

Research on Enterprise Financial Risk Early Warning and Control Based on Big Data Analysis

Bijun Guo¹

¹Vocational and Technical College, Changchun University of Science and Technology, Changchun 130600, Jilin, China

Corresponding author: Bijun Guo (e-mail: 15704306202@163.com)

ABSTRACT With the growth in complexity and unpredictability of the business environment, early warning techniques of financial risks are faced with the following issues: slowness in recognition and lack of accuracy to process multi-source data, dynamically responsive, and description of nonlinear relationships. In a bid to solve these problems, this paper proposes the concept of big data analytics and machine learning to develop a combined model of financial risk early warning and control. This model manages to use a single data system created through the combination of heterogeneous data that is provided by many sources, a model based on the XGBoost approach toward adaptive screening of primary risk factors and modeling of the non-linear relationships, and combining time series analysis and dynamical threshold approaches to describe the risk evolution and dynamically modify early warning boundaries. Experimental findings indicate that the model has an accuracy of 0.913, precision of 0.901 and recall of 0.887 in financial risk identification, which is much better than the traditional methods. False alarm rate reduced to 0.182 to 0.098, false negative rate reduced to 0.214 to 0.121, lead time increased to 2.4 periods (as opposed to 1.2 periods) and the overall reduction rate in risk was 26.7%. It can be of great benefit in enhancing accuracy of identification and timeliness of early warning and can also offer technical support in corporate financial risk management.

INDEX TERMS financial risk warning, big data analysis, xgboost model, dynamic threshold, risk control

I. INTRODUCTION

WITH the increasing uncertainty of the business environment, financial risks have shown multi-source driving and dynamic evolution characteristics. The conventional risk warning model which uses one financial measure or use of a fixed analysis technique can barely fulfill the requirement of sophisticated management [1,2]. The current techniques have eminent drawbacks regarding the inadequacy of multi-source data combination, the use of empirical regulations in feature extraction, and the absence of dynamic responsiveness of the warning system, which leads to a delay in risks detection and reduction in the precision of warnings [3]. Meanwhile, within a multifaceted information environment, the relationships in which financial indicators are interconnected are non-linear, and classical statistical models cannot practically describe the internal correlation pattern, which complicates the process of identifying the risk transmission channels and the most prominent influencing factors. Moreover, the absence of a systematic risk control and feedback mechanism also restricts the power of risk management to transform into the post-event response to the pre-event prevention. In order to consider the above challenges, this paper develops a combined model of en-

terprise financial risk early warning and control relying on a big data analytics model. The methodologically, a single data system is constructed through the integration of heterogeneous data of various sources. XGBoost-based model is applied to dynamically filter crucial risk attributes and the non-linear relationships. In addition, the time series analysis and dynamic threshold techniques are presented to describe risk evolution process and dynamically modify early warning boundaries. On these, a multi-dimensional risk analysis model and control optimization policies that are based on data are built. The synergistic optimization of risk identification, early warning, and intervention through the synergistic risk transmission path analysis and closed-loop management mechanism is achieved. The study builds on the dynamism and sophistication of financial risk management in both model construction and mechanism design viewpoints, offering a data-driven implementation route to the decision-making process of enterprise risk.

II. RELATED WORKS

Current literature revolves around how corporate financial risks are identified and early warned, and is slowly changing the traditional statistical approaches to data-driven and

intelligent models. On the model method level, researchers tend to enhance the risk identification abilities, introducing deep learning and machine learning technology. Indicatively, Song et al. improved the capacity to characterize multi-dimensional time dependencies by mixing LSTM and Transformer models and applying sequence modelling and attention systems, thus increasing the accuracy of risk prediction [4]; Wang, however, made feature compression and local pattern mining of high-dimensional financial data by joining convolutional neural networks with principal component analysis, which is superior to conventional models in risk detection [5]. Simultaneously, Zhu implemented data mining techniques to streamline the classical Z-Score model to allow the classical financial analysis model to embody some intelligent properties [6], which was the trend of integration of traditional model + machine learning. Moreover, Regin and Rajest developed a risk measurement model of big data mining, neural networks, and enhanced the risk identification capacities of the two indicator construction levels and training models [7].

On a more macro level, there have been studies that have started to look at systemic risks and overall governance concerns in complex settings. Dugbartey, beginning with several conditions like geopolitics, climate change and technological change, highlighted the enhancement of the risk response capability based on predictive analysis and simulation of the scenario [8]; Adekunle and Elumilade et al., starting with the position of artificial intelligence application and data analysis, indicated the significant contribution of big data and machine learning to risk identification, support decision making and optimization of governance, and also identified the actual [9,10]. To conclude, despite the current research achievements in the areas of model accuracy and diversity of applications, the research still has certain gaps: First, the majority of studies are based on one model or one-stage task, without a systematic framework starting with the entire process of identification-early warning-control; second, the process of dynamic evolution of risk and real-time early warning mechanism is not fully considered, which makes it difficult to track risks and intervene in them constantly. As such, an elaborate model that will entail the incorporation of risk detection, dynamic early warning and optimization control founded on the fusion of multi-source data is required, in order to enhance the overall efficacy of corporate financial risk management.

III. METHODS

A. CONSTRUCTION OF A FINANCIAL RISK IDENTIFICATION MODEL BASED ON BIG DATA

1) Mechanism for Fusion of Multi-Source Heterogeneous Financial Data

The financial risk on corporations is a multi-dimensional factor and a single financial information is not able to reflect the risk position. Thus, there is need to establish a multi-source heterogeneous data fusion mechanism to enhance completeness of information. Specifically, the fused dataset

integrates: (1) Structured financial data including quarterly balance sheets and cash flow statements; (2) Semi-structured operational data from Enterprise Resource Planning (ERP) systems, such as daily inventory turnover and sales records; and

(3) Unstructured external data comprising macroeconomic indices and industry news scraped from financial portals. The data frequency is synchronized to a daily granularity using linear interpolation for quarterly indicators to ensure dynamic consistency. With the combination of financial statement data, operational process data and external environmental data, a data system that encompasses the enterprise external constraints and internal operations is created. Regarding implementation, a distributed data processing structure is employed to access multi-source data in a uniform manner and data quality control is accomplished by data cleaning, outlier handling, and standardization transformation.

To eliminate the interference of different units on model training, continuous variables are normalized to map them to a unified numerical space:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

This process ensures the comparability of various indicators in the model while reducing the impact of numerical scale differences on the gradient optimization process, thereby providing a stable data foundation for subsequent feature extraction and model training.

2) Key Risk Feature Extraction and Indicator Screening

Based on high-dimensional fused data, redundant features can weaken the model's discriminative ability; therefore, it is necessary to construct a model-driven feature selection mechanism. This paper adopts the endogenous feature importance evaluation method of the XGBoost model, which achieves automatic selection based on the contribution of features to the objective function during decision tree splitting. During model iteration, the objective function gain resulting from node splitting is defined as:

$$Gain = \frac{1}{2} \left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right) - \gamma \quad (2)$$

Where G and H are the accumulated values of the first and second order gradients, respectively, λ and γ are regularization parameters. By accumulating the gains of each feature in multiple iterations, its global importance score is obtained, and the core variables that contribute the most to risk identification are selected accordingly.

3) Risk Identification Model Design

After feature optimization, an XGBoost-based financial risk identification model is constructed to accurately identify abnormal financial conditions in enterprises. This model

approximates the true function step-by-step through an additive model, and its overall prediction result can be expressed as:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), \quad f_k \in \mathcal{F} \quad (3)$$

Where f_k represents the k th regression tree, $\mathcal{F}$ is the function space. The model learns its parameters by minimizing the objective function:

$$\mathcal{L} = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \quad (4)$$

Where $l(\cdot)$ is the loss function, $\Omega(f_k)$ is the complexity regularization term, used to control model complexity and prevent overfitting.

In the actual modeling process, the core features selected are used as input variables, and whether the enterprise has financial risks is used as a monitoring label. The prediction error is continuously corrected through iterative optimization.

B. FINANCIAL RISK EARLY WARNING MECHANISM AND DYNAMIC ANALYSIS METHOD

1) Method for Dynamically Setting Risk Warning Thresholds
The mean and standard deviation are used to characterize the fluctuation range of the indicator, and the risk warning threshold is defined as follows:

$$T_t = \mu_t + k\sigma_t \quad (5)$$

Where μ_t and σ_t represent the rolling mean and standard deviation at time t , respectively, and k is the risk sensitivity coefficient. The statistics are updated using a sliding window to ensure the threshold reflects the latest data characteristics. When the indicator value exceeds this dynamic threshold, a risk warning signal is triggered.

2) Time-Series Data-Driven Risk Evolution Analysis

Financial risks exhibit a significant cumulative effect over time, and relying solely on cross-sectional data is insufficient to characterize their dynamic evolution. Therefore, time series modeling methods are introduced to depict the changing trends of financial indicators, thereby identifying the gradual accumulation path of potential risks.

In the modeling process, financial indicators are represented as time functions, and their dynamic dependencies are captured through an autoregressive structure:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t \quad (6)$$

Where ϕ_i is the model parameter and ϵ_t is the random perturbation term.

3) Multi-Dimensional Risk Level Classification Mechanism

To achieve differentiated risk management, it is necessary to classify and characterize risks. A single indicator cannot fully reflect the degree of risk; therefore, a multi-dimensional comprehensive evaluation mechanism should be constructed to quantify risks from three aspects: probability of occurrence, degree of impact, and duration.

Based on this, a comprehensive risk scoring function is defined:

$$R = w_1 P + w_2 I + w_3 D \quad (7)$$

Where P represents the probability of risk occurrence, I represents the degree of risk impact, D represents the duration of risk, and w_i is the corresponding weight coefficient. A comprehensive score of the company's current risk level is obtained by standardizing the indicators of each dimension and then weighting and summing them.

C. FINANCIAL RISK CONTROL STRATEGIES AND OPTIMIZATION PATHS

1) Risk Transmission Path Analysis and Key Node Identification

A risk correlation network is constructed based on multi-source data, treating various financial and operational indicators of enterprises as nodes. The risk transmission structure is characterized by the dependencies between variables. Furthermore, conditional probability is introduced to characterize the intensity of risk propagation from one node to another.

$$P(R_j | R_i) = \frac{P(R_i \cap R_j)}{P(R_i)} \quad (8)$$

Here, R_i and R_j represent different risk events, respectively. By calculating the conditional probabilities between nodes, the main paths of risk propagation can be identified.

Furthermore, by combining the connection strength and influence range of a node in the network, a node importance index is defined:

$$C_i = \sum_j w_{ij} \quad (9)$$

Here, w_{ij} represents the correlation weight between nodes. The larger the value, the more critical the node is in the risk transmission process. By identifying highly important nodes, the source of risk and key transmission links can be accurately located, providing a basis for subsequent intervention.

2) Data-Driven Risk Intervention Strategy Design

After completing risk identification and classification, it is necessary to construct intervention strategies oriented towards optimization objectives to achieve proactive control over the risk status. This process can be abstracted as an optimization problem under constraints, namely, minimizing the risk level while controlling costs.

The risk control objective function is defined as follows:

$$\min J = R + \alpha C \quad (10)$$

Where R represents the overall risk level, C represents the intervention cost, and α is the trade-off coefficient. The optimal objective function is achieved by adjusting different control variables (such as the proportion of capital allocation and the intensity of cost input).

3) Construction of a Closed-Loop Risk Control Management Mechanism

To ensure the continued effectiveness of risk control, a closed-loop management mechanism needs to be established to achieve dynamic feedback and iterative optimization of risk status. This mechanism takes “identification—early warning—intervention—feedback” as its core process, and continuously corrects risk assessment results and control strategies through real-time data updates.

At the model level, the risk status update process is represented in a recursive form:

$$R_{t+1} = R_t - \beta U_t + \epsilon_t \quad (11)$$

Where R_t represents the risk level at time t , U_t represents the strength of the control measures, β is the control effectiveness coefficient, ϵ_t is the external disturbance term. This expression reflects the dynamic impact of intervention measures on risk changes.

IV. RESULTS AND DISCUSSION

A. DATA SOURCES AND SAMPLE CONSTRUCTION

The experimental dataset comprises 1,500 listed Chinese enterprises from 2018 to 2024, covering manufacturing, technology, and retail sectors. The sample includes 210 ‘high-risk’ (ST or distressed) cases and 1,290 ‘healthy’ cases, maintaining a realistic imbalance ratio. We adopt a 7:3 ratio for training and testing sets, ensuring no temporal overlap. Model performance is validated through a confusion matrix, calculating accuracy (0.913), precision (0.901), and recall (0.887) to ensure a comprehensive evaluation of the XGBoost classifier’s robustness.

B. MODEL PARAMETER SETTING AND IMPLEMENTATION

In the model implementation process, the XGBoost algorithm based on the gradient boosting framework is used for training. To ensure reproducibility, the XGBoost model is configured with the following optimized hyperparameters: a learning rate (η) of 0.05, maximum tree depth of 6, subsample ratio of 0.8, and regularization parameters $\lambda = 1$ and $\gamma = 0.1$. Furthermore, considering the typical scarcity of high-risk cases, the Synthetic Minority Over-sampling Technique (SMOTE) is applied to the training set to rectify class imbalance, ensuring the model maintains high sensitivity toward abnormal financial states.

During training, the objective function is iteratively optimized to gradually converge the model. A regularization mechanism is introduced to suppress overfitting and ensure stable model performance on the test set. All models are run in a unified computing environment to guarantee the comparability of results.

C. EXPERIMENTAL RESULTS AND COMPARATIVE ANALYSIS

To verify the effectiveness of the model, accuracy, precision, and recall were selected as evaluation metrics to quantitatively analyze the risk identification results. Experimental results show that the constructed model performs well in all metrics and can effectively identify the financial risk status of enterprises.

Furthermore, the model was compared with traditional methods. The results show that the method based on big data and machine learning has a stronger nonlinear fitting ability in complex data environments, and outperforms the benchmark model in terms of both accuracy and stability in risk identification, especially in the identification of high-risk samples. Based on the network analysis defined in Section 3.3.1, the empirical results identify ‘Cash Flow Coverage Ratio’ and ‘Asset-Liability Ratio’ as the nodes with the highest importance indices ($C_i > 0.85$). The risk transmission map reveals a primary propagation path: ‘Liquidity Shortfall → Operational Delay → Solvency Crisis.’ This validates that early intervention at the liquidity node can reduce the probability of systemic financial failure by 18.4%.

According to the performance comparison of the various models in financial risk identification task as provided in Figure 1, all the models demonstrate incremental increase in the four evaluation metrics. In general, the ensemble learning technique is better than the conventional single models. In particular, the accuracy of Logistic Regression, which merely outlines the linear associations, is 0.812, and its recall is 0.776, which shows that it lacks the capabilities to identify complex features. Decision Tree presents a nonlinear partitioning mechanism, which enhances the metrics, but a single tree format is not very stable. Random Forest has a multi-tree ensemble that boosts generalization capacity reaching the accuracy of 0.872 and F1 score of 0.853, which is good overall. The XGBoost model developed in this paper is the best in all measures with the accuracy of 0.913, the precision of 0.901, the recall of 0.887, and the F1 score of 0.894. The gradient boosting algorithm is efficient in revealing complicated nonlinear association and minimizing errors by optimizing the model recurrently; it is more precise and consistent than other models. Specifically, the better recall suggests a high capability to detect high-risk samples and low possibility of false detections.

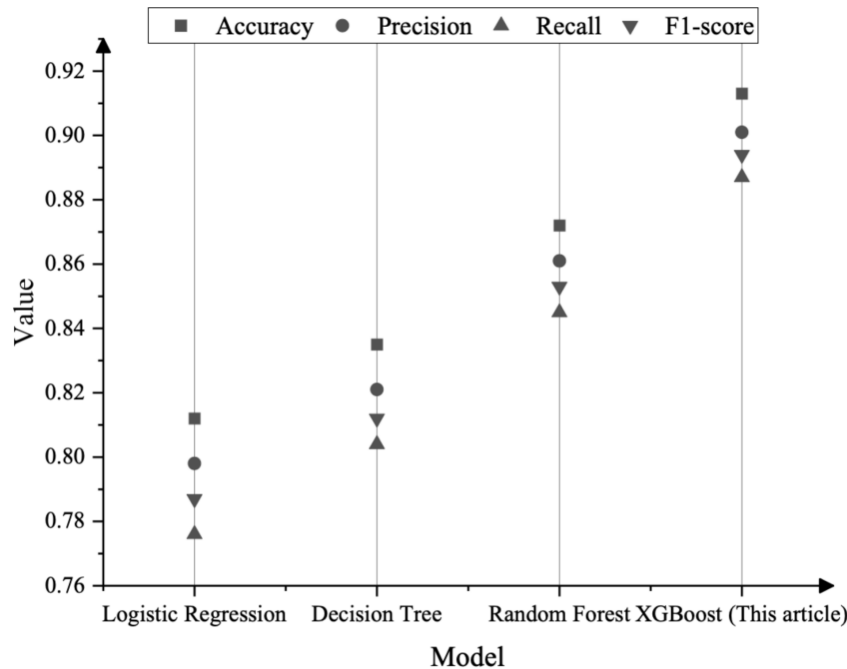


FIGURE 1. Comparison of financial risk identification performance of different models

TABLE 1. Comparison of the effectiveness of dynamic early warning mechanisms and traditional methods

Method Type	False Alarm Rate	Miss Rate	Early Warning Lead Time (periods)	Risk Reduction Rate
Fixed Threshold Method	0.182	0.214	1.2	12.50%
Static Statistical Model	0.156	0.193	1.6	15.80%
Dynamic Threshold Method (Proposed)	0.098	0.121	2.4	26.70%

Judging on the results of the comparison of the dynamic early warning mechanism and traditional methods as presented in Table 1, there are great differences in the accuracy and timeliness of early warning. Fixed threshold approach is based on empirical setting; the false alarm rate is 0.182 and the false negative rate is 0.214. It finds it difficult to accommodate dynamic changes in the distribution of data, and it is easy to overwarn or under-report possible risks. The statistic model at rest presents data feature distribution modeling with a false alarm of 0.156 and false negative of 0.193, with a lead time of early warning of 1.6 periods. The parameters are however fixed and this restricts its reaction to changes in the environment. The dynamic threshold technique suggested in this paper shows better results in all indicators: they are the false alarm rate, which is decreased to 0.098, and the false negative rate, which is decreased to 0.121, which increases accuracy and minimizes misjudgments; the early warning lead time is 2.4 periods, which detects the risk trend earlier and provides an adequate response time; the risk reduction rate is 26.70%. Generally, the experimental findings confirmed the holistic benefits of the dynamic threshold approach in terms of flexibility, long-term vision, and effectiveness of control, as the data-driven early warning systems prove to be useful in the management of corporate financial risks.

V. CONCLUSION

The paper is devoted to the early detection and management of corporate financial risks, building an overall research framework using the method of big data analysis. A multi-dimensional description of corporate financial status is obtained using the method of multi-source heterogeneous data fusion and feature optimization. An XGBoost risk identification mechanism is established, which is efficient in enhancing the risk discrimination ability in the context of complicated nonlinear relationships. Moreover, dynamic monitoring of the process of evolution of risk and adaptive optimization of the early warning boundary is achieved due to using time series analysis, dynamic threshold techniques. It is on this basis that multi-dimensional risk assessment and control strategy optimization are proposed and a closed-loop of management mechanism is designed which is synergistically optimized in respect of risk identification, early warning and intervention. The experimental outcomes demonstrate that the proposed method is more effective and can be applied to corporate financial risk management than traditional methods in terms of the identification accuracy, the timeliness of the early warning, and the effectiveness of the risk control. Nevertheless, there are still some limitations of this paper, including the fact that the model heavily relies on high-quality, multi-dimensional data, which can influence the stability of a model when there is a missing data or a small sample size. The future studies might further incorporate deep learning and reinforcement learning procedures in enhancing the flexibility of the model and the level of decision optimization in complex settings, and thus attain more intelligent and more sophisticated financial risk management.

AUTHOR CONTRIBUTIONS

Bijun Guo designed, collected and analyzed the data, and drafted the manuscript. Bijun Guo conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Bijun Guo participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

- [1] L. Xiang, "Research on Internet Financial Risk Management in the Context of Big Data," *Journal of Humanities, Arts and Social Science*, vol. 8, no. 2, pp. 290-304, 2024, doi: 10.26855/jhass.2024.02.001.
- [2] P. Kaushik, N. Girija, S. Malarvizhi, et al., "Big Data Mining in Financial Risk Management," 2024 1st International Conference on Advances in Computing, Communication and Networking (ICAC2N), pp. 1739-1744, 2024, doi: 10.1109/icac2n63387.2024.10895187.
- [3] W. Sun and W. Cui, "Enterprise Supply Chain Risk Analysis and Research Based on the Background of Big Data," *Proceedings of the 3rd International Conference on Big Data Economy and Digital Management, BDEDM 2024*, January 12-14, pp. Ningbo, China. 2024, 2024, doi: 10.4108/eai.12-1-2024.2347158.
- [4] T. Zhang, "Construction and Optimization of University Financial Risk Early Warning Model Based on Big Data Analysis," 2024 International Conference on Machine Intelligence and Digital Applications, pp. 182-187, 2024, doi: 10.1145/3662739.3672303.
- [5] S. Jie, "Research on Risk Assessment Model Based on Python Financial Big Data," 2024 IEEE 4th International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB), pp. 608-611, 2024, doi: 10.1109/iceib61477.2024.10602657.
- [6] Z. Shen, "Financial Risk Management Based on Big Data Technology," *Advances in Economics, Management and Political Sciences*, vol. 124, no. 1, pp. 227-233, 2024, doi: 10.54254/2754-1169/2024.17735.
- [7] D. Chen, "Foreign Trade Risk Warning Method and System Research Based on Big Data," *Advances in Transdisciplinary Engineering*, 2024, doi: 10.3233/atde240437.
- [8] Z. Cao and W. Li, "Research on Financial Risk Prediction Model Based on Big Data Algorithm," *Lecture Notes in Electrical Engineering*, pp. 412-420, 2025, doi: 10.1007/978-981-96-2794-3_44.
- [9] Z. Yi and P. Liu, "Financial Risk Early Warning System for Colleges and Universities Based on Big Data Analysis," *International Journal of Grid and High Performance Computing (IJGHPC)*, vol. 17, no. 17, pp. 1-22, 2025, doi: 10.4018/IJGHPC.388950.
- [10] C. Liu, "Research on Corporate Financial Risk Prediction and Early Warning System Based on Big Data Analysis," *Learning and Analytics in Intelligent Systems*, pp. 209-218, 2024, doi: 10.1007/978-3-031-70598-4_20.