ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

Innovative applications of big data in analyzing the financial situation of enterprises

Tongyu Xing[1,*], *Sisi Zhao*[2] *and Zhidan Zhu*[1]

[1] College of Finance and Economics, Shanxi College of Applied Science and Technology Taiyuan, Shanxi, 030062, China
[2] School of Economic and Management, Qinghai Minzu University, Xining, Qinghai, 810007, China

SUMMARY: *The paper initially uses cluster analysis to normalize and discover relationships between multi-source financial indicators, removing unnecessary data and choosing representative metrics. The extraction of common factors is performed through factor analysis, which reduces high dimensional indicators to five major dimensions. A more efficient ID3 decision tree algorithm and the Prophet time-series prediction model are subsequently incorporated to develop the financial condition early warning system. Empirical analysis relies on 16 essential financial indicators of a listed manufacturing company that existed between 2020 and 2024. The factor analysis model is 81.58percent cumulative variance contributor, and it has successfully derived five factors, which have clear economic meaning. The modified ID3 decision tree still has a general classification rate of at least 87 percent when given 50 percent category noise, and is significantly better than the traditional models. The overall score of the company in the financial field increased by 78.45 points or 6.42 points per year (1.68 per quarter) in the period of 2020-2024, which suggests the improvement of its financial performance throughout the years.*

KEYWORDS: *Financial condition analysis; Cluster analysis; Factor analysis; ID3 decision tree algorithm; Prophet model*

1 Introduction

Financial analysis is an important part of the management of corporations and cannot be overestimated [1]. It offers a data-based approach to making business decisions and is at the heart of managing capital and mitigating risks [2]. The emergence of big data opens new possibilities to the corporate finance analysis. By incorporating large volumes of both internal and external data and allowing real time processing and analysis, it renders the process of financial analysis more timely, comprehensive and detailed and changes and optimizes the existing situation and prospects of the future of corporate financial analysis [3, 4].

Big data has some particular features. The size of the data is extremely high, e.g., the amount of data produced by the world-wide web on a daily basis is increasing exponentially[5, 6]. To begin with, it is made up of various forms of data, including structured database data, semi-structured log data, and unstructured data like text, images, and videos. Second, it requires extremely high-speed processing; financial transaction data, to give another illustration, can be analyzed and processed in real-time. Third, it is of low value density, meaning that one needs

*13643474951@126.com
https://doi.org/10.65102/is2026403
Published: 30 April, 2026

to derive valuable insights out of large volumes of information. Fourth, it is produced by a variety of sources: internal business systems contain transaction and inventory data, social media regularly produces user reviews and shares, whereas the IoT devices keep sending operation and environment monitoring data.

Financial analysis is an organized and systematic approach. It analyzes the structure and size of assets, liabilities and equity owned by the owner to evaluate the financial wellbeing of a company [7]. It also breaks down the revenue, cost and profit compositions in operating results and observes the cash flow of operating, investing and financing activities to benefit the corporate stakeholders[8]. When making decisions, investment projects have to be analyzed through the expected returns and payback periods, whereas when choosing the sources of financing and their magnitude it is necessary to balance the costs and risks[9]. The measurement of capital utilization rates are used to improve efficiency in capital management where resources are tailored to the requirements of each department[10]. To control risks, the debt to equity ratio may act as an early warning signal of debt risks and profitability measures including gross profit margin are used to evaluate the quality of earnings that can help in actively preventing financial crises[11].

The use of big data has transformed the financial analysis as the combination of financial and non-financial data is currently possible. The market trend data provides businesses with an opportunity to take advantage of opportunities, the industry dynamics inform on the competitive landscape, and the customer feedback informs on how to improve the products and services[12]. In terms of process, big data utilizes sensors and network technologies to collect data in real time. It greatly reduces the cycle due to high-speed transmission and distributed computing to provide immediate analysis[13]. Due to the variety of analytical instruments, such as data mining software that uncovers the relationships among data points and visualization tools that can present complex information in graphs, results of financial analysis are increasingly detailed, accurate, and timely which strongly supports the decision-making of corporate management[14, 15].

The role of risk management in corporate accounting has gained a lot more attention. Successful risk management will allow companies to determine, measure and manage the different types of risk that could affect their accounting operations, which ensures the reliability and validity of financial information and thus contributes to the preservation of corporate financial soundness and stability of operations[16, 17]. In terms of financial risk evaluation in business processes, Jin et al. [18] used the advantages of data mining technology. Through the combination of internal and external factors influencing corporate financial risks and adding them to a time-series dynamic maintenance-based financial risk early warning model, they obtained real time warnings to corporate financial risks. Cao et al.[19] studied the practical application of big data analytics in corporate financial analysis, cutting down on the different financial audit areas of applicability. They found out that big data analytics is beneficial concerning the effectiveness of financial statement auditing. Saleh et al. [20] noted that big data is useful in enhancing the quality of corporate accounting statements and expert evaluations, and it helps build a strong financial position by customizing products, streamlining procedures, and improving the evaluation of risks, which ultimately increases the power of risk management.Chen et al. [21] employed artificial neural networks and data mining techniques to construct a financial condition prediction model. Through satisfaction-based empirical experiments, they demonstrated the model's feasibility, revealing that prediction accuracy increases as the time to actual financial risk occurrence decreases. Wang et al. [22] compared the financial risk early warning performance of three data mining techniques using 77 enterprises labeled with financial risks from 2005 to 2007. All three methods demonstrated

good short-term early warning effects, while neural networks and decision trees outperformed the logistic regression model in long-term early warning. Gao [23] employed data mining techniques to detect corporate financial risks, achieving an average detection accuracy of 90.27%. The study also identified that financial risk evaluation indicators across four dimensions—solvency, operational capability, profitability, growth potential, and cash flow capacity—influence a company's financial risk profile. Koyuncugil et al. [24] proposed a financial risk warning system based on data mining techniques. This system identified 31 risk vulnerabilities, 15 risk indicators, 2 warning signals, and 4 financial roadmaps for the Central Bank of Turkey, assisting the bank in assessing its financial condition.

Both the financial position and operational results of companies are carefully watched by various interested parties, namely, investors and creditors. In case a company is experiencing financial distress, it may cause both direct and indirect losses to many other stakeholders [25, 26]. Following this, Sun et al. [27] suggested a data mining technique that combines attribute-oriented induction, information gain, and decision trees to create a predictive model of corporate financial distress. This was empirically shown to be able to extract the most valuable information out of various databases. The study by Geng et al.[28] on 107 Chinese firms with special markers using data mining methods to create an early warning system of corporate financial distress. Through the examination of 31 financial indicators and three different time windows, they forecasted the negative influence of financial measures like net profit margin on total assets, return on total assets, and earnings per share on the worsening of profitability.

The financial position of a company is a reflection of both its profitability and risk level, which forms the foundation of an investor assessing value and making investment choices. That is why, stock prices play a crucial role in evaluating the financial soundness of corporations [29].. Artificial neural networks and data mining are the methods used by Xu et al. [30] to learn stock market trends. The method has shown some positive attributes in terms of small-scale networks, high rate of learning, and good prospects of financial analysis when tested, and it helped in the extraction and prediction of hidden information on corporate financial conditions. Kannan et al. [31] examined the trends in the stock prices with the help of big data and predictive methods, presenting the trends in the form of numbers and charts. This can lead companies to discover underlying trends in financial health based on historical data.

Cluster analysis is used in this paper to standardize raw financial indicators and mine correlations, removing dimensional differences. Factor analysis identifies common factors, which compresses financial measures into five fundamental components of profitability, solvency, operational efficiency, growth opportunity, and cash flow capacity. The solution to the problem of risk weighting is to mitigate the influence of too many features on the bias of the improved ID3 decision tree algorithm. It was combined with the Prophet time series model to create a dynamic early warning system which would detect outliers and changes in trends. Using a listed manufacturing company as a research subject, a factor analysis model was developed. Comparative experiments proved the effectiveness of the suggested improvement strategy. The appropriate ratios and indicators were investigated in order to reveal the main factors that impact corporate financial dynamics. To evaluate the integrated enhanced Wolters score technique, a simulation analysis system was applied.

2 Design of a Big Data-Driven Corporate Financial Condition Analysis Model

Conventional corporate financial analysis is based on structured financial statements data and evaluation of financial health using manual experience or elementary statistical approaches,

which have serious drawbacks. With businesses speeding up their digitalization, information produced by their activities has skyrocketed. This heterogeneous and multi-source data has deep financial reasoning that the conventional tools fail to reveal, which creates new opportunities as well as challenges to corporate financial analysis.

The core value of big data technology lies in revealing underlying patterns and latent rules through the collection, cleansing, modeling, and analysis of massive datasets. In the corporate finance domain, its innovative applications manifest across three dimensions: First, addressing redundant financial metrics and dimensionality issues through data preprocessing techniques. Second, leveraging machine learning algorithms to uncover nonlinear relationships among financial indicators. Third, utilizing real-time data streams and early-warning algorithms to enable dynamic monitoring and proactive intervention of financial risks. The integrated application of these technologies not only enhances analytical precision and efficiency but also empowers enterprises to identify potential crises in advance.

2.1 Preprocessing of Financial Data Based on Cluster Analysis

Data preprocessing is a crucial step in the data mining process. By preprocessing noisy, impure, and redundant data within the dataset, the quality of the data mining objects can be enhanced, ultimately improving the quality of the data mining results.

Many financial indicators exhibit significant interdependencies. Without further processing and dimensionality reduction, these indicators not only increase the complexity and time required for mining but also compromise the accuracy of results. Moreover, an excessive number of financial indicators can obscure the information provided to investors. Therefore, preprocessing these indicator data is essential.

(1) Data Standardization

This paper employs cluster analysis for data preprocessing, which identifies valuable relationships among data points within large datasets. To eliminate the impact of differing dimensions in the original data, standardization is required. The standardization formula is:

$$x_{ij} = \frac{y_{ij} - \overline{y}_j}{\sqrt{\mathrm{var}\left(y_j\right)}} (i = 1, 2, \cdots, P) \tag{1}$$

Moreover: $\overline{y}_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij}$, $sqrt\ \mathrm{var}\left(y_j\right) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_{ij} - y_j\right)^2}$

where $\overline{y}$ and $\sqrt{\mathrm{var}\left(y_j\right)}$ are the mean and standard deviation of the j th variable, respectively. y_{ij} is the value of the indicator before normalization and x_{ij} is the value of the indicator after normalization.

(2) Defining Distance

Given n indicators and p observations, with each indicator having p observations. Let the j th observation of the i th indicator be denoted as x_{ij}. Treating the n indicators as n points in a p-dimensional space, the closeness between two indicators can be measured by the distance between these two points in the p-dimensional space. Let d_{ij} denote the distance between indicators x_i and $\mathrm{x_j}$. The distance formula is defined as follows, employing the Euclidean distance in this paper.

$$d_{ij} = \sqrt{\sum_{k=1}^{n} \left| x_{ik} - x_{jk} \right|^2} \tag{2}$$

In this paper, a shorter distance indicates that the properties of two indicators are closer and can be grouped into the same category.

When performing cluster analysis in SPSS software, click “Classification” under the ‘Analyze’ module, then select “Systematically Classify.” Choose the appropriate clustering method and standardization method based on the clustering requirements. Under “Plot,” select to plot a dendrogram. The software will automatically perform the cluster analysis and output the results. The clinician then classifies the data based on the output.

Following the cluster analysis, more representative financial diagnostic indicators can be obtained, providing clearer analytical foundations for financial diagnosis.

2.2 Financial Condition Diagnosis Based on Factor Analysis

Despite the fact that preliminary processing of the data was carried out with the help of cluster analysis to choose representative indicators, it is necessary to conduct the further analysis to establish whether these indicators may be combined and whether more representative core indicators can be found. Hence, the factor analysis can also be used to extract further information on the data.

Factor analysis is a statistical technique that extracts representative common factors from a set of variables. Before applying factor analysis to financial diagnostics, its applicability must first be determined. Factor analysis can only be conducted if correlations exist between certain indicators while others remain uncorrelated. The analysis standardizes diagnostic indicators, eliminating correlations and non-comparability between them, thereby fundamentally ensuring diagnostic quality. Through factor analysis, indicators within grouped sets exhibit high internal correlation and low inter-group correlation, yielding a representative factor—the common factor—for each group. In financial diagnostics, each indicator can be expressed as the sum of a linear function composed of a few common factors and specific factors, with each factor being mutually independent. This eliminates information overlap between indicators and reduces dimensionality, facilitating the identification of primary issues. In financial diagnostics, the common factors derived from factor analysis provide the basis for determining diagnostic themes in subsequent solution generation. The process of extracting common factors in financial diagnostics broadly consists of:

(1) Standardizing raw data to eliminate differences in magnitude and units among variables.
(2) Calculate the correlation matrix of the standardized data;
(3) Determine the eigenvalues and eigenvectors of the correlation matrix;
(4) Compute variance contribution rates and cumulative variance contribution rates;
(5) Identify factors:

$$X_1 = a_{11}F_1 + a_{12}F_2 + \ldots + a_{1m}F_m + e_1 \tag{3}$$

$$X_2 = a_{21}F_1 + a_{22}F_2 + \ldots + a_{2m}F_m + e_2 \tag{4}$$

……

$$X_p = a_{p1}F_1 + a_{p2}F_2 + \ldots + a_{pm}F_m + e_p \tag{5}$$

(6) Factor rotation:

If the obtained m factors are ambiguous or lack clear practical significance, rotate the factors to derive more distinct practical meanings.

(7) Calculate factor scores using linear combinations of original indicators;

Employ the Warp scoring method to compute factor scores.

(8) Composite score

Taking the variance contribution ratio of each factor as the weight, the comprehensive evaluation index function is obtained from the linear combination of each factor. In the above equation, $e_1, e_2, e_3 \cdots\cdots e_p$ is the special factor, $a_{ij}(0 < i \leqslant p, 0 < j \leqslant m)$ is the factor loading, which, if it is close to the variance of X_i, can be used directly as a linear combination of $F_1, F_2, F_3 \cdots\cdots F_m$ as a linear combination of X_i, while ignoring the special factor e, of which the first n factors can be taken to reflect the original evaluation index when the cumulative contribution rate of the first n factors is not less than 80%;

$$F = \left(w_1 F_1 + w_2 F_2 + \ldots + w_m F_m\right) \Big/ \left(w_1 + w_2 + \ldots + w_m\right) \tag{6}$$

Here, w_i represents the variance contribution rate of the factor before or after rotation.

In an information-driven context, the SPSS software can be utilized to complete the aforementioned process. Within SPSS, the K-value test and sphericity test can be employed to verify the feasibility of factor analysis. When the K value exceeds 0.5 or the probability of the sphericity test is less than 0.05, it indicates that factor analysis can be conducted. The remaining calculations can be performed using the "Analysis" > "Dimensionality Reduction" > "Factor Analysis" module in SPSS. By selecting the appropriate options based on diagnostic requirements, indicators can be standardized to derive the correlation matrix, eigenvalues, eigenvectors, and contribution rates. Factor rotation is then applied to identify common factors, and composite scores are calculated.

The previous cluster analysis and factor analysis can yield more representative core diagnostic indicators. The use of these core indicators to conduct financial diagnosis analysis increases the accuracy of the diagnosis. This also increases the efficiency of diagnostics by reducing the number of indicators.

2.3 Financial Condition Early Warning Based on Data Mining

2.3.1 Early Warning System Function Settings

(1) Outlier Alert

When a company is experiencing financial problems or is at risk of them, monitoring and analyzing the financial and non-financial data will help in identifying outliers in some of those measures. It requires comparing the financial indicators of the company with the average level of the industry by using horizontal benchmarking to find metrics that are significantly higher or lower than the sector norm. The anomaly alerts make use of these outliers as reference points and depend on the financial indicator system of the company to determine the legitimacy of abnormal metrics. They also examine the reasons behind such anomalies by contrasting them with other financial indicators or business reports.

The outlier analysis approach should be based on the analysis of the financial indicators of the corporation. Reasonable indicators play a significant role in the efficiency of outlier analysis. According to the research conducted in the preceding section on visualizing financial data, which was informed by the financial analysis part of the Harvard analytical framework, a

financial indicator analysis system is developed using the five major functions of financial data. Then, the outlier analysis indicators mainly are:

1) Profitability indicators. Profitability is defined as the capability of a company to make profits within a certain time of its operations. Greater profits have a higher level of profitability. The health of a company development is an important indicator that corporate decision-makers use to measure the profitability of a company. Poor profitability is a sign of poor revenue generation and poor control over costs. Analysis of a companies profitability will be highly effective in predicting the future of the company. Moreover, when major changes in profitability are observed, it is possible to carry out detailed analysis of profitability or cost factors of the company and determine what caused the abnormalities. The previous academic studies have also indicated the fact that profitability is one of the main protection measures to evaluate a financial risk of a company.

2) Solvency Indicators. Solvency is the capability of a company to pay both long-term and short-term obligations with the help of its assets. The financial risk exposure of any company can be gauged directly by the capacity of the company to meet its debt obligations. Failure to repay debts is a sign of impending problems in maintaining operations and growth that may result in bankruptcy and liquidation. The solvency of an enterprise can also indicate the quality of its assets. High solvency means high asset liquidity, which allows quick conversion of the asset into cash and smooth capital flows. On the other hand, low solvency indicates low asset liquidity, which requires additional assessment of the valuation level of the assets.

3) Operational Efficiency Metrics. Operational efficiency refers to a company's ability to generate returns by utilizing its assets in business operations, fundamentally reflecting how effectively it leverages its assets. Strong operational efficiency indicates the company can rapidly realize asset value and demonstrates high asset liquidity. Weak operational efficiency suggests the company struggles to convert its resources into revenue, necessitating a multi-faceted analysis of management practices and business operations to identify causes. Since operational capability also reflects a company's debt-repayment capacity and profitability to some extent, it warrants close attention.

4) Growth Capability Indicators. Growth capability analysis assesses a company's capacity for business expansion. By examining this capability, one can better predict future development trends. Growth capability reflects the trajectory of a company's assets, including: trends in business scope, operational scale, and revenue/profit growth.

5) Cash Capacity Indicators. Cash capacity reflects the proportion of cash in a company's operating activities. As cash represents the most liquid asset, companies with strong cash capacity possess higher execution capabilities, enabling them to rapidly achieve development goals and fulfill strategic plans. Cash capacity analysis also serves as a complementary assessment to balance sheet and income statement operational analyses.

(2) Trend Early Warning

In some cases, a company's current financial or non-financial indicators may not show abnormalities, yet its future trajectory warrants management attention. The task entails applying longitudinal comparison techniques to predict future changes relying on past information and providing advance forecasts when the indicators indicate a decline. The present paper will be used to conduct time-series forecasting of different metrics with the help of the Python package known as Prophet and it will provide a warning of trends in future business development. It is also fast to detect indicators that demonstrate major change trends and explore the particular reasons why these changes occur, thus supporting decision making of the corporate management.

Prophet is an open-source model developed by Facebook focusing on large-scale time series analysis. It makes predictions of nonlinear trends by considering periodicity, additive impacts

and abrupt changes such as holidays. It is best at fitting data that is strongly periodic and has several cycles, but can also be used to fit data that has missing observations or outliers.

The main idea behind Prophet is the decomposition of a time series into three parts: trend, seasonality, and holidays. This decomposition is expressed as follows:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \tag{7}$$

$g(t)$ represents the trend component, used to fit the non-periodic components within the time series; $s(t)$ represents periodic variations (such as seasonal changes); $h(t)$ represents the impact of holidays (typically manifesting as special effects at certain time points); the error term ϵ_t accounts for all unaccounted-for error factors. This model transforms the forecasting problem into a curve-fitting problem, differing from models that analyze correlations within the data.

By forecasting various indicators using Prophet, early warnings are issued for metrics showing pronounced downward or upward trends, thereby fulfilling the trend alerting function.

2.3.2 Algorithm Support Based on Improved ID3 Decision Trees

(1) Fundamental Principles of Information Theory

Information theory views the communication process as the transmission of information in an environment subject to random disturbances. The calculation of mutual information primarily consists of two parts:

1) A posteriori entropy

A posteriori entropy is a measure of the information content of the input symbol U given the output symbol $V = V_j$ received at the channel receiver, defined as:

$$P\left(U \mid V_j\right) = \sum_j P\left(V_j\right) \sum_j P\left(U \mid V_j\right) \lg \frac{1}{P\left(U \mid V_j\right)} \tag{8}$$

2) Average Mutual Information

$H(U)$ represents the average uncertainty about the input symbol set U prior to receiving the symbol set V, while $H(U \mid V)$ represents the average uncertainty about the input symbol U after receiving the symbol set V. It is defined as:

$$I(U,V) = H(U) - H(U \mid V) \tag{9}$$

Among these: $I(U,V)$ is referred to as the average mutual information between U and V. It represents the amount of information about U obtained upon receiving the symbol set V.

(2) Design of the Weighted ID3 Improvement Algorithm

Research indicates a drawback of the traditional ID3 algorithm: it tends to favor attributes with more frequent values. This is because the weighted sum method causes the classification of the instance set to discard data tuples with low data volume. However, attributes with more frequent values are not always the optimal attributes.

To address the ID3 algorithm's tendency to select attributes with more frequent values as test attributes—which are not always optimal in practice—this paper introduces risk weighting:

Given $\mathrm{MAX}\left(-a_1, -a_i, \cdots, -a_n\right) \leqslant \beta \leqslant 0$

is referred to as the risk weight for uncertain knowledge. Here, a_i denotes the conditional

probability of an attribute, and the magnitude of β is determined by the conditional probabilities of attributes obtained in each iteration. If multiple conditional probabilities $a_1, a_i, \cdots, a_n$ exist during a computation, then $-\beta$ takes the minimum value among them, i.e., $0 \leqslant -\beta \leqslant \mathrm{MIN}\left(a_1, a_i, \cdots, a_n\right)$ or $\mathrm{MAX}\left(-a_1, -a_i, \cdots, -a_n\right) \leqslant \beta \leqslant 0$; β is a dynamically changing value.

When using risk weights, several aspects should be noted:

1) Select a valid β value where $\beta \leqslant 0$;

2) Choose appropriate risk weights for attributes to prevent data masking;

3) Risk weights are dynamically changing quantities that must vary with decision tree generation.

The improved ID3 algorithm enhances the rule generation method—specifically the attribute selection criteria algorithm. It strengthens attribute labeling while reducing labeling for non-important attributes. This prevents data tuples with fewer instances from being overwhelmed during decision tree generation. Ultimately, the decision tree reduces dependency on attributes with higher frequency values, thereby minimizing the occurrence of large data masking small data.

With the introduction of risk weights, the conditional entropy formula is calculated as:

$$H_\beta(U \mid V) = \sum_{j=1}^{n}\left(P\left(V_j\right) + \beta\right)\sum_{i=1}^{n} P\left(U_i \mid V_j\right) \lg \frac{1}{P\left(U_i \mid V_j\right)} \tag{10}$$

The formula for calculating corresponding mutual information is:

$$I_\beta(U,V) = H(U) - H_\beta(U \mid V) \tag{11}$$

The improved ID3 algorithm constructs decision trees by using Equations (10) and (11) as selection criteria for test attributes. In practical applications, one can first construct a decision tree using the ID3 algorithm. If the results show that important attributes with fewer values are farther from the root node than non-important attributes with more values, risk weights can be assigned. The decision tree can then be reconstructed using the improved ID3 algorithm to extract rules.

3 Practical Application of Big Data in Corporate Financial Condition Analysis

The empirical data analyzed in this paper is derived from the consolidated financial statements of a listed manufacturing company for the years 2020-2024. It encompasses the following financial ratios: quick ratio, debt-to-equity ratio, return on assets (ROA), return on equity (ROE), operating profit margin, accounts receivable turnover, inventory turnover, total asset turnover, operating revenue growth rate, net profit growth rate, total asset growth rate, operating cash flow per share, and net increase in cash and cash equivalents. All relevant data is sourced from the annual financial reports of listed companies disclosed on a specific website, and has been manually compiled and standardized.

3.1 Factor Analysis Model Construction

3.1.1 KMO and Bartlett's Sphericity Test

This study employed the KMO and Bartlett's sphericity test to conduct a preliminary evaluation of the selected data. Generally, a KMO value above 0.5 indicates suitability for factor analysis, while a KMO exceeding 0.7 signifies excellent suitability. The selected indicators achieved a KMO value of 0.692, exceeding the threshold of 0.5 and confirming the data's suitability for factor analysis. The Bartlett's sphericity test yielded a significance level below 0.05, with a Sig value of 2.944e-317. Both the KMO and Bartlett's sphericity test results are reasonably favorable, thus permitting the execution of factor analysis.

3.1.2 Factor Extraction

Standardization of raw data ensures the objectivity of factor analysis. Principal component analysis within factor analysis was subsequently employed to determine the number of principal factors. Calculations were performed for the pre-rotation variance contribution rate, pre-rotation eigenvalues, and pre-rotation cumulative variance contribution rate of the 16 financial indicators. The total variance explained is shown in Table 1. The extraction criteria for principal component analysis were: cumulative factor contribution exceeding 60% and consideration of eigenvalues. The cumulative contribution of the top five factors in the table reached 81.58%, meeting the requirements.

Table 1: Total variance explained

	Pre-rotation eigenvalue	Rotational front deviation contribution rate	Cumulative contribution rate of rotational forward deviation
1	6.1587	0.3849	0.3849
2	2.4375	0.1523	0.5372
3	2.0215	0.1263	0.6635
4	1.3286	0.0830	0.7465
5	1.1093	0.0693	0.8158
6	0.8937	0.0559	0.8717
7	0.6215	0.0388	0.9105
8	0.4011	0.0252	0.9357
9	0.3112	0.0195	0.9552
10	0.2735	0.0172	0.9724
11	0.1943	0.0121	0.9845
12	0.1029	0.0064	0.9909
13	0.0823	0.0051	0.9960
14	0.0521	0.0033	0.9993
15	0.0098	0.0006	0.9999
16	0.0021	0.0001	1

The visualization results of common factor extraction are shown in Figure 1. From left to right, the initial eigenvalues of each factor decrease progressively, exhibiting a diminishing trend.

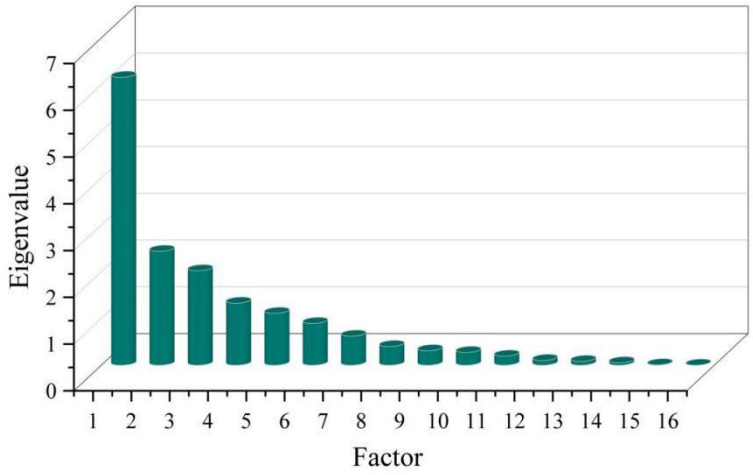

Figure 1: Visualized result of common factor extraction

3.1.3 Factor Rotation and Naming

To determine which aspects of corporate strength the extracted five principal factors represent, and to further classify the original 16 financial indicators into their respective principal factors, the original data was rotated using the maximum variance method. The loadings for each original factor are shown in Figure 2. The factor loading diagram reveals that the extracted principal component F1 exhibits high loadings on X1-X5, hence it is named the Profitability Factor. Principal component F2 shows high loadings on X6, X7, and X8, thus designated as the Solvency Factor. F3 demonstrates high loadings on X15 and X16, leading to its designation as the Operational Efficiency Factor. F4 exhibits high loadings on X11-X14, hence it is named the Growth Capability Factor; F5 shows high loadings on X9 and X10, both of which are related to net cash flow from operating activities, thus it is named the Cash Flow Capability Factor.

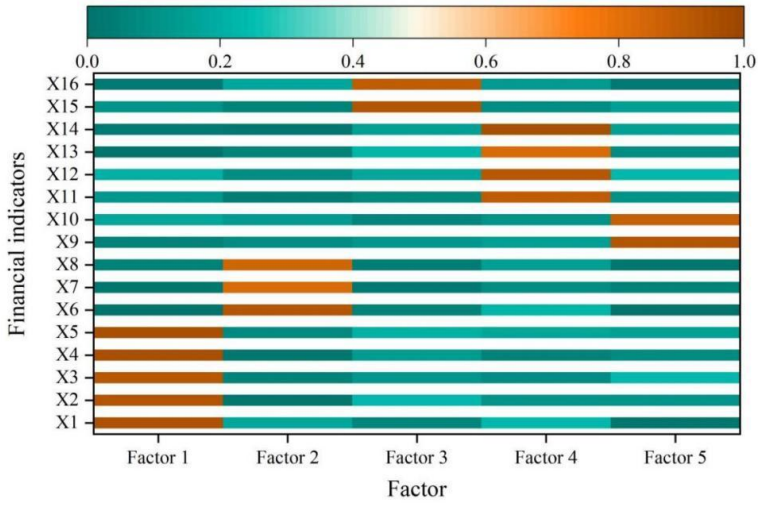

Figure 2: Loadings of each original factor

3.2 Financial Condition Early Warning

3.2.1 Comparative Analysis of Different Classification Models

For the original dataset, classification models based on both the ID3 decision tree and the improved ID3 decision tree were employed to predict corporate financial status. By altering the

category labels of training samples, varying levels of category noise were introduced into the original dataset to validate the performance of the proposed classification models. The prediction results of these models on datasets with different noise levels are presented in Table 2. It is evident that the proposed model effectively identifies and handles category noise samples. Specifically, on one hand, in terms of the accuracy rate for identifying noisy samples, the proposed model can accurately identify category-noise samples, with identification accuracy rates consistently above 0.86. On the other hand, by comparing the classification accuracy rates of different classification models, it can be observed that the proposed model effectively improves the overall accuracy rate, the accuracy rate for Class 1, and the accuracy rate for Class 2. Among these, for the initial state, since no category noise was added, the improvement in classification accuracy rate by the proposed model is relatively small. Such an upgrade is probably due to the greatest extent, to the location and treatment of the noise examples created at the case preservation stage. In the case of a dataset containing injected noise, the proposed model is very efficient in improving classification accuracy in all categories, and the most significant improvements were recorded in the presence of higher levels of categorical noise. The comparison made above shows that the enhanced classification model is effective since it correctly recognizes the presence of categorical noise samples as well as significantly improves the total classification accuracy.

Table 2: Prediction results on the original data sets under different noise levels

	ID3 Decision Tree			Improved ID3 Decision Tree			Accuracy rate
	Overall accuracy rate	First class accuracy	Second class accuracy	Overall accuracy rate	First class accuracy	Second class accuracy	
Initial	0.8901	0.9365	0.7562	0.8926	0.9711	0.7965	/
10% of the noise	0.8552	0.9236	0.7116	0.8887	0.9666	0.7816	0.8872
20% of the noise	0.8018	0.8804	0.6834	0.8863	0.9642	0.7624	0.8691
30% of the noise	0.7233	0.7962	0.6022	0.8852	0.9625	0.7573	0.8625
40% of the noise	0.6279	0.6517	0.5671	0.8827	0.9607	0.7554	0.8719
50% of the noise	0.4933	0.4983	0.4156	0.8716	0.9575	0.7361	0.8823

3.2.2 Comparative Analysis of Different Noise Levels

Further comparison reveals how the performance of the two classification models changes as noise levels gradually increase. The prediction results of the classification models on the original dataset at different noise levels are shown in Figure 3. As the level of class noise increases, the classification accuracy of the ID3 decision tree-based model drops significantly, while the classification accuracy of the proposed model remains relatively stable. Specifically, at the initial state and when 10% categorical noise was injected, the classification accuracy of the two models showed little difference. However, when 20% categorical noise was injected, the accuracy of the ID3 decision tree-based model began to decline noticeably, while the accuracy of the model proposed in this paper remained close to the initial state. As the category noise level further increased to 30%, the overall accuracy of the ID3-based decision tree classification model dropped to 0.7233. Furthermore, when 50% category noise was introduced, all three accuracy metrics of the ID3-based model fell below 0.5. In contrast, for the proposed model, although accuracy showed a slight decline as noise levels increased from 10% to 50%, its overall accuracy, Class 1 accuracy, and Class 2 accuracy remained above 0.87, 0.95, and 0.73, respectively.

The comparative analysis also indicates the high level of effectiveness and strength of the suggested CBR-based noise-resilient classification model that can demonstrate high

performance at datasets of different noise intensities.

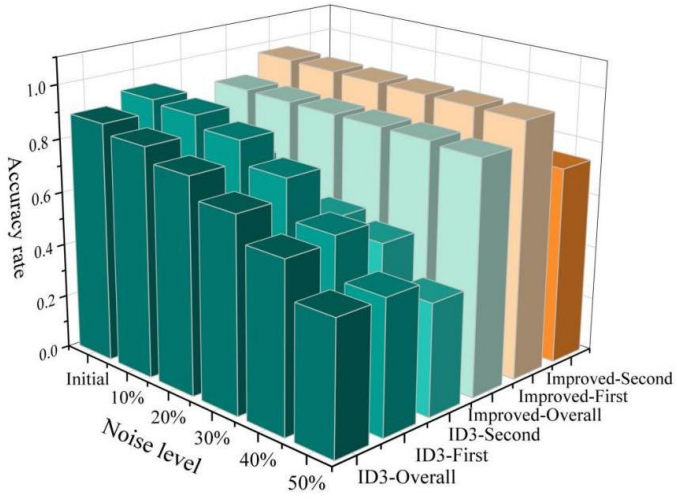

Figure 3: Prediction results on the original data sets under different noise levels

3.3 Analysis of Application Results

3.3.1 Decision Tree Generation

Having gathered, calculated, sorted out the necessary information, a classification model of the improved ID3 decision tree algorithm may be employed to continue examining relevant ratios and indicators that will help determine the key determinants of changes in the financial state of a company.

To simplify things, this paper will also choose three important financial ratios based on the 2024 Wall Score Method used by the company: debt-to-asset ratio, current ratio, and profit margin as the branch nodes and demonstrate how the proposed model groups financial status. In real-life application, every indicator in the Wall Score Method must be used as an attribute node to evaluate financial situation in general.

The calculation of the information entropy on each ratio shows that the information gain of the debt-to-asset ratio is highest. This ratio would serve as the root node of the decision tree depicted in Figure 4. If the debt-to-asset ratio is less than 40, and the current ratio is higher than 1.8 and the profit margin exceeds 1.4, then the financial state of the company is considered good.

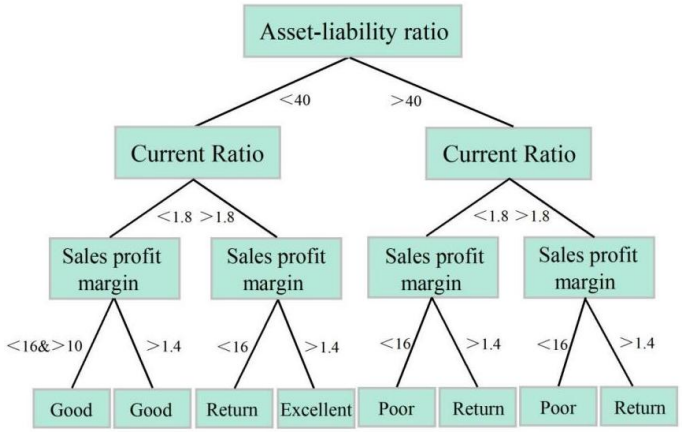

Figure 4: Decision tree

3.3.2 Results of Financial Indicator Analysis

The simulation analysis system has analyzed the company using the enhanced Volatility Score approach. The findings of the analysis of the financial indicators of the company in 2020 and 2024 are displayed in Table 3 and Table 4 respectively. Financial situation of the company as a whole showed a positive tendency of gradual improvement over the period between 2020 and 2024. The discrepancy between the current key financial indicators and the reference values was slowly reduced, which means better financial health. Debt-payment capability, operational effectiveness, and growth opportunity had made considerable progress especially, which is an indication of a step-by-step success in the process of financial control and strategy implementation. On total scoring, the company scored a total of 78.45 in 2020 and raised to 84.87 in 2024, a difference of 6.42 points. It means that there is a gradual increase in the overall financial performance of the company every year.

Table 3: Analysis results of financial indicators in 2020

Financial indicators	Actual ratio	Standard ratio	Difference	Adjustment factor	Standard rating	Score
Sales profit margin	15.386%	13.282%	2.104%	-0.9	5	4.05
Total asset return rate	12.433%	21.375%	-8.942	-0.7	5	4.32
Asset-liability ratio	42.195%	49.772%	-7.577	-0.73	10	8.87
Current ratio	2.7	1.6	1.1	-4.81	10	7.74
Accounts receivable turnover ratio	4.186	5.983	-1.797	1.03	20	15.15
Inventory turnover rate	1.842	7.992	-6.15	-5.22	20	14.47
Capital preservation and appreciation rate	128.375%	114.414%	13.961%	3.11	10	8.37
Sales revenue growth rate	12.753%	14.267%	1.514%	-2.98	10	8.26
Net profit growth rate	-4.963%	9.038%	-14.001%	-3.38	10	7.22
In total					100	78.45

Table 4: Analysis results of financial indicators in 2024

Financial indicators	Actual ratio	Standard ratio	Difference	Adjustment factor	Standard rating	Score
Sales profit margin	14.264%	13.282%	0.982%	-0.7	5	4.82
Total asset return rate	18.166%	21.375%	-3.209%	-0.8	5	4.61
Asset-liability ratio	47.355%	49.772%	-2.417%	-0.66	10	8.97
Current ratio	2.2	1.6	0.6	-4.23	10	8.74
Accounts receivable turnover ratio	4.962	5.983	-1.021	0.97	20	16.44
Inventory turnover rate	2.044	7.992	-5.948	-5.11	20	15.62
Capital preservation and appreciation rate	117.453%	114.414%	3.039%	2.45	10	9.04
Sales revenue growth rate	13.011%	14.267%	-1.256%	-3.03	10	8.62
Net profit growth rate	-1.024%	9.038%	-10.062%	-3.11	10	8.01
In total					100	84.87

4 Conclusion

The current paper will explore systematically the optimization directions of traditional financial analysis paradigms in the context of big data technology, and its main results will be as follows.

The factor analysis model had a cumulative variance contribution of 81.58 percent and was able to extract five common factors that have strong economic meaning. The model correctly classified categorical noise examples, and its recognition rate exceeded 0.86. With the rise in the injected noise level (10-50%), the total accuracy, Class 1 accuracy, and Class 2 accuracy were more than 0.87, 0.95, and 0.73, respectively.

When the debt-equity ratio of a company is less than 40 percent, the current ratio is more than 1.8 percent, and the profit margin is higher than 1.4 percent, it is said to be financially sound. All in all, the financial composite score of the company in 2020 was 78.45 points which increased to 84.87 points in 2024 (an improvement of 6.42 points) meaning that the financial performance of the company has been improving annually.

About the Authors

Tongyu Xing was born in Changzhi, Shanxi, P.R. China, in 1987. He obtained a Master's degree from Qinghai Minzu University in China.He is currently working at the College of Finance and Economics, Shanxi College of Applied Science and Technology. His main research direction is auditing and financial management.

Sisi Zhao was born in Baoji, Shaanxi, P.R. China, in 1989. She obtained a Master's degree from Qinghai Minzu University in China.She is currently working at the School of Economics and Management, Qinghai Minzu University. Her main research direction is financial accounting and financial management.

Zhidan Zhu was born in Fenyang, Shanxi, P.R. China, in 1986. She obtained a Master's degree from Shanxi University of Finance and Economics in China. She is currently working at the College of Finance and Economics, Shanxi College of Applied Science and Technology.Her main research direction is financial management.

References

[1] Zelgalve, E., & Zaharcenko, A. (2012). Transformation of the role of financial analysis in enterprise management. Management of Organizations: Systematic Research, 64, 147-167.

[2] Kumar, A. D., Boakye, M. M., & Celestin, M. (2021). The role of business mathematics in optimizing financial decision-making for SMEs in Ghana. Indo American Journal of Multidisciplinary Research and Review, 5(2), 47-52.

[3] Ahmadi, S. (2024). A comprehensive study on integration of big data and AI in financial industry and its effect on present and future opportunities. International Journal of Current Science Research and Review, 7(01), 66-74.

[4] Tang, J., & Karim, K. E. (2019). Financial fraud detection and big data analytics–implications on auditors' use of fraud brainstorming session. Managerial Auditing Journal, 34(3), 324-337.

[5] Zhu, B., Zheng, X., Liu, H., Li, J., & Wang, P. (2020). Analysis of spatiotemporal

characteristics of big data on social media sentiment with COVID-19 epidemic topics. Chaos, Solitons & Fractals, 140, 110123.

[6] Abbasi, A., Sarker, S., & Chiang, R. H. (2016). Big data research in information systems: Toward an inclusive research agenda. Journal of the association for information systems, 17(2), 3.

[7] Jackson, A. B. (2022). Financial statement analysis: a review and current issues. China Finance Review International, 12(1), 1-19.

[8] Zhao, L., & Huchzermeier, A. (2015). Operations–finance interface models: A literature review and framework. European Journal of Operational Research, 244(3), 905-917.

[9] Zopounidis, C., & Doumpos, M. (2002). Multi-criteria decision aid in financial decision making: methodologies and literature review. Journal of Multi-Criteria Decision Analysis, 11(4-5), 167-186.

[10] Putra, C. I. W., Soehaditama, J. P., Kurniawan, M. Y., Setyowati, T. M., & Sova, M. (2024). Fundamentals of Finance: Finance Management, Investment, Capital Market, and Funding. Dinasti Accounting Review, 2(1), 27-39.

[11] Chen, N., Ribeiro, B., & Chen, A. (2016). Financial credit risk assessment: a recent review. Artificial Intelligence Review, 45(1), 1-23.

[12] Hasan, M. M., Popp, J., & Oláh, J. (2020). Current landscape and influence of big data on finance. Journal of Big Data, 7(1), 21.

[13] Sohangir, S., Wang, D., Pomeranets, A., & Khoshgoftaar, T. M. (2018). Big Data: Deep Learning for financial sentiment analysis. Journal of Big Data, 5(1), 1-25.

[14] Liu, H., Huang, S., Wang, P., & Li, Z. (2021). A review of data mining methods in financial markets. Data Science in Finance and Economics, 1(4), 362-392.

[15] Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision support systems, 50(3), 559-569.

[16] Anton, S. G., & Nucu, A. E. A. (2020). Enterprise risk management: A literature review and agenda for future research. Journal of Risk and Financial Management, 13(11), 281.

[17] Crawford, J., & Jabbour, M. (2024). The relationship between enterprise risk management and managerial judgement in decision-making: A systematic literature review. International Journal of Management Reviews, 26(1), 110-136.

[18] Jin, M., Wang, Y., & Zeng, Y. (2018). Application of data mining technology in financial risk analysis. Wireless Personal Communications, 102(4), 3699-3713.

[19] Cao, M., Chychyla, R., & Stewart, T. (2015). Big data analytics in financial statement audits. Accounting horizons, 29(2), 423-429.

[20] Saleh, I., Marei, Y., Ayoush, M., & Abu Afifa, M. M. (2023). Big data analytics and

financial reporting quality: qualitative evidence from Canada. Journal of Financial Reporting and Accounting, 21(1), 83-104.

[21] Chen, W. S., & Du, Y. K. (2009). Using neural networks and data mining techniques for the financial distress prediction model. Expert systems with applications, 36(2), 4075-4086.

[22] Wang, A., & Yu, H. (2022). The construction and empirical analysis of the company's financial early warning model based on data mining algorithms. Journal of Mathematics, 2022(1), 3808895.

[23] Gao, B. (2022). The use of machine learning combined with data mining technology in financial risk prevention. Computational economics, 59(4), 1385-1405.

[24] Koyuncugil, A. S., & Ozgulbas, N. (2012). Financial early warning system model and data mining application for risk detection. Expert systems with Applications, 39(6), 6238-6253.

[25] Habib, A., Costa, M. D., Huang, H. J., Bhuiyan, M. B. U., & Sun, L. (2020). Determinants and consequences of financial distress: review of the empirical literature. Accounting & Finance, 60, 1023-1075.

[26] Abdu, E. (2022). Financial distress situation of financial sectors in Ethiopia: A review paper. Cogent Economics & Finance, 10(1), 1996020.

[27] Sun, J., & Li, H. (2008). Data mining method for listed companies' financial distress prediction. Knowledge-Based Systems, 21(1), 1-5.

[28] Geng, R., Bose, I., & Chen, X. (2015). Prediction of financial distress: An empirical study of listed Chinese companies using data mining. European Journal of Operational Research, 241(1), 236-247.

[29] Noviyanti, E. A., Rahayu, C. W. E., & Rahmawati, C. H. T. (2021). Financial performance and stock price: Another review on banks listed in Indonesia Stock Exchange. Journal of Management and Business Environment (JMBE), 3(1), 70.

[30] Xu, S., & Zhang, M. (2005, July). Data mining-an adaptive neural network model for financial analysis. In Third International Conference on Information Technology and Applications (ICITA'05) (Vol. 1, pp. 336-340). IEEE.

[31] Kannan, K. S., Sekar, P. S., Sathik, M. M., & Arumugam, P. (2010, March). Financial stock market forecast using data mining techniques. In Proceedings of the International Multiconference of Engineers and computer scientists (Vol. 1, No. 4, pp. 1-5).