Corporate Bankruptcy Prediction in the U.S. Using Random Forest and XGBoost Algorithms

Zihao Lan

School of Management Science and Engineering, Shandong University of Finance and Economics, Jinan, Shandong, China
202306140523@mail.sdufe.edu.cn

Abstract. Corporate bankruptcy prediction is crucial to investors, financial institutions and regulators, because it supports early risk warning, enhances capital allocation efficiency, and contributes to both financial market resilience and enterprise sustainability. Traditional statistical methods have limited performance in complex financial data. With the increasing uncertainty of the global economic environment and the growing complexity of enterprise operations, more robust predictive models are urgently needed. Machine learning has demonstrated significant advantages in capturing nonlinear relationships and handling high-dimensional financial indicators. This paper builds a high-precision bankruptcy prediction model based on Random Forest and eXtreme Gradient Boosting (XGBoost) algorithm to improve the prediction performance. The research uses open enterprise financial data, selects key indicators through feature engineering, and compares the performance of two integrated learning algorithms. The experimental results show that XGBoost is better than random forest in accuracy and Area Under Curve(AUC) value, and the feature importance analysis reveals the key factors affecting bankruptcy risk. The model in this paper provides a reliable tool for enterprise risk early warning, and provides an optimization direction for subsequent research. It has important practical and theoretical value.

Keywords: Bankruptcy Prediction, Random Forest, eXtreme Gradient Boosting, Machine Learning.

1 Introduction

In the United States, bankruptcy refers to the legal process [1] that individuals or corporations apply for when they are unable to repay their debts. With the intensification of global economic uncertainty in recent years, the business environment for corporations has become increasingly complex, ultimately leading to an increase in the risk of corporate bankruptcy. Corporation bankruptcy has a significant impact on the corporations themselves, investors, creditors, and even the entire financial market. Therefore, accurate and timely bankruptcy prediction has become an important research topic in the field of risk control, and numerous studies have been conducted on it.

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_13

In the 1960s, Beaver's single-factor discriminant method and Altman's multivariate discriminant model (MDA) were used for bankruptcy prediction [2, 3]. Although they have strong interpretability, they still cannot explain complex nonlinear feature relationships and high-dimensional data. In 1980, Ohlson first introduced logistic regression into bankruptcy prediction methods [4]. With the development of big data and machine learning theory, the focus of research has shifted more towards data-driven bankruptcy prediction methods. Decision Tree, Neural Network algorithm, Support Vector Machine, Random forest, and other methods have gradually been applied to bankruptcy prediction [5, 6, 7]. These machine learning models can automatically extract features and identify potential patterns by training on large amounts of historical data, providing higher accuracy and flexibility for bankruptcy prediction. Among them, Random Forest, as an ensemble learning method, is widely used in the field of financial risk prediction due to its robustness to high-dimensional features [8], strong generalization ability, and strong interpretability of feature importance. It achieves model integration by constructing multiple decision trees and using the Bagging strategy, effectively reducing the risk of overfitting and improving the stability of predictions [9].

The core theme of this study is company bankruptcy prediction based on the random forest algorithm, aiming to build a high-precision prediction model using historical financial data from US corporations' bankruptcies. The main methods of this paper include: first, collecting and organizing a dataset containing corporate financial indicators and business information, and conducting data cleaning and feature selection; second, training and predicting using Random Forest and eXtreme Gradient Boosting (XGBoost). Finally, evaluating model performance using indicators such as Area Under Curve(AUC) and F1-score.

2 Methods

2.1 Data Source

This dataset encompasses the fundamental accounting data of 8,262 distinct publicly traded US corporations listed on the New York Stock Exchange and National Association of Securities Dealers Automated Quotations(NASDAQ) from 1999 to 2018. It includes 19 basic financial indicators such as total liabilities, total revenue, and inventory. Due to the significant impact of company size on the original data size, directly comparing the net profits of large companies with those of small ones is meaningless. Moreover, financial ratios contribute significantly more to bankruptcy prediction than raw data [4]. To facilitate subsequent EDA and model application, based on the original dataset, 12 financial ratio feature information, such as the current ratio, net profit, and return on equity were constructed. Table 1 presents the complete features and their descriptive statistical information.

Table 1. Basic information of variables.

Variable Name	Count	Mean	STD
year	78682	2007.51	5.74
Current assets	78682	880.36	3928.56
Cost of goods sold	78682	1594.53	8930.48
Depreciation and amortization	78682	121.23	652.38
EBITDA	78682	376.76	2012.02
Inventory	78682	201.61	1060.77
Net Income	78682	129.38	1265.53
Total Receivables	78682	286.83	1335.98
Market value	78682	3414.35	18414.1
Net sales	78682	2364.02	11950.07
Total assets	78682	2867.11	12917.94
Total Long-term debt	78682	722.48	3242.17
EBIT	78682	255.53	1494.64
Gross Profit	78682	769.49	3774.7
Total Current Liabilities	78682	610.07	2938.39
Retained Earnings	78682	532.47	6369.16
Total Revenue	78682	2364.02	11950.07
Total Liabilities	78682	1773.56	8053.68
Total Operating Expenses	78682	1987.26	10419.63
equity	78682	1093.55	5516.83
Current ratio	78682	3.49	88.75
Net margin	78682	-11.48	279.73
Gross margin	78682	-6.32	208.7
Ebit margin	78682	-9.05	219.54
Ebitda margin	78682	-8.68	214.71
Liabilities to assets	78682	1.92	46.44
Debt to assets	78682	0.27	11.83
Debt to equity	78682	Inf	Nan
Debt to liabilities	78682	0.27	0.27
roa	78682	-0.69	13.65
roe	78682	-Inf	Nan

2.2 Introduction to the Method

Random Forest. The Random Forest algorithm employs the Bootstrapping under-sampling technique to obtain multiple training sets from the sample data set. Each training set corresponds to generating its own decision tree, and each decision tree is independent of the others. The growth of the decision tree stops when the maximum

depth of the tree is reached, and no pruning is required. Random sampling is used to partition the splitting nodes of the decision tree and select the optimal feature values[9].

The advantage of random forests lies in the fact that the training data used by the algorithm to train base learners is generated through random sampling of feature subsets, enabling effective handling of data with high-dimensional features.

XGBoost. eXtreme Gradient Boosting (XGBoost) is a highly efficient machine learning algorithm based on the gradient boosting framework, proposed by Tianqi Chen in 2016[10]. By integrating multiple weak learners (typically decision trees) and iteratively optimizing the model, it progressively reduces prediction errors. Compared to traditional gradient boosting methods, XGBoost introduces several key enhancements in both performance and efficiency. These include regularization terms to prevent overfitting, the utilization of second-order derivatives for faster convergence, optimized computation for feature splitting to handle large-scale datasets, and automatic missing value handling. These innovations make XGBoost particularly powerful for complex, high-dimensional machine learning tasks [11,12,13].

3 Result and Discussion

3.1 Exploratory Data Analysis

The dataset was plotted according to whether the company went bankrupt, as shown in the left circular chart of Figure 1. It was found that bankrupt companies accounted for only 6.6%, indicating a severe imbalance in the sample. Ultimately, Synthetic Minority Over-Sampling Technique (SMOTE) was considered to balance the sample. The right circular chart in Figure 1 shows the proportion of bankruptcies after using SMOTE oversampling, indicating that the sample is more balanced after applying SMOTE oversampling.

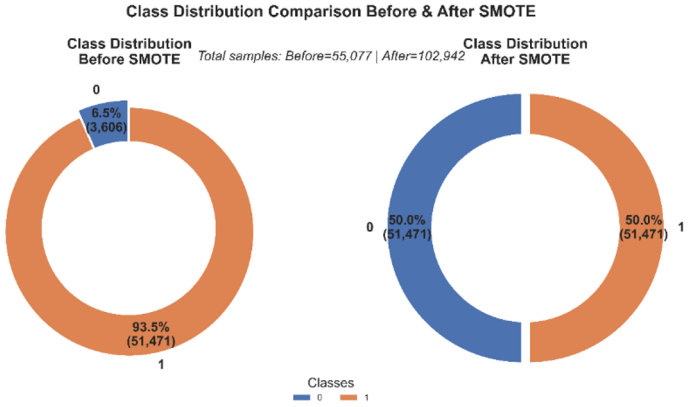

Fig. 1. Class Distribution Comparison Before & After SMOTE (Picture credit: Original)

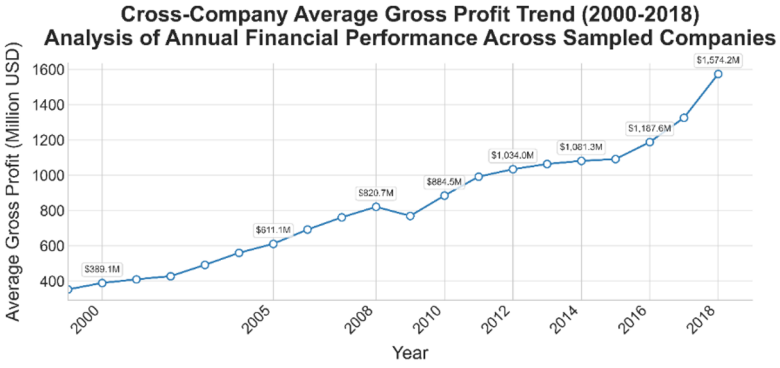

Fig. 2. Cross-Company Average Gross Profit Trend (2000-2018) (Picture credit: Original)

A line chart is plotted based on the average Gross Profit of all companies by year, as shown in Figure 2. The line chart illustrates the overall upward trend in the average total revenue across companies from 2000 to 2018. This is primarily due to the fact that the majority of the dataset consists of non-bankrupt companies, naturally leading to a gradual increase in overall profitability. This is consistent with the conclusion drawn from the previous pie chart. However, there was a significant decline after 2008. This phenomenon can be attributed to the most severe global economic crisis since the Great Depression in 1929, which occurred in the United States in 2008. This crisis had a profound impact on the global financial system and economy, subsequently leading to a significant reduction in the average total revenue of the companies in the dataset.

Figure 3 illustrates the trend of annual average changes in certain numerical characteristics from 2000 to 2018 through a line chart. The analysis results show that most characteristics, including current assets, total accounts receivable, net sales, total assets, gross profit, total revenue, and total liabilities, exhibit an overall upward trend,

which corroborates the conclusion drawn from Figure 2. The consistency of trends between Figure 2 and Figure 3 further validates the reliability of the dataset and the robustness of the analytical method employed.

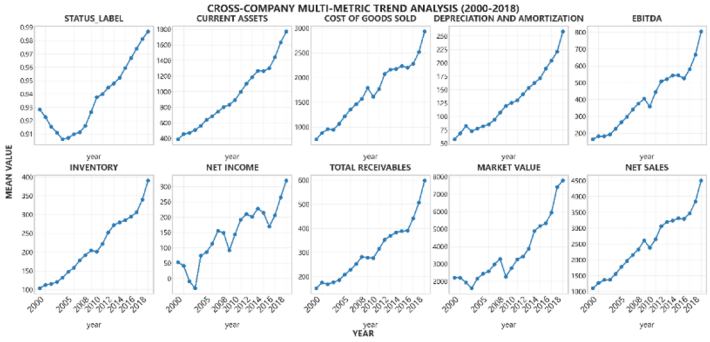

Fig. 3. Cross-Company Multi-Metric Trend Analysis (2000-2018) (Picture credit: Original).

Figure 4 presents the correlation coefficient matrix of various feature variables.

A strong positive correlation (0.99) is observed between Total Revenue and Total Operating Expenses, which aligns with financial logic as operating expenses typically increase proportionally with revenue growth.

Notably, Total Long-term Debts and Retained Earnings demonstrate a negative correlation (-0.32). This reveals a potential financial distress signal: when companies experience losses, their retained earnings decrease while long-term debts tend to increase to maintain operations, thereby forming this inverse relationship.

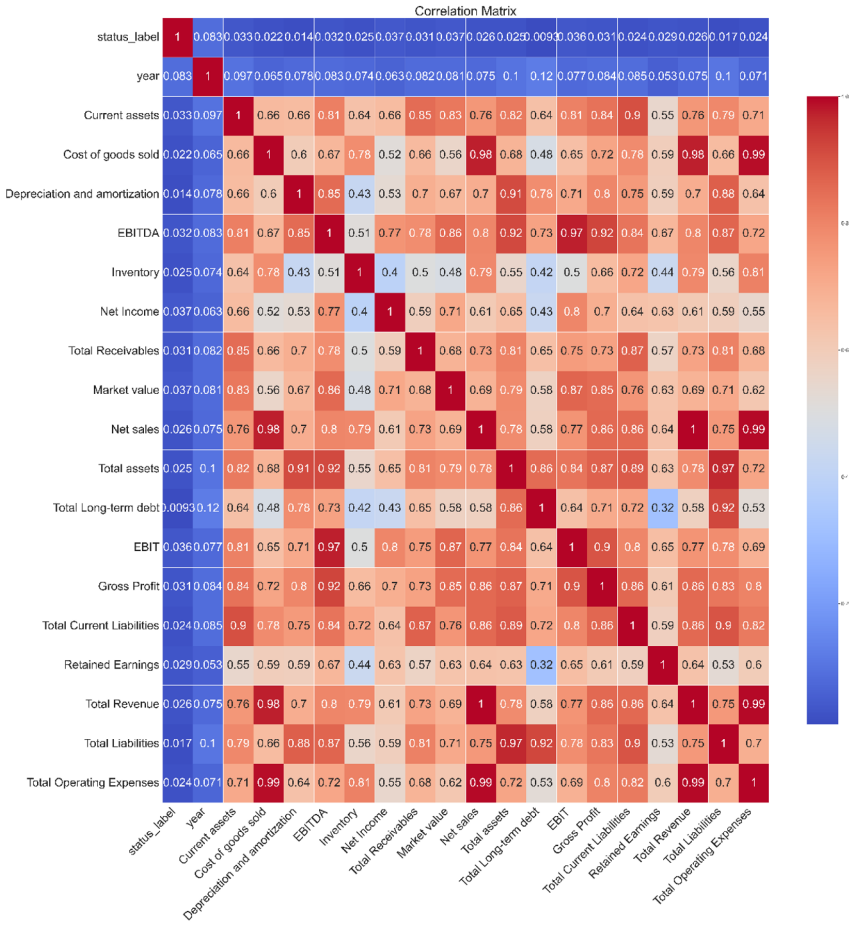

Fig. 4. Correlation Matrix (Picture credit: Original).

3.2 Model Evaluation

Based on the analysis of 78683 enterprise bankruptcy financial records, this study uses random forest and XGBoost algorithm to train the classifier for bankruptcy prediction, and then compares the performance of the two methods. Figure 5 shows the key performance indicators of Random Forest and XGBoost on this dataset.

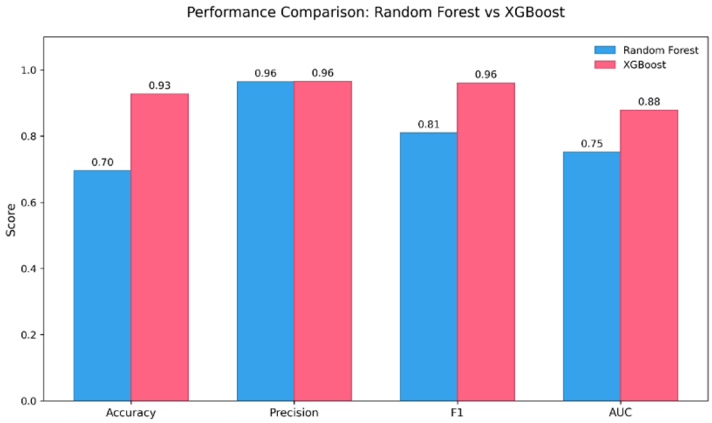

Fig. 5. Performance Comparison: RF vs XGBoost(Picture credit: Original)

Figure 5 shows the comparative performance evaluation of Random Forest (RF) and XGBoost between the four key indicators (accuracy, precision, F1 score and AUC). The analysis results show that XGBoost is superior to RF in all evaluation categories, showing a consistently higher score. It is worth noting that the accuracy of XGBoost is 0.93, the accuracy is 0.96, F1 scores are 0.96 and AUC are 0.88, while the corresponding scores of RF are 0.70, 0.96, 0.81 and 0.75 respectively. This remarkable advantage highlights the effectiveness of XGBoost on this specific dataset, and demonstrates its enhanced ability to capture complex feature interactions through gradient enhancement technology. Figure 6 and Figure 7 then show the five most important functions in Random Forest (RF) and XGBoost algorithms and their relative importance rankings.

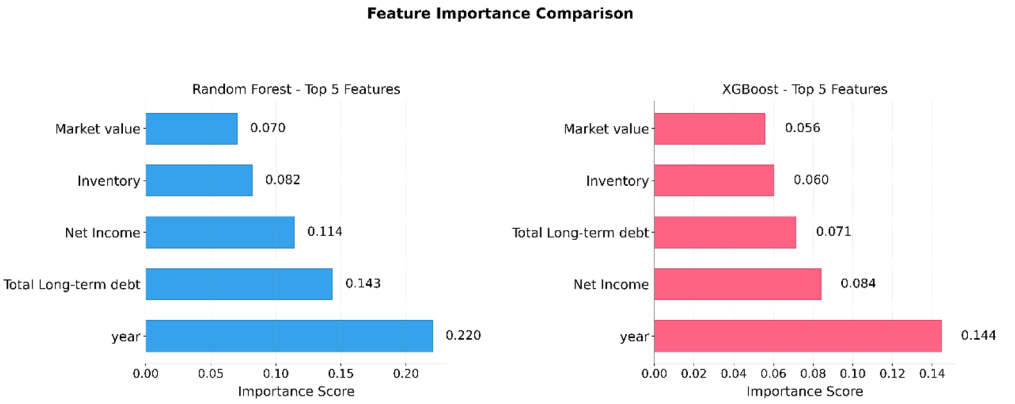

Fig. 6. Feature Importance Comparison (Picture credit: Original)

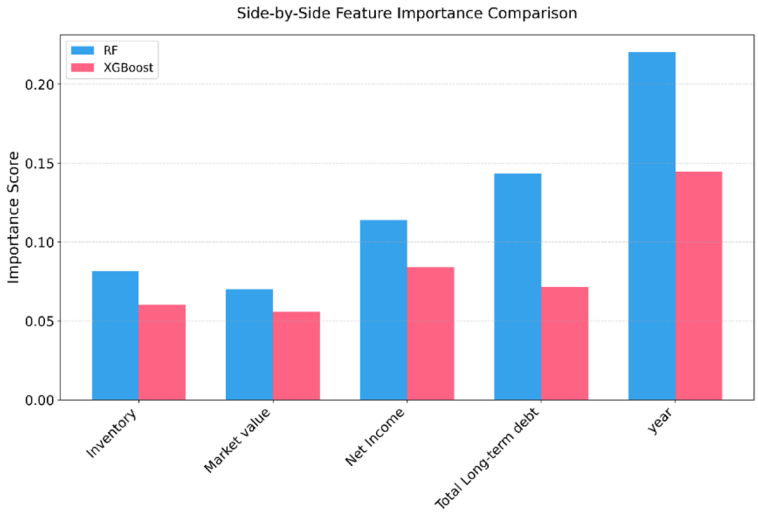

Fig. 7. Side-by-Side Feature Importance Comparison (Picture credit: Original)

The analysis shows that Year has the strongest prediction ability in the two machine learning models, reaching 0.22 importance score in the random forest model and 0.15 high score in XGBoost. This shows that macroeconomic conditions significantly affect the company's bankruptcy risk, and that systematic risk factors such as economic cycle fluctuations and industry wide trends play a key role in assessing the company's financial health.

Total long-term debt becomes the second most important predictor. Although its importance score is slightly lower than the annual factor, it is still higher than other financial indicators. This finding is strongly consistent with the classical financial theory, which believes that there is a positive correlation between financial leverage and corporate risk. Recognizing the level of debt, especially the long-term debt burden, is a key dimension in assessing corporate financial vulnerability.

Operating indicators, including net income, market value and inventory, showed moderate forecasting effectiveness, although their importance was significantly lower than that of macroeconomic and debt related factors.

4 Conclusion

This study uses the financial data of American companies to compare and analyze Random Forest and XGBoost algorithms used for binary bankruptcy prediction. These experimental results prove the outstanding performance of XGBoost, achieving 93% accuracy (32% higher than RF) and 96% F1 score on the test set. In view of the inherent category imbalance in the bankruptcy prediction task, F1's high score is particularly noteworthy, which confirms that XGBoost has enhanced its ability to model the

complex nonlinear relationship between financial indicators and bankruptcy risk through its advanced tree splitting mechanism. The feature importance analysis further reveals that macroeconomic factors are important determinants of enterprise bankruptcy.

The current research has two main limitations, which point out the direction for future work. First, relying only on internal financial statements ignores the broader economic background; Combining macroeconomic indicators can improve forecasting ability and identify specific external risk factors. Secondly, the static nature of XGBoost limits time analysis; Integrated time series models (such as Long Short-Term Memory algorithm) can capture evolving financial models and enhance the model's ability to interpret time-dependent bankruptcy processes. These extensions can provide more comprehensive insights into the prediction of corporate financial distress.

References

1. Loveland, F.O.: A Treatise on the Law and Proceedings in Bankruptcy. W.H. Anderson & Company, Cincinnati (1904)
2. Beaver, W.H.: Financial ratios as predictors of failure. Journal of Accounting Research 4(3), 71–111 (1966)
3. Altman, E.I.: Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. Journal of Finance 23(4), 589–609 (1968)
4. Ohlson, J.A.: Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research 18(1), 109–131 (1980)
5. Becerra-Vicario, R., Alaminos, D., Aranda, E., et al.: Deep recurrent convolutional neural network for bankruptcy prediction: A case of the restaurant industry. Sustainability 12(12), 5180 (2020)
6. Sun, J., Fujita, H., Zheng, Y., et al.: Multi-class financial distress prediction based on support vector machines integrated with the decomposition and fusion methods. Information Sciences 559, 153–170 (2021)
7. Barboza, F., Kimura, H., Altman, E.: Machine learning models and bankruptcy prediction. Expert Systems with Applications 83, 405–417 (2017)
8. Gurnani, I., Tandian, F.S., Anggreainy, M.S.: Predicting company bankruptcy using random forest method. In: Proceedings of the 2021 2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS), pp. 1–5. IEEE, Kuala Lumpur (2021)
9. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
10. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. ACM, San Francisco (2016)
11. Ben Jabeur, S., Stef, N., Carmona, P.: Bankruptcy prediction using the XGBoost algorithm and variable importance feature engineering. Computational Economics 61(2), 715–741 (2023)
12. Garg, K., Gill, K.S., Malhotra, S., et al.: Implementing the XGBoost classifier for bankruptcy detection and SMOTE analysis for balancing its data. In: Proceedings of the 2024 2nd International Conference on Computer, Communication and Control (IC4), pp. 1–5. IEEE, Jalandhar (2024)

13. Muslim, M.A., Dasril, Y.: Company bankruptcy prediction framework based on the most influential features using XGBoost and stacking ensemble learning. International Journal of Electrical and Computer Engineering (IJECE) 11(6), 5549–5557 (2021)