Bankruptcy Prediction

Akhilesh Negi
Roll No.: 221AI008

Sachin Choudhary
Roll No.: 221AI034

Maniyar Priyanshu Sanjay
Roll No.: 221AI023

Tushar Kanda
Roll No.: 221AI042

***Abstract*—**

[Importance] Bankruptcy prediction plays a vital role in financial risk management, offering early warnings that help stakeholders mitigate potential financial losses. Traditional bankruptcy models, such as the Altman Z-score, rely heavily on linear assumptions and historical financial ratios, limiting their effectiveness in dynamic financial environments.

[Research Gap] Traditional models fail to capture the non-linear relationships in complex financial data, which can reduce the precision of bankruptcy predictions. Existing machine learning models, while powerful, also face challenges in interpretability, data imbalance, and generalizability across industries.

[Objective] This study aims to create a more accurate and interpretable bankruptcy prediction model by combining multiple machine learning algorithms, including Logistic Regression, XGBoost, and Random Forest, in a Voting Classifier ensemble approach. The objective is to enhance prediction accuracy while addressing issues of data imbalance and ensuring broader applicability across various industry sectors.

[Methodology] Our methodology integrates techniques such as SMOTE (Synthetic Minority Over-sampling Technique) to address class imbalance in the dataset, alongside model performance evaluation metrics including accuracy, ROC AUC, and F1-scores. The ensemble Voting Classifier is designed to leverage the strengths of individual algorithms while minimizing their limitations.

[Results] The Voting Classifier outperformed traditional models like the Altman Z-score and other individual machine learning models, achieving an accuracy of 96.63% and a micro-averaged ROC AUC of 99.6%. This performance was consistent across all classes, demonstrating the model's capability to handle different levels of financial distress and bankruptcy risk effectively.

[Implications] This research contributes a highly accurate and interpretable bankruptcy prediction model that provides financial institutions with a practical tool for assessing risk. By enhancing model generalizability across various sectors, the study supports better risk management strategies applicable to different economic contexts. Additionally, the use of SMOTE and ensemble learning addresses common issues in bankruptcy prediction, including class imbalance and interpretability, making the model both adaptable and robust.

Overleaf Link.

***Index Terms*—Bankruptcy Prediction, Machine Learning, Natural Language Processing, data imbalance, SHAP, Ensemble Learning.**

I. Introduction

[Problem Statement] This project aims to predict bankruptcy for small and medium-sized enterprises (SMEs) based on financial ratios, focusing on interpretability and user-friendly reporting.

[Background] Bankruptcy prediction is a crucial challenge for SMEs, whose financial stability is often more volatile than that of larger firms. Traditional bankruptcy prediction models, such as the Altman Z-score [1], rely on static financial ratios to assess a company's likelihood of distress. While somewhat effective, these models fall short in capturing complex, non-linear relationships present in modern financial data, particularly for SMEs. Recent machine learning models, including Random Forest and XGBoost, have demonstrated improved predictive accuracy by identifying intricate patterns [2], [3]. However, their "black-box" nature limits their transparency for stakeholders who require interpretable reasoning behind predictions.

[Input and Output] In this project, the input data consists of financial ratios from a private dataset of 427 companies, with 66 features representing key financial metrics. The model outputs the predicted financial status of a company, classified as 'Healthy', 'Bankrupt', 'Probable Bankrupt,' or 'Financial Distress,' along with an interpretive explanation of the top three financial factors influencing the decision.

[Applications] This project has practical applications in the financial sector, specifically for business owners, financial analysts, and investors who require a reliable tool for predicting financial distress. The motivation behind this project is the growing demand for accurate and interpretable bankruptcy prediction tools that empower businesses to take proactive measures against financial collapse and help investors make well-informed decisions.

[Uniqueness] What sets this project apart is its integration of SHapley Additive exPlanations (SHAP) for interpretability, providing clear and detailed insights into why a particular company is classified as bankrupt, financially distressed, or healthy. Additionally, the model includes Natural Language Generation (NLG) to generate automated reports in plain language, making the results accessible and actionable for non-technical users.

[Contributions] The key contributions of this project include:

- A hybrid model leveraging ensemble learning methods (Logistic Regression, Random Forest, and XGBoost) to enhance accuracy in predicting the financial status of companies.
- The use of SHAP for interpretable insights into model decisions, improving transparency and trust [4], [5].
- An automated report generation feature via NLG, making the model's outputs understandable to non-technical users.

[Overview] The paper is structured as follows: *Section 2* reviews related works on bankruptcy prediction and machine learning methods, *Section 3* outlines the methodology, covering data preprocessing, model training, and SHAP analysis, *Section 4* discusses the results and evaluates the model's performance, and *Section 5* concludes with a summary of findings and future research directions.

II. Literature Review

The field of bankruptcy prediction has evolved significantly over the years, transitioning from traditional statistical models to more sophisticated machine learning techniques. This literature review compares and contrasts the different approaches, highlighting their strengths, weaknesses, and research gaps essential for advancing the field.

A. Traditional Approaches

Historically, bankruptcy prediction models have relied on statistical methods, particularly the use of financial ratios. Altman's Z-score model [1] is one of the most widely applied in this area, using a linear combination of various financial ratios to predict bankruptcy likelihood with considerable accuracy. However, this approach assumes linear relationships and relies heavily on historical financial data, limiting its effectiveness in dynamic markets with changing economic conditions [6]. While interpretable, these models lack the sophistication needed to capture complex, non-linear patterns in modern financial data.

B. Machine Learning Approaches

Machine learning has introduced new methods for handling non-linear relationships and complex interactions among financial metrics. Support Vector Machines (SVMs), Decision Trees, and Random Forests are notable techniques that have shown improved predictive power over traditional models. Studies by Zieba et al. [2] reveal that decision tree-based ensembles, such as Random Forests and Gradient Boosting, often outperform traditional models in accuracy. However, the "black-box" nature of machine learning models creates interpretability challenges, as their decision-making process is difficult to understand [7]. Techniques such as SHAP values and LIME are often required to explain feature importance, and issues like overfitting and large feature spaces persist, especially in datasets with high dimensionality.

C. Ensemble Models and Hybrid Approaches

Recent studies have focused on hybrid models that combine multiple algorithms to improve predictive accuracy. Ainan et al. [8] proposed a model combining machine learning methods with neural networks, enhancing predictive power by capturing more intricate patterns in financial data. Although these ensemble and hybrid approaches can yield high accuracy, they also require greater computational resources and can be more challenging to interpret. Additionally, overfitting is a concern, especially when these models are applied to smaller or imbalanced datasets [9].

D. Addressing Data Imbalance

Data imbalance poses a critical challenge in bankruptcy prediction, as bankrupt companies are often underrepresented in datasets. This imbalance can bias models toward non-bankrupt predictions. Techniques like oversampling, undersampling, and cost-sensitive learning have been applied to mitigate this issue. Zou et al. [4] utilized cost-sensitive learning to reduce bias toward the majority class, showing improvements in prediction accuracy. However, oversampling can introduce noise, while undersampling may lead to a loss of valuable information [5]. Finding an optimal balance without introducing bias remains a major challenge in the field.

E. Generalizability Across Industries

Ensuring model generalizability across industries and regions is a key issue. Many studies focus on a specific sector, leading to models that may not perform well outside their original domain. Khalil et al. [10] conducted cross-industry validation, applying models across various sectors to assess robustness. Results indicate that machine learning models require fine-tuning to perform consistently across industries, underscoring a gap in current research where adaptive models are needed for diverse economic contexts.

F. Gaps and Significance of the Research

[Literature Gap] From the reviewed literature, several research gaps are identified:

- Limited interpretability in complex machine learning models.
- Persistent data imbalance issues, with existing techniques affecting model performance due to trade-offs.
- Lack of generalizability, as models are often tailored for specific sectors or regions.
- Challenges in handling large feature spaces without overfitting, particularly with advanced machine learning models [3].

[Outcome of Literature Review] This study addresses these gaps by developing an ensemble-based machine learning model that emphasizes interpretable predictions, applies oversampling techniques for class balance, and uses cross-industry validation to enhance generalizability. The research thereby contributes to advancing bankruptcy prediction models with improved accuracy and adaptability across various financial settings.

III. Methodology

A. Data Collection and Preprocessing

We used a private dataset containing bankruptcy data for 427 companies, covering multiple years. The dataset consists of 3,576 rows and 66 financial features (x1, x2, …, x66). The target variable ("Label") contains four categories: 'Healthy', 'Bankrupt', 'Financial Distress', and 'Probable Bankrupt.'

To ensure data quality, we addressed missing values by imputing numeric values using the mean and categorical values

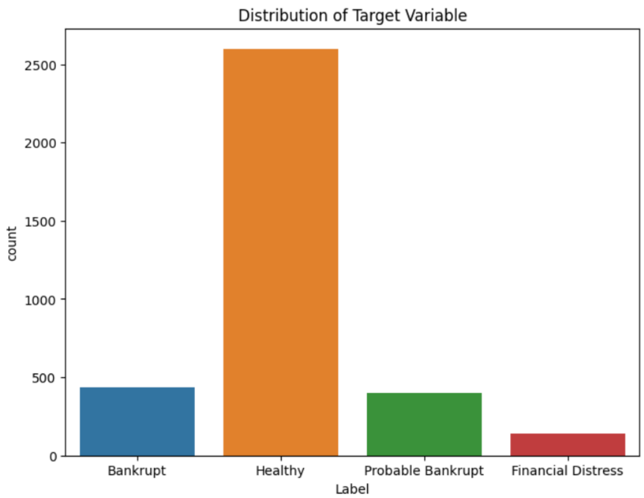

Fig. 1. This figure shows the highly skewed nature of dataset towards 'Healthy' class

using the mode. Outliers were detected and treated using z-scores and the IQR method. All numerical features were normalized using *StandardScaler* to maintain consistency.

Feature engineering focused on creating additional financial metrics to enhance the model's predictive power, particularly targeting profitability and liquidity ratios. Since the dataset showed imbalances (Fig. 1) between the four classes (with more 'Healthy' companies than those in other classes), we used the Synthetic Minority Oversampling Technique (SMOTE) to balance the class distribution and prevent model bias toward the majority class [8].

B. Model Training and Evaluation

We experimented with several machine learning models, including Logistic Regression, Random Forest, and XGBoost. These models were chosen for their ability to handle both linear and non-linear relationships in the financial data [11]. To improve accuracy, we combined predictions using a Voting Classifier, which aggregates the strengths of each individual model.

The dataset was split 80/20 for training and testing purposes. We evaluated the models using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to assess predictive performance. Cross-validation was employed to ensure that the models were robust and could generalize effectively to unseen data [7], [10].

C. Interpretability and User Interaction

To make the models more interpretable, we utilized SHAP (SHapley Additive exPlanations) to explain individual predictions. SHAP values provide insights into the most significant financial features influencing predictions, while user-specific SHAP analyses offer clear explanations for each company's bankruptcy risk classification. This helps businesses understand the key financial metrics impacting their predictions, ensuring both accuracy and interpretability.

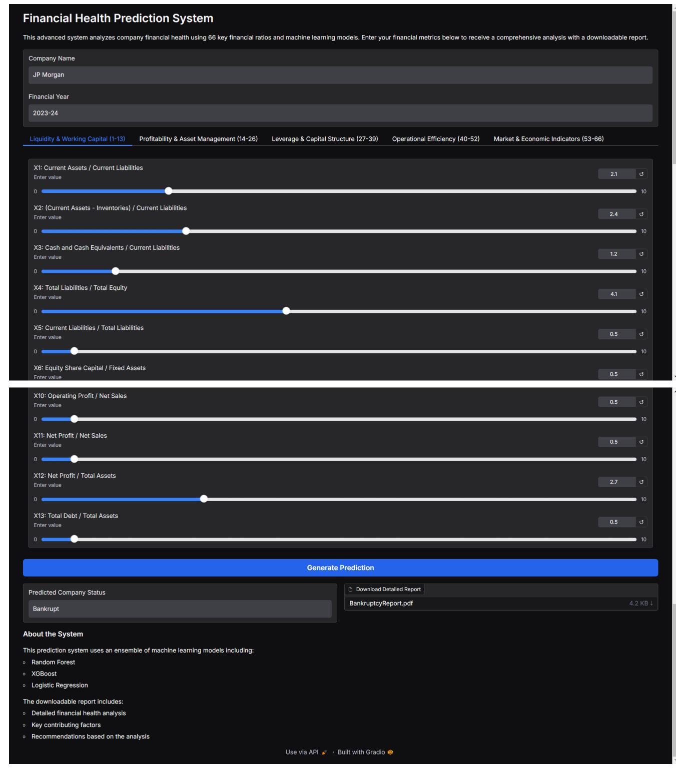

Fig. 2. User Interface developed using Gradio for bankruptcy prediction

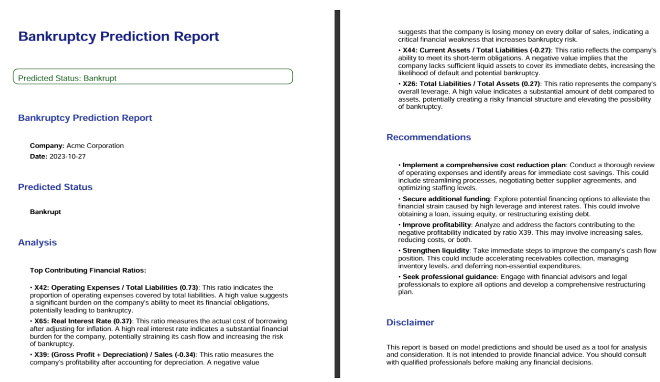

Fig. 3. The generated report contains the predicted status along with top influential ratios and recommendations.

We developed a user interface using Gradio (Fig. 2), creating an interactive web application where business users can input the required 66 financial metrics and receive instant predictions about their company's financial health. The interface provides an intuitive platform for users to enter their financial data and obtain comprehensive bankruptcy risk assessments. In addition to the prediction, SHAP-based explanations are provided to identify the top three financial features influencing the model's decision.

D. Automated Report Generation and Continuous Improvement

We integrated Natural Language Generation (NLG) technology to automatically generate narrative reports (3 summarizing

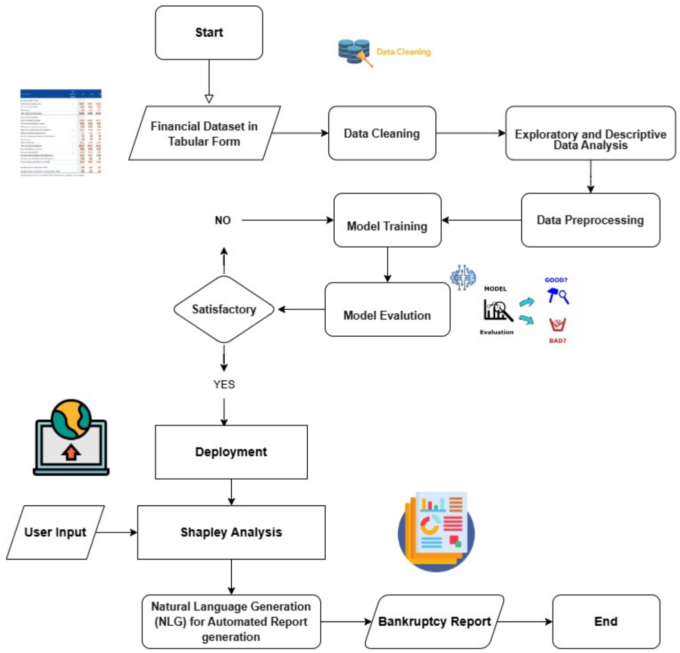

Fig. 4. Block Diagram showing the complete methodology workflow

the predicted bankruptcy status, key financial ratios affecting the outcome, and visualizations of SHAP values. We used Gemini API to generate recommendaions on how to deal with the current financial situation.

To ensure the system remains accurate and relevant, we implemented a feedback system for continuous improvement. In this, the users can flag any results which they deem as incorrect. These will be sent to the admistrator to further improve the model.

E. Flowchart Overview

The methodology for predicting bankruptcy in Small and Medium Enterprises (SMEs) follows a structured process, starting from data analysis to the generation of a final bankruptcy prediction report. This process is shown in Fig. 4. Below is a detailed breakdown of the steps involved in the bankruptcy prediction process:

1) Financial Dataset in Tabular Form: The process starts with the private dataset of 427 companies and 66 financial metrics that represent essential financial ratios. These metrics form the foundation for predicting bankruptcy by providing a comprehensive view of the company's financial health.

2) Data Cleaning: The raw data undergoes a thorough *data cleaning* process, eliminating inconsistencies, missing values, and outliers. This is crucial for ensuring the reliability of the dataset and boosting model performance.

3) Exploratory and Descriptive Data Analysis: Simultaneously with data cleaning, *exploratory and descriptive data analysis* is conducted. This involves visualizing trends, calculating summary statistics, and identifying patterns in the dataset, which guides the next steps in the model-building process.

4) Data Preprocessing: After cleaning, the data is prepared for model training through *data preprocessing*. Key techniques include normalization and SMOTE for class balancing, which help to enhance the model's prediction accuracy and efficiency.

5) Model Training: With the preprocessed data, different machine learning algorithms are applied during the *model training* stage. These include models like Logistic Regression, XGBoost, and Random Forest, all trained to detect patterns in financial data that indicate bankruptcy risk.

6) Model Evaluation: Following training, the models are assessed using various *evaluation metrics*, such as accuracy, precision, recall, F1-score, and ROC-AUC. This stage ensures that the model's performance meets the required standards before deployment.

7) Deployment: Once the model shows satisfactory results, it is *deployed* using Gradio for real-time predictions. Stakeholders can then input new financial data through the web interface and receive instant predictions on bankruptcy likelihood.

8) User Input: The *user input* stage allows financial analysts or business owners to input the 66 financial ratios through the Gradio interface. The model processes these inputs to make real-time bankruptcy risk assessments.

9) Shapley Analysis (SHAP): To increase transparency, *Shapley Analysis (SHAP)* is applied, giving users insight into the financial metrics that most influence the model's decisions. This interpretability makes the model more trustworthy and actionable.

10) Natural Language Generation (NLG): Once a prediction is made, *Natural Language Generation (NLG)* is used to automatically generate a human-readable report, explaining the prediction and highlighting the key financial features that contributed to the decision.

11) Bankruptcy Report: The final output is a comprehensive *bankruptcy report*, which summarizes the company's risk status, highlights significant financial factors, and offers recommendations for mitigating risk.

This multi-step approach ensures a robust, interpretable, and user-friendly bankruptcy prediction system. By combining machine learning with SHAP for interpretability and NLG for report generation, this project offers a comprehensive and transparent solution for predicting financial distress in SMEs.

IV. Results and Analysis

A. Model Performance Overview

We evaluated the performance of various machine learning models on the bankruptcy prediction task using accuracy, micro-averaged ROC AUC, and classification reports. The models tested include Logistic Regression, XGBoost, Random Forest, and a Voting Classifier, as well as a comparison with an SVM model and the traditional Altman Z-Score model. The dataset contained 3,576 rows, and an 80/20 train-test split was applied, resulting in 715 test instances for evaluation.

Table I summarizes the accuracy and micro-averaged ROC AUC scores for each model.

TABLE I
ACCURACY AND MICRO-AVERAGED ROC AUC SCORES FOR DIFFERENT MODELS

Model	Accuracy	Micro-Averaged ROC AUC
Logistic Regression	0.8072	0.953
XGBoost	0.9750	0.999
Random Forest	0.9663	0.997
Voting Classifier (Our Model)	0.9663	0.996
SVM	0.8572	0.971
Altman Z-Score Model	0.8072	0.953

Among the tested models, the Voting Classifier, which combines Logistic Regression, XGBoost, and Random Forest, achieved the an accuracy of 0.9663 and a micro-averaged ROC AUC of 0.996, demonstrating strong and consistent performance. The Random Forest classifier achieved an accuracy of 0.9663 and the best ROC AUC score (0.997). The XGBoost model performed exceptionally well with highest accuracy of 0.9750 and an ROC AUC of 0.999.

B. Phase-Wise Results

The results are divided into different phases for detailed evaluation of the model's performance:

Phase 1: Preprocessing and Feature Engineering In the first phase, preprocessing involved cleaning and normalizing the data. We also created additional financial metrics. During this phase, outliers were removed, and features were standardized. The preprocessing phase helped the models achieve higher accuracy by reducing noise and ensuring that data was presented uniformly.

Phase 2: Model Training In the second phase, we trained various models: Logistic Regression, Random Forest, XGBoost, and a Voting Classifier. The models were tested on the same dataset, and their performance was evaluated based on classification metrics.

Phase 3: Performance Evaluation In this phase, models were assessed using various metrics, including accuracy, micro-averaged ROC AUC, precision, recall, and F1-score. This detailed evaluation helped compare the models and select the best-performing model.

Phase 4: Model Comparison and Final Selection The final phase involved comparing the performance of the Voting Classifier with traditional models like the Altman Z-Score and SVM. The results showed that the Voting Classifier outperformed these models significantly, achieving an accuracy of 0.9663 and an F1-weighted score of 0.96.

Phase 5: Model Deployment The Voting Classifier model was deployed with the help of Gradio. The SHAP values were used for finding the most influential features. Gemini API was used to generate a suitable report along with recommendations on how to deal with the situation.

C. Classification Report Analysis

The detailed performance of the Voting Classifier, as shown in Table II, indicates that it performed consistently well across all four classes. The model achieved an overall accuracy of 0.96, with precision, recall, and F1-scores all maintaining balanced values across the different classes. This suggests that the Voting Classifier handles all classes with relatively equal effectiveness, minimizing bias towards any particular class.

TABLE II
CLASSIFICATION REPORT FOR VOTING CLASSIFIER

Class	Precision	Recall	F1-Score	Support
Bankrupt	0.97	0.94	0.96	529
Financial Distress	0.97	1.00	0.98	516
Healthy	0.98	0.95	0.97	522
Probable Bankrupt	0.95	0.97	0.96	513
Accuracy		0.96		2080
Macro Avg	0.97	0.97	0.97	2080
Weighted Avg	0.97	0.97	0.97	2080

The class-wise performance shows that the model is effective in predicting the 'Healthy' category, achieving the highest precision of 0.98, along with a recall of 0.95 and an F1-score of 0.97. The model also demonstrated exceptional performance in identifying companies under 'Financial Distress', with a perfect recall of 1.00 and a precision of 0.97, resulting in an F1-score of 0.98 — the highest among all classes.

The balanced distribution of support values across classes (ranging from 513 to 529 instances) validates our data preprocessing strategy for addressing class imbalance, further reinforcing the reliability of the model's strong performance metrics across all categories.

D. Comparison with Other Models

In comparison with other models, the Voting Classifier outperformed the SVM and Altman Z-Score models. Table III summarizes the accuracy and F1-scores for each model.

TABLE III
COMPARISON OF VOTING CLASSIFIER WITH OTHER MODELS

Model	Accuracy	F1-Score (Weighted Avg)
SVM	0.8572	0.89
Altman Z-Score	0.8072	0.82
Voting Classifier	0.9663	0.96

The SVM model achieved an accuracy of 0.8572, while the Altman Z-Score model had an accuracy of 0.8072. The Voting Classifier provided a significant improvement in performance, with a higher accuracy of 0.9663 and better F1-scores across all categories.

E. Discussion

The Voting Classifier, incorporating Logistic Regression, Random Forest, and XGBoost, offers a robust solution for predicting financial distress. Its balanced performance across multiple classes indicates its suitability for real-world applications, where identifying high-risk companies is crucial. Although XGBoost marginally outperformed the Voting Classifier in terms of accuracy (0.9750 vs 0.9663), the Voting Classifier's performance was consistent and interpretable across all categories, making it a reliable choice for stakeholders.

The comparison with traditional models like Altman Z-Score further demonstrates the advantages of using modern machine learning techniques for bankruptcy prediction. While Altman Z-Score remains a popular choice in financial analysis, it lacks the predictive power and adaptability of machine learning models.

In terms of improvements over existing work, our method significantly outperforms traditional models such as SVM and Altman Z-Score, demonstrating better accuracy and F1-scores. The use of ensemble learning in the Voting Classifier allowed us to capture the strengths of multiple models, leading to improved performance. Moreover, the integration of SHAP for model interpretability and the development of an interactive Gradio interface for end-user engagement represents a significant enhancement in terms of both transparency and usability. This makes the model more accessible and understandable for decision-makers in the financial sector.

V. Conclusion

[Key Findings] The Voting Classifier, a combination of Logistic Regression, XGBoost, and Random Forest, achieved an impressive accuracy of 0.9663 and a micro-averaged ROC AUC score of 0.996, significantly outperforming traditional models such as SVM and Altman Z-Score. This demonstrates the power of ensemble learning in bankruptcy prediction.

[Interpretations] The model's superior performance suggests that machine learning techniques, especially ensemble methods, offer a more accurate and adaptive approach to bankruptcy prediction. The Voting Classifier effectively identifies high-risk companies, particularly in the 'Financial Distress' and 'Probable Bankrupt' categories.

[Implications] This research provides financial institutions and analysts with a more reliable tool for assessing the financial health of companies. By integrating machine learning techniques, the model can enhance decision-making processes, reduce financial risks, and improve overall financial management strategies.

[Limitations] The dataset used in this study is limited in size and scope, potentially affecting the model's generalization to other sectors or economic conditions. Additionally, the model's reliance on historical financial data may not account for unforeseen economic shifts, which could impact its predictive power.

[Recommendations] Future research should focus on expanding the dataset to include more diverse company profiles and financial indicators. Additionally, incorporating macroeconomic factors and exploring hybrid models could further enhance the model's predictive accuracy and robustness.

[Future Enhancements] Future work could integrate real-time data and advanced machine learning models such as deep learning to improve predictive accuracy and provide dynamic bankruptcy risk assessments. The platform could be made more accessible through user interface improvements, focusing on interactive features and intuitive visualization tools. Additionally, future research could explore knowledge graph embedding techniques [12] and sophisticated information extraction pipelines [13] to incorporate broader financial and industry-specific data into the prediction models.

References

[1] E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *The Journal of Finance*, vol. 23, no. 4, pp. 589–609, 1968. [Online]. Available: http://www.jstor.org/stable/2978933

[2] M. Ziba, S. K. Tomczak, and J. M. Tomczak, "Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction," *Expert Syst. Appl.*, vol. 58, pp. 93–101, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:40512567

[3] H. Son, C. Hyun, D. Phan, and H. Hwang, "Data analytic approach for bankruptcy prediction," *Expert Systems with Applications*, vol. 138, p. 112816, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417419305123

[4] Y. Zou, C. Gao, and H. Gao, "Business failure prediction based on a cost-sensitive extreme gradient boosting machine," *IEEE Access*, vol. 10, pp. 42 623–42 639, 2022.

[5] W. Yotsawat, K. Phodong, T. Promrat, and P. Wattuya, "Bankruptcy prediction model using cost-sensitive extreme gradient boosting in the context of imbalanced datasets," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, p. 4683, 08 2023.

[6] Željko Đ. Vujovic, "Classification model evaluation metrics," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, 2021. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2021.0120670

[7] A. G. Kim and S. Yoon, "Corporate bankruptcy prediction with domain-adapted bert," in *Proceedings of the Third Workshop on Economics and Natural Language Processing*. Association for Computational Linguistics, 2021. [Online]. Available: http://dx.doi.org/10.18653/v1/2021.econlp-1.4

[8] U. H. Ainan, L. Y. Por, Y.-L. Chen, J. Yang, and C. S. Ku, "Advancing bankruptcy forecasting with hybrid machine learning techniques: Insights from an unbalanced polish dataset," *IEEE Access*, vol. 12, pp. 9369–9381, 2024.

[9] G. Perboli and E. Arabnezhad, "A machine learning-based dss for mid and long-term company crisis prediction," *Expert Systems with Applications*, vol. 174, p. 114758, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421001998

[10] A. A. Khalil, Z. Liu, A. Salah, A. Fathalla, and A. Ali, "Predicting insolvency of insurance companies in egyptian market using bagging and boosting ensemble techniques," *IEEE Access*, vol. 10, pp. 117 304–117 314, 2022.

[11] S. Shetty, M. Musa, and X. Brédart, "Bankruptcy prediction using machine learning techniques," *Journal of Risk and Financial Management*, vol. 15, no. 1, 2022. [Online]. Available: https://www.mdpi.com/1911-8074/15/1/35

[12] Q. Wang, Z. Mao, B. Wang, and L. Guo, "Knowledge graph embedding: A survey of approaches and applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 12, pp. 2724–2743, 2017.

[13] M. Y. Jaradeh, K. Singh, M. Stocker, A. Both, and S. Auer, "Information extraction pipelines for knowledge graphs," *Knowl. Inf. Syst.*, vol. 65, no. 5, p. 1989–2016, jan 2023. [Online]. Available: https://doi.org/10.1007/s10115-022-01826-x