: An overview, The Knowledge Engineering Review, 11(3), 205-244.

Ripley, B.D. (1994). Neural networks and related methods for classification, Journal of the Royal Statistical Society: Series B (Methodological), 56(3), 409-437.

Stiglitz, J.E. (1993). The role of the state in financial markets, The World Bank Economic Review, 7(1), 19-52.

Sachs, T., Tiong, R., & Wang, S. Qian. (2007). Analysis of political risks and opportunities in public private partnerships (PPP) in China and selected Asian countries: Survey results, Chinese Management Studies, 1(2), 126-148.

Shi, X., & He, X. (2015). The Study of Skew-Logistic Model and Its Application in Credit Scoring, Journal of Applied Statistics and Management, 34(6), 1048-1056.

Sohn, S.Y., Kim, D.H., & Yoon, J.H. (2016). Technology credit scoring model with fuzzy logistic regression, Applied Soft Computing, 43, 150-158.
Su, H. (2018). The research of personal credit risk assessment based on random forest model (Master’s thesis, Hunan University, Changsha, China). Retrieved from http://www.hnu.edu.cn/.

Shi, Q. (2005). Research on a Mixed Two-Phase Personal Credit Scoring Model Based on Neural Network-Logistic Regression, *Statistical Research*, 19(5), 45-49. DOI: 10.19343/j.cnki.11-1302/c.2005.05.011.

Thomas, L.C. (2000). A survey of credence and behavioural scoring: forecasting financial risk of lending to consumers, *International journal of forecasting*, 16(2), 149-172.

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

Wiginton, J.C. (1980). A note on the comparison of logit and discriminant models of consumer credit behavior, *Journal of Financial and Quantitative Analysis*, 15(3), 757-770.

Wu, W. (2008). Dimensions of social capital and firm competitiveness improvement: The mediating role of information sharing, *Journal of management studies*, 45(1), 122-146.

Wind, Y. (1978). Issues and Advances in Segmentation Research, *Journal of Marketing Research*, 15(3), 317-337.

Xia, Y., Liu, C., Da, B., & Xie, F. (2018). A novel heterogeneous ensemble credit scoring model based on bstacking approach, *Expert Systems with Applications*, 93, 183-199.

Yu, L., Wang, S., & Lai, K. (2009). An intelligent-agent-based fuzzy group decision making model for financial multicriteria decision support. *The case of credit scoring, European journal of operational research*, 194(3), 942-959.

Yang, C. (2018). Multi-dimensional Optimal Selection Strategy for Credit Evaluation Methods, *Statistics & Decision*, 34(21), 80-85. DOI: 10.13546/j.cnki.tjyjc.2018.21.019.

Yip, G.S., & McKern, B. (2016). China’s next strategic advantage: From imitation to innovation, *The MIT Press*.

**Appendix**

Library (t randomForest") options (max. Print = 1000000)

Data < -read.csv (t I New And... Csv", header = T, sep =", ")); data [, 6] < -as. Factor (data [, 6])

" Turn dependent variable y into factor."

Variable Class (data [, 6]) set.seed (100)

Ind = sample (1, nrow (data), replace = T, prob = c (0.8, 0.2))

" The whole sample into 8:2 training samples and forecast samples."

Data.rf There was no significant difference between the two groups (y ~., Fig.

Data [ind = = 1,], ntree = 50, nperm = 10, mtry = 3, stability = T, import = T)

Print (Fig. data.rf ) plot (i.e. data.rf )

data.pred The effect of (= predict (i.e. data.rf , data [ind = = 2,]) table (observed = data [ind = = 2, t y"],

predicted="data.pred )

Import (Fig. data.rf Type = 1 )

" The score of each variable."

Import (Fig. data.rf Type = 2)

VARIMPPLOT ( data.rf)

**Copyrights**

Copyright for this article is retained by the author(s), with first publication rights granted to the journal. This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).