

Research on Real-Time Processing and Risk Management Methods of Enterprise Financial Big Data Based on Distributed Computing Framework

Xin Zheng¹, Lei Zhang^{1,✉}, Chenlu Jia¹, Hongmei Yue¹

¹ Department of Management and Media, Shenyang Institute of Science and Technology, Shenyang, Liaoning, 110167, China

ABSTRACT

The risk of financial aspects intuitively reflects the development status and operating results of enterprises, enterprises must control the financial risk of this key link, so that the financial risk of a safe landing, to protect the stability and health of the enterprise. This paper selects the financial data of listed companies, and comprehensively analyzes the level of the company's financial performance from four aspects, namely, profitability, operating capacity, growth capacity and solvency indicators. Using Benford's law to test the quality of each data of each financial indicator, the Benford factor is introduced as a new explanatory variable, and combined with the company's financial risk early warning indicators to establish a random forest early warning model. The results show that profitability and growth capacity are the strengths of listed companies, while operational capacity and solvency are the weaknesses. The results analyzed by K-means clustering algorithm show that the sample companies are divided into 5 categories. And compared with the basic random forest model, the random forest model based on Benford's law can improve the accuracy of financial risk warning. Finally, the model with the best prediction effect is used to judge the financial status of G listed companies, get the early warning results, verify the accuracy and applicability of the model and put forward corresponding countermeasure suggestions.

Keywords: k-means cluster analysis, corporate risk early warning, random forest model, factor analysis

1. Introduction

The wave of digital transformation has swept across the world, profoundly changing the operation mode and business environment of enterprises. The rapid development of fintech has brought new risks, such as payment security risk, digital currency risk, blockchain technology application risk, etc [1-2]. These new types of risks are intertwined with traditional credit risks, market risks, operational

✉ Corresponding author.

E-mail address: yekongxingchen001@163.com (L. Zhang).

Received 05 November 2024; Revised 20 December 2024; Accepted 15 March 2025; Published Online 16 April 2025.

DOI: [10.61091/jcmcc127b-184](https://doi.org/10.61091/jcmcc127b-184)

© 2025 The Author(s). Published by Combinatorial Press. This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>).

risks, etc. to form a more complex risk matrix [3]. Compared with traditional risks, the occurrence of these new types of risks is often more insidious, difficult to predict, and faster propagation, wider impact, and more serious damage to the enterprise [4-6]. The traditional risk management model has been difficult to cope with the challenges brought by the digital era. Therefore, it is of great significance to construct an enterprise financial risk management system that can adapt to the characteristics of the digitalization era to safeguard the financial security of the enterprise, enhance the competitiveness of the enterprise and promote sustainable development [7-8].

A large number of scholars believe that the business analysis and decision-making support of enterprises rely heavily on big data technology, and through big data analysis, enterprises are able to deeply explore key information such as market trends, consumer behavior and operational efficiency. Literature [9] explores the application of decision tree algorithm and SVM method in enterprise financial risk management, which analyzes the potential financial risks of enterprises, and then stops the enterprise financial crisis from the source, and promotes the smooth development of enterprises in the market economic environment. Literature [10] for enterprise financial management in the face of massive accounting data information problems, designed a data mining algorithm, through the massive accounting data information clustering analysis processing to determine the enterprise financial risk index, significantly improve the efficiency of financial accounting management. Literature [11] constructed the Internet of Things big data application model framework applicable to the enterprise financial risk control information management system, which realizes the networking, informationization and interactive experience of the enterprise and meets the sustainable development needs of the enterprise. Literature [12] used the Internet of Things big data mining method based on fuzzy association rules to analyze multiple enterprise financial risk indicators and provide technical support for the early warning of financial crisis in enterprise operation. Literature [13] examines the role of big data technology in the construction of enterprise financing risk management system, establishes the financial risk analysis model based on principal component analysis and logistic regression algorithm, and effectively improves the ability of enterprises to resist financing risks. Literature [14] analyzes the impact of big data information processing technology on corporate financial decision-making, and the financial information management system based on big data analysis tools is conducive to breaking business and financial barriers, improving the efficiency and quality of corporate decision-making, and demonstrating a strong risk early warning capability. While big data analytics technology improves the efficiency of risk identification, it may also lead to wrong decisions due to algorithmic bias or data bias.

In addition, enterprises will be able to utilize the elastic resources of cloud computing platforms to achieve high efficiency in data processing and analysis. Literature [15] investigated the financial risk exposure analysis model of enterprises based on cloud computing, and its evaluation results were conducive to the reduction of enterprise inventory risk and investment risk, reflecting the characteristic advantages of efficiency, accuracy, effectiveness and security. Literature [16] shows that the application of cloud computing technology in the construction of enterprise financial management information system in the context of big data era can effectively reduce the application threshold of informationization construction and improve the return on investment in informationization, and it also shows a high degree of flexibility in meeting the needs of enterprises at different business stages. Literature [17] integrates financial shared service theory and information security theory to analyze the enterprise financial shared information security problem under cloud computing environment, which not only simplifies the enterprise financial service operation process, but also improves the data storage capacity and computational analysis capacity. Literature [18] proposes a design scheme for cloud computing financial management informatization oriented to the needs of enterprise business services, which improves the accuracy of identification of enterprise financial risks on cloud computing platforms by adding online financial management services and financial early warning modules. Although the wide application of cloud computing improves efficiency, it increases the

dependence on third-party providers, and will directly affect the business operations and data security of enterprises once the cloud service provider has problems. With technological advances, distributed computing has gradually become an important trend in enterprise business analysis and decision-making support. Distributed computing, by decentralizing processing tasks, improving computational speed and reliability, and reducing delays, will help enterprises better cope with the challenges of big data, improve the quality of decision-making, and enhance market competitiveness, however, the application of this technology to enterprise financial risk management is relatively understudied, and it has an important research value.

Combined with the characteristics of enterprises, this paper constructs a financial risk early warning indicator system suitable for manufacturing enterprises from the financial perspectives of profitability, operating ability, growth ability and solvency, R&D investment and organizational governance structure. Considering the test of financial data quality, Benford's factor is constructed, and the precision rate and accuracy rate are used as the evaluation indexes of financial risk early warning model. Secondly, Benford's law is applied to test the quality of each data of each financial indicator, and Benford factor is introduced as a new explanatory variable and combined with the company's financial risk early warning indicators to establish a random forest early warning model.

2. Enterprise financial management based on a distributed computing framework

2.1. Enterprise Finance Big Data Mining

2.1.1. Cloud Computing MapReduce Data Block Storage. Hadoop Distributed File System (HDFS) is utilized to store distributed large-scale dataset files, which is implemented through the Hadoop open source platform. The HDFS framework is shown in Figure 1.

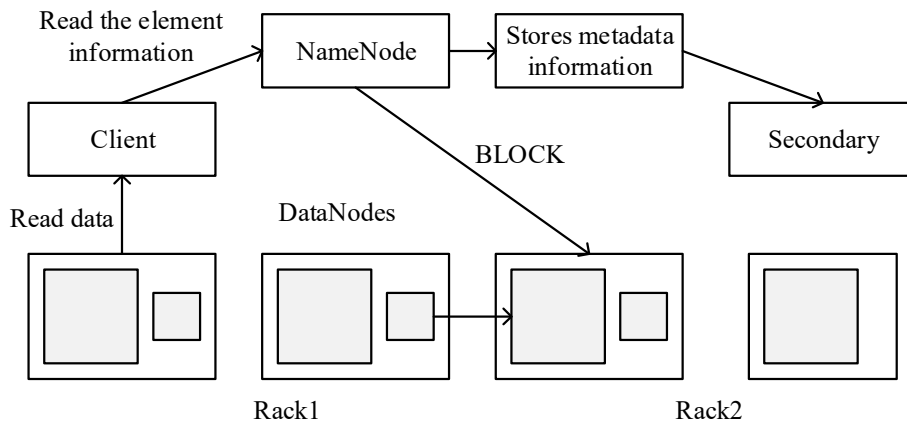


Fig. 1. HDFS framework

A master node NameNode is added to HDFS to store and manage the metadata in the directory files, saving the data after mapping it through storage nodes. At the same time, DataNodes store the BLOCK data in files according to the scheduling commands of the master node, and at the same time constantly report the status of the data blocks to the master node. To ensure the security of the storage environment, it is necessary to add corresponding redundant copies of different data blocks. When the loss of BLOCK in the data node is detected, the master node needs to mark the data block first, and at the same time generate a new data block according to the replica and then store it in the data node. Delete the failed data block and update the mapping relationship on the master node corresponding to the storage node to get a more complete data storage.

2.1.2. **Decision Tree Algorithm Enterprise Financial Data Mining.** In the economic business of the enterprise, financial data are analyzed according to the business characteristics of the enterprise to find the association between the data. Using the decision tree algorithm [19] to classify and organize the massive data, remove redundant data by pruning, and mine according to the analyzed data. Set the generalized value of e ensemble to form a decision tree, if the decision tree does not accurately divide all the data objects need to be repeated after training, in the massive data to extract new data to form a dataset, repeat the original data into the original data until the training of the data obtained by the accurate generalization. Use information gain to evaluate whether the classification of the sample is reasonable. The attribute nodes will be judged by the purity, if the purity in the nodes is higher, it means that the distribution of the nodes is more uniform, and better classification effect is obtained. Set the purity value as y , and the formula for measuring the purity value of the node is:

$$y = -\sum_i^{c-1} p \log_2 p(i|t) \quad (1)$$

where: i is the data of different categories. t is the node; c is the number of categories in the node. Setting the sample set and applying the test attributes to classify them into categories##, the corresponding desired results can be obtained as:

$$Info(S) = \sum_i^k \frac{N(v)}{N} I(v) \quad (2)$$

where: $I(v)$ is the purity value of the child node; N is the record in the sample set; v is the record after classification. The information gain is applied to determine the classification result, and the gain rate is calculated by the formula:

$$G = \frac{\Delta Info}{S} \quad (3)$$

where: S is the segmentation information; $\Delta Info$ is the expected result average. For test results in the same test set, after creating the root node, the test attribute set is viewed and the node is labeled if the input record results indicate the same class.

2.1.3. **Algorithm Parallelization Processing.** When constructing a decision tree, the calculation between attributes of the same node is more independent. In the process of selecting nodes, the information gain rate of all candidate attributes needs to be calculated. By calculating the statistical information of the candidate attributes of the same node in parallel, it is possible to divide the attributes in the sample data set. Parallel computing can shorten the branch attribute selection time and improve the efficiency of data mining. Before the iterative operation, read the division rules in the nodes of the previous layer and write the queue information into the HDFS file. Call MapReduce to label the attributes and categories of the subset in the current training sample set and apply the Hash table to store them, and use the Hash table to calculate the best branches of different nodes. The key values calculated in the Map phase are output according to the partition conditions of the decision tree and saved in the HDFS file directory according to the statistics of the same key values. After storing the data, the training dimension values in the sample subset are summed and the different subset statistics are summarized. Calculations are performed on different nodes and input to the classification sample set, and the data need to be normalized in order to facilitate the SPARK operation. Set the nearest neighbor value to f and cache it in the operation node. The result of the training sample data set is obtained by parallelizing the operation, and set the data points to be classified as d . The formula for sorting by training is:

$$d = \max [dist(k, f)] \quad (4)$$

where: k is the number of neighboring points in the sample data points. After sorting, it enables different computing nodes to carry out iterative operations independently, thus improving the classification efficiency of the algorithm.

2.2. Enterprise financial big data real-time processing model

2.2.1. Constructing the Sample Characterization Indicator System. Randomly sampling of finance, communications, manufacturing and other Shanghai listed companies 16, with a sample data set of listed companies is X , then the expression $X = \{x_1, x_2, \dots, x_n\}, n = 1, 2, \dots, 16$. Selected in turn, operating profit margin C1, return on assets C2, cost and expense margins C3 and so on a total of 12 specific indicators, specific as shown in Table 1. Through the above characteristics of the data, the establishment of sample feature matrix F in which N represents the number of companies, $C_{ij}, i = 1, 2, \dots, 16, j = 1, 2, \dots, 12$, variable i represents 16 rows of samples, variable j a column of sample characteristics, the establishment of 16×12 sample feature matrix. Using k-means clustering algorithm to cluster analysis of the sample feature matrix F , the defined sample feature matrix is shown in equation (5):

$$F = \begin{bmatrix} N & C_{ij} & C_{ij} & \dots & C_{ij} \\ 1 & C_{1,1} & C_{1,2} & \dots & C_{1,12} \\ 2 & C_{2,1} & C_{2,2} & \dots & C_{2,12} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 16 & C_{16,1} & C_{16,2} & \dots & C_{16,12} \end{bmatrix} \quad (5)$$

Table 1. Financial performance evaluation system

Evaluation content	Index	Symbol
Profitability	Operating margin	X ₁
	Asset yield	X ₂
	Cost margin	X ₃
Solvency	Asset ratio	X ₄
	Equity ratio	X ₅
	Cash flow ratio	X ₆
Operational capacity	Total asset turnover	X ₇
	Turnover of current assets	X ₈
	Equity turnover	X ₉
Growth ability	Revenue growth	X ₁₀
	Total asset growth rate	X ₁₁
	Equity growth rate	X ₁₂

2.2.2. Factor analysis model. Factor analysis is a multivariate statistical method, and its core idea is that there are complex interrelationships between the financial indicators of enterprises, rather than completely independent [20]. Therefore, it can effectively and quickly merge and categorize a large number of overlapping variables into a few major factors, maintain data integrity and accuracy, avoid data loss and change, and help to effectively monitor the financial status of enterprises. In the application of this method, the idea of dimensionality reduction is widely used, which divides the variables under study into several categories according to certain criteria, and each category contains different variables, and ultimately, through the extraction of a few public factors with low correlation, it can reflect most of the information of the original data, thus making this method both objective and scientific.

The analyzing steps of the factor analysis method are as follows:

1) Positive and standardized processing of raw data. When evaluating the indicators, in addition to the positive indicators, all the reverse indicators as well as the moderate indicators must be processed through the data, and the equation for the normalization of the reverse indicators is shown in formula (6).

$$x'_i = 1 / x_i \quad (6)$$

The normalization formula for the fitness indicator is shown in equation (7).

$$x'_i = 1 / (1 + |a_i - x_i|) \quad (7)$$

where a_i is the moderate value of indicator x_i .

By using SPSS to standardize the data of positive indicators, we can eliminate the influence of different data outlines and thus perform better multivariate statistical analysis. For example, we can standardize the data of earnings per share (yuan) and net sales margin (%) to better understand the relationship between them and ensure the smooth progress of factor analysis. In this paper, we use the Z-Score standardization processing technique in SPSS27.0 software in order to achieve effective standardization of the data. ZScore standardization processing is shown in equation (8).

$$z_{score}x = (x - \bar{x}) / \alpha \quad (8)$$

where \bar{x} is the mean of x and α is the variance of x .

2) Applicability test. Through the factor analysis method, we can screen out several independent evaluation indicators from the known raw data in order to better reflect the interrelationship between them.

3) Extracting public factors. The following three methods exist for extracting factors:

One is according to the size of the eigenvalue, if the eigenroot of the factor is greater than 1, then it can be selected as a common factor.

Second, by evaluating the cumulative contribution rate, we can determine which factors have the greatest impact on the data. When the cumulative variance contribution of these factors is more than 80% they can be taken as valuable public factors.

Third, through scatter analysis of the gravel plot, several factors before the inflection point appeared were found to have fairly high eigenvalues, which enabled us to better understand the raw data. Therefore, these factors are considered as public factors.

4) Naming the public factors. By combining the eigenvalues, cumulative contribution rate and fragmentation diagram, several key factors can be extracted. In order to better identify these key factors, they need to be named according to the economic connotation they represent.

5) Calculate the average score of each factor. When calculating the composite score for each public factor, the values of all the raw variables need to be linearly combined to more accurately reflect the relationship between them, and the data of these raw variables need to be standardized to better assess the composite financial performance. Therefore, the factor score coefficients and the standardized raw variable data are the important basis for calculating the composite score of each factor. As shown in formula (9).

$$F_i = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (9)$$

$\beta_i (i = 1, 2, \dots, n)$ is the score of factor F_i on variable x_n .

The composite score of the factors is obtained by multiplying the score of each principal factor with the contribution of the rotated principal factor separately. It can be expressed by the formula:

$$F = \alpha_1 F_1 + \alpha_2 F_2 + \cdots + \alpha_m F_m \quad (10)$$

$F_i (i = 1, 2, \dots, m)$ is the score of each factor and $\alpha_i (i = 1, 2, \dots, m)$ is the contribution of each factor.

2.2.3. K-means clustering algorithm. The sample feature matrix F is the data set of k-means clustering algorithm, and the number of clusters will be set to 6. In order to classify and explore the overall financial status and characteristics of similar listed companies, optimize the allocation of corporate resources, and make reasonable decisions and deployments for the future strategic development planning, six clustering centers $A = \{a_1, a_2, \dots, a_6\}$ are randomly selected, in the analysis of the company's financial status data, then the clustering algorithm's objective function is represented by L_n , and the mathematical expression is:

$$L_n = \sum_{i=1}^n \min_{1 \leq j \leq 19} |X_i - a_j|^2 \quad (11)$$

In Eq. (11), a_k is the cluster centroid of cluster A_k and the mathematical expression is shown below:

$$a_k = \frac{1}{|A_k|} \sum_{x_i \in A_k} x_i \quad (12)$$

First establish the sample feature matrix model, then abnormal data cleaning for missing data and error data, after cleaning to get the correct feature data, in the k-means clustering algorithm to arbitrarily set the cluster center point of the initial data clustering, and then according to the Euclidean distance formula to calculate the distance between the cluster center point and the sample data points, will be nearer to the sample data divided into clusters located in the center point of the clustering, then Judge whether to meet the iteration termination conditions, such as meet the output clustering labels, such as do not meet the update clustering center point recalculation; until the conditions meet the end of the k-means clustering algorithm.

2.3. Early warning model of RF financial risk based on Benford's law

2.3.1. Benford's Law. In 1938, American physicist Frank Benford observed that the frequency of the distribution of the first digit in natural data follows a stable pattern.

$$P(d_1) = \log_{10} \left(1 + \left(\frac{1}{d_1} \right) \right) \quad (13)$$

$$P(d_2) = \sum_{d_1=1}^9 \log_{10} \left(1 + \left(\frac{1}{d_1 \cdot d_2} \right) \right) \quad (14)$$

$$P(d_3) = \sum_{d_1=1}^9 \sum_{d_2=1}^9 \log_{10} \left(1 + \left(\frac{1}{d_1 \cdot d_2 \cdot d_3} \right) \right) \quad (15)$$

$$P(d_4) = \sum_{d_1=1}^9 \sum_{d_2=1}^9 \sum_{d_3=1}^9 \log_{10} \left(1 + \left(\frac{1}{d_1 \cdot d_2 \cdot d_3 \cdot d_4} \right) \right) \quad (16)$$

where d refers to the first non-zero digit counted from left to right in the natural data. For negative numbers, the observation is the first digit of the absolute value of the value.

Benford's law suggests that if the dataset is naturally occurring and the data is of high quality, then the frequency of the first digit in the dataset must be consistent with the above probability distribution. If the amount of data is larger, then the frequency distribution of the first digit in the general data set is more consistent with the theoretical frequency of Beford's law. Therefore, we can conclude that a deviation from Benford's law may imply an illegal forgery. Using this principle, Benford's law can be used to identify fraud in corporate financial data.

2.3.2. Benford's law test. This paper provides three methods for verifying that the distribution of the first digit of the sample data matches the theoretical observed frequency of Benford's law. In this subsection, e_i represents the actual observed frequency of the first digit i in the sample data, while p_i represents the theoretical frequency value of the first digit i in Benford's law.

1) χ^2 The expression for the goodness-of-fit test χ^2 statistic is:

$$\chi^2 = N \cdot \sum_{i=1}^9 \left[\frac{(e_i - p_i)^2}{p_i} \right] \quad (17)$$

When the χ^2 statistic exceeds these thresholds, this may suggest the potential for fraud in the sample data.

2) Modified K-S goodness-of-fit test. The execution of the K-S test involves the following steps: first, the cumulative distribution function of the first digit frequencies of the sample data is calculated and

then compared to the theoretical distribution function of Benford's Law. This comparison process involves calculating the differences between the two distribution functions, subsequently taking the absolute values of these differences, and ultimately determining the maximum of these, which is often referred to as the statistical D-value. By comparing the magnitude of the statistical D-value with a pre-determined critical value, it is possible to assess how well the distribution of the first digits of the sample data fits a particular theoretical distribution function, in this case Benford's law. If the statistical D-value is greater than the set critical value, this indicates that the distribution of the first digits of the sample data is significantly different from the theoretical distribution of Benford's law. The test is:

$$V_n = \max [F_e(x) - F_p(x)] + \max [F_p(x) - F_e(x)] \tag{18}$$

where $F_e(x)$ is the cumulative distribution function of the frequency of the first digit of the sample data, and $F_p(x)$ represents the cumulative distribution function of a particular theoretical distribution. Giles (2007) corrected the statistic in the K-S goodness-of-fit test, and the expression of the corrected statistic is as follows:

$$V_n^* = V_n \left[N^{\frac{1}{2}} + 0.155 + 0.24 \times N^{\frac{1}{2}} \right] \tag{19}$$

3) PERSON CORRELATION COEFFICIENT TEST The correlation coefficient r is calculated by the formula:

$$r = \frac{\sum (p_i - \bar{p}_\tau)(e_i - \bar{e}_\tau)}{\sqrt{\sum (p_i - \bar{p}_\tau)^2 (e_i - \bar{e}_\tau)^2}} \tag{20}$$

where r takes values between $[0, 1]$ and correlation coefficient r values close to 1 indicate that the frequency observations of the sample data are more consistent with Benford's theoretical values.

2.3.3. Constructing the Benford Factor. First, by analyzing the distribution of the first digit in the dataset, the frequency corresponding to the first digit 1-9, i.e., the observed probability of the first digit in the group, was obtained. The difference between the observed frequency of the first digit and the theoretical frequency of Benford's law was used in this study to construct the Benford factor. The specific procedure is described below:

Let $X_j \{j=1,2,3,\dots,k\}$ denote the financial indicator variable. Denote by $r_d^{(j)}$ the difference between the observed frequency of the first digit of indicator X_j , d , and the theoretical frequency of Benford's law, and the expression for $r_d^{(j)}$ is shown in (21):

$$r_d^{(j)} = f_d^{(j)} - f_{B,d}^{(j)}, j = 1, 2, 3, \dots, k \tag{21}$$

where $f_d^{(j)}$ denotes the observed frequency of the first digit d of indicator X_j and $f_{B,d}^{(j)}$ denotes the theoretical frequency of Benford's law while satisfying the following constraint $\sum_{d=1}^9 f_d^{(j)} = 1, \sum_{d=1}^9 f_{B,d}^{(j)} = 1$. Then there:

$$\sum_{d=1}^9 r_d^{(j)} = 0, j = 1, 2, 3, \dots, k \tag{22}$$

The greater the absolute value of the difference $r_d^{(j)}$, the greater the likelihood that financial fraud has occurred in the indicator, if the data have been tinkered with and manipulated. The number that maximizes the absolute value of the frequency difference of the numbers is $a^{(j)}$, then there is:

$$a^{(j)} = \text{ard max}_d |r_d^{(j)}|, j = 1, 2, 3, \dots, k \quad (23)$$

B_j is the Benford factor for indicator $X_j \{j = 1, 2, 3, \dots, k\}$. It has the following expression:

$$B_j = \begin{cases} 1 & X_{i,j} \text{ first digit header } a^{(j)} \\ 0 & \text{other} \end{cases} \quad (24)$$

If the first digit of indicator $X_{i,j}$ is $a^{(j)}$, then the value of B_j is taken as 1; otherwise, it takes the value of 0. This paper converts the problem of financial data quality into a measurable dichotomous variable, and finally realizes the comprehensive consideration of financial data quality.

2.3.4. Random forest early warning model based on Benford's law. This paper proposes a random forest early warning model based on Benford's law. The model takes into account the quality factor of the financial data, which improves the predictive ability of the operational risk early warning model.

It is assumed that after k training, k classification models i.e. decision trees can be obtained finally, which will be somewhat different as they are trained on different subsets of data, denoted as $\{h_1(x), h_2(x), \dots, h_k(x)\}$. These decision tree models are then utilized to form an integrated learning model, which can be finally expressed as in Eq:

$$H(x) = \arg \max \sum_{i=1}^k I(h_i(x) = Y) \quad (25)$$

where $H(x)$ denotes the integrated classification model, Y denotes the categorical variables, $h_i(x)$ denotes a single decision tree model, and $I(\cdot)$ is the schematic function.

This paper chooses Random Forest as the base model, but also integrates the advantages of Benford's law, combines it organically with the Random Forest model, converts the data quality problem into a classification problem, and ultimately achieves the dual effect of business risk identification and early warning. The specific steps of constructing RFB early warning model:

1) The data quality will be examined and Benford factors will be constructed. Let the dataset $D = \{(X_1, Y_1), (X_2, Y_2), \dots, (X_N, Y_N)\}$, where X_i denotes the independent variable, $X_i = (X_{i,1}, X_{i,2}, \dots, X_{i,k}) (i = 1, 2, 3, \dots, N)$, Y_i denote the categorical variables. N denotes the amount of data, and k denotes the number of variables. Then the difference between the observed and theoretical frequencies of the first digit is calculated and the Benford factor is labeled using equation (25) B_j . The constructed Benford factor is added to the random forest model as a new independent variable, denoted as $X_i^B = (X_{i,1}, X_{i,2}, \dots, X_{i,k}, B_{i,1}, B_{i,2}, \dots, B_{i,k}) (i = 1, 2, 3, \dots, N)$, and the dataset is denoted as $D^B = \{(X_1^B, Y_1), (X_2^B, Y_2), \dots, (X_N^B, Y_N)\}$.

2) Resample dataset D^B and construct a decision tree model. n sample data sets are drawn, which are recorded as $D^{B(s)} \{s = 1, 2, \dots, n\}$. The initial number of decision trees is set according to the number of samples, and then the optimization of model parameters is carried out. Finally, the n sample data sets are selected to build n decision trees, and the decision tree model sequence is $\{h_1(X^{B(1)}), h_2(X^{B(2)}), \dots, h_n(X^{B(n)})\}$.

3) Combine the n decision trees to obtain the random forest model, as shown in Equation (26). $h_s(X^{B(s)})$ represents the s th decision tree model, $H(X^B)$ represents the combination of the random forest model, $I(\cdot)$ for the indicative function.

$$H(X^B) = \arg \max \sum_{s=1}^n I(h_s(X^{B(s)}) = Y) \quad (26)$$

The RFB model construction process is shown in Figure 2. The RFB model combines the advantages of Benford's Law and Random Forest Model, which is highly practical, especially suitable for the field of business risk early warning, with the following significant advantages:

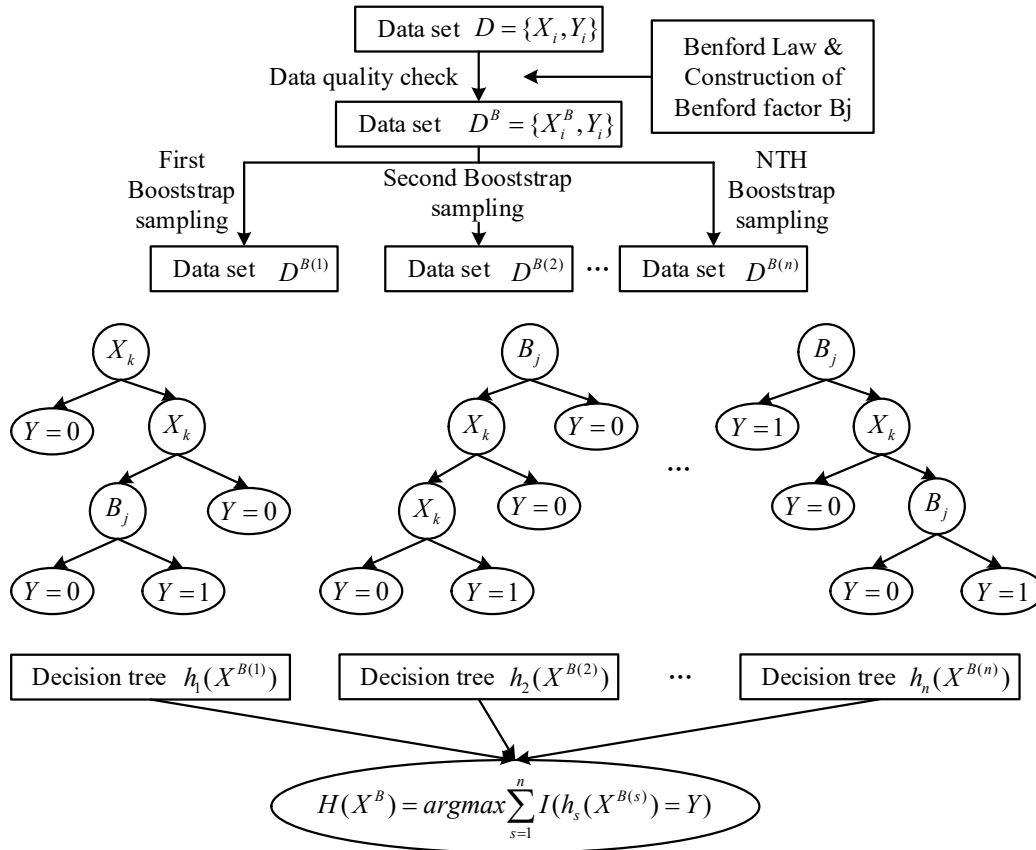


Fig. 2. RFB model construction procedure

3. Empirical analysis of real-time processing of big data in corporate finance

The time span of this study is the whole year of 2023, and the raw data are mainly from the 2023 annual financial statements of listed companies published on the website of InWealth. Considering that the fiscal year-end date of individual sample companies is not uniform with the sample subject, this paper also appropriately refers to the relevant data in the quarterly financial statements of these companies, in order to ensure as much as possible the consistency of the time period of the research study. For the use of measurement software, this paper utilizes SPSS software for data processing.

3.1. Factor analysis process and results

3.1.1. Model Testing and Common Factor Extraction. In this paper, KMO test and Bartlett's sphericity test were conducted on 12 indicators of the sample company to verify the applicability of the factor analysis method to the sample variables, and the KMO test was 0.528 (>0.5), and the value of the Bartlett's sphericity test was 258.172, with P=0.000 (<0.05), so it can be stated that the original variables are suitable for the factor analysis.

In this paper, principal component analysis is used to extract factor variables based on the correlation coefficient, and the factors with eigenvalues greater than 1 are extracted as the common factors. Table 2 shows the variance of the common factor, which indicates the degree of explanation of the extracted common factor on each original variable, and the larger the extracted value indicates that the variable is expressed better by the common factor. From the table, it can be seen that the

extracted values of each variable are high and above 0.5, which indicates that the extracted public factors have stronger explanatory power for the original variables.

Table 2. Shows the factorial variance

Index name	Initial	Extraction
Operating margin	1.000	0.978
Asset yield	1.000	0.949
Cost margin	1.000	0.958
Asset ratio	1.000	0.988
Equity ratio	1.000	0.988
Cash flow ratio	1.000	0.895
Total asset turnover	1.000	0.928
Turnover of current assets	1.000	0.947
Equity turnover	1.000	0.952
Revenue growth	1.000	0.638
Total asset growth rate	1.000	0.879
Equity growth rate	1.000	0.685

Table 3 shows the total variance of the interpretation, according to the standard of eigenvalue greater than 1, this paper extracted a total of 4 male factors, namely F_1 , F_2 , F_3 , F_4 , the percentage of variance indicates the contribution rate of the variance, that is, the amount of information represented by each male factor, and it is generally believed that the cumulative contribution rate of extracted male factors should be higher than 80%. From the table, it is easy to see that the cumulative contribution rate of the variance of the four common factors reaches 89.30%, which means that the four factors reflect 89.84% of the information of the original index, and it can be considered that the four common factors contain most of the information of the original variables, so it is feasible to use these four factors to evaluate the financial performance level of the sample company.

Table 3. The total variance explained in Table

Factor	Initial eigenvalue			Extract the sum of squares and load			Rotate the squares and load		
	Total	Percentage of variance	Cumulative percentage	Total	Percentage of variance	Cumulative percentage	Total	Percentage of variance	Cumulative percentage
1	3.75	31.25	31.25	3.75	31.25	31.25	3.068	25.59	25.59
2	3.278	27.32	58.57	3.278	27.32	58.57	2.891	24.13	49.72
3	1.974	16.45	75.02	1.974	16.45	75.02	2.628	21.92	71.64
4	1.714	14.28	89.30	1.714	14.28	89.30	2.179	18.20	89.84
5	0.646	5.38	94.68						
6	0.302	2.52	97.20						
7	0.109	0.91	98.11						
8	0.073	0.61	98.72						
9	0.042	0.35	99.07						
10	0.023	0.19	99.26						
11	0.089	0.74	100						
12	0.000	0.000	100						

3.1.2. Naming of common factors. In order to better understand the economic meaning of each factor, this paper uses the maximum variance method to rotate the factors orthogonally so that the loadings

on each factor are as close as possible to 0 or ± 1 . Table 4 shows the rotated factor loading matrix, reflecting the loadings of the original indicators on each of the public factors, and the rotation process makes each public factor have a clearer economic meaning. As can be seen from the table, the common factor F_1 in the operating profit margin, return on assets and cost and expense margin has a higher load, the above indicators mainly reflect the profitability of the enterprise, so it is called the common factor F_1 for the profitability factor; the common factor F_2 in the total asset turnover, current asset turnover and shareholders' equity turnover has a higher load, the above indicators mainly reflect the enterprise's operating ability, so it is called the common factor for the operation factor; the common factor 4 in the growth rate of operating income, total asset growth rate and the growth rate of shareholders' equity, so it is called the common factor F_3 in the operating factor. Factor 4 has higher loadings on operating income growth rate, total assets growth rate and shareholders' equity growth rate, the above indicators mainly reflect the growth ability of the enterprise, so it is called Factor F_3 as the growth factor; Factor F_4 has higher loadings on assets and liabilities ratio and equity ratio, the above indicators mainly reflect the ability of the enterprise to pay off its debts, so it is called Factor F_4 as the debt-servicing factor.

Table 4. Factor load matrix after rotation

Index name	F ₁	F ₂	F ₃	F ₄
Operating margin	0.969	-0.169	0.071	-0.079
Asset yield	0.849	0.285	0.331	0.215
Cost margin	0.914	-0.228	0.259	-0.105
Asset ratio	-0.021	0.078	0.022	0.993
Equity ratio	-0.019	0.064	0.023	0.995
Cash flow ratio	0.464	-0.058	0.806	-0.178
Total asset turnover	-0.167	0.921	0.233	0.049
Turnover of current assets	-0.087	0.949	0.095	0.177
Equity turnover	0.011	0.963	-0.165	-0.034
Revenue growth	0.184	0.107	0.714	0.296
Total asset growth rate	0.132	-0.003	0.928	-0.048
Equity growth rate	0.539	-0.211	-0.591	0.000

3.1.3. **Public Factor Score and Comprehensive Evaluation.** In this paper, regression analysis was used to linearly regress the factors on the original variables to produce least squares estimates of the coefficients, and the matrix of factor score coefficients was calculated as shown in Table 5.

Table 5. Factor score coefficient matrix

Index name	F ₁	F ₂	F ₃	F ₄
Operating margin	0.341	0.011	-0.083	-0.022
Asset yield	0.299	0.133	0.013	0.084
Cost margin	0.292	-0.029	0.013	-0.032
Asset ratio	0.005	-0.047	-0.014	0.468
Equity ratio	0.004	-0.052	-0.012	0.469
Cash flow ratio	0.066	-0.029	0.296	-0.094
Total asset turnover	-0.025	0.319	0.062	-0.051
Turnover of current assets	0.031	0.335	-0.016	0.015
Equity turnover	0.097	0.378	-0.136	-0.082
Revenue growth	-0.014	-0.016	0.273	0.123
Total asset growth rate	-0.072	-0.051	0.385	-0.041
Equity growth rate	0.262	0.000	-0.311	0.034

Based on the matrix of factor score coefficients, the following four common factor expressions can be obtained (X_1, X_2, \dots, X_{12} is the standardized data for each indicator):

$$F_1 = 0.341X_1 + 0.299X_2 + 0.292X_3 + 0.005X_4 + 0.004X_5 + 0.066X_6 - 0.025X_7 + 0.031X_8 + 0.097X_9 - 0.014X_{10} - 0.072X_{11} + 0.262X_{12} \quad (27)$$

$$F_2 = 0.011X_1 + 0.133X_2 - 0.029X_3 - 0.047X_4 - 0.052X_5 - 0.029X_6 + 0.319X_7 + 0.335X_8 + 0.378X_9 - 0.016X_{10} - 0.051X_{11} + 0.000X_{12} \quad (28)$$

$$F_3 = -0.083X_1 + 0.013X_2 + 0.013X_3 - 0.014X_4 - 0.012X_5 + 0.296X_6 + 0.062X_7 - 0.016X_8 - 0.136X_9 + 0.273X_{10} + 0.385X_{11} - 0.311X_{12} \quad (29)$$

$$F_4 = -0.022X_1 + 0.084X_2 - 0.032X_3 + 0.468X_4 + 0.469X_5 - 0.094X_6 - 0.051X_7 + 0.015X_8 - 0.082X_9 + 0.123X_{10} - 0.041X_{11} + 0.034X_{12} \quad (30)$$

And then based on the variance contribution of each public factor by the weighting of the cumulative contribution of the variance of the four sub, the formula for the total financial performance score of the sample public factor firms was weighted:

$$\begin{aligned}
 F &= 26.271\% / 89.84\% F_1 + 25.174\% / 89.84\% F_2 \\
 &\quad + 22.728\% / 89.84\% F_3 + 19.272\% / 89.84\% F_4 \\
 &= 0.281 F_1 + 0.265 F_2 + 0.241 F_3 + 0.201 F_4
 \end{aligned}
 \tag{31}$$

3.2. Cluster analysis process and results

In order to further explore the commonalities exhibited by the sample companies in terms of financial performance, this paper conducts a systematic cluster analysis of the sample companies on the basis of the four public factor scores, seeking to classify the sample companies into a smaller number of relatively homogeneous groups and to analyze their characteristics in depth on the basis of the clusters. Specifically, four public factor scores are selected as clustering indicators, the clustering method adopts the intergroup linkage method, and the distance test selects the squared Euclidean distance method, and the clustering spectrogram obtained is shown in Figure 3.

As can be seen from the figure, the clustering of the sample companies has a strong hierarchical nature, in order to ensure that the classification has obvious differences without being too fragmented, this paper believes that the use of 5-class clustering is more appropriate, that is, the sample companies are divided into 5 categories.

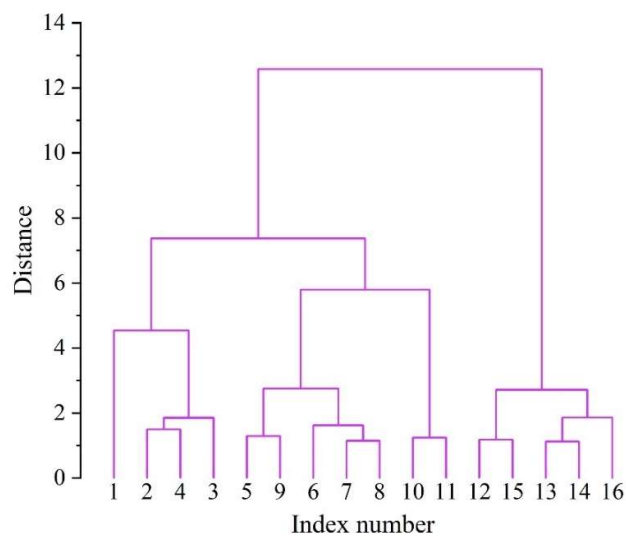


Fig. 3. Sample company cluster lineage

The results of the cluster analysis are shown in Table 6. As can be seen from the table, the profitability, growth and solvency of these companies are lower than the sample average, but their overall financial performance is slightly higher than the sample average, which is mainly due to the fact that the above companies have a very high operating capacity, which to a certain extent makes up for their shortcomings in other areas. Secondly, the profitability factor and the debt service factor are the company's strengths, whereas the operational capacity and the growth capacity are its weaknesses in comparison.

Table 6. Results of the cluster analysis

Categories	Company number	F ₁ equipartition	F ₂ equipartition	F ₃ equipartition	F ₄ equipartition	F equipartition

First class	1	-0.105	0.264	0.063	3.693	0.802
Second type	3	1.108	-0.358	1.392	-0.239	0.511
Third class	5	-0.284	0.909	-0.248	-0.338	0.031
Class four	2	2.001	-0.778	-2.221	0.001	-0.179
Fifth class	5	-0.801	-1.191	-0.084	-0.141	-0.599

4. Early warning analysis of enterprise financial risks

On the basis of the financial indicator system in Table 1, cash flow (cash flow from operating activities per share X13, cash current liabilities ratio X14, and cash flow growth rate per share X15) is introduced to constitute the financial risk early warning indicator system.

4.1. Sample selection and data sources

Table 7 shows the descriptive statistical analysis of various financial indicators of listed companies. As can be seen from the table, there is a general difference between the individual financial indicators of ST and normal operating companies. In particular, the mean value of total asset turnover X_6 is only 0.67 for ST companies and 15.11 for normal companies, which is a significant difference. Total asset turnover measures the long-term solvency of a company, which suggests that there are serious problems with the solvency of ST companies.

Table 7. Descriptive statistical analysis of the financial indicators of the listed companies

Financial index	Symbol	ST company		Normal company	
		Mean	Standard deviation	Mean	Standard deviation
Operating margin					
Asset yield	X_1	-0.09	0.59	-0.02	0.44
Cost margin	X_2	-0.04	0.15	0.06	0.12
Asset ratio	X_3	-0.33	0.62	-0.06	0.49
Equity ratio	X_4	0.99	0.77	1.61	1.04
Cash flow ratio	X_5	0.76	0.68	1.18	0.98
Total asset turnover	X_6	0.67	16.72	15.11	26.41
Turnover of current assets	X_7	-0.02	0.32	0.19	0.29
Equity turnover	X_8	-0.51	2.88	-0.11	3.19
Revenue growth	X_9	6.14	7.92	5.07	5.77
Total asset growth rate	X_{10}	13.47	18.65	11.29	16.77
Equity growth rate	X_{11}	4.71	0.45	0.83	0.52
Cash flow per share	X_{12}	0.49	0.43	0.81	0.51
Cash current liability ratio	X_{13}	0.13	0.53	0.25	0.71
Cash flow rate per share	X_{14}	0.09	0.32	0.17	0.32
Financial index	X_{15}	1.33	6.17	0.83	4.79

4.2. Effectiveness of Benford's Law-based Random Forest Model for Early Warning

4.2.1. Constructing the Benford Factor. First of all, Benford's law is used to test the data quality of financial indicators of A-share listed companies. The theoretical frequency of Benford's law, the observed frequency of the first digit of the financial indicators of A-share listed companies, as well as the difference between the observed frequency and the theoretical frequency and the results of the test are shown in Table 8. At the significance level of 0.01, the critical value of the χ^2 test is 20.09.

Based on the last column of Table 8, we can see that there are 6 indicators whose χ^2 test values are larger than the critical value, namely quick ratio X_5 , total asset turnover X_6 , accounts receivable

turnover X_{10} , total asset turnover X_{12} , cash flow from operating activities per share X_{13} and cash flow growth rate per share X_{15} . It means that they fail the χ^2 goodness-of-fit test, and it is considered that there is a significant difference between the observed frequency of the first digit of these indicators and the theoretical frequency of Benford's law. There is a significant difference and these indicators have higher financial risk. The remaining nine financial indicators pass the χ^2 goodness-of-fit test, but the actual observed frequencies are still different from the theoretical frequencies, and they also have a certain level of financial risk, which is lower than the six financial indicators that do not pass the goodness-of-fit test.

Benford factors are constructed using the above indicators. From the table it can be seen that the indicators of higher financial risk $X_5, X_6, X_{10}, X_{12}, X_{13}, X_{15}$ the first digit of the largest absolute value of the difference between the observed frequency and the theoretical frequency of the first digit of the largest number of numbers are 1, 2, 1, 3, 1, 2, the difference of 10.92, -8.07, 8.31, 8.89, -7.61, -6.51, respectively. the indicators of relatively low financial risk $X_1, X_2, X_3, X_4, X_7, X_8, X_9, X_{11}, X_{14}$ the first digit of the absolute value of the difference between the observed frequency and the theoretical frequency of the largest number of 1, 2, 9, 2, 3, 1, 1, 9, 3, the difference is -3.51, 6.97, 3.42, 2.55, 1.13, 7.41, 4.47, 2.53, -2.49. 15 Benford factors were established according to the formula and recorded as $B_j \{j = 1, 2, \dots, 15\}$, respectively.

Table 8. Frequency of observed first numbers and comparison with theoretical frequencies

Index		1	2	3	4	5	6	7	8	9	χ^2
Benford	(1)	31.22	18.41	13.11	10.02	8.31	7.27	6.19	6.08	4.97	
X_1	(2)	26.64	17.1	15.22	9.75	7.32	6.77	7.05	4.36	5.71	6.74
	(3)	-3.45	-0.51	2.74	0.08	-0.58	0.1	1.26	-0.78	1.13	
X_2	(2)	29.87	24.47	11.95	7.61	5.44	5.73	4.06	4.9	5.99	19.22
	(3)	-0.2	6.87	-0.55	-2.08	-2.47	-0.99	-1.71	-0.17	1.41	
X_3	(2)	30.16	17.12	10.87	6.79	8.71	8.18	3.8	6.51	7.89	19.08
	(3)	0.09	-0.46	-1.6	-2.89	0.77	1.45	-2	1.41	3.29	
X_4	(2)	31.5	20.1	11.68	10.63	8.45	5.14	5.14	3.77	3.5	5.91
	(3)	1.39	2.51	-0.81	0.92	0.5	-1.56	-0.62	-1.32	-1.02	
X_5	(2)	41.01	12	7.85	5.71	8.14	7.07	5.14	5.17	7.87	43.27
	(3)	10.92	-5.66	-4.63	-3.97	0.26	0.39	-0.63	0.04	6.02	
X_6	(2)	29.63	9.51	9	11.71	11.98	8.15	5.45	8.68	6	38.68
	(3)	-0.5	-8.07	-3.5	1.96	4.02	1.46	-0.34	3.59	1.4	
X_7	(2)	29.07	16.86	13.61	10.05	8.42	5.71	5.45	5.96	4.86	2.11
	(3)	-1.06	-0.75	1.07	0.36	0.52	-0.98	-0.36	0.86	0.29	
X_8	(2)	37.5	14.65	10.87	8.18	10.61	5.46	4.86	4.04	3.79	16.24
	(3)	7.39	-2.94	-1.62	-1.55	2.68	-1.24	-0.9	-1.04	-0.77	
X_9	(2)	34.49	15.51	11.13	6.23	7.06	6.8	7.3	4.66	6.77	15.47
	(3)	4.38	-2.12	-1.35	-3.44	-0.84	0.13	1.53	0.47	2.21	
X_{10}	(2)	38.04	13.85	9.74	10.32	5.13	5.74	4.63	5.4	7.07	23.09
	(3)	8.31	2.25	0.81	-2.06	-1.38	0.91	-1.46	-1.83	2.49	
X_{11}	(2)	25.28	14.38	21.49	10.87	8.17	7.05	4.88	4.05	3.79	13.18
	(3)	-4.82	-3.21	8.99	1.21	0.22	0.4	-0.89	-1.01	-0.75	
X_{12}	(2)	23.11	16.3	14.42	11.65	7.34	6.48	10.29	5.69	4.63	31.45
	(3)	-6.97	-3.27	8.89	1.99	-0.57	-0.16	4.54	0.58	0.07	
X_{13}	(2)	30.72	18.22	10.03	10.09	7.88	7.89	4.58	5.43	5.15	23.65
	(3)	-7.61	-1.59	1.45	0.36	-0.03	1.2	-1.19	0.31	0.59	
X_{14}	(2)	28.81	11.38	11.15	8.42	10.34	8.42	6.8	8.71	6.21	4.09
	(3)	-1.31	-6.19	-1.32	-1.27	1.42	2.43	1.77	0.99	3.6	
X_{15}	(2)	26.64	17.1	15.22	9.75	7.32	6.77	7.05	4.36	5.71	25.37
	(3)	-3.45	-6.51	2.74	0.08	-0.58	0.1	1.26	-0.78	1.13	

Note: (1) in the second column indicates the theoretical frequency of the first digit under Benford's law, (1) indicates the observed frequency of the first digit of each financial indicator, and (3) indicates the difference between the observed frequency and the theoretical frequency of the first digit under Benford's law.

4.2.2. Constructing a random forest model based on Benford's law. The random forest model with the addition of Benford factor variables includes 15 financial indicator variables and 15 Benford factor variables, totaling 30 variables. A random forest model based on Benford's law is established, here, $X_i^{B(x)} = (X_{i,1}^{B(s)}, X_{i,2}^{B(s)}, \dots, X_{i,15}^{B(s)}, B_{i,1}, B_{i,2}, \dots, B_{i,15}) (i = 1, 2, \dots, 368)$.

According to the idea of cross-validation, 80% of the dataset is randomly selected as the training set and the remaining 20% as the test set. Combined with the sample size of the dataset, the number of initial decision trees is set to 100. The initial model based on Benford law random forest is established through the training set, and the prediction accuracy of the test set is used to judge the model's merits. The prediction accuracy rate is the probability that all classifications are correct, i.e., the proportion of actual normal companies being judged as normal companies and actual ST companies being judged as ST companies. The initial prediction accuracy of the Benford law-based random forest model built

using the financial data of listed companies is 88.91%. The model parameters were tuned using the learning curve.

The results of the learning curve are shown in Figure 4, the horizontal coordinate of the learning curve indicates the number of decision trees, the vertical coordinate indicates the model prediction accuracy, and the parameter with the highest prediction accuracy is selected as the optimal parameter. From the figure, it can be seen that when the model parameter values are around 35~55, the model has a higher prediction accuracy in the test set. Further refinement analysis of the learning curve reveals that the model has the highest prediction accuracy of 89.78% in the test set when the model parameter value is 40. The prediction accuracy improved by 1.35 percentage points after parameter tuning.

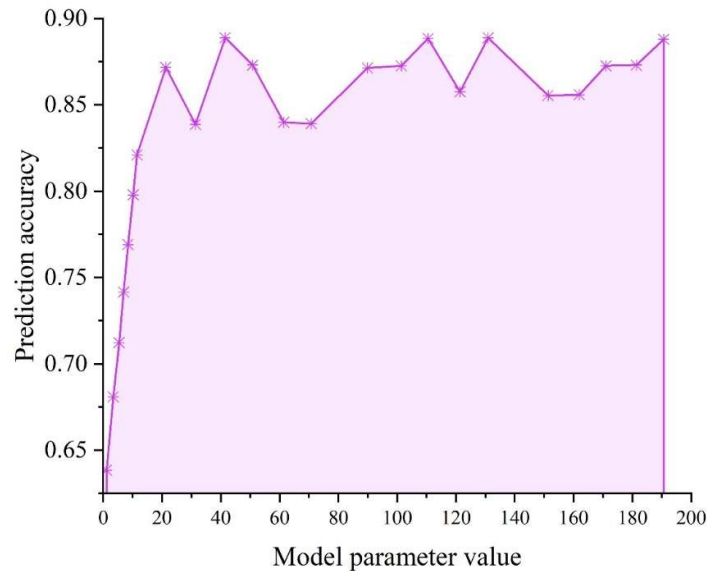


Fig. 4. Parameter tuning of the random forest model based on the Benford-law

4.2.3. Comparative analysis of model prediction effects. In order to compare the model effect, the random forest model without adding Benford factor variables is constructed, i.e., only the 15 financial indicators of the financial risk early warning indicator system are included. To build a random forest model without adding Benford factor variables, here, $X_i^{(o)} = (X_{i,1}^{(o)}, X_{i,2}^{(o)}, \dots, X_{i,15}^{(o)}) (i = 1, 2, \dots, 368)$. Setting the initial decision tree tree as 100, the initial prediction accuracy of the random forest model is 75.68%, using the learning curve for parameter tuning, the results are shown in Figure 5.

The figure shows that the model has the highest prediction accuracy when the model parameter values are between 15 and 30. Further refinement of the learning curve analysis found that the highest prediction accuracy of 79.45% was achieved in the test set when the model parameter value was 19. The prediction accuracy of the random forest model is improved by 2.88% before and after parameter tuning, which is greater than that of the random forest model based on Benford's law, but the prediction accuracy of the random forest model based on Benford's law is much higher than that of the random forest model at 89.78%, and the difference between the two is 9.81%, which fully demonstrates that the introduction of Benford's factor variable is able to effectively identify the sample point data of listed companies with financial risk, which is of great significance to improve the financial risk early warning ability of the model.

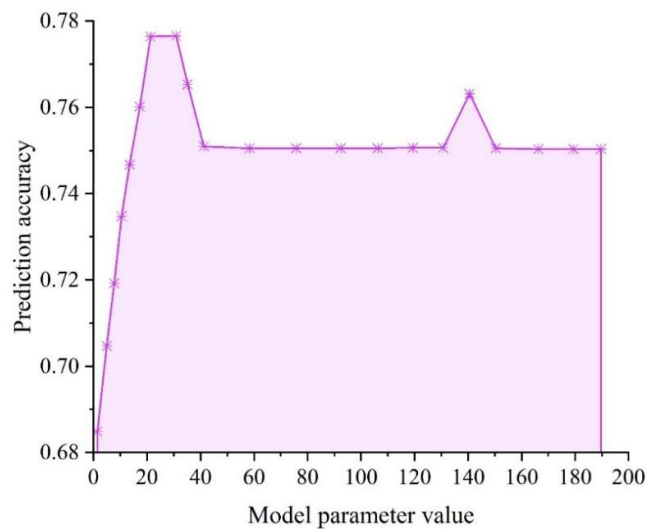


Fig. 5. Parameter tuning of the random forest model of the financial data of listed companies

4.3. Analysis of the application of early warning models

After completing the standardization process, the financial and non-financial index data of listed companies in 2021 are brought into the random forest model with the best model prediction effect selected in the previous section, i.e., the data of year T-3 is used to predict the financial risk status of the enterprise in the base period (i.e., year 2023), and the experimental results are visualized using Python and seaborn visualization libraries, as shown in Fig. 6.

In this paper, the random forest model based on Benford's law is used as the benchmark model for early warning monitoring of financial risk of listed companies, and P is used to denote the probability of the financial risk situation of listed companies, and the P value ranges from 0 to 1. G listed companies are judged to be abnormal with a P value greater than 0.5, and are judged to be normal with a P value less than 0.5. In this paper, for the data of listed companies in the validation set, 200 repeated experiments with different random numbers are carried out, the horizontal coordinate in the graph indicates the number of experiments, and the vertical coordinate is the probability value of predicting the result of 1 in the output of each experimental model, and its average probability value is 0.724. Listed company G is ST in 2021, and the prediction result is consistent with the actual status of listed companies in 2023, i.e., in the 200 repeated experiments 196 correctly predicted and 4 incorrectly predicted, obtaining 98% prediction accuracy, which proves that the model is able to effectively discriminate between ST and non-ST companies.

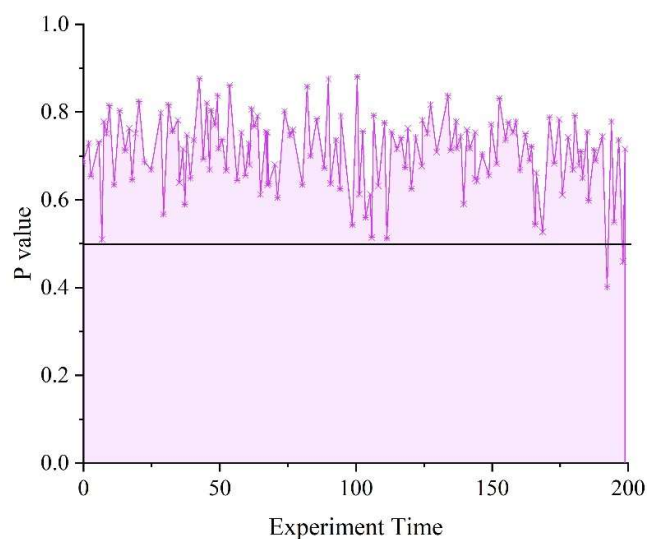


Fig. 6. Financial risk early warning results of the company

5. Conclusions and recommendations

5.1. Conclusion

This paper constructs a financial performance evaluation system with profitability, solvency, operating ability and growth ability as the core, evaluates and compares the financial performance of listed companies in 2021 based on factor analysis and cluster analysis methods, evaluates the quality of the data by introducing Benford factors, and proposes a random forest model based on the Benford law in conjunction with the random forest model to give a specific application study in financial specific application research in early warning of risk. The research results are as follows:

1) From the process of factor analysis, the profitability factor, operation factor, growth factor and debt service factor together determine 89.30% of the content of the financial performance of the listed company, and the four male factor variance contribution rate is 26.271%, 25.174%, 22.728% and 19.272% respectively, which indicates that the relative importance of each male factor in the process of financial performance evaluation of the listed company of sports goods From high to low are: profitability factor, operation factor, growth factor and debt service factor, it can be inferred that in terms of the impact on the financial performance of sporting goods listed companies, the company's profitability has the greatest degree of influence, followed by the operating ability, the growth ability again, and the last is the ability to pay off debt.

2) Selected listed companies' financial data to establish a random forest model based on Benford's law, and calculated the model prediction accuracy and precision rate. The results show that the Benford factor can fully utilize the information about data quality and effectively identify the specific sample points with high financial risk. Secondly, the prediction accuracy of the random forest model with Benford's factor is significantly higher than that of the random forest model.

5.2. Recommendations

The financial situation of the enterprise will affect investors, banks and other institutions on the company's investment attitude, which in turn affects the financing plan and business development, the bankruptcy and closure of many enterprises often stems from the blind confidence in their own financial situation, without proper control of financial risk, which ultimately led to the demise of the enterprise. According to the former manufacturing listed companies financial risk characteristics and analysis of the financial situation of listed companies, for the summarized part of the business operations, put forward the following suggestions.

1) The establishment of financial early warning system, sound risk prevention and control. Listed companies in the manufacturing industry in the state of continuous operation, both to focus on immediate economic benefits, but also to establish a sound financial early warning system, the two complement each other is the plan for long-term development. As the management should start from the source of risk, a comprehensive grasp of financial risk, to facilitate the discovery and avoidance of risk; as a financial management personnel, should be armed with professional knowledge, strengthen the supervision consciousness, find out the financial risk in advance and actively respond to reduce the impact of financial crisis on the company, reduce losses and losses.

2) Optimize the internal control system and improve the governance structure. Strengthening the construction of enterprise internal control system is one of the means to control financial risks. Focusing on the key aspects of the industry's internal control, and linking with the actual situation of the Company, under the standardized internal control system enterprises should strengthen the implementation and supervision of their own internal control, establish an internal control department independent of other departments, and the independent internal control department

regularly issues internal control evaluation reports to ensure the transparency and efficiency of the internal control system.

3) Enhance R&D and innovation capabilities to improve core competitiveness. At a time when information technology is developing at a high speed, technology is the basic element of enterprise competitiveness, and the level of the company's scientific research and technology measures its core competitiveness. In order to improve its core competitiveness, the company is supposed to increase its investment in scientific research manpower and capital, strengthen technological innovation and upgrading, and continuously innovate in order to maintain its competitive advantages. From the perspective of the market, targeted research and development of best-selling products, to enhance the technology of similar products relative to the industry, and continuously increase market share. In addition, we are concerned about the risk of inventory decline, avoiding inventory backlog, improving inventory management and efficiently converting inventory into operating profit.

References

- [1] Liu, Z. (2023). Research on the risk management of enterprise cloud accounting application under the background of big data. *Advances in Economics and Management Research*, 6(1), 122-122.
- [2] Cong, X. (2021, June). Research on financial risk management of E-commerce enterprises in the era of big data. In *Proceedings of the 7th International Conference on Frontiers of Educational Technologies* (pp. 195-199).
- [3] Song, Y., & Wu, R. (2022). The impact of financial enterprises' excessive financialization risk assessment for risk control based on data mining and machine learning. *Computational Economics*, 60(4), 1245-1267.
- [4] Tkachuk, A., & Dankevich, A. (2018). Theoretical essence of risks, hazards and threats in the context of enterprise economic safety provision. *University Economic Bulletin*, (38), 118-125.
- [5] Campbell, R. (2020). The need for cyber resilient enterprise distributed ledger Risk Management Framework. *The Journal of The British Blockchain Association*.
- [6] Shi, W. (2021). Analyzing enterprise asset structure and profitability using cloud computing and strategic management accounting. *PloS one*, 16(9), e0257826.
- [7] Ouyang, Z. (2021). Risk Control of Virtual Enterprise Based on Distributed Decision-Making Model. *Complexity*, 2021(1), 5535753.
- [8] Feng, Z. (2022). Simulation Analysis of Artificial Intelligence in Enterprise Financial Management Based on Parallel Computing. *Mobile Information Systems*, 2022(1), 2958176.
- [9] Wang, X., Luo, X., & Hu, Y. (2021, September). Enterprise accounting and financial risk analysis system based on decision tree and SVM. In *2021 4th International Conference on Information Systems and Computer Aided Education* (pp. 2015-2018).
- [10] Yao, L. (2019). Financial accounting intelligence management of internet of things enterprises based on data mining algorithm. *Journal of Intelligent & Fuzzy Systems*, 37(5), 5915-5923.
- [11] Ren, D., & Wu, H. (2022). Design and Implementation of Enterprise Financial Risk Control Information Management System Based on Big Data of Internet of Things. *Mobile Information Systems*, 2022(1), 5677870.
- [12] Shang, H., Lu, D., & Zhou, Q. (2021). Early warning of enterprise finance risk of big data mining in internet of things based on fuzzy association rules. *Neural Computing and Applications*, 33(9), 3901-3909.

-
- [13] Wang, Y. (2023, December). Construction of enterprise financing risk management system under big data technology. In 3rd international conference on digital economy and computer application (DECA 2023) (pp. 29-35). Atlantis Press.
- [14] Ren, S. (2022). Optimization of Enterprise Financial Management and Decision-Making Systems Based on Big Data. *Journal of Mathematics*, 2022(1), 1708506.
- [15] Han, W. (2023). Enterprise financial risk model based on cloud computing in age of big data. *Soft Computing*, 1-11.
- [16] Chen, X., & Metawa, N. (2020). Enterprise financial management information system based on cloud computing in big data environment. *Journal of Intelligent & Fuzzy Systems*, 39(4), 5223-5232.
- [17] Chen, Y. (2022). Enterprise Financial Data Sharing Based on Information Fusion Cloud Computing Environment. *Wireless Communications and Mobile Computing*, 2022(1), 5994628.
- [18] Zhou, X. (2022, June). Enterprise Financial Management Informatization under Cloud Computing Environment. In 2021 International conference on Smart Technologies and Systems for Internet of Things (STS-IOT 2021) (pp. 101-106). Atlantis Press.
- [19] Liliana Miranda Aragón & Alejandro Ivan Aguirre Salado. (2025). An analysis of the spatial distribution of NO₂ extremes in the metropolitan area of the valley of Mexico using a decision tree. *Stochastic Environmental Research and Risk Assessment*(prepublish),1-13.
- [20] Qian Zheng, Annan Zhou & Shui Long Shen. (2025). Mapping bushfire risk based on scale division and factor analysis: A case study from Victoria, Australia. *International Journal of Disaster Risk Reduction*105222-105222.