International Journal for Housing Science and Its Applications
Publish August 6, 2025. Volume 46, Issue 3 Pages 4684-4696

https://doi.org/10.70517/ijhsa463401

Research on the application of corporate financial forecasting model based on gradient elevator in the economic landscape of the new era

Tong Xie[1,*]
[1] PricewaterhouseCoopers Zhong Tian LLP, Beijing Branch, Beijing, 100000, China
Corresponding authors: (e-mail: tong.xie.phd@163.com).

Abstract With the transformation, upgrading and continuous development of China's economy, the competition among enterprises is becoming more and more intense, and the financial risks they face are also increasing. Therefore, in order to ensure the sustainable and healthy development of enterprises, it is particularly important to strengthen the prediction of financial risk. This paper takes listed transportation enterprises as research samples, selects 13 indicators from 4 aspects, and reduces the dimensionality of financial indicators through factor analysis, and finally extracts 11 principal components for model fitting. Then the combined model SMOTE-Light GBM for financial prediction of medium enterprises is proposed, and the seed of random numbers is selected as 1326, and the model's accuracy, precision, recall, f1_score, ROC curve, and AUC evaluation indexes are all over 95%, and the classification prediction effect is excellent. Operating profit margin, total asset turnover, total asset growth rate, operating income growth rate, and accounts receivable turnover have significant effects on corporate financial forecasting.

Index Terms corporate financial forecasting, factor analysis, Light GBM, SMOTE oversampling

I. Introduction

Enterprise financial forecasting is based on sales forecasts, based on sales forecasts to prepare production forecasts, and then prepare forecasts of inventory, investment, costs and expenses, profits, cash flow and so on [1]. The projected financial report includes projected balance sheet, projected income statement and projected cash flow statement, and the projected financial indicator system mainly includes four aspects: profitability, solvency, operating ability and growth ability [2], [3]. Enterprise financial forecasting based on the projected financial statements, centered on the financial indicator system, through the comprehensive analysis of financial indicators, forecasting, timely reflection of the changes in the financial situation of the enterprise and the business situation, and the various aspects of the enterprise or the possible occurrence of business risks to issue a risk early warning signal, to provide the management with a basis for decision-making [4]-[6].

Traditional financial forecasting lacks dynamism, the most direct manifestation is its static nature [7]. Specifically, including, the static nature of the premise, the premise of the traditional financial plan is a stable environment, so that the future predictability is large. The static nature of the implementation, with the plan requires the responsible unit or responsible person, rather than the other way around. Insufficient attention to changes in the environment and its impact, the traditional financial forecasting, although the theory also includes a feedback mechanism, but not enough research on how to reflect the difference between the actual and forecast in a timely manner to the financial forecast [8]-[10]. Due to the insufficient reflection of the dynamics of strategic decision-making, it leads to the enterprise capital decision-making is not only a huge amount of money, but also has the irrevocability of inputs, so that the capital investment decision must be very cautious [11], 12]. In summary, the development direction of financial forecasting innovation is to enhance the dynamics of financial forecasting, and the problem of financial forecasting dynamics has become a key issue in the study of financial forecasting models.

Since the 1930s, scholars at home and abroad have made many research results in financial forecasting. From the perspective of the timeline, the research methods from univariate analysis to the introduction of Logistic regression to the current data mining algorithms such as neural networks, support vector machines, decision trees and other data mining algorithms and combinatorial models, the accuracy of corporate financial forecasting is constantly improving [13]-[15]. Hassan, E et al. constructed a financial forecasting model using logistic regression analysis with 18 financial ratios as independent variables and validated it with Pakistani stock exchange data for the period of 2005-2020 and the predictions were fair [16]. With the development of machine learning algorithms,

Mishra, N et al. compared the performance of Logistic regression, LDA (Implicit Dirichlet Distribution) and ANN (Artificial Neural Networks) models in predicting finances and found that ANNs have better predictive accuracy than traditional statistical models such as logistic and LDA models [17]. Ali, M. M et al. utilized Long Short-Term Memory (LSTM) recurrent neural networks to predict financial indicators of industrial companies and found that LSTM models performed better than classical statistical tools on this task [18]. Chen, M. Y et al. developed a financial forecasting model combining particle swarm optimization (PSO) and support vector machine (SVM), and the PSO-SVM model outperformed other machine learning models in predicting financial risks on a standard data set [19]. Senoguchi, J A two-layer Genetic Algorithm (GA) approach for optimizing the decision tree structure of complex financial big data prediction models, a combined approach that outperforms traditional machine learning methods and supports small-sample predictions [20]. Wang, Z constructed a financial forecasting system for Chinese listed companies based on analyzing the financial indicators of Chinese listed companies from 2013 to 2018, using a combination of random forest model and decision tree model [21].

With the depth of the study, researchers are no longer limited to the study of financial data, but also focus on the financial text and mining of financial text [22]. Researchers' mining of financial texts generally focuses on the texts on company performance disclosure, audit reports and company annual reports [23]. Yang, R et al. introduced a text-mining approach to assess financial, strategic, operational and hazard risks in corporate annual reports and investigated the relationship between these financial risk indicators and auditing, leading to financial risk prediction [24]. Kanungsukkasem et al. proposed a topic modeling method called Financial Latent Dilichorean Allocation (FinLDA), which extracts features from textual and financial time series data to improve the accuracy of financial forecasting models [25]. García-Méndez et al. combined topic segmentation, co-reference resolution, LDA topic modeling and discourse temporal analysis to automatically detect relevant financial data in unstructured financial news texts, and used natural language processing to perform text processing and text mining on the texts [26]. In summary, carrying out research on financial prediction modeling can effectively resolve and prevent business risks and financial risks, which not only contributes to the normal development of enterprises, but also has an inestimable role in the stability of the financial order and the healthy development of society.

In this paper, listed transportation enterprises are selected as research cases, and the financial evaluation index system is constructed from four aspects of profitability, solvency, operating ability, and development ability, and the factor analysis method is introduced to analyze each factor, and then evaluate the enterprise finance. The data indicators are summarized into 11 principal components through principal component analysis, and SMOTE-Light GBM is used as a prediction classification model to realize the enterprise financial prediction. Subsequently, in order to verify the effectiveness of the algorithm, the accuracy, precision, recall, f1_score, ROC curve and AUC value are empirically investigated, and the effect of SMOTE-Light GBM model is comparatively analyzed with other models, and the characteristic importance chart is obtained.

II. Sample Selection and Corporate Financial Forecasting Evaluation Index System

II. A. Sample selection and data sources

II. A. 1) Principles of Sample Selection

(1) Clarify the industry classification of the transportation industry

Based on the industry classification standards in China's National Economic Classification (GB/T4754-2017) to confirm the transportation industry, the railroad transportation industry with code beginning with G53, the road transportation industry with code beginning with G54, the water transportation industry with code beginning with G55, and the air transportation industry with code beginning with G56 are selected as the main objects of research.

(2) Selection of A-share listed transportation companies

Given the availability and typicality of the data, transportation companies currently listed on China's stock exchanges in A-shares are selected.The investors of A-shares (RMB ordinary shares) are individuals, organizations or institutions in China, and the auditing is performed by domestic accounting firms, applying domestic accounting standards; whereas, the nominal value of B-shares is denominated in RMB, but the purchase and sale need to be conducted in foreign currencies.The B-share companies are audited by foreign accounting firms, adopting international accounting standards. International Accounting Standards (IAS) are applied. Since domestic accounting standards and international accounting standards have some differences in the implementation rules, the data information of A-share and B-share cannot be put together for comparison. Therefore, this paper only selects A-share listed companies when selecting research samples.

(3) Select companies with complete data

The data of some listed companies are missing in some years or some indicators due to statistical or other reasons. Some statistical methods and software allow missing data values when using them, while some do not, so

in order to avoid unnecessary trouble in the later empirical analysis, this paper directly excludes listed transportation companies with incomplete data when selecting samples.

II. A. 2) Sample selection and data sources

Due to the small number of listed companies in the transportation industry, and the emergence of more data omissions and missing data in the latest financial data, so according to the above principles of sample selection, considering the integrity of the data, as well as trying to make the sample set have more data integrity of listed transportation companies, this paper finally selects 66 transportation companies listed on the A-share market in the financial data of 2016-2023 for analysis and research, and the data source is CSMAR database.

II. B.Preliminary selection of forecasting indicators

II. B. 1) Principles for selecting financial indicators

(1) Effectiveness. The ultimate goal of this paper is to reasonably and effectively predict the financial risk of transportation enterprises, so the selected indicators must be able to reflect the financial situation of the enterprise in some way, in order to detect changes in financial risk through the changes in the indicators in a timely manner.

(2) Reasonability. In the process of enterprise operation, many factors will trigger or deepen the financial risk of the enterprise, only to analyze the causes and influences of financial risk scientifically and reasonably, in order to find out the key factors affecting the financial risk, so as to determine the financial indicators applicable to financial risk prediction.

(3) Completeness. The causes of enterprise financial risk are complex and variable, and may be the result of the interaction of a variety of uncertain factors. The comprehensiveness of the indicator system can avoid the omission of important indicators.

(4) Accessibility and operability. The number of indicators to measure the financial situation is large and intricate, if the study of the company's undisclosed indicator data, then the cost of obtaining the data is too high or even impossible, the study will no longer be meaningful.

II. B. 2) Indicators affecting finance

(1) Solvency indicators

Solvency is closely related to an enterprise's financing risk. Enterprises that are able to make timely debt repayment when the debt is due have a strong solvency.

(2) Operating capacity indicators

Operating capacity determines the operating risk of the enterprise, and is mainly used to evaluate the efficiency level of the company in using the existing means of production in the process of daily operation. Generally speaking, the faster the speed of capital turnover, the higher the efficiency of its use. The efficiency of capital utilization determines the reproduction capacity and profitability of the enterprise.

(3) Profitability indicators

The profitability of the enterprise determines the size of the investment risk. Profitability, also known as profitability, is the ability of an enterprise to sell its products within a certain period of time to obtain income.

(4) Development capacity indicator

Development capacity is used to measure the future growth potential of an enterprise. Because with the development of science and technology, enterprises are faced with unpredictable social and market environment, how to adapt to different environments to maintain sustained strong growth is very critical for enterprises. Generally speaking, the total assets growth rate, fixed assets growth rate, gross operating income growth rate, net profit growth rate and so on in the enterprise financial indicators can be used to measure the enterprise's development capacity.

II. B. 3) Establishment of a system of preliminary financial indicators

According to the analysis of the indicators affecting the financial risk of enterprises in the previous section and the adjustments made according to the characteristics of transportation enterprises, in accordance with the principle of selecting financial indicators, under the condition of ensuring that the data of indicators are complete and accessible, this paper selects a total of 20 most representative preliminary financial indicators from five aspects, as shown in Table 1.

II. C.Factor analysis modeling

II. C. 1) Factor Analysis Algorithm Flow

Factor analysis is a statistical technique for extracting common factors from groups of variables for research. When performing calculations, factor analysis adopts a dimensionality reduction technique, based on the internal logic of the original variables with each other, transforming complex and numerous indicators into a number of mutually

independent representative common factors, which, for multivariate variables, is an effective statistical method for simplifying and analyzing high-dimensional data. After the dimensionality reduction process, factor analysis method can be evaluated and analyzed with more streamlined variables.

Table 1: Financial forecasting index system

Categories	Serial number	Index name
Profitability	X1	Net profit
	X2	Cost margin
	X3	Operating margin
	X4	Mobility ratio
Solvency	X5	Speed ratio
	X6	Asset ratio
	X7	Cash ratio
Development ability	X8	Revenue growth
	X9	Net equity growth rate
	X10	Total asset growth rate
Operational capacity	X11	Receivable turnover
	X12	Turnover of current assets
	X13	Total asset turnover

The mathematical model of factor analysis method is as follows:

$$\begin{aligned} X_1 &= a_{11}F_1 + a_{12}F_2 + \cdots + a_{1m}F_m + a_1\varepsilon_1 \\ X_2 &= a_{21}F_1 + a_{22}F_2 + \cdots + a_{2m}F_m + a_2\varepsilon_1 \\ &\vdots \\ X_p &= a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pm}F_m + a_p\varepsilon_1 \end{aligned} \tag{1}$$

where $X_1, X_2, ...X_p$ are the p original variables, which are standardized variables with mean 0 and standard deviation 1, $F_1, F_2, F_3, ...F_P$ are m factor variables, m is less than p, expressed in matrix form as:

$$X = AF + a\varepsilon \tag{2}$$

For the analysis of high-volume data indicators, the biggest advantage of the factor analysis method is that it can not only reduce the burden brought about by the overly large amount of data, but also ensure the reasonableness and objectivity of the analysis results, and also better explain the role of the extracted common factor on the research object. The application of factor analysis to performance evaluation is an effective method, which can fit the data from multiple perspectives and conduct comprehensive performance evaluation, and at the same time, with the help of factor analysis for weighting analysis, it can clarify the key to influence the comprehensive performance. However, when conducting factor analysis method, the initial data requirements are relatively high, and if the correlation between the evaluation indicators is insufficient or the sample size is not large, it may have an adverse effect on the accuracy and authenticity of the evaluation results.

The basic operational steps of the factor analysis method are as follows:

(1) Select the corresponding indicator variables in accordance with the object and goal of the study, and ensure that there is correlation between each variable.

(2) Identify the nature of each indicator by using autonomous judgment method or SPSS software, and positively treat the moderate and inverse indicator variables therein, and subsequently standardize the operation of these indicators.

(3) SPSS software was used to test the adaptability of the factor analysis method by selecting the KMO test and Bartlett's spherical test. When the value of KMO exceeds 0.5 and is close to 1, the fewer the statistical observations of the spherical test, the more its accuracy is less than 0.05 and tends to 0, which makes it more suitable for factor analysis.

(4) The common degree of each variable was analyzed using the principal component analysis method of SPSS software, thus identifying the factors of the male.

(5) To ensure that the common factors had real meaning, the component matrix was rotated with the help of Kaiser's standardized maximum difference method of SPSS software, and the number of common factors was calculated.

(6) Naming the extracted male factors and providing detailed descriptions of them.

(7) Calculate the score of each factor and compare it with the overall score to arrive at the ranking.

II. C. 2) Factor analysis applicability test

First of all, the applicability of the research data is analyzed, and then the factor analysis method is applied to refine the information, that is, the applicability of the selected indicators and data is first evaluated and judged by KMO and Bartlett's spherical test as shown in Table 2. As can be seen from the table, the KMO value is 0.713, which is greater than 0.6, indicating that there is a correlation between these indicators. From the SPSS test results, the significance of Bartlett's test of sphericity is 0, which indicates that there is no assumption of independence and there is a strong correlation between the variables. Therefore, the data is well suited for the use of factor analysis.

Table 2: KMO and bartlett sphericity test

KMO sampling availability number		0.713
Bartlett sphericity test	Approximate card	732.425
	Freedom	81
	Significance	0.000

Under the premise that the selected variables meet the requirements of the test, the public factor can be extracted by using factor analysis, using SPSS26.0 software to analyze the effect of extracting the main factor on the original variables, and the variance of the public factor is shown in Table 3. The closer the extracted value is to 1.000, the stronger the interpretability is. Through Table 3, it can be learned that in the 13 financial indicators, the amount of information extracted is more than 70%, indicating that the degree of deficiency of the information of all the original variables is very small, and all of them can be interpreted on the original variables.

Table 3: Common factor variance

Variable	Initial	Extraction
Net profit	1.000	0.987
Cost margin	1.000	0.972
Operating margin	1.000	0.992
Mobility ratio	1.000	0.923
Speed ratio	1.000	0.947
Asset ratio	1.000	0.819
Cash ratio	1.000	0.895
Revenue growth	1.000	0.749
Net equity growth rate	1.000	0.709
Total asset growth rate	1.000	0.869
Receivable turnover	1.000	0.699
Turnover of current assets	1.000	0.878
Total asset turnover	1.000	0.728

II. C. 3) Extracting the common factor

The total variance explained is shown in Table 4. There is a significant difference in the contribution of different principal components to the total variance of the dataset, which directly affects their ability in explaining the data. As can be seen from the table, the contribution of variance of the first four factors reaches 85.42%, indicating that the first four factors extracted contain most of the financial information of the original data.

In addition, the gravel plot provides an intuitive way to assess the relative importance of different factors, which in turn helps authors decide how many common factors should be retained. The fact that the decline in eigenvalues flattens out after a certain point indicates that factors beyond that point may no longer provide meaningful information. The gravel plot is shown in Figure 1. When the number of factors on the horizontal coordinate corresponding to the transition point of the gravel plot is 4, it indicates that among all the possible factors, the first four factors are the most critical, which together constitute the core of the data structure and can effectively reflect the main features and information of the data. Therefore, it is more appropriate to select 4 male factors.

Table 4: Total variance interpretation

Constituent	Initial eigenvalue			Extracting the load of the load			Rotational load squared		
	Total	Percentage of variance	Cumulation %	Total	Percentage of variance	Cumulation %	Total	Percentage of variance	Cumulation %
1	5.005	38.5	38.5	5.005	38.5	38.5	3.713	28.562	28.562
2	2.466	18.969	57.469	2.466	18.969	57.469	3.178	24.446	53.008
3	2.12	16.308	73.777	2.12	16.308	73.777	2.44	19.308	72.315
4	1.467	11.285	85.062	1.467	11.285	85.062	1.657	12.746	85.062
5	0.584	0.04492	89.554						
6	0.489	0.03762	93.315						
7	0.265	0.02038	95.354						
8	0.236	0.01815	97.169						
9	0.126	0.00969	98.138						
10	0.135	0.01038	99.177						
11	0.087	0.00669	99.846						
12	0.011	8.46154E-4	99.931						
13	0.009	6.92308E-4	100						

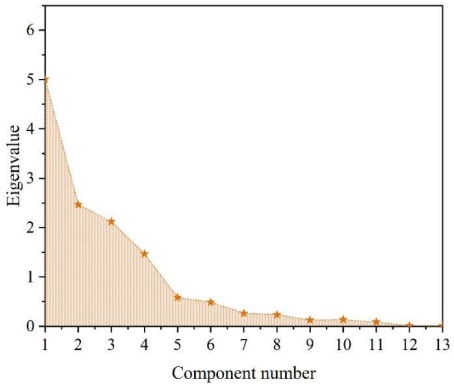


Figure 1: Rubble map

II. C. 4) Naming the common factor

By analyzing the selected common factors, the information embodied in each of them can be seen very directly, and the variables contained in each of them can be used to quickly analyze the problem and make the calculation results better reflect the actual situation. Due to the ambiguous meaning of the factor loadings, the use of the component matrix to name the male factors is not applicable, which requires the rotation of the loading matrix by the maximum variance method to better understand the factors. The factor loading matrix effectively reveals the strength of association between the original variables and each of the common factors. If an initial variable has a loading value close to 1 or -1 on a factor, this indicates that the variable is highly correlated with that factor, and it can almost be said that the variability of the variable is largely explained by that factor. On the contrary, if the loading value is close to 0, this means that the variable is minimally related to this factor, and the factor has almost no explanatory power for the variability of this variable the rotated component matrix is shown in Table 5. As can be seen from the table:

The male factor 1 has higher loadings on the indicators current ratio X4, quick ratio X5, gearing ratio X6 and cash ratio X7. These indicators themselves are important tools to measure the solvency of enterprises, so the male factor 1 is named as solvency factor F1.

Common factor 2 has a higher loading on net sales margin X1, cost and expense margin X2, and operating profit margin X3. These indicators reflect the profitability of the enterprise, so the public factor 2 is named as profitability factor F2. public factor 3 has larger loadings on accounts receivable turnover X11, total asset turnover X12, current asset turnover X13, which reflect the operating efficiency of the enterprise, so the factor 3 is named as operating capacity factor F3.

The indicators with larger loadings on the male factor 4 are operating income growth rate, net assets growth rate, total assets growth rate, these indicators reflect the development ability of the enterprise, so factor 4 is named development ability factor.

Table 5: The component matrix after rotation

Variable	F1	F2	F3	F4
X1	0.183	0.986	-0.037	-0.001
X2	0.163	0.975	-0.117	0.048
X3	0.18	0.989	-0.051	0.013
X4	0.918	0.23	-0.215	-0.017
X5	0.959	0.162	-0.122	0.017
X6	0.84	0.271	-0.104	-0.23
X7	0.948	0.024	0.036	0.032
X8	-0.395	-0.273	0.414	0.548
X9	0.33	0.312	0.225	0.659
X10	-0.143	0.012	-0.126	0.896
X11	-0.083	0.127	0.805	-0.083
X12	-0.182	-0.08	0.896	0
X13	0.056	-0.167	0.787	0.209

II. C. 5) Factor scores

After extracting the four public factors, the paper uses regression analysis to calculate the component score coefficient matrix, which is shown in Table 6. The score function of each public factor can be derived from the table. The score function of each public factor can be derived from the table:

$$F_1 = -0.076*X1 - 0.057*X2 - 0.05*X3 + 0.267*X4 + 0.294*X5 + 0.236*X6 + 0.316*X7 - 0.041*X8 + 0.114*X9 - 0.014*X10 + 0.014*X11 + 0.014*X12 + 0.101*X13 \quad (3)$$

$$F_2 = 0.343*X1 + 0.337*X2 + 0.343*X3 - 0.028*X4 - 0.055*X5 + 0.008*X6 - 0.099*X7 - 0.048*X8 + 0.068*X9 - 0.006*X10 + 0.095*X11 + 0.033*X12 - 0.039*X13 \quad (4)$$

$$F_3 = 0.023*X1 - 0.021*X2 + 0.021*X3 - 0.029*X4 + 0.009*X5 + 0.032*X6 + 0.012*X7 + 0.095*X8 + 0.062*X9 - 0.159*X10 + 0.358*X11 + 0.378*X12 + 0.315*X13 \quad (5)$$

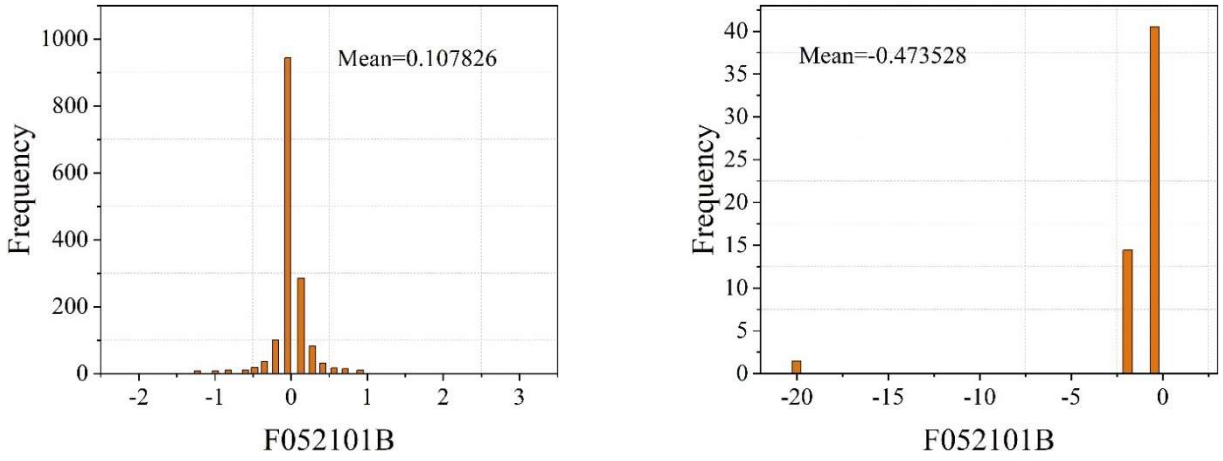
$$F_4 = -0.038*X1 + 0.019*X2 - 0.009*X3 + 0.042*X4 + 0.053*X5 - 0.121*X6 + 0.049*X7 + 0.293*X8 + 0.387*X9 + 0.563*X10 - 0.134*X11 - 0.092*X12 + 0.058*X13 \quad (6)$$

Table 6: Component score coefficient matrix a

Variable	F1	F2	F3	F4
X1	-0.049	0.343	0.023	-0.038
X2	-0.057	0.337	-0.021	0.019
X3	-0.05	0.343	0.021	-0.009
X4	0.267	-0.028	-0.029	0.042
X5	0.294	-0.055	0.009	0.053
X6	0.236	0.008	0.032	-0.121
X7	0.316	-0.099	0.012	0.049
X8	-0.041	-0.048	0.095	0.293
X9	0.114	0.068	0.062	0.387
X10	-0.014	-0.006	-0.159	0.563
X11	0.014	0.095	0.358	-0.134
X12	0.014	0.033	0.378	-0.092
X13	0.101	-0.039	0.315	0.058

II. C. 6) Factors influencing enterprise financial indicators

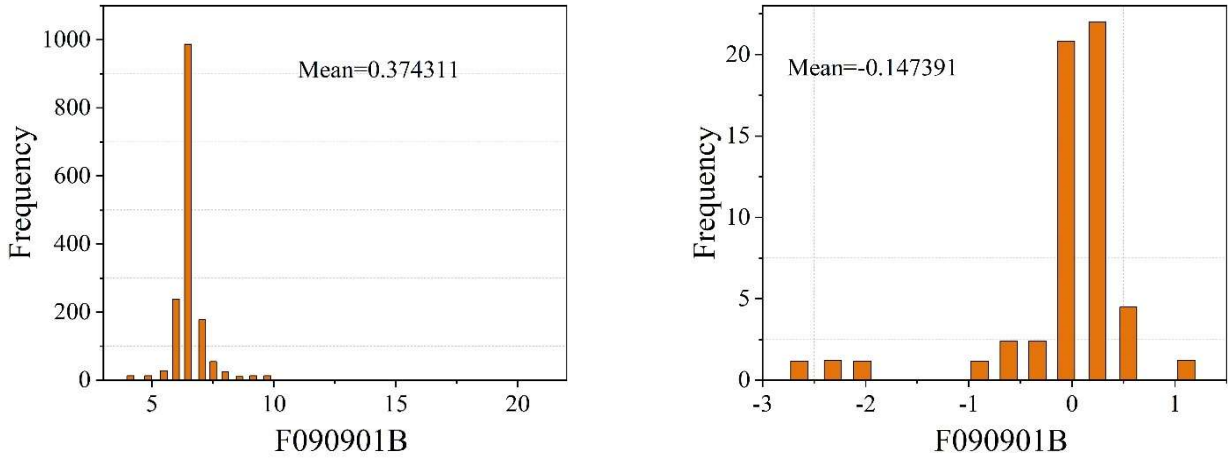
Figure 2 shows the comparison of cost and expense margins in corporate financial forecasts and non-corporate financial forecasts, from which it can be seen that the mean value of cost and expense margins in corporate financial forecasts is -0.4735, while that in non-corporate financial forecasts is 0.1078, which indicates that the cost and expense margins attributed to the profitability of the enterprise are positive and the closer they are to 1, which means that the enterprise is very likely to belong to non-corporate financial forecasts.



(a) Non-financial forecast (b) Corporate financial forecast

Figure 2 Enterprise cost margin

Figure 3 shows the comparison of operating profit margin per share, from the figure we get the operating profit margin per share of the enterprise company is -0.1442, while the non-enterprise financial forecast of operating profit per share is 0.3564. It shows that in the case of per-share indicators among the operating profit per share indicator of the value of the larger and positive, the company's financial situation is better, the more likely that the company will not be in financial difficulties.



(a) Non-financial forecast (b) Corporate financial forecast

Figure 3: Business margins per share

III. Evaluation model based on SMOTE-Light GBM algorithm

III. A. Construction of SMOTE-Light GBM Algorithm

III. A. 1) SMOTE algorithm

The SMOTE oversampling technique synthesizes new minority class samples through a nearest neighbor algorithm, which is particularly suitable for unbalanced datasets in the evaluation of financial forecasting models for small and medium-sized enterprises [27]. The technique first calculates the Euclidean distance between each sample and other samples in the minority class dataset to determine the k nearest neighbors; second, the upsampling

multiplier N is chosen, and N samples are randomly selected from the nearest neighbors, and randomly and linearly combined with the original samples, to generate the new samples $x^{'}(i,j)$, and the following is the related calculation method:

$$x^{'}(i,j) = x_i + rand \cdot (x_j - x_i) \tag{7}$$

where i denotes the original sample index (1 to M), j denotes the randomly selected neighboring sample index (1 to N), and rand is a random number between 0 and 1. This method fuses the new samples with the existing minority class samples to create a balanced dataset.

III. A. 2) LightGBM Algorithm

LightGBM is an efficient gradient boosting framework designed and open-sourced by Microsoft, similar to the GBDT technique, which uses the negative gradient of the loss function as the residual to construct a new decision tree and optimize the model prediction to be closer to the actual value. The specific implementation steps are:

In the first step, the goal of each iteration is to construct an auxiliary model that aims to minimize the iterative loss function $L[y, F_t(x)]$, with the formula:

$$L[y, F_t(x)] = L[y, F_{t-1}(x) + h_{t(x)}] \tag{8}$$

In this process, $F_{t-1}(x)$ and $L[y, F_{t-1}(x)]$ represent the model and loss function obtained from the previous iteration, respectively.

In the second step, the negative gradient from the previous step is utilized to fit the loss function approximation of the current iteration with the expression:

$$i_{ij} = \frac{\partial F[y_i, F_{t-1}(x_i)]}{\partial F_{t-1}(x_i)} \tag{9}$$

In the third step, the optimal segmentation point is found by minimizing the squared error:

$$h_i(x) = \arg\min_{h \in H} \sum [r_{ij} - h(x)]^2 \tag{10}$$

In the fourth step, the final model of the current iteration is updated to:

$$F_t(x) = F_{t-1}(x) + h_t(x) \tag{11}$$

LightGBM is an efficient gradient enhancement technique that improves the accuracy and robustness of the model by preventing overfitting through histogram optimization dealing with continuous feature segmentation, leaf node growth methods and limiting the maximum depth of the tree as compared to the standard gradient boosted tree.LightGBM also improves on feature processing and parallel processing with faster computation speeds and higher accuracies. Therefore, this study chose to apply it to the classification task.

III. B. Evaluation indicators

Accuracy represents the ratio of the sum of the number of non-firm financial forecasts judged by the model to be non-forecasts and the number of firm financial forecasts judged to be in default to the total number of firms:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

Recall Ratio:

The first type of error is the ratio of misclassifying non-firm financial forecasts as samples of firm financial forecasts:

$$TypeI - error = \frac{FN}{TP + FN} \tag{14}$$

The second type of error is the ratio of misclassifying a sample of corporate financial forecasts as a sample of non-corporate financial forecasts:

$$TypeII-error=\frac{FP}{TN+FP} \tag{15}$$

The F1 value is the reconciled average of precision and recall, which balances the model's checking accuracy and checking completeness, and is an important indicator for evaluating the model's performance.The AUC, as the area under the ROC curve, can make up for the deficiency of the ROC curve, and the larger the AUC is, the stronger the model's performance in identifying the state of credit default is.

III. C. Analysis of empirical results of the SMOTE-LightGBM model

III. C. 1) Proportional division of training and test sets

In order to experiment the most suitable training set and test set ratio division for the data set of this paper, this paper first carries out three sets of different ratio divisions for the data set, and then experiments its experimental results under the SMOTE-Light GBM algorithm, so as to judge which kind of ratio division for the training set and the test set is most suitable for the empirical analysis of this paper, as shown in Table 7. The proportional divisions for the data sets in the table are as follows: using 70% of the data as the training set and 30% of the data as the test set; using 75% of the data as the training set and 25% of the data as the test set; and using 80% of the data as the training set and 20% of the data as the test set. For Python code implementation, the random number seed selected in this paper is 1326.

As can be seen from the table, in terms of the evaluation metrics such as accuracy, precision, recall, f1_score and AUC value, using 75% of the data as the training set and 25% of the data as the test set, compared to using 70% of the data as the training set and 30% of the data as the test set, performs better in terms of accuracy, precision, recall and f1_score evaluation metrics and is AUC value is the same; compared with using 80% of the data as the training set and 20% of the data as the test set, the performance in accuracy, precision and f1_score evaluation indexes are better, and the performance in recall and AUC value is worse but not much different, so overall, the experimental effect of using 75% of the data as the training set and 25% of the data as the test set for the model in this paper is is better than using 80% of the data as the training set and 20% of the data as the test set. Overall, for this paper's dataset and the application of the constructed model, using 75% of the data as the training set and 25% of the data as the test set in the three sets of training set and test set ratio division of the evaluation results of each model is seen to be the best.

Table 7: The results of the data set in different training sets and test set proportions

	Accuracy rate	Accuracy rate	Recall rate	F1-score	AUC
70% training set, 30% test set	0.9781	0.9583	0.9747	0.9792	0.9808
75% training set, 25% test set	0.9734	0.9629	0.9712	0.9786	0.9808
80% training set and 20% test set	0.9746	0.9594	0.9821	0.9782	0.9811

III. C. 2) Analysis of the results of the SMOTE-LightGBM model application

This paper establishes a financial prediction assessment model for ascending enterprises based on SMOTE-Light GBM algorithm for empirical research, after re-establishing the financial prediction index system of ascending enterprises after feature screening, data preprocessing such as missing value processing, min-max standardization, SMOTE oversampling processing (random number seed set to 0) are carried out on the data set, and 75% of the data are used as the training set and 25% of the data as a test set (random number seed set to 1234), in the training and prediction of SMOTE-LightGBM model, the parameter optimization of LightGBM model is carried out, and the LightGBM classifier is constructed by grid searching through cross-validation for 5 times (cv=5), and the input data is searched for suitable hyperparameters, and the optimal value of the parameter is obtained as: Learning rate learning_rate=0.2, number of iterations for boosting n_estimators=50, number of leaf nodes in a tree num_leaves=25. After the implementation of the Python program for the classification model, the results of the evaluation experiments such as the confusion matrix, the evaluation report of the classification model and the ROC curve were obtained, as shown in Table 8.

From the table, it can be seen that: the dataset of this paper has a classification accuracy of 0.9609, precision of 0.9521, recall of 0.9885, and f1_score of 0.9634 after training and prediction of SMOTE-Light GBM model. The model has high accuracy in terms of these four metrics, with a high ability to correctly predict categorized samples. The macro avg in the table denotes a macro average, which is an arithmetic average of each type, i.e., the evaluation indicators are summed and averaged. weighted avg denotes a weighted average, which is calculated by using the proportion of each category's sample size to the total number of samples in all categories as a weight to calculate the weighted average of each evaluation indicator.

Table 8: The smote-lightgbm model evaluation report

	Precision	Recall	F1-score	Support
0	0.9879	0.9518	0.9618	310
1	0.9521	0.9885	0.9634	315
Accuracy			0.9629	618
Macro avg	0.9617	0.9603	0.9622	618
Weighted avg	0.9611	0.9609	0.9622	618

The ROC curve obtained from the financial prediction evaluation model of listed companies based on the SMOTE-LightGBM algorithm is shown in Fig. 4.The ROC curve is relatively steep, indicating that the number of positive samples after the model classification and prediction is high. The area enclosed by the ROC curve, i.e., the AUC value, is 0.9879, which is much larger than 0.8, and the AUC score of the ROC curve is used as the model evaluation criterion, indicating that the LightGBM model is well suited for the application of the dataset in this paper, and the prediction effect is very good.

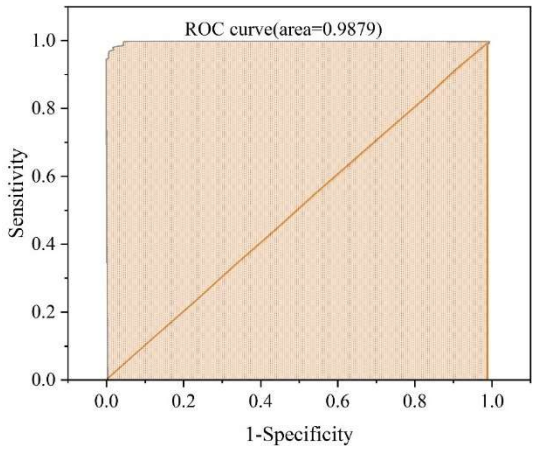


Figure 4: The smote-lightgbm model roc curve

III. C. 3) Evaluation of the importance of indicators for assessing models

The degree value of the importance of each evaluation indicator is obtained according to the Python program, and the contribution degree of each indicator in the classification model is analyzed.The indicator importance ranking diagram of the SMOTE-LightGBM model is shown in Figure 5. According to the quick ratio ($X5$), net sales margin ($X1$), gearing ratio ($X6$), cost-expense margin ($X2$), quick ratio ($X5$), net assets growth rate ($X9$), and current ratio ($X4$), the indicators ranked at the bottom of the ranking do not have much influence on the financial forecast of listed companies, while the operating profit margin ($X3$), Total Assets Turnover Ratio ($X13$), Total Assets Growth Rate ($X10$), Operating Income Growth Rate ($X8$), Accounts Receivable Turnover Ratio ($X11$), which are the top-ranked indicators, have a greater impact on the classification of financial forecasts of listed companies.

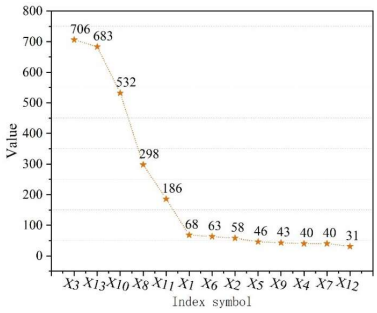


Figure 5: The smote-lightgbm model is a sort diagram of the importance of the measure

IV. Application of the economic landscape in the new era

(1) The establishment of an industry-specific “accounting management report” system helps to enhance the unity of accounting management over production and business operations, thereby realizing the role of accounting management activities in constraining other production and business functions at the institutional level, which also helps the transition of accounting management from the microeconomic level to the macroeconomic level.

(2) Monthly, semi-annual or continuous disclosure of “accounting management reports” can reduce surplus manipulation by improving the timeliness and intensity of information disclosure. Moreover, the economic information of the industry is often directly related to various complex businesses, which makes it easier to see through the lies and more difficult to make them up while highlighting the characteristics of the business.

(3) The “accounting management report” and the “general accounting report” are emphasized together to form a multi-level financial reporting system. By providing rich qualitative data on the operating performance, output scale, management efficiency, etc. of different enterprises in the same industry, it can reflect the real operating ability of enterprises and help statement users make horizontal comparisons and discover information that affects the development of enterprises and enhances their intrinsic value. Moreover, improving the transparency of accounting information also helps market regulation and market valuation.

V. Conclusion

The study uses factor analysis to identify key corporate financial predictors by extracting the common factors, and applies SMOTE-Light GBM to evaluate the financial predictions of listed companies. Through the experimental results of SMOTE-Light GBM model, it is found that smote oversampling treatment is more suitable to deal with the unbalanced data in this paper than not dealing with unbalanced data and undersampling treatment. And in the case of random number seed are set to 1326, through the experimental results analysis and comparison, the most suitable training set test set division of this paper's dataset is selected, using 75% of the data as the training set and 25% of the data as the test set, in which the accuracy rate, precision rate, recall rate, fil score and are 0.9609, 0.9521, 0.9885, respectively, 0.9634, which shows that the SMOTE-Light GBM model performs well in all evaluation indexes and has good classification prediction effect.

References

[1] Chen, L. H., & Guo, T. Y. (2011). Forecasting financial crises for an enterprise by using the Grey Markov forecasting model. Quality & Quantity, 45(4), 911-922.

[2] Ma, Z., Wang, X., & Hao, Y. (2023). Development and application of a hybrid forecasting framework based on improved extreme learning machine for enterprise financing risk. Expert Systems with Applications, 215, 119373.

[3] Lan, X. (2021, October). Intelligent forecast model of enterprise financial risk from the perspective of budget constraint. In 2021 3rd International Conference on Artificial Intelligence and Advanced Manufacture (pp. 1922-1925).

[4] Qin, W. (2022). Research on financial risk forecast model of listed companies based on convolutional neural network. Scientific Programming, 2022(1), 3652931.

[5] Tan, L., Zhang, W., & Liu, B. (2013). Forecast Management Based on Enterprise Financial Accounting Report. Management & Engineering, (13), 60.

[6] Tao, Y. (2022, February). Financial Risk Forecast Model of Small and Medium-Sized Enterprises Based on Neural Network Algorithm. In The International Conference on Cyber Security Intelligence and Analytics (pp. 905-912). Cham: Springer International Publishing.

[7] Yousaf, U. B., Jebran, K., & Wang, M. (2022). A comparison of static, dynamic and machine learning models in predicting the financial distress of chinese firms. Romanian Journal of Economic Forecasting, 25(1), 122.

[8] Bousbaa, Z., Sanchez-Medina, J., & Bencharef, O. (2023). Financial time series forecasting: a data stream mining-based system. Electronics, 12(9), 2039.

[9] Jansen, W. J., Jin, X., & de Winter, J. M. (2016). Forecasting and nowcasting real GDP: Comparing statistical models and subjective forecasts. International Journal of Forecasting, 32(2), 411-436.

[10] Korol, T. (2018). The implementation of fuzzy logic in forecasting financial ratios. Contemporary Economics, 12(2), 165-187.

[11] Smith, W. K. (2014). Dynamic decision making: A model of senior leaders managing strategic paradoxes. Academy of management Journal, 57(6), 1592-1623.

[12] Iastremska, O., Tryfonova, O., Mantaliuk, O., & Baranets, H. (2023). The impact of strategic decisions on the future development of organisations and economic dynamics. Futurity Economics&Law, 3(4), 117-133.

[13] Chen, S., Goo, Y. J. J., & Shen, Z. D. (2014). A hybrid approach of stepwise regression, logistic regression, support vector machine, and decision tree for forecasting fraudulent financial statements. The Scientific World Journal, 2014(1), 968712.

[14] Okasha, M. K. (2014). Using support vector machines in financial time series forecasting. International Journal of Statistics and Applications, 4(1), 28-39.

[15] Pietukhov, R., Ahtamad, M., Faraji-Niri, M., & El-Said, T. (2023). A hybrid forecasting model with logistic regression and neural networks for improving key performance indicators in supply chains. Supply Chain Analytics, 4, 100041.

[16] Hassan, E., BIRAU, R., AWAIS-E-YAZDAN, M. U. H. A. M. M. A. D., & Paliu-Popa, L. (2023). Anticipating financial distress in monster sectors of Pakistan’s economy: an application of logit. Industria Textila, 74(3), 363-370.

[17] Mishra, N., Ashok, S., & Tandon, D. (2024). Predicting financial distress in the Indian banking sector: A comparative study between the Logistic Regression, LDA and ANN models. Global Business Review, 25(6), 1540-1558.

[18] Ali, M. M., Babar, M. I., Hamza, M., Jehanzeb, M., Habib, S., & Khan, M. S. (2019). Industrial financial forecasting using long short-term memory recurrent neural networks. International Journal of Advanced Computer Science and Applications, 10(4), 88-99.
[19] Chen, M. Y. (2014). Using a hybrid evolution approach to forecast financial failures for Taiwan-listed companies. Quantitative Finance, 14(6), 1047-1058.
[20] Senoguchi, J. (2021). Forecast of complex financial big data using model tree optimized by bilevel evolution strategy. Journal of Big Data, 8(1), 116.
[21] Wang, Z. (2022). A study on early warning of financial indicators of listed companies based on random forest. Discrete Dynamics in Nature and Society, 2022(1), 1314798.
[22] Türegün, N. (2019). Text mining in financial information. Current analysis on economics & finance, 1(647), 18-26.
[23] Ashtiani, M. N., & Raahemi, B. (2023). News-based intelligent prediction of financial markets using text mining and machine learning: A systematic literature review. Expert Systems with Applications, 217, 119509.
[24] Yang, R., Yu, Y., Liu, M., & Wu, K. (2018). Corporate risk disclosure and audit fee: A text mining approach. European Accounting Review, 27(3), 583-594.
[25] Kanungsukkasem, N., & Leelanupab, T. (2019). Financial latent Dirichlet allocation (FinLDA): Feature extraction in text and data mining for financial time series prediction. IEEE Access, 7, 71645-71664.
[26] García-Méndez, S., de Arriba-Pérez, F., Barros-Vila, A., González-Castaño, F. J., & Costa-Montenegro, E. (2023). Automatic detection of relevant information, predictions and forecasts in financial news through topic modelling with Latent Dirichlet Allocation. Applied Intelligence, 53(16), 19610-19628.
[27] Jia chuang Wang & Long jun Dong. (2024). Risk assessment of rockburst using SMOTE oversampling and integration algorithms under GBDT framework. Journal of Central South University,31(8),2891-2915.