

Research on Intelligent Early Warning of Corporate Bond Default Based on Machine Learning

Pengfang Gao^{1,*}, Nuolin Yu¹, Jingwen Gao²

¹ School of Science, Shandong Jianzhu University, Jinan 250101, China.

² College of Literature and History (College of Culture and Tourism), Weifang University, Weifang 261061, China.

* Corresponding author: Pengfang Gao (Email: encounter117026@163.com)

Abstract: This study aims to construct an efficient and accurate intelligent early warning model for corporate bond default with the help of machine learning tools. The study takes the real default data from 2017-2022 as samples, uses random undersampling and SMOTE oversampling to deal with the unbalanced data, and applies random forest for feature screening. The data covers the market conditions in recent years, includes the actual performance of defaulted and non-defaulted enterprises, and effectively reflects the default characteristics of corporate bonds. The sample covers 12 industries, including transportation, non-ferrous metals, fossil energy, etc., to avoid single-industry bias and capture the similarities and differences of default risks in different industries. Three machine learning algorithms, namely XGBoost, logistic regression and decision tree, are selected to construct the early warning model, and XGBoost is determined to be the optimal model through the comparative analysis of accuracy, recall, AUC and F1 value. The core indicators include ROA, net cash flow from financing activities, etc. The XGBoost model integrates multiple decision trees to form a powerful prediction capability, and it can accurately identify 86% of bond default risk enterprises. The model provides a data-driven early warning tool for regulatory and investment decisions, verifies the effectiveness of integrated learning in unbalanced data scenarios, and is of great practical value for improving the risk-return control ability of the bond market.

Keywords: XGBoost, Unbalanced data processing, Intelligent early warning, Bond defaults.

1. Introduction

In recent years, with the rapid development of China's bond market, the scale of corporate bond issuance has continued to grow, and bond defaults have shown a rising trend. Corporate bond defaults not only affect the reputation and credibility of enterprises but also cause significant losses to investors and the entire financial market. Hothis studyver, traditional early warning methods for corporate bond defaults often have the following two limitations: firstly, the financial indicators mainly focus on the historical operating data of the enterprise, ignoring the impact of external factors such as the market environment and industry trends; secondly, although the expert judgment is subjective and empirical to a certain extent, it is difficult to accurately predict the future default risk of the enterprise due to the limitations of personal knowledge and experience. With the continuous development of big data, artificial intelligence, and other technologies, multi-dimensional information integration has become a new research trend. At the same time, machine learning algorithms have powerful data processing and analyzing capabilities, which can dig out the complex relationships and laws bet this stud yen the data, and provide new methods and means for the early warning of corporate bond default [1].

In terms of the relevant influencing factors of bond default, macroeconomic conditions, industry environment, and the condition of the main body of debt issuance and decision-making have a close relationship to bond default [2]. The regional economic environment, the degree of economic development, the financial cycle, and other factors have a certain relationship with the enterprise bond default. From different aspects of the enterprise's operating conditions, the probability of enterprise default also has a certain impact [3]. In other aspects, the strategies implemented by enterprises,

entrepreneurial characteristics, and rational prediction of the market directly affect the subsequent decisions of enterprises, and have a more direct impact on corporate bond default [4].

In the research in the field of bond disclosure, the transparency and completeness of information play a significant role in enhancing the overall transparency of the bond market [5]. High-quality bond disclosure not only reduces the financing costs of bond issuers but also makes the market system better as the scope of disclosure expands [6]. In addition, this comprehensive disclosure also has an impact on the risk of bond defaults and helps reduce the likelihood of defaults occurring [7]. By analyzing other literature, it is found that most of the literature has limitations, and this study applies by analyzing a particular case.

Therefore, this study aims to utilize machine learning algorithms to construct an intelligent model that can more comprehensively and accurately warn of corporate bond defaults. By integrating data from different sources and types and combining the advantages of machine learning algorithms, this study will dig deeper into the potential risk factors of corporate bond defaults and provide valuable reference information for enterprises, investors, and regulators. At the same time, this study will also compare and analyze the performance of different models and algorithms to provide theoretical support and practical guidance for the construction of a more efficient and accurate corporate bond default early warning system.

2. Research methodology

2.1. Data Acquisition and Preprocessing

2.1.1. Data sources

This study uses the CHOICE Financial Terminal as the data source to collect the data of bond default enterprises from

2017-2022, and screens out 42 default enterprises and 200 non-default enterprises, involving 83 bonds. The sample data covers 12 industries, including transportation, transportation equipment, nonferrous metals, and fossil energy. In this study, the firms that have experienced bond defaults are labeled as 1, and otherwise as 0.

2.1.2. Data pre-processing

In the preliminary stage of data analysis, this study exhaustively cleaned and preprocessed the raw data to ensure the validity and accuracy of the subsequent analysis. To address the problem of missing values, this study adopted two main strategies: firstly, this study deleted bonds with too little data, and secondly, this study applied interpolation using the LOCF method. The LOCF method is a commonly used missing value treatment, which estimates missing values by replacing the missing value with the most recent valid observation prior to the missing value. The LOCF method is a commonly used missing value processing method that estimates missing values by replacing the missing value with the most recent valid observation before the missing value. This method assumes that the observation before the missing value is representative of the state of the missing value, thus providing an efficient way of handling missing values [8]. By applying the LOCF method, this study has successfully reduced the number of missing values in the data and improved the completeness and usability of the dataset.

There are relatively few listed companies with bond defaults in the study of corporate bond defaults, and the collected dataset presents a strong imbalance. Traditional classifiers generally prioritize the accuracy of the majority class samples in order to ensure the overall prediction accuracy of the training model, which can lead to a higher classification error of the model on the minority class samples of defaulting enterprises. Therefore, before applying the model for early warning, this study performs random undersampling treatment for the majority class samples and SMOTE oversampling treatment for minority class samples [9].

2.2. Research methodology

2.2.1. Selection of indicators

Based on the online literature on bond default risk metrics, the indicators are selected from both financial and non-financial perspectives, and a total of 30 indicators, 21 financial and 9 non-financial, are selected. The 21 financial indicators correspond to the firm's solvency, development ability, operating ability, and cash flow analysis, respectively. The 9 non-financial indicators are selected from the directions of bond attributes and macro indicators.

2.2.2. Model construction

XGBoost (Extreme Gradient Boosting) is a state-of-the-art integrated learning model constructed based on the Gradient Boosting Decision Tree (GBDT) framework, which optimizes the loss function by iteratively training a series of decision trees [10]. Its innovative introduction of regularization terms, the use of additive ideas and forward optimization, iteratively generates the next round of classifiers based on the classification training results of the previous round of models in the construction, and integrates the results of a single decision tree according to the rules of Boosting. The model complexity is effectively controlled, the generalization ability is significantly enhanced, and the risk of overfitting is avoided. The model is highly scalable, supports parallel computing, and can efficiently handle large-scale datasets, while it is

robust to missing values and can automatically perform feature selection and importance assessment. The model is widely used in the fields of financial prediction, medical diagnosis, image recognition, etc. By virtue of its efficiency and accuracy, it has become an important tool in the field of machine learning with both theoretical depth and practical value, providing a powerful solution for complex data modeling.

The key hyperparameters within the model that need to be tuned include the number of trees in the forest, $n_estimators$, and the maximum depth of the tree, max_depth . And this study relies heavily on the grid search method to optimize the parameters in XGBoost. The optimal parameter values are finally determined as shown in Table 1, and this combination of parameters allows the XGBoost model to achieve the highest prediction accuracy.

Table1. XGBoost model parameter settings

Parameter name	Parameter Symbol	Parameter Value
Number of trees	$n_estimators$	1000
Maximum depth	max_depth	3
Learning rate	$learning_rat$	0.5

In this study, the XGBoost model is used for prediction while comparing and analyzing the effect with the decision tree and binary logistic regression model.

3. Modeling and solving

3.1. Feature Importance Analysis and Model Interpretation

3.1.1. Characteristic importance analysis

This study calculated the characteristic importance score of each indicator by constructing the XGBoost model, the higher the score, the higher its relative importance, after excluding the indicators with a score of 0, the results of the top ten indicators of the characteristic importance rankings selected for this study are shown in Figure 1 and Table 2, and the results of the table show that coupon interest rate, self-inspection of the debt ratio, the year-on-year growth rate of the total profit, and the total amount of issuance are the factors that have high impacts on the defaults of the enterprise factors.

Table 2. Top 10 Feature Importance Results

norm	Characteristic importance assessment value
Equity Ratio(ER)	0.159029566
Coupon Rate(CR)	0.112016472
Debt-Asset Ratio(DAR)	0.110735948
Year-on-Year Growth Rate of Total Profit(YTYG_TP)	0.077465733
Return on Assets(ROA)	0.048944972
Total Shareholders' Equity(TSE)	0.045295942
Year-on-Year Growth Rate of Net Assets(YTYG_NA)	0.020351966
Net Cash Flow from Financing Activities	0.020012442
Net Profit Margin on Total Assets(ROA)	0.01598107
Return on Equity(ROE)	0.011122791

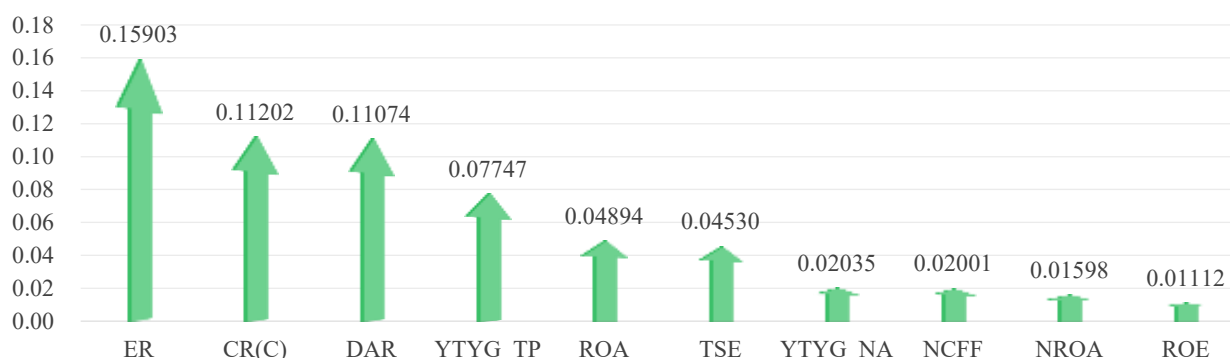


Figure 1. Feature Importance Presentation

3.1.2. Model interpretation

The equity ratio, also known as the debt-to-equity ratio, measures the degree of a firm's financial leverage. A higher equity ratio increases the risk of default; Coupon rate (current) this is the interest rate specified at the time of bond issuance, too high a coupon rate will result in a heavier interest burden for the enterprise; Gearing ratio shows the ratio of total liabilities to total assets and is a key indicator of the financial stability of an enterprise. A high gearing ratio indicates that the enterprise faces greater financial pressure; the year-on-year growth rate of total profit affects debt servicing capacity; and the return on total assets (ROA) measures the enterprise's ability to utilize its assets to generate net profit. Low ROA affects the solvency and default risk of enterprises; Total shareholders' equity indicates the net equity of the firm's owners in the firm's assets; Year-on-year growth rate of net assets shows the year-on-year growth of a firm's net assets, which affects its credit rating and risk of default; Net cash flow from financing activities. This reflects the net cash that a firm receives through financing activities. Persistent negative cash flow indicates that a company is experiencing difficulties in raising funds; lot his study r total net asset margin (ROA), an important measure of a company's profitability, leads to a decline in profitability; and return on equity (ROE), which measures the efficiency of a company's ability to utilize shareholders' capital to generate profits. A decline in ROE indicates that a company's ability to create value for its shareholders has this studyakened. Changes in these ratios are of particular concern because they can quickly affect a firm's bond value and probability of default.

3.2. Model Solving

In order to compare the fit of XGBoost, Decision Tree, and Binary Logistic Regression models, the three models this studyre modeled separately in this study using Python on the same data. In this study, the overall dataset was divided in the ratio of training set: test set = 7:3 to select the overall optimal model. The test set is used to assess the model's performance and robustness, and the average of 10 rounds of validation results is used as the experimental results and the number of indicators in the test model is 10 for both normal and default groups. The model training effect is shown in Table 3: From the results in the table, it can be seen that the results of Accuracy, Recall, AUC, and FI-Score indicators of the corporate bond default early warning model based on the XGBoost model are superior, so the XGBoost model is chosen to be constructed.

Table 3. Comparison of XGBoost, Decision Tree, and Binary Logistic Regression Model Fitting Effectiveness

	Logistic	decision tree	XGBoost
Accuracy	0.78	0.86	0.89
Recall	0.82	0.78	1.00
AUC	0.82	0.79	0.86
FI-Score	0.77	0.83	0.83
G-mean	0.67	0.73	0.76

4. Conclusions and recommendations

In this study, a corporate bond default early warning model based on XGBOOST is constructed, which effectively improves the ability of recognizing a few types of samples while providing an interpretable index system for risk assessment by using real default sample data and imbalance processing technology. The research results not only make up for the shortcomings of traditional methods but also enrich the methodological theory of machine learning in the early warning application of financial risks. The research results show that the XGboost model demonstrates strong prediction capability by integrating decision trees and excels in accuracy and other metrics, with a recall rate of 86%. The model can comprehensively capture the default risk characteristics of enterprises, assist regulators, and remind enterprises to take precautions. In terms of corporate capability, a high equity ratio, gearing ratio, year-on-year decline in profit, year-on-year decline in net assets, negative funding cash flow, and high coupon rate increase default risk, while a large shareholders' equity is low risk. In response to the above conclusions, this study puts forward the following suggestions: enterprises should strengthen the deep mining analysis of financial data, bond attribute data, and macroeconomic data. In the era of the digital economy, massive data can assist in recognizing the risk of corporate bond default, and relevant data mining should be strengthened with the help of big data technology. At the same time, enterprises and society should cultivate composite talents who combine computer technology with the direction of corporate finance and bonds, so as to better utilize artificial intelligence technology to solve the problem of corporate bond default early warning.

Although this study has achieved certain results, there is still room for further improvement in the future. At the data level, there is insufficient mining of unstructured data such as textual disclosure and market public opinion, and in the future, this study can expand the direction to incorporate diversified unstructured data and build a real-time dynamic monitoring mechanism.

References

- [1] Lee, Y.C. Application of support vector machines to corporate credit rating prediction[J]. *Expert Systems with Applications*, 2007, 33(1): 67-74.
- [2] Xiang Shi, Zeng Yinqiu, Yan Xinguo, Liu Yuyin. Research on bond default risk monitoring and early warning based on the support vector machine method[J] *Financial Economy*, 2022(01): 40-50.
- [3] Luo Chaoyang, Li Xuesong. Financial cycle, total factor productivity, and bond default[J]. *Economic Management*, 2020, 42(2): 5-22.
- [4] Qianyun Ji. Managerial overconfidence, corporate strategic aggressiveness and debt default [D]. Inner Mongolia University of Finance and Economics, 2022.
- [5] Healy M P, Palepu G K. Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature[J]. *Journal of Accounting and Economics*, 2001, 31(1): 405-440.
- [6] W. C S, B. J W. On financial contracting: An analysis of bond covenants[J]. *Journal of Financial Economics*, 1979, 7(2): 117-161.
- [7] XU Kun, MENG Huazhe, YANG Xiaozhou. Influencing factors and formation mechanism of corporate bond defaults--Based on rooted theory[J]. *Finance and Accounting Monthly*, 2022(2): 114-122.
- [8] Wang K, Vos D J, Smart M, et al. Explaining Youth Driver Licensing Determinants Using XGBoost and SHAP[J]. *Transport Policy*, 2025, 16887-100.
- [9] Zhao G, Li L, He H, et al. LGSMOTE-IDS: Line Graph based Weighted-Distance SMOTE for imbalanced network traffic detection[J]. *Expert Systems With Applications*, 2025, 281127645-127645.
- [10] W. S, Agus S, Andrew M, et al. Bigotry in last observation carried forward (LOCF) analysis[J]. *Research*, 2022, 10(1): 19-20.