J. COMBIN. MATH. COMBIN. COMPUT. 127b (2025) 9537--9555

Journal of Combinatorial Mathematics and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Construction of Corporate Financial Risk Prediction Model Based on Random Forest Algorithm

Ke Sun[1], Yupeng Li[2,✉]

[1] *School of Accountancy, Guangzhou College of Technology and Business, Guangzhou, Guangdong, 510850, China*

[2] *School of Accountancy, Anyang Institute of Technology, Anyang, Henan, 455000, China*

ABSTRACT

Financial risk has a greater impact on the operation and development of enterprises, and accurate prediction of financial risk has become an industry demand, so as to better help enterprises avoid possible financial risk. The article establishes an enterprise financial risk prediction model based on the random forest algorithm, and fills in the oversampling of financial data through the SMOTENC algorithm, and realizes the downsizing of financial data by combining with the KPCA algorithm. Based on the enterprise financial risk characterization index system, the financial data of 358 listed enterprises were selected to carry out model validation and application analysis. The accuracy of corporate financial risk prediction based on Random Forest can reach up to 94.17%, and the average value of the overall time efficiency of the model is 0.68%, which is faster than the comparison algorithm in terms of financial data processing capability. Based on the results of financial risk prediction, the changes in corporate profitability, operating ability, solvency and development ability can be analyzed in depth, providing data support for enterprises to formulate preventive measures for corporate financial risk.

Keywords: random forest, SMOTENC algorithm, KPCA algorithm, financial risk prediction

1. Introduction

Enterprise financial risk refers to the enterprise in the whole process of financial activities, due to a variety of uncertainty caused by the enterprise loss opportunities and possibilities [1]. Enterprise financial risk throughout the whole process of production and operation, can be divided into: financing risk, investment risk, capital recovery risk and income distribution risk in four aspects [2-4]. Its main features include objectivity, comprehensiveness, uncertainty and coexistence of gain and loss [5-7]. Objectivity means that the risk exists everywhere and all the time. Financial risk is not transferable by human will, people can not avoid it, can not eliminate it, only through a variety of technical means to

✉ Corresponding author.
E-mail address: kelly_sun2024@163.com (Y. Li).
Received 05 March 2024; Revised 25 May 2024; Accepted 05 December 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-522

cope with the risk, and then avoid the risk. Comprehensiveness refers to the financial risk exists in all aspects of enterprise financial management work, in the fund-raising, capital utilization, capital accumulation, distribution and other financial activities will produce financial risk. Uncertainty refers to the financial risk in certain conditions, a certain period of time may occur, or may not occur. The coexistence of gain and loss means that the risk is directly proportional to the gain, the greater the risk, the higher the gain, and vice versa, the lower the gain.

However, the macro-environment of enterprise financial management is complex and changeable, and the financial management system cannot adapt to the complex and changeable macro-environment. The complexity of the macro environment of financial management is the external causes of financial risk. The external environment of financial management involves a wide range, of which the most important are the legal environment, financial environment and economic environment [8-11]. Revision of national laws, interest rate fluctuations, inflation, the government's economic policy adjustments and competition between enterprises may bring favorable aspects of the enterprise, but also may make the enterprise face certain threats, an enterprise if you can not adapt to these complex and changing external environment, will inevitably bring difficulties to the enterprise. Lack of scientific financial management decisions lead to decision-making errors and financial risk. Financial decision-making errors is the most direct cause of financial risk [12]. At present, the financial decision-making of Chinese enterprises generally exists in the phenomenon of empirical decision-making and subjective decision-making, which leads to decision-making errors often occur, resulting in the generation of financial risk. For this reason, it is necessary to predict the financial risk of enterprises in order to reduce the risk of loss and bankruptcy risk, etc.

Most of the current research on enterprise risk prediction is enterprise credit risk prediction [13] and risk assessment [14], and the prediction of enterprise financial risk is not in-depth. Among them, enterprise financial risk evaluation is an important basis for identifying and preventing potential enterprise risks and risk decision management. Literature [15] optimized financial risk evaluation using a fuzzy multi-criteria decision making (MCDM) algorithm under a comprehensive compromise scheme (CoCoSo). Similarly, literature [16] utilized MCDM for the evaluation of unbalanced classifiers in credit and bankruptcy risk prediction using multiple financial metrics, as classifiers as well as started to be widely used in financial risk prediction, the study optimized the classifiers for this purpose. Logistic regression, artificial neural networks, and random forest algorithms were also used to evaluate classifiers in financial risk prediction, with the best prediction accuracy of the random forest algorithm under additional factor conditions [17]. Among them, logistic regression and support vector machine models have also been applied to the construction of financial early warning models, but the accuracy effect is lower than the back propagation neural network model [18]. However, the one-to-one decomposition fusion method combined with support vector machine can predict multiple types of financial distress forecasts and outperforms the combination of support vector machine with two decomposition fusion methods, one-to-one residual and error-correcting output coding [19].

There are also methods for corporate financial risk prediction and early warning based on deep learning [20], a combination of multilayer perceptron artificial neural networks and traditional Altman Z-Score models [21], reinforcement learning [22], LightGBM machine learning algorithm [23], and regression analysis [24]. Deep learning has also been used for investment risk in retail financial risk [25]. In addition, there are studies on simulation modeling of corporate bankruptcy financial risk and considering the dynamics of financial performance of economic entities to achieve accurate bankruptcy financial risk prediction [26]. However, regardless of early warning or prediction, enterprises should pay attention to the risk of bad debt losses to ensure the normal operation of enterprises.

Introducing the random forest algorithm into enterprise financial risk prediction can improve the accuracy and generalization ability of financial risk prediction and make the results of enterprise financial risk prediction more accurate and reliable. This paper proposes a corporate financial risk prediction model based on the random forest algorithm, and selects the financial data of 358 listed enterprises as the research object. SMOTENC algorithm is used to oversample the financial data to optimize the data structure, and combined with KPCA algorithm to downsize the high-dimensional financial data. The normality test and non-parametric test are carried out for the validity of the data of corporate financial risk indicators, and the validity of this paper's model is verified through the comparison of the model's prediction effect, time efficiency and security. We also analyzed the financial risk factors of Company W as an example to provide data support for optimizing financial risk management measures.

2. Enterprise financial risk prediction model

As the economic environment becomes more and more market-oriented and internationalized, the financial risks faced by enterprises are also increasing day by day, and the number of enterprises in operational difficulties or even bankruptcy and liquidation due to financial risks is also increasing. Enterprises, especially listed companies, have become increasingly concerned about financial risks, and accordingly began to use some economic and financial indicators to predict and evaluate financial crises and issue early warning signals. Although the general prediction algorithm has the ability to parallel computing, multi-layer feedback, adaptive learning, but its operation has the disadvantages of the black box, the need for a large number of training samples, slow training speed, etc., relative to the ordinary means of multivariate discriminant analysis of the accuracy of its discrimination has not been significantly improved. Therefore, it is necessary to find new early warning indicators and prediction methods to discover the real operating conditions of enterprises, in order to prevent the occurrence of enterprise financial crisis.

2.1. Random Forest Algorithm

Random Forest (RF) is an integrated learning method that combines Bagging algorithm with decision trees [27]. Random Forest is to generate an algorithmic forest by randomization, each tree of this forest is a decision tree and uncorrelated with each other. After the forest is constructed, when a new sample enters this forest, it will be categorized and discriminated by each tree, and then statistically which category is selected the most in the end is decided to be that category. The whole algorithmic process can be specifically divided into three parts:

1) Construct k subsample set from the original sample set using the Bagging algorithm, with each sample set having a fixed m indicators randomly selected from the full set of indicator variables.

2) After obtaining the subsample set, construct a series of decision trees using complete splitting.

3) For a new sample, all decision trees in the forest are allowed to participate in the decision making and the category with the highest score is selected as the result.

Generalization error is an important evaluation criterion to measure the generalization ability of the model, usually the overfitting model, its generalization ability is poor, then the generalization error is generally higher, on the contrary, the smaller the generalization error, which indicates that the model's generalization ability is stronger. When the decision tree of Random Forest increases to a certain degree, its generalization error will converge to an upper bound, i.e., the Random Forest algorithm has a good ability to prevent overfitting.

Let there be a total of k decision tree in the forest $\{f_1(x), f_2(x), \ldots\ldots, f_k(x)\}$, x is the input of the samples to be classified, y is the actual class of the samples, $I(\Theta)$ is the schematic function, $av_k(\Theta)$ is the mean of the results, and the interval function is:

$$mr(x, y) = av_k I(f_k(x) = y) - \max_{j \to y} av_k I(f_k(x) = j) \tag{1}$$

The above interval function measures the predictive ability of the random forest model by calculating the difference between the number of decision trees classified correctly and those classified incorrectly, with larger values indicating better predictive ability.

Let PE^* be the generalization error of the random forest model, then:

$$PE^* = P_{x,Y}(mr(x,y) < 0) \tag{2}$$

$$\lim_{L\to\infty} PE^* = P_{X,Y}(P_\theta(f(X,\theta)=Y) - P_\theta(f(X,\theta)=j) \prec 0) \tag{3}$$

L is the number of decision trees, and the above equation shows that as the number of decision trees increases, the generalization error, i.e., obeys the law of large numbers and converges to an upper bound when the number of decision trees is sufficiently large.

Define the strength of the random forest $\{f(x,\theta)\}$ as the expectation of its interval function, i.e:

$$s = E_{x,y}mr(x,y) \tag{4}$$

Let the intensity $s>0$, according to Chebyshev's inequality, viz:

$$PE^* \le \mathrm{var}(mr)/s^2 \tag{5}$$

Setting $\hat{j}(x,y) = \arg\max_{j\cdot y} p_s(f(x,\theta)=j)$, the interval function is:

$$\begin{aligned} mr(x,y) &= P_\theta(f(x,\theta)=y) - P_\theta(f(x,\theta)=\hat{j}(x,y)) \\ &= E_\theta[I(f(x,\theta)=y) - I(f(x,\theta)=\hat{j}(x,y))] \end{aligned} \tag{6}$$

Define the original interval function as:

$$rmg(\theta,x,y) = I(f(x,\theta)=y) - I(f(x,\theta)=\hat{j}(x,y)) \tag{7}$$

That is, $mr(x,y)$ is the expected value of $mng(\theta,x,y)$, then:

$$mr(x,y) = E_\theta rmg(\theta,x,y) \tag{8}$$

To $\forall h$, both:

$$[E_{\hat{\theta}}h(\theta)]^2 = E_{\hat{\theta},\hat{\theta}}h(\theta)h(\theta') \tag{9}$$

where θ and θ' are independently and identically distributed, then:

$$mr(x,y)^2 = E_{\theta,\theta}rmg(\theta,x,y)rmg(\theta',x,y) \tag{10}$$

From the above equation $\mathrm{var}_{X,Y}\, mr(x,y)$ can be expressed as:

$$\begin{aligned} \mathrm{var}_{X,Y}\, mr(x,y) &= E_{X,Y}(mr(x,y))2 - (E_{X,Y}mr(x,y))2 \\ &= E_{X,Y}(E_\theta rmg(\theta,x,y))2 - (E_\theta E_{X,Y}rmg(\theta,x,y))2 \\ &= E_{\theta,\theta'}(\mathrm{cov}_{X,Y}(mg(\theta,x,y),rmg(\theta',x,y))) \\ &= E_{\theta,\theta'}(\rho(\theta,\theta')sd(\theta)sd(\theta')) \end{aligned} \tag{11}$$

When θ,θ' is constant, the correlation coefficient of $mg(\theta,x,y),mg(\theta',x,y)$ is $\rho(\theta,\theta')$; and $sd(\theta),sd(\theta')$ is the standard deviation of $rmg(\theta,x,y),rmg(\theta',x,y)$, respectively, thus:

$$\mathrm{var}_{X,Y}\, mr(x,y) = \bar{\rho}(E_\theta sd(\theta))^2 \le \bar{\rho}E_\theta\,\mathrm{var}(\theta) \tag{12}$$

$$\bar{\rho} = \frac{E_{\theta,\theta'}(\rho(\theta,\theta')sd(\theta)sd(\theta'))}{E_{\theta,\theta'}(sd(\theta)sd(\theta'))} \tag{13}$$

In summary, the upper bounds for $\mathrm{var}_{X,Y}\, mr(x,y)$ and PE^* are:

$$\mathrm{var}_{X,Y}\, mr(x,y) \le \bar{\rho}(1-s^2) \tag{14}$$

$$PE^* \le \bar{\rho}(1-s^2)/s^2 \tag{15}$$

From the above process, it can be obtained that in order to make the generalization ability of the random forest strong, the upper bound of its generalization error should be low. The generalization error is related to the correlation between the decision trees and the strength of the decision trees, i.e., the higher the strength of each decision tree in the random forest and the lower the correlation between the decision trees, the stronger the generalization ability of the random forest model.

2.2. SMOTENC algorithm

The SMOTENC method is an improved version of the SMOTE algorithm. The SMOTENC algorithm can deal with mixed datasets containing discrete and continuous variables, i.e., datasets containing both quantitative and fixed class variables. The algorithm solves the problem of unbalanced datasets by generating new synthetic samples to expand the few class samples in the dataset [28]. When generating new samples, the SMOTENC algorithm takes into account the distance relationship between the samples along with the category differences between the samples, thus ensuring that the generated synthetic samples can maintain the category distribution in the dataset. At the same time, the SMOTENC algorithm also uses a special interpolation method, i.e., linear interpolation between the samples, to ensure that the newly generated samples can maintain the discrete nature of the original dataset. The SMOTENC algorithm is able to better deal with imbalanced datasets with mixed data and achieve better performance in classification models. The principle of the SMOTENC algorithm is as follows:

Suppose the original dataset is:

$$D = (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \tag{16}$$

where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is the p attribute feature of the ind sample and y is the category label of the ith sample.The main steps of the SMOTENC algorithm are as follows:

For each minority class sample x_i, calculate the distance d_{ij} between it and all the samples in the original dataset, and sort them from smallest to largest distance to obtain the set of nearest k nearest neighbor samples as:

$$N_i = x_{i1}, x_{i2}, \ldots, x_{ik} \tag{17}$$

For each minority class sample x_i, n_i synthetic samples are generated based on the k nearest neighbor samples. Specifically, for the jth nearest neighbor sample x_{ij}, a random number between 0 and 1 is generated r. The eigenvalue of the synthetic sample z_j is then calculated, i.e:

$$z_j = x_i + r \times (x_{ij} - x_i) \tag{18}$$

where r is a randomly generated coefficient that controls the location of the generated synthetic samples, and x_i and x_{ij} are the feature vectors of sample x_i and its jth nearest neighbor sample, respectively.

The generated n_i synthetic samples are added to the original dataset to obtain a new dataset as:

$$D' = (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n), z_1, z_2, \ldots, z_{n_i} \tag{19}$$

Repeat the above steps for all the minority class samples to perform the sampling and generation operations to obtain the final expanded dataset D''.

In summary, the SMOTENC algorithm is an oversampling algorithm to solve the category imbalance problem by generating synthetic samples to expand the minority class samples and improve the prediction accuracy of the classifier.

2.3. Financial risk projection model

Based on the enterprise financial data sampled by SMOTENC algorithm, the model framework applied to enterprise financial risk prediction is proposed in combination with Random Forest algorithm as shown in Figure 1 [29]. Its specific process is as follows:
1) Data processing. , i.e., collect data related to enterprise financial risk, including solvency, operating capacity, cash flow, development capacity and profitability and other related data, and utilize SMOTENC algorithm for oversampling, data cleaning, missing value filling and outlier processing.
2) Use the random forest algorithm for feature variable screening, i.e., extract the feature variables that have an important impact on the prediction of corporate financial risk from the feature variables.
3) Use the random forest algorithm to train the screened feature variables, and conduct cross-validation during the training process to find the optimal parameter combinations, so as to improve the model's generalization ability and prediction accuracy.
4) After the model training is completed, the model is validated and evaluated using test data to ensure its reliable prediction.

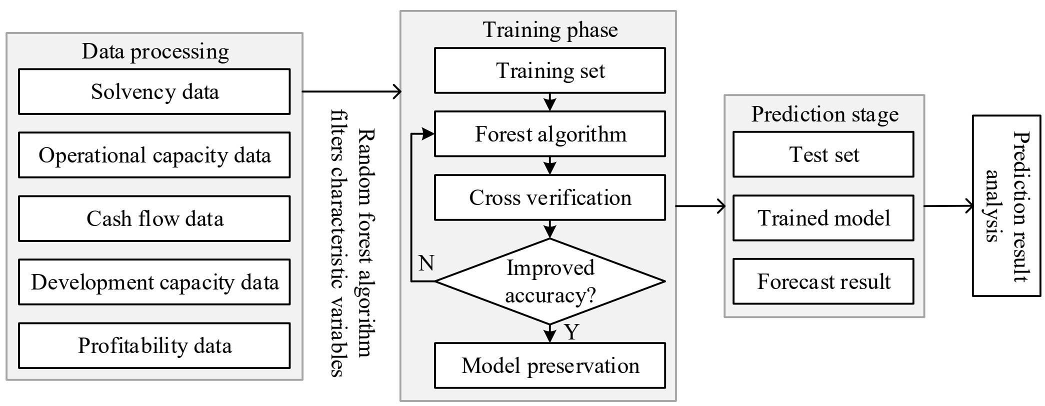


Fig. 1. Enterprise financial risk forecasting model

3. Financial risk characterization and treatment

The 21st century is an era of highly developed and rapid dissemination of information, as the most important micro-market subject of the enterprise, the expansion of its overall scale, the innovation of business management mode, and the improvement of social and economic benefits can not be separated from the efficient management of enterprise financial information. For the enterprise subject, the cognitive ability of financial management information, the enterprise's own economic strength, information access and management capabilities, and the number of effective information resources are not the same. Therefore, in the micro market, the distribution of enterprise financial information is obviously unbalanced, and enterprises will face various risks in the process of information management. In such a context, accurate prediction of the degree of risk of various types of financial management information faced in the process of enterprise operation and management, in order to eliminate various unfavorable factors in the operation process, reduce the business risk of the enterprise. Improving the quality of internal and external financial information of enterprises and realizing the accurate prediction of enterprise financial risk can minimize various risks in operation and increase the future development of the enterprise's income.

3.1. Sources for the selection of financial risk indicators

3.1.1. Selection of indicators of financial characteristics. In order to achieve a good prediction effect, the selection of enterprise financial risk prediction indicators abide by the rules of validity, rationality, completeness, accessibility and operability. Based on this, this paper finds that most scholars start from the enterprise's solvency, operating ability, profitability, cash flow ability and development ability to build the index system of enterprise financial risk prediction, because these five aspects of the ability to reflect the enterprise's financial risk situation more comprehensively. As a result, the enterprise financial risk prediction indexes constructed in this paper are shown in Table 1, mainly including profitability, operating ability, solvency, development ability and cash flow ability.

Table 1. Financial risk forecasting index

Primary indicator	Secondary indicator	Symbol
Profit ability	Operating margin	PA1
	Return on equity	PA2
	Total asset returns	PA3
Operational ability	Receivable turnover	OA1
	Inventory turnover	OA2
	Total asset turnover	OA3
Solvency ability	Mobility ratio	SA1
	Speed ratio	SA2
	Asset ratio	SA3
Development ability	Revenue growth	DA1
	Net profit growth rate	DA2
	Total asset growth rate	DA3
Cash flow ability	Cash current liability ratio	CA1
	Total cash debt ratio	CA2
	Sales cash ratio	CA3

The data of some indicators in the table can be obtained directly from the daily operation data of the enterprise, and some of the data can be obtained by simple arithmetic. Among them, the total return on assets is calculated by the formula:

$$PA3 = (Z_l + C_l) / \bar{Z} * 100\% \tag{20}$$

In the formula, Z_l represents total profit, C_l represents interest expense and $\bar{Z}$ represents average total assets. In the first-level indicators of development capacity, the growth rate of operating income, the growth rate of net profit and the growth rate of total assets are calculated as follows:

$$\begin{cases} DA1 = (Y_b - Y_s) / Y_s * 100\% \\ DA2 = (J_b - J_s) / J_s * 100\% \\ DA3 = (Z_b - Z_s) / Z_s * 100\% \end{cases} \tag{21}$$

In the formula, Y_b represents the current period's operating income, Y_s represents the previous period's operating income, J_b represents the current period's net profit, J_s represents the previous period's net profit, Z_b represents the current period's total assets, Z_s represents the previous period's total assets. According to the above operation, to obtain the specific value of each index, for the follow-up of the enterprise financial crisis risk prediction provides the basis.

3.1.2. Sources of data for the study sample. In this paper, based on the same industry and the same period, whether a listed company is directly ST (special treatment) due to “abnormal financial condition" is taken as the criterion of whether it is in financial risk, and Chinese listed companies whose listed companies are ST for the first time from 2016 to 2023 are selected as a sample of financial risk, of which the year in which a listed company is ST is denoted as T year. Due to the characteristics of the disclosure system of listed companies, the SEC determines whether a listed company is in an abnormal financial condition and decides whether it should be given special treatment based on its performance announced in its annual report of the previous year. So using its T-1 year data to judge whether the company will fall into financial risk in that year is not very meaningful and may bring the problem of exaggerating the model's predictive ability. At the same time, earlier accurate prediction can help companies take measures to cope with the crisis in advance, so in order to ensure the reliability and foresight of the prediction, this paper chooses the data set of companies in year T-3 as the research sample. In addition, in order to rigorously screen the process of data selection, the listed companies that are ST are processed according to the following methods and criteria after being screened out from them:

1) Excluding the listed companies that are directly ST due to "other abnormal conditions".
2) Excluding the sample of financial companies.
3) In the treatment of missing values of indicators, in order to retain the original information as much as possible, the samples with a large number of statistically significant missing values are excluded, otherwise, the interpolation method is used to deal with the missing values.
4) Companies that have been listed for less than 2 years, i.e., companies listed in 2022 and thereafter, are deleted.
5) In accordance with the principle of the same industry, the same period and similar asset size, a financial normal sample is randomly selected for each financial risk sample for control, so that the influence of industry attributes and asset size on financial risk prediction can be excluded.

After a series of sample selection and data processing as described above, 358 sample companies are finally obtained, including 179 positive and negative samples each. In terms of data sample division, the training set (80%) and test set (20%) are randomly selected in the ratio of 8:2. The data in this paper mainly comes from Cathay Pacific (CSMAR) database and RESSET database.

3.2. Data processing for financial risk indicators

3.2.1. Sample financial data processing. In the process of predicting the risk of corporate financial crisis, because there is some correlation and overlap between each financial index variable, and the dimensionality of the financial data is relatively high, this will lead to a significant increase in the calculation of the subsequent prediction model, which in turn affects the efficiency and accuracy of the prediction. In order to solve this problem, it is particularly important to introduce the principal component analysis (KPCA) method to reduce the dimensionality of the data.

The basic idea of KPCA lies in the fact that the method utilizes nonlinear mapping technology to map the sample points of the original high-dimensional data space into a high-dimensional or even infinite-dimensional feature space, so that the data are linearly differentiable in this new space. The advantage of this is that the dimensionality of the data can be reduced without losing too much information, thus simplifying the computational complexity of the subsequent prediction model [30]. The dimensionality reduction process of KPCA can be divided into the following steps:

First, a suitable kernel function (e.g., Gaussian kernel function or polynomial kernel function) is selected, based on which each sample point of the original dataset is corresponded to a higher dimensional feature space, so as to obtain the covariance matrix of a high-dimensional sample. I.e:

$$j(X)j(X)^T \omega_j = l_j \omega_j,\ j = 1, 2, \ldots, L, d \tag{22}$$

where $j(X)$ denotes the vector of high-dimensional mapped samples of dimension $D * l(D \square d)$, l_j denotes the eigenvalues, and ω_j denotes the vector of dimensions corresponding to the eigenvalues.

Next, the kernel matrix, which is the inner product representation of each sample point in the new space, is computed, and then the kernel matrix is centered to eliminate the mean shift in the data.

Then, the eigenvectors and eigenvalues are obtained by eigenvalue decomposition. On this basis, the largest set of each eigenvector is used as the eigenspace after dimensionality reduction.

Finally, on this basis, the original sampling points are mapped to obtain the dimensionality reduced data as:

$$x^{new} = \omega_i^T j(x) \tag{23}$$

where x^{new} represents the data after dimensionality reduction.

Through the dimensionality reduction process of KPCA, the redundant information and noise in financial data can be effectively removed and the most representative features can be retained. At the same time, the dimensionality reduced data is easier to understand and visualize, which helps to better analyze and explain the patterns and laws in the financial data.

In addition to sample characteristic dimensionality reduction, sample balance is also an important influencing factor. In practical applications, the uneven distribution of the number of samples of different categories of data is encountered, which triggers the phenomenon of overfitting. Overfitting may make a model perform well in the training set but poorly in the test set or in reality, thus reducing the prediction accuracy. In order to solve this problem, the SMOTENC algorithm based on the previously proposed is handled to carry out the data oversampling process. The SMOTENC algorithm can effectively increase the number of samples in a few categories, making the number of samples in different categories more balanced.

3.2.2. Financial risk indicator tests. In order to determine whether the data between the 15 financial indicators of the selected 358 enterprises are consistent with normal distribution, the normality test is carried out first. If it meets, T-test can be carried out on the sample indicators to determine whether there is a significant difference between the underlying indicators, and if not, non-parametric tests can be used to verify whether there is a significant difference between the indicators. Only the existence of significant differences between the screened indicators can improve the feasibility of financial early warning. The normality test results of the financial indicator data are shown in Table 2. As can be seen from the table, the significance level of the 15 financial risk indicators designed in this paper is lower than 0.05, indicating that none of the 15 indicators conform to the normal distribution, and non-parametric tests are needed.

Table 2. The normal test results of the financial index data

Secondary indicator	Symbol	Z value	P value
Operating margin	PA1	5.474	0.026
Return on equity	PA2	3.332	0.012
Total asset returns	PA3	6.675	0.047
Receivable turnover	OA1	5.438	0.009
Inventory turnover	OA2	8.911	0.034
Total asset turnover	OA3	6.837	0.011
Mobility ratio	SA1	6.216	0.033
Speed ratio	SA2	6.932	0.006
Asset ratio	SA3	4.206	0.015
Revenue growth	DA1	4.095	0.018
Net profit growth rate	DA2	6.598	0.009
Total asset growth rate	DA3	7.132	0.028
Cash current liability ratio	CA1	4.737	0.004

Total cash debt ratio	CA2	5.116	0.000
Sales cash ratio	CA3	8.689	0.001

Non-parametric test is a test method to judge the overall parameters of the sample data using all the sample data in the process of data analysis when a clear judgment cannot be made on the overall distribution pattern of the sample data, with the purpose of judging whether there is a significant difference between the indicators. In this paper, the Mann-Whitney U test is used for non-parametric testing, and the test results of financial risk indicators are shown in Table 3. As can be seen from the table, the 15 financial risk evaluation indicators selected in this paper all pass the Mann-Whitney U test, and their significance is lower than 0.05. Therefore, the financial risk evaluation indicators designed in this paper can effectively reflect the financial risk situation of the enterprise, and can provide effective data support for the prediction of the enterprise's financial risk.

Table 3. The results of the financial risk index

Secondary indicator	Symbol	P value	Symbol	Symbol	P value
Operating margin	PA1	0.012	Revenue growth	DA1	0.005
Return on equity	PA2	0.003	Net profit growth rate	DA2	0.001
Total asset returns	PA3	0.015	Total asset growth rate	DA3	0.009
Receivable turnover	OA1	0.007	Cash current liability ratio	CA1	0.015
Inventory turnover	OA2	0.013	Total cash debt ratio	CA2	0.005
Total asset turnover	OA3	0.000	Sales cash ratio	CA3	0.013
Mobility ratio	SA1	0.002	-	-	-
Speed ratio	SA2	0.008	-	-	-
Asset ratio	SA3	0.016	-	-	-

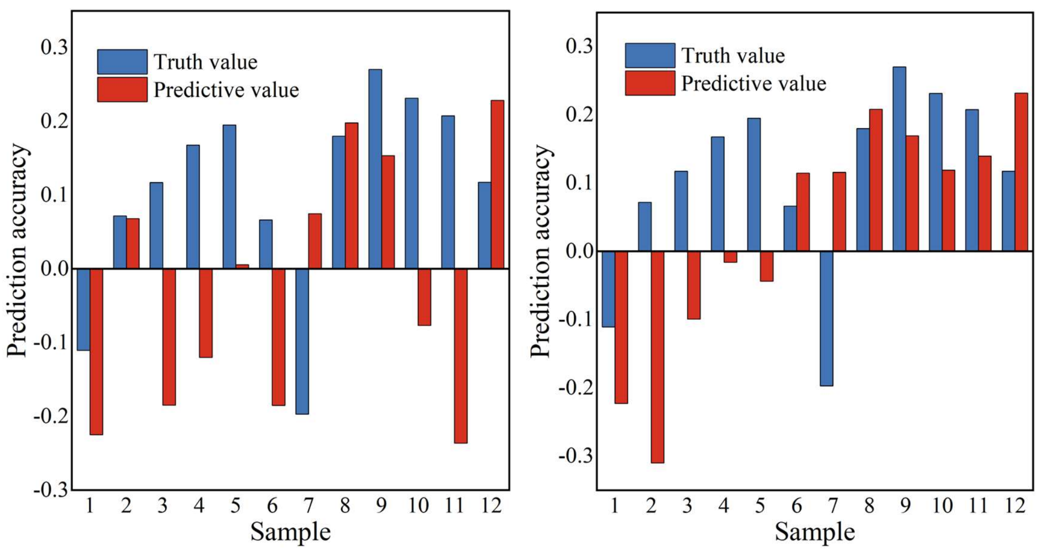
4. Validation of enterprise financial risk projections

In the era of "Internet +", the world economy is becoming increasingly globalized and informatized, and the business development of enterprises is facing unprecedented opportunities, but at the same time, it is also subject to the constraints of environmental factors, such as the economic market factors, laws and regulations, social and cultural factors, and the policy environment and other unpredictable environmental factors, which have brought uncertainty to the enterprise's financial situation. Uncertainty, the financial risks faced by enterprises are increasing. Most enterprises encounter serious financial crisis or even bankruptcy in the late stage because they do not pay enough attention to the financial problems in the early stage and take effective means to deal with the crisis situation in time, which is unfavorable to the subsequent development of the enterprise. Therefore, how to accurately identify enterprise financial risk through a reasonable enterprise financial risk prediction model has become a research hotspot to ensure the stable operation and development of enterprises.

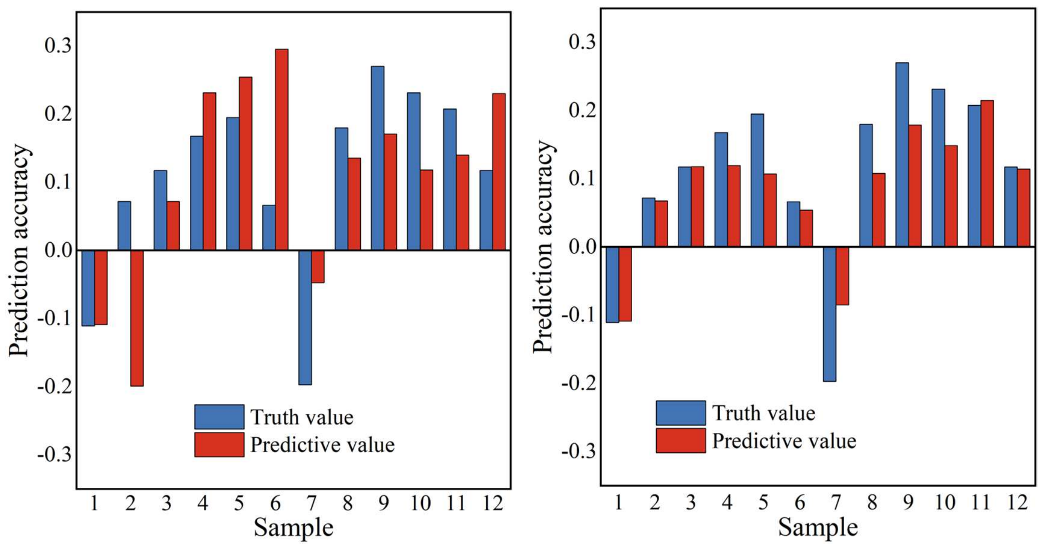
4.1. Financial risk prediction model validation

4.1.1. Comparison of model prediction effects. In order to verify the effectiveness of this paper's model in making enterprise financial risk prediction, three machine learning algorithms, namely, Support Vector Machine (SVM), Nearest Neighbor (KNN), and Linear Regression (LR), are selected as comparison algorithms, and 12 sets of test samples are randomly selected from the data set and input into each algorithmic model respectively for comparison. The predicted values are compared with the actual values to predict the judgment results of financial risk. Figure 2 shows the prediction results of the four algorithms, of which Figures 2(a)~(d) show the prediction results of SVM, LR, KNN, and RF algorithms, respectively.

From the figure, it can be found that all four algorithms have different effects in the comparison of predicted and actual values. In the financial risk prediction results of SVM algorithm, except for samples 2 and 8, the predicted value of each sample has a certain gap with the actual value, especially samples 3, 4, 10 and 11, where there is a large gap between the predicted value of financial risk and the actual value. In the prediction results of the LR algorithm, the predicted values of samples 2, 3, 4, 5, and 7 have a certain gap with the actual values, especially the predicted values of samples 2, 3, and 7 have a large gap with the actual values, while in samples 9 to 11 the predicted values are a certain distance away from the real values, and the predicted values of samples numbered 8 and 12 are completely more than the real values. In the prediction results of the KNN algorithm, the predicted value of each financial data sample has a certain gap with the real value, only the predicted value of sample 2 has a large gap with the actual value, and the other samples do not have a large gap with the real value, which indicates that KNN has a certain advantage in financial risk prediction. Random Forest's financial risk prediction effect is the best, are closer to the actual value, although the predicted values of samples 5, 7, 8, 9 and 10 have a certain gap with the real value, but the results of these predictions are not much different from the real value. The synthesis of the above algorithm prediction results, indicating that the random forest algorithm proposed in this paper has a better prediction effect in the prediction of corporate financial risk, and can effectively improve the prediction accuracy of financial data.



(a) SVM (b) LR



(c) KNN

(d) RF

Fig. 2. The prediction of four algorithms

In addition, in order to further validate the effectiveness of the RF model, this paper classifies four categories of financial risks in the dataset, i.e., no risk (T1), slight risk (T2), moderate risk (T3), and severe risk (T4). The three machine learning algorithms mentioned above are still chosen as a comparison, and the confusion matrix is used to verify the accuracy of each algorithm in making financial risk predictions. Figure 3 shows the confusion matrices of different algorithms, where Figures 3(a)~(d) show the confusion matrices of SVM, LR, KNN, and RF algorithms, respectively.

It can be seen from the figure:

1) The RF classifier has the best prediction accuracy for samples with different levels of corporate financial risk, with prediction accuracies of 91.23%, 92.35%, 90.13%, and 94.17% for heavy, moderate, slight, and no risk samples, respectively. Even the misclassified samples were only misclassified as neighboring categories.

2) SVM classifier and LR classifier have similar prediction accuracies and are better than KNN classifier for slight and no financial risk samples. Among them, the prediction accuracy of RF and KNN classifiers for no financial risk is higher than 90%, and the prediction accuracy of SVM and LR for no financial risk is lower than 90%. Except for RF, KNN, SVM and LR are less effective in predicting moderate financial risk, and their prediction accuracies are 78.95%, 48.54% and 38.24%, respectively.

3) Overall, the prediction accuracy of the KNN classifier for corporate financial risk is low relative to the RF algorithm, and samples with moderate, slight and no risk are misclassified as moderate financial risk, while the prediction effect of the sample with heavy financial risk is second only to the RF classifier, and its prediction accuracy is 92.75%.

In summary, the enterprise financial risk prediction model proposed by this paper based on the random forest algorithm has superior prediction effect, can be more accurately predicted to the enterprise financial risk level, and provide accurate identification results for enterprises to take reasonable measures to avoid financial risks.

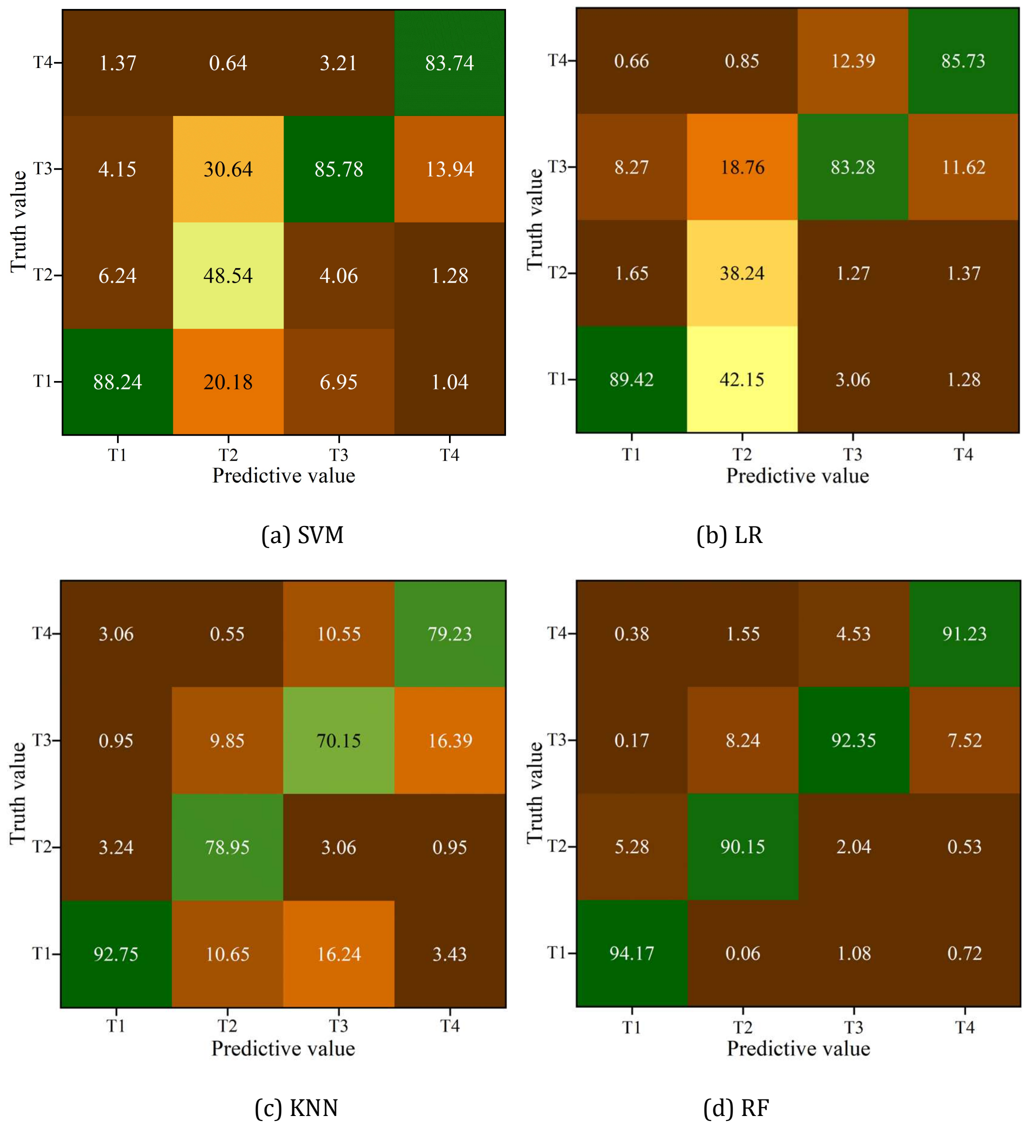


Fig. 3. Confusion matrix of different algorithms

4.1.2. Time efficiency and security. Among the model input factors, the actual situation, profitability, operation and growth have a stronger relationship with time efficiency, and the time efficiency of prediction is faster, which can timely understand the overall situation of corporate finance, thus enabling enterprises to quickly make risk prevention and control measures. So in the time efficiency analysis mainly for these four factors for prediction model input. In the experimental process, the data of enterprise finance are grouped and organized into 20 groups, and the test data are used to compare and analyze the time spent by SVM, LR, KNN and RF algorithms in predicting enterprise financial risks, and their specific results are shown in Figure 4.

As can be seen from the figure, the RF algorithm proposed in this paper has an average value of time efficiency of 0.68% when predicting corporate financial risks, which is 0.09, 0.07 and 0.17 percentage points higher than the KNN, SVM and LR algorithms, respectively. This shows that this paper has higher efficiency in dealing with enterprise financial risk data, and has better time efficiency on the basis of ensuring the accuracy of financial risk prediction, which further validates the effectiveness of this paper's method.

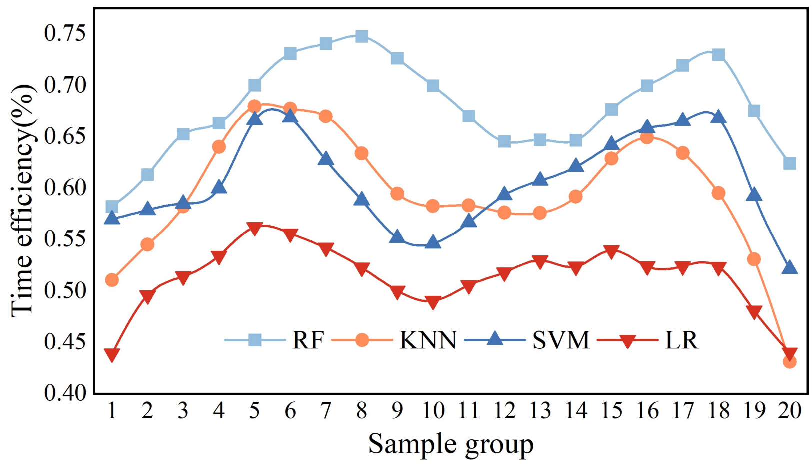


Fig. 4. Algorithm time efficiency

In the model input factors, the actual situation, growth situation and cash flow have a strong relationship with the security of the enterprise financial risk, because of the existence of confidentiality of the enterprise financial data, the model should minimize the occurrence of security problems, so in the security analysis can be predicted from these 3 factors model input. In the experimental process, the data of enterprise finance is grouped and organized into 20 groups, and the test data is used to compare and analyze the problem situation of the 4 algorithms, and its specific results are shown in Figure 5.

According to the experimental results, it can be seen that the model proposed in this study has fewer security problems in the process of enterprise financial risk prediction, and only 4 problems occur in 20 groups of tests, which is better than other models, and to a certain extent, it can reduce the occurrence of security problems. While realizing the prediction of enterprise financial risk, the security of enterprise financial data is better guaranteed.

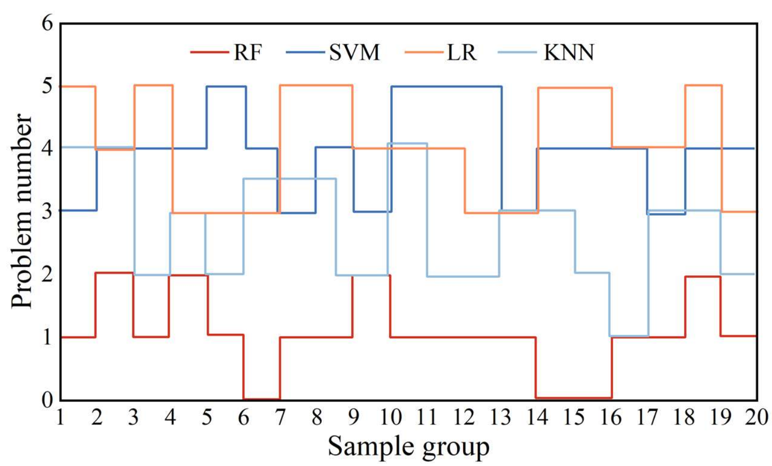


Fig. 5. Safety analysis chart

4.2. Example analysis of financial risk forecasting

4.2.1. Debt service and operational capacity. Based on the corporate financial risk prediction model established in the previous section, the early warning indicator of Company W in 2022 was selected as its future prediction target through risk assessment of the company's financial situation. The financial situation of Company W in 2023 is predicted to be a slight financial risk using the financial risk prediction model, which is consistent with the actual financial risk situation of the company in 2023. This shows that the enterprise financial risk prediction model based on the random forest algorithm has high prediction accuracy and can provide an effective method for the financial risk prediction of listed enterprises.

In this paper, we will analyze the financial situation of Company W to find the specific financial risk factors of the company. This paper takes the financial risk indicators of Company W from 2016 to 2022 as an example, and compares and analyzes the solvency and operating capacity. Table 4 shows the changes of solvency and operating capacity indicators of Company W between 2016~2022.

As can be seen from the table, Company W's current ratio has been decreasing every year, and the main factor for its reduction is the continuous expansion of its debt, which reached $50 million in 2021. The turnover in 2022 will be 96.4857 million, an increase of 120.42% over 2020, and the company's current debt is also rising. This index reflects the company's short-term debt repayment ability, which means that the company is stronger, but if it is too high, if the company's short-term debt exceeds a certain threshold, it will lead to the company's working capital not working smoothly, which will reduce the company's profits. Judging from the current situation, the short-term debt capacity of Company W is very normal. In addition, the proportion of receivables of Company W is too high, and if the proportion of receivables is too high, there is a possibility of bad debts, so it is necessary for the company to improve its short-term debt repayment ability from the perspective of the current ratio index. The shareholding structure reflects not only the shareholding structure of the company, but also the company's debt profile. The index reflects the long-term debt burden, which is too high, the debt ratio is too high, the demand for more interest payments, and the rising level of fees. This ratio is on the low side, making it impossible for companies to fully leverage their financing and invest more capital in their daily operations. It is necessary to combine the actual situation of enterprises and appropriately borrow in order to create greater economic benefits for enterprises. Although Company W has a smaller percentage of equity ownership, the overall growth rate is relatively fast, reaching 58.6% in 2022, indicating that Company W's debt level is gradually increasing.

In addition, between 2016 and 2022, Company W's inventory turnover has been decreasing steadily, from 16.014 in 2016 to 2.728 in 2022, and it has decreased by 13.286 percentage points over the past seven years. The production of inventory is faster than the rate of sales, which indicates that Company W has some inventory turnover mismatch on a daily basis, which results in the business having too high a percentage of working capital and a lack of liquidity as it grows. Such liquidity will affect the capital flow of the business, indicating that the business is not doing well. Over the past seven years, the fixed asset turnover index has gradually declined from 3.112 in 2016 to 2.024 in 2022, and although the decline is not significant, there is no sign of growth in this index over the past seven years. It indicates that the fixed assets of the firms are not being used efficiently and the management class is not high. The increase in 2021 is higher than that of 2020, which is mainly due to the fact that the revenue outweighs the increase in fixed assets.

Table 4. Solvency and operational capacity

Year	Mobility ratio	Equity ratio	Inventory turnover	Fixed asset turnover
2016	13.213	0.028	16.014	3.112
2017	12.651	0.013	15.276	3.055
2018	3.682	0.427	3.769	2.328
2019	2.577	0.235	3.782	2.139
2020	1.832	0.468	3.035	1.957
2021	1.859	0.543	2.816	2.146
2022	1.794	0.586	2.728	2.024

4.2.2. Profitability and development capacity. For the profitability index of Company W, it is mainly shown through operating profit margin, return on net assets and total return on assets, and by studying the level of its profitability, it is able to understand the profitability level of the company.

Figure 6 shows the results of profitability development of Company W. As can be seen from the figure, Company W's operating profit margin, return on net assets, and total return on assets have been decreasing from 2016 to 2022, and although each indicator has increased in 2021 and 2022, the enterprise's operating profit margin is much lower than the historical level of 2019, and the total return on assets is much lower than the historical average level before 2019, which indicates that Company W still needs to continuously improve the enterprise's profitability and that there is still a lot of room for the business to make a profit.

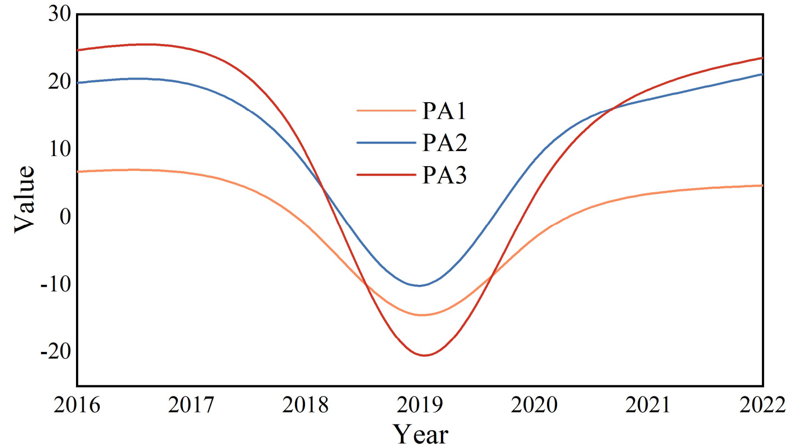


Fig. 6. Profitability development results

This paper measures the development ability of Company W based on the three financial indicators selected above, the growth rate of operating income reflects the growth rate of the company's operating income compared with last year, and the growth rate of total assets reflects the expansion rate of the company's capital scale. Figure 7 shows the development capacity of Company W in 2016~2022. As can be seen from the figure, the operating income of Company W will decline sharply in 2020~2021, and the operating income will increase slightly in 2022, but the growth rate is far lower than the industry average of 1.87%. At the same time, the net profit growth rate of W company in 2019~2022 is less than 0, indicating that the company's shareholder value is gradually decreasing. In 2021, the growth rate of W's operating income will increase compared with 2020, but the operating income is far lower than the industry average of 2.67*1.04 million yuan, and its operating income growth rate will decrease in 2022. This shows that Company W lacks competitiveness and insufficient development potential.

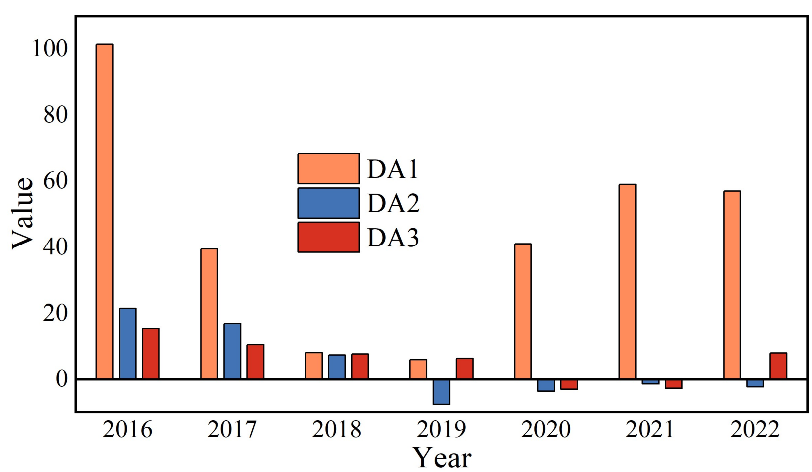


Fig. 7. The ability of enterprises to develop

In summary, analyzing the development situation of Company W from four aspects, it is not difficult to find that Company W does have certain financial risks. For solvency, Company W's short-term solvency is insufficient, its gearing ratio is low and its capital structure is unreasonable. From the analysis of profitability, the return on corporate net assets has been falling, indicating that compared with other companies, Company W's profitability is weak. Analyzing from the operating ability, the company's inventory turnover ratio is far below the industry average, and the accounts receivable turnover ratio is also decreasing. Analyzing from growth ability, the overall net asset growth rate of the company declines year by year, the net profit growth rate is much lower than the industry average, and the growth potential of the enterprise is insufficient. The model prediction results are consistent with the actual situation of the company, and it can be proved that the random forest model constructed in this paper can effectively predict the degree of financial risk that exists in W company.

5. Conclusion

The article constructs an enterprise financial risk prediction model based on the random forest algorithm, oversamples and downsizes the enterprise financial data through SMOTENC algorithm and KPCA algorithm, and verifies the effectiveness of the enterprise financial risk prediction model through examples.

1) Under the sample data of different enterprise financial risk levels, the financial risk prediction accuracy based on random forest is the best, and the prediction accuracy of heavy, medium, slight and no risk samples is 91.23%, 92.35%, 90.13% and 94.17%, respectively. The RF algorithm has a mean value of 0.68% of time efficiency in enterprise financial risk prediction, and the security problem appears only 1~2.

2) Based on the results of financial risk prediction, the enterprise financial risk factors can be excavated, and quantitative data analysis can help enterprises better avoid financial risks, and provide a guarantee for the stable and healthy development of enterprises.

About the Author

Ke Sun was born in Anyang, Henan. P.R. China, in 1988. She obtained a doctoral degree from Philippine Christian University in the Philippines. I am currently working at the school of accountancy, Guangzhou College of Technology and Business. My main research direction is corporate finance and financial analysis.

Yupeng Li was born in Anyang, Henan. P.R. China, in 1985. He obtained a doctoral degree from Jose Rizal University in the Philippines. I am currently working at the school of accountancy, Anyang Institute of Technology. My main research direction is corporate finance and financial analysis.

References

[1] Fedoryshyna, L., & Todosiychuk, V. (2019). Analiz controlling financial risks of enterprise. Polish journal of science.-2019.-№ 21, Vol. 2.-P. 30-42.

[2] Xu, J. (2018). Enterprise Risk Management and the Cost of Capital. Journal of Risk & Insurance, 85(1), 159-201.

[3] Faisal, F., Abidin, Z., Haryanto, H., & Seetharam, Y. (2021). Enterprise risk management (ERM) and firm value: The mediating role of investment decisions. Cogent Economics & Finance, 9(1), 2009090-2009090.

[4] Hunziker, S. (2021). Enterprise Risk Management. Springer Books.

[5] Akash, T. R., Reza, J., & Alam, M. A. (2024). Evaluating financial risk management in corporation financial security systems. World Journal of Advanced Research and Reviews, 23(1), 2203-2213.

[6] Wen, F., Li, C., Sha, H., & Shao, L. (2021). How does economic policy uncertainty affect corporate risk-taking? Evidence from China. Finance Research Letters, 41(C).

[7] Hung, S., & Qiao, Z. (2017). Shadows in the sun: crash risk behind earnings transparency. Journal of Banking & Finance, 83, 1-18.

[8] Yahyoyevna, O. F., & Tatulovna, Z. M. (2024). MECHANISM OF ORGANIZING FINANCIAL MANAGEMENT IN ENTERPRISES. Miasto Przyszłości, 47, 866-872.

[9] Mazur, N., Khrystenko, L., Pásztorová, J., Zos-Kior, M., Hnatenko, I., Puzyrova, P., & Rubezhanska, V. (2021). Improvement of Controlling in the Financial Management of Enterprises. TEM Journal: Technology, Education, Management, Informatics, 10(4), 1605-1609.

[10] Nizam, E., Ng, A., Dewandaru, G., Nagayev, R., & Nkoba, M. A. (2019). The impact of social and environmental sustainability on financial performance: A global analysis of the banking sector. Journal of Multinational Financial Management, 49(C), 35-53.

[11] Cameron, C. (2017). The strategic and legal risks of work-integrated learning: An enterprise risk management perspective. International Journal of Work-Integrated Learning, 18(3), 243.

[12] Williams, J. R., Haka, S. F., Bettner, M. S., & Carcello, J. V. (2018). Financial & managerial accounting: The basis for business decisions. McGraw-Hill.

[13] Uddin, M. S., Chi, G., Al Janabi, M. A., & Habib, T. (2022). Leveraging random forest in micro‐enterprises credit risk modelling for accuracy and interpretability. International Journal of Finance & Economics, 27(3), 3713-3729.

[14] Javaid, H. A. (2024). Improving Fraud Detection and Risk Assessment in Financial Service using Predictive Analytics and Data Mining. Integrated Journal of Science and Technology, 1(8).

[15] Peng, X., & Huang, H. (2020). Fuzzy decision making method based on CoCoSo with critic for financial risk evaluation. Technological and Economic Development of Economy, 26(4), 695.

[16] Song, Y., & Peng, Y. (2019). A MCDM-Based Evaluation Approach for Imbalanced Classification Methods in Financial Risk Prediction. IEEE Access, 7, 84897-84906.

[17] Malakauskas, A., & Lakštutienė, A. (2021). Financial distress prediction for small and medium enterprises using machine learning techniques. In Inžinerinė ekonomika (No. 32, pp. 4-14).

[18] Sun, X., & Lei, Y. (2021). Research on financial early warning of mining listed companies based on BP neural network model. Resources Policy, 73(C).

[19] Sun, J., Fujita, H., Zheng, Y., & Ai, W. (2021). Multi-class financial distress prediction based on support vector machines integrated with the decomposition and fusion methods. Information Sciences, 559, 153-170.

[20] Cao, Y., Shao, Y., & Zhang, H. (2022). Study on early warning of E-commerce enterprise financial risk based on deep learning algorithm. Electronic Commerce Research, 22(1).

[21] Wu, D., Ma, X., & Olson, D. (2022). Financial distress prediction using integrated Z-score and multilayer perceptron neural networks. Decision Support Systems, 159, 113814-113814.

[22] Ju, C., & Zhu, Y. (2024). Reinforcement Learning-Based Model for Enterprise Financial Asset Risk Assessment and Intelligent Decision-Making. Applied and Computational Engineering, 97, 181-186.

[23] Wang, D. N., Li, L., & Zhao, D. (2022). Corporate finance risk prediction based on LightGBM. Information Sciences, 602, 259-268.

[24] Valaskova, K., Kliestik, T., Svabova, L., & Adamko, P. (2018). Financial Risk Measurement and Prediction Modelling for Sustainable Development of Business Entities Using Regression Analysis. Sustainability,

10(7), 1-15.

[25] Kim, A., Yang, Y., Lessmann, S., Ma, T., Sung, M. C., & Johnson, J. E. V. (2020). Can deep learning predict risky retail investors? A case study in financial risk behavior forecasting. European Journal of Operational Research, 283(1), 217-234.

[26] Mayovets, Y., Vdovenko, N., Shevchuk, H., Zos-Kior, M., & Hnatenko, I. (2021). SIMULATION MODELING OF THE FINANCIAL RISK OF BANKRUPTCY OF AGRICULTURAL ENTERPRISES IN THE CONTEXT OF COVID-19. Journal of Hygienic Engineering & Design, 36.

[27] Abiona Qozeem O. & Head Monique H. (2025). Determination of Bridge Elements' Weights Using the Random Forest Algorithm. Journal of Performance of Constructed Facilities(1).

[28] Jin Yupeng,Yang Junfeng,Yang Xu & Liu Zhongchao. (2024). Cross-domain bearing fault diagnosis method based on SMOTENC and deep transfer learning under imbalanced data. Measurement Science and Technology(1).

[29] Fengyiyi Chen. (2024). Financial Risk Prediction Model Based on Categorical Regression Tree Algorithm from the Perspective of Education Management. Accounting, Auditing and Finance(3).

[30] Chunli Liu,Xiaolong Zhang & Jiarui Bai. (2024). Fault Diagnosis of Aircraft Hydraulic Pipeline Clamps Based on Improved KPCA and WOA–KELM. Processes(11),2572-2572.