




## Research on Financial Credit Evaluation and Early Warning System of Internet of Things Driven by Computer-Aided Technology

Guicang Yang<sup>1</sup>

<sup>1</sup>Department of Finance, Zhengzhou College of Finance and Economics, Zhengzhou 450000, China  
[yang2207652983@126.com](mailto:yang2207652983@126.com)

Corresponding author: Guicang Yang, [yang2207652983@126.com](mailto:yang2207652983@126.com)

**Abstract.** As China's economy becomes more credit oriented, the role of credit consumption in driving economic growth is further highlighted, and the willingness and ability of credit consumption of residents are steadily increasing. Domestic Internet financial institutions have taken personal consumption loan business as one of their future development strategies. However, the risk management level of domestic Internet financial institutions for personal consumer loans is relatively low, and the management methods and methods are still relatively backward. Besides, Internet financial institutions do not have effective personal credit evaluation methods, which seriously hinders the development of personal credit business. How to implement data-driven service upgrade technology to assist market regulation has become a hot research issue. In this context, this paper realizes the integration of user authentication information, financial information and behavior information through data processing and feature engineering, and uses deep forest algorithm to build a model for credit risk assessment. The results show that this method can effectively mine the potential value of data and improve the effect of credit risk assessment and warning. Finally, this paper puts forward suggestions for the credit risk early warning and market supervision of Internet finance.

**Keywords:** Internet finance; Deep forest; Financial credit evaluation assessment; Risk warning; Computer aided technology

**DOI:** <https://doi.org/10.14733/cadaps.2022.S6.158-169>

### 1 INTRODUCTION

With the development of science and technology and the development of economic globalization, the financial industry has undergone historic changes and Internet finance has emerged. The emergence of Internet finance has played a huge role in promoting the development of micro, small and medium-sized enterprises and opened the door for mass entrepreneurship and innovation. With the advent of the era of big data in the Internet of Things and the widespread

promotion of intelligent sensing technologies such as RFID NFC, it is possible to bring the third reform to the financial industry, and a new model has emerged: Internet of Things finance [1]. So far, the connotation of Internet finance is not clear, there are many views, but scholars have reached a consensus on the core elements of Internet finance, the two most important elements of Internet finance are financial function and Internet platform. The forms of Internet finance in China mainly include third-party payment online lending platform and crowd funding. In the new financial ecological environment, the development of digital inclusive financial products to meet the consumer financing needs of customers, and the establishment of a sound risk management system and regulatory norms has become the top priority in the development of Internet finance.

Traditional finance and current Internet finance are both subjective credit systems, and their credit evaluation mainly relies on manual evaluation, while Internet of Things finance is an objective credit system, which mainly uses intelligent sensing devices to collect objective information and constitute an objective credit system, which can comprehensively resolve the credit crisis existing in current Internet finance. Internet of Things finance is a new mode based on Internet finance and traditional finance, which is an innovation of Internet finance and a new application field. The financial model of Internet of Things is mainly based on cloud big things transfer technology, realizing the integration of capital flow, information flow and physical flow of small, medium and micro enterprises, comprehensively reducing financial risks brought by virtual economy, providing consumers with more convenient financial services and providing more financial service support for the development of real industry economy. The emergence of the financial mode of the Internet of Things solves the problem of the innovation of the mode of the Internet of Things and can promote the deep integration of the financial field.

China attaches great importance to the development of digital finance, emphasizes the in-depth integration of big data technology and business, strengthens the innovation of products and services, and improves service efficiency. Data-driven business upgrading and technology-assisted credit decision-making will provide better support and services for the credit financing needs of individual private economy and small and micro enterprises. Under the trend of the continuous close integration of fintech and banking business, how to apply Internet big data and algorithms to risk assessment and market regulation has become a hot topic in current research. Therefore, the application of big data analysis technology in credit risk assessment and early warning of Internet finance is of great significance to promote the transformation and upgrading of Internet financial products and services and the healthy development of the industry. Firstly, it provides Internet financial institutions with effective means to avoid credit risks and reduces the credit losses of Internet financial institutions. Secondly, it improves the work efficiency of Internet financial institutions and expands their business volume. Thirdly, credit evaluation model can assist Internet financial institutions to tap the long tail market. Therefore, it is very important and necessary to use computer aided technology to study the financial credit evaluation and warning of Internet of Things.

## 2 RELATED STUDIES

Modern credit evaluation methods originated in the United States, which has formed a mature personal credit investigation system Experian, Equifax and TransUnion are the three major personal credit data collection companies in the United States, and they mainly use FICO method to conduct credit scores. In 2006, the People's Bank of China established a national personal credit investigation system. Subsequently, the People's Bank of China issued a notice on preparing for credit investigation, and eight institutions, including Zhema Credit, Tencent Credit and Pengyuan Credit, were allowed to prepare for personal credit investigation. There are a lot of researches on credit evaluation methods at home and abroad. Internet finance has injected new vitality into China's financial market, and users' demands for consumer finance are becoming more and more diverse. Through the survey of financial product specifications on the official websites of major banks and other public materials, the results show that major banks are increasing the

development and investment in financial products. Zhou et al. [3] pointed out that in the era of big data, Internet financial credit risks are characterized by strong concealment and wide spread, and traditional regulatory means are difficult to implement. Computer-aided technology can promote the industry management level effectively. Therefore, the application of Internet technology to the field of credit risk has become a research hotspot. Sun [4] takes default probability as the credit risk evaluation standard, the logistic regression model of the borrower's credit risk is constructed, and the loan amount, loan term and the ratio of the amount repaid are obtained through empirical study. The recent repayment is the important index affecting credit risk. Recently, random forest training is used to extract effective predictive variables as the input of neural network to establish a combination model and Lending Club data sets verify that the method has good learning ability and classification accuracy [5].

With the development of computer technology, scholars have found that artificial intelligence methods can efficiently process unstructured data and greatly enhance the prediction performance of models. Neural Network (NN) is a common artificial intelligence method. It generally has an input layer, an output layer and hidden layer. The feature vectors of the input layer are transformed to the output layer through the hidden layer, and the classification results are achieved by the output layer indicated by Deng et al [6]. the results of [7] show that the practicability of SVM is better than that of neural network model, and SVM is considered to be a suitable method for developing credit scoring model. Risk assessment system is an important support to reduce the cost of credit approval and control loan risks. For the loans of small, medium and micro businesses and modern financial credit, efficient credit evaluation can effectively realize the batch release of loans and credit payment products. Based on the financial and financial transaction data accumulated by traditional banks, Zhang et al. [8] obtained customers' credit data from Internet platforms, integrating objective credit data based on Internet of Things finance, a evaluation and warning analysis system oriented to Internet of Things finance is constructed. Therefore, computer aided technology can not only process high latitude data, but also has perfect theoretical support, which has been proved by the industry to have good practical effect.

Single credit evaluation model has been thoroughly studied by scholars, and the room for improvement of single model is limited, especially for big data or complex objects. More and more scholars turn to hybrid methods and integrated learning methods. Studies show that they can greatly improve the performance of credit evaluation system. Hybrid model is a prediction model that combines two or more different models together. Common integration methods include Bagging, Random Subspace and Boosting. Bagging processes sample points of the initial training set to generate multiple training subsets. Random subspace Randomly selects the training mode of feature subset; Boosting can combine a series of weak classifiers into strong ones. Belgiu et al. [9] proposed a credit evaluation model based on fuzzy mathematics and neural network, and finds that the combination of finance and Internet big data can improve the level of financial services and better realize the inclusive value of digital finance. The integration model is established by the authors of [10]. The experimental results show that the integration method improves the prediction accuracy of neural network with better financial credit evaluation and early warning performance than single model.

To sum up, due to the low entry threshold and low default cost of Internet finance, its credit risks are characterized by strong concealment and wide spread. Research mainly risk assessment model, at present stage in the development of a new round of financial innovation, financial technology and machine learning techniques such as increasingly involved in the actual business, collecting, processing and analysis data to enhance the ability, through the big data technology application in business improve the level of risk control is the inevitable trend of future development. At present, the domestic financial supervision is delayed, supervision measures are not timely, supervision means are single, low efficiency, supervision is difficult, there is no evidence to rely on the difficulty of auditing and operation of large problems such as, for example, local small, medium and micro enterprises have difficulty in funding, deliberately apply for credit financing to the bank, if the bank credit is not approved, easy If the bank audit is not strict, it is

very easy to accumulate financial risks. In this paper, the method of data analysis and feature engineering is adopted to construct behavioral features in the data information to enhance the analysis of user behavior habits. The deep forest model is used to realize the feature combination of multidimensional data information. Combined with empirical analysis, the evaluation effect of this method on credit risk of Internet finance is explored

### 3 FINANCIAL CREDIT EVALUATION AND EARLY WARNING SYSTEM BASED ON DEEP RF MODEL

#### 3.1 Random Forest (RF) Model

RF is a new integrated machine learning algorithm, which achieves better feature extraction effect by establishing multiple forests and integrating and synthesizing multiple forests. RF is a new integrated machine learning algorithm, which achieves better feature extraction effect by establishing multiple forests and integrating and synthesizing multiple forests. The Bagging method, also known as bootstrap aggregating, primarily uses put-back selection data to train models and build sub-classification models, and build enhanced overall classification models from multiple sub-classification models. The algorithm has been proved to have strong generalization ability, simple implementation and no over-fitting. The generalization error of random function of RF algorithm is:

$$PE = P_{X,Y}(\mathbf{K}(\mathbf{X}, \mathbf{Y}) < 0) \quad (1)$$

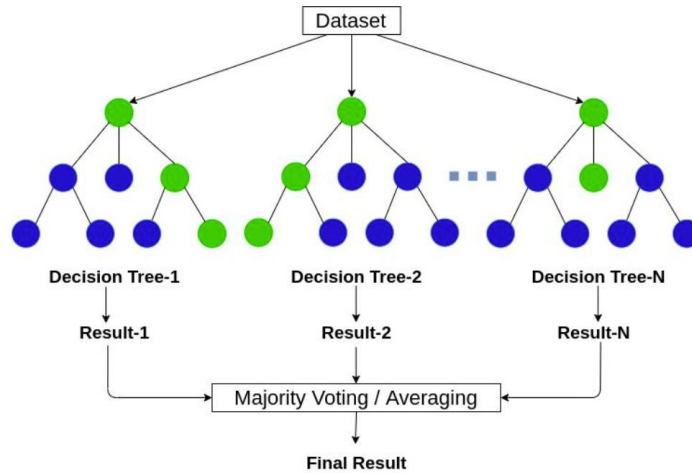
Where  $P_{X,Y}$  is a misclassification function for a given input  $X$ ,  $\mathbf{K}(\mathbf{X}, \mathbf{Y})$  is the edge function of random forest. As the number of trees increases, for all random variables  $\theta$ ,  $PE$  is convergent to

$$P_{X,Y}(P_{\theta}(h(\mathbf{X}, \theta) = \mathbf{Y}) - \max P_{\theta}(h(\mathbf{X}, \theta) = j) < 0) \quad (2)$$

Define RF generalization error satisfy:

$$PE \leq \rho^{(1-)} / s^2 \quad (3)$$

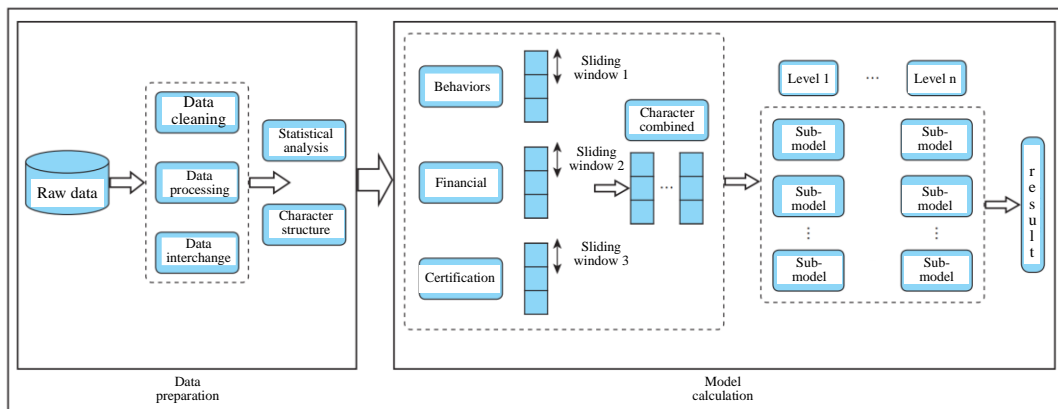
The generation process of the random decision tree in RF is as follows: when a subset of data is selected from the original data set, the probability that each sample is not selected is  $(1 - \frac{1}{N})^N$  ( $N$  is the total number of samples). Then  $K$  variables are randomly selected from the original set of variables to get the best segmentation point. After the above steps are repeated, the expected decision tree can be generated. VarSelRF algorithm is a variable (feature) selection method under the FRAMEWORK of RF algorithm. VarSelRF is mainly based on the reverse deletion strategy of OOB, and sorts according to the score of each feature in RF, and deletes a certain number of low-scoring features, so as to retain more important features and realize feature selection, which is especially applicable. In the case of large amount of process data and complex data characteristics. In conclusion, there are many classifications tree randomly in the forest. We will classify input samples, we need to input sample input to classify each tree make an image of the parable: the meeting among the forests, talk about whether an animal is mouse or squirrel, every tree should be independently express his points about the question, every tree that is voting. Hence, this algorithm can achieve better classification and prediction performance.



**Figure 1:** Flow chart of Random forest algorithm.

### 3.2 Gcforest Model

Deep forest is a new integrated learning algorithm proposed by Professor Zhou Zhihua and Dr. Feng Ji from Nanjing University. The algorithm has two core parts: multi-granularity scanning and cascade forest structure. In the stage of multi-granularity scanning, sliding Windows are used according to feature types to construct feature subsets to realize feature combination and enhance the utilization of multidimensional data information. Each layer of the cascaded forest structure is composed of multiple random forests as sub-models. According to the training results of each layer, the weight of the sub-model is adjusted, that is, the sub-model with high accuracy is given a larger weight, and the final result is obtained through integration and fusion. First, data cleaning and other pre-processing works are carried out on the original data, and then statistical analysis and feature engineering are carried out on the data set to complete data preparation. In the multi-granularity scanning stage of deep forest, the feature fusion of behavior information, financial information and authentication information is realized, and the integrated computing is realized in the cascade forest structure. Finally, the research framework of deep forest applied to credit risk assessment and early warning is shown in Figure 2.



**Figure 2:** Credit risk assessment and early warning research framework.

## 4 ANALYSIS OF RESULTS

### 4.1 Dataset Introduction

In conclusion, there are many classification trees randomly in the forest. We will classify an input sample, we need to input sample input to classify each tree. Make an image of the parable: the meeting in the forest, discuss whether an animal is mouse or squirrel, every tree should be independently published his views on the question, every tree that is voting. Hence, This algorithm can achieve better classification and prediction performance. The experimental data used in this paper are credit data from the Give Me Some Credit competition on Kaggle. Data acquisition of the web site is <http://www.kaggle.com/c/GiveMeSomeCredit/data> online lending and collecting the data for the bank to carry out the data, according to the second chapter is given in section 1 of the Internet financial righteousness, this data can be determined for the Internet financial credit data. The experimental dataset is given in Table 1:

Ordinal variable	Variable name	Type
1	SeriousDlqin2yrs	character type
2	Revolving Utilization Of Unsecured Lines	percentage
3	Age	integer
4	Number Of Time 30-59 Days Past Due Not Worse	integer
5	Debt Ratio	percentage
6	Monthly Income	real number
7	Number Of Open Credit Lines And Loans	integer
8	Number Of Times 90 Days Late	integer
9	Number Real Estate Loans Or Lines	integer
10	Number Of Time 60-89 Days Past Due Not Worse	integer
11	Number Of Dependents	integer

**Table 1:** Kaggle credit database data dictionary.

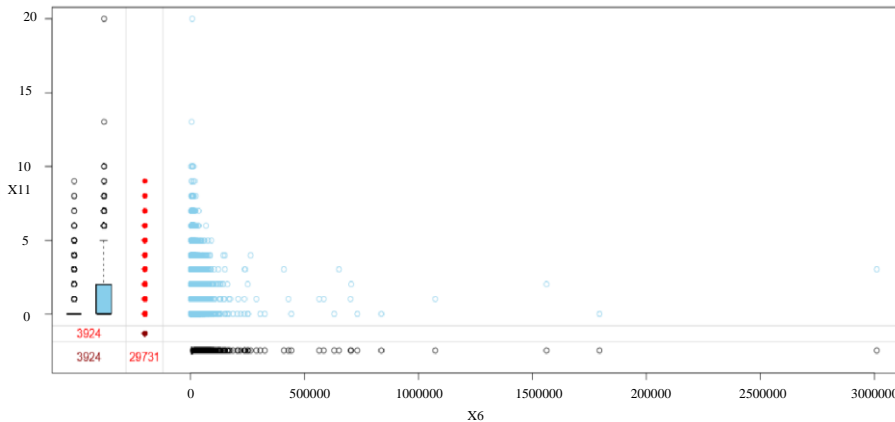
There are two data files on the website, training and testing dataset, among which the training dataset contains 150,000 sample and the test data set has 101,503 sample points. Each sample point contains 11 variables, including 10 characteristic variables for calculation and 1 classification variable. X1(SeriousDlqin2yrs) is a classification variable, one class is 0(non-defaulting customer) and one class is 1(defaulting customer). Because we do not know the value of the X1 classification variable of the test data set. In addition, the Give Me Some Credit competition has ended, the answer submission channel has been closed, and the predicted results of the model on the test set cannot be submitted. Therefore, the experimental data in this paper only uses the training data set on the Website of Kaggle, Table 1, to illustrate the characteristic variables in the Give Me Some Credit database, which is referred to as Kaggle credit data later.

The English variable names in the data dictionary are very long. In this paper, variable serial numbers are used to replace variable names. For example, X1 is used to replace SeriousDlqin2yrs. This will be more concise in the later model construction process. X1 variation is the classification variable of credit level. Customers who are more than 90 days overdue are defined as Bad (default) customers, and those who are less than 90 days overdue are defined as Good (non-default) customers. In order to match the formula derivation of RF, the value of X1 is changed here. The value of defaulting customers is changed from 1 to -1, and the value of high-quality customers is changed from 0 to 1. Among 150,000 sample points in the training set, there are 139,974 non-default sample points, accounting for 93.32%; There were 10,026 default sample points, accounting for 6.68%.

**4.2 Analysis of Experimental Results**

The kaggle credit data used in this paper is not complete, and there are several variables missing. In order to overcome the deficiency of deleting data directly, this paper adopts data filling method to fill missing data. According to the number of constructed filling values, missing data filling method can be divided into single value filling method and multiple filling method. Single value filling method only estimates one possible value for each missing value, and common interpolation methods include random interpolation method mean interpolation method and regression interpolation method. This method has two disadvantages such as changing the distribution of initial sample data and failing to explain the uncertainty of missing values. Multiple filling method can overcome the above two shortcomings, it is widely used in missing value filling.

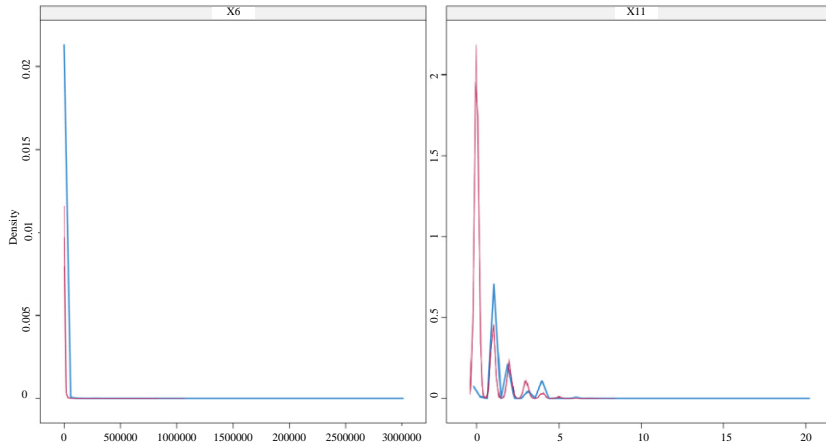
In a total of 120260 complete records, 25,800 sample points were missing variable X6, and 3924 sample points were missing variable X6 and X11. According to the analysis of variable angle of credit data, there are 3922 missing values in variable X11, with a missing ratio of 2.6%. There are 29731 missing values in variable X6, and the missing ratio is 19.8%. Now the analysis of variables X6 and X11 is focused. Figure 3 is the scatter diagram of two variables. The blue point in the middle is the complete record point, and the red point on the left corresponds to the known point X11 and the unknown point X6. There's no point where X6 is known and X11 is unknown indicates that There is no point where X6 is known and X11 is unknown. The red dot at the bottom and the intersection to the left corresponds to X11 unknown X6 unknown.



**Figure 3:** Scatter plot of variables X6 and X11.

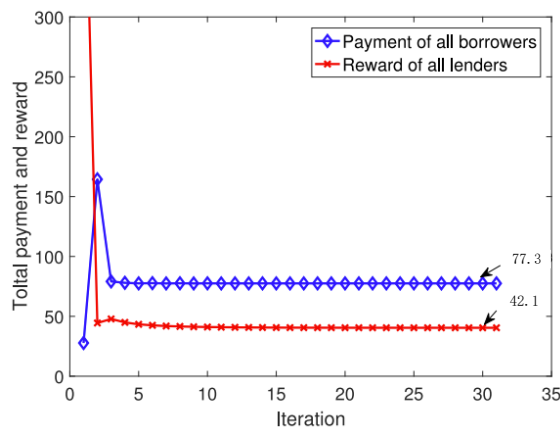
Figure 3 also shows the edge distribution of each variable. The blue boxplot on the left represents the edge distribution of 120,629 blue points (known points of both variables) of variable X11, and the red boxplot represents the edge distribution of 25,807 red points (unknown points of X6 and known points of X11) of variable X11. Similarly, the black dots at the bottom are the marginal

distribution of 120,629 blue dots (known points for both variables) for variable X6. The above steps can facilitate subsequent modeling. Figure 4 shows the density distribution curves of the original complete data and estimated data of the variable with missing values. The blue curve is the observed data curve of the variable, and the estimated data curve of the variable with red color is very similar to the blue curve, indicating that the estimated value of the missing value is very close to the real value.



**Figure 4:** Density distribution curve of observed and estimated values.

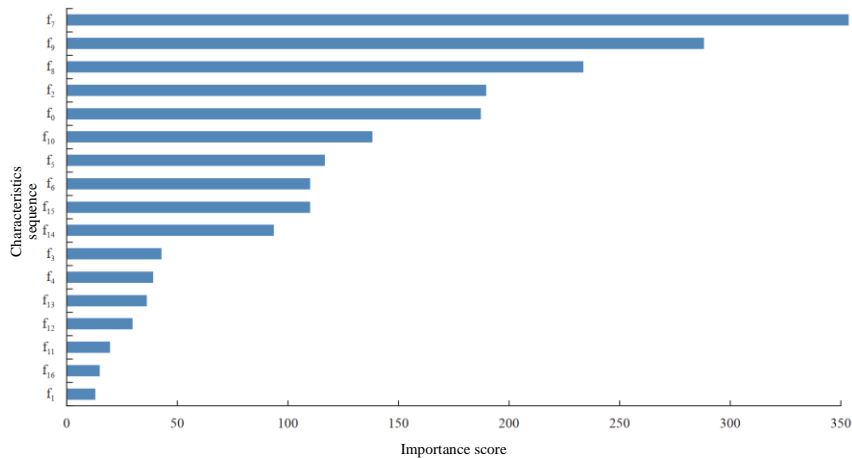
The user's authentication and financial information can be understood through the user's attribute loan records, but the repayment willingness and repayment ability of the user cannot be directly understood through this information. Therefore, feature extraction and feature construction should be carried out on the original data, and users' behavior habits can be understood indirectly through users' use of products and borrowing and returning records. This paper is based on the existing borrowing times, borrowing amount, borrowing success date "and other data to enrich user behavior information through statistical calculation. Finally, the authentication information, behavior information and financial information of users are classified into the characteristic indexes of credit risk assessment. Furthermore, you can see from Figure 5 that total payments and total returns have apparently stabilized in value.



**Figure 5:** Payments of all borrowers and rewards of all lenders.



After feature engineering is completed, feature importance can be measured according to The Times of feature splitting or information gain value. In this paper, the times of feature splitting, that is, the times of feature being used as splitting node, are selected to calculate the importance score of feature, as shown in Figure 6. According to the analysis of characteristic importance, loan interest rate, loan amount and total outstanding principal are important factors in financial information. In the behavioral information, the characteristic importance score of the number of successful repayment in history is the highest, and the number of successful repayment can indirectly reflect the behavior habits and performance ability of customers. In the authentication information, the account and the flow authentication is the important credential.

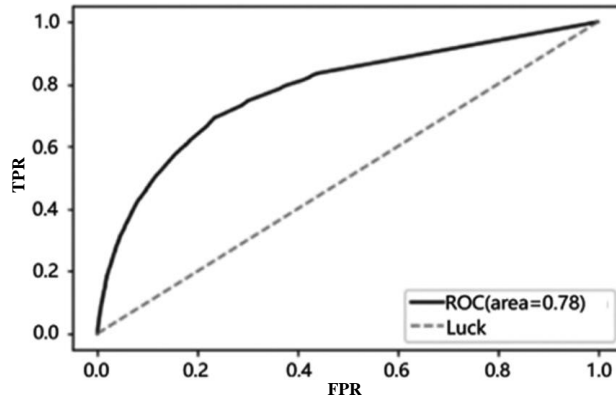


**Figure 6:** Test sample identification results.

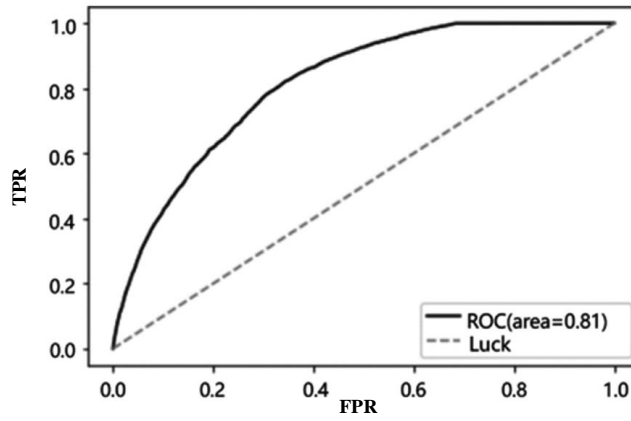
In the financial field, ROC curve, KS value and PSI index are often used to evaluate the accuracy, differentiation and stability of models respectively. ROC curve can balance accuracy and recall rate well, and is a comprehensive evaluation index of model accuracy. The area under the curve is the AUC value. The larger the AUC, the better the overall accuracy of the model. KS value represents the differentiation degree of negative and positive samples. The higher KS value is, the better the differentiation degree of samples is. The PSI stability index reflects the stability degree of the model. The lower the PSI value is, the higher the stability of the model is. In order to verify the application effect of the deep forest model, a control experiment was conducted on the same experimental data set according to the model and feature categories. Random forest, xgboost, and the deep forest model are adopted respectively in the dimension of financial information and authentication information. The ROC curve ratio of each model is shown in Figure 7(a)- Figure 7(c). The evaluation effect of each model is analyzed through the area under the ROC curve, that is, AUC value.

According to the comparative analysis in Figure 7(a), Figure 7(b), and Figure 7(c), under the same information dimension, deep forest has better overall prediction performance than random forest xgboost. By Comparing Figure 7(a), and Figure 7(c), it can be seen that the introduction of behavioral information can improve the accuracy of the model to a certain extent in the same deep forest algorithm. Finally, the prediction performance of the proposed method is given in Figure 8.

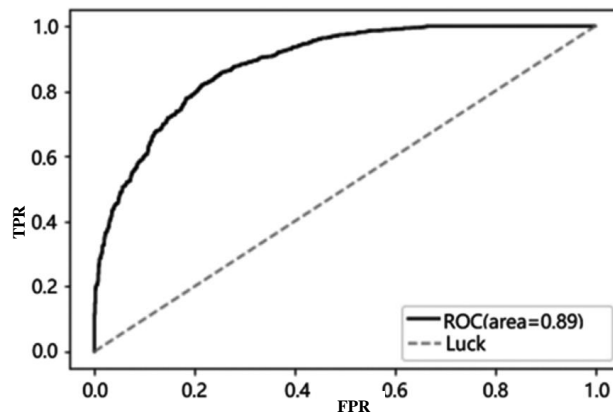
In this paper, credit risk assessment of Internet finance is studied, and a credit risk assessment model is established by using deep forest algorithm. Deep forest model has good overall prediction effect in credit risk assessment. The supplement of behavioral information can make the model not only guarantee the accuracy of prediction, but also effectively improve the ability to distinguish positive and negative samples, which has good applicability.



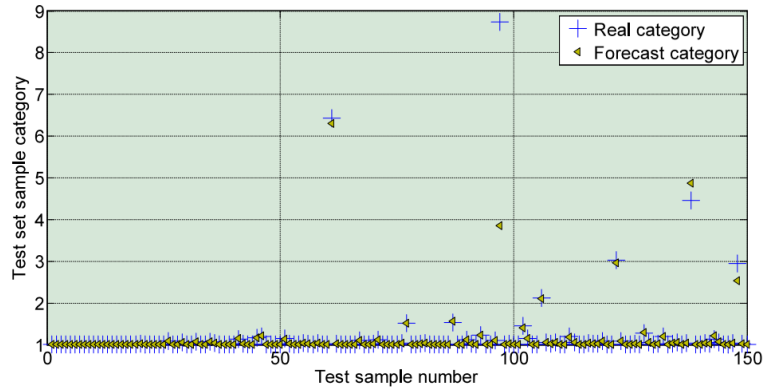
**Figure 7(a):** ROC curve of Random forest.



**Figure 7(b):** ROC curve of xgboost.



**Figure 7(c):** ROC curve of xgboost deep forest model.



**Figure 8:** The prediction performance of the proposed method

Through numerical experiments and comparison of results, the method in this paper can realize effective assessment and early warning of credit risks of Internet finance by using the feature combination of multidimensional data information, and realize multi-dimensional supervision by using data mining, which can improve the early warning ability of credit risks.

Guicang Yang, <https://orcid.org/0000-0002-7905-4743>

## REFERENCES

- [1] Glova, J.; Sabol, T.; Vajda, V.: Business models for the internet of things environment, *Procedia Economics and Finance*, 15, 2014, 1122-1129. [https://doi.org/10.1016/S2212-5671\(14\)00566-8](https://doi.org/10.1016/S2212-5671(14)00566-8)
- [2] Du, G.; Liu, Z.; Lu, H.; Application of innovative risk early warning mode under big data technology in Internet credit financial risk assessment, *Journal of Computational and Applied Mathematics*, 386, 2021, 113260. <https://doi.org/10.1016/j.cam.2020.113260>
- [3] Zhou, G.; Zhang, Y.; Luo, S.: P2P network lending, loss given default and credit risks, *Sustainability*, 10(4), 2018, 1010. <https://doi.org/10.3390/su10041010>
- [4] Sun C.: The Study on Calculation of Credit Risks in China's Internet Finance. *Management & Engineering*, 22, 2016, 1-43. <https://doi.org/10.5503/J.ME.2016.22.009>
- [5] Millard, K.; Richardson, M.: On the importance of training data sample selection in random forest image classification: A case study in peatland ecosystem mapping, *Remote sensing*, 7(7), 2015, 8489-8515. <https://doi.org/10.3390/rs70708489>
- [6] Deng, C.; Zu, M.: A new co-training-style random forest for computer aided diagnosis, *Journal of Intelligent Information Systems*, 36(3), 2011, 253-281. <https://doi.org/10.1016/j.cam.2020.113260>
- [7] Kim, S.; Kwak, S.; Ko, C.: Fast pedestrian detection in surveillance video based on soft target training of shallow random forest, *IEEE Access*, 7, 2019, 12415-12426. <https://doi.org/10.1109/ACCESS.2019.2892425>
- [8] Zhang, J.; Yin, Z.; Chen, P.: Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review, *Information Fusion*, 59, 2020, 103-126. <https://doi.org/10.1016/j.inffus.2020.01.011>
- [9] Belgiu, M.; Drăguț, L.: Random forest in remote sensing: A review of applications and future directions, *ISPRS journal of photogrammetry and remote sensing*, 114, 2016, 24-31. <https://doi.org/10.1016/j.isprsjprs.2016.01.011>

- [10] Maas, A.; Rottensteiner, F.; Heipke, C.: A label noise tolerant random forest for the classification of remote sensing data based on outdated maps for training, *Computer Vision and Image Understanding*, 188, 2019, 102782. <https://doi.org/10.1016/j.cviu.2019.07.002>