Vol. 19 (2026): Proceedings of 2026 International Conference on Financial Technology, Mathematical Statistics, and Business Analytics (ICFMB 2026)
E-ISSN: 3105-0522 P-ISSN: 3105-0514
DOI: https://doi.org/10.70267/icfmb.2026196167

Predicting Bond Default Risk Based on the XGBoost Model

Shurui Xie*

School of Finance, Shandong University of Finance and Economics, Jinan, China

**Corresponding author: Shurui Xie.*

Abstract

At present, China faces a large scale of bond defaults. To address the problems of low prediction accuracy and limited indicator systems in traditional models, this paper conducts research on bond default risk prediction based on the XGBoost model. This study selects 616 valid samples of credit bonds from entity industries in the Wind database from 2019 to 2024, and constructs a multi-dimensional indicator system covering bond characteristics, financial leverage, and other factors. After data cleaning and feature selection, the dataset is divided into training and test sets at a 7:3 ratio, and the XGBoost model is trained using 3-fold cross-validation. The results show that the sustainability of corporate profitability is the core influencing factor of default. The model achieves an AUC of 0.997 on the test set, with no severe overfitting, demonstrating excellent generalization ability and prediction accuracy. This provides methodological support for the prevention and control of bond default risk.

Keywords

bond default, XGBoost, default prediction

1. Introduction

With the continuous development of the bond market, the scale of bond defaults has gradually expanded. Bond default prediction plays a crucial role in controlling or reducing the scale of defaults. Accurate prediction of default risk can help maintain financial market stability and, to a certain extent, reduce investment losses. In today's diversified financial market, single traditional economic models (such as Z-score and Merton models) struggle to achieve precise default risk prediction and exhibit weak generalization ability. Wang Lisha et al. constructed a Stacking heterogeneous ensemble model that integrates the advantages of various base models, achieving good predictive performance on listed companies' bond default data, with an overall AUC improvement of 0.7% to 11.3% [1]. This result highlights the effectiveness of machine learning models in improving the accuracy of bond default prediction. Therefore, how to better leverage machine learning models for bond default prediction has become critically important.

Since the "Chaori Bond" incident in 2014 broke the implicit guarantee of rigid payment, the number of bond defaults in China showed a gradually rising trend from 2014 to 2019, reaching a peak of 144.9 billion yuan in 2019. It then declined briefly, falling to 31.819 billion yuan in 2023. However, with the explosive growth of bond extensions, the form of corporate bond default risk exposure has shifted from substantive default to extension. In 2024, the default scale rebounded sharply to a high of 71.929 billion yuan [2]. This indicates that bond defaults involve large scales and amounts, which have a certain impact on investors and market stability, necessitating efforts to avoid bond default risks as much as possible. Currently, academia has developed various default prediction methods, with mainstream approaches including traditional statistical models, machine learning algorithms, deep learning optimization, and hybrid models (balancing interpretability and accuracy). However, the indicator systems in existing studies still have limitations: on the one hand, there is insufficient coverage of management risk attitudes and unstructured textual information. Structured indicators have long update cycles and respond relatively slowly, making it difficult to capture changes in corporate risk conditions in a timely manner. In contrast, high-frequency textual information containing professional insights and market expectations is more forward-looking [3]. On the other hand, macroeconomic indicators mostly use provincial-level data, which lacks variation in values. Therefore, a bond default prediction indicator system must simultaneously ensure information sufficiency and differentiation to provide a more reliable predictive basis for the model [4].

Addressing the above gaps, this paper constructs a multi-dimensional bond default risk prediction indicator system that reflects corporate finance, capital structure, liquidity levels, and credit ratings, and employs the XGBoost model for bond default prediction. The marginal contributions of this paper are as follows: first, it conducts feature importance analysis on the sample to identify the key factors triggering bond defaults; second, through multiple sets of data experiments, it demonstrates the model's advantages and accuracy in default early warning. This provides investors and regulatory authorities with a more precise default risk prediction tool and offers new ideas and empirical support for solving current credit risk assessment challenges.

2. Main Influencing Factors of Bond Default Risk and Prediction Content

Bond defaults are caused by multiple factors; therefore, a comprehensive consideration of multi-dimensional influencing factors is essential when conducting bond default prediction. From the perspective of internal corporate management, first, the basic influencing factors of an enterprise should include its financial condition, such as the asset-liability ratio, corporate profitability, and liquidity, as well as management capability and industry position. Previous studies have shown that higher management capability not only serves as an important indicator of the effectiveness of corporate social responsibility but also effectively reduces bond default risk during periods of economic uncertainty [5]. These factors constitute important bases for assessing a company's debt repayment capacity. Second, the rigid strategic mindset of some enterprises prevents them from gaining a foothold in markets with rapid technological change. This is evidenced by the high incidence of defaults in the internet software services sector (accounting for 27.09% of default amounts) and the automobile manufacturing industry (15.23%). Due to insufficient compensation incentives and performance evaluation pressure during their tenure, management tends to maintain traditional business paths rather than pursue technological innovation, resulting in lagged strategic adjustments [6]. Third, high-quality corporate development can reduce bond default risk by optimizing the information environment and improving credit ratings [7].

From the perspective of the external economic environment, factors such as the macroeconomic cycle, interest rate levels, monetary policy, industry prosperity, and market liquidity significantly affect corporate operations and financing. For example, China's economic growth rate declined from 6.9% to 5.2% between 2015 and 2024, with both aggregate demand and capacity utilization contracting simultaneously. The average capacity utilization rate in the industrial sector fell to 75.8%. The macroeconomy has shifted from a high-speed growth phase to a medium-speed growth phase, entering a period of cyclical transition. The contraction in aggregate demand and the decline in capacity utilization are typical characteristics of an economic downturn, directly leading to substantial revenue declines in cyclical industries. This confirms that fluctuations in the macroeconomic cycle directly impact corporate fundamentals [6]. From the perspective of public opinion and behavior, non-structural information such as negative news sentiment, regulatory penalties, external evaluations, and changes in senior executives can reflect potential operational risks of enterprises and signal the default risk of their bonds. Yu Yunpeng's research also confirmed this point: he processed textual data from news and self-media through the BERT model to convert it into structured inputs capable of representing sentiment and viewpoints, thereby significantly improving the performance of prediction models [8].

Through comprehensive analysis of various influencing factors, corresponding prediction results are generated. The content of prediction usually focuses on assessing, within a specific time frame, the likelihood that the bond issuer will be unable to pay interest or principal on time and in full, as well as the potential losses if default occurs. This provides decision-making support for investors, financial institutions, and regulatory authorities, enabling them to avoid risks as much as possible.

3. Data and Variables

The research data in this paper are primarily sourced from the Wind database, with the time span focused on credit bonds issued and outstanding between 2019 and 2024, including corporate bonds, medium-term notes, directional instruments, and other types. Financial enterprises were excluded, with the focus placed on issuers in the real economy sectors to avoid interference from special capital structures. To ensure data integrity, samples with complete issuance information, financial indicators (such as asset-liability ratio and current ratio from the 2021–2024 annual reports), and default status were retained. For samples with partial missing values, industry median values were used for imputation, while samples with severe data missing were excluded. Ultimately, 616 valid samples were obtained, including 174 default samples. The samples cover multiple real economy sectors such as manufacturing, construction, and wholesale and retail, and span the full credit rating spectrum from AAA to C. This ensures sample diversity and rating representativeness, which helps improve the model’s generalization ability. Default is defined as a binary 0-1 dummy variable (1 = default, 0 = no default).

In terms of research variables, this paper sets whether default occurs as a binary 0-1 dummy variable, serving as the core dependent variable in the study. If a sample enterprise experiences a substantive default event during the study period, the default variable (Default) is assigned a value of 1. Substantive default includes the following scenarios: First, overdue and unpaid principal or interest on bonds, loans, or other debts issued by the enterprise, including principal overdue, interest overdue, or failure to repay after extension. Second, the enterprise is determined to be in debt default, or a default announcement is disclosed by exchanges, banks, or other financial institutions following a bond or credit default. Third, the enterprise is listed as a dishonest judgment debtor due to debt default, or shows signs of bankruptcy reorganization or liquidation. If a sample enterprise does not experience any of the above substantive default events during the study period, repays all debt principal and interest on time and in full, and has no default-related announcements or adverse credit records, the default variable (Default) is assigned a value of 0.

To more accurately identify the impact of explanatory variables on bond default and mitigate omitted variable bias and endogeneity issues, this paper draws on research in the field of bond default prediction and selects control variables from four dimensions: bond characteristics, financial leverage, debt repayment capacity, and profitability.

The control variables for bond characteristics are as follows:

Issue amount (in 100 million yuan), which measures the initial issuance scale of the bond. A larger issue amount implies more concentrated debt repayment pressure during the period and may also reflect the enterprise’s financing needs. It is expected to be positively correlated with bond default risk.

Bond balance, where a higher value indicates a larger scale of outstanding debt, heavier long-term debt burden, and increased cash flow pressure, expected to raise the probability of bond default.

Latest bond rating, which reflects the market’s comprehensive assessment of the bond’s repayment capacity. In the analysis, credit ratings are converted into ordinal numerical values: AAA/AAAsf=5, AA+/AA+sf=4.5, AA/AAsf=4, AA-/AA-sf=3.5, A+/A+sf=3, A/Asf=2.5, A-/A-sf=2, BBB+/BBB+sf=1.5, BBB/BBBsf=1, BBB-/BBB-sf=0.8, BB+/BB+sf=0.6, BB/BBsf=0.5, BB-/BB-sf=0.4, B+/B+sf=0.3, B/Bsf=0.2, B-/B-sf=0.1, CCC and below (including sf)/C/Csf=0. For samples with missing rating data, the default value is set to 2.5 (medium credit rating). Higher ratings indicate stronger repayment capacity and higher creditworthiness, corresponding to lower default risk; thus, they are expected to be negatively correlated with default risk.

For financial leverage and debt repayment capacity, the control variables include the asset-liability ratio, current ratio, their trend indicators (asset-liability growth rate and the decline in current ratio), as well as short-term debt and the ratio of short-term debt to monetary funds.

For profitability, the variables include return on assets (ROA) and the number of periods in which ROA is negative.

4. Model Construction

Ensemble learning methods possess strong predictive performance. Therefore, this study employs the XGBoost algorithm from the ensemble learning family to predict bond defaults based on the data. XGBoost belongs to the Boosting class of algorithms. Its core idea is to iteratively construct multiple weak learners (typically CART regression trees) and then linearly combine them into a strong learner. Unlike traditional GBDT, XGBoost introduces further improvements in the definition of the objective function, optimization strategy, and regularization mechanism.

Unlike traditional methods that use only first-order derivatives, the XGBoost objective function approximates the loss function using a second-order Taylor expansion, thereby more accurately locating the optimal split points. Among these, the L2 regularization mechanism effectively prevents model overfitting and improves generalization ability. During tree construction, XGBoost adopts a greedy algorithm to find the best split points. By transforming the objective function into a quadratic function with respect to leaf node scores, the optimal weights and the corresponding minimum objective function value can be directly derived. In practical computation, the algorithm traverses all features and their possible split points, selecting the feature that causes the largest reduction in the objective function for splitting.

Overall, XGBoost incorporates techniques such as second-order derivative optimization, regularization terms, and column sampling. It maintains high prediction accuracy while controlling model complexity. Its efficient engineering implementation enables the algorithm to run quickly on large-scale datasets, making it one of the most influential algorithms in the current machine learning field. Previous studies have also utilized the XGBoost model and achieved superior results compared to traditional economic models. For example, Li (2018) constructed an ensemble learning model with XGBoost as the main component to predict default risk in China's P2P lending, achieving higher prediction accuracy than traditional machine learning models [9]. Xiao Yanli (2021) built a combined early-warning model for bond default risk based on GWO-XGBoost, and the results showed that the GWO-XGBoost model exhibited superior prediction accuracy and stability [10].

Before the start of the research, the raw data related to bonds were cleaned. For missing values, quantitative indicator data were imputed using the median of the indicator, while categorical variables were imputed using the mode. To mitigate the impact of extreme values on the overall results, the IQR rule combined with boxplot visualization was used to identify outliers. Obvious extreme values, such as asset-liability ratios greater than 300% and short-term debt/monetary funds greater than 500%, were replaced with the median. After data cleaning, to improve the model's predictive performance and optimize the feature indicator system, Pearson correlation analysis was applied to eliminate the original indicators of interest-bearing liabilities, monetary funds, and the bond rating at issuance. Finally, Z-score standardization was performed on all continuous feature variables to eliminate scale differences between variables.

To ensure the effectiveness of model training and validation, the data were randomly split into training and test sets at a 7:3 ratio. Data shuffling was applied to eliminate ordering bias, and 3-fold cross-validation was used for hyperparameter tuning and stability verification. The base learner was set to gbtree (gradient boosted tree). Considering the bond default prediction scenario, the following parameters were determined: number of base learners=100, learning rate=0.1, L1 regularization term=0.5, L2 regularization term=2, sample feature sampling rate=0.8, tree feature sampling rate=0.8, node feature sampling rate=0.7, maximum tree depth=3, and minimum sample weight per leaf node=0. Through regularization constraints and multi-dimensional sampling settings, model overfitting was effectively suppressed, ensuring strong generalization ability. The entire model training process took 4.681 seconds, demonstrating high computational efficiency.

5. Empirical Results

The empirical results reveal the degree of influence of each indicator on bond default risk, i.e., feature importance. The XGBoost model identified the most important feature as “Number of years with negative ROA” (42.40%), indicating that the sustainability of corporate profitability is a key indicator for judging future default risk. Other important features include the 2024 asset-liability ratio (12.20%), the latest bond rating (13.70%), and the 2024 return on total assets (11.30%). The remaining features, such as current ratio-related indicators and bond balance, also contribute to the model’s prediction but with lower importance. Under the XGBoost model, corporate profitability sustainability and leverage levels are the primary bases for distinguishing credit quality.

Table 1: Performance Metrics

	Accuracy	Recall	Precision	F1	AUC
Training Set	0.995	0.995	0.995	0.995	1
Cross-Validation Set	0.963	0.963	0.963	0.963	0.99
Test Set	0.962	0.962	0.962	0.962	0.997

According to the XGBoost analysis of the indicator data, the model’s evaluation metrics on the training set, cross-validation set, and test set are shown in the table above. On the training set, accuracy, recall, precision, and F1 score all reached 0.995, with an AUC value of 1.000, indicating that the model perfectly fitted the data during the training phase. The metrics on the cross-validation set and test set are relatively similar. Accuracy is around 0.962–0.963, and the AUC values are 0.990 and 0.997, respectively. The close performance between the training and test sets strongly indicates that the model does not suffer from severe overfitting and possesses excellent generalization ability.

In summary, the XGBoost model demonstrates high training efficiency, with feature selection centered on profitability sustainability and leverage levels, which is reasonable. It also achieves extremely high accuracy and AUC values on the test set, indicating that the model has strong application value and reliability in real-world scenarios.

6. Conclusion

This paper takes relevant bond information from 2019 to 2024 as the research sample. After appropriate data cleaning and processing, 616 valid samples were obtained. The dependent variable in this study is whether default occurs, set as a 0-1 categorical variable. The explanatory variables were selected from four dimensions—bond characteristics, financial leverage, debt repayment capacity, and profitability—to construct an XGBoost-based bond default risk prediction system. Subsequently, data preprocessing was performed: missing values were imputed, outliers were removed using the IQR rule, features were screened using Pearson correlation analysis, and Z-score standardization was applied. During model training, the data were divided into training and test sets at a 7:3 ratio, and 3-fold cross-validation was used for model tuning and optimization. The bond default risk prediction results based on the XGBoost model show that feature importance analysis successfully identified key influencing factors. The high precision and AUC values reflect the model’s high computational efficiency and accurate prediction capability. This provides effective machine learning methodological support for bond default risk prediction and compensates for the weak discriminative power of traditional economic models.

For future bond risk prediction and prevention, it is recommended that investors utilize the precise predictions of the XGBoost model to optimize and adjust investment strategies. Regulatory authorities may incorporate the XGBoost model into their systems and conduct risk monitoring by combining financial indicators and bond characteristics. In subsequent research, the indicator system can be enriched by incorporating non-structural textual information such as corporate negative public sentiment and enterprise management information to make the information more comprehensive. In addition, comparative studies can be conducted using ensemble learning models such as Stacking or deep learning models such as LSTM, or by fusing the advantages of multiple algorithms to further improve prediction accuracy. Based on the bond default prediction system constructed in this paper, future studies can expand the research sample and time span, conduct dynamic analysis across different market cycles, and provide more targeted empirical references for credit risk prevention and control in the bond market.

References

[1] Wang, L. S. & Guo, S. J. (2025). Bond default prediction of listed companies based on Stacking ensemble. *Modern Information Technology*, 9(20), 171-177. https://doi.org/10.19850/j.cnki.2096-4706.2025.20.031.

[2] Zhu, L., & Du, X. Q. (2025). The impact of green innovation on corporate bond default risk. *Modern Finance*, (12), 48-56+29.

[3] Zhong, N. H., Hao, Y. T., & Liu, Y. Y. (2025). Research on early warning of real estate bond defaults based on unstructured text. *Journal of Sun Yat-sen University (Social Science Edition)*, 65(04), 359-374. https://doi.org/10.13471/j.cnki.jsysusse.2025.04.030.

[4] Gao, L. Y. (2025). Research on bond default prediction of listed companies based on LDA-XGBoost [Master's thesis, Dalian Jiaotong University]. https://doi.org/10.26990/d.cnki.gsltc.2025.000434.

[5] Suganda, T. R., & Kim, J. (2023). An empirical study on the relationship between corporate social responsibility and default risk: Evidence in Korea. *Sustainability*, 15(4), Article 3644. https://doi.org/10.3390/su15043644.

[6] Chen, R. (2025). Analysis of the causes of bond defaults in state-owned enterprises and research on improvement strategies. *Business Manager*, (06), 96-98.

[7] Li, C., & Qu, Y. H. (2025). Research on the relationship between high-quality corporate development and bond default risk. *Friends of Accounting*, (20), 22-31.

[8] Yu, Y. P. (2024). Research on bond default risk prediction model based on multi-source heterogeneous data [Master's thesis, Sichuan University]. https://kns.cnki.net/kcms2/article/abstract?v=OGEEvuKkrhYhOLT1kNnqxIBGlVdhkKYN4yn982zLSC5bohlBvUswyLG5VTVFLpy8DNMKowzr_cb4oZfSQHPJ-MrSoDMK8wO0h7IfY9MiBPNT8LfcLe-4BQbH2JMp8Dvl7Ow5o-ehs2Qx0NMfu86wnnqmmeLZTGSVkTZz-EuK1hg51kny_5NXGQ==&uniplatform=NZKPT&language=CHS.

[9] Li, W., Ding, S., Chen, Y., & Yang, S. (2018). Heterogeneous ensemble for default prediction of peer-to-peer lending in China. IEEE Access. Advance online publication. https://doi.org/10.1109/ACCESS.2018.2810864

[10] Xiao, Y. L., & Xiang, Y. T. (2021). Early warning of corporate bond default risk: Based on the GWO-XGBoost method. *Shanghai Finance*, (10), 44-54. https://doi.org/10.13910/j.cnki.shjr.2021.10.005

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.