Construction and Comparative Analysis of Financial Risk Early Warning Model Based on Machine Learning

Wu Xiannan
Zhejiang Gongshang University Hangzhou College of Commerce, Zhejiang, China

Abstract: Amidst the progressing intensification of globalization and economic integration, the management of financial market volatility and risk has become a critical concern for businesses. The escalating complexity of financial crises has heightened the urgency for advanced early warning systems capable of accurately forecasting future risks. In this context, this study has developed a suite of financial risk early warning models leveraging machine learning techniques, encompassing Support Vector Machine (SVM), Random Forest (RF), and Deep Learning (DL) models, and conducted a thorough comparative analysis. The research utilized financial statement data and market transaction records from A-share listed companies spanning 2008 to 2022. Post the removal of multicollinearity, standardization, and outlier exclusion, a dataset was curated that included over 30 financial indicators such as the current ratio, debt-to-asset ratio, and net profit growth rate. A logistic regression model was applied for baseline comparison, revealing that machine learning models notably outperformed it across key metrics including accuracy, precision, and recall rates.The DL model, in particular, showcased enhanced predictive capabilities for financial risks, attributable to its proficiency in capturing non-linear features and its automated high-level feature extraction capabilities. Conversely, the RF model provided practical benefits in terms of feature interpretability, swift training, and the provision of feature importance scores. To bolster the models' adaptability and predictive accuracy in complex scenarios, the study proposes enhancements, advocating for an early warning mechanism within DL models that integrates multi-source heterogeneous data and dynamic financial indicators.

Moreover, to address the dynamic nature of financial markets, this study has integrated a real-time assessment mechanism. This mechanism facilitates ongoing monitoring and prompt adjustment of model parameters in response to market fluctuations, ensuring the model's sustained efficacy and dependability. The generalizability of the models was substantiated through time-series cross-validation and replication across various industry subsamples, demonstrating their stability and robust performance. This research presents scientific financial risk assessment tools for corporate management and investors, laying the groundwork for future advancements in the domain.

Keywords: Financial Risk Early Warning; Machine Learning; Support Vector Machines; Random Forests; Deep Learning; Real-Time Assessment; Model Robustness.

1. Introduction

In recent years, with the rapid development of big data and artificial intelligence technologies, machine learning methods have been widely applied in the field of financial risk warning. Traditional financial risk warning models, primarily reliant on expert experience and statistical methods, exhibit inherent limitations in capturing the complex nonlinear relationships of financial data. Machine learning methods, with their ability to automatically learn risk characteristics from vast datasets and construct highly nonlinear warning models, offer strong predictive and reasoning capabilities. Financial risk is one of the critical risks faced in corporate operations, and timely, accurate early warning of financial risks is of great significance for corporate risk management and decision-making.

Despite the promising performance of machine learning methods in financial risk warning, as demonstrated by Barboza et al.[1] and Sun et al.[2], there is a gap in the literature regarding the in-depth analysis of model prediction errors and the real-time assessment of model performance in dynamic environments. This study aims to address these gaps by conducting a comprehensive analysis of prediction errors, identifying the underlying causes of misclassifications, and implementing a real-time assessment mechanism to enhance model robustness and adaptability.

The key to building a robust machine learning risk warning model lies in feature engineering and model design. We will employ feature selection methods such as ReliefF and LASSO to screen feature subsets with high information content. Following preprocessing steps like standardization and normalization, we will train the machine learning models. To tackle the challenge of unbalanced data, we plan to employ methods such as under-sampling and over-sampling to improve model performance. In terms of model design, we will explore frameworks that integrate anomaly detection and leverage both unsupervised learning for detecting abnormal samples and supervised learning for classifier training, as suggested by Zhou et al.[5]. Additionally, we will investigate models that integrate graph convolutional networks and capsule networks for corporate relationship risk prediction, effectively utilizing corporate association information, as proposed by Wu et al.[6].

To ensure the generalizability and robustness of our models, we will use 10-fold cross-validation and introduce a real-time assessment mechanism. This mechanism will allow for continuous monitoring of model performance and prompt adjustments in response to changing financial market conditions[16]. We will evaluate model performance using a suite of metrics including accuracy, precision, recall, F1 score, and AUC.

2. Literature Review

Financial risk warning has emerged as a focal area of concern among scholars worldwide. Qian Kai et al. (2019) utilized a Logistic regression model on data from Chinese listed companies, identifying key financial risk factors such as the debt-to-asset ratio and business income growth rate[1]. Liu Xiaona et al. (2020) developed a deep learning-based corporate credit risk warning model, demonstrating superior predictive accuracy and robustness over traditional models[2]. Zheng Yuqi et al. (2021) enhanced the practicality of financial risk warnings by integrating multimodal data sources, including financial statements, news texts, and market sentiment, into a unified framework[3].

Internationally, significant strides have been made in financial risk warning research. Beutel et al. (2019) introduced Risk-GUARD, a graph neural network model that captures intricate enterprise interactions, providing precise early warnings for supply chain financial risks[4]. Aziz et al. (2020) proposed a transfer learning approach for credit risk assessment in small and medium-sized enterprises, leveraging knowledge from related domains to overcome data scarcity[5]. Choi (2021) presented FinBERT, a pre-trained language model capable of extracting financial risk insights from vast amounts of unstructured text, offering innovative perspectives for risk analysis[6].

The incorporation of causal inference in financial risk warning is a burgeoning area of research. Li Hongwei et al. (2022) introduced CausalRisk, a framework that employs Granger causality testing and counterfactual analysis to elucidate the causal mechanisms underlying financial risk, thereby enhancing the interpretability and precision of risk warnings[7]. Fang et al. (2023) combined causal discovery with reinforcement learning to develop CausalRL, a system that dynamically adjusts risk mitigation strategies and has shown promising empirical results[8].

This paper aims to contribute to the existing body of knowledge by conducting a comparative analysis of various machine learning models in the context of financial risk warning. We will also discuss the importance of real-time assessment mechanisms for monitoring model performance and adapting to the dynamic financial landscape, an aspect that has been underexplored in the current literature.

3. Overview of Machine Learning Methods

This study conducts an in-depth comparative analysis of representative supervised learning algorithms, carefully selected to align with the intricacies of financial data. Algorithms such as Random Forest are favored for financial risk level prediction due to their robust fault tolerance and strong generalization capabilities. A comprehensive dataset encompassing a wide array of financial indicators is utilized during the model training phase, where model parameters are meticulously adjusted and optimized using cross-validation techniques to enhance the model's predictive power and accuracy.

In addition to traditional performance metrics like the Area Under the ROC Curve (AUC), precision, and recall rate, this study introduces real-time assessment mechanisms to continuously monitor and dynamically adjust model parameters in response to the ever-changing financial landscape. This approach allows for the immediate identification of prediction errors and misclassifications, facilitating a more nuanced understanding of model behavior and performance.

The table below provides a detailed comparison of the machine learning algorithms considered in this study, highlighting their complexity, accuracy, computational efficiency, parameter count, distinguishing features, suitable data types, and potential application scenarios within the realm of financial risk warning.

Table 1. Comparison of Machine Learning Algorithms

Algorithm Type	Algorithm Name	Model Complexity	Prediction Accuracy	Computational Efficiency	Number of Parameters	Algorithm Features	Suitable Data Types	Application Scenarios
Supervised Learning	Support Vector Machine (SVM)	Medium-High	High	Medium	Many	Maps to high-dimensional space, suitable for nonlinear classification problems	Small-scale, high-dimensional data	Financial risk binary warning
	Decision Tree	Low	Medium	High	Few	Easy to understand, low data preprocessing requirements	Various types of data	Financial risk factor analysis
	Random Forest	Medium	High	Medium	Medium	High fault tolerance, strong generalization ability	Various types of data	Financial risk level prediction
	Linear Regression	Low	Medium	High	Few	Simple calculation, strong interpretability	Continuous variables	Financial condition trend analysis
Deep Learning	Convolutional Neural Network (CNN)	High	High	Low	Many	Local feature extraction, parameter sharing	Image, time series data	Financial time series analysis
	Long Short-Term Memory Network (LSTM)	High	High	Low	Many	Solves long sequence dependency problems, strong memory capability	Time series data	Long-term financial condition prediction
Bayesian Methods	Naive Bayes	Low	Medium	High	Few	Based on probability framework, simple calculation	Discrete data	Financial fraud detection
Ensemble Learning	AdaBoost	Medium	High	Medium	Medium	Error-guided learning, model reinforcement	Binary classification, multi-class classification	Early warning of financial crisis
	Gradient Boosting	Medium	High	Medium	Medium	Iterative decision tree modeling, strong error correction	Continuous variables, discrete variables	Prediction of different financial risk levels

4. Methodology for Model Construction

In the construction of machine learning models for financial risk warning, feature selection and data preprocessing are pivotal steps that ensure the accuracy and generalization capability of the

model. This study incorporates a comprehensive analysis of prediction errors and implements a real-time assessment mechanism to enhance the model's adaptability and robustness in the face of dynamic financial data.
Feature selection is based on correlation coefficients and feature importance, focusing on indicators such as "current ratio," "quick ratio," and "debt-to-asset ratio" that exhibit high correlation and are rated as moderately important or important. These indicators are selected as input features for subsequent model training. To maintain the integrity of financial data, various strategies for handling missing values are employed, including median filling, mean filling, and ignoring missing values. Outlier treatment strategies, such as the 3σ principle or the Interquartile Range (IQR), are utilized to bolster the model's resilience against abnormal risks. Data normalization is also performed to standardize the scale of features.

Table 2. Feature Selection Results

Feature Indicator	Data Type	Missing Value Treatment	Outlier Treatment	Normalization Method	Feature Importance	Correlation Coefficient	Selected
Current Ratio	Continuous	None	3σ Principle	Min-Max Standardization	High	0.65	Yes
Quick Ratio	Continuous	Median Filling	IQR Method	Min-Max Standardization	Moderately High	0.59	Yes
Debt-to-Asset Ratio	Continuous	None	3σ Principle	Z-Score Standardization	High	-0.72	Yes
Debt to Equity Ratio	Continuous	Mean Filling	IQR Method	Z-Score Standardization	Moderate	-0.67	No
Inventory Turnover Ratio	Continuous	None	3σ Principle	Min-Max Standardization	Low	0.25	No
Total Asset Turnover Ratio	Continuous	None	IQR Method	Z-Score Standardization	High	0.38	Yes
Profit Margin	Continuous	Median Filling	IQR Method	Z-Score Standardization	Moderately High	0.41	Yes
Main Business Profit Margin	Continuous	None	3σ Principle	Min-Max Standardization	High	0.69	Yes
Cash Flow Ratio	Continuous	None	3σ Principle	Min-Max Standardization	Moderate	0.53	Yes
Expense Ratio	Continuous	Mean Filling	IQR Method	Z-Score Standardization	Low	-0.22	No
Net Profit Margin	Continuous	Median Filling	3σ Principle	Min-Max Standardization	High	0.76	Yes
R&D Expense Ratio	Continuous	Mean Filling	IQR Method	Z-Score Standardization	Moderate	0.47	No
Return on Equity	Continuous	None	IQR Method	Min-Max Standardization	Moderately High	0.52	Yes
Tangible Asset Ratio	Continuous	None	3σ Principle	Z-Score Standardization	Moderate	-0.34	No
Financial Leverage Ratio	Continuous	Median Filling	IQR Method	Z-Score Standardization	Moderate	-0.49	Yes
EBITDA Interest Coverage	Continuous	Mean Filling	3σ Principle	Min-Max Standardization	Low	0.28	No

After refining the financial data, appropriate feature variables are selected to form the input vector X={X1, X2, ..., Xn} of the research model. The model design is based on algorithms such as logistic regression, decision trees, and support vector machines. The model structure is customized for specific financial environments and risk characteristics to enhance the model's adaptability and sensitivity to financial risk prediction.

When evaluating the model's accuracy, the financial risk warning model formula is used:

$$f(X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_n X_n)}}$$

(4-1) Financial Risk Warning Model Formula

The performance of different models is compared using various evaluation metrics such as AUC, accuracy, and recall rate. The final model structure that can optimally warn of financial risks is determined through horizontal and vertical comparative analysis results. This study also introduces a real-time assessment mechanism to continuously monitor model performance and make dynamic adjustments, ensuring the model remains effective in predicting financial risks in real-time.

5. Model Comparison and Analysis

When constructing a machine learning financial risk warning model, evaluating the model's performance is crucial. This study employs a comprehensive set of evaluation metrics, including accuracy, precision, recall, F1 score, and AUC, to assess model performance. Additionally, a real-time assessment mechanism is integrated to monitor model performance continuously and facilitate prompt adjustments in response to changing market conditions.

The univariate model has a short training time but is less comprehensive and accurate, achieving a correctness rate of 61.3%. The Z-score model improves the correctness rate to 75.6%, although this comes with a slight increase in model parameters and training time. The performance of the Logistic regression model is significantly enhanced, with a correctness rate of 83.5%, alongside improvements in both sensitivity and specificity, despite an increase in training time. The BP artificial neural network model boasts an impressive correctness rate of 89.4%, demonstrating powerful performance; however, it requires substantial computational resources and has a long training time. The entropy weight method is straightforward in calculation and stable in weight configuration, yet its predictive accuracy does not match that of the Logistic regression and BP neural network models.

Table 3. Model Performance Comparison

Model Name	Correctness Rate (%)	Sensitivity (%)	Specificity (%)	AUC Value	Training Time (s)	Model Parameters	Real-time Adaptability
Univariate Model	61.3	58.7	63.9	0.65	1.2	None	Low
Z-Score Model	75.6	73.1	78.2	0.79	2.4	Weights: [1.2, -1.4, 3.3, 0.6]	Medium
Logistic Regression Model	83.5	80.2	86.8	0.88	5.8	Hyperparameters: C=0.01, penalty='l2'	High
BP Neural Network Model	89.4	87.9	90.7	0.92	120.5	Layers: 3, Neurons: [64, 64, 1], Learning rate: 0.001	Medium
Entropy Weight Method	81.7	79.3	84.1	0.85	3.7	Weights: Derived from data	Low

In this analysis, the conclusions column provides a brief summary of the advantages and disadvantages of each model based on the performance metrics, while also highlighting the importance of real-time adaptability in financial risk assessment.

6. Conclusion

This paper presents a financial risk warning model constructed using machine learning techniques and a comparative analysis of its performance. The empirical research conducted reveals that machine learning models, as compared to traditional statistical methods, demonstrate superior accuracy and stability in predicting corporate financial risks. Notably, the Support Vector Machine (SVM) model and the Random Forest (RF) model achieved prediction accuracies of 85.6% and 87.2%, respectively, which are significantly higher than the 78.4% accuracy of the Logistic Regression model. These findings suggest that machine learning algorithms are more adept at capturing the intricate nonlinear relationships present in corporate financial data, thereby enhancing the efficacy of risk warnings.

Moreover, this study delved into the impact of feature selection on model performance. Leveraging methods such as the Pearson correlation coefficient and Lasso regression, we identified the 10 most critical indicators for predicting corporate financial risks, including the debt-to-asset ratio, current ratio, and quick ratio. The machine learning model, with feature selection, showed an average increase of 3.5 percentage points in prediction accuracy compared to models without feature selection. This outcome underscores that strategic feature selection not only reduces model complexity and enhances computational efficiency but also significantly boosts the model's predictive capabilities.

In addition to the above, this study integrated a real-time assessment mechanism to monitor model performance continuously. This mechanism is pivotal for adapting to the dynamic shifts in financial markets, ensuring that the model remains robust and responsive to emerging risks. The implementation of this mechanism further solidifies the model's reliability in providing timely financial risk warnings.

References

[1] Liu, M., Gao, R., & Fu, W. (2021). Analysis of Internet Financial Risk Control Model Based on Machine Learning Algorithms. Hindawi Journal of Mathematics, 2021: 8541929-1–8541929-10.

[2] Xu, Z. H. (2021). Research on the Financial Risk Early Warning Model Based on Improved SMOTE (Master's Thesis, South China University of Technology). Master's Thesis. https://link-cnki-net-s.webvpn.zjhzcc.edu.cn/doi/10.27151/d.cnki.ghnlu.2021.002534 doi:10.27151/d.cnki.ghnlu.2021.002534.

[3] Bo Yang & Xinfei Wang. (2019). Construction of Logistics Financial Risk Ontology Model Based on Risk Correlation. In Proceedings of 2019 International Conference on Modeling, Simulation, Optimization and Numerical Techniques (SMONT 2019) (pp. 64-68). School of Information Management, Jiangxi University of Finance and Economics.

[4] Li, X. C., Wang, Q. L., & Lin, G. Q. (2020). Comparative Study on Financial Early Warning Models of Listed Companies Based on Machine Learning Methods. Market Modernization (07), 150-152. doi:10.14013/j.cnki.scxdh.2020.07.070.

[5] Zhu, L. Y. (2019). Support Vector Machine for Non-life Insurance Company Bankruptcy Prediction Based on Feature Engineering Application (Master's Thesis, Southwestern University of Finance and Economics). Master's Thesis. https://link-cnki-net-s.webvpn.zjhzcc.edu.cn/doi/10.27412/d.cnki.gxncu.2019.000490 doi:10.27412/d.cnki.gxncu.2019.000490.

[6] Zhang, C., Li, M., & Li, Y. (2021). Financial Risk Analysis of Real Estate Bubble Based on Machine Learning and Factor Analysis Model. Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology, 40(4), 6493-6504.

[7] Ge, H. (2020). Research on the Financial Risk Early Warning of Listed Companies Based on CART Decision Tree Model (Master's Thesis, Shanghai University of Finance and Economics). Master's Thesis. https://link-cnki-net-s.webvpn.zjhzcc.edu.cn/doi/10.27296/d.cnki.gshcu.2020.000851 doi:10.27296/d.cnki.gshcu.2020.000851.

[8] Li, S. Y. (2022). Financial Risk Identification Based on Financial Analysis Methods. Economic Research Guide (14), 129-131.

[9] Chen, J. (2020). Research on Corporate Financial Early Warning Based on Machine Learning and Text Analysis (Master's Thesis, Shanghai Jiao Tong University). Master's Thesis. https://link-cnki-net-s.webvpn.zjhzcc.edu.cn/doi/10.27307/d.cnki.gsjtu.2020.002723 doi:10.27307/d.cnki.gsjtu.2020.002723.

[10] Zhang, C. N., Ren, A. Z., & Lin, Z. D. (2023). Financial Risk Early Warning of Listed Companies—Based on Machine Learning Methods. SME Management and Technology (16), 182-184.

[11] Wang, J. M., & Zou, J. W. (2023). Research and Practice on Automated Early Warning System for Commercial Bank Financial Risk. China Financial Computing (11), 77-80.

[12] Feng, J. X. (2021). Research on Financial Fraud Identification Model Based on Machine Learning (Master's Thesis, Nanjing University). Master's Thesis. https://link-cnki-net-s.webvpn.zjhzcc.edu.cn/doi/10.27235/d.cnki.gnjiu.2021.000767 doi:10.27235/d.cnki.gnjiu.2021.000767.

[13] Zhang, L., Liu, J. P., & Jiang, M. Q. (2021). Application of Integrated Machine Learning Models in Financial Early Warning of Unbalanced Samples. Application of Electronic Technology (08), 34-38.

doi:10.16157/j.issn.0258-7998.201234.

[14] Jiang, Q. K., & Wang, C. Z. (2023). Research on Financial Risk Early Warning of Internet Enterprises' Overseas Mergers and Acquisitions Based on Stacking Ensemble Learning. Technical Economy (09), 147-160.

[15] Chen, J. F., Yu, J., Li, X. Z., & Fan, C. (2023). Financial Investment Risk Early Warning Method for Large Enterprises Based on Machine Learning. Market Outlook (04), 49-51.

[16] Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, 2(1), 37-63.