ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)

https://doi.org/10.17559/TV-20221003064737
Original scientific paper

A Random Forest Approach to Appraise Personal Credit Risk of Internet Loans

Haining YANG

Abstract: In view of the fact that in recent years, internet loan business has gradually exposed that the pre prevention and management of risks are not comprehensive enough, which has led to the untimely response of most platforms to the consequences of the borrower's breach of contract, resulting in insufficient cash flow on the platform, resulting in a series of problems such as cash withdrawal difficulties and serious runs. In this study, the borrower's personal credit risk identification is studied, and the data mining process and method of credit data risk are proposed. Select the Internet loan data of a domestic city commercial bank, and use random forest algorithm and decision tree algorithm to identify and predict the risk. The research results show that the prediction accuracy of the random forest model built in this paper reaches 97% through the high-quality pre-processing of the original credit data, indicating that the model has a high reliability and can well identify the risks related to Internet loans of commercial banks. At the same time, the research also finds that several interesting characteristics, such as the borrower's balance, amount and fund use, are crucial to identify whether the borrower defaults. In general, the research in this paper can improve the level of commercial banks' lending decision-making, and contribute to the sound development of commercial banks.

Keywords: credit identification; decision tree; internet loan; random forest

1 INTRODUCTION

With the continuous development of high technology and the popularization of the Internet, in this context, the combination of finance and technology, the Internet finance industry has received more extensive attention. Internet finance provides a simpler and more convenient loan business. Compared with traditional bank loan business, more and more people choose Internet loans to facilitate their daily needs. Internet loan refers to a commercial bank that uses information and communication technologies such as the Internet and mobile communications to conduct cross-validation and risk management based on risk data and risk models, automatically accept loan applications and conduct risk assessments online, and complete credit approval, contract signing, and loan approval. Core business operations such as payment and post-loan management, personal loans and working capital loans for consumption, daily production and operation turnover, etc. are provided to eligible borrowers.

With the rapid development of the Internet financial industry [1], the number of credit data is also growing. The asymmetry of information and the lack of information are relatively serious. Most Internet financial platforms lack a relatively perfect personal credit evaluation system [2], which makes it difficult to identify personal credit risks. Even if some financial platforms have their own personal evaluation system, the accuracy is relatively poor. From the existing research, there are few studies to build a data dictionary. There is a lack of detailed understanding of data samples. Most studies directly apply all the data [4], but lack understanding of the data samples themselves. Therefore, this study builds a data dictionary based on data samples, conducts high-quality pre-processing of sample data, and then uses data mining methods to analyse and research the data of a domestic commercial bank. It is of great significance for the Internet financial platform to put forward useful suggestions for personal credit risk identification, which can promote the development of the Internet financial platform to a certain extent.

2 RELATED WORKS

In general, the main methods of data research are model construction and data analysis [3, 5]. In the field of data analysis, in terms of credit evaluation indicators, the empirical results of Collier & Hampshire [6] show that the larger the loan amount of the borrower, the higher the ratio of the loan amount to its annual income, the lower the possibility of obtaining a loan, and the higher the loan interest rate. High. Lin, Li, and Zheng [7] summed up a series of specific evaluation indicators through empirical research methods, including: gender, age, marital status, loan amount, company size and so on.

Emekter, Tu, Jirasakuldech and Lu [8] pointed out that the more important factor in the risk of online loan default is credit rating, and the higher the credit rating, the lower the probability of default risk of borrowers. Mild, Waits and Wöckl [9] used decision support tools to analyse credit data, established a credit risk assessment model, and used a linear regression model to study and analyse the credit risk of borrowers. Among decision support tools, loan interest rates, lender annual income and borrower repayment period are the main drivers of credit risk. Lin, Prabhala and Viswanathan [10] analysed the influence of social relations in credit evaluation, and found that social relations play an important role in the success of borrowers' borrowing and the risk of default of borrowers. The degree of friendship of applicants can improve the success of financing probability, and can also reduce the interest rate of the loan when financing.

Everett [11] proposed that the loan group and more soft information would indeed reduce the credit risk of online lending, but the leader of the loan group, because of having more information and certain lending decision-making power, would often lead to a certain monopoly rent, which would increase the cost and default risk of the borrower. Dorfleitner et al. [12] proposed that soft information can reflect the borrower's education level to a certain extent, so the probability of spelling errors in soft information is negatively related to the default rate. C. Serrano-Cinca, Gutiérrez-Nietoand López-Palacios [13] believed that the borrower's credit rating may have a certain impact on whether normal repayment is possible, and made analysis and prediction to reduce the degree of information asymmetry between the borrower and the borrower through credit rating.

Le [14] believed that investors had different sensitivities to the borrower's personal information and order information in the process of risk investment. In the classification of risk sensitivity, the information with higher sensitivity was the borrower's loan term, credit rating, amount and marital status. The information with lower risk sensitivity was the borrower's industry, working time, age and education level. Wen, Zhang and Wu [15] found that the main factor affecting platform credit in platform transactions is information transparency in transaction credit, and the second important factor is compliance in brand credit. The platform should focus on it. Luo and Chen [16] used the sample data of a bank to assess the credit risk of micro credit. Through research and analysis, they found that the natural information of customers, such as gender, education background, age, income, occupation and territory, had a significant impact on credit risk. Wang and Liu [17] established a credit risk assessment model based on actual data and found that the important factors affecting personal credit risk are the borrower's age, marital status, education level and other six indicators.

In terms of credit evaluation methods, Ala'raj and Abbod [18] proposed a new combination method based on classifier consistency to combine each classifier, and verified the high prediction performance of the model through credit score data. Malekipirbazari and Aksakalli [19] proposed a random forest classification algorithm to model and predict the borrower's borrowing status based on the data obtained from the Lending Club platform. Research shows that the random forest prediction results have a higher accuracy and higher precision. Hayashi and Oishi [20] applied a j48-graft continuous rule extraction method to credit scoring, which improved the accuracy and interpretability of the model. Rajamohamed and Manokaran [21] proposed an improved rough K-means algorithm, and used SVM, random forest, decision tree, and other algorithms to classify data to predict credit card loss. Plawiak, Abdar, and Acharya [22] proposed a classifier based on the deep genetic cascade ensemble of different support vector machine classifiers to be used in the prediction of credit scores. The results show that the model has the highest prediction accuracy. Danenas, Paulius, Gintautas and Garsva [23] constructed a support vector machine model, which is based on particle swarm optimization, and compared with Logistic regression and RBF neural network, and found that this method has higher classification accuracy in credit evaluation degree, but it is insufficient in terms of performance stability. Malekipirbazari and Aksakalli [24] used the random forest classification algorithm to conduct credit evaluation analysis on the borrower's data. The research results show that the random forest algorithm has a higher accuracy on the borrower's data with high credit. Wang, Han, Liu, and Luo [25] proposed a deep learning algorithm based on the attention mechanism LSTM for consumer credit scoring, and the results show that this method has higher prediction accuracy. Kw, Ml, Jc, Xz,and Gang [26] applied the integration method to the field of credit evaluation, proposed a support vector machine integration model to improve DS evidence theory and used it to build a personal credit evaluation model, and incorporated attribute reduction into the modelling process.

It can be seen from the relevant research work that the current research on Internet personal credit risk identification rarely conducts comprehensive and high-quality processing on the samples, lacks the understanding of the data samples themselves, does not accurately and comprehensively explain the factors that affect the borrower's default, and does not fully understand the data, which leads to low prediction accuracy of the model and is difficult to meet the needs of the borrower's credit analysis and identification under big data. Therefore, this research establishes a data dictionary based on data samples. The data dictionary describes what data the data sample contains, which can clearly and intuitively see the detailed description of each element of the sample data, which is conducive to the understanding of the sample data itself. The data dictionary can also have a clear understanding of the description of the statistical values of complex and large number of samples, as well as the types and contents of the variables included, it can quickly grasp the information of data, providing great convenience for data analysis, processing and subsequent modelling.

3 RESEARCH DATA AND METHOD

3.1 Data Acquisition

This study selects personal loan data of a domestic commercial bank to compare domestic and foreign credit data. In this data sample, there are 7313 records in the original data, and one record is a sample number. The remaining 33 valid attributes of the research sample data are shown in Tab. 1.

Table 1 Valid property sheet

Number	Property Name	Number	Property Name
1	Business types.	17	Overdue penalty interest balance (¥).
2	Whether the business is under the credit line.	18	Compound interest balance (¥).
3	Industry type (largest class).	19	Margin amount (¥).
4	Industry type (big class).	20	Margin percentage.
5	Industry type (middle class).	21	Execution rate.
6	Industry type (sub-class).	22	Contract start date.
7	Currency.	23	Contract expiry date.
8	Amount (¥).	24	Term (months).
9	Actual billing amount (¥).	25	Main guarantee method.
10	Balance (¥).	26	Occurrence type.
11	Normal balance (¥).	27	Whether to issue an agent.
12	Overdue balance (¥).	28	Interest rate floating method.
13	Sluggish balance (¥).	29	Floating value of interest rate.
14	Bad debt balance (¥).	30	Use.
15	On-balance sheet interest (¥),	31	Whether online loan business.
16	Off-balance sheet interest (¥).	32	Whether paper files are handed over.
		33	Days past due.

The data in this study is the loan information data from a domestic commercial bank to identify the credit of user characteristics. Therefore, the most important thing in the data attribute variable is whether the borrower's repayment period is overdue. Through the decision tree algorithm and random forest, the algorithm establishes a model for classification research, takes the number of days overdue for the principal as the target variable, and changes the variable to a new name for whether there is overdue behaviour. In the sample value, the number of overdue days is 0, there is no overdue behaviour, and the number of overdue days greater than 0 is regarded as overdue behaviour. The remaining 32 attribute variables were selected as the characteristic variables of this study.

3.2 Attribute Category Division

According to the data dictionary and data attribute characteristics, the data is classified into attribute categories. The data record information of the commercial bank mainly includes the borrower's industry, contract term, loan purpose, and repayment balance data after the loan. The basic information and credit information attributes are divided into:

The basic information attributes of the borrower, mainly including the type of loan business, the industry type of the borrower, the contract period and the main guarantee methods, as shown in Tab. 2.

Table 2 Borrower basic information attribute table

Number	Property Name	Number	Property Name
1	Business types.	10	Term (months)
2	Whether the business is under the credit line.	11	Main guarantee method.
3	Industry type (largest class).	12	Occurrence type.
4	Industry type (big class).	13	Whether to issue an agent.
5	Industry type (middle class).	14	Interest rate floating method.
6	Industry type (sub-class).	15	Use.
7	Currency.	16	Whether online loan business.
8	Contract start date.	17	Whether paper files are handed over.
9	Contract expiry date.		

Table 3 Borrower loan information attribute table

Number	Property Name	Number	Property Name
1	Amount (¥).	9	Off-balance sheet interest (¥).
2	Actual billing amount (¥).	10	Overdue penalty interest balance (¥).
3	Balance (¥).	11	Compound interest balance (¥).
4	Normal balance (¥).	12	Margin amount (¥).
5	Over due balance (¥).	13	Marginpercentage.
6	Sluggish balance (¥).	14	Execution rate.
7	Bad debt balance (¥).	15	Floating value of interest rate.
8	On-balances heet interest (¥).	16	Days past due.

The attributes of the borrower's credit information, mainly include the borrower's balance after repayment, overdue interest and execution interest rates, etc., as shown in Tab. 3.

3.3 Descriptive Statistical Analysis

In order to better understand the data and prepare for the subsequent model building, all the above sample variables were initially screened out, leaving 33 attribute variables. The following uses SPSS statistical analysis software to carry out descriptive statistics on continuous attribute variables, mainly including data of the maximum value, minimum value, average value and standard deviation of the variable, as shown in Tab. 4.

Table 4 Descriptive statistical analysis table (*N* = 7313)

(¥)	Min	Max	Average	Std
Amount	2000	9800000	272707.79	430586.241
Actual billing amount	2000	9800000	273141.32	431057.464
Balance	2000	9800000	237548.8380	426602.46926
Normal balance	0	9800000	234331.6531	413299.27491
Overdue balance	0	24820.47	157.6818	1355.26821
Sluggish balance	0	8000000.00	3059.5032	112328.92311
Bad debt balance	0	0	.00	.000
On-balance sheet interest	0	137264.11	41.2294	1617.57111
Off-balance sheet interest	0	962146.29	314.4047	12823.79607
Overdue penalty interest balance	0	75.83	0.2178	2.66404
Compound interest balance	0	282713.54	61.2004	3386.89082
Execution rate	4.35	15.00	7.7580	0.67700
Term (months)	6	144	92.68	39.923
Floating value of interest rate	0	50.00	3.0433	0.85513
Days past due	0	1579	2.21	44.542
valid N (listwise)				

3.4 Data Cleaning

In order to avoid data redundancy, which is not conducive to the subsequent model research and analysis, after descriptive statistics of the data, the data samples are screened for data variables. Since the four types of data about the industry type except the largest category, the remaining three sub-categories exceed 50 items, so it is replaced with the largest class, and three of its items are deleted. The two variables of contract start date and expiry date can be replaced by duration and deleted. Whether the credit business, currency, interest rate floating method, whether the agency is issued, whether the online loan business, whether the paper file is handed over to the sample value, etc., are meaningless to the model decision, delete it, and the bad debt balance, margin amount and margin ratio sample, the data is all 0, and it is deleted. There are 14 deleted data, and 19 variables remain.

From the sample data and the actual meaning of the variables, it can be seen that the amount is consistent with the sample value of the actual outgoing amount attribute, so delete the actual outgoing amount and retain the attribute of the amount instead. The balance consists of the normal balance and the overdue balance, so delete the normal balance and use the balance to replace the borrower's normal balance. Since this paper studies whether the principal of the loan will be overdue and default, so the overdue balance, sluggish balance, and on-balance sheet interest owed, Off-balance sheet interest, overdue fine amount, compound interest balance, these six variables are all record variables that the principal has overdue behaviour. They belong to the overdue variables and are deleted, leaving only 11 valid data variables.

Table 5 Valid attribute variable table after cleaning

Number	Property Name	Number	Property Name
1	Business	6	Term
2	Industry type (largest class)	7	Main guarantee method
3	Amount	8	Type of occurrence
4	Balance	9	Floating value of interest rate
5	Execution rate	10	Use
		11	Principal overdue days

It can be seen from descriptive statistics that continuous data has no missing values, discrete data only has missing values in industry type (largest class), and the number of missing value samples is 1118. Because of industry type (largest class), this attribute variable has practical significance and cannot be arbitrary. Therefore, the sample value of the missing data value is filled with 0, which is recorded as the missing data sample value.

Since the classification model is used for modelling in this study, the attributes of continuous values will affect the classification results, reduce the accuracy of the classification model, and reduce the efficiency of the model, so the continuous attributes in the experimental data are discretized.

3.5 Research Method

(1) Decision Tree:

The decision-making process of the decision tree needs to start from the root node of the decision tree [24], analyse and compare the data to be tested with the characteristic nodes in the decision tree, and then select the next comparison branch according to their analysis and comparison results, and so on, until the leaf node is used as the final result of the decision selection. The key of the decision tree is to measure the attribute selection. The attribute measurement of the decision tree is to use the information gain to select the attribute information, and select the highest information gain as the test attribute of the current node.

(2) Random Forest:

Random forest is a combination algorithm of multiple decision trees. CART decision tree is used as the meta classifier of random forest, and its way is to build a forest randomly. In machine learning, random forest (Zhao, 2016) is a classifier containing multiple decision trees. It is an integrated learning model that uses multiple trees to train and predict samples.

Each decision tree in a random forest is independent of each other, and the output category is determined by the mode of the output category of the individual tree. It includes multiple decision trees trained by Bagging integrated learning [28] technology, input feature sample data to be classified, and the final result is determined by voting the output of a single decision tree. This method combines the ideas of Bagging (Bootstrap aggregating) and (random subset method) to build a set of decision trees. It uses bootstrap to randomly and repeatedly extract sample data, and uses random subspace method to randomly select $m(m < M)$ (characteristics of training data set) feature variables to generate a decision tree [30].

4 EXPERIMENT RESULTS AND DISCUSSION

4.1 Model Construction Based on Random Forest Algorithm

Use the Random Forest Classifier in the python sklearn.ensemble module to implement the random forest model, use the train_test_split method in the sklearn.model_selection module to divide the sample data into training sets: the test set is 6:4, set oob_score to True, and then pass forest. feature_importance_output feature importance ranking: Use the borrower data samples to build a random forest model, the accuracy of the random forest test set is 0.978, and the model out-of-bag estimation accuracy score oob_score: 0.97, the confusion matrix of the model sample obtained below is shown in Tab. 6.

Table 6 Confusion matrix table of model samples

Forecast Result	0	1
0	2856	11
1	52	7

The output random forest model variable importance ranking is shown in Tab. 7.

Table 7 Random Forest model variable importance ranking table

Sort by name	Property Name	Score
1)	Industry type (largest class)	0.188386
2)	Balance	0.178934
3)	Amount	0.176340
4)	Main guarantee method	0.114700
5)	Business	0.106230
6)	Use	0.089293
7)	Interest rate floating method	0.072453
8)	Execution rate	0.037372
9)	Term	0.034466
10)	Type of occurrence	0.001826

The ROC curve of random forest is shown in Fig. 1. According to the model results, it can be seen that the accuracy of the model is good, and from the results of the variable importance output by the random forest model, the difference between the top five variables is not too big, and they are all the largest categories of industry types, balance, Amount, main guarantee methods and business types, and the last one is the occurrence type, loan period and interest rate, the degree of influence is not too big.

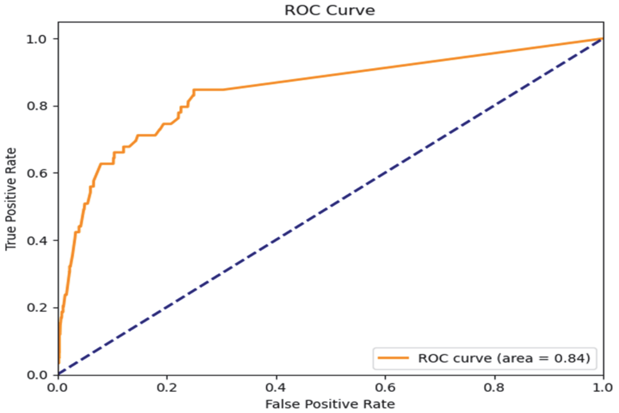

Figure 1 Random forest ROC curve

When borrowing a loan, you need to pay attention to the first few variable information, whether there is any overdue behavior on the principal. Personal credit identification of user characteristics is a more important information prompt. The text has to be organized in the following order:

4.2 Model Construction Based on Decision Tree Algorithm

By constructing the most important feature analysis, deleting the low-importance features, and selecting the top eight attribute feature variables to construct the decision tree model, the experiment first uses the pandas module in the python language to input and read data, and then use the train_test_split method in the sklearn.model. Selection module divides the test set of the training set. In this experiment, the original data is divided into the training data set and the test data set according to the ratio of 6:4. Using the DecisionTreeClassifier function in the sklearn.tree module, the criterion of the value of the parameter is set to "gini" to implement the CART decision tree model, and the value of the max_depth parameter is set to 15, and pruning is performed in the process of generating the decision tree. Then, through the dec.feature_importance_output feature importance ranking, the accuracy of the model test set is 97.5, and finally the Graphviz package is used to visualize the decision tree. Some screenshots of the generated CART decision tree model are shown in Fig. 2.

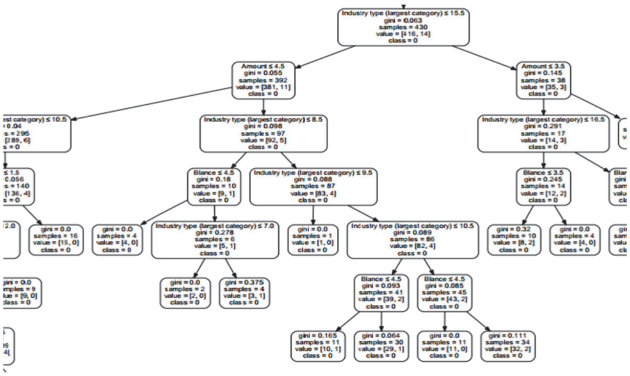
Industry type (largest category) ≤ 15.5 gini = 0.063 samples = 430 value = [416, 14] class = 0
Amount ≤ 4.5 gini = 0.055 samples = 392 value = [381, 11] class = 0
Amount ≤ 3.5 gini = 0.145 samples = 38 value = [35, 3] class = 0
Industry type (largest category) ≤ 8.5 gini = 0.098 samples = 97 value = [92, 5] class = 0
Industry type (largest category) ≤ 16.5 gini = 0.291 samples = 17 value = [14, 3] class = 0
Blance ≤ 4.5 gini = 0.18 samples = 10 value = [9, 1] class = 0
Industry type (largest category) ≤ 9.5 gini = 0.088 samples = 87 value = [83, 4] class = 0
Blance ≤ 3.5 gini = 0.245 samples = 14 value = [12, 2] class = 0
gini = 0.0 samples = 16 value = [15, 0] class = 0
gini = 0.0 samples = 4 value = [4, 0] class = 0
Industry type (largest category) ≤ 7.0 gini = 0.278 samples = 6 value = [5, 1] class = 0
gini = 0.0 samples = 1 value = [1, 0] class = 0
Industry type (largest category) ≤ 10.5 gini = 0.089 samples = 86 value = [82, 4] class = 0
gini = 0.32 samples = 10 value = [8, 2] class = 0
gini = 0.0 samples = 4 value = [4, 0] class = 0
gini = 0.0 samples = 2 value = [2, 0] class = 0
gini = 0.375 samples = 4 value = [3, 1] class = 0
Blance ≤ 4.5 gini = 0.093 samples = 41 value = [39, 2] class = 0
Blance ≤ 4.5 gini = 0.085 samples = 45 value = [43, 2] class = 0
gini = 0.165 samples = 11 value = [10, 1] class = 0
gini = 0.064 samples = 30 value = [29, 1] class = 0
gini = 0.0 samples = 11 value = [11, 0] class = 0
gini = 0.111 samples = 34 value = [32, 2] class = 0

Figure 2 Decision tree classification model diagram

The order of importance of variables in the output decision tree model is shown in Tab. 8.

Table 8 Decision tree model variable importance ranking table

Sort by name	Property Name	Score
1)	Amount	0.186389
2)	Balance	0.181578
3)	Use	0.165627
4)	Industry type (largest class)	0.159915
5)	Main guarantee method	0.097008
6)	Business	0.090706
7)	Interest rate floating method	0.077130
8)	Execution rate	0.041645

According to the results of the model, it can be seen that the accuracy of the model is good. In terms of the importance of variables output by the decision tree model, the top five variables are the largest categories of balance, amount, purpose and industry type, as well as the main guarantee methods, while the floating value of the interest rate and the execution interest rate are the last, and the impact is not too great.

4.3 Experiment Results Analysis

The above data results show that, in terms of the accuracy of the model, the accuracy of the model construction of the personal lending data of a bank in China shows that the accuracy of the model is as high as about 97%. From this, it can be concluded that the establishment of a data dictionary is very necessary for the later model construction, the accuracy of the model performance results is high, and the classification algorithm used is applicable to the domestic and foreign data in this study. However, from the perspective of the model confusion matrix, the model still performs poorly in identifying defaulting borrowers. The main reasons for this may be as follows:

(1) In terms of data samples, according to the personal loan credit data information obtained by a domestic commercial bank, the number of information indicators is small and the information is incomplete. There are 49 data variables in total, but only about 10 valid information data are available for modelling, and the number of data set samples is small, which has an impact on the model fitting and training effects.

(2) The data is unbalanced. From the data dictionary established and the descriptive statistics of the data, we can clearly see that the number of data samples obtained by a domestic commercial bank is small, and the target variables of the data set are seriously unbalanced. The imbalance of the data set has a certain impact on the training process of domestic and foreign data modelling and the training results of the model.

(3) The process of data processing affects the desensitization of the acquired data. The hidden information may also contain information that is important to the decision variables. It is possible to affect the accuracy of the model results as well as the deviation of the confusion matrix.

5 CONCLUSIONS

This paper selects the Internet loan data of a city commercial bank in China, and uses random forest algorithm and decision tree algorithm to carry out risk identification and prediction research. The research structure shows that:

(1) Statistical analysis of the data, as well as pre-processing such as cleaning and transformation, can effectively improve the accuracy of the prediction model.

(2) The prediction accuracy rate of the random forest model constructed in this paper reaches 97%, indicating that the model has high credibility and can well identify the risks related to Internet loans of commercial banks. When borrowing loans, you need to focus on the balance, amount, purpose and the largest category in the industry and the main guarantee methods are important information prompts for whether the principal is overdue and for personal credit identification of user characteristics.

6 REFERENCES

[1] Yao, Q., Hou, D., & Cheng, L. (2021). Precautionary saving, inequality and fiscal policy: a hank model. *Economic computation and economic cybernetics studies and research/Academy of Economic Studies*, *55*(1/2021), 57-72. https://doi.org/10.24818/18423264/55.1.21.04

[2] Cai, Y. X., Gong, Y. L., & Sheng, G. Y. (2021). The gold price and the economic policy uncertainty dynamics relationship: the continuous wavelet analysis. *Economic computation and economic cybernetics studies and research/Academy of Economic Studies*, *55*(1/2021), 105-116. https://doi.org/10.24818/18423264/55.1.21.07

[3] Xu, W., Sun, H., Awaga, A., Yan, Y. S., & Cui, Y. (2022). Optimization approaches for solving production scheduling problem: A brief overview and a case study for hybrid flow shop using genetic algorithms. *Advances in Production Engineering & Management*. https://doi.org/10.14743/apem2022.1.420

[4] Julio, C. & Karina, T. K. (2021). Manipulation power in bargaining games using machiavellianism. *Economic computation and economic cybernetics studies and research, Academy of Economic Studies*, 55(2/2021), 299-313. https://doi.org/10.24818/18423264/55.2.21.18

[5] Zhang, L., Yan, Y., Xu, W., Sun, J., & Zhang, Y. (2022). Carbon Emission Calculation and Influencing Factor Analysis Based on Industrial Big Data in the "Double Carbon" Era. *Computational Intelligence and Neuroscience, 2022*. https://doi.org/10.1155/2022/2815940

[6] Collier, B. & Hampshire, R. C. (2010). Sending mixed signals: multilevel reputation effects in peer-to-peer lending markets. *Conference on Computer Supported Cooperative Work*. https://doi.org/10.1145/1718918.1718955

[7] Lin, X., Li, X., & Zheng, Z. (2017). Evaluating borrower's default risk in peer-to-peer lending: evidence from a lending platform in China. *Applied Economics*, *49*, 3538-3545. https://doi.org/10.1080/00036846.2016.1262526

[8] Emekter, R., Tu, Y., Jirasakuldech, B., & Lu, M. (2015). Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. *Applied Economics*, *47*, 54-70. https://doi.org/10.1080/00036846.2014.962222

[9] Mild, A., Waitz, M., & Wöckl, J. (2015). How low can you go? Overcoming the inability of lenders to set proper interest rates on unsecured peer-to-peer lending markets. *Journal of Business Research*, *68*, 1291-1305. https://doi.org/10.1016/j.jbusres.2014.11.021

[10] Lin, M., Prabhala, N. R., & Viswanathan, S. (2011). Judging Borrowers by the Company They Keep: Friendship Networks and Information Asymmetry in Online Peer-to-Peer Lending. *ERPN: Industry Studies (Sub-Topic)*. https://doi.org/10.2139/ssrn.1355679

[11] Everett, C. R. (2015). Group Membership, Relationship Banking and Loan Default Risk: The Case of Online Social Lending.*ERPN: Information Asymmetry (Sub-Topic)*.

[12] Dorfleitner, G., Priberny, C., Schuster, S., Stoiber, J., Weber, M., Castro, I. D., & Kammler, J. (2016). Description-text related soft information in peer-to-peer lending-Evidence from two leading European platforms. *Journal of Banking and Finance*, *64*, 169-187. https://doi.org/10.1016/j.jbankfin.2015.11.009

[13] Serrano-Cinca, C., Gutiérrez-Nieto, B., & López-Palacios, L. (2015). Determinants of Default in P2P Lending. *PLoS ONE*, *10*. https://doi.org/10.1371/journal.pone.0139427

[14] Le, Z. H. (2015). Identity Discrimination:A Study on Efficiency of Internet Financial Innovation Based on P2P Network Lending. *Economic Management Journal*.

[15] Wen, X., Zhang, Z., & Wu, X. (2017). A Research on the Influence Factors of P2P Lending Market. *Advances in economics and business*, *5*, 11-17. https://doi.org/10.13189/aeb.2017.050102

[16] Luo, F. & Chen, X. (2017). Credit risk assessment of personal small loan based on logistic regression model and its application. *The Theory and Practice of Finance and Economics*.

[17] Wang, J., Liu, L., & Finance, S. O. (2017). Research on real estate credit risk of commercial banks in china based on logistic model. *Review of Economy and Management.*

[18] Ala'Raj, M., & Abbod, M. F. (2016). Classifiers consensus system approach for credit scoring. *Knowledge-Based Systems*, *104*(jul.), 89-105. https://doi.org/10.1016/j.knosys.2016.04.013

[19] Malekipirbazari, M. & Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2015.02.001

[20] Hayashi, Y. & Oishi, T. (2018). High accuracy-priority rule extraction for reconciling accuracy and interpretability in credit scoring. *New Generation Computing*, *36*(4), 393-418. https://doi.org/10.1007/s00354-018-0043-5

[21] Rajamohamed, R. & Manokaran, J. (2017). Improved credit card churn prediction based on rough clustering and supervised learning techniques. *Cluster Computing*, *21*(3), 1-13. https://doi.org/10.1007/s10586-017-0933-1

[22] Pawiak, P., Abdar, M., & Acharya, U. R. (2019). Application of new deep genetic cascade ensemble of svm classifiers to predict the australian credit scoring. *Applied Soft Computing*, *84*(2019). https://doi.org/10.1016/j.asoc.2019.105740

[23] Danenas, P. & Garsva, G. (2015). Selection of support vector machines based classifiers for credit risk domain. *Expert Systems with Application*, *42*(6), 3194-3204. https://doi.org/10.1016/j.eswa.2014.12.001

[24] Malekipirbazari, M. & Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2015.02.001

[25] Wang, C., Han, D., Liu, Q., & Luo, S. (2019). A deep learning approach for credit scoring of peer-to-peer lending using attention mechanism lstm. *Quality Control Transactions*, *7*, 2161-2168. https://doi.org/10.1109/ACCESS.2018.2887138

[26] Kw, A., Ml, B., Jc, B., Xz, B., & Gang, L. B. (2022). Research on personal credit risk evaluation based on xgboost. *Procedia Computer Science*,*199*, 1128-1135. https://doi.org/10.1016/j.procs.2022.01.143

[27] Zhao, F. Z. (2016). Research on reputation evaluation model of mixed enterprises based on decision tree and asvm. *Computer Knowledge and Technology*.

[28] Rao, C., Liu, Y., & Goh, M. (2022). Credit risk assessment mechanism of personal auto loan based on pso-xgboost model. *Complex & Intelligent Systems*, 1-24. https://doi.org/10.1007/s40747-022-00854-y

[29] Li, S. & Ji, X. (2019). The application of Lgb bag in credit risk assessment of p2p network borrowers *Technical Economy*, *38* (11).

[30] Jin-Wang, W. U. & Zhou-Yi, G. U. (2018). Credit risk assessment of commercial banks based on non-equilibrium samples: taking bank a as an example. *Financial Theory & Practice*.

Contact information:

Haining YANG, PhD
Department of Management Science and Engineering, School of Economics and Management, University of Science and Technology Beijing,
Department of Management Science and Engineering, School of Economics and Management, University of Science and Technology Beijing, Beijing 100083, China
E-mail: yanghaining@apiins.com