Factors and Impacts of Financial Crisis Formation Based on Big Data and RF Early Warning Model

Liyan Jiang
Anyang University, Anyang 456550, China
E-mail: jiangliyan12345678@163.com

Keywords: financial crisis, big data, RF early warning model, financial indicators, SMOTE algorithm

Received: June 16, 2025

With the continuous changes and globalization of financial markets, the factors that shape financial crises have become more complex. To solve the unbalanced data, the study introduces the synthetic minority over-sampling technique with editorial nearest neighbour method to improve the performance of the model. By examining the financial indicators in detail, different aspects of ST and non-ST enterprises are compared. The results show that the synthetic minority over-sampling technique and the editorial nearest neighbour method perform well in financial risk prediction, improving the correct classification rate of the "1" sample by 2.26% and the "0" sample by 3.46%. In addition, the study observes that the "cash content of operating income" of special treatment enterprises is significantly higher than that of non-special treatment enterprises, and that there are significant differences in debt service capacity, overhead growth rate, and profitability. The study provides a powerful methodology for corporate financial risk management and a more effective risk monitoring tool for government regulators and financial institutions.

Povzetek: Prispevek se ukvarja z analizo napovedi finančnih kriz z metodami velikih podatkov in strojnega učenja. Predlaga rabo SMOTE-ENN za uravnoteženje in naključne gozdove za zgodnje opozarjanje; izpostavi pomembnost kazalnikov denarnih tokov ter razlikovanje ST/ne-ST podjetij na podatkih A-trga.

1 Introduction

As the global economy continues to develop, enterprises are facing an increasingly complex and changeable market environment [1]. The occurrence of financial crises is often accompanied by various factors, including market fluctuations, management decisions, economic cycles, etc. Traditional financial analysis methods may be inadequate in complex changes. Therefore, there is an urgent need to combine big data and advanced machine learning algorithms to accurately understand and predict the factors causing financial crises. Traditional financial analysis methods have a series of shortcomings when dealing with financial crisis prediction [2]. These methods usually rely too much on historical data and analysis based on statistical indicators, and cannot fully consider the dynamic changes in the market and the uncertainty of the external environment. This makes it relatively rigid in rapidly changing market conditions and difficult to capture emerging risk factors in a timely manner. In addition, traditional methods are relatively weak in processing large-scale, high-dimensional data, and the explosive growth of data brought about by the current information age makes traditional methods unable to handle massive amounts of data [3]. This limits its ability to comprehensively analyze various potential risk factors, necessitating the need for more flexible and powerful analytical tools. The innovation of the research is to make full use of big data technology and Random Forest (RF) early warning model to model the formation factors of financial crisis. The introduced big data technology is expected to provide more powerful support in data acquisition, storage, and processing, and the random forest early warning model stands out due to its good adaptability to high-dimensional data and non-linear relationships. By combining the two, the research is expected to reveal the intrinsic mechanism of financial crises more comprehensively and accurately, and provide enterprises with more accurate early warning and risk management strategies.

The study consists of four sections. The first section summarizes the research on big data prediction and RF early warning model prediction in financial crises. The second section analyzes the financial crisis data from RF in a big data environment. The third section evaluates the financial early warning model with a comparative multi-model analysis. The fourth section is a summary of the full paper.

2 Related works

Big data is often used to analyse corporate financial problems. Many experts have explored. To establish a healthy, orderly, safe and sustainable development environment, Zhang built an innovative risk early warning model based on big data technology for the financial risk assessment of Internet credit. This model aims to improve the use efficiency of internal and external data and ensure that the early warning system is timely, accurate and effective [4]. However, the modeling framework remains relatively conventional and lacks mechanisms to address imbalanced data distributions, which can lead to biased

early warning outcomes. To improve the financial and economic development and enhance the competitiveness of financial institutions, Luo firstly understood the statistical characteristics of financial data and the design principles of the risk early warning system in the era of big data. Subsequently, the research on the design structure and implementation of financial data statistics and risk early warning analysis system was conducted [5]. However, the study does not offer concrete implementation paths for modeling and algorithmic application, and lacks empirical validation, limiting its practical guidance for financial risk prediction. To reduce the financial risk loss of commercial banks in the context of Internet finance, Lin applied the BP neural network to determine the number of nodes, activation function, learning rate and other parameters of each layer of the neural network. A large number of data samples was used to construct an Internet credit risk early warning model [6]. However, BP neural networks are prone to over-fitting and sensitive to feature noise. In addition, the study does not implement systematic optimization for class imbalance, which affects model stability. Yang discussed how to implement data-driven service upgrading techniques to assist in market regulation, and integrated user authentication information, financial information, and behavioral information through data processing and feature engineering. A credit risk assessment model was constructed using a forest algorithm to exploit the potential value of the data [7]. However, the feature selection process lacks transparency, and the model shows limited interpretability. Moreover, it does not adequately address issues such as small sample sizes and data imbalance common in financial scenarios. Gautama discussed how to develop a fraud alert system using the Software Development Lifecycle (SDLC) approach, building an integrated big data system and fraud alert system to identify potential fraudulent behavior. The fraud alert system integrated LHKPN data and comparative technology data using specific metrics and variables to assess the amount of risk in identifying potential fraud [8]. However, the research focuses on system design and implementation, with limited analysis of model performance and prediction effectiveness, making it difficult to generalize to broader financial risk warning tasks.

Early warning models based on RF algorithms have significant advantages in financial early warning. Jarmulska et al. aimed to answer whether it is possible to design effective and useful machine-learning based early warning systems. The author designed and compared multiple models based on econometric Logit model and RF model to evaluate their performance in predicting financial stress risk. The results showed that it obtained effective early warning models that could correctly predict 70-80% of fiscal stress events and calm periods [9]. However, the study does not test the model generalization across different sample environments and lacks assessment of its adaptability to diverse financial crisis scenarios. Cong discussed how to build financial early warning models for listed enterprises based on multiple different integrated classification models. The experimental results showed that they were effective in predicting financial data for listed enterprises [10]. However, the study primarily focuses on accuracy metrics, without considering comprehensive evaluation indicators such as recall and F1 value under imbalanced sample conditions, and lacks interpretability analysis of the model. Yang et al. explored how to establish financial risk early warning measures to cope with the severe financial situation. A prediction model was established using RF to assess financial risks. Although the model was highly accurate, it had limited interpretability. The Fisher Discriminant Method based on RF was used to improve the interpretability and accuracy of the model [11]. However, the study does not further address feature redundancy or class imbalance, thus achieving only limited improvements in model performance. Wang et al. proposed the "Expert Voting EWS" framework to predict the impending systemic banking crisis. Machine learning algorithms, especially RF classifier, were used to build the EWS. Classifiers, to construct the EWS. The region under the receiver operating characteristic curve was analyzed to compare the generalization ability of different classifiers [12]. Although the system design is practical, the methodology relies heavily on manual expert integration, lacking robust self-adaptive generalization capabilities. Liu et al. used logistic models and seven different machine learning methods, including RF, gradient boosting decision trees, and integrated models to avert the catastrophic effects of the financial crisis. Shapley value decomposition and Shapley regression were used to analyze the causality and impact of machine learning models [13]. However, the study does not incorporate data balancing or sample optimization strategies during the preprocessing stage, which affects the precision of risk sample identification.

In summary, although existing research has made progress in model construction, data integration, and interpretability analysis, especially the powerful performance of RF models in handling nonlinear feature recognition, there remain several limitations. These include the lack of systematic oversampling or noise-reduction strategies for imbalanced data, insufficient generalization across different datasets or enterprise types, limited exploration of financial indicator differences between Special Treatment (ST) and Non-special Treatment (Non-ST) enterprises, and inadequate use of comprehensive performance metrics such as F1 value and AUC. The novelty of this study lies in integrating the Synthetic Minority Over-sampling Technique with Editorial Nearest Neighbour method (SMOTE-ENN) sampling strategy with the RF model, which systematically enhances the model robustness and classification performance under imbalanced samples. Additionally, based on empirical data from the A-share market, the importance of key financial indicators for ST and non-ST enterprises is compared. This study extends the applicability of financial risk prediction models in identifying risk heterogeneity and improving adaptive modeling capabilities.

3 Research on random forest-based financial crisis data analysis in big data environment

The first subsection explores the factors that shape the financial crisis of an enterprise and its impact on the business. The second subsection describes the SMOTE algorithm to deal with unbalanced data, particularly in the training and evaluation of the RF tree dataset. The third subsection explores how the RF model can be used to select important financial warning features as well as improve the generalization of the model.

3.1 Research on the formation factors of corporate financial crisis and its impact in the context of big data

The formation mechanism of financial crisis is a complex process of multi-factorial interaction. Firstly, poor business decisions are a key factor, including unwise strategic planning and resource allocation, wrong capital budgeting decisions, and failure to adapt to market changes or choosing inappropriate product mix [14]. In addition, financial opacity and fraudulent behaviour are important factors contributing to crises, such as false financial statements and improper fund flows. High leverage and debt problems also often trigger financial crises, especially when enterprises are too highly leveraged to carry their debt burden. Liquidity problems may lead to debt default. Macroeconomic factors, such as economic downturns, inflation, or exchange rate fluctuations, as well as competitive and market factors, such as market share losses and unstable market demand, also have a significant impact on financial crises. In addition, corporate governance issues, including inadequate board supervision and management ethics, as well as balance sheet problems such as improper asset valuation and holding of illiquid assets, may have adverse effects on the financial condition of the enterprise. In this complex context, big data analysis plays a crucial role. By applying advanced data analysis and warning models, enterprises can better monitor these factors, track changes in financial indicators, and identify potential risk signals in advance. Big data technologies are able to process large-scale financial data to identify anomalous patterns, predict future trends, and simulate different economic scenarios. This helps organizations assess their financial risks more accurately and take timely action to avoid financial crises from developing. In addition, big data can help businesses better understand market fluctuations, including prices and market volatility, in order to adjust strategies and resource allocation more flexibly, thereby reducing financial pressure [15].

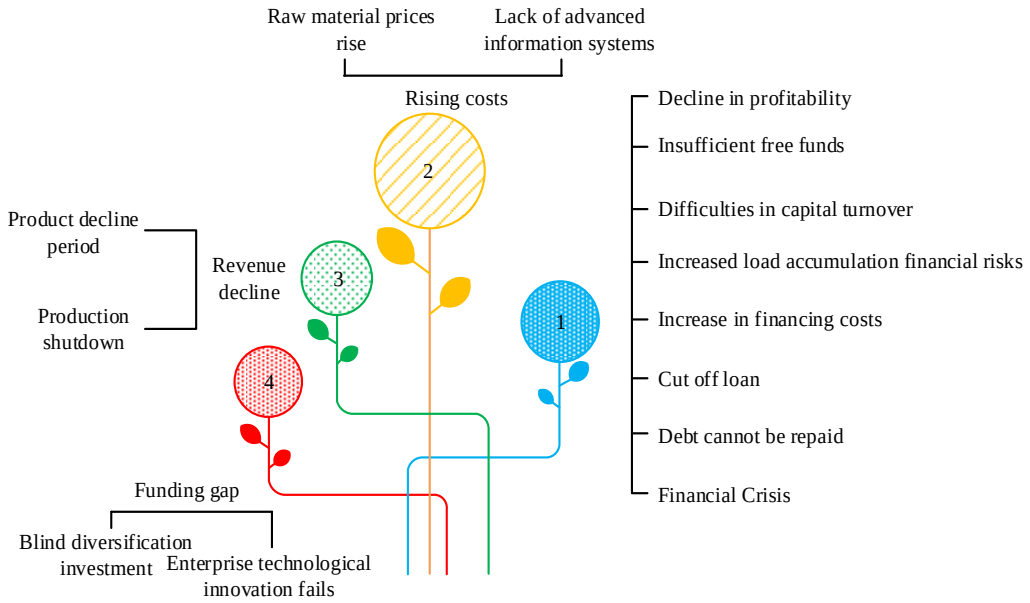


Figure 1: Accounting framework for corporate financial distress identification

The fundamental cause of financial crisis is the accumulation of debt, which leads to the inability to repay debts on time. The initial sign is a decline in income from main business, which may be due to reasons such as stagnant products, rising costs, and failed investments. Blind diversification and over-expansion lead to difficulties in returning capital. Enterprises may attempt to borrow from external sources, but the increase in financing costs can lead to poor financial conditions, making it difficult to obtain new financing, ultimately resulting in a financial crisis.

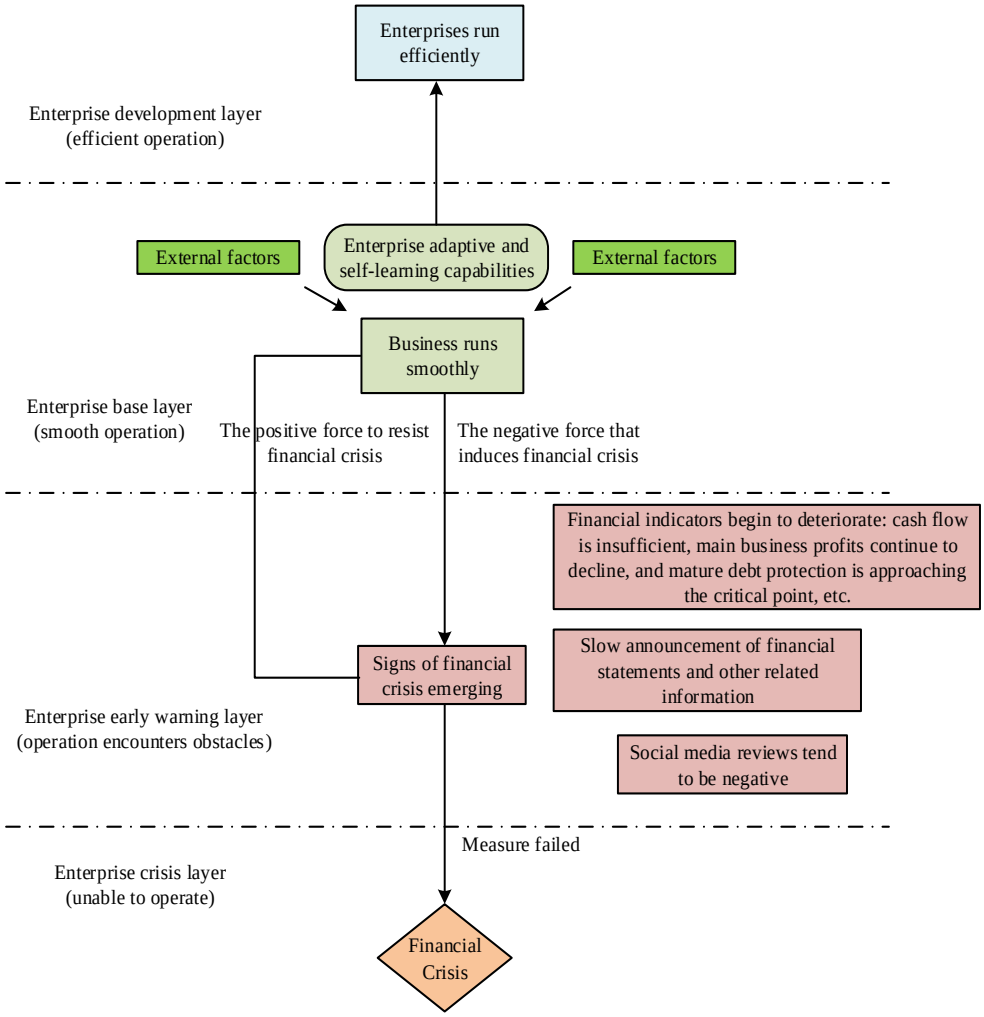
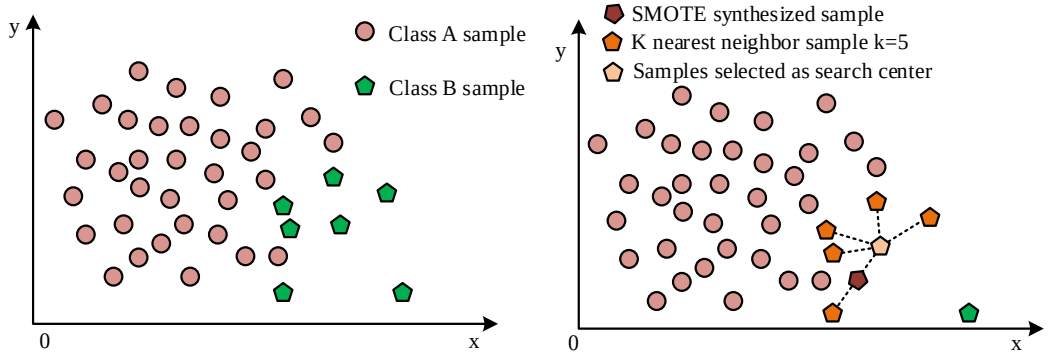


Figure 2: Mechanism diagram of financial distress formation in enterprises

This study conceptualizes enterprise operations as a dynamic regulatory system influenced by both internal mechanisms and external factors. As illustrated in the figure, enterprises operate efficiently under normal conditions, supported by adaptive and self-learning capabilities that help maintain operational stability. Faced with external disturbances or internal inefficiencies, the ability of an enterprise to identify early signals and respond effectively determines whether it can sustain equilibrium or move towards risk accumulation. The enterprise life cycle depicted in the figure includes four phases: stable operation, early anomalies, risk buildup, and crisis onset. In the initial stage, financial indicators are generally healthy [16]. However, as the pressure increases, such as tightening cash flow, declining profits, delayed financial disclosures, and increasingly negative social media sentiment, early signs of financial stress begin to emerge. If these signals are not addressed in time, they may evolve into a full-blown financial crisis. The chart emphasizes that financial crises are rarely sudden events. On the contrary, they are the result of gradual deterioration and failure of corrective mechanisms. Establishing a financial warning system that integrates multiple data sources can help identify risks in the pre-crisis stage and intervene in a timely manner, enhancing the resilience and stability of the enterprise.

3.2 SMOTE algorithm based training and evaluation method for random forest tree dataset

Synthetic Minority Over-sampling Technique (SMOTE) is a technique used to deal with unbalanced datasets by generating synthetic samples to optimize the dataset to improve the performance of a machine learning model. The core idea of SMOTE is to improve the machine learning model by inserting new synthetic samples in the feature space, balancing the sample distribution between different categories.



(a) Calculate K nearest neighbor samples of X (b) Synthesize new samples

Figure 3: Conceptual structure of the SMOTE algorithm

The original dataset is split into a feature matrix (x) and a target vector (y), and the data is explored to understand the imbalance and determine which category is the minority category and which is the majority category. For the minority category, synthetic samples are generated using SMOTE to balance the category distribution. Meanwhile, the generated balanced dataset (including the original samples and the generated synthetic samples) is used to train the RF model. The hyperparameters of the RF can be adjusted as needed to further improve the model performance [17], as shown in equation (1).

$$X_{new} = X_i + (\bar{X} - X_i) \times \delta \quad (1)$$

In equation (1), $X_i \in S\min$. The sample X_i is randomly selected from the nearest neighbors of K=5 to generate a new sample X_{new} . δ denotes the random number. Before determining the model, the model effect is evaluated, and the evaluation indicators are the criteria for objectively assessing the performance and effectiveness of the model. For the classification task, the common evaluation metrics of the classification model are Precision and Recall, as shown in equation (2).

$$R = \frac{TP}{TP + FN} \quad (2)$$

In equation (2), TP denotes the number of samples correctly classified as positive examples by the model. That is to say, the number of actual positive examples correctly predicted by the classifier as positive examples. FN denotes the number of samples misclassified as positive by the model, i.e., the number of actual negative samples that the classifier incorrectly predicts as positive. The precision can be determined using the confusion matrix, which is shown in Figure 4.

Confusion matrix		Actual value	
		Positive	Negative
Predictive value	Positive	TP	FP
	Negative	FN	TN

Figure 4: Confusion matrix structure for binary classification evaluation

Since precision and recall are often opposed to each other, i.e., a high precision is associated with a low recall, and a high recall is associated with a low precision. When there is an imbalance between positive and negative samples, relying solely on precision or recall cannot effectively measure the strengths and weaknesses of the model. To consider the precision and recall comprehensively, the F-Measure (also known as F-Score) is used as a comprehensive evaluation index [18]. F-Measure is the weighted summed average of the precision and recall, which can consider the precision and coverage ability of the classifier comprehensively [9]. The specific formula is shown in equation (3).

$$F = \frac{\left(a^2 + 1\right) \times P \times R}{a^2 \times \left(P + R\right)} \quad (3)$$

Micro-F1 is a comprehensive evaluation metric based on a global overall calculation. It first calculates the True Positive (TP), False Positive (FP), and False Negative (FN) for all categories, and then uses these statistics to calculate the overall precision and recall. Finally, a comprehensive performance evaluation is computed by F1 value [19]. Micro-F1 is suitable for cases where the sample categories are unevenly distributed, as it considers the weights of all category samples in the calculation. The precision is specifically shown in equation (4).

$$P_{micro} = \frac{\sum_{i=1}^{M} TP_i}{\sum_{i=1}^{M} \left(TP_i + FN_i\right)} \quad (4)$$

The recall rate is specifically shown in equation (5).

$$R_{micro} = \frac{\sum_{i=1}^{M} TP_i}{\sum_{i=1}^{M} \left(TP_i + FN_i\right)} \quad (5)$$

Equations (6) and (7) are substituted into equation (5) to obtain the micro-F1 value, as shown in equation (6).

$$F1_{micro} = \frac{2 \times P_{micro} \times R_{micro}}{P_{micro} + R_{micro}} \quad (6)$$

The results of the corresponding F1 value are calculated according to equation (6). Then, the arithmetic mean is obtained, as shown in equation (7).

$$F1_{micro} = \frac{\sum_{i=1}^{M} F1_i}{M} \quad (7)$$

In Eq. (7), M denotes the number of target categories. In practical multi-classification problems, there is usually no definite conclusion on whether to choose to use micro-F1 values or macro-F1 values as an evaluation metric for model strengths and weaknesses. Each metric has its own characteristics and applicable scenarios, depending on the specific problem and needs.

3.3 Random forest-based financial early warning feature selection and generalisation capability study

RF is a powerful machine learning algorithm that reduces the over-fitting by randomly selecting features and combining multiple decision trees. These trees are constructed independently and predictions are made using majority voting (classification) or averaging (regression). RF evaluates performance through cross-validation and also estimates the importance of features to help identify key features.

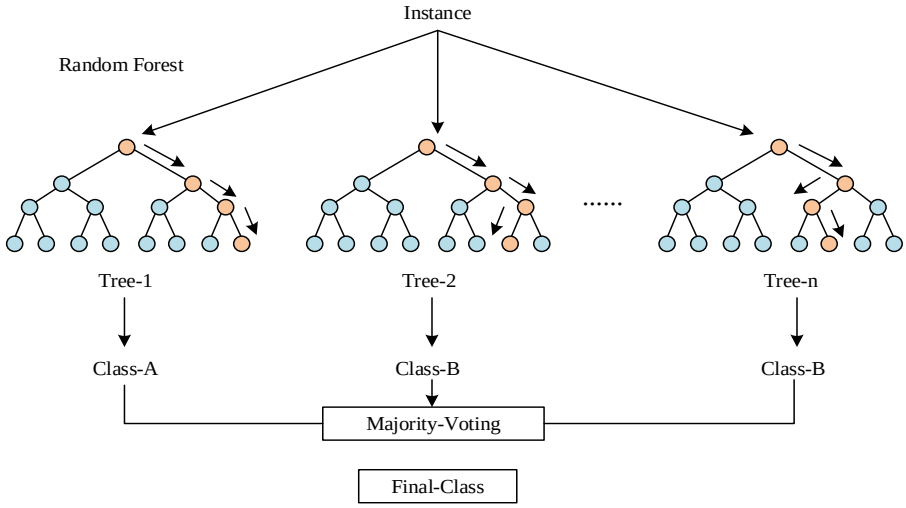


Figure 5: Design logic of the financial safety verification boundary

The application of RF in financial early warning research has significant advantages. The RF model introduces randomness. Firstly, it improves prediction performance by preferring the most valuable features. Secondly, it ranks the importance of each feature to help select the most influential factors that contribute to financial early warning. In addition, it evaluates the importance of individual indicators, estimates the model's generalization error, and deepens the understanding of model stability and generalization performance, which is crucial for financial crisis prediction and risk management. In addition, RF excels in dealing with classification problems, providing reliable prediction results that are not limited by the amount of data. Most importantly, it is suitable for large-scale and high-dimensional data without the need for feature selection and data standardization, simplifying data preparation and accelerating model construction and deployment, making it a powerful tool for financial early warning research [20]. In the above process, it is crucial to determine the best feature and attribute selection when constructing the decision tree model. The core principle of the ID3 decision tree algorithm is to maximize the information gain. The information gain is calculated by analyzing the information gain of the dataset D under the feature A to determine the priority of the best features and attributes to divide the dataset. This is specifically shown in equation (8).

$$Gain(D,A)=Ent(D)-Ent(D|A)=-\sum_{k=1}^{|y|} p_k \log_2 p_k - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v) \quad (8)$$

In Eq. (8), $Ent(\ )$ denotes the information entropy. p_k denotes the class sample percentage. D^v denotes the data ensemble when the v-th attribute is taken on the feature A. To prevent the over-fitting phenomenon, the improved information gain expression is shown in Eq. (9).

$$Gain\_ratio(D,A)=\frac{Gain(D,A)}{IV(D,A)}=\frac{Gain(D,A)}{-\sum_{v=1}^{V}\frac{|D^v|}{|D|}\log_2\frac{|D^v|}{|D|}} \quad (9)$$

The pre-cutting method and post-cutting method are shown in equation (10).

$$Gini(D)=\sum_{k=1}^{|y|}\sum_{k`\neq k} p_k p_{k`} = 1-\sum_{k=1}^{|y|} p_k^2 \quad (10)$$

Equation (3) represents the Gini index of the dataset D under the feature A. The feature division point is shown in equation (11).

$$Gini\_index(\mathrm{D,A}) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} Gini(D^v) \quad (11)$$

The minimization criterion selects the optimal division features and division points and minimizes the squared error, as shown in equation (12).

$$\min_{A,a}[\min_{c_1} \sum_{xi \in D_1(A,a)} (y_i - c_1)^2 + \min_{c_2} \sum_{xi \in D_2(A,a)} (y_i - c_2)^2] \quad (12)$$

In Eq. (12), $D_1(A,a) = \{x \mid x^{(A)} \le a\}$, ; $D_2(A,a) = \{x \mid x^{(A)} > a\}$, y_i denotes the true value of the sample. c_1, c_2 denote the average sample prediction of the sub dataset.

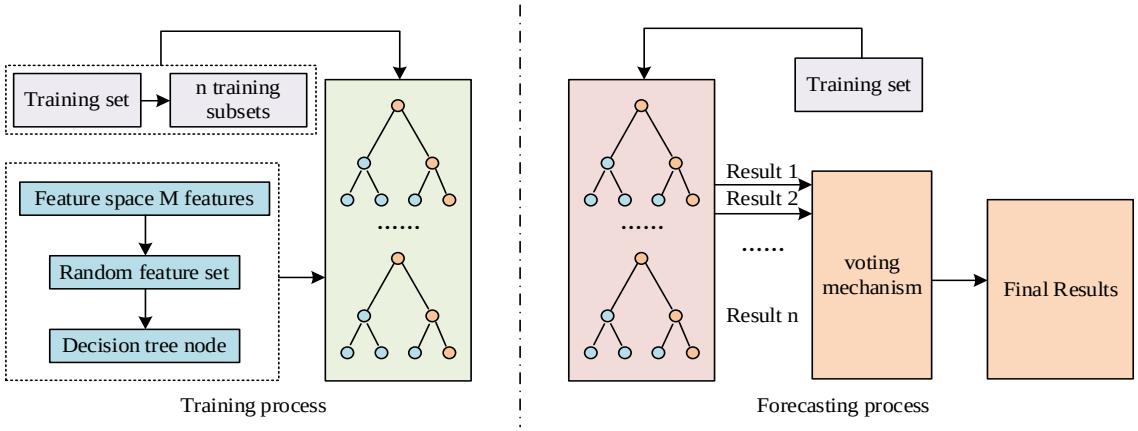


Figure 6: Workflow of random forest model training and prediction

The RF model is unique in two key ways. Firstly, it takes a putative back sampling strategy, leading to out-of-packet data, which is used to test the generalization ability. Secondly, RF introduces randomness in feature selection, combined with an independent base learner and parallel computing, ensuring strong generalization ability. The change in Gini index of the feature variable X_j before and after the above node branching is calculated as shown in equation (13).

$$VIM_{im}(X_j) = Gini(D,A) - Gini\_ratio(D,A) - Gini(D) \quad (13)$$

The sum of the changes in the Gini index of the variable X_j over all the decision nodes of the first i tree is calculated, as shown in equation (14).

$$VIM(X) = \sum_{v=1}^{c} VIM(X_v) \quad (14)$$

In Eq. (14), c denotes the number of feature variables.

$$VIM(X_j) = \frac{VIM(X_j)}{VIM(X)} \quad (15)$$

Eq. (15) represents the normalized total Gini change for a feature variable to obtain the final RF importance score.

4 Evaluation of financial early warning models and comparative multi-model analysis

This section is to develop financial early warning models and conduct comparative analyses to determine which method or model performs better in predicting financial crises. The first subsection integrates the financial early warning model and the performance. The second subsection focuses on modelling corporate financial crisis prediction using the RF.

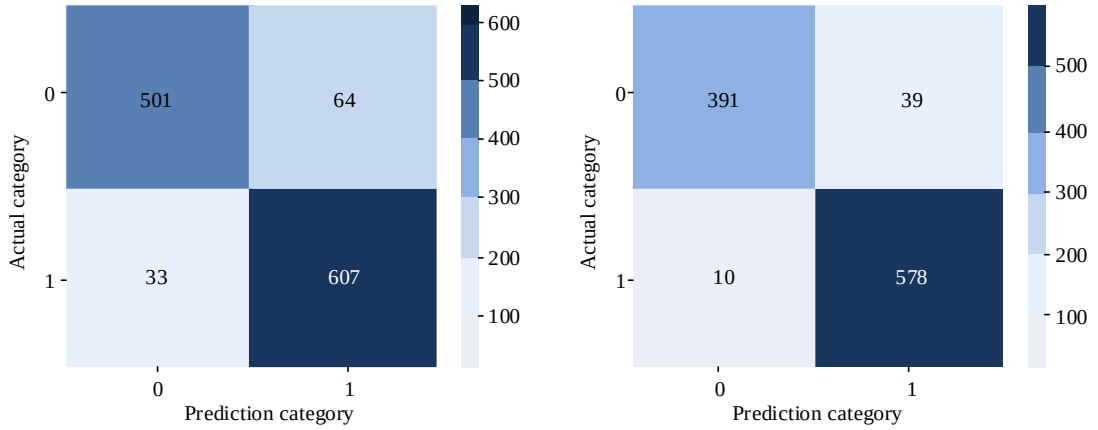
4.1 Development of an integrated model for financial early warning and its comparative cross-analysis

Before training, missing values are filled with medians and outliers are removed using the IQR method. The dataset is split into 70% training and 30% testing sets using stratified sampling. SMOTE-ENN is applied on the training set to handle class imbalance, with SMOTE (k=5) generating synthetic minority samples and ENN removing noisy majority samples, resulting in a near 1:1 class ratio.

RF hyperparameters are tuned using GridSearchCV with five-fold cross-validation. The search covers n_estimators (100-300), max_depth (10, 20, None), and min_samples_split (2, 5, 10). To prevent over-fitting, Out-of-bag (OOB) scoring is used and redundant features are removed based on importance rankings. All experiments are conducted with random_state=42 using scikit-learn (v1.2.2) and imbalanced-learn (v0.10.1) for reproducibility.

Cross-sectional comparative analyses are crucial in the field of financial early warning to help assess the capability and effectiveness of different integrated models for financial early warning in practical applications. A series of standardized experiments and performance evaluations are conducted to compare the performance of different models on the same dataset. These experiments include different performance evaluation metrics such as accuracy, check accuracy, check completeness, F1 value,

AUC under ROC curve, etc. to quantify the performance of each model.



(a) Test set confusion matrix 1 (b) Test set confusion matrix 1

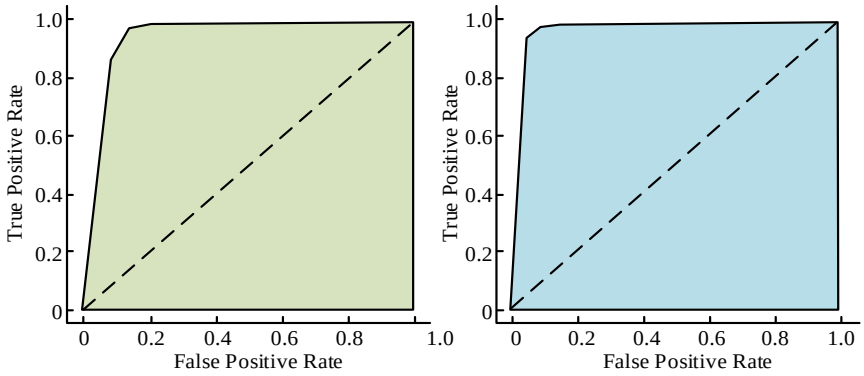
Figure 7: Confusion matrices of the random forest model under two sampling methods

As shown in Figure 7, two data balancing methods are used in the study, namely SMOTE-ENN integrated sampling method and Borderline-SMOTE oversampling method. In Figure 7(a), for "0" category samples, the correct classification rate is 88.67%, while for "1" samples, the correct classification rate is 94.84%. In Fig. 7(b), the correct classification rate is 90.93% for "0" samples and 98.30% for "1" samples. In conclusion, no matter which data balancing method is used, the RF model has a higher correct classification rate for "1" samples than "0" samples. The model based on the SMOTE-ENN integrated sampling method performs better on the correct classification rate, which is 2% higher than the correct classification rate of the "0" samples. The model based on the SMOTE-ENN performs better in the correct classification rate of both types of samples, which is 2.26% and 3.46% higher, respectively. The evaluation indexes of the two models under the RF tree model are shown in Table 1.

Table 1: Evaluation indicators of the random forest model under two data balancing methods

Evaluation index	**Data balancing method used**	
	Borderline-SMOTE oversampling (%)	SMOTE-ENN integrated sampling (%)
Recall	94.85	98.31
Accuracy	90.45	93.69
Precision	91.96	95.20
Out-of-Bag Score	92.49	94.61
F1 value	92.61	95.94

According to the data in Table 1, the accuracy of financial risk identification of the RF model under the Borderline-SMOTE oversampling data balancing method reaches 91.96%. Regarding the recall rate, 94.85% of the enterprises that actually have financial risks can be warned in advance, but there are still 5.15% of enterprises that cannot be identified in advance. When using the SMOTE-ENN integrated sampling data balancing method, the financial risk identification accuracy using the RF model reaches 95.20%. In terms of recall, 98.31% of the enterprises that actually have financial risks can be warned in advance, but 1.69% of the enterprises cannot be identified in advance. To further verify the reliability of the above-mentioned performance differences, the paired t-test is conducted on the results of the two sampling methods. The results show that the SMOTE-ENN method is significantly superior to the Borderline-SMOTE method in multiple indicators (p<0.01), indicating that this method has stronger statistical advantages and practical applicability in financial risk prediction. The performance differences of the model shown in Table 1 fundamentally stem from the differences in the processing mechanisms of the training sample structures between the two data balancing methods. Although Borderline-SMOTE enhances the sample density of the minority classes in the boundary region, it does not clean up the noise samples of the majority classes, resulting in confusion in the model in boundary learning and limiting the precision and F1 value. While the SMODS-ENN method generates minority class samples, it introduces the editing nearest neighbor strategy to effectively remove the outliers located near the boundaries in the majority classes, significantly optimizing the class discrimination structure. This improvement enables the model to learn clearer classification boundaries during the training process and enhances the ability to identify high-risk enterprises.



(a) Test set ROC curve 1 (b) Test set ROC curve 2

Figure 8: ROC curves and AUC values under SMOTE and SMOTE-ENN sampling strategies

From Fig. 8(a) and Fig. 8(b), the ROC curves of the RF model show different trends among different data balancing methods. The ROC curve of the RF model shows a significant improvement when using the SMOTE-ENN sampling method. This method makes the ROC curve closer to the upper left corner and closer to the AUC value of 1, indicating significantly better model performance. This trend suggests that SMOTE-ENN integrated sampling is critical to improving the classification performance of the RF model. The success of the integrated sampling approach suggests that it can play a key role in improving the prediction accuracy when dealing with unbalanced data.

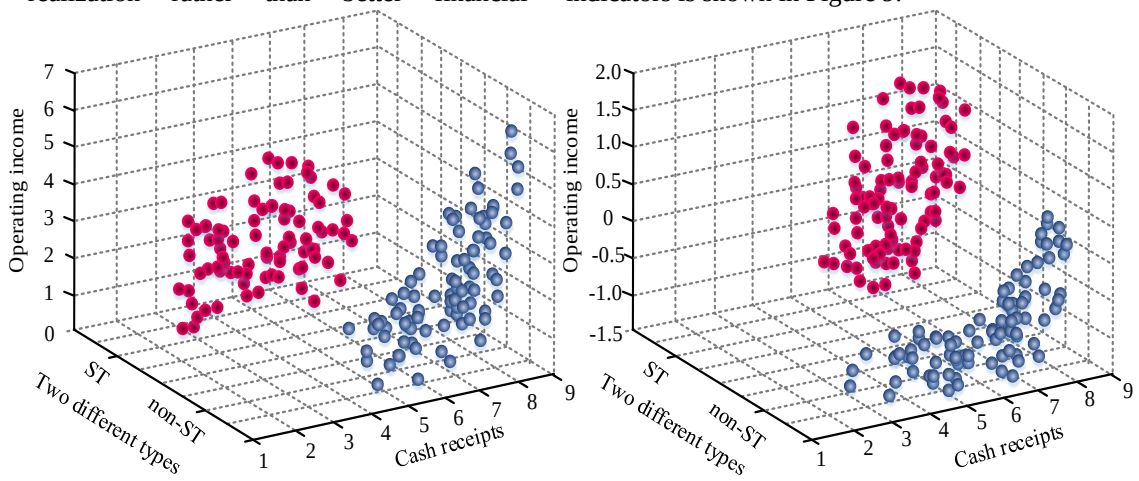
4.2 Analysis of financial crisis prediction of enterprises based on random forest algorithm

RF combines multiple decision trees to improve forecasting accuracy by combining the results of multiple models. It can also provide the importance ranking of each feature. By analyzing the importance ranking of these indicators, the study can identify the factors that have the greatest impact on the financial stability. This study is based on data from the A-share market and selected six ST enterprises. These enterprises refer to those that are given special treatment by the stock exchange in certain special circumstances when the enterprise encounters financial or operational problems. Non-ST enterprises refer to those enterprises that are not classified as ST enterprises, which do not have special financial or operational problems, or at least have not been publicly disclosed. The importance ranking of these indicators is shown in Table 2.

Table 2: Index importance ranking

Indicator name	Weights	Indicator name	Weights
Cash content of operating income	0.332	Operating income growth rate	0.010561
Cash flow from operating activities/interest-bearing liabilities	0.095	Cash reinvestment ratio	0.009645
Operating profit margin	0.075	Total cash recovery rate	0.009438
ROE	0.074	Operating margin	0.009221
Net profit margin on total assets	0.054	Current ratio	0.007136
Rate of return on intangible assets	0.040	Sustainable growth rate	0.00701
Cost profit margin	0.034	Net profit growth rate	0.00631
Overhead rate	0.030	Net cash flow per share	0.006115
Total asset turnover ratio	0.028	Total assets growth rate	0.005007
Assets and liabilities	0.023	Basic earnings per share growth rate	0.004527
Intangible assets growth rate	0.022	Management expense growth rate	0.004512
Net operating cash flow/current liabilities	0.021	Price to sales ratio	0.002465
Corporate free cash flow per share	0.019	Cash ratio	0.001823
Growth rate of net flow from operating activities	0.017	/	/

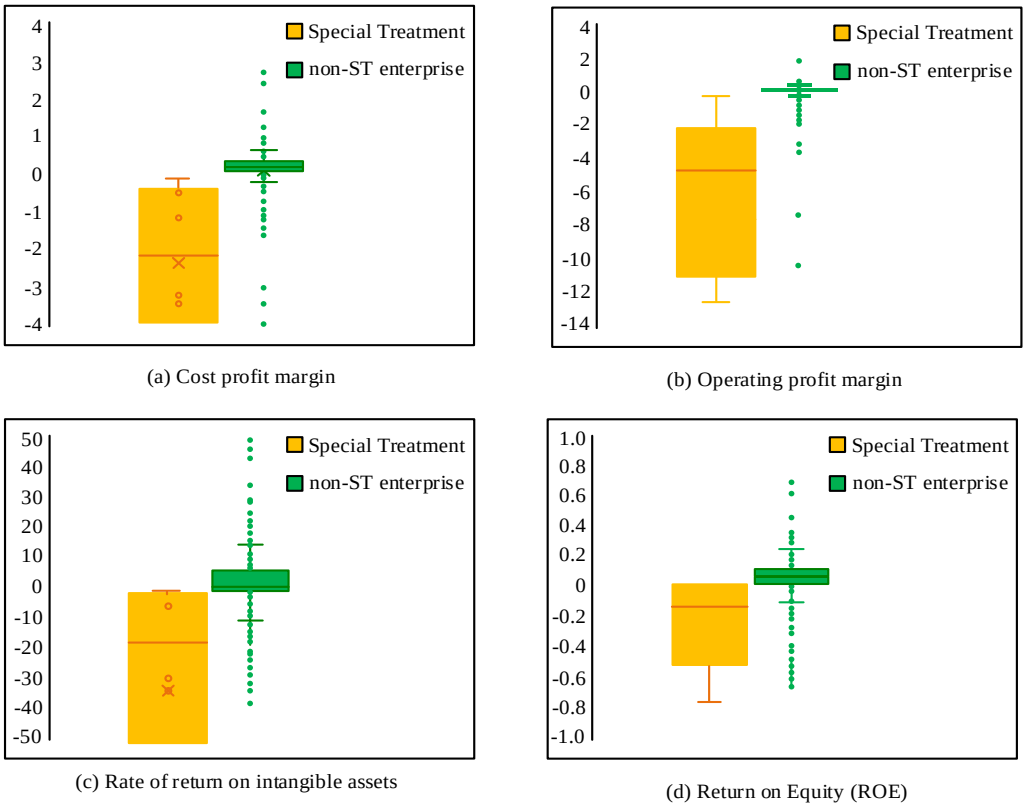
Table 2 ranks financial indicators by their importance within the RF model, calculated through the average reduction in Gini impurity across-decision trees. A higher weight reflects a greater contribution to the classification accuracy. The indicator "Cash content of operating income" shows the highest importance value of 0.332, indicating its strong discriminative ability between ST and non-ST enterprises. Although ST enterprises have a higher average for this indicator, the result may be due to delayed revenue realization rather than better financial performance. Indicators such as "Cash flow from operating activities/interest-bearing liabilities" and "Operating profit margin" also rank highly, reflecting the model's focus on liquidity and profitability. In contrast, traditional static metrics such as the "Cash ratio" have lower weights, suggesting limited influence in predicting financial distress. This importance ranking reflects the model's preference for operational and cash flow efficiency indicators, providing practical guidance for financial risk assessment. The comparison of key financial indicators is shown in Figure 9.



(a) Comparison of cash content of operating income (b) Operating cash flow to interest-bearing debt ratio

Figure 9: Comparison of key financial indicators between ST and non-ST enterprises

Figure 9 shows the differences between ST enterprises and non-ST enterprises in terms of key financial indicators, mainly focusing on two dimensions: solvency and the growth rate of administrative expenses. Solvency is measured by the ratio of cash flow from operating activities to interest-bearing liabilities. A high ratio indicates a stronger debt paying ability of the enterprise. The results show that the overall solvency of non-ST enterprises is higher than that of ST enterprises, reflecting the stability of the former on cash liquidity and debt repayment structure. In terms of the growth rate of administrative expenses, ST enterprises show a significant negative growth, with an average of -0.95, while the average of non-ST enterprises is 0.22, mainly distributed in the positive growth range. This indicates that ST enterprises often cut management expenses under financial pressure. Further, the independent sample t-test is used to compare the two groups of samples. The results show that the difference is statistically significant (p<0.01), indicating that the growth rate of management expenses has a strong discriminatory ability in identifying the financial health status of enterprises. The significant differences in the above-mentioned key indicators not only reveal the structural weaknesses of ST enterprises on debt-paying ability and operating expenditure management, but also provide a solid empirical basis for risk identification and modeling based on financial characteristics.



(a) Cost profit margin

(b) Operating profit margin

(c) Rate of return on intangible assets

(d) Return on Equity (ROE)

Figure 10: Comparison of profitability indicators for ST and non-ST enterprises

Figure 10 shows the comparison between ST enterprises and non-ST enterprises in multiple profitability indicators, including cost profit margin, operating profit margin, return on intangible assets, and return on equity. Firstly, in terms of the cost profit margin indicator, the ST enterprises is relatively high, indicating that the cost expenditure per unit of income is larger, further confirming the insufficiency of their profitability. The operating profit margin reflects an enterprise's ability to generate profits through its main business. ST enterprises perform significantly worse, with an average of -6.24, which is much lower than the average of 0.06 for non-ST enterprises. This indicates that their main business is generally in a loss-making state. According to independent sample t-test, the difference is statistically significant (p<0.01). In terms of the return on intangible assets, the average level of ST enterprises is significantly lower than that of non-ST enterprises, reflecting their shortcomings in the utilization efficiency of intangible resources such as brand and technology. The return on equity is used to measure the ability of an enterprise to create value for shareholders. The average value of ST enterprises is -0.22, which is lower than 0.01 of non-ST enterprises. This indicates that ST enterprises not only have weaker profitability, but also have lower returns to shareholders. This difference is also statistically significant (p<0.05). The differences in the above-mentioned multiple dimensions jointly reveal that ST enterprises have systematic disadvantages on operational efficiency and profit creation ability. The significant differences in these profit indicators not only verify the disadvantaged position of ST enterprises in financial performance, but also provide quantifiable discrimination basis for the risk identification model.

5 Conclusion

Financial crisis is a phenomenon that may have a serious impact on the survival and development of an enterprise. This study delves into the factors that lead to financial crises in enterprises, making full use of big data techniques and RF models to perform detailed analyses and financial risk prediction. Meanwhile, the study uses SMOTE-ENN and Borderline-SMOTE to cope with the unbalanced data problem. The RF model is used for financial risk prediction, where the correct classification rate of the "1" sample is higher than that of the "0" sample. The SMOTE-ENN method performs better, increasing the correct classification rate of the "1" sample. The SMOTE-ENN method performs even better, improving the correct classification rate by 2.26% for the "1" sample and 3.46% for the "0" sample. The ROC curves show a significant performance improvement when using SMOTE-ENN. The study also focuses on financial indicators and finds that the "cash content of operating income" is significantly higher for ST enterprises than for non-ST enterprises. There are also significant differences in debt service capacity, overhead growth rate, and profitability, highlighting the differences in financial indicators between different types of enterprises. Especially for ST enterprises, it is examined in depth to have a more comprehensive understanding of their financial situation and risks. The shortcoming of the study is that it focuses on the current financial prediction. Afterwards, it is

necessary to consider the long-term trend of the financial crisis and the impact of macro factors.

Funding

Research Project on Teacher Education Curriculum Reform in Henan Province in 2022, stage results of exploration and practice of teaching reform of local education majors from the perspective of "curriculum Ideology and politics", Project NO: 2022-JSJYZD-059.

References

[1] Ge J, Wang F, Sun H, Lu F. Research on the maturity of big data management capability of intelligent manufacturing enterprise. Systems Research and Behavioral Science, 2020, 37(4): 646-662. DOI: 10.1002/sres.2819

[2] Zeng H. Influences of mobile edge computing-based service preloading on the early-warning of financial risks. The Journal of Supercomputing, 2022, 78(9): 11621-11639. DOI: 10.1007/s11227-022-04329-2

[3] Zhang Z, Chen Y. Tail risk early warning system for capital markets based on machine learning algorithms. Computational Economics, 2022, 60(3): 901-923. DOI: 10.1007/s10614-022-10260-8

[4] Zhang B. Application of Innovative Risk Early Warning Model Based on Big Data Technology in Internet Credit Financial Risk. Journal of Information Technology Research (JITR), 2022, 15(1): 1-12. DOI: 10.4018/JITR.2022010101

[5] Luo P. Design and implementation of financial data statistics and risk early warning analysis system in the era of big data. Advances in Economics and Management Research, 2023, 6(1): 364-364. DOI: 10.2991/978-94-6463-142-5_53

[6] Lin M. Innovative risk early warning model under data mining approach in risk assessment of internet credit finance. Computational Economics, 2022, 59(4): 1443-1464. DOI: 10.1007/s10614-021-10180-z

[7] Yang G. Research on Financial Credit Evaluation and Early Warning System of Internet of Things Driven by Computer-Aided Technology. Comput. Aided. Des. Appl, 2022, 19(S6): 158-169. DOI: 10.1080/16864360.2022.2147043

[8] Gautama B H, Hanif R, Maretaniandini S T. Fraud Early Warning System: Identifikasi Potensi Fraud dalam Pelaporan Harta Kekayaan Penyelenggara Negara Berbasis Big Data. Innovative: Journal of Social Science Research, 2023, 3(4): 3117-3131. DOI: 10.36315/2023/v3i4.1243

[9] Jarmulska B. Random forest versus logit models: Which offers better early warning of fiscal stress? Journal of Forecasting, 2022, 41(3): 455-490. DOI: 10.1002/for.2806

[10] Cong W. Study of financial warning ensemble model for listed companies based on unbalanced classification perspective. International Journal of Intelligent Information Technologies (IJIIT), 2020, 16(1): 32-48. DOI: 10.4018/IJIIT.2020010103

[11] Yang L, Zhong Z. Research on Early Warning of Financial Risk of Local Financial Enterprises. Financial Engineering and Risk Management, 2022, 5(6): 27-33. DOI: 10.4236/firm.2022.56003

[12] Wang T, Zhao S, Zhu G, Zheng H. A machine learning-based early warning system for systemic banking crises. Applied economics, 2021, 53(26): 2974-2992. DOI: 10.1080/00036846.2020.1870657

[13] Liu L, Chen C, Wang B. Predicting financial crises with machine learning methods. Journal of Forecasting, 2022, 41(5): 871-910. DOI: 10.1002/for.2840

[14] Sun J, Yin F, Altman E, Makosa L. Effects of corporate financial distress on peer firms: do intra-industry non-distressed firms become more conditionally conservative? Accounting and Business Research, 2023, 53(6): 646-670. DOI: 10.1080/00014788.2023.2221560

[15] An B, Suh Y. Identifying financial statement fraud with decision rules obtained from Modified Random Forest. Data Technologies and Applications, 2020, 54(2): 235-255. DOI: 10.1108/DTA-05-2020-0113

[16] Park D, Ryu D. A machine learning-based early warning system for the housing and stock markets. IEEE Access, 2021, 9(6): 85566-85572. DOI: 10.1109/ACCESS.2021.3077962

[17] Chen J. Construction and Application of an Economic Intelligent Decision-making Platform Based on Artificial Intelligence Technology. Informatica, 2024, 48(9). https://doi.org/10.31449/inf.v48i9.5705

[18] Liu H. Enhanced CoCoSo method for intuitionistic fuzzy MAGDM and application to financial risk evaluation of high-tech enterprises. Informatica, 2024, 48(5). https://doi.org/10.31449/inf.v48i5.5169

[19] Zhang P. Big Data-Driven Threat Intelligence Analysis and Early Warning Model Construction. Journal of Global Humanities and Social Sciences, 2023, 4(04): 171-175. https://doi.org/10.61360/BoniGHSS232014160804

[20] Rehman H. Financial Risks Classification Early Warning Analysis of Data Mining Technology. Journal of Global Humanities and Social Sciences, 2022, 3(3): 57-60. DOI:10.47852/bonviewGHSS2022030305