# Financial Risk Early Warning Model of Internet Financial Companies Based on SMOTE-Random Forest

Yan Qu[1, *], Pinghong Ji[2]
[1]Qinghai Communications Techical College, Xining 810003, China
[2]Qinghai Province Transportation Holding Group Co., Ltd., Xining 810001, China
*Corresponding email: quyan_lunwen@163.com

**Abstract**

Taking my country's Internet finances, based on the home and abroad, combined with the characteristics of Internet financial companies, the SMOTE algorithm is used and combined with random forests to establish the financial management of Internet financial companies. Risk early warning model. Research shows that the random forest early warning model has stable recognition accuracy and good prediction performance, so it has a wide range of practical value. The improved SMOTE algorithm based on PCA can realize the equalization of unbalanced data sets and use random forest as a classifier to classify and predict geological data. Because the noise data in the original data set may cause the change of the data distribution after interpolation, it is proposed to combine the PCA algorithm and the SMOTE algorithm, first perform noise reduction and dimension reduction, and then perform data interpolation to improve the classification performance of imbalanced data sets. My country's Internet financial listed companies conduct experiments on research samples; algorithms can better improve classification accuracy and provide new ideas for the classification and prediction of unbalanced data.

**Keywords**

Smote algorithm, Random forest, Financial risk early warning

## Introduction

Internet finance is a tool that relies on electronic payment and third-party platforms to realize a new financial model of financing [1]. Compared with the traditional financial business model, Internet finance is obviously different [2].

It utilizes powerful back-office and information technology to have a wider business scope and more optimized services and products [3].

However, due to the short development time of my country's Internet finance and prominent network insecurity problems, this has led to the relatively backward risk response of Internet finance, and various supervision and management systems are not perfect, which fully demonstrates the development of my country's Internet finance. There is more room for healthy development [4]. Since the rise of finance, traditional finance has been leading the development of finance, but it has always been limited by its own characteristics [5]. A traditional financial institution, such as a bank, has a fixed transaction location, but cannot realize transactions in different locations [6].

It can handle business for users, but it needs to charge users a certain amount of service [7]. However, services have also been innovated and developed.

Through the network platform, traditional financial transactions can be simplified, such as order transactions, transaction information query, online customer service, etc., which greatly reduces the cost of financial institutions cost.

The operation mode of the Internet-based platform of traditional financial business includes one is an independent operation mode, which mainly

cooperates with other enterprises through the Internet; the other is a platform operation mode, which is mainly related to e-commerce websites. Mainly based on website operation [8].

The transaction process of the financial service model based on the Internet platform is completely dependent on the Internet [9]. Its online trading platform is huge and has many users, from companies to individuals [10]. On this network platform, network information is constantly developing, which allows investors and financiers to conduct financial transactions conveniently [11]. This model completely changes the traditional financial model, trades in a brand-new way, and relies entirely on online transactions, such as P2P network credit [12]. This model is free and convenient to trade and is a good choice for investors and financing investors [13].

The new Internet financial model is not actually a financial business, it is a third-party payment platform, but it is Internet-based financial support [14]. There are various types of financial services, both offline and online, and offline financial service functions are transformed into the Internet [15]. When customers receive information about various transaction services, their understanding of transaction services also increases [16]. This kind of integration and coordination between financial institutions and customers has greatly improved the efficiency of financial support for Internetization [17]. This efficient Internet-based financial support model will also reduce the information asymmetry between institutions and customers, promote innovation in Internet financial services, and optimize the structure of Internet finance [18].

Because in Internet finance, the information protection of both financial institutions and users is more stringent, and the flow of personal assets and funds cannot be completely transparent [19]. At this time, Internet finance is equivalent to online banking, which limits the role and power of banks [20]. It makes the bank unable to grasp the flow of funds in detail, which affects the bank's capital operation and affects the bank's capital structure. Information disclosure is an important part of Internet financial security. Without information disclosure, there will be no security in the Internet financial environment. To protect the security of users' personal information, the information disclosure of Internet finance needs to be highly transparent.

However, due to the lack of transparency in information disclosure, the problem of information asymmetry in the process of Internet financial transactions is very prominent, and some companies use this information asymmetry to defraud the interests of customers, which affects the Internet financial industry. healthy development. Ambiguity of its regulatory body has always existed. Multiple departments are endowed with direct supervision and management capabilities, and each institution has different management responsibilities, but there are overlapping parts, which also restrict the rapid development of Internet finance. Because Internet finance has the characteristics of complicated and easy access to information, and it can also for transactions, which leads some illegal users to conduct illegal operations through loopholes in Internet laws, for example, stealing customers through non-compliant financial ports Information conducts cybercrime, which ultimately damages the interests of customers and disrupts the order of my country's Internet finance. The current lack of self-supervision of Internet financial companies and platforms has directly resulted in many violations and illegal operations not being detected, warned and dealt with in a timely and effective manner.

From an essential point of view, there is a certain connection between Internet finance and finance, which inevitably involves the core topic of risk control. There are three main risks involved in the Internet finance industry: first, market risk; second, network technology risk; third, management and operational risk.

In terms of market risk, it is divided into two points: First, the government's macro-policy intervention in finance, such as monetary policy. Secondly, the corresponding risks brought by the market's own

competition and bubbles to Internet finance. Because China's capital market is not developed enough, compared with other European and American countries, my country's market is very prone to bubbles and vicious competition. For cyber risk, it is the centre of Internet financial risk, mainly because cyber risk is closely related to Internet security and reliability. Internet finance is an emerging industry based on the Internet. If the security problems existing on the Internet cannot be solved and controlled, it will have a negative impact on the entire financial industry. In addition, in terms of managing operational risks, it is the easiest to control and solve, mainly including reputation risk, specific operational risk and customer credit risk.

## State of the art

### *Overview of internet finance*

In a broad sense, Internet finance generally refers to Internet financial tools with cloud computing, cloud payment and search engines as the core, which can not only realize the integration and interoperability of funds but also use information exchange and payment as intermediary business. Under normal circumstances, Internet finance generally refers to the construction of open and new financial formats and service systems with cloud computing and big data technology as the background, with the Internet as the core, and based on the Internet platform, and formed on this basis. financial activity.

The characteristics of Internet finance are as follows:

(1) Innovation. Under the background of science and technology and the information age, people's daily life has undergone relatively great changes. Whether it is daily work or entertainment, the Internet has been widely used, and the financial industry is no exception, which makes Internet finance emerge as the times require. In fact, Internet finance relies on social networks, cloud computing and other emerging technology platforms to carry out financing, payment and other businesses, which is a financial innovation model developed based on traditional financial institutions.

(2) High efficiency. Internet finance generally uses computers to carry out related business operations, and the operating procedures are gradually becoming standardized and standardized, which greatly improves the efficiency of business processing.

(3) Low cost. Internet finance can use the network platform to match, trade, identify and price information on both sides of the supply and demand of funds, and has the characteristics of no transaction costs and no monopoly profits. At the same time, Internet finance can also get rid of the operating costs and capital investment required to open business outlets.

### *Common risks of internet finance*

In the development process of the information age, Internet financial risks are two-sided. There are not only traditional financial risk types, but also Internet risk types. In the development process of Internet finance, the common types of risks are as follows:

(1) Credit risk. In the process of development of Internet finance, it will not only encounter various credit risks existing in traditional finance, but also new credit risks. The payment and transaction of Internet finance are generally completed through the network platform. This form will lead to the lack of the identity of the trader, increase the difficulty of information verification, and further increase the credit risk due to information asymmetry. My country's existing Internet financial credit risk prevention system is not perfect, and it relies too much on guarantees and rigid payment, which further exacerbates Internet financial credit risks.

(2) Technical risk. Since Internet finance can simply be regarded as the combination of the financial industry and the Internet industry, both customers and enterprises have paid attention to capital gains. In the process of choosing investment, most people value the quality of financial services, and therefore, put forward higher requirements for Internet finance. At this time, the insecurity of the Internet may cause serious risk losses, and the defects of Computer and Internet technology will also increase the risk of viruses infringing on the information system, thereby inducing technical risks.

(3) Manage risks. Since the Internet is borderless,

there are certain differences in the management systems of different countries, it is very likely to cause management conflicts. At present, Internet finance needs continuous improvement and improvement in many places. For example, the service system on the Internet cannot meet the basic needs of customers, the main responsibilities are not clear, and the obligations and rights of both parties to the transaction are unclear, etc. All these will increase financial risks. incidence.

***Reasons for the formation of internet financial risks***

The current progress of interconnected technology, the reform and the change of customer groups have promoted the rapid. At the same time, the combination of the financial industry and national policies reflects the concept and value of inclusive finance, satisfying all Potential needs of customer groups. Internet finance can services, combine finance and the Internet to meet the needs of the public for new types of finance, solve the problems of high financing threshold and high financing costs, and indirectly promote the development of the real economy.

In order to avoid the occurrence of Internet financial risks, it is necessary to do a good job in the prevention of Internet financial risks according to the actual situation, and to refine and improve the Internet financial supervision system. First, it is necessary to analyze the causes of Internet financial risks and gradually analyze and put forward reasonable suggestions. To analyze Internet financial risks, we can analyze the causes of Internet financial risks from macro, meso and micro perspectives, and risks from the perspectives of national attitudes and regulatory policies, the characteristics of the Internet industry, and the individual level of the public. Explore.

(1) Macro level

In the process of analyzing the risks of Internet finance, it is necessary to give a comprehensive and systematic analysis of the development process of the Internet and consider my country's attitude towards Internet regulation and the basis of Internet development from a historical perspective and seek suggestions from existing experience.

Analysis from a macro perspective shows that my country's Internet finance emerged relatively late, developed rapidly, and achieved rapid development and improvement in a short period of time. However, there is a certain gap between the formulation of relevant laws and regulations and the development of the industry. Although the Chinese government has issued many policies and suggestions, it has further clarified the risks and incentives of Internet finance, and put forward corresponding countermeasures, but under the rapid development of Internet finance, its own shortcomings and drawbacks are difficult to solve, and related loopholes are not easy to solve. The lack of improvement and treatment, coupled with the lag of national policies and laws and regulations, makes Internet finance difficult to supervise on a macro level, which provides conditions for the formation of Internet financial risks.

(2) Meso-level

Internet finance draws on the basic forms of the Internet and finance, uses the Internet as a tool to spread and develop finance, innovates business, technology, and transaction methods, and increases the number of customers. The main functions of Internet finance are still financing, price discovery, payment and settlement, etc., which are completely consistent with traditional finance. Therefore, it also has great risks. Even due to technical limitations and imperfect management systems, it faces local risks. much larger than traditional finance. Therefore, at the mesa level, we should combine the characteristics of the Internet industry to explore the reasons for the formation of Internet financial risks. Under normal circumstances, the Internet industry itself is virtual, and it is easy to amplify the risk of trust in exchanges and transactions; Internet technology is fragile and easily attacked by network hackers and viruses, and finance involves a large amount of corporate information and personal information. Improvements can easily leak information and exacerbate systemic financial risks.

Internet encryption technology may impact the original ecological pattern and lead to the reconstruction of the original system's credit.

(3) Micro level

Through the analysis of Internet financial risks from the micro level, it is found that it generally analyses the public's personal behaviour in all aspects and obtains the incentives for Internet financial risks. It is found that the following factors will aggravate financial risks: 1. Lack of relevant knowledge. With the continuous capital market, although it has achieved certain results, it is still in its infancy. The public's own capabilities are limited, and their limited financial management knowledge leads to an incomplete understanding of the risks and benefits of the capital market. My country's basic deposit interest rate is low, the public to increase income through assets. Therefore, when high-yield Internet financial products appear, most of the public will use bank deposits to purchase high-yield Internet financial products but ignore the Internet. The risks of financial products cause most of the public to have low returns or even losses. The herd effect of the public.

Internet finance shows its good development by attracting some public investment online. Although some public have some doubts about its income, it will be purchased and recommended by the surrounding public, pushed by relevant APPs, and updated in real time. and other influences, making more customers join the crowd buying Internet financial products. To sum up, under the combined action of various factors from macro, mesa and micro perspectives, Internet financial risks are gradually generated and formed.

## Methodology

### *Weighted oversampling*

In the classic oversampling class sample is the same, which will generate many samples with little value for class distinction. Because the samples closer to the center of the category can better represent the characteristics of this category. These two samples contain more classification information and is more valuable to classification. It is hoped that the class and boundary can be fully utilized. the samples close to the class center and boundary can generate more new samples. For this, the Euclidean distance of each minority class sample relative to the remaining minority class samples is used to determine the relative position of each sample, assigning a different weight. Make Samples close to the class center and class boundary have greater weights.

Weight calculation steps

Let the contains C features. Each sample and other samples, as shown in formula (1):

$$D_{ij}(x_i, x_j) = \sqrt{\sum_{k=1}^{C}(x_{i,k} - x_{j,k})^2} \quad (1)$$

In the formula, i=1, 2…, M, j=1, 2…, M, i≠j, Di, j (xi, xj).

Calculate the sum Di of the distances from the sample xi to other samples. The larger Di is, the closer xi is to the boundary, and the smaller Di is, the closer xi is to the center, as shown in formula (2):

$$D_i = \sum_{j=1, j\neq i}^{M} D_i(x_i, x_j), i = 1,2,\cdots,M \quad (2)$$

Normalize Di, as shown in formula (3):

$$ND_i = \frac{D_i - D_{min}}{Dmin_{max}, i = 1,2,\cdots,M} \quad (3)$$

Calculate RNDi, which is the absolute value of the difference between each element in ND and the mean value of ND. The larger the RNDi.

The number of new samples should be larger, as shown in formula (4):

$$RND_i = \left| ND_i - \frac{\sum_{i=1}^{M} ND_i}{M} \right|, i = 1,2,\cdots,M \quad (4)$$

Calculate the weight of each sample, as shown in formula (5):

$$W_i = \frac{RND_i}{\sum_{i=1}^{M} ND_i}, i = 1,2,\cdots,M \quad (5)$$

Wi is the weight value of the ith sample. multiplied by this weight is the final number of new samples generated from this sample.

Weighted SMOTE algorithm steps:

(1) For the ith sample xi neighbor samples.

(2) For the ith sample xi, randomly select Nw=[N×Wi] samples {x1, x2…, xn} from its k nearest neighbor samples and is rounded down.

(3) Among the Nw samples selected in (2), generate Nw new samples about the minority sample xi according to formula (6):

$$x_{new} = x_i + rand(0,1) \times |x_i - x_n| \quad (6)$$

The schematic diagram of weighted SMOTE is shown in Figure 1.

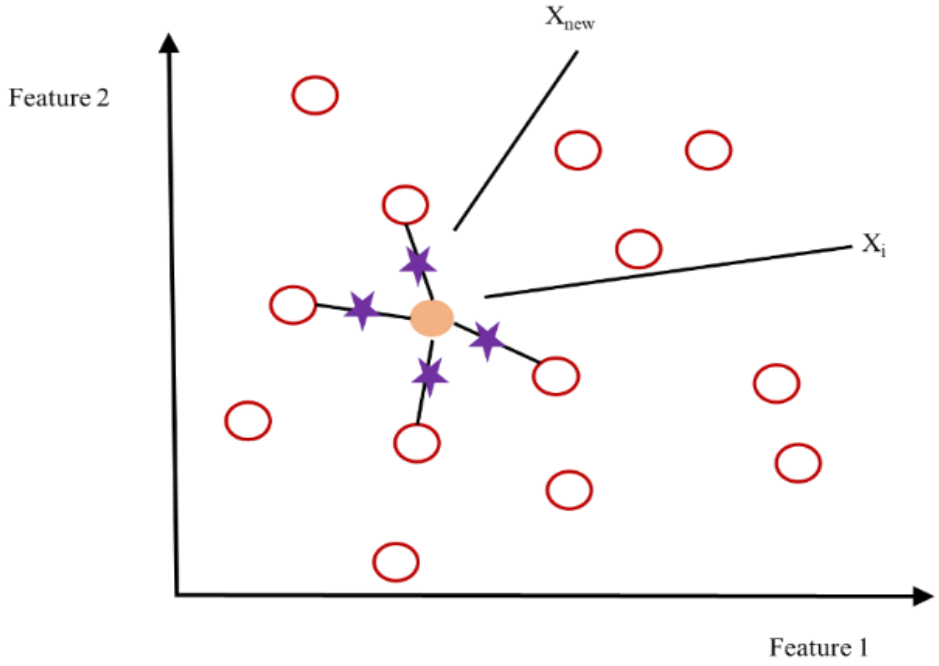


Figure 1. Schematic diagram of weighted SMOTE.

***Weighted random forest***

(1) Random forest algorithm

The random proposed by BreimanL. It is an integrated machine learning method. The essence is to combine the Bagging algorithm and the randomsubspace algorithm to build a classifier composed of multiple uncorrelated decision trees. To avoid the shortcoming that decision trees are prone to over-fitting, the training samples adopt the Bootstrap technique.

From the original data set, N samples are randomly selected and repeatedly selected as the training set of a tree, and each time the training set is randomly selected. A part of the sample features constructs a decision tree, and each decision tree is not pruned during the training and growth process, and finally the final result of the classifier is determined by voting. For example, for a trained random forest model, the test set is X, the number of categories is C, and the number of decision trees is T, then the output of the model is formula (7):

$$H(X) = argmax\{\sum_{t=1}^{T} I(h_t(X) = y)\} \quad (7)$$

Among them, and I(.) is an indicator function.

The Seg, glass and wine data sets in the UCI database are used to verify the algorithm, and the selected data sets show obvious imbalance. Algorithm tests are performed on the Seg, glass and wine datasets, and the datasets are subjected to random forest classification on the direct initial data, the SMOTE processed data, and the PCA-SMOTE processed data, and the AUC values of the experimental results (described in detail in Section 3.3) are analyzed. Analysis (Figure 2), Seg, glass and wine data sets after PCA-SMOTE processing, the classification results are better than the classification results after SMOTE processing. Therefore, the random forest algorithm based on PCA-SMOTE algorithm has better classification performance.

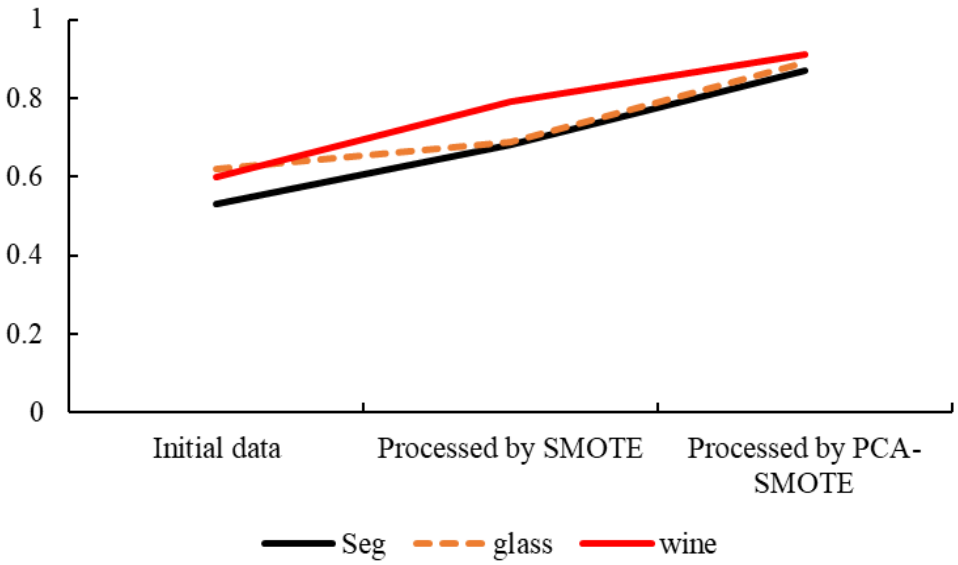


Figure 2. The AUC values of the experimental results of the Seg, glass and wine datasets.

(2) Weighted random forest algorithm

It can be seen from formula (7). It will further affect the classification effect of the decision tree, making some trees with poor effect cast the wrong number of votes, thus affecting the classification ability of the random forest. For this reason, this paper proposes a weighted random forest model. The main method is to evaluate the classification effect of the decision tree in the decision tree training stage and assign weight to each tree. When the random forest algorithm votes, each tree must be multiplied by the corresponding Weight value, which can reduce the influence of the decision tree with low training accuracy on the entire model. Therefore, the model output in Equation (7) is rewritten as formula (8):

$$\widehat{H}(X) = argmax\{\sum_{t=1}^{T} I(h_t(X) = y) \times w_t\} \quad (8)$$

Where wt is the weight value of t-th. decision tree. Replacement and the number of samples is equal to the number of samples in the original training set. Since there is This part is called the bag The number of out-of-bag samples is usually one third of the original number of samples. The out-of-bag samples are used as the test set to make the classification performance better. have greater weight.

For classifiers using imbalanced data, the commonly used evaluation metric of classification accuracy is not a good measure of classification ability, because it only considers the situation of correctly classified samples and considers that the classification errors of the majority class and the minority class are equally important. Therefore, the Kappa coefficient (Kappa Coefficient, CK) is used to evaluate the overall classification ability of the decision tree. CK is an index proposed by Cohen et al. in 1960 to evaluate the degree of consistency of judgment. It also considers various missed and misclassified samples. Represents the ratio of classification and completely random classification to produce error reduction, and its calculation result is (−1, 1), but usually CK falls at (0, 1), and the larger the CK value, the more consistent the predicted result and the actual result. The higher the sex, the better the classifier performance. The calculation of CK is shown in formula (9):

$$CK = \frac{ACC - CK_C}{1 - CK_C} \quad (9)$$

Among them, ACC (accuracy) is the classification accuracy, indicating the actual consistency ratio of the classification, and CKc is the accidental consistency ratio of the classification.

In the confusion matrix, TP (TruePositive) indicates the actual positive class, and the prediction is also positive class. As shown in formulas (10) and (11):

$$ACC = \frac{TP + TN}{TP + FN + TN + FP} \quad (10)$$

$$CK_C = \frac{(TP+FN)(TP+FP)(FP+TN)(FN+TN)}{(TP+FN+TN+FP)^2} \quad (11)$$

To assign larger weights to more capable classifiers, literature studies have shown that: if a set of independent classifiers L1, L2..., LM are independent of each other, and the accuracy is p1, p2..., pM, then each the relationship between the classifier weight and the corresponding accuracy is shown in formula (12):

$$w_t \propto ln\frac{p_t}{1-p_t}, t = 1,2,\cdots,M \quad (12)$$

Replace pt in formula (12) with CK. Since the value range of CK is (−1, 1), formula (12) is rewritten as formula (13):

$$w_t^{CK} = ln\frac{1+CK_t}{1+CK_t}, t = 1,2,\cdots,M \quad (13)$$

For unbalanced datasets, classification accuracy cannot comprehensively evaluate the classification effect, specificity, CK, and G-mean are used to evaluate the classification accuracy. The specificity is used to evaluate the correct rate of classification of minority samples. CK and Gmean are used to evaluate the overall classification performance of the imbalanced dataset. Which maximizes the two classes while maintaining the balance of the classification accuracy of the majority class and the minority class. The relationship is shown in formulas (14) and (15):

$$specificity = \frac{TN}{FP+TN} \quad (14)$$

$$G-mean = \sqrt{\frac{TP \times TN}{(TP+FN)(TN+FP)}} \quad (15)$$

**Resylt analysis and discussion**

***Experimental data and environment***

This paper selects a total of 63 Internet financial listed companies including 3 ST companies and 60 normal operating companies as the research objects and uses the financial data of each quarter of 63 Internet financial companies from 2017 to 2019 as the research sample. After deleting some missing values, a total of 752 sets of data were obtained, including 32 sets of ST companies and 720 sets of normal companies. ST company refers to a company that has been specially treated by the stock exchange because the company has suffered losses for two consecutive years, which can be regarded as having a high financial risk (Rise database).

When selecting financial indicators, this paper first refers to the research of Zhao Nan et al. The 18 indicators in this paper have passed the significance test; then 23 financial indicators selected by Yang Shu's and Wang Leping are considered; finally, combined with the random forest itself algorithmic properties. This paper finally uses a total of 27 financial indicators in seven categories as research variables. These 27 financial indicators reflect the per-share indicators, operating capacity, profitability, solvency, cash flow, capital structure and growth capacity of Internet finance companies respectively. It can fully reflect the financial status of Internet financial companies.

Because the data of 3 ST companies and the data of 60 normal companies in the early warning indicators are relatively serious unbalanced data, in order to solve the impact of unbalanced data on the random forest model, this paper uses the SMOTE algorithm to balance the unbalanced data. After that, random forest is used to carry out the data of the division. The test set contains the data of 220 groups of normal companies and the data of 12 groups of ST companies. After the data is balanced by the SMOTE algorithm, the new training set contains data from 270 groups of normal companies and 200 groups of data from ST companies, and the ratio is close to 1:1.

The significance of variables was analyzed using the reduction of average accuracy and the reduction of average impurity, respectively. The results show that the net profit margin of sales, net assets per share and growth rate of net assets are the top three in the two-importance analysis, and the importance is higher. Therefore, for enterprises, they should focus on these three financial indicators to accurately reflect the company's financial status.

***Experimental results and analysis***

From Table 1, we can see that the overall prediction accuracy of the random forest model constructed

from unbalanced data reaches 95.27%, while the prediction accuracy for ST Company is only 46.17%. According to the random forest model, the prediction accuracy of ST company is as high as 76.41%, and the overall accuracy is 97.35%. Judging from the prediction results, this prediction model, as the financial risk early warning model of Internet financial companies, is an ideal financial risk early warning model with good stability and practical value.

At the same time, the overall accuracy of the model is further enhanced, showing that the optimized random forest model not only reduces the bias caused by data imbalance but also strengthens the reliability of financial risk assessment. Therefore, it provides strong technical support for enterprises and regulatory institutions in preventing potential financial crises.

Table 1. Prediction results of training samples and test samples.

| Group | | Category | Predicted value/piece | | Accuracy/% |
|---|---|---|---|---|---|
| | | | Normal | ST | |
| Unbalanced | Actual value | Normal | 212.62 | 4.38 | 97.98 |
| | | ST | 6.46 | 5.54 | 46.17 |
| | Total | | 219.08 | 9.92 | 95.27 |
| Balance | Actual value | Normal | 213.76 | 3.24 | 98.51 |
| | | ST | 2.83 | 9.17 | 76.41 |
| | Total | | 216.59 | 13.41 | 97.35 |

The training set and the test set are obtained, and the PCA-SMOTE algorithm is applied to the training set to achieve a data balance state, and the random forest is used as the classifier for classification. At the same time, direct sample classification, SMOTE algorithm post-processing, PCA algorithm post-classification are performed on the training set, and the results are compared with PCA-SMOTE classification, including the majority class misjudgment rate, minority class misjudgment rate, Fmeasure value, AUC value, Gmean value and ROC curve, as shown in Figure 3.

If the original data set is directly classified by random forest without any operation, the classification effect is relatively poor, Gmean = 0.783, AUC = 0.53; after the SMOTE algorithm balances the data set, the classification result Gmean = 0.8, AUC = 0.67; and after PCA, after the SMOTE algorithm balances the dataset, the Gmean=0.962 and AUC=0.90 of the classification result.

The analysis shows that the classification effect of the PCA-SMOTE algorithm is significantly higher than that of the other two algorithms, which indicates that the improvement of the algorithm proposed in this paper is valuable.

During the experiment, the change of parameters will also cause large fluctuations in the results of SMOTE and random forest. In the random forest algorithm, ntree = 500 is used in this experiment. Since the new samples in the SMOTE method are interpolated, non-real samples, perc.over and perc.under should be appropriately selected to avoid the degradation of the classification quality of the dataset.

Reasonably improve the classification accuracy. Therefore, careful parameter tuning is essential to ensure the stability and reliability of the experimental results. By selecting appropriate values for perc.over and perc.under, the risk of overfitting can be reduced while maintaining the representativeness of the synthetic samples. This optimization not only enhances the robustness of the PCA-SMOTE algorithm but also provides a more accurate and practical solution for financial risk early warning in imbalanced datasets.

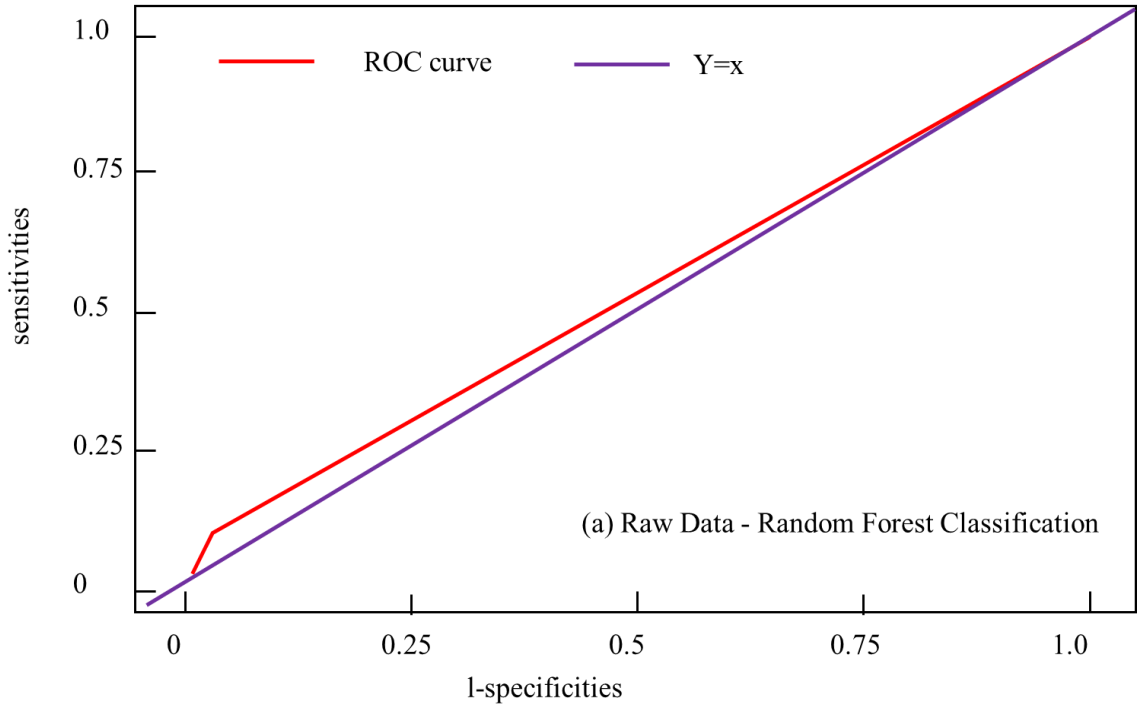


Figure 3. Actual output and expected output.

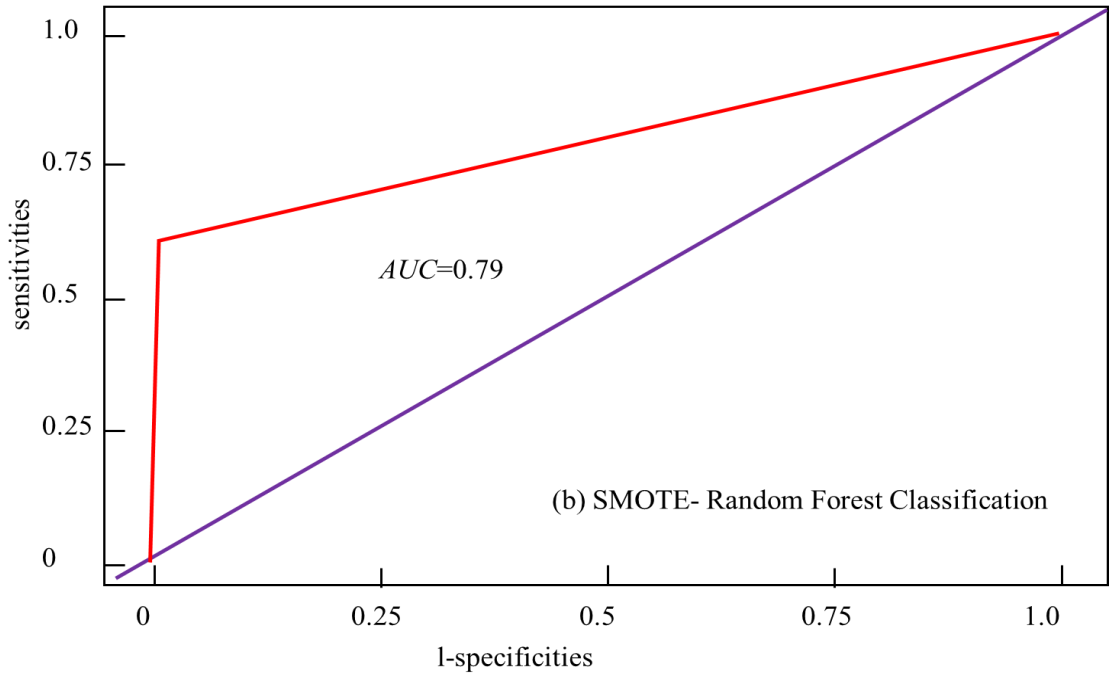


Figure 4. Output value of the test sample.

After feature selection, the random forest. In the process of model building, two parameters, the maximum number of features, need to be determined. Figure 5 shows the error distribution of models with different number of decision trees. the error of the model gradually decreases. When the number of decision trees is equal to 20, the error in the model is basically stable. For the sake of safety, the decision-making is determined. The number of trees is 40. predicted as complaint with the decision tree, and the green curve represents the change of the error predicted as non-complaint with the decision tree. This indicates that increasing the number of decision trees beyond a certain threshold does not significantly improve model performance, but instead increases computational cost and complexity.

Therefore, setting the number of decision trees to 40

not only ensures the stability of the model but also balances efficiency and accuracy.

This choice provides a reliable foundation for subsequent experiments and enhances the robustness of the classification results. This choice provides a reliable foundation for subsequent analysis and helps to improve the robustness of the prediction results.

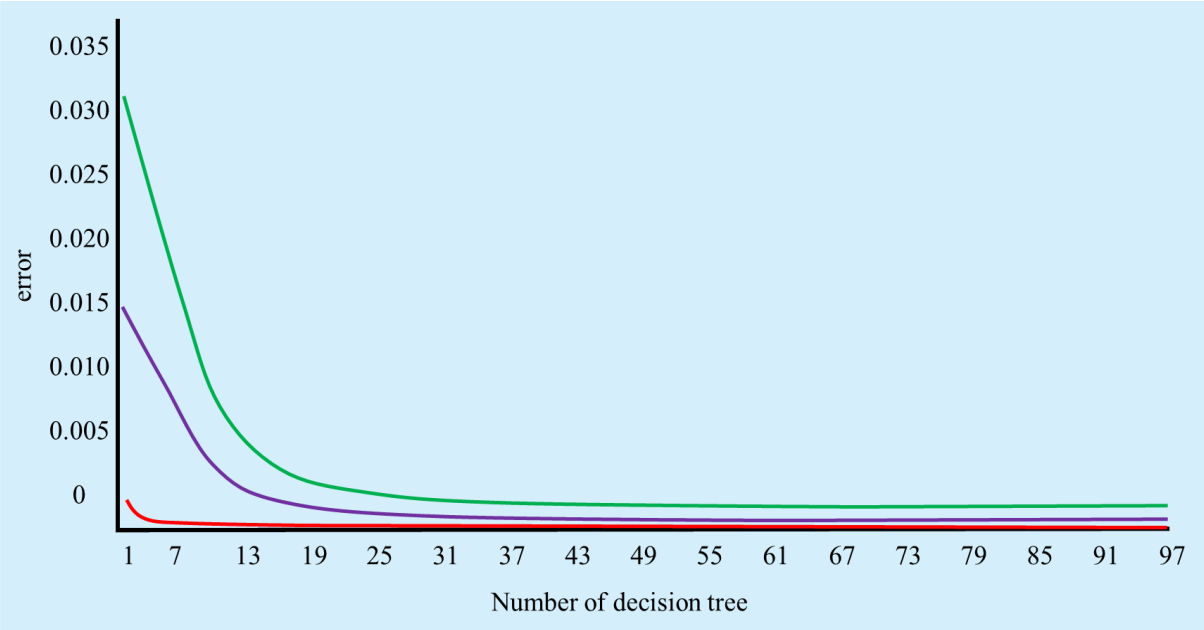


Figure 5. Error distribution of models with different number of decision trees.

Algorithms formed by different combinations of SMOTE (SM), weighted SMOTE (WSM), random forest (RF), and weighted random forest (WRF), namely SM+RF, WSM+RF, SM+WRF and the algorithm in this article (WSM+WRF).

For comparison, experiment the classification results of these four algorithms on the vehicle0 dataset. In the data set, 70% of the majority class samples and minority class samples are taken as the training set and 30% as the test set. Two common parameters in the random forest algorithm are the number of deciding trees (ntree) and the number of randomly selected features (mtry). In this experiment, ntree and mtry are carefully adjusted to evaluate their impact on model performance and to ensure a fair comparison among the four algorithms. With the increase of ntree, the classification error of random forest will tend to be stable. Since random forest will not overfit, set ntree large enough. mtry is used to reconcile the balance between classification performance and diversity, and mtry is set as the square root of the total number of features. Figure 6 and Figure 7 show the specificity and G-mean comparison results of the four algorithm combinations on the vehicle0 dataset. It can be seen from the figure that when the number of decision trees in the random forest algorithm increases to about 100, the specificity and Gmean tend to be stable. In the performance of these two indicators, the WSM+RF algorithm is better than the SM+RF algorithm, the WSM+WRF algorithm is better than the WSM+RF algorithm, and the SM+WRF algorithm is better than the WSM+RF algorithm.

The results show that the random forest classifier is used. In the case of, weighted SMOTE is more effective than SMOTE in processing unbalanced data. After training, more minority class samples can be correctly distinguished, which helps to improve the overall classification ability of the classifier. The weighted random forest has better classification ability than the random forest when both use weighted SMOTE to process the data.

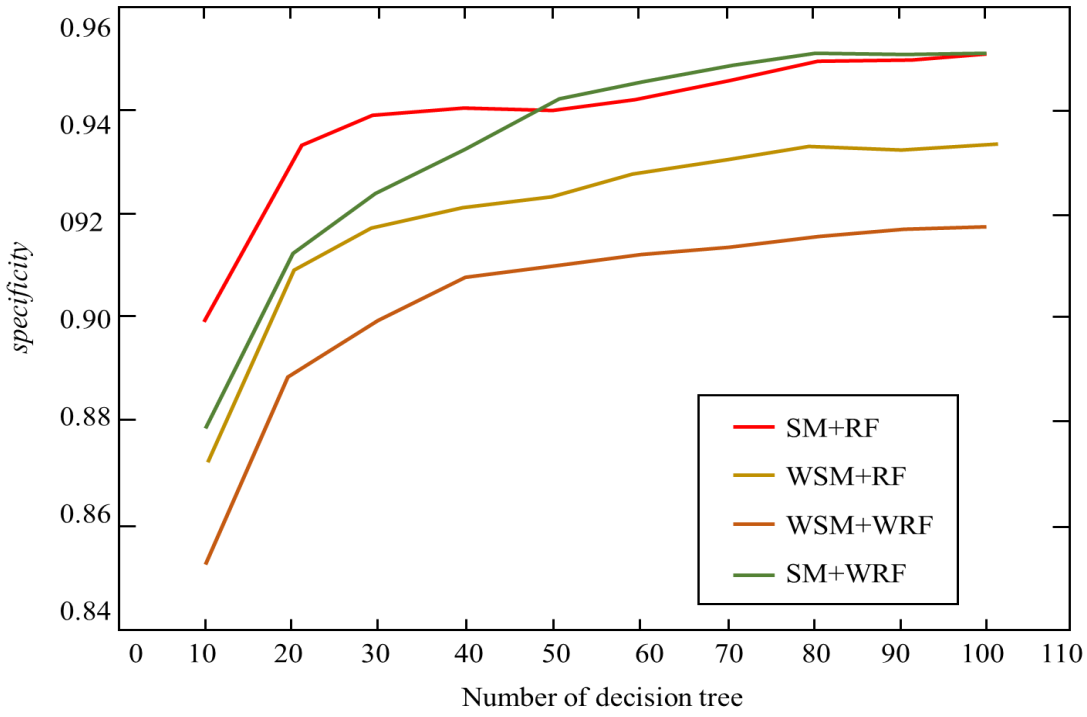


Figure 6. Vehicle0 dataset specificity comparison chart.

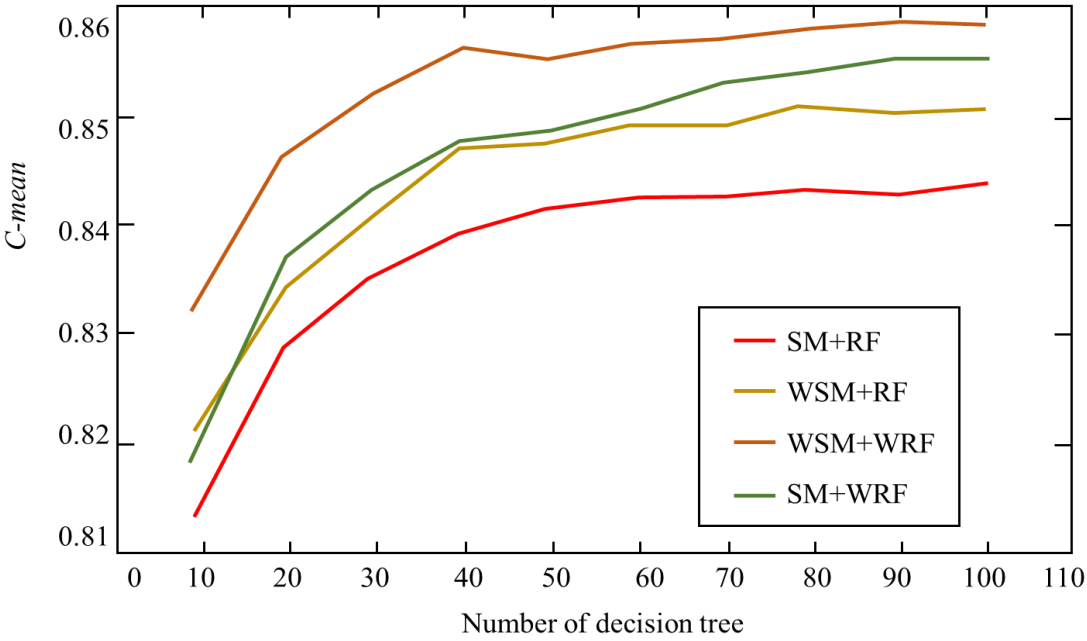


Figure 7. G-mean comparison chart of vehicle0 data set.

**Conclusion**

With the popularization of Internet technology, Internet platforms meet consumer demand, the Internet-based financial system has developed rapidly, and the model of Internet finance is constantly innovating. but in the meanwhile, A series of questions about Internet finance also followed. The paper mainly starts with the development model of my country's Internet finance and then analyses the current situation of my country's Internet finance development and analyses the existing problems. Through the financial risk early warning model of Internet financial companies established by SMOTE-random forest, the following conclusions and suggestions can be drawn. First, by referring to the financial risk early warning model established in this article, regulators and investors can take the financial risk status of Internet finance companies

as a reference and then make choices to reduce some financial losses; banks can also selectively lend. Second, the relevant government regulatory departments should strengthen supervision and information disclosure mechanisms and improve relevant laws and regulations. For example, the company is required to publish information such as the company's shareholders and operating conditions, which does not involve the company's confidentiality, and can also provide data support and basis for later financial risk early warning.

**Funding**

This work was not supported by any funds.

**Acknowledgements**

The authors would like to show sincere thanks to those techniques who have contributed to this research.

**Conflicts of Interest**

The authors declare no conflict of interest.

**References**

[1] Xu, J., Yang, T., Zhuang, S., Li, H., Lu, W. (2024) AI-based financial transaction monitoring and fraud prevention with behaviour prediction. *Applied and Computational Engineering*, 77, 218-224.

[2] Feng, R., Qu, X. (2022) Analyzing the Internet financial market risk management using data mining and deep learning methods. *Journal of Enterprise Information Management*, 35(4/5), 1129-147.

[3] Ionescu, S. A., Diaconita, V. (2023) Transforming financial decision-making: the interplay of AI, cloud computing and advanced data management technologies. *International Journal of Computers Communications & Control*, 18(6).

[4] Joloudari, J. H., Marefat, A., Nematollahi, M. A., Oyelere, S. S., Hussain, S. (2023) Effective class-imbalance learning based on SMOTE and convolutional neural networks. *Applied Sciences (Basel, Switzerland)*, 13(6), 4006.

[5] Karthik, M. G., Krishnan, M. M. (2021) Hybrid random forest and synthetic minority over-sampling technique for detecting Internet of Things attacks. *Journal of Ambient Intelligence and Humanized Computing*, 1-11.

[6] Li, X. (2022) Research on the influencing factors of Internet financial risk and its prevention mechanism. *Modern Economics & Management Forum*, 3(1), 20-24.

[7] Lin, M. (2022) Innovative risk early warning model under data mining approach in risk assessment of Internet credit finance. *Computational Economics*, 59(4), 1443-1464.

[8] Liu, M., Gao, R., Fu, W. (2021) Analysis of Internet financial risk control model based on machine learning algorithms. *Journal of Mathematics*, 2021, 1-10.

[9] Sadhu, P. K., Yanambaka, V. P., Abdelgawad, A. (2022) Internet of things: Security and solutions survey. *Sensors*, 22(19), 7433.

[10] Limanto, S., Buliali, J. L., Saikhu, A. (2024) GLoW SMOTE-D: Oversampling technique to improve prediction model performance of students failure in courses. *IEEE Access*, 12, 8889-8901.

[11] Liang, F., Zhao, P., Huang, Z. (2023) Financial technology, macroeconomic uncertainty, and commercial banks' proactive risk-taking in China. *China Economic Quarterly International*, 3(2), 77-87.

[12] Zheng, Z. (2024) Financial risk early model warning combining SMOTE and random forest for Internet finance companies. *Journal of Cases on Information Technology*, 26(1), 1-10

[13] Tan, Y. (2022) Financial Risk Management of Small and Medium Sized Enterprises in the Internet *Environment*, 2022, 12(5), 1-12.

[14] Ahirwar, A., Sharma, N., Bano, A. (2023) Enhanced SMOTE & Fast Random Forest Techniques for Credit Card Fraud Detection. *Solid State Technology*, 64(1), 1234-1245.

[15] Chen, L. (2024) Internet Financial News Text Classification Algorithm Based on Blockchain Technology. *Springer, Cham*, 12(2), 56-67.

[16] Han, J., Cheng, H., Shi, Y., Wang, L., Song, Y., Zhang, W. (2023) Connectivity analysis and application of fracture cave carbonate reservoir in Tazhong. *Science Technology and Engineering*, 17(6), 156-167.

[17] Hou, Z. K., Cheng, H. L., Sun, S. W., Chen, J., Qi, D. Q., Liu, Z. B. (2024) Crack propagation and hydraulic fracturing in different lithologies. *Applied Geophysics*, 17(3), 256-267.

[18] Li, L., Li, H. (2024) Analysis of Financing Risk and Innovation Motivation Mechanism of Financial Service Industry Based on Internet of Things. *Complexity*, 2024(4), 1-10.

[19] Li, S., Liu, X., Li, C. (2023) Research on Risk Prediction Model of Internet Finance Based on Cloud Computing. *Journal of Mathematics*, 1853(5), 052033-052038.

[20] Liu, M., Gao, R., Fu, W. (2023) Analysis of Internet Financial Risk Control Model Based on Machine Learning Algorithms. *Journal of Mathematics*, 4(1), 78-89.