Research on Financial Credit Risk Prediction under the Internet Background

Sunhan Rao [a], Guopeng Chen [b]

College of Business and Public Management, Wenzhou-Kean University, 325060, China

[a] 1129950@wku.edu.cn, [b] 1129092@wku.edu.cn

Abstract. As a new type of financial business model linked to Internet technology, Internet finance can solve the financing difficulties of individuals and small and medium-sized enterprises to a certain extent. With the continuous development of Internet technology, the data sources of Internet finance are more extensive and the scale of data is getting bigger and bigger. This paper combines two two-step sub-sampling algorithms with logistic regression model, and analyzes the prediction effect of Internet financial credit risk through the results of numerical simulation and empirical analysis. The logistic regression model based on the two-step subsampling algorithm maintains high accuracy and saves time significantly. Therefore, the regression model based on the two-step subsampling algorithm can make good predictions on the credit risk of Internet finance in both high-dimensional and low-dimensional Internet financial credit data.

Keywords: Internet finance; random forest-logistic model; algorithm.

1. Introduction

1.1 Research background

Internet finance is a new type of financial business model. It is a double-edged sword. On the one hand, Internet finance serves the real economy and promotes the development of traditional finance. On the other hand, the complexity of Internet finance has led to more risks in Internet finance. sex. Hong et al. divided Internet financial risks into operational risks, virtual risks, technical risks, legal and regulatory risks, and regulatory coverage risks. Song Yang and Wang Xiao pointed out that Internet finance has not changed the essence of finance, so Internet finance still has risks in traditional finance, such as credit risk, liquidity risk and operational risk. However, the foundation of Internet finance is modern information technology, so Internet financial risks are more difficult to predict than traditional financial risks [1].

1.2 Research significance

Scholars have long debated the nature of Internet finance. Xie believes that Internet finance is different from traditional finance and belongs to a third-party financial financing model [2]. Zhou Yu believes that the emergence of Internet finance has promoted the development of traditional physical finance and will not have an essential impact on physical finance. Yin holds a neutral attitude. It is believed that Internet finance mainly transforms the financial service mode, does not produce new financial products, and substantially expands financial functions [3]. In short, the development of Internet finance is closely related to physical finance, and credit risk control is equally important to Internet finance [4].

Therefore, this paper considers using a two-step sub-sampling algorithm to approximate the default probability of Internet financial credit risk [5]. This method firstly uses the simple random sampling method without replacement to extract part of the sample size, then uses the maximum likelihood estimation method to calculate the preliminary parameter estimates, and finally uses the obtained parameter estimates and the A-optimal criterion to calculate the sampling probability and further Samples are drawn, and a model is established for the samples obtained by the two samplings to predict the credit risk of Internet finance. The two-step sub-sampling algorithm solves the problem that the scale of big data is too large and difficult to handle to a certain extent [6]. It is more practical

in the context of big data and can better control the occurrence of credit defaults in Internet finance from the source [7].

1.3 Research ideas and methods

This paper firstly combines the two-step sub-sampling algorithm with the logistic regression model, and studies the accuracy and CPU running time of the logistic regression model based on the two-step sub-sampling algorithm through numerical simulation. Secondly, when there are many variables, the effect of the two-step sub-sampling algorithm is studied [8]. Taking the online lending data set as an example, the prediction effect of the random forest-logistic model and the Lasso-logistic model on the credit risk of online lending is compared, and the optimal model is selected to combine with the sampling algorithm, and then the accuracy and time cost of the sampling algorithm are compared[9]. Finally, when there are few independent variables, the prediction effect of the logistic regression model based on the two-step subsampling algorithm is studied. Taking the credit card fraud dataset as an example, the logistic regression model is combined with a two-step sub-sampling algorithm to compare the accuracy and time cost between the sampling algorithms [10].

2. Concepts and Basic Models of Internet Finance Credit Risk

The basic introduction to the credit risk model of Internet finance mainly consists of the following three parts: the related concepts of Internet finance credit risk, the basic model and the theoretical overview of evaluation indicators. Among them, the basic models of Internet financial credit risk include: random forest-logistic regression model, Lasso-logistic regression model, optimal sub-sampling algorithm and two-step sub-sampling algorithm [11].

2.1 Random Forest-Logistic Regression Model

The random forest method includes two algorithms: random forest regression and random forest classification. It is an algorithm that integrates many trees through the idea of integration. The main application of this paper is the random forest classification algorithm [12]. The basic steps are to use the Bootstrap method to randomly select n subsamples from all known samples, and then model each subsample as a decision tree. The prediction results of the model are summarized, and the sample is divided into categories with more votes, and the expression is as follows:

$$S(x) = \arg\max_{Y} \sum_{i=1}^{n} I(s_i(x) = Y)$$

The logistic regression model is a generalized linear model, which is widely used in many scientific research fields and can handle the problem of binary and multi-category dependent variables [13]. This article uses the binary logistic regression model, hereinafter referred to as the logistic regression model. The main difference from the linear regression model is that the dependent variable of the logistic regression model is a categorical variable, including two categories. Generally, the key type is recorded as 1, and the other type is recorded as 1. The independent variables of the logistic regression model can be continuous variables or discrete variables [14].

Let the observed value of the i-th sample be $X_i = \left(x_{i1}, \cdots x_{ip}\right)^T$, where X is the observed value of the independent variable, and Yi is the i-th dependent variable, taking the value 1 or . The standard logistic distribution is as follows:

$$P(Y_i = 1 \mid X_i) = p_i(\beta) = \frac{\exp(\beta^T X_i)}{1 + \exp(\beta^T X_i)}, i = 1,2, \dots, n$$

The deformation can get the linear form of the logistic regression model:

$$\log \frac{p_i(\beta)}{1 - p_i(\beta)} = \beta_0 + \beta_1 x_{i1} + \cdots \beta_p x_{ip} = \eta_p(X_i), i = 1,2, \cdots, n$$

Then the log-likelihood function of this distribution is:

$$l(\beta) = \sum_{i=1}^{n} [Y_i \log p_i(\beta) + (1 - Y_i) \log\{1 - p_i(\beta)\}]$$
$$= \sum_{i=1}^{n} \{Y_i \eta_p(X_i) - \log[1 + \exp(X_i)]\}$$

2.2 Lasso-logistic regression model

Lasso was originally developed based on the non-negative strangulation method, and was later improved by Tibshirani to propose the Lasso method. The objective function of the non-negative strangulation method is shown in Eq [15].

$$\sum_{i=1}^{n} \left(Y_i - a - \sum_{j} c_j \hat{\beta}_j^0 X_i^j \right) \text{ S.t. } c_j \geq 0, \sum c_j \leq \boldsymbol{t}$$

The Lasso-logistic regression model applies the Lasso method to the logistic regression model, that is, adding a penalty function to the formula to find the minimum value of the convex function formula.

$$S_\lambda(\beta) = -l(\beta) + \lambda \sum_{j} |\beta_j|$$

Get the coefficient estimates for the Lasso-logistic regression model:

$$\hat{\beta}_\lambda = \arg min \sum_{i=1}^{n} \left\{ y_i \eta_p(X_i) - \log\{1 + \exp(X_i)\} \right\} + \lambda \sum_{j} |\beta_j|$$

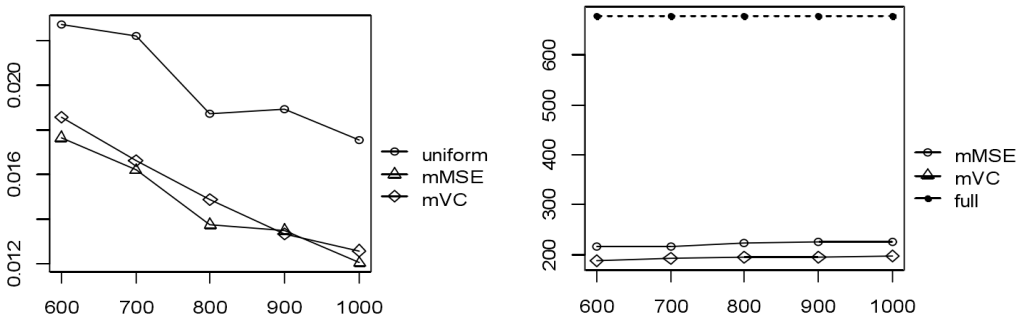
3. Empirical Analysis of Internet Finance Credit Risk

The two-step subsampling algorithm is often used in big data processing, and the extracted samples have good properties. Firstly, starting from the accuracy and time cost of the two-step sub-sampling algorithm, the two-step sub-sampling algorithm is compared with the full-sample and simple random sampling methods by means of numerical simulation. Predict the effect of analysis.

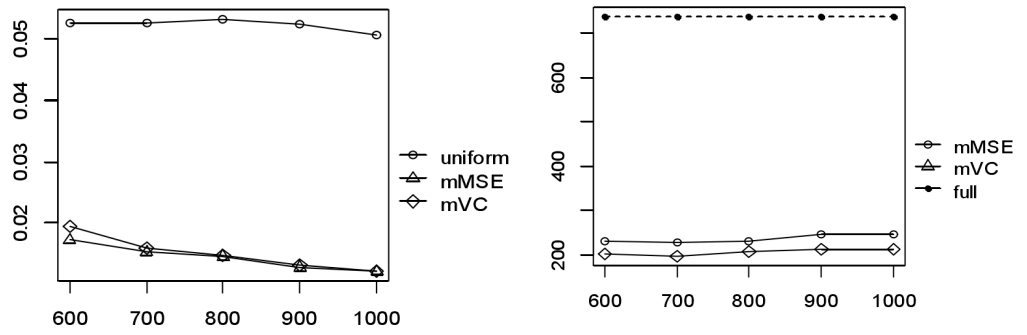
3.1 Numerical simulation

In order to introduce the two-step sub-sampling algorithm into the prediction of Internet financial risks, this paper firstly generates four distributions of data for the simulation, and compares the algorithm with the simple random sampling algorithm to test the accuracy and time cost of the algorithm[16]. In order to ensure the reliability of the results, the programs are run on Windowsl0, i5 processor, 4G memory laptop devices. Assume that the dependent variable Y in the model has two values, respectively. and 1; X is an independent variable, which is a 3-dimensional variable; the true value of R is (1, 1, 1), and this paper simulates N-1,000,000 data sets for the different distributions of X, where N represents the amount of data generated [17].

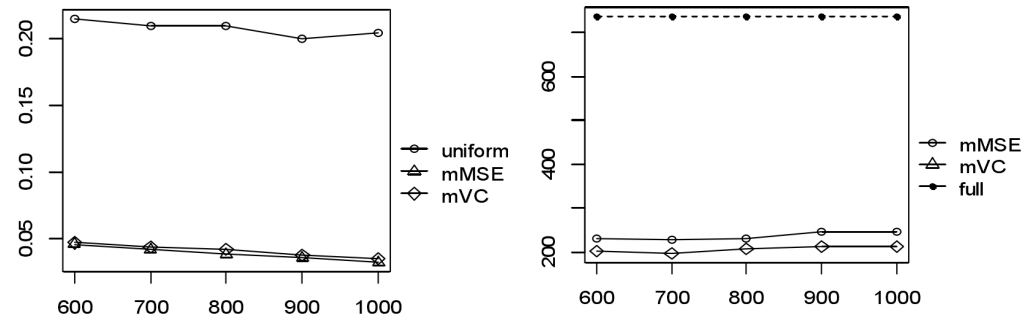
In order to evaluate the influence of the sample size of the second sampling on the two-step sub-sampling algorithm, this paper fixed the sample size of the first sampling r0 = 500, changed the sample size of the second sampling: r = 600, 700, 800, 900, and repeated K2 1000 times. Due to limited resources, it is impossible to use the existing computer to obtain 1000 prediction results. Therefore, the average mean square error of the model parameters is used to represent the accuracy of the model, and the total CPU running time represents the time cost of the algorithm[18]. For the distribution (1), (2) , (3), (4) Use different sampling methods to build logistic regression models and compare them. The average mean square error reflects the difference between the parameters solved after sampling and the parameters under the full sample to a certain extent. The total CPU running time refers to the time occupied by the program running 1000 times, which represents the running of the program to a certain extent. efficiency. The specific results are shown in Figure 1. The uniform in the legend represents simple random sampling [19].



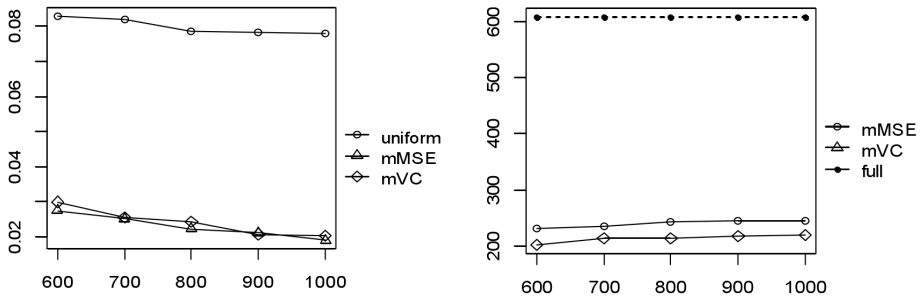
(a) Normal distribution



(b) t distribution



(c) Exponential distribution



(d) Mixed normal distribution

Figure 1. Average mean square error and total CPU running time of each sampling algorithm under different distributions

Through the simulation data, it is found that compared with the simple random sampling method, the average mean square error of the parameters of the two-step sub-sampling algorithm is smaller, indicating that the sample information extracted by the two-step sub-sampling algorithm is closer to the full sample information; The difference in the average mean square error of the algorithms is small, indicating that π_i^{mMSE} maintains the approximate accuracy of π_i^{mMSE} without calculating the inverse of the matrix; compared with the full sample, the two-step subsampling algorithm saves

more time. The above results show that the two-step subsampling algorithm saves time and cost while maintaining high accuracy [20].

3.2 Empirical results of RF-logistic model

The RF-logistic model first needs to use the random forest algorithm to select important variables, and then the logistic regression model can be established. 70% of the data set was randomly selected as the training set and 30% as the test set, and the random forest algorithm was used to screen important variables. After preliminary data cleaning, there are still many variables in the data set. If all variables are used to directly build a model, it will not only waste unnecessary time, but also may cause collinearity, resulting in inaccuracy of the model. According to Occam's razor principle, the model is not the more variables the better, but under the same conditions, the less variables the simpler the model, the better. Therefore, all the independent variables cannot be directly used to build the model, but the dimensionality of the variables should be reduced first, and the model should be built using the data after dimensionality reduction.

Table 1. RF-logistic regression model training set confusion matrix

	predict default	Predict compliance	total
actual default	34934	5801	40735
actual performance	3439	145424	148863
total	38373	151225	189598

In the test set, there are a total of 81,257 pieces of data, of which 17,225 pieces of data were actually breached, and 64,032 pieces of data were actually performed. Among the 16,296 samples that are based on default, when the 64,961 RF-logistic regression model predicted to be performance is used to predict the test set data, there are 14,841 samples that predict actual default, and 62,577 of the remaining 1,455 are spline samples that are actual performance. 2,384 of the samples actually performed the contract, and 2,384 actually breached the contract.

Using the established RF-logistic regression model to predict the test set data, the results obtained are shown in Table 2.

Table 2. Confusion Matrix of RF-logistic Regression Model Test Set

	predict default	Predict compliance	total
actual defaul	14841	2384	17225
actual performance	1455	62577	64032
total	16296	64961	81257

In the test set, there are a total of 81,257 pieces of data, of which 17,225 pieces of data were actually breached, and 64,032 pieces of data were actually performed. Among the 16,296 samples that are based on default, when the 64,961 RF-logistic regression model predicts the test set data, there are 14,841 samples that predict actual default, and 62,577 of the remaining 1,455 samples are spline samples that actually perform. 2,384 of the samples actually performed the contract, and 2,384 actually breached the contract.

The data set is increased from 67 variables to 81 variables, of which the repayment status is the dependent variable, 20 dummy independent variables, and 60 numerical independent variables. According to the division method of the data set by the RF-logistic model, 70% of the data set is used as the training set and 30% as the test set, and the Lasso-logistic model is established by using the

data of the training set, and the AUC value is used as the standard for screening variables. The results are shown in Figure 2.

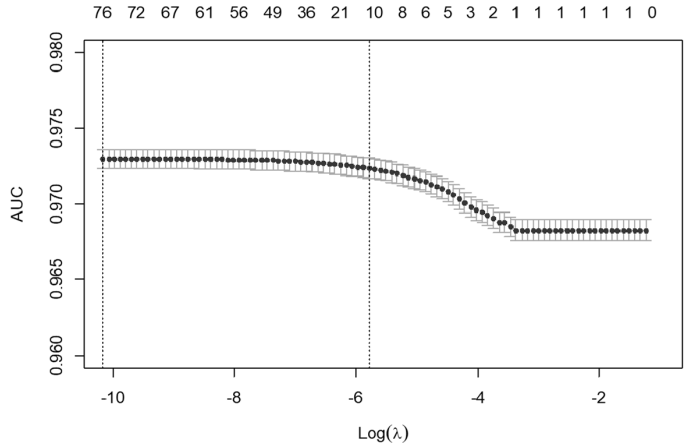


Figure 2. Selection of the optimal number of variables for the Lasso-logistic model

3.3 Empirical results of Lasso-logistic model

Among the variables screened by the Lasso-logistic model, 9 variables are the same as those selected by the RF-logistic model, and the regression coefficients of the same variables have the same sign. The number of bankruptcy records and the loan repayment interval are two new variables that are different from the Lasso-logistic model and the RF-logistic model, and the coefficients are all positive numbers, which means that the more bankruptcy records, the longer the repayment interval. Easier to default.

Table 3. Lasso-logistic regression model training set confusion matrix

	predict default	Predict compliance	total
actual default	34758	5977	40735
actual performance	3298	145565	148863
total	38056	151542	189598

Table 3 shows that in the training set, there are a total of 189,598 pieces of data, of which 40,735 pieces of data are actually breached, and 148,863 pieces are actually performed. Using the Lasso-logistic regression model to predict the training set, among the 38,056 samples predicted to be defaults, there are 34,758 actual defaults and 3,298 actual performances; among the 151,542 samples predicted to be performances, 145565 samples actually performed. , there are 5977 samples of actual default.

The established RF-logistic regression model is used to predict the test set data, and the obtained results are shown in Table 4.

Table 4. Lasso-logistic regression model test set confusion matrix

	predict default	Predict compliance	total
actual default	14711	2454	17225
actual performance	1406	62626	64032
total	16117	65080	81257

In the test set, there are a total of 81,257 pieces of data, of which 17,225 pieces of data were actually breached, and 64,032 pieces of data were actually performed. When making predictions

based on the Lasso-logistic model, among the 16,117 samples predicted to be defaults, 14,711 samples actually defaulted, and the remaining 1,406 samples were samples of actual performance; among the 65,080 samples predicted to be performance, 62,626 samples actually Performance, 2454 actual breach of contract.

As can be seen from Table 4, the accuracy, recall and value of the RF-logistic model on the test set are higher, and its accuracy, recall and Fi value are 0.11%, 0.76% and 0.00300RF higher than those of the Lasso-logistic model, respectively. The accuracy of the -logistic model is 0.21% lower than that of the Lasso-logistic model on the test set. On the whole, on the test set, the prediction effects of the two models on the credit risk of online lending are also similar. The ROC curve can intuitively show the classification performance of the model, as shown in Figure 3.

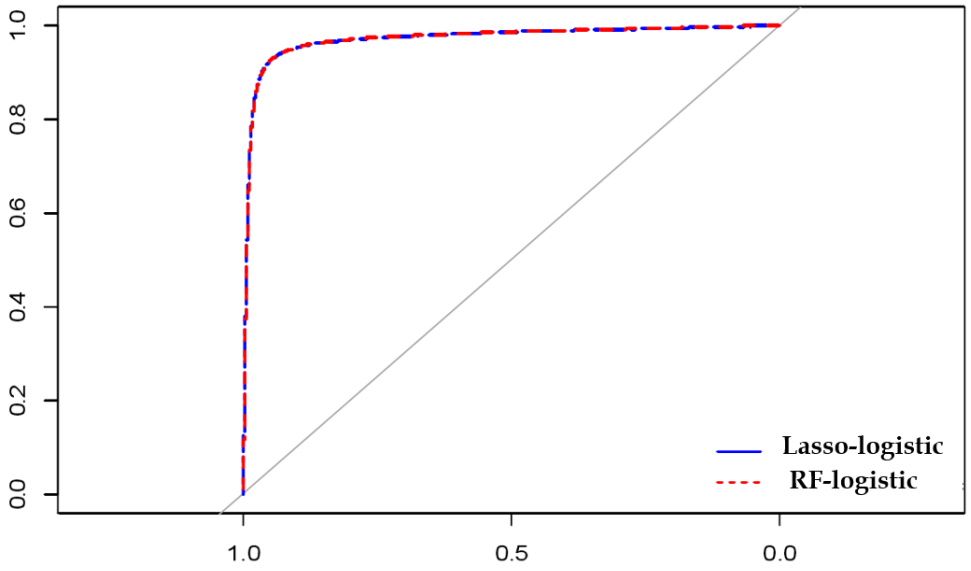


Figure 3. ROC curves of RF-logistic model and Lasso-logistic model

4. Conclusion

This paper combines the two-step sub-sampling algorithm with the logistic regression model, and analyzes the effect of the two-step sub-sampling algorithm in the credit risk prediction of Internet finance through the results of numerical simulation and applied empirical results. Through numerical simulation comparison, the prediction effect of the logistic regression model based on the two-step subsampling algorithm and the logistic regression model based on the simple random sampling method. Compared with the logistic regression model based on simple random sampling, the logistic regression model based on the two-step subsampling algorithm is closer to the results obtained by all samples. The logistic regression model established based on the two-step subsampling method is close to the logistic regression model established based on all samples in terms of precision, prediction accuracy, and classification performance. It saves a lot of time, which can well solve the problems of large data scale, computer can't process it or it takes too long in the background of big data.

Referencnes

[1] The Influence of Internet Credit on College Students' Consumption Behavior [J]. Zhang Yi. Cooperative Economy and Technology. 2021(19).

[2] The Development Model and Governance Reflection of China's Internet Consumer Finance——Based on the Experience of Countries Along the "Belt and Road" [J]. Cheng Xuejun. Consumer Economy. 2021(04).

[3] Research on Financial Economic Cycle and Bank Credit Risk Management [J]. Ren Baojun. Business Exhibition Economy. 2021(12).

[4] Research on the Influence of Internet Consumer Credit on the Consumption Behavior of Higher Vocational College Students [J]. Fan Pingping, Xu Wang. Neijiang Science and Technology. 2021(06)

[5] Opportunities and Challenges for the Development of Internet Consumer Finance [J]. Wang Jingli. Financial Theory and Teaching. 2021(03).

[6] Analysis of the Development Status of Internet Consumer Finance in Heilongjiang Province——Taking the Internet Consumer Finance of College Students as an Example [J]. Wang Xiaojia, Zhang Dechun. China Collective Economy. 2021(16).

[7] Research on the Consumer Behavior of College Students in the Internet Credit Environment [J]. Zhang Ling. Computer Knowledge and Technology. 2021(09).

[8] Discussion on the governance of college students' campus loans from the perspective of information asymmetry [J]. Li Yulin. Higher Education Forum. 2021(01).

[9] Research on the Effect of Internet Consumer Finance on the Consumption Demand of Urban Residents——Analysis Based on the Perspective of Liquidity Constraints [J]. Quiet. Science and Technology and Industry. 2020(12).

[10] The development of Internet consumer finance in the new era [J]. Yu Lun. Modern Marketing (late issue). 2020(11).

[11] Discussion on the Consumer Financial Market of College Students under the Internet Background [J]. Dai Hsinchu, Shao Hantong. Industry and Technology Forum. 2020(21).

[12] Cultivation of college students' consumption concept under the background of Internet finance [J]. Li Chang. Science and Education Wenhui (late issue). 2020(04).

[13] Research on the Influence of Internet Finance on College Students' Consumption [J]. Ma Yu, Wu Weirong. China Collective Economy. 2020(12).

[14] Research on the problems and countermeasures of college students' online loan [J]. Zhang Fufu, Wu Xiaorong. Jiang Science Academic Research. 2019(04).

[15] Research on the problems and countermeasures of college students' online loan [J]. Zhang Fufu, Wu Xiaorong. Jiangxue Academic Research. 2019 (04).

[16] Strengthening consumer ethics education in the new era [J]. Zhou Zhongzhi. China Moral Education. 2019(20).

[17] Analysis of the Risks and Countermeasures of Internet Consumer Finance in my country [J]. Li Jinqiu. Liaoning Economic Vocational and Technical College. Journal of Liaoning Economic Management Cadre College. 2019(04).

[18] On the Risk Management of Internet Consumer Finance [J]. Xia Zhongxi. Modern Marketing (Management Edition). 2019(05).

[19] Research on the credit risk management of college students' consumer finance under the background of the Internet - Taking Anhui Province as an example [J]. Niu Yuxin, Zhang Zhenxing, Wu Fan. Modernization of shopping malls. 2018(24).

[20] Research on the consumer behavior of college students in the Internet era [J]. Wang Fengshuang. Modernization of shopping malls. 2018(13).