International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 8413-8425

https://doi.org/10.70517/ijhsa464716

# Application of Random Forest Algorithm in Personalized Financial Decision-Making

**Yulin Lan[1], Haili Lang[1,*] and Lulu Lan[1]**
[1] Weifang Engineering Vocational College, Qingzhou, Shandong, 262500, China
Corresponding authors: (e-mail: c18853617626@163.com).

**Abstract** To assist enterprises in making personalized financial decisions, this paper designs a big data-based financial decision support platform based on the design concept of "data processing-data analysis-data presentation-data decision-making," providing a decision support environment for financial decision-makers. To optimize personalized financial decisions, a random forest algorithm is used to construct an enterprise financial data risk warning model. Sample data and financial risk warning indicators are selected, and the random forest algorithm is used to estimate feature importance. The confusion matrix is employed as the metric standard for financial warning results. The hyperparameters of the random forest model are optimized, including n-tree optimization and mtry selection. Financial indicator data from T-1 year, T-2 year, and T-3 year are extracted separately for risk prediction analysis, and corresponding random forest classification models are constructed based on this. Compare the financial risk prediction accuracy rates of each model to validate the feasibility of the random forest algorithm as a key technology for a big data financial decision support platform. For T-1 year data, the enterprise financial data risk warning model based on the random forest algorithm demonstrates the best predictive performance, with accuracy and recall rates both exceeding 90%, with accuracy as high as 97%.

**Index Terms** random forest algorithm, financial decision-making, financial risk warning, confusion matrix

## I. Introduction

In the era of big data, faced with significant changes in the social and business environments, many companies, institutions, and individuals are actively seeking breakthroughs and digital transformation to achieve value creation and efficiency improvements [1], [2]. In the modern business environment, sound financial decisions can help businesses, institutions, families, and individuals increase revenue, reduce costs, improve profits, and drive development [3]-[5]. In the context of financial digital transformation, the role of finance needs to transition from that of an accountant to a strategic advisor, and the function of finance needs to shift from measuring value to creating value. Therefore, how to conduct in-depth analysis of massive, multi-source, and heterogeneous business and financial data, making "data assets" a key component in optimizing resource allocation, and ultimately achieving the goals of enhancing financial insight, improving financial operational efficiency, and better empowering business operations, is a critical issue that needs to be addressed in the financial decision-making process [6]-[10].

With the continuous development of technologies such as the internet, the internet of things, and mobile internet, a massive amount of data has been generated globally. In corporate and institutional financial decision-making, traditional manual analysis models can no longer adapt to the growing volume of data and complex business environments. By rapidly storing, processing, and analyzing large amounts of data, more accurate and comprehensive information support can be provided to identify market trends, uncover business opportunities, and achieve optimal allocation of financial resources and value maximization [11]-[14]. In household and personal financial decision-making, the accuracy of decisions is significantly impacted by the expanding scale of relevant decision data and insufficient understanding, leading to financial losses for individuals [15], [16]. Personalized financial decision-making can enhance decision accuracy, strengthen financial risk management, and maximize value, which is of great significance for improving competitiveness and achieving sustainable development.

Literature [17] describes a personalized financial planning system with network attack detection capabilities built in a cloud environment using algorithms such as autoencoders, support vector machines, collaborative filtering, and reinforcement learning, thereby enhancing financial security and user service satisfaction. Literature [18] proposes applying artificial intelligence and machine learning algorithms to comprehensively analyze relevant investment information, providing personalized financial plans and financial services for personal finances. Literature [19] developed a personalized financial configurator based on a three-tier distributed system for personal pension-related decisions, enabling the rapid generation of recommendations such as retirement decision advice, future

asset changes, and investment decision advice. Current personalized financial decision-making lacks targeted financial advice and insufficient differentiation of customer groups, limiting the accuracy of financial decisions for both businesses and individuals.

The random forest algorithm is a powerful machine learning technique widely applied in classification, regression, and other machine learning tasks. Its core idea is to combine multiple decision trees to enhance the model's accuracy and stability, achieving more precise predictions [20]. In finance, the random forest algorithm plays a significant role. Literature [21] improves the random forest algorithm using pruning methods and oversampling techniques, thereby constructing a system to identify and warn of corporate financial distress, promoting sustainable corporate development. Literature [22] analyzed the performance of the random forest algorithm in predicting long-term and short-term stock price trends and used the prediction results for stock selection. Literature [23] applied the random forest algorithm to detect credit card fraud, achieving 98.6% accuracy due to its multi-decision tree characteristics. Literature [24] explored the application performance of the random forest algorithm in investment decision-making at the Damascus Stock Exchange, which can not only effectively classify investment decisions but also identify investment opportunities.

This paper combines the development and application of big data in financial decision-making, utilizing the random forest algorithm to construct a financial data risk warning model for enterprises, and designing an enterprise financial decision support platform composed of data processing, data analysis, data presentation, and data-driven personalized decision-making. The paper analyzes the process and generalization error of the random forest algorithm and optimizes the hyperparameters of the random forest algorithm model. It screens sample data, constructs a risk feature model for corporate financial data based on the random forest algorithm, analyzes financial risk warning indicators, and proposes measurement indicators for financial risk warning results. It also analyzes the effectiveness of financial risk warning based on the random forest algorithm and compares the prediction accuracy of various models.

## II. Development of corporate financial decision-making under big data technology conditions

### *II. A. Application of Big Data in Financial Decision-Making*

In recent years, big data has brought revolutionary changes to various industries and had a profound impact on financial decision-making. Its role in shaping the future of finance is undeniable. The ability to collect, process, and analyze large amounts of data has opened up numerous applications in financial decision-making.

#### II. A. 1) Risk Management

Risk management is an important aspect of financial work, and the application of big data has completely transformed the way financial institutions identify, assess, and mitigate risks. Big data analysis has become an indispensable tool for risk managers, enabling them to make more informed decisions and enhance the overall stability of the financial system [25].

(1) Identifying potential risks. Big data allows financial institutions to identify potential risks by analyzing large datasets from various sources.

(2) Predictive modeling for risk assessment. Big data analytics employs predictive modeling techniques to assess risks more accurately. Machine learning algorithms can analyze historical data to predict future market trends and potential disruptions, enabling institutions to anticipate market crashes, credit defaults, or other adverse events and take proactive measures.

(3) Developing effective risk mitigation strategies. Big data-driven risk management enables organizations to develop more effective risk mitigation strategies. By understanding the underlying factors contributing to risks, institutions can adjust their risk management approaches.

(4) Stress testing. Stress testing is a critical component of risk management, particularly for regulatory compliance. Big data facilitates advanced stress testing by simulating various scenarios and assessing their impact on financial portfolios, helping institutions evaluate their resilience to adverse conditions and make necessary adjustments to mitigate potential losses.

(5) Fraud detection and prevention. In addition to market-related risks, big data also plays a significant role in fraud detection and prevention. By analyzing transaction data and detecting abnormal patterns or anomalies, financial institutions can quickly identify fraudulent activities and take immediate action to prevent losses.

(6) Compliance and regulatory reporting. Adhering to regulatory requirements is an important aspect of risk management. Big data solutions simplify compliance efforts by automating data collection, verification, and reporting, ensuring that institutions effectively and accurately fulfill their regulatory obligations.

#### II. A. 2) Financial Forecasting and Budgeting

Financial forecasting and budgeting are fundamental processes for managing an organization's financial health and making strategic decisions. Big data has revolutionized these areas by providing massive and diverse datasets, enabling more accurate predictions, optimizing resource allocation, and enhancing financial planning. The following are some key applications of big data in financial forecasting and budgeting.

First, predicting market trends and future financial conditions.
Second, optimizing budget allocation.
Third, scenario planning and sensitivity analysis.
Fourth, real-time financial monitoring.
Fifth, customer segmentation and revenue forecasting.
Sixth, supply chain optimization.

#### II. A. 3) Fraud Detection and Prevention

Fraud remains a major challenge for businesses across all industries, resulting in financial losses and damage to reputation. The application of big data has become a game-changer in the field of fraud detection and prevention. By leveraging large datasets and advanced analytical techniques, organizations can proactively identify and reduce fraudulent activities. The following are some key ways in which big data is applied to fraud detection and prevention.

First, identifying patterns and anomalies.
Second, real-time monitoring.
Third, machine learning and predictive modeling.
Fourth, behavioral analysis.
Fifth, network analysis.
Sixth, continuous improvement.

### *II. B.Design of a Big Data-Based Financial Decision Support Platform*

#### II. B. 1) General Requirements for Platform Design

The Financial Decision Support System Platform is a decision support environment designed to assist financial decision-makers in problem analysis, model construction, decision-making process simulation, and evaluation of decision outcomes. It primarily revolves around human-computer interaction systems.

This system requires businesses to incorporate various technologies, such as model libraries, neural networks, and artificial intelligence, to analyze valuable information within accounting systems and construct various economic models. Based on experts' past experience, the system uses technologies like model libraries to make accurate predictions about financial conditions with a focus on intelligence.

With this system, managers not only feel less pressure, but they can also make decisions more efficiently and with higher quality. Plus, the decision support system helps companies stay on top of market trends and boost their competitiveness in the market.

#### II. B. 2) Key Technologies in Platform Design

First, massive data storage and computing. In today's era of information explosion, data types are no longer limited to structured and semi-structured data. The scale of unstructured data continues to grow, and companies need to pay attention to and process not only internal data but also external data.

Second, data visualization is where data assets are truly transformed into value. Big data analysis systems should not only be able to display data, but also identify data characteristics and intelligently recommend relevant charts to users. They can also assemble different visualizations by business to form reporting and decision-making recommendations.

Third, text mining, natural language processing, and machine learning must fully explore the intrinsic characteristics of data to achieve the effects of speed, comprehensiveness, novelty, and accuracy. This requires not only the stacking and listing of algorithms themselves, but also an understanding of the essential characteristics of the enterprise's business and an iterative machine learning system.

Additionally, the construction and management of big data analytics systems involve more than just the installation and deployment of server hardware, networks, and software. More importantly, they require the operational platform capabilities to support the deployment, scheduling, monitoring, and troubleshooting of large-scale clusters.

#### II. B. 3) Overall framework of platform design

This paper is based on the specific functions achieved by big data technology in the application of corporate financial decision-making, combined with the current status of existing intelligent big data platforms (business intelligence)

and big data analysis clouds. The concept of constructing a financial decision support system for big data platforms is: “data processing-data analysis-data presentation-data decision-making.”

The big data financial decision-making support system platform model primarily includes the following components:

(1) Data Processing

The front-end component of data processing is data acquisition. Big data technology enables enterprises to obtain more useful data information, primarily through the integration of multi-source heterogeneous data.

(2) Data Analysis

Data analysis is a critical component. It involves converting external data obtained through big data technology into usable data through certain preprocessing methods, thereby transforming it into potentially valuable data. This process also involves deep data mining.

(3) Data Presentation

Data-driven decision-making is the pinnacle of platform design. Before making financial decisions, it is essential to first understand what is happening and why. This is descriptive analysis, which involves basic statistical analysis of the company's financial statements and daily operational data. It also includes comparisons of metrics under the same dimensions, with common financial metric comparisons including time-based comparisons, spatial comparisons, and standard comparisons. In this paper, the random forest algorithm is used to analyze data for early warning of financial risks, and based on this, personalized financial decisions aligned with the company's development are made.

## III. Construction and application of enterprise financial data risk warning models

### *III. A. Random Forest*

The Random Forest algorithm (RF) was proposed by combining the Classification and Regression Tree (CART) algorithm with the Bagging algorithm. The Random Forest algorithm is an ensemble learning algorithm based on decision trees, combining the Bagging ensemble learning theory with the random subspace method. It involves training multiple CART models using the Bagging ensemble technique and forming a collection of these models. When inputting samples to be classified, the final classification result is determined by voting on the output results of each decision tree. The Random Forest algorithm achieves high accuracy, exhibits excellent scalability and parallelism for high-dimensional data classification problems, demonstrates strong tolerance for outliers and noise, and is less prone to overfitting [26], [27].

#### III. A. 1) Algorithm Flow

A random forest is a classifier composed of a series of CARTs. The algorithm flow is described as follows:

Step 1: Use a random sampling method with replacement to extract K training sample sets from the original sample set of size N, with each sample set having the same number of samples as the original sample set.

Step 2: Randomly select m attributes (m << M) from the attribute set M. Train the decision trees using the optimal feature variables selected from the variables.

Step 3: Train the selected K training sample sets. Each sample generates one CART, resulting in a total of K decision trees.

Step 4: Combine the classification results of the K decision trees and use simple majority voting to determine the final classification result.

#### III. A. 2) Generalization error

Generalization error is the error of a model on a new sample set (test set) and is an important indicator for evaluating the quality of a model.

(1) Definition and proof of convergence

Let $\{h_l(X), h_2(X), ..., h_k(X)\}$ is a sequence of $k$ given classifier combinations, $x$ is the input vector, and $Y$ is the corresponding output random vector, then the margin function of the sample point $\{x, y\}$ is defined as follows:

$$mg(X,Y) = av_k I(h_k(X) = Y) - \max_{j \neq Y} av_k I(h_k(X) = j) \quad (1)$$

Among these, $I(\cdot)$ is the characteristic function, and $av_k(\cdot)$ is the average value of the obtained function values. $av_k I(h_k(X) = Y)$ is the average number of votes for assigning the unclassified sample $x$ to the correct category. $av_k I(h_k(X) = j)$ is the average number of votes for assigning the unclassified sample $x$ to the incorrect category.

Therefore, the margin function $mg(X,Y)$ is the minimum difference between the average number of votes for correct classification and the average number of votes for incorrect classification in the classifier set. The smaller

the value of the margin function, the worse the classification performance of the random forest model, and vice versa.

Let the generalization error of the classifier combination be $PE^*$, then it can be defined as:

$$PE^* = P_{X,Y}(mg(X,Y) < 0) \tag{2}$$

Among these, $P_{x,y}(\cdot)$ represents the probability value obtained in the $X$ and $Y$ spaces.

Let $\Theta$ denote the random vector corresponding to each decision tree, and let $h(X,\Theta)$ denote the classifier output vector of $X$ and $\Theta$. Then, in the random forest algorithm, $h_k(X) = h(X,\Theta_k)$. When the number of trees in the forest becomes sufficiently large, equation (2) follows the strong law of large numbers. Thus, the following theorem can be obtained:

Theorem 1: As the number of trees increases, for all parameter sequences $\{\Theta_1,\Theta_2,\Theta_3,\cdots,\Theta_k\}$, the generalization error $PE^*$ converges everywhere to:

$$P_{X,Y}((P_\Theta h(X,\Theta) = Y - \max_{j\neq Y} P_\Theta(h(X,\Theta) = j)) < 0) \tag{3}$$

This indicates that the generalization error of random forests converges to a limit value, meaning that overfitting does not occur as the number of trees increases.

(2) Classification performance and correlation

The upper bound of the generalization error of random forests is primarily influenced by two parameters: the classification strength of a single tree and the degree of correlation between trees. Specifically, the generalization error is inversely proportional to the classification strength and directly proportional to the correlation. A theoretical analysis of these two interacting factors can provide a better understanding of how random forests work.

Definition 1: The margin function expression for a sample point $(x,y)$ in the random forest algorithm is:

$$mr(X,Y) = P_\Theta(h(X,\Theta) = Y) - \max_{j\neq Y} P_\Theta(h(X,\Theta) = j) \tag{4}$$

Classification efficiency can be expressed as:

$$s = E_{X,Y} mr(X,Y) \tag{5}$$

Among them, $E_{x,y}(\cdot)$ is the expected value obtained in the $X,Y$ space.

Assuming that $s \geq 0$, it can be seen from Chebyshev's inequality that:

$$PE^* \leq var(mr)/s^2 \tag{6}$$

Among them, $var(mr)$ is the variance of the random forest margin function $mr(X,Y)$ on the random forest. $var(mr)$ also has a more intuitive expression, which is derived as follows:

Let $\hat{j}(X,Y) = \arg\max_{j\neq Y} P_\Theta(h(X,\Theta) = j)$, therefore:

$$\begin{aligned} mr(X,Y) &= P_\Theta(h(X,\Theta) = Y) - P_\Theta(h(X,\Theta) = \hat{j}(X,Y)) \\ &= E_\Theta[I(h(X,\Theta) = Y) - I(h(X,\Theta) = \hat{j}(X,Y))] \end{aligned} \tag{7}$$

Definition 2: The margin function of the meta-classifier is:

$$rmg(\Theta,X,Y) = I(h(X,\Theta) = Y) - I(h(X,\Theta)) = \hat{j}(X,Y) \tag{8}$$

Therefore, $mr(X,Y)$ is the expected value of $rmg(\Theta,X,Y)$ on $\Theta$.

For any function $f$, the following identity holds:

$$[E_\Theta f(\Theta)]^2 = E_{\Theta,\Theta'} f(\Theta) f(\Theta') \tag{9}$$

Among them, $\Theta,\Theta'$ are independently and identically distributed.

Therefore:

$$mr(X,Y)^2 = E_{\Theta,\Theta'} rmg(\Theta,X,Y) rmg(\Theta',X,Y) \tag{10}$$

From equation (10), we obtain:

$$\begin{aligned} var(mr) &= E_{\Theta,\Theta'}(cov_{X,Y} rmg(\Theta,X,Y) rmg(\Theta',X,Y)) \\ &= E_{\Theta,\Theta'}(\rho(\Theta,\Theta') sd(\Theta) sd(\Theta')) \end{aligned} \tag{11}$$

Among them, $\rho(\Theta,\Theta')$ is the correlation between $rmg(\Theta,X,Y)$ and $rmg(\Theta',X,Y)$ when $\Theta,\Theta'$ are fixed, and $sd(\Theta)$ is the standard deviation of $rmg(\Theta,X,Y)$ when $\Theta$ is fixed.

Therefore:

$$var(mr) = \bar{\rho}(E_\Theta sd(\Theta))^2 \leq \bar{\rho} E_\Theta var(\Theta) \tag{12}$$

Among them, $\bar{\rho}$ is the average value of the correlation coefficient $\rho$:

$$\bar{\rho} = E_{\Theta,\Theta}(\rho(\Theta,\Theta')sd(\Theta)sd(\Theta')) / E_{\Theta,\Theta'}(sd(\Theta)sd(\Theta')) \tag{13}$$

$E_\Theta var(\Theta)$ satisfies the following inequality:

$$E_\Theta var(\Theta) \le E_\Theta(E_{X,Y} rmg(\Theta, X, Y))^2 - s^2 \le 1 - s^2 \tag{14}$$

Combining (14) and (15), we obtain an upper bound for the variance $var(mr)$:

$$var(mr) \le \bar{\rho}(1 - s^2) \tag{15}$$

Combining (4) and (15), we obtain the following theorem.

Theorem 2: The upper bound formula for generalization error is:

$$PE^* \le \bar{\rho}(1 - s^2) / s^2 \tag{16}$$

In the formula, when the classification strength s of a single tree is increased and the correlation $\rho$ between decision trees is decreased, the generalization performance of the classifier can be improved. The above derivation proves that when the number of trees in a random forest increases, the algorithm will not exhibit overfitting, and the generalization error of the algorithm will converge to a limit value.

### *III. B. Construction of a risk characteristic model for corporate financial data*

#### III. B. 1) Financial Data Risk Model Based on Random Forest

In order to more accurately identify financial data risk information for enterprises, it is necessary to collect different samples of financial information data. Based on the different attribute characteristics of the samples, the random forest algorithm is introduced to construct a feature model for the financial samples.

After the samples are extracted, a financial feature model is constructed based on the random forest algorithm. According to the attribute differences between different data, data risk components are constructed separately. The information features corresponding to each set of component data can be expressed as:

$$x_i = x_i^N + x_i^{AN} \tag{17}$$

In the equation, $x_i$ represents the characteristic component of one type of financial information report. $x_i^{AN}$ represents the financial information component with risk characteristics. $x_i^N$ represents the financial data component with multiple attribute characteristics under normal conditions. $N$ represents the total number of different attribute financial information characteristic reports included in the model.

To more accurately identify the distribution range of financial information risk characteristics, in the decision space constructed by the random forest, different attribute financial feature information components are clustered using random clustering, resulting in a corporate financial data risk feature model based on the random forest algorithm:

$$\mu = \frac{1}{|S_i|}\sum_{t \in S_i} t \tag{18}$$

In the formula, $\mu$ represents the results obtained after clustering the components of financial information characteristics with different attributes. $S_i$ represents the set of financial risk information characteristics obtained after clustering decisions under the random forest algorithm. $t$ represents the set obtained after dividing the financial risk feature information differences.

#### III. B. 2) Financial risk warning indicators based on random forests

To make risk prediction more accurate, we use random sampling to pick decision samples when building the model, and then use the sampled data to build a decision tree. Based on the classification results of the different branches of the decision tree, we bid on financial risk feature information to determine the early warning indicators under the decision tree random forest model. That is:

$$H_r = argmax\sum_{m=I}^{M}[Ih(X, \Theta_m) \times j] \tag{19}$$

In the formula, $H_r$ represents the random forest risk feature decision model indicator. $h$ represents the financial risk prediction result of the decision tree. $M$ represents the maximum number of features of the decision tree. $x$ represents the maximum number of leaf nodes of the decision tree, $\Theta_m$ represents the maximum depth of the decision tree, and $j$ represents the number of decision trees.

To avoid the issue of training homogenization caused by identical training samples each time, during the training process, a portion of the sample subset is retained and not included in the training, and this retained sample subset is defined as out-of-circle data. Based on this rule, each training randomly selects $m$ decision feature samples,

and the remaining decision feature data are all out-of-circle data. After training, a high-precision financial risk warning indicator is obtained, and the calculation formula is:

$$A = (T + G) / (T + F + G + N^{'}) \tag{20}$$

In the formula, $A$ represents the warning indicator characteristics obtained after training. $T$ represents the normal indicator coefficient, $G$ represents the risk indicator coefficient. $F$ represents the reverse risk decision indicator, and $N^{'}$ represents the forward risk decision indicator.

#### III. B. 3) Financial risk warning based on random forests

To ensure the accuracy and stability of risk prediction, simplify the random forest structure. Set the total amount of risk feature data to be optimized as $N$, extract $Y$ risk features to form the training set, and use the data with the most obvious risk features as the regression tree splitting features for splitting training. After training, the risk feature sets correspond to different feature regression trees, and the number of sets is the same as the number of extracted features, i.e., $Y$.

After multiple splitting training sessions, the important feature functions of the financial risk data are obtained as follows:

$$I(x_i) = \sum_{i=1}^{Y} \frac{errOOB2(x_i) - errOOB1(x_i)}{Y} \tag{21}$$

In the equation, $Y$ represents the number of regression trees. $errOOB1(x_i)$ represents the error obtained from the first split of the training data. $errOOB2(x_i)$ represents the error obtained from the training data after perturbation. $I(x_i)$ represents the importance of the feature $x_i$ corresponding to the financial risk data. The coefficient corresponding to $I(x_i)$ is positively correlated with the importance level. When $I(x_i) < 0$, the corresponding feature $x_i$ is the optimal coefficient for the warning output. At this point, the equation represents the output function for corporate financial risk prediction under the random forest algorithm.

### *III. C. Sample Data and Selection of Early Warning Indicators*

As of 2024, there are more than 3,000 companies listed on the Shanghai Stock Exchange and Shenzhen Stock Exchange, with over 1,500 listed on the Shanghai Stock Exchange and over 2,000 listed on the Shenzhen Stock Exchange. This chapter selects sample companies experiencing financial and non-financial difficulties among listed companies based on principles of objectivity and systematicity.

This paper selects the year in which the listed company was subject to special treatment as the base period, i.e., period t. The base period data for this paper is sourced from 2024. Therefore, the year prior to ST designation, 2023, is designated as t-1, 2022 as t-2, and 2021 as t-3.

Based on the basic information of listed companies in the Guotai An database and the 2024 trading situation announcements of listed companies, there were a total of 180 listed companies subject to special treatment in 2024. After handling missing values, only 132 ST companies had complete data. These 132 ST companies span 13 industries classified by the China Securities Regulatory Commission.

This paper uses financial data from the t-3, t-2, and t-1 periods to predict the financial distress status of listed companies in the base period. A matching sampling method is employed to select non-distressed companies in the same industry as the ST companies. However, due to the extremely small number of ST samples, the number of companies subject to special treatment in 2024 was less than 200, accounting for approximately 6% of all listed companies, indicating a significant disparity in the number of companies between the two categories.

Therefore, to reduce the disparity in sample sizes, this paper ultimately selected 300 non-financially distressed companies matched by industry to pair with the 132 financially distressed companies.

#### III. C. 1) Random Forest Estimation of Feature Importance

Random forest models are not only applicable to classification and regression problems but also perform exceptionally well in feature selection. This is due to the random forest model's ability to measure the importance of features after training the model. The bootstrap sampling method used in random forests involves randomly selecting a portion of the samples each time, with the unselected samples constituting the out-of-bag (OOB) data. Each decision tree in the random forest can calculate its error using the corresponding OOB data. At this point, noise is randomly added to a specific feature in the out-of-bag data, and the out-of-bag error of the decision tree is recalculated. If the out-of-bag accuracy significantly decreases after randomly adding noise to a specific feature, it indicates that this feature is highly important for the classification results of the samples. The sum of the importance

scores of all features is 1, and features with higher importance scores are more critical for the accuracy of the prediction results.

**III. C. 2) Selection of early warning indicators based on random forests**

Feature selection based on random forests involves the following steps:

First, train the model using the original feature set and calculate the importance of each feature.

Second, sort the features by importance and remove the least important feature or features.

Third, retrain the model using the updated feature set.

Fourth, repeat the above three steps until the number of features is reduced to the specified number.

As the number of features continues to decrease, the predictive accuracy of the randomly trained forest model will also change accordingly. For a given task, features can be categorized based on their relationship to the task into: relevant features, irrelevant features, and redundant features. Relevant features have a certain association with the task and can provide effective information for achieving the task objectives. Irrelevant features, on the other hand, are completely different from relevant features and have no connection to the task objectives. Redundant features contain information that is covered by other features, and the presence of such features may reduce the model's predictive performance. Ideally, as the number of features continues to decrease, redundant features and irrelevant features are gradually removed, making the training model simpler and improving the model's predictive accuracy. However, when the number of features is reduced to a certain extent, relevant features are deleted, leading to a continuous decline in model accuracy.

After training the model in this paper, the least important features are removed. Each time the model is trained, the training dataset is used to test the model's predictive accuracy.

The relationship between the number of features and the accuracy of the early warning model is shown in Figure 1. The following figure shows the predictive accuracy of the early warning model trained using data from 2023 and 2024 at different feature counts.

From the trend of the bar chart, it can be seen that as the number of samples decreases, the model accuracy continues to decline. The two highest values in 2024 occurred when the number of features was 23 and 10, with accuracy values of 0.782 and 0.757, respectively.

The 2023 data showed that the accuracy of the early warning model began to decline overall when the number of features was 9, which is similar to the test results of the 2024 data. Based on this, the paper selected the 9 most important features in 2024 to construct the dataset.

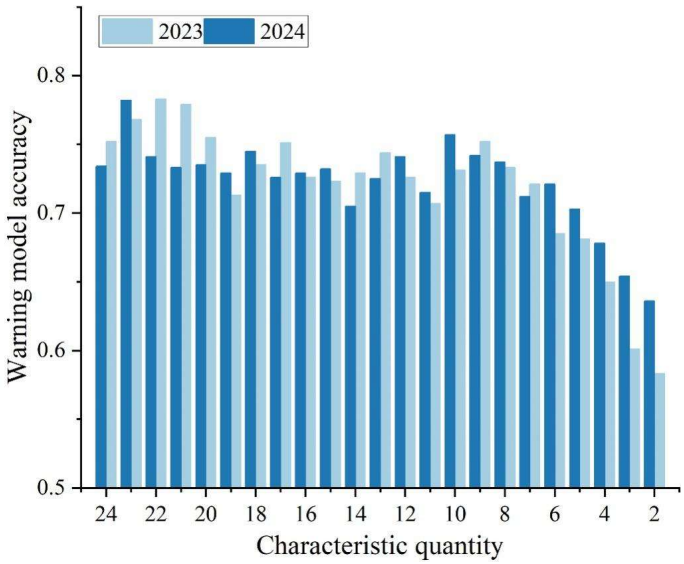


Figure 1: The number of characteristics and the accuracy of the warning model

The feature importance and descriptive statistics are shown in Table 1.

Table 1: Characteristic importance and descriptive statistics

| Index code | Importance | Mean | Standard deviation | Minimum value | Maximum value |
|---|---|---|---|---|---|
| D1 | 0.1536 | -0.0556 | 5.2631 | -134.5124 | 23.2546 |
| D2 | 0.1332 | 2.1361 | 182.4424 | -2526.2416 | 5.6934 |
| D3 | 0.1015 | 1.3348 | 60.9613 | -142.8693 | 1.2215 |
| D4 | 0.0967 | 0.0621 | 0.3629 | -20.5549 | 3.5249 |
| D5 | 0.0723 | 0.1175 | 0.7548 | -0.7823 | 263.5887 |
| D6 | 0.0691 | 0.4593 | 0.2433 | 0.0125 | 110.2234 |
| D7 | 0.0565 | 0.0522 | 12.5423 | 23.7432 | 3.5046 |
| D8 | 0.0332 | 1.9614 | 1.5961 | -182.0015 | 2301.0093 |
| D9 | 0.0214 | 0.8002 | 4.7443 | 0.03249 | 261.0248 |

#### III. C. 3) Early warning result measurement indicators

A confusion matrix, also known as a probability table or error matrix, is a matrix used to indicate the performance of a warning model or classifier. Take binary classification in a classification model as an example, where the established model judges the result of a sample to be either 0 or 1.

Accuracy rate: The accuracy rate refers to the proportion of all correct judgments in a classification model out of the total number of observations. The higher the accuracy rate of a model, the better its performance. The calculation formula is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{22}$$

Precision rate: The precision rate refers to the proportion of correct predictions among all results predicted as “1” by the model. The higher the value of this indicator, the better. In this paper, considering the weights, it can be expressed as the ratio of correct predictions to the total number of predictions for that class. The calculation formula is as follows:

$$Precision = \frac{TP}{TP + FP} \tag{23}$$

Recall rate: The recall rate refers to the proportion of results correctly predicted by the classification model among all results with a true value of “1.” It represents the probability that an event actually occurred and was correctly predicted. The higher this metric, the better.

This metric can be expressed as: the proportion of correctly predicted results among the total number of true results for a given class, taking weights into account. The calculation formula is as follows:

$$Recall = \frac{TP}{TP + FN} \tag{24}$$

F1-Score: This metric is the harmonic mean of precision and recall, combining the results of precision and recall. The n-Score ranges from 0 to 1, with higher values indicating a more robust model and better classifier performance. That is:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{25}$$

### *III. D. Analysis of a Random Forest-Based Financial Risk Early Warning Model for Companies*

#### III. D. 1) Random Forest Hyperparameter Optimization

Here, this paper optimizes the hyperparameters of the random forest.

(1) ntree optimization

In the RF algorithm, the number of trees (ntree) determines the upper limit of the generalization error and prevents the model from overfitting. Therefore, selecting an appropriate ntree value can improve classification performance and reduce model training time. Set each ntree value to 20-100 with a step size of 20, build the model, and observe the changes in the AUC value. The AUC value assesses the performance of the classifier, with a range of [0,1], where higher values are better.

In this paper, the GridSearchCV module in Python is used to set the corresponding grid parameters for n-tree optimization.

The relationship between the NTREE parameter values and the AUC area is shown in Figure 2. The horizontal axis represents the number of trees, and the vertical axis represents the AUC value of the model samples under different parameters.

It is observed that from 0 to 100 trees, the AUC trend first increases and then decreases, indicating that the number of trees should be increased. From 150 to 200, the AUC increases significantly, and the slope of the curve steepens sharply. At 200, the value exceeds 0.875. At this point, increasing the number of trees does not change the AUC value. From 200 to 1000, the AUC value remains unchanged and tends to stabilize. Thus, n-tree is set to 200.

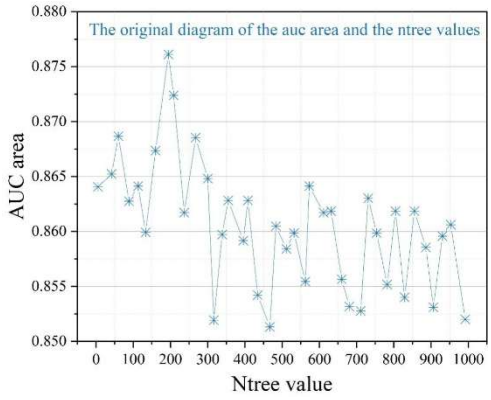


Figure 2: Parameter Ntree values and AUC area value diagram

(2) mtry optimization

In each branch, mtry variables are randomly selected, which increases the diversity between trees and improves the model's noise tolerance and generalization ability. Therefore, this parameter should be selected appropriately.

In the previous section, we optimized ntree and used its results to optimize mtry. We still selected the parameter based on the maximum AUC value and adjusted the mtry value to detect changes in the AUC value.

We continued to use the GridSearchCV module in Python and set the corresponding grid parameters to optimize mtry. The relationship between the MTRY value and the AUC area is shown in Figure 3.

Observing the figure, when mtry is in the range [0,5], the AUC area decreases, indicating that increasing the number of variables within this range does not enhance generalization or robustness. Therefore, the mtry value should be increased. When mtry is 6, the AUC reaches its maximum value of 0.896, and as mtry increases, the AUC value decreases. Thus, mtry is set to 6.

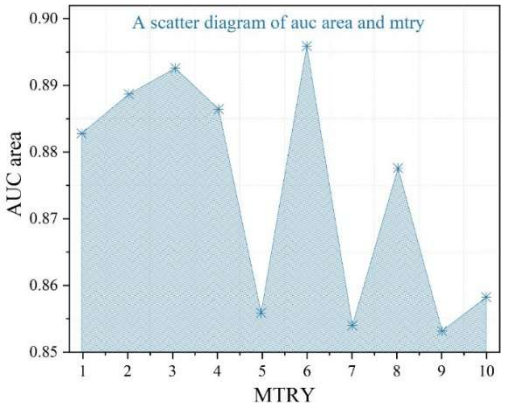


Figure 3: Parameter MTRY values and AUC area value diagram

**III. D. 2) Comparison of prediction results for different intervals**

This paper uses whether a company is labeled as “ST” as an indicator of its financial risk. Assuming that the time point at which a company is labeled as “ST” is T, financial indicator data from the previous three years (T-1, T-2, and T-3) are extracted to conduct risk prediction analysis, and a corresponding random forest classification model is constructed based on this data.

The prediction results of the random forest models with different time intervals are shown in Figure 4, which displays the prediction results on the test set. The model using data from T-1 (i.e., the year before the designation)

demonstrates the best prediction performance, with an accuracy rate as high as 97% and a recall rate of 91%. For T-2, the model's accuracy rate and recall rate are 0.9552 and 0.4869, respectively.

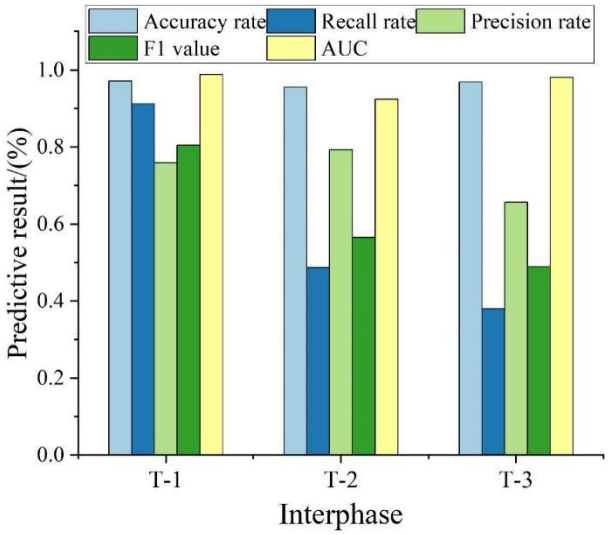


Figure 4: The results of the random forest model of different interphase periods

**III. D. 3) Evaluation of model results**

In addition to tuning the ntree and mtry parameters, this paper continues to use GridSearchCV to tune other detailed parameters. A final random forest early warning model for the financial risk of listed companies is formed, and graphviz in Python is used to draw each decision tree.

The analysis of the RF model's early warning results is shown in Table 2. As can be seen from the table, the model demonstrates excellent overall classification capability. The classification accuracy on the training set is 99.07%, which validates the RF model's high fitting power. The comprehensive accuracy on the prediction set is 97.01%, indicating that the trained model possesses strong identification and generalization capabilities and is suitable for widespread use.

Observing the detailed classification metrics, the RF model's true positive rate is 100% in both the training and test sets, indicating that it can effectively identify financially healthy companies. Although there are cases in the test set where financially stable companies are classified as crisis-prone enterprises, from a practical perspective, over-identification is beneficial for enterprise risk warning, prompting company management and investors to strengthen risk control.

Table 2: RF model warning analysis

| | Training RF model | | Total | Prediction RF model | | Total |
|---|---|---|---|---|---|---|
| | Crisis company | Normal company | | Crisis company | Normal company | |
| Crisis company | 128 | 4 | 132 | 30 | 4 | 34 |
| Normal company | 0 | 300 | 300 | 0 | 100 | 100 |
| Accuracy % | 96.97% | 100% | 99.07% | 88.24% | 100% | 97.01% |
| Miscalculation rate % | 3.03% | 0.00% | 0.93% | 11.76% | 0.00% | 2.99% |

**III. D. 4) Comparison of model test results**

The financial risk warning system for listed real estate companies was input into support vector machine and decision tree models to build corresponding warning models, and the prediction results of different models were obtained for comparison and analysis.

First, a confusion matrix was used for comparison and analysis, which directly showed the prediction effects of different models. In terms of misclassifying positive and negative samples as negative and positive samples, respectively, the RF model performed best, with each type of error occurring only once, while the decision tree model had the highest number of misclassifications, with each type of error occurring four times.

The comparison of the confusion matrices for each model is shown in Table 3. Comparing the model evaluation metrics obtained from the confusion matrix, the support vector machine, decision tree, and random forest models were evaluated. The random forest model achieved the highest accuracy, precision, recall, F1 score, and AUC value,

while the decision tree model had slightly poorer predictive performance. Overall, the RF model outperformed other models across all metrics. Therefore, the random forest algorithm can be incorporated into the design of a big data-based financial decision support platform to meet the service requirements of the financial decision platform and make critical judgments for personalized decision-making.

Table 3: Comparison of the model confusion matrix

| Confusion matrix | | Prediction category | |
|---|---|---|---|
| | | Positive class | Negative class |
| Decision tree | Positive class | 25 | 15 |
| Real category | Negative class | 15 | 25 |
| SVM | Positive class | 29 | 11 |
| Real category | Negative class | 11 | 29 |
| Logistic regression | Positive class | 30 | 10 |
| Real category | Negative class | 10 | 30 |
| K neighbor | Positive class | 32 | 8 |
| Real category | Negative class | 8 | 32 |
| Random forest | Positive class | 36 | 4 |
| Real category | Negative class | 4 | 36 |

The prediction accuracy of different models is shown in Figure 5. In practical applications, the logistic regression method has the lowest accuracy in predicting corporate financial data risks, while the random forest model has an accuracy rate of over 95% in predicting corporate financial data risks, making it the most effective in terms of prediction performance.

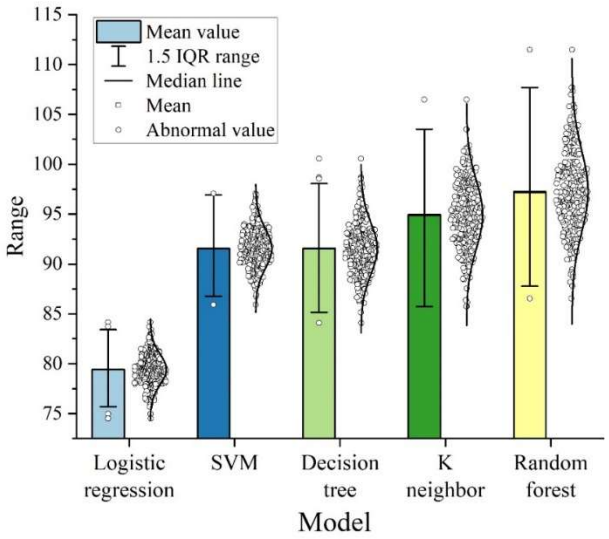


Figure 5: Prediction of different models

## IV. Conclusion

This paper employs the random forest algorithm to establish a financial data risk warning model for enterprises, utilizes the predictive results of the random forest algorithm to formulate financial decisions, and constructs a financial decision support platform for enterprises based on big data technology.

To mitigate the disparity in sample sizes, 300 non-financially distressed companies were selected to match the 132 financially distressed companies. Among these, the random forest algorithm model demonstrated more stable predictions for corporate financial data in 2024, with a maximum prediction accuracy of 0.782. By optimizing the random forest hyperparameters, the corporate financial risk warning model based on the random forest algorithm achieved optimal predictive performance for data intervals of T-1 years, with an accuracy rate as high as 97% and a recall rate of 91%.

Comparing the confusion matrices and prediction accuracy rates of various models, the enterprise financial risk warning model based on the Random Forest algorithm achieves superior risk prediction efficiency. It can be

incorporated into the design of a big data-based financial decision support platform to assist in the personalized formulation of enterprise financial decisions.

## References


[1] Kraus, S., Durst, S., Ferreira, J. J., Veiga, P., Kailer, N., & Weinmann, A. (2022). Digital transformation in business and management research: An overview of the current status quo. International journal of information management, 63, 102466.

[2] Cosa, M. (2024). Business digital transformation: strategy adaptation, communication and future agenda. Journal of Strategy and Management, 17(2), 244-259.

[3] Kim, J., Gutter, M. S., & Spangler, T. (2017). Review of family financial decision making: Suggestions for future research and implications for financial education. Journal of Financial Counseling and Planning, 28(2), 253-267.

[4] Zhou, J., San, O. T., & Liu, Y. (2023). Design and implementation of enterprise financial decision support system based on business intelligence. International Journal of Professional Business Review: Int. J. Prof. Bus. Rev., 8(4), 24.

[5] Mahjub, H., Naderi, A., Kharazi, S. K., & Entezari, Y. (2023). Strategic financial decision making in comprehensive public universities. Quarterly Journal of Research and Planning in Higher Education, 24(2), 53-83.

[6] D'Acunto, F., & Rossi, A. G. (2023). IT meets finance: financial decision-making in the digital era. In Handbook of financial decision making (pp. 336-354). Edward Elgar Publishing.

[7] Rauf, M. A., Shorna, S. A., Joy, Z. H., & Rahman, M. M. (2024). Data-driven transformation: Optimizing enterprise financial management and decision-making with big data. Academic Journal on Business Administration, Innovation & Sustainability, 4(2), 94-106.

[8] Bisht, D., Singh, R., Gehlot, A., Akram, S. V., Singh, A., Montero, E. C., ... & Twala, B. (2022). Imperative role of integrating digitalization in the firms finance: A technological perspective. Electronics, 11(19), 3252.

[9] Thu, N. A., & Quan, T. T. (2023). Impact of digital transformation on financial decision making at Big4 banks in Vietnam. International Journal of Advanced Multidisciplinary Research and Studies, 3(1), 757-766.

[10] Huang, S., & Shi, B. (2024, April). Role and Challenges of Intelligent Data Analysis in Enterprise Financial Decision-Making. In International Conference on Computational Finance and Business Analytics (pp. 15-24). Cham: Springer Nature Switzerland.

[11] Ren, S. (2022). Optimization of Enterprise Financial Management and Decision-Making Systems Based on Big Data. Journal of Mathematics, 2022(1), 1708506.

[12] Li, Z. M., & Liu, Z. A. (2024). Research on intelligent visualization analysis of enterprise financial big data in the age of data intelligence. Proceedings Series, 4(1).

[13] Zhao, X., & Saeed, O. (2022). Intelligent Financial Processing Based on Artificial Intelligence-Assisted Decision Support System. Mobile Information Systems, 2022(1), 6974246.

[14] Wong, A., Holmes, S., & Schaper, M. T. (2018). How do small business owners actually make their financial decisions? Understanding SME financial behaviour using a case-based approach. Small Enterprise Research, 25(1), 36-51.

[15] Salleh, M. C. M., Chowdhury, M. A. M., Nasarudina, A. F. B. M., & Ratnasari, R. T. (2020). The impact of cognitive factors on individuals' financial decisions. Management and Accounting Review, 19(3), 69-88.

[16] Wijayanti, T. C., Naim, S., Hendayani, N., Alfiana, A., & Hanum, F. (2024). Identify the use of economics for family financial management in digital days. Indonesian Interdisciplinary Journal of Sharia Economics (IIJSE), 7(1), 325-345.

[17] Vadisetty, R. (2024). Machine Learning for Personalized Financial Planning on Cloud. Revista de Inteligencia Artificial en Medicina, 15(1), 478-515.

[18] Devan, M., Tillu, R., & Shanmugam, L. (2023). Personalized Financial Recommendations: Real-Time AI-ML Analytics in Wealth Management. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 2(3), 547-559.

[19] Shafiee, S., Zhang, L. L., & Rasmussen, K. M. (2024). Improving financial literacy and supporting financial decisions: Developing a personalized configurator. Journal of the Knowledge Economy, 15(3), 14256-14285.

[20] Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random forest algorithm overview. Babylonian Journal of Machine Learning, 2024, 69-79.

[21] Zhang, J. (2024). Impact of an improved random forest-based financial management model on the effectiveness of corporate sustainability decisions. Systems and Soft Computing, 6, 200102.

[22] Tan, Z., Yan, Z., & Zhu, G. (2019). Stock selection with random forest: An exploitation of excess return in the Chinese stock market. Heliyon, 5(8).

[23] Niveditha, G., Abarna, K., & Akshaya, G. V. (2019). Credit card fraud detection using random forest algorithm. Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol, 5(2), 301-306.

[24] Alakkari, K., & Ali, B. (2025). Using Random Forest Algorithm to Improve Investment Decision Making in Damascus Stock Exchange. EDRAAK, 2025, 57-61.

[25] Miaoyi Zhang. (2025). Analysis of the Dilemma and Countermeasures of Financial Risk Management. Journal of Economic Research,2(2),32-34.

[26] Jia Zhang,Yadong Ge,Yibo Wang,Junyu Tao,Zaixin Li,Shuang Fu... & Guanyi Chen. (2025). Photovoltaic power plants in mountainous area: Environmental impacts analysis based on random forest algorithm. Renewable Energy,254,123670-123670.

[27] Chen Xiangyu & Guo Yonggang. (2025). Seismic Hazard Loss Assessment of Reservoir Dams Based on Random Forest Algorithm. Natural Hazards Review,26(3).