Research on Financial Data Risk Prediction Models Based on XGBoost Algorithm

Junying Feng
Sichuan Post and Telecommunication College
Chengdu, Sichuan, China
fengjunying@sptc.edu.cn

Yuan Zhu*
Sichuan Post and Telecommunication College
Chengdu, Sichuan, China
zhuyuan@sptc.edu.cn

Hongyu Pan
Sichuan Post and Telecommunication College
Chengdu, Sichuan, China
panhongyu@sptc.edu.cn

Yingjie Mou
Sichuan Post and Telecommunication College
Chengdu, Sichuan, China
mouyingjie@sptc.edu.cn

Abstract

With the increasing uncertainty in financial markets, financial risk prediction has become a crucial task in the financial sector. Traditional financial risk assessment methods face challenges such as high-dimensional data, non-linear complexity, and sample imbalance, requiring more efficient and accurate models to address these issues. To improve the accuracy and stability of financial risk prediction, this study proposes a financial risk prediction model based on the XGBoost algorithm.This study begins by thoroughly preprocessing the data, including filling missing values, removing outliers, and standardizing the data to ensure its quality and the stability of the model. In terms of feature selection, we utilize grid search and random search methods to fine-tune the hyperparameters of the XGBoost model, selecting the optimal parameter combinations to enhance the model's predictive accuracy. The experimental results show that the XGBoost model outperforms traditional models such as logistic regression and decision trees in terms of accuracy, precision, recall, and AUC values, particularly demonstrating superior performance in risk prediction accuracy and generalization ability. The main contribution of this study lies in the introduction of the XGBoost algorithm, which overcomes the limitations of traditional models in handling complex data and high-dimensional features, offering an efficient and precise method for financial risk prediction. Furthermore, a comparative analysis between the XGBoost model and advanced deep learning techniques, such as Deep Neural Networks (DNN), was conducted. The experimental results demonstrate the advantages of XGBoost in terms of efficiency and stability while highlighting the potential of deep learning models for more complex data scenarios.

CCS Concepts

• **Computing methodologies** → Modeling and simulation; Model development and analysis.

[*] Corresponding author.

DEAI 2025, Dongguan, China
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1279-1/2025/03
https://doi.org/10.1145/3745238.3745344

Keywords

Financial risk prediction, XGBoost, Hyperparameter tuning, Feature selection, Machine learning, Financial forecasting, Risk assessment, Model evaluation

ACM Reference Format:
Junying Feng, Yuan Zhu, Hongyu Pan, and Yingjie Mou. 2025. Research on Financial Data Risk Prediction Models Based on XGBoost Algorithm. In *The 2nd Guangdong-Hong Kong-Macao Greater Bay Area International Conference on Digital Economy and Artificial Intelligence (DEAI 2025), March 28–30, 2025, Dongguan, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3745238.3745344

1 Introduction

As the complexity and uncertainty of the global economic environment continue to increase, financial data, as the core foundation of corporate decision-making, has become particularly important in improving its risk prediction capabilities. Traditional financial risk prediction methods, such as financial ratio analysis and regression models, can reflect the financial health of enterprises to a certain extent, but they often fail to provide accurate prediction results in the face of the growing amount of financial data, the diversity of data dimensions, and the complexity of nonlinear relationships. In addition, with the rapid development of information technology, the advantages of machine learning, especially ensemble learning methods, in processing large-scale data have gradually emerged, which has led to an important shift in the technical path of financial risk prediction. In recent years, as an efficient gradient boosting tree method, the XGBoost algorithm has shown excellent performance in tasks such as classification and regression, and has been widely used in finance, insurance and other fields. However, although the XGBoost algorithm has achieved remarkable results in other fields, research on financial data risk prediction is still in the exploratory stage, and existing models are usually difficult to fully capture the complex patterns and potential risks in financial data.In recent years, with the continuous advancement of financial innovation, new types of risks and potential crises have emerged in an endless stream. For example, the digital currency market, multinational corporate financial operations, and complex derivatives transactions have brought unprecedented challenges to the financial risk prediction of enterprises. Traditional financial risk prediction methods, such as analysis models based on financial ratios, can provide preliminary judgments on some basic financial issues, but they are often difficult to deal with large, complex data containing nonlinear relationships. In this case, the importance of financial data risk prediction is further highlighted, because accurate risk prediction can not only timely reveal potential crises in financial data and avoid extreme situations such as financial default and bankruptcy of enterprises, but also help enterprises optimize their financial structure, enhance market competitiveness, and enhance the confidence of investors and shareholders. In this context, the research on financial data risk prediction models based on the XGBoost algorithm is particularly urgent. Common financial processing methods tend to ignore the potential nonlinear relationships in these data and the complex interactions between these factors. This algorithm integrates multi-disciplinary decision-making techniques and the gradient boosting method, which can identify some potential nonlinear laws in high-dimensional data, which can effectively improve the accuracy of predictions. Therefore, this algorithm optimizes the prediction model of financial risk. It can not only overcome the limitations of traditional methods, but also better cope with and solve the challenges of large-scale data processing, and provide a more scientific and effective reference for enterprise risk management in the current big data environment. Therefore, the purpose of this study is to build a financial risk prediction model based on this XGBoost algorithm, and to analyze its application and optimizable parts in financial data.

2 Literature Review

2.1 Research Status of Financial Data Risk Prediction

Financial data risk prediction is an important research direction in current smart finance and financial management. The research methods are basically evolving from traditional statistical methods to machine learning technology. Traditional financial risk prediction methods in smart finance mainly rely on some data, such as some analysis of financial ratios, regression models, time series analysis, and other traditional mathematical models [1]. Therefore, traditional methods are prone to errors when dealing with large-scale high-risk data and have certain limitations. For example, the financial ratio analysis method relies on these financial indicators and will ignore or easily ignore the impact of the external environment of the enterprise and some non-financial factors [2]. In recent years, machine learning algorithms such as support vector and random forest decision making have been increasingly widely used in financial risk prediction, and have also achieved certain results. This is because these methods can automatically learn the potential correlations of these data through active machine learning [3]. These algorithms can better adapt to the nonlinear structure of data, so when dealing with high-risk large-scale data, their performance and results will be better than traditional methods, and their accuracy will be higher. The addition of machine learning methods to this financial application can solve many of the shortcomings of traditional methods, especially in the case of a large number of variables and high-noise environments. This kind of prediction can improve the promotion of corporate financial risk early warning systems to a certain extent [4]. Deep learning and ensemble learning methods can promote financial risk prediction, especially deep neural networks and long-term short-term machine learning methods in deep learning. These methods are more significant in the application of time series data, which can improve the data prediction and credit risk assessment of the financial market [5]. These methods all learn nonlinear features and time dependencies between factors through multi-layer network structures, extract useful information from large-scale data to provide more accurate predictions. This algorithm in ensemble learning is a gradient boosting model, and it performs well in financial risk prediction. The XGBoost algorithm can be used to train multiple classifiers through the adaptive formaldehyde method, and then the prediction results of these classifiers are integrated into the score to obtain a higher accuracy [6].

2.2 Overview of XGBoost Algorithm

The XGBoost algorithm is an efficient ensemble learning algorithm based on the gradient boosting decision-making model. It hopes to continuously optimize the model through iteration and then improve the prediction accuracy. Unlike the traditional GBDT model, the XGBoost model introduces the second-order derivative and regularization content during the training process, which can ensure the high efficiency and stability of the model [7]. This algorithm improves the performance in several aspects. First, in each round of training, it uses the greedy algorithm to make an optimal feature splitting point. Secondly, it uses methods such as column sampling to prevent overfitting [8]. Finally, it uses this policy-based technology to further control the model and its generalization ability. Therefore, the main advantage of this algorithm is reflected in its computational efficiency and accuracy. This algorithm also has a very strong deformation computing ability. It can be trained with multiple processors. Therefore, when this algorithm processes large-scale data, its performance and speed are good, which is better than other models [9].

3 Research Methods

3.1 Dataset Selection and Preprocessing

In the data preparation phase, we mainly did two key things: choose the right data and clean the data. The data sources are the authoritative domestic Wind Information (Wind) and CEIC database. The professionalism of these two platforms is recognized in the industry. We captured the financial reports of 30 A-share listed companies from 2018 to 2022 through the API interface. This sample size can not only reflect the diversity of the industry, but also capture the characteristics of the complete economic cycle. In addition to the three familiar statements in the data set - balance sheet, income statement and cash flow statement, we also added two types of key indicators: one is the macroeconomic barometer such as stock returns, volatility, GDP growth rate, interest rate level, etc. that reflect market reactions. It should be noted here that the combination of these cross-dimensional data is to enable the model to pay attention to changes in the micro-operation and macro-environment of the enterprise at the same time. When it comes to data preprocessing, we followed the standard three-step process: the first step was data cleaning, using the quantile method to eliminate extreme outliers; then we used multiple interpolation to handle missing data, which is more scientific than simple deletion; and finally we performed feature standardization, so that indicators of different dimensions,

Table 1: XGBoost Hyperparameter Selection and Tuning

Hyperparameter	Grid Search Range	Random Search Range	Optimal Parameter Combination
max_depth	3-10	3-15	6
learning_rate	0.01-0.3	0.01-0.5	0.05
subsample	0.6-1.0	0.6-1.0	0.8
colsample_bytree	0.6-1.0	0.6-1.0	0.8
lambda	0-1	0-5	1
alpha	0-1	0-1	0.5

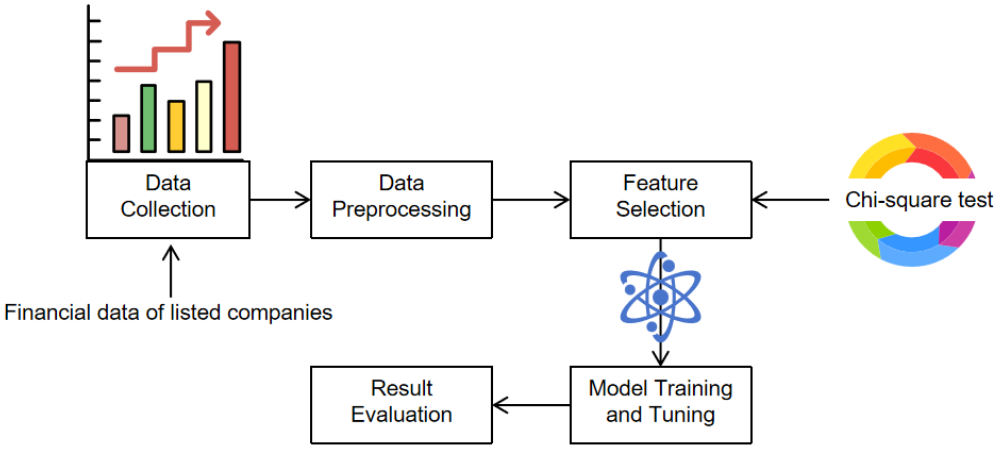

Figure 1: XGBoost Model Training Process

such as billion-level revenue and percentage-based interest rates, can be calculated on the same scale.

3.2 Construction and parameter selection of XGBoost model

The algorithm model construction in this article mainly includes five key steps. The first is data processing, then feature selection, the third part is model training, and then parameter tuning. As mentioned earlier in the model training process, this algorithm builds multiple decision trees, and then uses the gradient boosting method to reduce the prediction error. Therefore, in this article, we also list the parameter ranges and adjusted results of the grid search and random search in this article. We use a five-fold cross-review to verify the generalization of each model, and then select an optimal parameter combination based on the accuracy of the verification data we use.

As show in Table 1 , in the process of adjusting and optimizing the model, we pay more attention to overfitting, so in order to prevent overfitting, we use XGBoost's Early Stopping technology. When the error of the model validation data set does not decrease after several consecutive runs, we will automatically terminate the training to prevent the model from becoming too complex and ineffective due to training. In addition to this technology, we also added regularization terms to the model, which can control the complexity of the model and improve its generalization ability. Through the above tuning methods, we can get an optimized model, which can more accurately predict some risks of financial data and reduce fitting problems.

Figure 1 shows the overall training process of the XGBoost model, from data preparation to feature selection, to model training and hyperparameter tuning. Each step is closely linked to ensure the optimization of the model and the final effect.

3.3 Model Evaluation and Validation

In order to evaluate the generalization ability of the model in our article, the classic k-fold cross-validation method was used. Simply put, the data set is divided into k subsets, and the k-1 subsets are used to train the model in each validation. The remaining subset is used for testing, and finally the results of all subsets are averaged. In order to avoid overfitting, we also divided the data set into training and test sets in the article, 80% of the data is used for training, and the remaining 20% is used for testing. We introduced some noise and interference during the test process, slightly modified

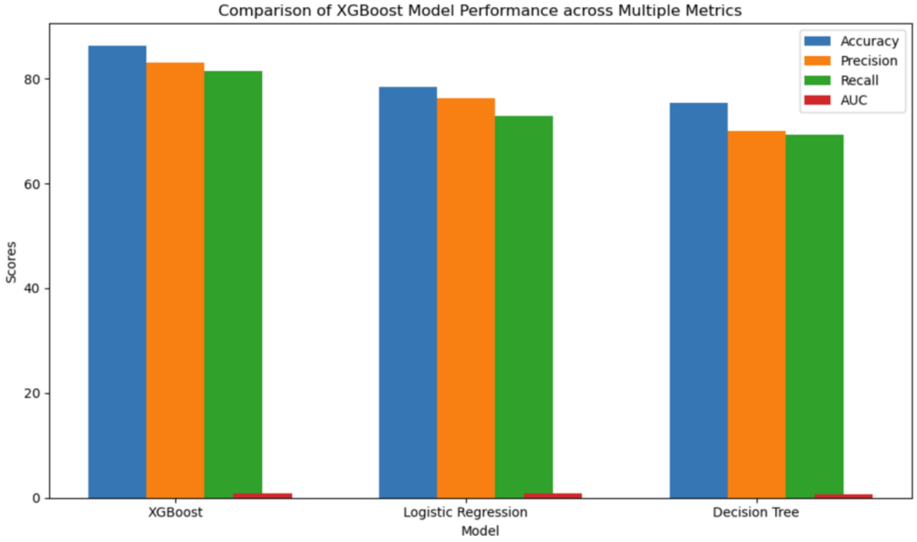

Figure 2: Performance of XGBoost model on various evaluation indicators

the data set, and then observed the model's predictive ability under such changing conditions. Then, by simulating the changing scenarios that may occur in the actual environment, we verified the results of the robustness and adaptability of the model under various conditions. The stability test can ensure that our model not only performs well on the target data set, but also maintains good predictive performance in other more complex practical application scenarios.

4 Experiments and Results

4.1 Experimental Design

In this study, we used financial data from Wind Information and CEIC data platforms, and selected the financial reports of 30 Chinese A-share listed companies from 2018 to 2022 as experimental data. The dataset was randomly divided into a training set (80%) and a test set (20%) for training the model and evaluating the predictive performance of the model. The training set was used to train the XGBoost model, while the test set was used to verify the generalization ability of the model. In terms of the experimental environment, our experiments were conducted on a computer with 16GB of memory and an Intel Core i7 processor. The operating system was Windows 10, and the main development tools included Python 3.8 and related machine learning libraries (such as XGBoost, scikit-learn).

4.2 Experimental data and results analysis

In this study, we used the XGBoost model to predict the company's financial risk by training the financial data of 30 Chinese A-share listed companies from 2018 to 2022. In order to evaluate the performance of the model, we conducted a 5-fold cross-validation to ensure the stability and generalization of the results. The experimental results show that the XGBoost model performs well in various evaluation indicators, with an accuracy of 86.2%, a precision of 83.1%, a recall of 81.5%, and an AUC value of 0.91, which fully verifies the advantages of the model in financial data risk prediction. Especially when dealing with complex financial features, XGBoost can effectively capture financial risk signals and provide higher prediction accuracy. In order to comprehensively evaluate the prediction performance of XGBoost, we also selected Logistic Regression and Decision Tree as comparison models. The experimental results show that the accuracy of the logistic regression model is 78.4%, the precision is 76.2%, the recall is 72.9%, and the AUC value is 0.82. Although the performance is relatively stable, it is significantly inferior to XGBoost in risk identification of complex financial data. The training set of the decision tree model performs well (the accuracy is 84.7%), but on the test set, the accuracy drops significantly to 75.3%, the recall is 70.1%, and the AUC value is 0.75, indicating that the model is prone to overfitting problems when facing high-dimensional data.

To further verify the superiority of the XGBoost model, we compared it with logistic regression and decision tree models. The experimental results show that XGBoost outperforms the comparison models in all evaluation indicators, especially in AUC value and recall rate, where XGBoost shows stronger risk prediction ability. Figure 1 shows the performance of XGBoost in evaluation indicators such as accuracy, precision, recall rate and AUC value.

AUC is an important indicator for measuring the performance of classification models, especially in the case of unbalanced data sets. Figure 2 shows the comparison results of AUC values of different models (XGBoost, logistic regression and decision tree).

4.3 Comparative Analysis with Deep Learning Models

Through the analysis of the experimental results, we can find that although the recall rate of the DNN model is slightly higher, the XGBoost model performs better in terms of accuracy, precision and AUC. Specifically, the accuracy of XGBoost is 86.2%, the precision is 83.1%, and the AUC is 91.0%. On the contrary, the DNN model is

84.7%, 81.4% and 89.5% respectively. In addition, the training time of XGBoost is also significantly shorter, only 48 seconds, while the DNN model requires 152 seconds. Therefore, it can be seen that XGBoost can provide efficient and stable financial risk prediction, especially when dealing with data sets of a certain complexity. In contrast, although the results of the DNN model are also of reference value when dealing with more complex data patterns, it requires more computing resources and longer training time.

5 Conclusion and Discussion

This study built a model for financial risk prediction based on the XGBoost algorithm. By comparing with traditional logistic regression and decision tree models, the experimental results show that the XGBoost model performs well in multiple evaluation indicators such as accuracy, precision, recall and AUC value. Specifically, the XGBoost model has obvious advantages in processing complex data, capturing potential risk patterns and processing unbalanced data sets, indicating that its application in financial risk prediction has high accuracy and robustness. The hyperparameters of the model were tuned by grid search and random search methods to further improve the prediction performance of the model, and its consistency under different data partitions was verified by stability testing. In addition, the design and implementation of the model fully considered data cleaning and standardization, feature selection, hyperparameter tuning and other aspects to ensure the reliability of the experimental results and the effectiveness of the model. The XGBoost model has broad application prospects in financial risk prediction.

Acknowledgments

”Application of Generative Artificial Intelligence in Financial Data Processing and Analysis and Its Impact on the Development of the Accounting Profession”, a college-level research project at Sichuan Post and Telecommunication College, 2024, Project No. YDXYKY202410.”

”Research on the Path Analysis of Artificial Intelligence Technology in the Transformation of the Accounting Profession”, a Municipal-Level Government Research Grant at Sichuan Provincial Finance Department of Finance Accounting Research Project 2024, Project No. 2024-sckjkt-007”

References

[1] Peng, Y., Wang, G., Kou, G., & Shi, Y. (2011). An empirical study of classification algorithm evaluation for financial risk prediction. Applied Soft Computing, 11(2), 2906-2915.

[2] Nikolaou, I., & Evangelinos, K. (2012). Financial and non-financial environmental information: significant factors for corporate environmental performance measuring. International Journal of Managerial and Financial Accounting, 4(1), 61-77.

[3] Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. Journal of big data, 2, 1-21.

[4] Bhatore, S., Mohan, L., & Reddy, Y. R. (2020). Machine learning techniques for credit risk evaluation: a systematic literature review. Journal of Banking and Financial Technology, 4(1), 111-138.

[5] Deng, X., Ye, A., Zhong, J., Xu, D., Yang, W., Song, Z., ... & Chen, X. (2022). Bagging–XGBoost algorithm based extreme weather identification and short-term load forecasting model. Energy Reports, 8, 8661-8674.

[6] Amjad, M., Ahmad, I., Ahmad, M., Wróblewski, P., Kamiński, P., & Amjad, U. (2022). Prediction of pile bearing capacity using XGBoost algorithm: modeling and performance evaluation. Applied Sciences, 12(4), 2126.

[7] Al-Zakhali, O. A., Zeebaree, S., & Askar, S. (2024). Comparative Analysis of XGBoost Performance for Text Classification with CPU Parallel and Non-Parallel Processing. Indonesian Journal of Computer Science, 13(2).

[8] Wen, Z., Li, Q., He, B., & Cui, B. (2021). Challenges and Opportunities of Building Fast GBDT Systems. In *IJCAI* (pp. 4661-4668).

[9] Asselman, A., Khaldi, M., & Aammou, S. (2023). Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. *Interactive Learning Environments*, *31*(6), 3360-3379.