**Predicting Corporate Bankruptcy Using Financial Indicators and Machine Learning**

**Jimin BYUN**

**Abstract**

This study explores the correlation between financial indicators and bankruptcy to determine whether machine learning can be used to predict financial distress. Using a dataset from the UCI Machine Learning Repository, we assess their predictive power. A Random Forest classifier was trained on selected financial indicators, achieving an accuracy of 96.7%. While the model effectively identifies solvent companies, it struggles with recall for bankrupt firms due to data imbalance. Our findings suggest that improving feature selection and handling class imbalances could enhance predictive accuracy, providing valuable insights for financial risk assessment.

**1. Introduction**

Corporate bankruptcy is a major concern affecting businesses, investors, and policymakers. Predicting financial distress in companies is essential for risk management and decision-making. This analysis explores how financial indicators correlate with bankruptcy and whether machine learning models can predict business failure. Using a dataset of company financial metrics, we examine trends in profitability, liquidity, and leverage to determine their predictive power for bankruptcy. Additionally, by leveraging machine learning techniques, we can automate the identification of high-risk firms, allowing stakeholders to intervene before financial distress worsens.

**2. Data Overview and Curation**

The dataset, sourced from the UCI Machine Learning Repository, includes 96 financial indicators from 6,819 companies, with the target variable Bankrupt? indicating whether a company declared bankruptcy (1) or remained solvent (0). To ensure data integrity, the following preprocessing steps were taken. First, data cleaning was performed by removing unnecessary spaces from column names to avoid indexing issues. Next, the dataset was checked for missing values, but none were found. Out of 96 financial indicators, six were chosen for analysis based on domain knowledge and correlation strength. These indicators include ROA(A) before interest and % after tax, Operating Profit Rate, Debt ratio %, Current Ratio, Quick Ratio, and Net Income to Total Assets. Finally, the selected numerical features were standardized using StandardScaler() to normalize differences in scale across variables. The data was then split into training and test sets to evaluate the predictive capability of machine learning models.

**3. Key Findings**

Exploratory Data Analysis revealed significant financial differences between bankrupt and non-bankrupt companies. Companies that went bankrupt exhibited significantly lower profitability, as evidenced by lower ROA and Operating Profit Rate values. Bankrupt companies had significantly higher debt ratios, indicating over-leverage and financial instability. Firms that went bankrupt had much lower Current Ratio and Quick Ratio, suggesting weaker short-term financial health. A strong correlation was observed between Net Income to Total Assets and bankruptcy, where lower net income signaled a higher likelihood of failure.

A heatmap of feature correlations showed strong negative correlation between Debt Ratio % and ROA, indicating that highly indebted firms struggle to generate returns. Quick Ratio and Current Ratio were positively correlated with financial stability, reinforcing the importance of liquidity. Net Income to Total Assets was one of the most predictive indicators of bankruptcy. Additionally, companies with higher levels of financial leverage tended to be more vulnerable to economic downturns, leading to increased risk of bankruptcy.

도표, 평면도, 기술 도면, 개략도이(가) 표시된 사진

자동 생성된 설명사각형, 도표, 스크린샷, 평면도이(가) 표시된 사진

자동 생성된 설명

**4. Machine Learning Model Performance**

To evaluate whether these financial indicators could predict bankruptcy, a Random Forest Classifier was trained and tested. The model achieved an accuracy of 96.7%, indicating a high level of correct classifications. The precision for bankrupt firms was 46.2%, meaning the model has moderate confidenc..e in predicting actual bankruptcies. The recall for bankrupt firms was 13.6%, suggesting that while the model is precise, it misses many bankrupt companies. The confusion matrix showed that the model correctly identified most non-bankrupt firms but struggled with bankrupt ones due to data imbalance, as bankrupt firms account for only about 3.2% of the dataset. Despite these limitations, the model provides valuable insights into financial risk assessment and highlights areas where further refinement is necessary.

**5. Discussion**

While the model demonstrated strong overall accuracy, its recall for predicting bankrupt firms remained low. This suggests that class imbalance in the dataset significantly impacts predictive performance. Addressing this issue through techniques such as SMOTE (Synthetic Minority Over-sampling Technique) could enhance model effectiveness. Additionally, testing alternative machine learning models, such as XGBoost or Logistic Regression, could yield better recall rates. Incorporating external macroeconomic factors, such as interest rates, GDP growth, and industry trends, could further improve the predictive power of the model. Understanding these financial signals allows businesses and investors to take early action in mitigating risks associated with financial distress.

**6. Conclusion and Recommendations**

This analysis highlights key financial indicators associated with corporate bankruptcy and demonstrates the potential of machine learning in predicting financial distress. However, despite high accuracy, the model's recall for bankrupt firms is relatively low due to data imbalance. Future improvements could include using oversampling techniques such as SMOTE to better represent bankrupt firms in the dataset. Exploring alternative machine learning models such as XGBoost and Logistic Regression could also improve recall. Incorporating external macroeconomic factors such as interest rates and market trends could enhance predictive power. Additionally, firms can leverage these findings to implement early intervention strategies, such as restructuring debt or optimizing cash flow management, to reduce the risk of financial distress. By understanding financial warning signs, businesses, investors, and regulators can take proactive measures to prevent corporate failures, making data-driven insights invaluable in financial risk assessment.

**7. Appendices**

Additional details on data preprocessing steps, feature selection methodologies, and model hyperparameter tuning can be found in the appendix. Supporting visualizations and extended correlation matrices are also included to provide further insights into the analysis.

ChatGPT was used to assist with code generation, data processing, and structuring the written report. It provided relevant dataset recommendations, generated exploratory data analysis scripts, and helped streamline the machine learning workflow. While the AI-generated code was mostly correct, minor adjustments were necessary, such as handling column inconsistencies and improving model evaluation methods. Overall, the AI significantly accelerated the analysis process, but human oversight was required to ensure accuracy and relevance.