Research on Factor Identification and Contribution of Bank Nonperforming Loans Based on Lightgbm Algorithm

Siqin Shu[1, #], Yidi Sun[2, #] and Yu Zhou[3, *, #]

[1]Department of International Business, Tianjin Foreign Studies University, Tianjin, China, 300270

[2]School of Statistics and Data Science, Xinjiang University of Finance and Economics, Xinjiang, China, 830012

[3]School of Management Science and Engineering, Central University of Finance and Economics, Beijing, China, 100081

*Corresponding author: 2020311565@email.cufe.edu.cn

[#]These authors contributed equally.

Abstract. In recent years, financial risks and crises have occurred frequently, which has had a far-reaching impact on the world economic pattern. Among them, the bank non-performing loan ratio (NPL) is usually used as a barometer of financial risks, which is used to identify the factors of bank non-performing loans. Combined with the current macro and micro economic data in China, this paper selects 17 indicators from four representative commercial banks from 2010 to 2021 as samples and uses the Light Gradient Boosting Machine (LightGBM) algorithm to empirically analyze the influencing factors of commercial banks' non-performing loan ratio, to quantify the sensitivity of each factor to the non-performing loan ratio. The results show that the influence of the bank's internal finance (0.7874) on the bank's non-performing loan ratio is much greater than the bank's external macro factors (0.2126), and the secondary index with the highest weight in the internal finance category is the medium and long-term loan interest rate (0.3029). Banks should pay attention to endogenous variables, especially the changes in medium and long-term indicators, and timely predict and reduce the risks caused by non-performing loans.

Keywords: LightGBM; Nonperforming loan ratio; Financial risks; bank; China.

1. Introduction

As an important part of the financial industry, the continuous and stable operation of commercial banks is inseparable from the development of the entire financial industry. Therefore, financial risk has always been the focus of the global financial community. In recent years, financial crises have erupted in many countries around the world. One of the direct causes is the decline in the quality of credit assets in the banking system, which increased the attention on banks' nonperforming loans. A large number of exogenous non-performing loans will not only provide essential financial support for China's gradual reform but also increase the systematic risk of the banking industry. At present, there are not a few studies on non-performing loans of commercial banks, and many innovative research models have also been born.

Gu and Zhang [1] established a decision-making model based on macroeconomic theory and concluded that the exogenous non-performing loans of commercial banks will increase the operating burden of commercial banks, reduce the management level of commercial banks, and drive the endogenous non-performing loan rate to rise. Svetozar et al. [2] used the static panel model in the research process and believed that the increase in GDP was negatively correlated with the rise in NPL ratio, while the foreign currency loan ratio and exchange rate level were positively correlated with it. Wang et al. [3] used the dynamic GMM regression method to discuss and showed that financial repression led to the rise of bank non-performing loan ratio, and direct government intervention can weaken the negative impact of financial repression on bank non-performing loan ratio. Wu and Wang [4] combined mixed OLS regression and fixed effect regression methods to conclude that the credit environment has a significant impact on the non-performing loan ratio *(NPL)* of commercial banks. Further analysis found that it is significantly affected by the level of economic development and the

degree of government intervention. Xiao et al. [5] through the ANP network learning model and grounded theory, put forward that the lack of internal control management has a significantly greater impact on bank non-performing loans than external defensive factors, and the bank loan credit process and professional ethics of business personnel are the key issues.

However, there are some deficiencies in previous studies: for example, at this stage, there are certain differences between the research conclusions of exploring the influencing factors of non-performing loans; And when the amount of data is large, whether it is traditional regression analysis and macroeconomic models, or machine learning methods such as an/gbdt, it is easy to cause problems such as time-consuming, low accuracy, over fitting and so on, which is also the possible reason why research conclusions are often inconsistent.

Therefore, this paper selects the *LightGBM* algorithm, which is more advanced in the field of deep machine learning in recent years, and integrates the previous relevant literature to conduct factor analysis on the correlation between the macro environment, the external situation of the bank, and the internal finance of the bank, three primary indicators and 17 secondary indicators and the bank's non-performing loan ratio. The research of this paper is conducive to the expansion of methods for commercial banks to prevent the risk of non-performing loans and to explore the mechanism of various factors inside and outside the market on the non-performing loan rate.

2. Construction of index system for influencing factors of non-performing loan ratio

2.1 Non-performing loan ratio

The calculation formula of non-performing loan ratio *(NPL)* is: *NPL= non-performing loan / total loan*. Its application scope includes banks, small and medium-sized enterprises, financial institutions, etc. this time, the non-performing loan of banks is taken as the research object.

From the perspective of banks, when evaluating the quality of bank loans, the non-performing loan ratio will divide loans into five categories according to the risk basis: normal, concerned, secondary, suspicious, and loss. It is an important indicator to measure the quality of bank assets, a reflection of the level of bank credit management, and an important factor to measure the stability of the financial system. The increase in non-performing loans will have an adverse impact on the profitability, liquidity, and solvency of banks.

In recent ten years, many scholars have studied the influencing factors and emerging problems of bank non-performing loans; However, the existing research methods are judged by a single regression model, and the limitations of the algorithm type are likely to cause the demonstration results to shift in the same direction. Therefore, this paper introduces the *LightGBM* algorithm, a mapping algorithm with strong nonlinearity, to further explore the impact mechanism of various factors on non-performing loans of commercial banks.

2.2 Explanatory variables

Referring to the literature around the world in recent years, in the previous research on the influencing factors of non-performing loans, scholars generally believe that the formation of non-performing loans is more significantly affected by banks' behavior and other factors than macro factors. Gross domestic product growth (GDP), unemployment rate (unem), and inflation rate (INF) are considered to be the main macroeconomic factors affecting non-performing loans; The provision coverage ratio, capital adequacy ratio, and bank asset size are considered to be the main micro factors affecting non-performing loans.

Based on the above research results, this paper divides the macro factors into the macro environment and bank external conditions selecting the bank non-performing loan ratio as the explanatory variable and selects the macro environment, bank external conditions, and bank internal finance as three primary indicators and 17 secondary indicators as the explanatory variables to explore

the impact of different factors on the non-performing loan ratio. Among them, GDP growth rate (GDP), inflation rate (CPI), total retail sales of social consumer goods (SR), and money supply growth rate (M2) are selected as macroeconomic indicators that affect the rate of bank non-performing loans; Select bank size (size), number of banks (Num), commercial credit environment index (CEI) and government financial intervention (GFI) as the external situation indicators that affect the bank's non-performing loan ratio; Bank size (size), number of banks (Num), commercial credit environment index (CEI), government financial intervention (GFI), bank asset liability ratio (DAR), capital adequacy ratio (car), annual interest rate of bank loans (LR), deposit loan ratio (LDR), non-performing loan provision coverage (PCR), cost income ratio (CI), total social fixed asset investment (Tifa), loan / total liability (PB) and loan concentration (LC) are used as external situation indicators that affect the bank's non-performing loan ratio. To sum up, the index system of influencing factors of non-performing loan ratio constructed in this paper is shown in Table 1:

Table 1. Summary of index selection

Primary index	Secondary indicators	Index unit	Reference
Macro Environment U1	GDP growth rate (GDP)	%	[6]
	Inflation rate (CPI)	%	
	Total retail sales of social consumer goods (SR)	100 million yuan	
	The growth rate of money supply (M2)	%	
Bank External Situation U2	Bank size (SIZE)	100 million yuan	[6] [7] [8]
	Number of banks (NUM)	Number	
	Business credit environment index (CEI)	/	
	Government financial intervention (GFI)	/	
Bank Internal Situation U3	Bank asset-liability ratio (DAR)	%	[6] [7]
	Capital adequacy ratio (CAR)	%	
	The annual interest rate of a bank loan (LR)	%	
	Deposit loan ratio(LDR)	/	
	Nonperforming loan provision coverage(PCR)	%	
	Cost income ratio(CI)	%	
	Total investment in social fixed assets(TIFA)	100 million yuan	
	Loans / total liabilities(PB)	/	
	Loan concentration(LC)	%	

3. LightGBM Algorithm

The Light Gradient Boosting Machine (LightGBM), first proposed by the Microsoft team in 2017, is an improved optimization algorithm for *Gradient Boosting Decision Tree (GBDT)*.GBDT, a popular machine learning algorithm, is effectively implemented in *eXtreme Gradient Boosting (XGoost)* and *Preimplantation Genetic Testing (pGBRT)*. However, when the feature dimension is large and the data is large, GBDT is difficult to achieve high efficiency. The proposed LightGBM can solve GBDT's problems in the mass data processing.LightGBM has the advantages of fast training speed, less memory, high accuracy and not being easy to over-fit.LightGBM is a decision tree ensemble model, which belongs to Boosting algorithm. In principle, it is similar to XGBoost. In each iteration, the negative gradient of the loss function is used as the residual approximation of the current decision tree to fit the new decision tree.

The main improvements of LightGBM on GBDT include decision tree algorithm based on Histogram and Leafwise algorithm with depth limitation, *Gradient-based One-side Sampling (GOSS)*, etc. The histogram algorithm is to discretize the continuous floating-point eigenvalues into K integers and construct a histogram of width K using these data in the training process. When traversing the data, the discretized value is used as the index to accumulate statistics in the histogram. After traversing the data, the optimal segmentation point is found according to the discrete value of the histogram. The cost of establishing the histogram is O (data×feature), and the cost of finding the segmentation point is O (bin×feature). Since the bin is typically much smaller than Data, building the histogram will reduce computational complexity. The schematic diagram is shown in Figure 1.

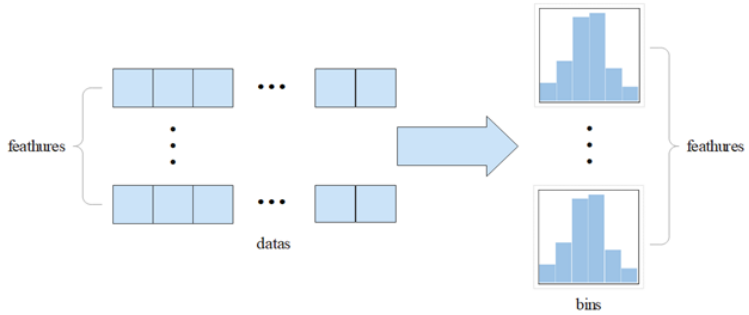

Fig. 1 Histogram Algorithm

Leaf-wise algorithm with depth limitation is to find the Leaf with the maximum splitting gain from all the current leaves to split each time, and so on. At the same time, the depth of the tree is controlled and the amount of data per leaf is limited to prevent over-fitting. The schematic diagram is shown in Figure 2.

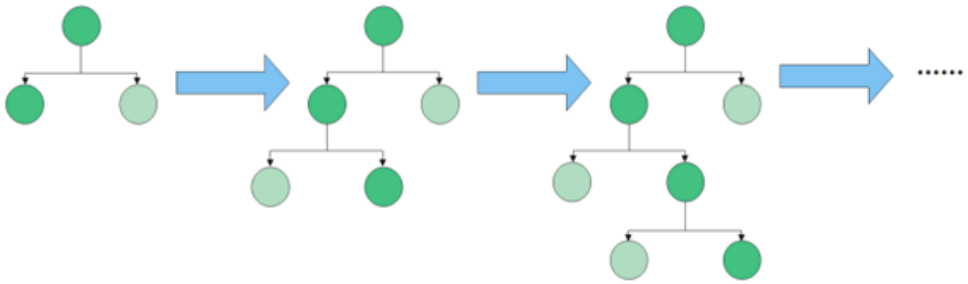

Fig. 2 Leaf-wise tree growth

The unilateral gradient sampling algorithm (GOSS) reduces the number of data and the complexity of computation by reserving large gradient data and discarding small gradient data. When calculating the information gain, sort in descending order according to the absolute value of the data gradient, select the $a*100\%$ data with the largest absolute value, then randomly select $b*100\%$ data in the remaining data, and then enlarge the randomly selected small gradient data by $\frac{1-a}{b}$ times, so as not to change the distribution of the original data too much and affect the accuracy of the training model.

The gradient lifting decision tree in the LightGBM algorithm is obtained through several iterations based on a large number of training data. During each iteration, the negative gradient of the loss function is used as the residual approximation value of the current decision tree to fit the new decision tree and add it to the tree group that has been iterated before. When the marginal increment of the accuracy of the tree is less than a fixed threshold, the iteration stops, and the LightGBM model composed of M decision trees is the output:

$$F(x_i)=\sum_{n=1}^{M} f_n(x_i) \tag{1}$$

Where x_i is the eigenvector of the input; $f_n(x_i)$ is the NTH decision tree of iteration.

4. Setting of inspection indicators

Checking the algorithm is helpful to measure the gap and connection between multiple algorithms, evaluate the robustness and generalization performance of each algorithm, find the model with the best training effect and the highest accuracy, and finding the optimal algorithm for expressing data. In addition, the model testing can also evaluate the merits of a single algorithm to ensure that the algorithm used has high precision while avoiding over-fitting and is only effective for training data, which can ensure that the results of the model output have more use-value. Therefore, root means square error (RMSE) and absolute coefficient (R^2) [9] were selected as the algorithm test indexes in this paper.

$$RMSE = \sqrt{\frac{1}{R}\sum_{k=1}^{R}(y_{k\_real} - y_{k\_predicted})^2} \tag{2}$$

$$R^2 = 1 - \frac{\sum_{k=1}^{R}(y_{k\_real} - y_{k\_predicted})^2}{\sum_{k=1}^{R}(y_{k\_real} - y_{k\_mean})^2} \tag{3}$$

Where R is the size of test sample data, y_{k_real} is the real value of the KTH sample data point, $y_{k\_predicted}$ is the predicted value of the KTH sample data point, and y_{k_mean} is the average value of the KTH sample data point.

R^2 is a relative measure, which compares the predicted value with the mean value to evaluate the merits of the algorithm. Its value range is [0, 1]. Generally, the closer R^2 is to 1, the smaller the error between the predicted value and the actual value, namely, the better the fitting degree of the selected model to the data.

RSME is the square root of the mean square difference between the predicted value and the actual value, and its value range is [0, +∞). RSME reflects the deviation degree between the predicted value and the real value. The smaller RMSE is, the smaller the deviation between the predicted value and the real value is, that is, the model has higher accuracy and better effect.

5. Empirical Analysis

5.1 Research object selection and data collection

The nonperforming loan ratio of Chinese commercial banks is at a relatively high level. The quality of credit assets decreases and the risk increases. There are certain differences in the nonperforming loan ratios of various types of banks. City commercial banks have a high rate of nonperforming loans, but they are small in size and have regional differences. The amount of nonperforming loans of joint-stock commercial banks is relatively small, and there has been a downward trend in the past two years. The number of nonperforming loans of state-owned commercial banks are huge, exceeding the sum of the two. In addition, the state-controlled holding has more diverse management methods and means, and the formulation and implementation of policies are more credible. Therefore, the nonperforming loan ratios of four banks including Bank of China, Industrial and Commercial Bank of China, China Construction Bank, and Agricultural Bank of China are selected as the main research objects. From

the WIND database, CEIE, China Banking Regulatory Commission, and other websites, we collected relevant data from 2010 to 2021, and made descriptive statistics on the data, as shown in Table 2.

Table 2. Data Descriptive Statistics

Name	Max	Min	*Mean*	*Middle*
NPL	2.39	0.85	1.39	1.43
GDP	8.10	2.20	6.27	6.85
CPI	2.90	0.92	1.92	2.01
SR	440823.00	300930.80	374016.09	377783.10
M2	12.32	8.28	9.92	9.32
GFI	0.26	0.22	0.24	0.24
DAR	94.76	90.69	92.51	92.33
CAR	18.02	11.59	14.59	14.43
LDR	87.47	55.77	71.88	71.50
PCR	367.04	136.69	216.85	208.37
CI	38.59	22.30	29.32	28.57
LC	7.94	5.29	6.31	6.17
L/L	0.68	0.51	0.59	0.59
SLR	6.35	4.35	5.08	4.63
LR	6.43	4.75	5.36	5.03
TIFA	552884.00	218834.00	402174.33	420146.00

5.2 LightGBM Algorithm

Some scholars such as XIE [10] have compared the prediction accuracy of algorithms Xgboost and LightGBM. Use software such as SPSS and Python to set γ, the maximum growth depth of the tree, the minimum number of samples (weight) of leaves, and the number of subtrees. A larger number of subtrees will make the model more stable. Since there are missing values in individual samples in the data, the minimum sample weight value of the leaf node is quoted here to limit the minimum value of the sum of the sample weights. If it is less than this value, it will be pruned together with the sibling nodes. This model adopts the ten-fold cross-validation method, and the data set is randomly divided into ten parts, of which the training set accounts for 7 parts, and the remaining 3 parts are used as the test set. The training results are shown in Table 3.

Table 3. Model Training Results

Weight $(a=1,2,3,6)$	Estimators $(b=6,9,12)$	R2	RMSE
1	6	0.1952	0.2731
1	9	0.4471	0.2256
1	12	0.4822	0.2155
2	6	0.1186	0.2813
2	9	0.4081	0.2374
2	12	0.4765	0.2211
3	6	0.1117	0.2811
3	9	0.4228	0.2349
3	12	0.4399	0.2276
6	6	0.2095	0.2699
6	9	0.4415	0.2268
6	12	0.4629	0.2205

When γ=0.1, the depth of the tree is 10, the samples are equally weighted, and the number of subtrees is 12, the accuracy is the best at this time R^2=0.4822. When the sample weight is greater than

1, the accuracy diverges, so we select the results of equal weights and 12 iterations for analysis. At this time, the overall training set of the model R^2=0.9067, RMSE=0.1029; the test set results R^2=0.8481, RMSE=0.1007, the overall fitting degree is high.

5.3 Model Results Analysis

Based on the LightGBM algorithm constructed above, we found by adjusting the parameters that when each sample has the same weight and the number of subtrees is 12, the accuracy is the best, and the overall model fits better. The importance of the indicators of the results at this time is analyzed, and then the important factors that affect the bank's non-performing loan ratio are found. The specific results are shown in Figures 3 and 4. Among them, the inflation rate, total retail sales of consumer goods, the growth rate of money supply, total social fixed asset investment, loans/total liabilities, and loan concentration indicators are relatively weak characteristics. get its importance to 0.

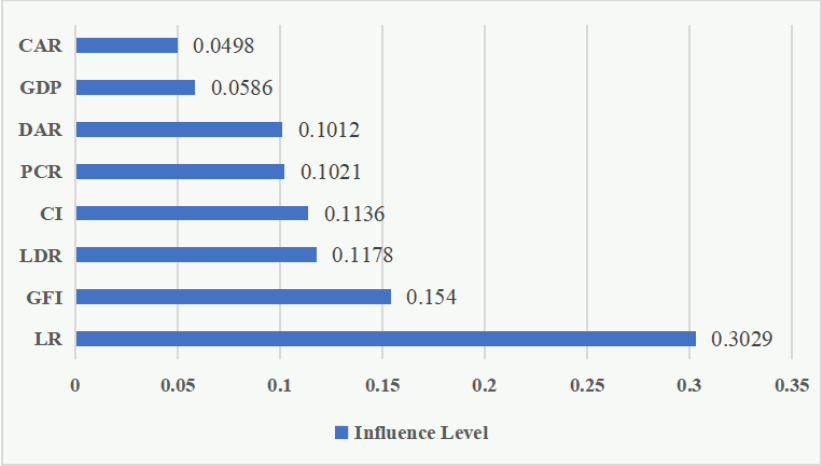

Fig. 3 Factor Importance Ranking

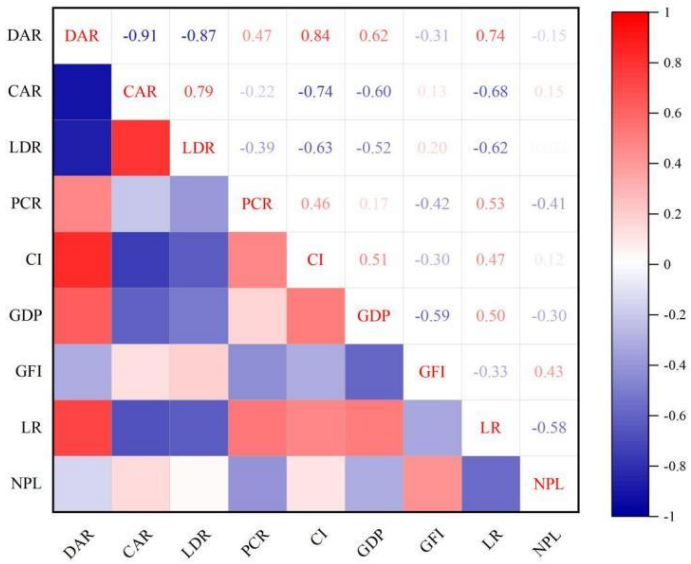

Fig. 4 Correlation Heat Map

(1) It can be seen from the above results that in the macro-environment, the GDP growth in different years has a certain correlation with the non-performing loan ratio of banks. (0.0586) To control the nonperforming loan ratio, the Chinese government has issued programmatic documents many times, so the nonperforming loan ratio of various banks has been significantly reduced. However, from 2019 to 2021, due to the impact of the epidemic, most areas of China have stopped production, which has seriously affected people's income, and the GDP of various regions has

generally decreased, and the nonperforming loan ratio has fluctuated compared with previous years. A decline in GDP will lead to an increase in a certain percentage of bank bad debts, which is negatively correlated with the nonperforming loan ratio.

(2) Secondly, in the external situation of the bank, the main influencing factor is the government's fiscal intervention. (0.154) The government increases fiscal spending to stimulate economic growth, which affects bank loans to a certain extent. At the same time, the government has increased its investment in supervision departments, established supervision departments, strengthened supervision, and formulated supervision policies to affect the social credit environment, which has a positive impact on the reduction of nonperforming loans.

Various factors affect the internal financial situation of banks, and the impact of medium and long-term loan interest rates is the most significant. (0.3029) Medium- and long-term loans can bring more profits to banks, but for a longer period, the liquidity of the overall loan assets is weak, which has a greater impact on the risk behavior of banks and is more likely to lead to the generation of nonperforming loans.

In addition, the loan-to-deposit ratio (0.1178), the cost-to-income ratio (0.1136), the non-performing loan provision coverage ratio (0.1021), and the bank's asset-liability ratio decreased in order. The loan-to-deposit ratio policy was originally used to control the scale of loan issuance and limit the expansion of loan scale by commercial banks in pursuit of profits. However, due to the cancellation of the policy, the banks have not yet found the best loan-to-deposit ratio, which leads to an increase in credit risk and aggravates the rise in the nonperforming loan ratio. Some scholars such as LIANG [11] also pointed out that the provision coverage ratio has also been increasing in recent years, which shows that the current non-performing loan ratio does not truly reflect the quality of the bank's assets and affects the bank's ability to resist risks.

5.4 Policy Suggestion

Based on the above analysis and research, the following policy recommendations are put forward:

First, the government should give full play to the leading and supervising role of the main body in the construction of the social credit environment system. Improve the social credit system to improve the overall authenticity of information disclosure. Clear communication barriers and information barriers between regulatory bodies. At the same time, the improvement of the credit environment can also ease the financing constraints of SMEs. Appropriately increase the proportion of fiscal expenditure, and balance the allocation and effective use of financial resources in various regions. Optimize and adjust the financial system structure, establish credit supervision departments, formulate supervision policies, establish credit supervision platforms, etc. They are all focusing on promoting the major project of improving and optimizing the financial market environment.

Second, the banking industry needs to strengthen the control of loan risks. Reasonably control the scale of loans, strictly examine the credit and repayment ability of lenders, and lend prudently. Improve the customer structure and quality from the source, improve the overall asset quality, and alleviate the impact of the increase in the provision for non-performing loans on the operating efficiency of commercial banks. It is also necessary to adapt to the policy change in the loan-to-deposit ratio. Find the most appropriate loan-to-deposit ratio according to your actual situation, as well as your critical value of the loan-to-deposit ratio, to reduce the dependence on the loan-to-deposit spread. Reasonably avoid its negative impact on profitability. At the same time, it also needs to expand a variety of profitable businesses.

6. Conclusion

(1) by constructing the LightGBM algorithm, it is found that the model has the highest accuracy and the best fitting effect when each sample has equal weight and the number of subtrees is 12. At this time, the model's overall training set R^2 is 0.9067, RMSE is 0.1029; Test set result R^2 is 0.8481, and RMSE is 0.1007.

(2) Among the first-level indicators, the influence of the bank's internal finance on the non-performing loan ratio is 0.7874, the second is the external situation of the bank is 0.154, and the influence of the macro-environment is 0.0586.

(3) Among the secondary indicators, the annual interest rate of bank loans has the highest impact on the non-performing loan ratio, with an impact of 0.3029, while the capital adequacy ratio has the lowest impact of 0.0498.

References

[1] Gu Feng, Zhang Chao. Internal driving effect of exogenous non-performing loans of commercial banks [J]. Journal of Shanghai Jiaotong University, 2006(04): 611-614.

[2] Svetozar Tanaskovi ć, Maja Jandri ć. Macroeconomic and Institutional Determinants of Non-performing Loans. [J] Journal of Central Banking Theory and Practice, 2015, 1, pp. 47-62 Received: 10 December 2014; accepted: 26 December 2014.

[3] Wangxiaorao, Pangzhi & zhangxiru. Direct government intervention, financial repression, and non - performing bank loans [J]. -- Based on inter - Provincial Dynamic GMM estimation method Journal of Beijing Industrial and Commercial University (SOCIAL SCIENCE EDITION), 2018 (01): 97-104.

[4] Wu Jingmei & Wang Ping. Analysis and suggestions on the impact of credit environment on non-performing loan ratio of commercial banks [J]. Investment Research, 2022(01): 4-17.

[5] Xiao Yue, Xiao Binqing, Wang Jie. A Study on the Causes of Non-performing Loans in Regional Small and Medium-sized Banks from the Perspective of Financial Security: "Internal Control" or "External Defense" [J]. Financial Forum, 2022, 27(04): 70-80.

[6] Yindelei. Analysis of the influencing factors of non - performing loan ratio of commercial banks in China Chinese business theory, [J] 2016 (28): 79-80.

[7] Tu Yuhang. Research on the Influencing Factors of the Non-performing Loan Ratio of Commercial Banks in my country [J]. Productivity Research, 2018(12): 40-44+126.

[8] Ling Jianghuai, Kuang Yawen The impact of credit environment on the financing constraints of small and medium-sized enterprises -- An Empirical Study Based on the survey data of Chinese enterprises of the world bank [J] Journal of South China Normal University (SOCIAL SCIENCE EDITION), 2016 (03): 127-132.

[9] Huan Zhang, Peisong Yang, Duli Yu, Kunfeng Wang & Qingyuan Yang. Prediction of methane storage in covalent organic frameworks using a big-data-mining approach. [J] Chinese Journal of Chemical Engineering, 2021(11): 286-296.

[10] Xie Yong, Xiang Wei, Ji Mengzhong, Peng Jun, Huang Yihuai.Application analysis of forecasting monthly housing rent based on Xgboost and LightGBM algorithm [J]. Computer Applications and Software, 2019, 36(09): 151-155+191.

[11] Liang Hao, Yu Yongsheng. The Potential Impact of Limiting the Provision Coverage Ratio of Commercial Banks——An Analysis of "Financial Rules for Financial Enterprises (Draft for Comment)" [J]. New Finance, 2020, (02): 30-35.