Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 5986295, 11 pages
https://doi.org/10.1155/2022/5986295

Research Article

Research on Efficiency in Credit Risk Prediction Using Logistic-SBM Model

Dongmei Li and Liping Li

School of Management, Shanghai University, 333 Nanchen Road, Baoshan District, Shanghai 200444, China

Correspondence should be addressed to Liping Li; liliping@shu.edu.cn

Received 1 March 2022; Accepted 12 April 2022; Published 3 June 2022

Academic Editor: Maode Ma

Network lending, an innovative financial lending product, is separated from traditional financial media and implemented on the Internet platform. We study the credit risk prediction of online loan based on risk efficiency analysis. Moreover, we put forward the concept of borrower risk efficiency and apply it to risk prediction. The main task of this study is to establish risk efficiency characteristics on the basis of referring to various risk characteristics and carry out risk prediction after passing the screening of a series of features. The framework is realized by combining logistic regression and slack-based measure (SBM), and feature selection and verification are carried out through machine learning and statistics. Firstly, the efficiency risk characteristics are extracted and the risk efficiency is calculated by MaxDEA. Secondly, the features are screened and verified by Python. Then, the efficiency value obtained by SBM method is used as a new index for the training and testing of logistic model together with the initial related indexes. Moreover, in order to prove the effectiveness of the proposed credit risk prediction control scheme based on risk efficiency, the research compares the prediction before and after adding the risk efficiency feature. The simulation results demonstrated that the logistic-SBM model is more suitable for credit risk prediction than the commonly used logistic method, which realized the efficient prediction of credit risk based on the logistic-SBM model. Finally, some suggestions are put forward to China's regulatory authorities and the platform itself to control the credit risk of Internet lending industry.

1. Introduction

In "Interim Measures for the Management of Business Activities of Online Lending Information Intermediaries" promulgated in 2016, online lending is defined as direct lending between individuals including natural persons, legal persons, and other organizations through the Internet platform. Internet finance peer-to-peer (P2P) network finance is a branch of Internet finance, which is the product of the combination of Internet and finance. The academic definition of Internet finance has something in common with Internet finance, which is a new financial business model for traditional financial institutions and Internet enterprises to achieve financing. Davis and Gelpern and Slattery believe that P2P online lending has injected fresh vitality into the traditional lending market to meet the needs of investors and consumers [1, 2]. Financial technology based on P2P is one of the new breakthroughs in financial service institutions [3]. The main business models of Internet finance include Internet payment, online lending, equity crowd funding, Internet fund sales, Internet insurance, Internet trust, and internet consumer finance. Lenders have a greater impact on borrowers than do borrowers on lenders [4]. As technologies of big data and block chain advance, the financial credit risk in the context of the Internet has become a popular research subject [5]. P2P online lending originated in foreign countries. The earliest P2P online lending platform in the world is Zopa in the UK, which was established in London in March 2005. The new financial industry represented by peer-to-peer lending has gradually become a new source of volatility due to the increasing complexity of the Chinese financial market [6]. In 2007, China established its first P2P network lending enterprise. P2P lending platforms have different backgrounds and transparency [7]. Platform

background is related to operational risk [8]. The embryonic period of the development of P2P online lending financial enterprises is from 2007 to 2012. From 2013 to 2015, the development of P2P online lending financial enterprises has entered a period of vigorous expansion. From 2017 to now, it is a period of consolidation and standardization of P2P online lending financial enterprises. There are more than 10000 P2P online lending financial enterprises, in which more than 5000 were operated at the same time. The annual transaction scale is about 3 trillion yuan, and the bad debt loss rate is very high. Through continuous rectification, the People's Bank of China issued the "fintech development plan (2019-2021)" in September 2019, proposing to "further enhance the technology application ability of the financial industry and realize the deep integration and coordinated development of Finance and technology." By the beginning of 2020, there are already a lot fewer P2P online lending institutions in operation.

In China, the scope of definition of online lending includes both individual-to-individual lending, individual-to-business lending, and corporate organization-to-business organization lending. Since the birth of the first P2P in China in 2007, online lending has developed rapidly. To a certain extent, it is not only the result of the continuous advancement of modern information technology but also the inevitable product of the diversification of lending needs. However, the problems exposed have become more prominent during the development. Investors should pay attention to information asymmetry and credit risk impact [9]. Therefore, the problems of online lending industry in China have not only the common problems of other countries' online lending but also the specific problems of our country. Internet financial risk is not only directly related to the operation and development of the Internet financial system itself but has also a very important impact on the country's macroeconomic operation because of its rapid development speed and growing scale of development [10].

2. Literature Review

Since 2013, innovative Internet financial services such as Yirendai, Crowdfunder, and Renrendai have been born in China, promoting the reform of financial service models and accelerating financial marketization. Although there are a large number of online lending investors, they basically lack professional lending knowledge [11]. Moreover, the amount of online lending is small. When the lender lacks the effective information of the borrower, the bidding will often follow suit blindly and other irrational behaviors, which will inevitably increase the credit risk of online lending [12, 13]. However, the risk of the industry has also become obvious. The theory of information asymmetry was first put forward by Akerlof (1970) [14] by observing the phenomenon of used car market. In online lending, information asymmetry can also lead to the possibility of borrowers' default [15, 16]. The imbalance of these factors will lead to the platform's resources, and opportunities cannot play a role, resulting in the collapse of the platform [17]. The survival of the platform depends on the age, scale, and life cycle of the enterprise [18]. The management ability of platform operators plays a key role in the success or failure of small and micro platforms [19, 20]. For instance, in February 2017, 55 problematic platforms were involved in illegal fundraising, difficulty with cash withdrawal, fraud, absconding with money, and loss of connection and other risky breaches. Recent years have seen the rapid development of Internet finance in China, and various peer-to-peer (P2P) lending platforms have been released [8]. There is diversity of default behaviors of borrowers with different credit grades in online P2P loan market [21]. Reputation plays an important role in the long-term development of P2P lending platform [22]. These negative news have greatly affected investors' investment confidence and have had a very bad impact on the social reputation of the entire industry. Therefore, it is particularly important to scientifically evaluate Internet financial risks. The issue of risk and regulation of P2P lending platform in China is taken seriously. The P2P industry has promulgated the regulation that online loan platform must be online for fund deposit business, which makes bank deposit gradually normalized [23]. The difference between P2P online loan and traditional financial institutions lies in the transaction system of P2P online loan, which adopts the interest rate auction system when the transaction is concluded. Herzenstein and Barasinska [24] studied the interest rate of the American prosper online lending platform in 2011 and 2014, respectively. They found that the borrowers would set the maximum interest rate they were willing to pay for borrowing the funds, and then, the investors would decide whether to borrow according to the loan information provided by the prosper online lending platform. This innovative financial lending model provides investors with a new way of financial management. Liu et al. mainly find that investors' herd behavior exists significantly [25]. P2P mode can make the idle funds of investors not only increase in value but also meet the borrower's demand for funds to increase a loan channel. In this lending mode, the lending process no longer depends on offline financial institutions but relies on the network lending platform to match the needs of both sides and to realize the transaction. The reasons for choosing logistic-SBM model are as follows: DEA can be used to explore the new intersecting fields including management science, mathematics, mathematical economics, and operations research. DEA uses multiple inputs and outputs to measure the relative efficiency of each DMU. In the process of risk management for borrowers of Internet financial loan products, the DEA method can take each borrower as each DMU to obtain its efficiency value, rather than just studying the traditional indicators of the borrower. At present, there are few researches on the real customer credit data in China. Therefore, this study selected the logistic regression method for big data analysis through the comparison of different mathematical model methods. In this study, according to the characteristics of the source data, data envelopment analysis was used to process the source data and then, the data was trained in the logistic regression model to improve the accuracy of the model prediction. This method not only provides an innovative method to study the credit risk analysis of

Internet Financial borrowing customers but also expands the research space in this field, which has both theoretical and practical significance. Based on the present situation of the P2P lending platform development in our country, its development in the process of credit risk, transaction risk, legal risk, and so on is analyzed. In addition, corresponding regulatory measures were put forward to strengthen the development of P2P lending platform in China, which is greatly important.

3. Methods

The notion of probability is very closely related to the notion of symmetry [26]. Credit risk prediction is essential to predict the probability of default of borrowers. The specific research methods are as follows.

First is data preparation. This study divides the credit data of Internet financial technology companies into a sample set and a test set.

Then, SBM-DEA model was established. According to the above five indicators, the efficiency value of each customer was obtained by using DEA model through MaxDEA software. DEA_score was added to the next dataset.

The third step is feature processing. The feature processing methods include feature binning, correlation coefficient, IV, and random forest model.

The fourth step was to test the logistic-SBM model. The prediction results of the model are observed directly through the mixed matrix diagram. The AUC value of the model was calculated and tested. The model was tested by the K-S test.

In the last step, we compared the values of corresponding evaluation measures of two models.

The logistic-SBM model was established through MaxDEA and Python software.

3.1. Data Source Preparation. We used the real credit data of an Internet financial technology company as the analysis object. The company is mainly engaged in small loans, online finance, and other Internet financial products. The platform has a variety of data sources, high data quality, and rich data information. The loan customer risk management model to be studied in this paper selected the loan records of the platform. The sample population data was sampled and divided into a sample set and test set.

For the data selection, the loan with the end of repayment and the loan with default were selected for modelling. The target variable was selected according to the user's "repayment status" characteristics. If the loan has been repaid in which the default has not occurred, the value is 0. If there is overdue loan in which default occurs, the value is 1. Finally, 14028 transaction data that have been paid off were selected as the sample set, among which 10237 cases have been successfully paid off, accounting for 72.98% of the total number of samples. Besides, 3791 cases have overdue loans, accounting for 27.02% of the total number of samples.

3.2. SBM-DEA Method. This study used the SBM-DEA method (short for SBM method) to preprocess the data, because it can distinguish each customer to measure their respective efficiency value, rather than dividing them into different categories. This method can improve the prediction accuracy of the model and make the prediction of the initial logistic model more effective. The nonoriented SBM model is used in this study. The nonoriented SBM model is as follows:

$$\begin{aligned} \min \quad & \rho = \frac{1-(1/m)\sum_{i=1}^{m}(s_i^-/x_{ik})}{1+(1/q)\sum_{r=1}^{q}(s_r^+/y_{rk})} \\ \text{s.t.} \quad & X\lambda + s^- = x_k \\ & Y\lambda - s^+ = y_k \\ & \lambda, s^-, s^+ \geq 0 \end{aligned} \tag{1}$$

The SBM model uses $p*$ to represent the efficiency value of the evaluating DMU. It measures the inefficiency from both input and output, which is called the nonoriented model. In the unsupervised SBM model, there is no zero in the input and output data. In the SBM model, the inefficiency of input and output is reflected as follows:

$$\begin{aligned} & \frac{1}{m}\sum_{i=1}^{m}\frac{s_i^-}{x_{ij}}, \\ & \frac{1}{q}\sum_{r=1}^{q}\frac{s_r^+}{y_{rk}}. \end{aligned} \tag{2}$$

If the efficiency value ($p*$) of the SBM model is equal to 1, it means that the DMU evaluated is strongly efficient, while the efficiency of radial model is weakly efficient. The projection value (target value) of the evaluated DMU is

$$\begin{aligned} & \widehat{x_k} = x_k - s^-, \\ & \widehat{y_k} = y_k + s^+. \end{aligned} \tag{3}$$

The reasons for SBM indicator selection are as follows: the input indicators include borrower's liability information, credit risk score, and income information. These three indicators can mainly summarize the borrower's asset flow and external risk evaluation information. The output indicators are the borrower's loan amount and period, which are the most important indicators to describe the borrower's loan situation. Input and output indicators of the SBM method are shown in Table 1.

According to the correlation of indicators obtained in the initial stage of logistic regression and the experience summary in daily business, three input indicators and two output indicator were finally selected. Therefore, the following five indicators were selected as the input and output indicators of the SBM method.

According to the above five indicators, the efficiency value of each customer was obtained by using the SBM method through MaxDEA software. DEA_score was added to the next dataset. DEA_score distribution diagrams are shown in Figure 1.

Table 1: Input and output indicators of the SBM method.

Indicators		Indicator description
Input indicator	Income per month	Monthly income amount of the borrowing customer
	M_final_score	Credit risk score of external credit institutions to the buyer
	External_debt	Amount of external liabilities of the borrower
Output indicator	Loan amount	Loan amount of the borrower
	Product period	Number of loan periods of the borrower

3.3. Logistic-SBM Modelling Process. Due to the wide and complex dimensions of the data used in this study and the large amount of data involved, the logistic DEA model consisted of a series of steps. The logistic DEA model selected the input and output values of the DEA model according to the initial index of the logistic model method. Then, MaxDEA software is used to calculate the efficiency value of each customer as a decision unit (DMU). As a new index, the efficiency value obtained by DEA would be used in the training and testing of the logistic model together with the initial relevant index. Finally, the model was used to test the default probability of loan customers, which verifies the effectiveness and accuracy of the model. It was helpful to analyze the contribution of DEA index to the accuracy of the logistic regression model.

3.3.1. Feature Binning. Through the observation of the collected datasets, it was found that many data types are inconsistent, in which many of them were character type. Because these character indicators may play a great role in the model, we used weight of evidence (WOE) to transform many character indicators into measurable numerical indicators. According to the chi-square value of each pair of adjacent intervals, the two intervals with the smallest value are combined. The formula used in this step is as follows:

$$x = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(A_{ij}-E_{ij}\right)^2}{E_{ij}},$$
$$E_{ij} = \frac{N_i \times C_i}{N}. \tag{4}$$

A_{ij} is the ith interval and the number of jth instances, E_{ij} is the desired frequency of A_{ij}, N is the total number of samples, N_i the number of samples in the ith group, and C_i is the proportion of the jth sample in the whole.

Feature information table is shown in Table 2. The continuous characteristic variable was discrete. Discrete feature states were often merged to reduce the number of states. It is convenient to transform all variables to similar scales. At the same time, some missing features will be brought into the model as an independent box. The reduction of extreme values and meaningless fluctuations in characteristics have an impact on the score and increase the stability and robustness of the model.

3.3.2. Correlation Coefficient. The correlation coefficient was obtained by calculating the correlation of each feature. The correlation coefficient formula is as follows:

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}}. \tag{5}$$

Among them, $\mathrm{Cov}(X, Y)$ is the covariance of X and Y; $\mathrm{Var}[X]$ and $\mathrm{Var}[Y]$ are the variance of X and Y, respectively.

If the absolute value of characteristic correlation coefficient was greater than 0.7, it was considered as a strong correlation feature. If there was strong correlation between features, some features can be deleted and one of them can be retained, as shown in Table 3. Delete the total debt ratio indicator.

3.3.3. IV. IV (Information Value) measures the amount of information about a variable. From the formula, it is equivalent to a weighted sum of the WOE values of the independent variables, in which the size of the value determines the influence of the independent variables on the target variables. The feature Information Value (IV) index can measure the concentration of the feature containing predictor variables. Weight of evidence (WOE) is a supervised coding method. The calculation formula is

$$\mathrm{WOE}_i = \log\left(\frac{G_1/G_{\mathrm{total}}}{B_1/B_{\mathrm{total}}}\right). \tag{6}$$

The IV is mainly used to code the input variables and evaluate the predictive ability. The value of characteristic variable IV indicates the predictive ability of the variable. The feature information degree of the remaining features was calculated, including the IV of the other features. After grouping, the formula for calculating the IV of each group is shown in Table 4.

According to the reference threshold of IV, the features with IV less than or equal to 0.02 are defined as nonpredictive features. Therefore, all features of this class were deleted. According to the characteristic IV shown in Table 5, "Marriage" and "Birth_month" features were deleted.

3.3.4. Random Forest Model. Random forest model is an integrated algorithm, which generates many trees and gets the result by voting or calculating the average. For grouped variables, cart Gini value is used as the evaluation standard. The steps of random forest model feature importance

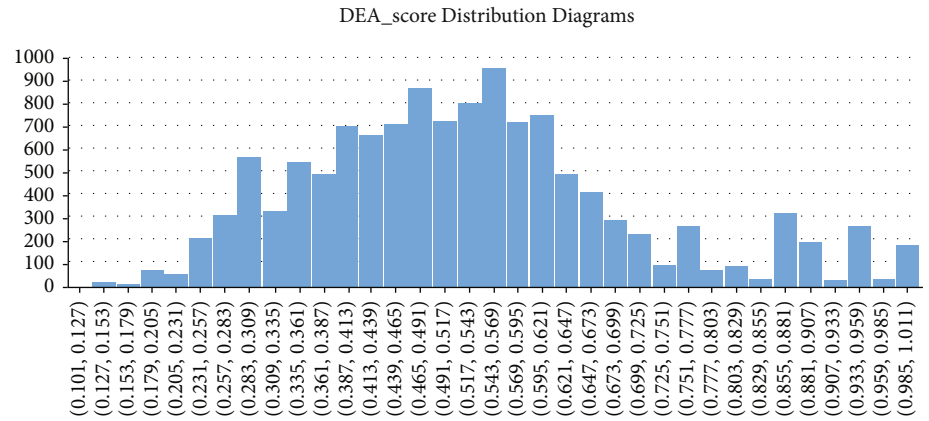

Figure 1: DEA_score distribution diagrams.

Table 2: Feature information table.

No.	Features	Feature interpretation	Class number
1	DEA_score	Efficiency score of borrowing	5
2	Education	Borrower's highest education	5
3	Marriage	Marital status of the borrower	4
4	Home type	Type of residence of the borrower	4
5	Company	Type of work unit of the borrower	5
6	Pay method	How the borrower pays wages	3
7	Job type	Job type of the borrower	4
8	Product name	Types of borrowers' lending products	4
9	Sales department	Which business department is responsible for the borrower's lending behavior	3
10	Bank	Ownership of bank card signed by the borrower	5
11	Family aware	Is the borrower aware of his borrowing behavior	3
12	Pro_id	The registered residence of a borrower	5
13	Birth month	Month of birth of the borrower	3
14	Birthday	Date of birth of the borrower	4
15	Inapv_edr	External debt ratio of borrowers	5
16	Inapv_idr	Internal debt ratio of the borrower	5
17	Inapv_tdr	Total debt ratio of the borrower	5
18	Age	Age of borrower	6
19	Entry date	Working days of the borrower	5

selection were as follows. The formula for calculating the Gini index is

$$\mathrm{GI}_m = \sum_{k=1}^{|k|} \sum_{k \neq k} p_{mk} p'_{mk} = 1 - \sum_{k=1}^{|k|} p_{mk}^2. \tag{7}$$

The meaning of each indicator in the formula is as follows: k means that there are k categories.

P_{mk} means the proportion of the category k in the node m.

The importance of the feature x-j at the node m is the Gini exponential change before and after the node m branch and is calculated as follows:

$$\mathrm{VIM}_{jm}^{(\mathrm{Gini})} = \mathrm{GI}_M - \mathrm{GI}_l - \mathrm{GI}_r. \tag{8}$$

Among them, GI_l and GI_r, respectively, represent the Gini index of the two new nodes after branching.

When the node where the feature$x_j$appears in the decision tree$i$is in the set$M$, the calculation formula of the importance of$x_j$in the$i$th tree is

$$\mathrm{VIM}_{ij}^{(\mathrm{Gini})} = \sum_{m \in M} \mathrm{VIM}_{jm}^{(\mathrm{Gini})}. \tag{9}$$

Assuming that there are n trees in the RF, then the importance of x_j in the nth tree is

$$\mathrm{VIM}_{j}^{(\mathrm{Gini})} = \sum_{i=1}^{n} \mathrm{VIM}_{ij}^{(\mathrm{Gini})}. \tag{10}$$

Table 3: Index correlation.

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	DEA_score	1.00	0.03	0.01	0.04	0.12	-0.16	-0.04	-0.29	0.00	0.08	-0.25	0.04	0.01	0.00	0.25	-0.12	0.14	0.02	0.07
2	Education	0.03	1.00	0.06	0.11	0.18	-0.16	-0.01	-0.04	-0.04	0.05	0.03	0.06	-0.01	0.01	0.10	-0.02	0.10	-0.10	0.10
3	Marriage	0.01	0.06	1.00	-0.04	0.03	-0.02	-0.03	0.02	-0.06	0.03	0.04	0.02	0.01	0.00	0.00	0.01	0.01	-0.37	-0.14
4	Home type	0.04	0.11	-0.04	1.00	0.20	-0.14	-0.05	-0.16	-0.05	0.01	0.02	0.21	-0.01	0.01	0.05	-0.04	0.04	0.08	0.17
5	Company	0.12	0.18	0.03	0.20	1.00	-0.41	-0.16	-0.30	-0.18	0.12	0.02	0.24	-0.02	0.02	0.23	-0.07	0.19	0.04	0.35
6	Pay_method	-0.16	-0.16	-0.02	-0.14	-0.41	1.00	0.11	0.29	0.25	-0.15	0.03	-0.19	-0.01	0.00	-0.32	0.09	-0.23	0.01	-0.26
7	Job type	-0.04	-0.01	-0.03	-0.05	-0.16	0.11	1.00	0.09	0.06	0.00	0.00	-0.03	0.00	-0.01	0.00	-0.04	-0.04	0.02	-0.05
8	Product name	-0.29	-0.04	0.02	-0.16	-0.30	0.29	0.09	1.00	0.05	-0.10	0.27	-0.13	-0.02	-0.02	-0.34	0.25	-0.17	-0.13	-0.18
9	Sales department	0.00	-0.04	-0.06	-0.05	-0.18	0.25	0.06	0.05	1.00	-0.06	-0.31	-0.33	0.00	0.01	-0.09	0.01	-0.11	0.03	-0.08
10	Bank	0.08	0.05	0.03	0.01	0.12	-0.15	0.00	-0.10	-0.06	1.00	-0.02	0.09	-0.01	0.00	0.18	-0.07	0.10	0.00	0.09
11	Family aware	-0.25	0.03	0.04	0.02	0.02	0.03	0.00	0.27	-0.31	-0.02	1.00	0.10	0.00	-0.03	-0.26	0.19	-0.05	-0.06	-0.04
12	Pro_id	0.04	0.06	0.02	0.21	0.24	-0.19	-0.03	-0.13	-0.33	0.09	0.10	1.00	0.00	0.01	0.09	-0.03	0.08	0.05	0.17
13	Birth month	0.01	-0.01	0.01	-0.01	-0.02	-0.01	0.00	-0.02	0.00	-0.01	0.00	0.00	1.00	0.02	0.01	0.00	0.00	-0.01	-0.01
14	Birthday	0.00	0.01	0.00	0.01	0.02	0.00	-0.01	-0.02	0.01	0.00	-0.03	0.01	0.02	1.00	0.01	0.00	0.00	0.02	0.00
15	Inapv_edr	0.25	0.10	0.00	0.05	0.23	-0.32	0.00	-0.34	-0.09	0.18	-0.26	0.09	0.01	0.01	1.00	-0.20	0.72	0.06	0.19
16	Inapv_idr	-0.12	-0.02	0.01	-0.04	-0.07	0.09	-0.04	0.25	0.01	-0.07	0.19	-0.03	0.00	0.00	-0.20	1.00	0.35	-0.09	-0.11
17	Inapv_tdr	0.14	0.10	0.01	0.04	0.19	-0.23	-0.04	-0.17	-0.11	0.10	-0.05	0.08	0.00	0.00	0.72	0.35	1.00	0.04	0.14
18	Age	0.02	-0.10	-0.37	0.08	0.04	0.01	0.02	-0.13	0.03	0.00	-0.06	0.05	-0.01	0.02	0.06	-0.09	0.04	1.00	0.33
19	Entry date	0.07	0.10	-0.14	0.17	0.35	-0.26	-0.05	-0.18	-0.08	0.09	-0.04	0.17	-0.01	0.00	0.19	-0.11	0.14	0.33	1.00

Table 4: Group IV calculation formula.

Group	WOE	IV
Group 1	$\log\left(\frac{G_1/G_{\text{total}}}{B_1/B_{\text{total}}}\right)$	$\left(\frac{G_1}{G_{\text{total}}}-\frac{B_1}{B_{\text{total}}}\right)\log\left(\frac{G_1/G_{\text{total}}}{B_1/B_{\text{total}}}\right)$
Group 2	$\log\left(\frac{G_2/G_{\text{total}}}{B_2/B_{\text{total}}}\right)$	$\left(\frac{G_2}{G_{\text{total}}}-\frac{B_2}{B_{\text{total}}}\right)\log\left(\frac{G_2/G_{\text{total}}}{B_2/B_{\text{total}}}\right)$
……	……	……
Group n	$\log\left(\frac{G_n/G_{\text{total}}}{B_n/B_{\text{total}}}\right)$	$\left(\frac{G_n}{G_{\text{total}}}-\frac{B_n}{B_{\text{total}}}\right)\log\left(\frac{G_n/G_{\text{total}}}{B_n/B_{\text{total}}}\right)$
Total		$\sum\left(\frac{G_i}{G_{\text{total}}}-\frac{B_i}{B_{\text{total}}}\right)\log\left(\frac{G_1/G_{\text{total}}}{B_1/B_{\text{total}}}\right)$

Table 5: IV of each feature.

Feature No.	Features	IV
1	DEA_score	0.104
2	Education	0.025
3	Marriage	0.011
4	Home type	0.117
5	Company	0.083
6	Pay method	0.079
7	Job type	0.067
8	Product name	0.796
9	Sales department	0.204
10	Bank	0.054
11	Family aware	0.319
12	Pro_id	0.067
13	Birth_month	0.004
14	Birthday	0.020
15	Inapv_edr	0.168
16	Inapv_idr	0.168
17	Age	0.149
18	Entry date	0.046

Finally, the importance scores obtained through normalization are processed. The formula is as follows:

$$\text{VIM}_j = \frac{\text{VIM}_j}{\sum_{i=1}^{c}\text{VIM}_i}. \tag{11}$$

The variable importance score is represented by VIM, and the Gini index is represented by GI. Assuming that there are*m*features*x*_1,*x*_2, and*x*_*m*, the Gini index score of each feature*x*_*i*is now calculated. Features are ranked from high to low according to their importance, and the top *n* features are selected.

Order of feature importance is shown in Table 6. Firstly, the feature variables in the random forest were sorted in descending order according to VI (variable importance). Then, the indexes with unimportant proportion were removed from the current feature variables to obtain a new feature set.

Table 6: Order of feature importance.

No.	Features	Importance	Cum_importance
1	Product name	0.204	0.204
2	Family aware	0.086	0.29
3	Age	0.066	0.356
4	Inapv_idr	0.065	0.420
5	Entry date	0.060	0.480
6	Birthday	0.060	0.540
7	Inapv_edr	0.058	0.599
8	DEA_score	0.056	0.655
9	Home type	0.054	0.709
10	Pro_id	0.047	0.756
11	Sales department	0.046	0.803
12	Job type	0.046	0.849
13	Education	0.044	0.893
14	Company	0.043	0.936
15	Bank	0.041	0.977
16	Pay_method	0.023	1.000

The result of feature importance was obtained by the random forest algorithm. The results would be retained three decimal places and sorted according to the importance from high to low. At the same time, the cumulative importance was calculated. According to the feature importance ranking, it was obvious that the feature of "Pay_method" showed the low importance.

3.3.5. Logistic-SBM Model Variables. In the application of P2P network credit loan, the logistic model was adopted due to its high discrimination ability in the field of default loan customer identification. The logistic formula is

$$\begin{aligned} E(p) &= f\left(\beta_0+\sum\beta_i x_i\right), \\ f(x) &= \frac{\exp(x)}{1+\exp(x)}. \end{aligned} \tag{12}$$

The overdue status of a group of applicants in the performance period is $\{y_1, y_2, \cdots, y_n\}$ and $y_i \in \{0, 1\}$. The likelihood function and log likelihood function are

$$\begin{aligned} \mathrm{L}(p) &= \prod P(Y=y_i) = \prod p^{y_i}(1-p)^{1-y_i}, \\ l(p) &= \log(\mathrm{L}(p)) = \log\left(\prod P(Y=y_i)\right) = \sum(y_i \log(p) \\ &\quad + (1-y_i)\log(1-p)) = \sum\left(y_i\left(\beta_0+\sum\beta_i x_{ij}\right)\right. \\ &\quad \left.-\log\left(1+\exp\left(\beta_0+\sum\beta_i x_{ij}\right)\right)\right). \end{aligned} \tag{13}$$

The parameter estimation formula is as follows:

$$\widehat{p} = \operatorname{argmax} l(p),$$

$$\widehat{p} = \frac{\sum y_i}{n},$$

$$\begin{aligned} l(p) &= \sum (y_i \log (p) + (1 - y_i) \log (1 - p)) \\ &= \sum \left(y_i \left(\beta_0 + \sum \beta_i x_{ij} \right) - \log \left(1 + \exp \left(\beta_0 + \sum \beta_i x_{ij} \right) \right) \right). \end{aligned} \tag{14}$$

The parameter estimation formula is as follows:

$$\frac{\partial \mathrm{l}}{\partial \beta_q} = \sum \left(y_i - \frac{1}{\exp \left(-\beta_0 - \sum \beta_i x_{ij} \right)} \right) x_{iq}. \tag{15}$$

Estimate the β_q by the gradient descent method; the formula is as follows:

$$\begin{aligned} \beta_q^{r+1} &= \beta_q^r - h\delta, \\ \delta &= \frac{\partial l}{\partial \beta_q} \Big| \beta_q = \beta_q^r. \end{aligned} \tag{16}$$

It is very important to select variables from the dataset. Considering the correlation coefficient, validity, and importance of index data, 15 variables were selected in the final logistic-SBM model for empirical study. Logistic-SBM model variables are shown in Table 7.

Table 7: Logistic-SBM model variables.

Feature No.	Features
1	DEA_score
2	Education
3	Home type
4	Company
5	Job type
6	Product name
7	Sales department
8	Bank
9	Family aware
10	Pro_id
11	Birthday
12	Inapv_edr
13	Inapv_idr
14	Age
15	Entry date

4. Result Analysis and Inspection

Model verification is used to measure the predictive ability of the developed model, including internal and external tests. The internal test is the comparison between the prediction situation of the test set in the sample and the actual situation. The external test is the comparison between the prediction situation and the actual situation of the dataset except the model after passing the model. The primary goal of the developed model is to distinguish whether the borrower is in default. Besides, the accuracy of model prediction, confusion matrix analysis, and the Kolmogorov-Smirnov test can all be used as criteria for judging the quality of this model.

4.1. Confusion Matrix Analysis. Accuracy is an important concept and indicator in model evaluation. The performance of the resulting classier can then be evaluated in terms of the recall (or sensitivity) and precision of the classier on an evaluation dataset. Recall and precision are defined in terms of the number of true positives (TP), misses (FN), and false alarms (FP) of the classier (cf. Table 8).

In Table 7, the first line expresses prediction results from the prediction model; the first column expresses the actual results in the original data. True positive (TP) expresses the amount that the positive samples are correctly classified as positive; false negative (FN) expresses the amount that the positive samples are misclassified as negative; false positive (FP) expresses the amount that the negative samples are misclassified as positive; true negative (TN) expresses the amount that the negative samples are correctly classified as negative. As the common evaluation measures, the accuracy-specific expressions are shown as follows:

$$A(\text{accuracy}) = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}. \tag{17}$$

The borrower results predicted by the model were compared with the marked good and bad borrowers. From this result, the model has a strong predictive ability. 77.49% of borrowers were accurately predicted, and only 22.51% of borrowers were incorrectly predicted. Among them, the first quadrant is the number of borrowers that the model predicts to be nondefaulting and actually not defaulting. In the second and third quadrants, the number of errors is predicted. The fourth quadrant indicates that the model predicts the number of defaults and actual defaults. The accuracy of the model was that the ratio of the number of accurate predictions to the total number was 77.49%, in which the accuracy rate was high.

4.2. AUC-ROC Curve Observation. The AUC-ROC curve is a performance measurement for classification problems under various threshold settings. ROC (receiver operating characteristic curve) is a probability curve, and AUC (area under the curve) represents the degree or measure of separability which represents how many models can distinguish categories. The higher the AUC, the better the model predicts 0 as 0 and 1 as 1. The ROC curve of the logistic-DEA model is shown in Figure 2.

4.3. K-S Test. The KS indicator measures the largest gap between the cumulative distribution of responding

Table 8: Confusion matrix for binary classification.

		Prediction positive 1	Prediction negative 0	Total N
Actually positive	1	True positives (TP)	False positives (FN)	N-pos
Actually negative	0	False negatives (FP)	True negatives (TN)	N-neg
Total M		M-pos	M-neg	

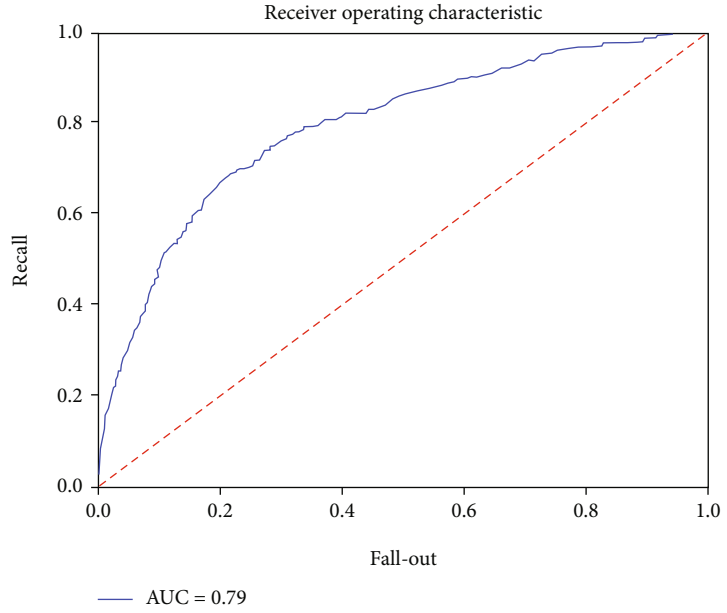

Figure 2: AUC-ROC curve.

customers and nonresponding customers. The calculation formula was as follows:

$$\mathrm{KS} = \mathrm{MAX}(\mathrm{ABS}(\mathrm{CPD}(S_i) - \mathrm{CPG}(S_i))). \tag{18}$$

$\mathrm{CPD}(S_i)$ is the proportion distribution of accumulated good customers, $\mathrm{CPG}(S_i)$ is the proportion distribution of accumulated bad customers.

Firstly, the scores of samples were ranked from large to small and then, the cumulative proportion of good and bad samples in each quantize interval was calculated. The larger the distance between the two, the higher the KS value, indicating that the model area has the ability to distinguish good and bad customers. In the actual business, if the KS value is less than 20%, the accuracy of the model is poor. If the KS value is between 20% and 30%, it means that the model discrimination effect is general. If the KS value is between 30% and 60%, the model is very effective.

The KS value was obtained by the logistic-SBM model, as shown in Figure 3. The KS value of the logistic-SBM model is 33.3%, indicating the good prediction effect and the better effect of distinguishing default customers of the model.

4.4. Comparison of Model Evaluation. Precision, specificity, and recall are important concepts and indicators in model evaluation too. As the common evaluation measures, sensitivity, specificity, G-Measure, and F-Measure are used to make the evaluation. F-Measure is also called F-Score. F-Measure is the weighted harmonic average of precision (P) and recall (R). It is an evaluation standard of the model and is often used to evaluate the quality of the classification model. The F-Measure function synthesizes the results of P and R when the parameter $\alpha = 1$; the weight of P and R is the same. When F-Measure is higher, the model is more effective. Their specific expressions are shown as follows:

$$R(\mathrm{recall}) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}},$$

$$S(\mathrm{specificity}) = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}},$$

$$P(\mathrm{precision}) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}},$$

$$G\text{-Measure} = \sqrt{\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}} = \sqrt{RS},$$

$$F\text{-Measure} = F_\alpha = \frac{(1+\alpha^2)(\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})\mathrm{TP}+\mathrm{FN}) \times (\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})\mathrm{TP}+\mathrm{FP})}{\alpha^2(\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})\mathrm{TP}+\mathrm{FN}) + (\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})\mathrm{TP}+\mathrm{FP})} = \frac{(1+\alpha^2)PR}{\alpha^2 P + R},$$

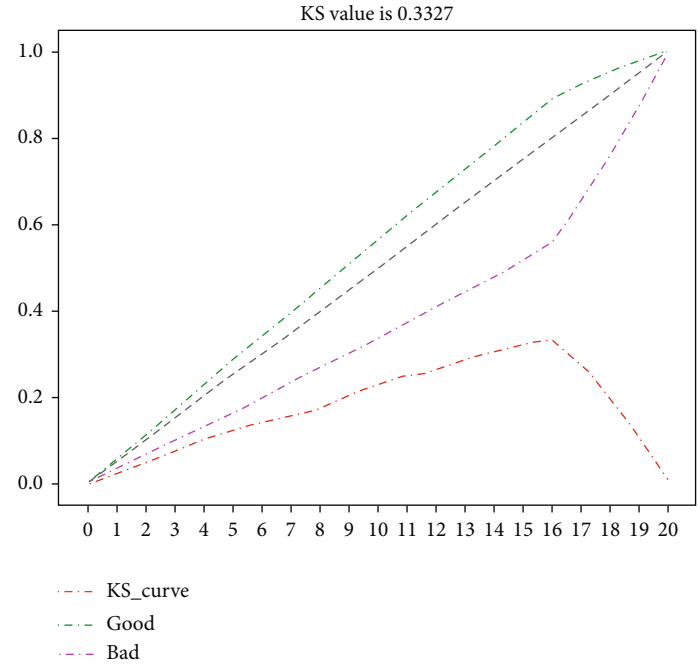

Figure 3: K-S value.

Table 9: Performance comparison of two models.

	Logistic-SBM	Logistic
Accuracy (%)	77.49	76.88
Recall (sensitivity) (%)	92.18	91.7
Precision (%)	79.87	79.54
Specificity (%)	38.73	37.78
G-Measure (%)	59.75	58.86
F-Measure (%)	85.59	85.19

$$F_1 = \frac{(\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})\mathrm{TP}+\mathrm{FN}) \times (\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})\mathrm{TP}+\mathrm{FP})}{(\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})\mathrm{TP}+\mathrm{FN}) + (\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})\mathrm{TP}+\mathrm{FP})} = \frac{PR}{P+R}. \quad (19)$$

We compare the mean values of corresponding evaluation measures of two models. The performance comparison of two models is shown in Table 9.

The relationship of two models can fully explain that the logistic-SBM model presented by this article has the optimal performance relative to the logistic model. The higher the value of related evaluation indicators, the better the effect of the model. Simulation results show that the logistic-SBM is more suitable for credit risk evaluation than the popularly used logistic with consideration of related evaluation indicators. According to the above research results, it can be known that using data envelopment analysis to preprocess the data and increase the efficiency value in the logistic regression model can improve the accuracy of the model.

5. Concluding Remarks

With the rapid development of the Internet, P2P has been applied in various fields [27]. At present, the risk management of borrowers in the P2P network lending platform mainly includes the following: first, the basic information authentication of borrowers. Mine their identity information and credit level from many aspects, and rate the borrowers. Feature variables are extracted from the basic information to determine the characteristics of credit management. The second is the combination of credit line management and credit risk. The loan limit of the borrower corresponds to the corresponding credit risk level.

Credit risk has four main characteristics: asymmetry, accumulation, unsystematic, and endogenous. The good operation of a platform requires strict audit of borrowers. Only through high-quality borrowers to minimize the risk of P2P network credit transactions can the P2P platform maintain stable operation. The grade assigned by the P2P lending site is the most predictive factor of default, but the accuracy of the model is improved by adding other information, especially the borrower's debt level [28]. The results suggest that borrower's social information can be used not only for credit screening but also for default reduction and debt collection [29].

Relevant suggestions have been put forward, which provide reference for the credit management of the P2P network lending industry in China. Regulatory authorities and the platform itself should take some measures to control the credit risk of the P2P Internet lending industry. The specific recommendations were as follows: (1) improve the social credit investigation system, and realize information sharing; (2) improve and implementation of policies; and

(3) undertake social responsibility, and actively develop through innovation.

Data Availability

This study collected partial loan records from an inclusive finance platform in China from 2014 to 2018.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] K. E. Davis and A. Gelpern, "Peer-to-peer financing for development: regulating the intermediaries," *NYUJ Int'l L. & Pol.*, vol. 42, p. 1209, 2009.

[2] P. Slattery, "Square pegs in a round hole: SEC regulation of online peer-to-peer lending and the CFPB alternative," *Yale J. on Reg.*, vol. 30, p. 233, 2013.

[3] B. Budiharto, S. N. Lestari, and G. Hartanto, "The legal protection of lenders in peer to peer lending system," *Law Reform*, vol. 15, no. 2, pp. 275–289, 2019.

[4] Q. Wang, X. Xiong, and Z. Zheng, "Platform characteristics and online peer-to-peer lending: evidence from China," *Finance Research Letters*, vol. 38, article 101511, 2021.

[5] X. Liu, "A visualization analysis on researches of internet finance credit risk in coastal area," *Journal of Coastal Research*, vol. 103, no. sp1, pp. 85–89, 2020.

[6] X. Fang, B. Wang, L. Liu, and Y. Song, "Heterogeneous traders, the leverage effect and volatility of the Chinese P2P market," *Journal of Management Science and Engineering*, vol. 3, no. 1, pp. 39–57, 2018.

[7] W. Zhang, Y. Zhao, P. Wang, and D. Shen, "Investor sentiment and the return rate of P2P lending platform," *Asia-Pacific Financial Markets*, vol. 27, no. 1, pp. 97–113, 2020.

[8] L. Ma, Y. Li, D. Li, H. Li, Y. Wang, and C. Ren, "Risk identification and decision making for P2P companies: an empirical study in the Bohai coast regions," *Journal of Coastal Research*, vol. 106, no. sp1, pp. 191–196, 2020.

[9] Z. Abdul Halim, J. How, P. Verhoeven, and M. K. Hassan, "Asymmetric information and securitization design in Islamic capital markets," *Pacific-Basin Finance Journal*, vol. 62, p. 101189, 2020.

[10] X. Lv, L. Zhou, and X. Guo, "Research on P2P network loan risk evaluation based on generalized DEA model and R-type clustering analysis under the background of big data," *Journal of Financial Risk Management*, vol. 6, no. 2, pp. 163–190, 2017.

[11] Y. Guo, W. Zhou, C. Luo, C. Liu, and H. Xiong, "Instance-based credit risk assessment for investment decisions in P2P lending," *European Journal of Operational Research*, vol. 249, no. 2, pp. 417–426, 2016.

[12] M. Herzenstein, U. M. Dholakia, and R. L. Andrews, "Strategic herding behavior in peer-to-peer loan auctions," *Journal of Interactive Marketing*, vol. 25, no. 1, pp. 27–36, 2011.

[13] J. H. Zeng and S. Yang, "Herding behavior of lenders in P2P lending markets and its rational test: evidence from PaiPaiDai market," *Modern Finance and Economics (Journal of Tianjin University of Finance and Economics)*, p. 7, 2014.

[14] G. A. Akerlof, "The Market for Lemons: Quality Uncertainty and the Market Mechanism," *Quarterly Journal of Economics*, vol. 84, no. 3, pp. 488–500, 1970.

[15] S. C. Berger and F. Gleisner, "Emergence of financial intermediaries in electronic markets: the case of online P2P lending," *BuR Business Research Journal*, vol. 2, no. 1, pp. 39–65, 2009.

[16] G. N. Weiss, K. Pelger, and A. Horsch, *Mitigating adverse selection in P2P lending–empirical evidence from prosper. com*Available at SSRN 1650774, 2010.

[17] M. A. Razi, J. M. Tarn, and F. A. Siddiqui, "Exploring the failure and success of DotComs," *Information Management & Computer Security*, vol. 12, no. 3, pp. 228–244, 2004.

[18] G. D. Bruton and Y. Rubanik, "Resources of the firm, Russian high-technology startups, and firm growth," *Journal of Business Venturing*, vol. 17, no. 6, pp. 553–576, 2002.

[19] Y. Honjo, "Business failure of new firms: an empirical analysis using a multiplicative hazards model," *International Journal of Industrial Organization*, vol. 18, no. 4, pp. 557–574, 2000.

[20] R. Sullivan, "Entrepreneurial learning and mentoring," *International Journal of Entrepreneurial Behavior & Research*, vol. 6, no. 3, pp. 160–175, 2000.

[21] M. Lan, *Online P2P lending industry: an international analysis*, University of Nottingham, 2019.

[22] X. Chen, X. Hu, and S. Ben, "How do reputation, structure design and FinTech ecosystem affect the net cash inflow of P2P lending platforms? Evidence from China," *Electronic Commerce Research*, vol. 21, no. 4, pp. 1055–1082, 2021.

[23] M. H. Akhtar, I. S. Chaudhry, M. R. Sheikh, and A. Shahzadi, "Business model, risk and financial stability of banks: a multi-country analysis," *Pakistan Journal of Social Sciences (PJSS)*, vol. 40, no. 1, pp. 401–414, 2020.

[24] N. Barasinska and D. Schäfer, "Is Crowdfunding Different? Evidence on the Relation between Gender and Funding Success from a German Peer-to-Peer Lending Platform," *German Economic Review*, vol. 15, no. 4, pp. 436–452, 2014.

[25] R. Liu, N. Chen, and Y. Li, *"The Herd Behavior on Peer-to-Peer Online Lending Markets: Evidence from China," Discrete Dynamics in Nature and Society, Vol*, vol. 2021, 2021.

[26] J. D. Velimirovic and A. Janjic, "Risk assessment of circuit breakers using influence diagrams with interval probabilities," *Symmetry*, vol. 13, no. 5, p. 737, 2021.

[27] H. Wang, K. Fan, H. Li, and Y. Yang, "A dynamic and verifiable multi-keyword ranked search scheme in the P2P networking environment," *Peer-to-Peer Networking and Applications*, vol. 13, no. 6, pp. 2342–2355, 2020.

[28] C. Serrano-Cinca, B. Gutierrez-Nieto, and L. Lopez-Palacios, "Determinants of default in P2P lending," *PLoS One*, vol. 10, no. 10, article e0139427, 2015.

[29] R. Ge, J. Feng, B. Gu, and P. Zhang, "Predicting and deterring default with social media information in peer-to-peer lending," *Journal of Management Information Systems*, vol. 34, no. 2, pp. 401–424, 2017.