Advances in Economics, Business and Management Research, volume 178
Proceedings of the 2021 International Conference on Enterprise Management and Economic Development (ICEMED 2021)

The Application of Big Data in Preventing Financial Risks of P2P Network Loan

ZHIQING LI[1]

[1]*College of Applied Science and Technology，University Of Hainan，Haikou，571700，China*
Email：BaileyGong@163.com

ABSTRACT
With the increasing impact of the Internet on people's lives, Internet financial platforms are also emerging, According to the data published on zero one, By the end of 2020, P2P the number of online lending platforms has reached 6063, Among them, the number of platforms operating in normal state is 1185, 19.5 per cent of the total number of platforms, The year-on-year decline was 46.8 per cent, The number of platforms in an abnormal state is 4672, The proportion of the total number of platforms is 77.1. Although less, But the number of problem platforms is still huge. The number of problem platforms is enough to worry, along with credit risk, A lot of platform loans can't be recovered, The non-performing loan rate has risen. From the beginning of online shopping, To Internet banking, third party payments, On the Internet, the rapid development of Internet finance takes only a few years. Especially in P2P, the development of recent years has reached an alarming rate, but there are also a series of credit risk problems.
In order to ensure that the credit risk of Internet financial platform can be effectively reduced, it is necessary to use modern big data technology to rate the risk. This paper mainly studies a large amount of data formed by users on the Internet, and uses data crawler technology to collect, store and transmit data. After that, big data analysis software RapidMiner big data analysis software are used to preprocess the crawling data, and random forest algorithm is used to match the best data and rule scheme in the software. Different from the traditional research on Internet financial credit risk, this paper uses logical regression analysis to evaluate the default probability of users according to the rules, and predicts the default and fraud of users. The results are analyzed. So as to effectively reduce the credit risk of Internet financial platform enterprises.

Keywords: *big data; Internet finance; credit risk*

1. BIG DATA APPLICATION PATTERN ANALYSIS

1.1. Big data credit model

In Internet finance, it can not only improve the speed of business processing, but also improve the efficiency of work. It also brings more problems to risk control. For example P2P the most important feature of online lending is personal credit as collateral. The application of big data technology is likely to make efficient and rapid personal credit information become a reality. For network lending, the most remarkable feature is efficiency, collateral is personal credit, which reduces a lot of more complex links and processes, but also because of the characteristics of the network. Big data technology can play its role. For example, in the network quickly grab some basic information of users, ip assets and so on. At present, the global practical American personal credit consumption score FICO model is obtained by big data technology. For example, when calculating the credit score of FICO, it will collect more than 1 million data information, quantify the credit qualification, literacy, consumption level, repayment level and other content of the user, fully understand the basic information of the user, and then get the final score in these indicators, such as Sesame credit score, which is a new way of personal credit based on Internet and big data using corresponding means and techniques.[1]

The biggest difference with foreign countries is that our country has adopted the government-led credit information model for a long time, which means that government departments can basically obtain all the company and personal credit information. The credit system in the economic market is very imperfect.[2]

2. ESEARCH ON USER CLASSIFICATION METHOD OF BIG DATA TECHNOLOGY

2.1. RapidMiner

2.1.1RapidMiner overview

RapidMiner, also known as the YALE Learning Environment, was born at Dortmund University of Technology in Germany and started with a project jointly developed by Ingo Mierswa 、RalfKlinkenberg and Simon Fischer in the artificial intelligence department. At the end of 2014, RapidMiner authorized the general agent in China to formally enter the predictive analysis market in China. Provide predictive analysis solutions, technical support, training and certification services for Chinese users.

RapidMiner the analysis structure of the corresponding components is similar to the tree structure in Microsoft system, each operator is marked on each node of the tree graph. The software contains various operators such as data processing, data modeling, data transformation, data exploration and data evaluation.RapidMinner have extended suites Rhadoop, can be integrated with Hadoop to run tasks on a cluster. Its data mining process is simple, powerful and intuitive, and has rich data mining analysis and algorithm functions. It is often used to solve various business key problems, such as resource planning, early warning prediction, social media monitoring and so on.[3]

2.1.2 RapidMiner Modeling Analysis Method and Process

Some inconsistent, incomplete and noisy problem data will increase with the increase of data acquisition dimension. Therefore, in order to ensure the high quality of the data, the excavated data should be preprocessed first. in this paper, we first import a series of data captured by crawler technology into the RapidMiner for dimensionality reduction, missing value processing, data variable conversion, bad data processing, data standardization, principal component analysis, attribute selection, data protocol and so on to obtain high quality data modules. The most extensive use of big data technology for Internet financial credit risk identification is the use of random forest and logical regression. In this paper, we combine the two to get new results.[4]

These include the following steps:

1) selective extraction and new data extraction from crawler data sources form historical data and incremental data respectively.

2) data exploration analysis and preprocessing are carried out on the two data sets formed by 1), including the exploration and analysis of data missing values and outliers. At the same time, the attribute specification, cleaning and transformation of the data are carried out.

3) part 2) the more complete data formed by preprocessing are modeling data, modeling based on random forest, logical regression and designed rules, training models, and identifying default risk.

4) different risk types are obtained according to the model results, and the corresponding processing results are obtained.[5]

2.2 Data extraction and preprocessing

Big data modeling analysis is the same as RapidMiner modeling analysis process. Before data analysis, we still need to preprocess the acquired data, some useless data, some missing data, repetitive data and some data containing smooth noise in the original data, etc.

2.2.1 Missing Value Processing

Statistically speaking, the data estimation will be biased by the missing value, and the representativeness of the sample data will also decrease. The missing value often exists in each big data, so it is also very important to deal with the missing value of the original data to some extent.

At present, the effective treatment of missing value is divided into two main aspects. One is to identify missing data, the other is to process missing data. Before the missing value processing, we should make an effective judgment. There are three methods of missing value processing, namely, data replacement method, direct deletion method and data interpolation method. The most common and effective method is data interpolation, which is generally divided into multiple interpolation and regression interpolation. Multiple interpolation refers to the automatic generation of a set of completed data for the missing value of the data, and the process is carried out several times, and the random sample data of the missing value is finally obtained. The regression interpolation rule is mainly to apply the regression model. The dependent variable of the model is the missing value that needs interpolation, and the independent variable is other related variables in the data set. The regression function is used as a tool to predict the dependent variables that need interpolation effectively.[6]

2.2.2 exception handling

In addition to the processing of missing values, the outliers in the original data also need to be processed by the corresponding program. Before processing, people need to be effectively identified. The common methods usually include variable box diagram and univariate scatter diagram.

The processing of outliers in the original data is not a simple elimination, because some abnormal data may contain valuable data, so the processing of outliers needs to be specific according to the situation, as shown in TABLE 1.

2.2.3 data transformation

Data change usually refers to the discretization of continuous variables, normalization and construction of the attributes of variables, the ultimate purpose of which is to make the data can meet the needs of the algorithm and further apply to data mining.

(1) Simple functional transformation

The simple transformation of function refers to the simple transformation of raw data by using common mathematical formulas, as shown in formulas 1 to 4.

$$x' = x^2 \quad (1)$$

$$x' = \sqrt{x} \quad (2)$$

$$x' = \log(x) \quad (3)$$

$$\nabla f(x_k) = f(x_{k+1}) - f(x_k) \quad (4)$$

The above method is usually used to deal with some raw data with non-normal distribution, which is transformed by simple mathematical function to make it have normal distribution state. In the process of time series analysis, simple function transformation can sometimes transform some abnormal sequences into stationary sequences which are easy to observe and calculate.[7]

(2) Standardization

The standardized processing in the process of data mining is the basic content of the whole work. The difference of different data values may vary greatly because of the different dimensions of the data. The final results of data analysis without standardized processing are likely to be affected. Therefore, when the data is normalized, the data is usually scaled according to a certain proportion, and a special falling interval is specified before scaling, so as to facilitate the comprehensive analysis of all the data. There are three common methods:

The first method is zero-difference standardization, also known as minimum-maximum standard method. Generally used for linear transformation of raw data,[0,1] interval is the most commonly used mapping interval in this method. The specific conversion process is shown in formula 5:

TABLE 1 Common Methods for handling outliers

Exception handling method	Method description
Delete records with outliers	Delete raw data with outliers directly
As missing values	Apply the missing value processing method to handle outliers
Average revision	Using the mean value of data before and after outliers to correct
Not processed	Direct modeling of data containing outliers

$$x^* = \frac{x - \min}{\max - \min} \quad (5)$$

The maximum value of the sample is expressed by max, the minimum value is expressed by min, the difference between the maximum and the minimum value of the sample is expressed by max-min, and the relationship between the original data will be expressed by deviation standardization.At the same time, if the system encounters a region beyond the expected range [min,max], it needs to redefine the range of regional values.

The second method is standard deviation standardization, which is also called zero-means standardization. The mean value of the data is equal to 0 and the standard deviation is equal to 1.Its specific transformation formula is:

$$x^* = \frac{x - \bar{x}}{\sigma} \quad (6)$$

In formula 6, the mean value of the original data is used to represent the standard deviation of the initial data. Standard deviation standardization method is one of the most widely used data standardization methods so far, but because the influence of data outliers on standard deviation and mean value is generally great, formula 6 needs to be modified accordingly. As shown in formula 7:

$$\sum_{i=1}^{i=n} |x_i - W| \quad (7)$$

Formula 7 uses M to represent the median to replace the mean in formula 6, and W to represent the absolute standard deviation to replace the standard deviation in formula 6.

The third method is decimal calibration standardization, which is easy to understand. Usually, the attribute value decimal of the target data is moved to make the target data fall within the range of [-1], and the maximum absolute value of the target data will determine the number of decimal moving digits. The conversion formula is shown in formula 8:

$$x^* = \frac{x}{10^k} \quad (8)$$

3. EMPIRICAL TEST OF THE MODEL ALGORITHM

3.1 Application of data crawler in this article

In this paper, we apply crawler technology to the Internet financial industry. In general, when evaluating the credit of the lending industry in the Internet, the common means include: querying the account flow of the loan customer, shopping record and habit and other basic information. Based on this, we can understand the customer's consumption and credit. For example, when the account flow in the lending customer's bank is authentic and reliable, and the purchase record is kept when shopping on the network platform, the customer's credit is good, reliable and economical. At the same time, this customer has the ability to repay the loan. This information can be obtained by crawling using crawler technology and achieving faster efficiency.[8]

On the other hand, the Internet lending industry, default, fraud, not only after the completion of their loans, there is a record of their default probability. Instead, it can be expected to default before its loan. The credit information of the Internet industry is obtained by focusing on crawler technology, and the data information is stored in the blacklist of enterprises, and the similarity is compared. Compare potential high risk default customers. It is not difficult to see that crawler technology support can achieve similar exposure sites crawling. In the rules of blacklist database collision, some data information is crawled and obtained on this series of websites.[9]

To sum up, crawler technology can not only crawl and obtain basic information such as account flow, shopping record and habit of loan customers, but also grab default, credit fraud blacklist and historical record information

3.2 Data preprocessing

In this study, the missing value processing, data transformation and other data protocols, cleaning and transformation solutions and processing methods. The data information which is consistent with and satisfied with the cleaning conditions is eliminated by the method of "Filter Examples"" in the RapidMiner, and the processing method is as follows: a series of data information satisfied with the cleaning conditions is eliminated. Then delete the attributes that are not related to credit risk identification, weak correlation or redundancy. At the same time, the data is transformed into "appropriateness" format, and the requirements of mining task content and calculation method are satisfied. Specific use of attribute structure and data standardization.[10] In this paper ,20% of the samples are randomly selected as test samples, and the remaining 80% are used as training samples. The Split Data" segmentation data operator in the RapidMiner is called to complete the grouping of test samples and training samples.

3.3 Model empirical test

This part mainly carries on the model algorithm empirical application. Sample usage is as follows: Import incremental data in the data module in the model. This article continues to crawl January 2019 data through crawler technology, with 1399 incremental data. After periodic collision testing and data cleaning extraction ,480 data were used for incremental testing. Finally, the possibility and probability of fraud are predicted, and the probability is transformed into credit value, rather than simply expressed as default value (1) or non-default case (0). This is also a special way to identify and judge fraud customers in combination with modeling.[11]

The default probability is set to 0.03/0.05/0.1/0.3, and the credit score of each user is calculated according to the piecewise function, which is divided into five parts :870-900/830-870/680-830580-680,<580. Define the credit rating for each part as A,B,C,D and E. The formula for calculating credit ratings is:

$$\text{Score} = \begin{cases} 900 - 1000 * p, p \le 0.03 \\ 870 - 2000 * (p - 0.03), p \le 0.05 \\ 830 - 3000 * (p - 0.05), p \le 0.1 \\ 680 - 500 * (p - 0.1), p \le 0.3 \\ 580 - 1000 * (p - 0.3), p > 0.3 \end{cases} \quad (9)$$

The credit rating and credit rating results of some users are shown in TABLE 2.

In this paper ,150 sample data were randomly selected from 480 sample data and calculated as credit score and rating. Combined with each credit rating, the sample size and fraud number of each part are calculated, as shown in TABLE 3 below.Thus, the fraud rate of this part increases relatively with the decrease of credit rating. And when the credit rating is reduced to the lowest stage, that is, the E segment, the proportion of fraudsters in this part is 26.67.Compared with the traditional credit rating, big data technology has an important auxiliary role. It can reduce the

errors and deviations caused by human causes while reducing the input of manpower, improving the efficiency of

TABLE 2 User Credit Score and Rating

Number	Non-compliance	Predicted default probability	Credit rating	Credit rating	Segment
1	0	0.00786311	892	A	860-900
2	0	0.00842175	891	A	860-900
3	0	0.00856324	891	A	860-900
4	0	0.01065312	890	A	860-900
5	0	0.01149537	889	A	860-900
6	0	0.01172415	888	A	860-900
7	0	0.01217532	887	A	860-900
8	0	0.01228656	887	A	860-900
9	0	0.01230691	887	A	860-900
10	0	0.01269543	887	A	860-900
11	0	0.01386513	886	A	860-900
12	0	0.01387129	886	A	860-900
13	0	0.01468221	885	A	860-900
14	0	0.01483293	885	A	860-900
15	0	0.01496787	884	A	860-900
16	0	0.01532594	884	A	860-900
17	0	0.01554981	884	A	860-900
18	0	0.01607932	883	A	860-900
19	0	0.01608328	883	A	860-900
20	0	0.01537623	884	A	860-900

work and audit, and rationally allocating personnel resources.[12]

With the help of credit rating model, when incremental data or new customers appear, we can calculate the corresponding credit rating, and combine the interval of credit rating to implement targeted processing methods. For example, when the customer's credit rating is very high, that is, A grade, it can be considered that there is no credit risk and can pass the review link directly. When the customer's credit rating is very low, that is, E grade, it can be considered that the credit risk is very high and can be directly rejected; When the customer's credit rating is in the middle stage, it is necessary to evaluate the credit rating, combined with manual examination or collection of other information.[13]

TABLE 3 Training Set Group Forecast

Sample number	Number of samples	Includes default and fraud users	Actual default rate	Projected default rate
1	295	121	0.4102	0.4074
2	295	43	0.1458	0.1448
3	295	39	0.1017	0.1010
4	295	26	0.088	0.0875
5	295	20	0.0678	0.0673

6	295	17	0.0576	0.0547
7	295	15	0.0508	0.0512
8	295	11	0.0373	0.0362
9	295	9	0.0305	0.0543
10	295	5	0.0169	0.0219

TABLE 4 Test Set Group Forecast

Sample number	Number of samples	Includes default and fraud users	Actual default rate	Projected default rate
1	295	113	0.3831	0.3671
2	295	41	0.1390	0.1359
3	295	40	0.1356	0.1137
4	295	35	0.1186	0.0975
5	295	24	0.0814	0.0773
6	295	16	0.0542	0.0497
7	295	14	0.0475	0.0412
8	295	8	0.0271	0.0362
9	295	7	0.0237	0.0343
10	295	4	0.0136	0.0119

TABLE 5 Forecast Default Rate and Actual Default Rate

Group number	Projected default rate	Real default rate
1	0.03942	0.02537
2	0.05072	0.04068
3	0.07136	0.05927
4	0.11231	0.12057
5	0.20731	0.25754

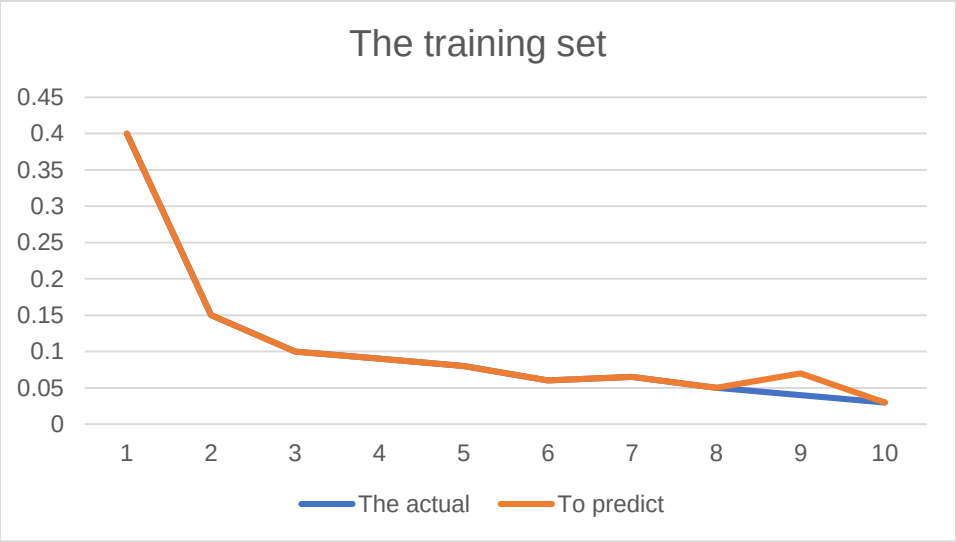

Figure 1 Group Forecast

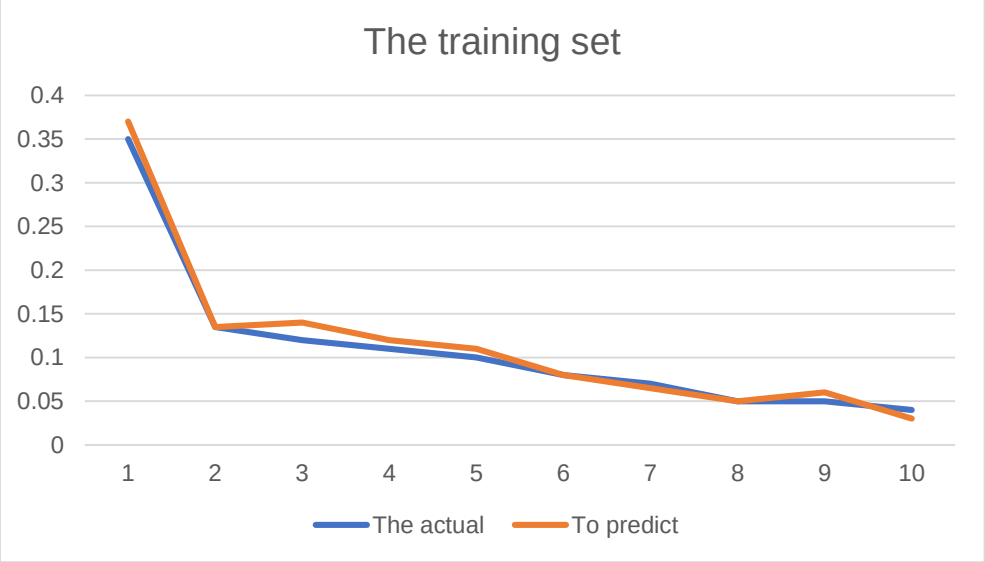

Figure 2 Packet Forecast

Combined with this series of methods, customer credit rating can be quantified, but also can save a lot of labor costs.

3.4 model validation

In this part, this paper mainly calculates the actual and predicted fraud rate in each data group in the training set and the prediction set, and judges the prediction ability of the model. It shows that the actual fraud rate is the ratio of the fraud customer and the sample data in the sample group, while the predicted fraud rate is the average of the predicted fraud probability.The actual fraud rate and predicted fraud rate values for each group are compared with the broken line diagram as follows (see TABLE 4, TABLE 5, Figure 1, Figure 2).Through comparative analysis, it is found that whether the data allocation is balanced or extremely unbalanced, the prediction ability of the model is better. This part does not use the traditional prediction error rate, that is, calculate the error ratio between the prediction structure and the actual category, and test the function and effect of the model.[14] Combined with the calculation of default probability, it can be transformed into credit score in the subsequent process, and then combined with credit score to evaluate the credit rating of its lending customers. For customers with different credit ratings to form a different loan process, that is: based on the final risk control strategy.[15]

Sort the probability of default, divide it into five equal parts, get the average of expected default rate of each part, and calculate the average of partial response variables, that is, predict the probability and status of each part, and compare the proportion of default. There is a significant positive correlation between actual default ratio and prediction probability (specific data and comparison data are shown in TABLE 5).[16

3.5 Comparison with traditional methods

At present, the measurement of Internet financial credit risk mainly depends on the logical regression model. After the logical regression of the data in this paper, the predicted default rate and the real default rate of five parts are obtained. Accuracy = true default rate / predicted default rate /5

That is, the total accuracy of the model.

The accuracy of this model is 0.91841, while that of logical regression model is 0.91534. On the whole, the difference is small and close to 1, but the accuracy of this model is slightly higher than that of logical regression model.

4. RESEARCH FINDINGS AND CONCLUSION

In the process of this study, the rule importance index is obtained by calculation, and the relevant mathematical model is established based on the screening rules, which is used to calculate the probability of user default. Combined with the calculated default probability, the final credit score is obtained and the credit rating is obtained. The purpose of this study is not to calculate whether the user has default probability —— that is, the output result is or not, but this general judgment classification often has the wrong result, the credibility is not high. For this problem, the improvement is carried out in the study, and the credit status of the user is scored by calculation.

Based on the relationship between the score interval and the probability of historical default, the credit score is divided into five grades, which are A、B、C、D、E grades respectively. The reason for this approach is that it

brings it closer to actual needs and is practical. If you can accurately calculate whether the user is a default user based on simple operation, of course, it is gratifying for the whole research, but in the process of actual operation, if there is a serious misjudgment, misjudgment or serious opportunity cost problem, you need to change the way. The specific measures adopted in the study are to formulate a targeted process scheme based on the predicted users of different credit ratings. If the credit rating of the tested object is at a high level, it can be subject to a more relaxed wind control audit and agree to its application; if the measured object's credit rating results are poor, The application can be rejected based on the automatic judgment of the computer system. Based on this processing operation to solve the resources at the same time, improve the efficiency of work, reasonably reduce the possibility of misjudgment, misjudgment, help Internet financial enterprises to improve their own level of risk prevention and control. Detailed analysis of this study and the conclusions obtained are of guiding significance to P2P how network lending enterprises can efficiently apply big data processing technology to improve their risk management level and business development efficiency.

Based on the empirical operation and results of this paper, big data technology has the characteristics of rapid data processing and clear reflection in Internet financial credit risk identification. Moreover, the data processing method adopted in this paper can be compatible with other factors, which is convenient for latecomers to increase or reduce the influencing factors and embodies the characteristics of strong practicability.

Finally, the application of risk control using big data technology is very popular at present. Relying on big data to establish wind control model has strong pertinence, comprehensive data and scientific modeling, which greatly reduces the possibility of misjudgment. Thus reducing the risk of enterprise losses, for the safe development of Internet finance to provide a solid guarantee. From this aspect, the practicability of this paper is also greatly improved.

REFERENCES

[1] Andrew Marshall,Alistair Milnc.Variable reduction,sample selection bias and bank credit scoring.Journal of Empirical Financc ,2010,17：501-512.

[2] Haddad,Gholam Reza Keshavarz,Gazar ,Ho sein Ayati.A Comparison between Logit Model and Classification Regression Trees (CART) in Customer Credit Scoring Systems. Quarterly Journal of the Economic Rescarch ,2007

[3] Galindo J,Tamayo P.Credit Risk Assessment UsingStatistical and Machine Lcarning Basic Mcthodology and Risk Modeling Application [J]. Computational Economics ,2000,15(1-2)：107-143.

[4] Hand D.J,Henley W.E.Statistical Classification Mcthods in Consumer Credit Scoring：A Revicw [J]. Journal of the Royal Stati stical Socicty ,1997,Series A 160(3)：523-541.

[5] Guo Y,Zhou W,Luo C,et al.Instance-based credit risk assessment for investment decisions in P2P lending [J]. European Journal of Operational Research ,2016,249(2)：417-426.

[6] Harris T.Quantitative credit risk assessment using support vector machines：Broad versus Narrow default definitions [J]. Expert Systems with Applications ,2013,40(11)：4404-4413.

[7] Koutanaci Fatemch Nemati,Sajedi Hedich,Khanbabaci Mohammad.A hybrid data mining model of feature selection algorithms and cnsemble learning classifiers for credit scoring.[J]. Journal of Rctailing &Consumer ,2015,27(C):11-23...

[8] Li Jiao. Research on Big Data Wind Control System of P2P Network Lending in China [J].]1 Operations and Management ,2021(01):161-165.

[9] Sun Ping. Research on P2P Network Lending Success Rate Model and Simulation J]. Driven by Big Data Journal of Ningbo Institute of Engineering 32(03):20-26+71.

[10] Ren Ronghuan. A probe into the risk and Supervision of P2P Network Lending in the era of big data [J]. Management and Technology for Small and Medium-sized Enterprises (forthcoming),2020(08):80-81.

[11] Ma Xiaojun, Song Yanqi, Chang Baishu, Yuan Mingyi, Su Heng. Study on the Application of P2P default Prediction Model based on CatBoost algorithm [J].].13 Forum on Statistics and Information ,2020,35(07):9-17.

[12] Kishino. Research on the Application of Big Data Technology in Credit Information System of P2P Network loan platform [J].]1 Foreign trade 2020(05):75-77+83.

[13] Strong South Korea. An Analysis of P2P Industry Risk Control Path —— Taking Guangdong Province as an example [J].].10 Special Economic Zone 2020(04):9-16.

[14] Huang Xiaohong, Fan Yantian, Liu Xiang. Analysis of Internet Financial Market based on SCP Model —— taking P2P Network loan Industry as an example Economy and Management ,2020,34(03):44-51.

[15] Cui Yan, Liu Lixin. Risk Evaluation of P2P Network Lending platform based on big data [J].]1 Forum on Statistics and Information ,2020,35(04):42-51.

[16] Li Yingchun, Peng Rui, Gao Kaiye. A study on P2P Network loan risk in the context of big data [J].]1 China Management Informatization 22(21):144-146.