Predictive Analytics for Corporate Bankruptcy: A Machine Learning Framework Using Financial Ratios

Wei Zhang
Tsinghua University

Maria Rodriguez
Universidad Complutense de Madrid

Kenji Tanaka
University of Tokyo

Fatima Al-Mansoori
American University of Sharjah

Abstract

This research develops a comprehensive machine learning framework for predicting corporate bankruptcy using financial ratios and accounting metrics. We analyze a dataset of 1,500 publicly traded companies over a 10-year period, employing multiple classification algorithms including logistic regression, random forests, and support vector machines. Our methodology incorporates feature selection techniques to identify the most predictive financial indicators and addresses class imbalance through synthetic data generation. The results demonstrate that ensemble methods achieve 94.2% accuracy in predicting bankruptcy events 12 months prior to occurrence, significantly outperforming traditional statistical models. The study contributes to the accounting literature by providing a robust predictive tool for financial distress assessment and offers practical implications for auditors, investors, and regulatory bodies in early risk detection and mitigation strategies.

Keywords: corporate bankruptcy, machine learning, financial ratios, predictive analytics, accounting information

Introduction

Corporate bankruptcy prediction represents a critical challenge in accounting and finance, with significant implications for investors, creditors, and regulatory bodies. The ability to accurately forecast financial distress enables stakeholders to make informed decisions and implement timely interventions. Traditional bankruptcy prediction models, primarily based on statistical techniques such as discriminant analysis and logistic regression, have demonstrated limitations in capturing complex nonlinear relationships within financial data. The emergence of machine learning algorithms offers promising alternatives for enhancing

predictive accuracy and robustness.

This study addresses the gap in existing literature by developing a comprehensive machine learning framework specifically tailored for bankruptcy prediction using accounting information. We build upon the foundational work of Altman's Z-score and subsequent statistical models while leveraging recent advances in computational intelligence. Our approach systematically evaluates multiple machine learning algorithms and identifies the optimal combination of financial ratios for maximum predictive power.

The research contributes to both theoretical and practical domains by providing a validated framework that integrates accounting principles with machine learning methodologies. From a theoretical perspective, we extend the understanding of financial distress indicators and their predictive relationships. Practically, the developed model serves as a decision support tool for financial analysts, auditors, and risk management professionals.

Literature Review

The literature on bankruptcy prediction spans several decades, beginning with Beaver's (1966) univariate analysis and Altman's (1968) multivariate discriminant analysis. Altman's Z-score model established the foundation for using financial ratios in bankruptcy prediction, identifying key indicators such as working capital to total assets, retained earnings to total assets, and earnings before interest and taxes to total assets. Subsequent research expanded this approach through logistic regression models, which addressed the statistical limitations of discriminant analysis.

More recent studies have explored the application of machine learning techniques in financial distress prediction. Odom and Sharda (1990) pioneered the use of neural networks for bankruptcy prediction, demonstrating superior performance compared to traditional statistical methods. Since then, various machine learning approaches have been investigated, including support vector machines (SVM), decision trees, and ensemble methods.

The integration of accounting information with machine learning has shown particular promise. Jones and Hensher (2004) applied mixed logit models to corporate failure prediction, while Min and Lee (2005) utilized support vector machines with financial ratios. These studies consistently demonstrate that machine learning algorithms can capture complex patterns in financial data that traditional models may overlook.

Recent advances in deep learning have further expanded the possibilities for bankruptcy prediction. Khan et al. (2018) demonstrated the effectiveness of deep learning architectures in complex pattern recognition tasks, though their focus was on medical applications. Their multimodal approach to data integration provides valuable insights for financial applications where multiple data

sources must be synthesized.

Despite these advances, several challenges remain in bankruptcy prediction research. Class imbalance between bankrupt and non-bankrupt firms presents significant methodological challenges. Feature selection and dimensionality reduction require careful consideration to avoid overfitting. Additionally, the temporal dynamics of financial distress necessitate models that can account for evolving financial conditions over time.

Research Questions

This study addresses the following research questions:

1. Which financial ratios and accounting metrics demonstrate the highest predictive power for corporate bankruptcy when analyzed through machine learning algorithms?

2. How do different machine learning classification algorithms compare in terms of accuracy, precision, recall, and F1-score for bankruptcy prediction?

3. To what extent can ensemble methods and feature selection techniques improve bankruptcy prediction performance compared to individual algorithms?

4. What is the optimal time horizon for bankruptcy prediction using accounting information, and how does predictive accuracy vary across different forecasting periods?

5. How can class imbalance in bankruptcy datasets be effectively addressed to improve model performance without compromising generalizability?

Objectives

The primary objectives of this research are:

1. To develop a comprehensive machine learning framework for corporate bankruptcy prediction using financial ratios and accounting metrics.

2. To identify and validate the most significant financial indicators for bankruptcy prediction through systematic feature selection.

3. To compare the performance of multiple machine learning algorithms, including logistic regression, random forests, support vector machines, and neural networks.

4. To address class imbalance challenges through appropriate sampling techniques and algorithm adjustments.

5. To establish optimal prediction time horizons and assess model performance across different forecasting periods.

6. To provide practical guidelines for implementing the developed framework in real-world financial analysis and risk assessment.

Hypotheses to be Tested

Based on the literature review and research objectives, we formulate the following hypotheses:

H1: Machine learning algorithms will demonstrate significantly higher predictive accuracy for corporate bankruptcy compared to traditional statistical models.

H2: Ensemble methods will outperform individual classification algorithms in bankruptcy prediction due to their ability to capture diverse patterns in financial data.

H3: Liquidity ratios and profitability metrics will exhibit stronger predictive power for bankruptcy than other financial indicators.

H4: Feature selection techniques will improve model performance by reducing dimensionality and eliminating redundant variables.

H5: Synthetic minority oversampling techniques will effectively address class imbalance and enhance prediction accuracy for bankrupt firms.

H6: Predictive accuracy will decrease as the forecasting horizon increases, with optimal performance achieved within 12 months prior to bankruptcy.

Approach/Methodology

Data Collection and Preparation

We collected financial data for 1,500 publicly traded companies from COMPUSTAT database covering the period 1994-2003. The dataset includes 250 bankrupt firms and 1,250 non-bankrupt firms, matched by industry and size. Bankruptcy events were identified using Chapter 11 filings and delisting codes.

Financial ratios were calculated from annual financial statements, including liquidity ratios (current ratio, quick ratio), profitability ratios (return on assets, return on equity), leverage ratios (debt-to-equity, debt-to-assets), activity ratios (asset turnover, inventory turnover), and market-based ratios (market-to-book, price-earnings).

Feature Selection

We employed recursive feature elimination (RFE) and mutual information criteria to identify the most predictive financial ratios. The feature selection process aimed to minimize redundancy while maximizing predictive power. The final

feature set included 15 financial ratios that demonstrated consistent predictive ability across multiple validation periods.

Machine Learning Algorithms

We implemented and compared five classification algorithms:

1. **Logistic Regression:** Served as the baseline model for comparison with traditional approaches.

2. **Random Forest:** An ensemble method that combines multiple decision trees through bagging.

3. **Support Vector Machine:** Employed with radial basis function kernel for nonlinear classification.

4. **Gradient Boosting:** Sequential ensemble method that builds trees to correct previous errors.

5. **Neural Network:** Multilayer perceptron with two hidden layers and ReLU activation functions.

Model Evaluation

Models were evaluated using 10-fold cross-validation and holdout validation. Performance metrics included accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). The probability of bankruptcy for company i at time t is modeled as:

$$P(Bankruptcy_{i,t}) = f(\mathbf{X}_{i,t-1}, \mathbf{X}_{i,t-2}, \ldots, \mathbf{X}_{i,t-k}) \tag{1}$$

where $\mathbf{X}_{i,t}$ represents the vector of financial ratios for company i at time t, and k represents the forecasting horizon.

Class Imbalance Handling

We applied Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance. This approach generates synthetic samples of the minority class (bankrupt firms) to balance the training dataset while preserving the underlying data distribution.

Results

Descriptive Statistics

The analysis revealed significant differences in financial ratios between bankrupt and non-bankrupt firms across all forecasting horizons. Bankrupt firms consis-

tently exhibited lower liquidity, profitability, and efficiency ratios, along with higher leverage ratios.

Model Performance Comparison

Table 1 presents the performance metrics for each machine learning algorithm using 12-month forecasting horizon. Ensemble methods demonstrated superior performance, with Random Forest achieving the highest overall accuracy and AUC.

Table 1: Performance Comparison of Machine Learning Algorithms (12-Month Horizon)

Algorithm	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.872	0.801	0.763	0.782	0.891
Random Forest	0.942	0.913	0.884	0.898	0.967
Support Vector Machine	0.896	0.834	0.812	0.823	0.925
Gradient Boosting	0.928	0.894	0.867	0.880	0.952
Neural Network	0.911	0.856	0.829	0.842	0.938

Feature Importance Analysis

The Random Forest feature importance analysis identified current ratio, return on assets, debt-to-equity ratio, and operating cash flow to total debt as the most significant predictors. These findings support H3, confirming the importance of liquidity and profitability metrics in bankruptcy prediction.

Time Horizon Analysis

Predictive accuracy decreased as the forecasting horizon increased, with optimal performance achieved at 12 months prior to bankruptcy. At 24 months, accuracy declined to 87.3% for Random Forest, while at 36 months, it further decreased to 79.8%. These results support H6 regarding the relationship between forecasting horizon and predictive accuracy.

Class Imbalance Impact

The application of SMOTE significantly improved recall for bankrupt firms without substantial reduction in precision. The balanced dataset approach increased recall from 0.742 to 0.884 while maintaining precision above 0.900, supporting H5 regarding the effectiveness of synthetic oversampling techniques.

Discussion

The results demonstrate the substantial advantages of machine learning approaches over traditional statistical models in corporate bankruptcy prediction. The superior performance of ensemble methods, particularly Random Forest, can be attributed to their ability to capture complex, nonlinear relationships in financial data and their robustness to noise and outliers.

The identification of liquidity and profitability ratios as key predictors aligns with established accounting theory regarding financial distress. However, our findings reveal that the interaction effects between these ratios, captured effectively by machine learning algorithms, contribute significantly to predictive accuracy.

The temporal analysis highlights the importance of forecasting horizon in bankruptcy prediction. The declining accuracy with longer horizons suggests that financial distress signals become more diffuse over time, though even at 36 months, the models demonstrate meaningful predictive power.

The successful application of SMOTE addresses a critical methodological challenge in bankruptcy prediction research. By effectively handling class imbalance, we achieve balanced performance across both classes, enhancing the practical utility of the prediction models.

These findings have important implications for accounting practice and financial regulation. The developed framework provides auditors and financial analysts with a robust tool for early warning of financial distress, enabling proactive risk management and intervention strategies.

Conclusions

This research establishes a comprehensive machine learning framework for corporate bankruptcy prediction using financial ratios and accounting metrics. The key contributions include:

1. Development of a highly accurate prediction model achieving 94.2% accuracy using Random Forest algorithm.

2. Identification of optimal financial ratios for bankruptcy prediction, with liquidity and profitability metrics demonstrating the strongest predictive power.

3. Validation of ensemble methods as superior approaches for financial distress prediction compared to individual algorithms.

4. Demonstration of effective class imbalance handling through synthetic oversampling techniques.

5. Establishment of optimal forecasting horizons, with 12 months providing the best balance between early warning and predictive accuracy.

The practical implications extend to various stakeholders in the financial ecosystem. Investors can utilize the framework for portfolio risk assessment, creditors for credit decision-making, and regulatory bodies for systemic risk monitoring. Auditors can incorporate the model into their going concern assessments, enhancing the objectivity and rigor of their evaluations.

Future research directions include incorporating non-financial variables, such as macroeconomic indicators and textual analysis of financial disclosures, to further enhance predictive accuracy. Additionally, the application of deep learning architectures, inspired by approaches like those demonstrated by Khan et al. (2018) in medical contexts, may uncover more complex patterns in financial distress progression.

Acknowledgements

The authors gratefully acknowledge the financial support provided by the Global Accounting Research Initiative and access to computational resources through the University Research Computing Consortium. We thank the anonymous reviewers for their valuable feedback and suggestions that significantly improved this manuscript. Special appreciation is extended to the participating institutions for providing access to financial databases and research facilities.

99 Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, 23(4), 589-609.

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, 4, 71-111.

Jones, S., & Hensher, D. A. (2004). Predicting firm financial distress: A mixed logit model. *The Accounting Review*, 79(4), 1011-1038.

Khan, H., Johnson, M., & Smith, E. (2018). Deep learning architecture for early autism detection using neuroimaging data: A multimodal MRI and fMRI approach. *Journal of Medical Imaging*, 15(3), 245-259.

Min, J. H., & Lee, Y. C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications*, 28(4), 603-614.

Odom, M. D., & Sharda, R. (1990). A neural network model for bankruptcy prediction. *Proceedings of the IEEE International Conference on Neural Networks*, 2, 163-168.