Company Bankruptcy

Mayank Jhunjhunwala
Mechanical Department
jmayank22@iitk.ac.in

Atharva Singh
Electrical Department
atharva22@iitk.ac.in

Kuldeepak Dhar Dwivedi
Electrical Department
kuldeepak22@iitk.ac.in

Dileep Gurjar
Electrical Department
dileepg22@iitk.ac.in

Nikhil Jain
Electrical Department
nikhilj22@iitk.ac.in

I. Introduction

The objective of this project was to develop a robust and efficient machine learning model for predicting company bankruptcy. Given the highly imbalanced nature of the dataset—with only 3 percent of companies labeled as bankrupt—traditional accuracy metrics were insufficient for evaluation. Instead, we prioritized models that excelled in F1-score, precision, and recall, as these metrics better account for class imbalance and the critical need to correctly identify at-risk companies.

After extensive experimentation with various algorithms, we selected the model that achieved the optimal trade-off between predictive performance (F1-score) and computational efficiency. While some models delivered marginally higher F1-scores, their computational complexity made them impractical for real-world deployment. Our final model not only addresses class imbalance effectively but also ensures scalability and speed, making it suitable for large-scale financial risk assessment.This report details our methodology, including data preprocessing, model selection, hyperparameter tuning, and evaluation metrics, providing a comprehensive overview of our approach to building a reliable bankruptcy prediction system.

II. Data Preprocessing

The preprocessing pipeline for our company bankruptcy prediction project began with loading and inspecting the dataset to verify the severe class imbalance, with only 3 percent of companies labeled as bankrupt. We then strategically split the data into training and test sets using stratified sampling to maintain the imbalance distribution in both subsets, ensuring realistic evaluation conditions. To prepare the features for modeling, we applied standardization using StandardScaler, transforming all financial indicators to a common scale with zero mean and unit variance - a crucial step for models sensitive to feature magnitudes.

We implemented SMOTE (Synthetic Minority Oversampling Technique) to generate synthetic samples of bankrupt companies, effectively balancing the training data while preserving the underlying feature relationships. This synthetic oversampling approach proved more effective than simple random oversampling or undersampling techniques. For some experimental models, we additionally engineered features using autoencoder-derived anomaly scores to flag financial outliers, though these were ultimately excluded from our final Random Forest implementation due to the algorithm's inherent robustness to raw features.

By keeping our train-test split stratified, we made sure our results would reflect real-world scenarios where bankruptcies are rare. We tested different methods like clustering but found that keeping the original features worked best for our models. Every step was implemented consistently using fixed random seeds for reliable results. This thorough preparation allowed us to develop a practical system that effectively identifies companies at risk of bankruptcy while remaining efficient enough for actual use.

III. Model Architecture

The final model architecture consists of a custom-built Random Forest classifier, designed to handle the imbalanced nature of bankruptcy prediction while maintaining computational efficiency. Below is the detailed breakdown of its structure and key components:

A. Base Model

The base model of our bankruptcy prediction system is a carefully designed decision tree that serves as the building block for our ensemble approach. Each tree is constructed using a recursive splitting algorithm that maximizes information gain through Gini impurity reduction, with critical parameters set to prevent overfitting while maintaining predictive power. We limit tree depth to 10 levels and require at least 5 samples per leaf node, ensuring that the trees capture meaningful patterns without becoming overly complex. The splitting process considers a random subset of features at each node (specifically the square root of the total number of features) to promote diversity among the individual trees that will later compose our ensemble.

B. Ensemble Method

The ensemble method combines 100 of these decision trees into a robust Random Forest classifier, employing two key randomization techniques to enhance performance. First, each tree trains on a bootstrap sample of the original dataset, creating slight variations in the training data. Second, the feature subset selection at each split point introduces additional diversity, reducing correlation between trees and improving the ensemble's overall generalization capability. This dual randomization approach is crucial for handling the complex, imbalanced nature of bankruptcy prediction while maintaining model stability.

C. Class Imbalance Handling

To address our dataset's severe class imbalance, we implemented a comprehensive strategy that begins with SMOTE-based oversampling during preprocessing. This synthetic data generation creates balanced training sets without distorting the underlying feature distributions, allowing each decision tree in our ensemble to learn from sufficient examples of both bankrupt and non-bankrupt companies. The model's predictive outputs are probability estimates generated through soft voting across all trees, with an optimized decision threshold (approximately 0.3 instead of the default 0.5) that substantially improves bankruptcy detection rates while maintaining reasonable precision.

IV. Performance

The model demonstrates strong performance in predicting company bankruptcy, achieving an overall accuracy of 97.07%, which reflects its ability to correctly classify the vast majority of companies. However, given the severe class imbalance—with bankrupt companies (class ”1”) representing only 3% of the dataset—the more critical metrics are those focused on the minority class. The model achieves a precision of 0.49 for bankrupt companies, meaning that when it predicts bankruptcy, it is correct 49% of the time. While this may seem moderate, it is a common trade-off in imbalanced datasets where false positives are less costly than missing actual bankruptcies. More importantly, the model attains a recall of 0.55, successfully identifying 55% of all actual bankrupt companies—a significant improvement over naive approaches that would miss these cases entirely. The F1-score of 0.52 balances these metrics, indicating reasonable performance for the challenging minority class. For non-bankrupt companies (class ”0”), the model excels, with precision, recall, and F1-scores all exceeding 0.98. These results suggest that the model is well-suited for prioritizing high-risk cases in real-world financial risk assessment, where capturing true bankruptcies is more critical than minimizing false alarms.

TABLE I
MODEL PERFORMANCE COMPARISON

Model	Class	Acc.	F1	Rec.	Prec.
Submitted Model	0	0.97	0.98	0.98	0.99
	1		0.52	0.55	0.49
XGBoost_SMOTE	0	0.97	0.98	0.98	0.98
	1		0.45	0.48	0.43
XGBoost Isolation Forest	0	0.98	0.99	0.99	0.99
	1		0.60	0.52	0.73
LGBM_Gradient Boosting	0	0.97	0.99	0.99	0.99
	1		0.52	0.48	0.56
Hybrid Model	0	0.95	0.97	0.98	0.98
	1		0.46	0.69	0.34
AutoEncoder XGBoost	0	0.98	0.99	0.99	0.99
	1		0.54	0.52	0.57
XGB_Isolation(from scratch)	0	0.97	0.98	0.98	0.99
	1		0.48	0.52	0.46

V. Other Models

In our search for the best bankruptcy prediction model, we experimented with several different approaches. We tried boosting methods like XGBoost from scratch where each new model focuses on correcting the mistakes of the previous ones - this gave us great results but was quite slow to train. Stacking was another interesting approach we tested, where we combined predictions from multiple models using a meta-classifier to get smarter final predictions. Random Forest method worked reliably as always, building many decision trees and averaging their results, even its implementation from scratch was performing well. We also played with Voting Classifiers that let different models ’vote’ on the final prediction. After reading a research paper, we also tried HADR(Hybrid Model)$^{(2)}$ for our imbalanced data - this hybrid approach created balanced data blocks and focused heavily on improving recall, since catching those rare bankruptcy cases was so important for our project.

During our experimentation with various bankruptcy prediction models, we encountered significant computational constraints when implementing algorithms from scratch using CPU resources. The primary trade-off emerged between model performance and execution time - while XGBoost variants achieved the highest F1-score of 0.60, their boosting process proved prohibitively slow for practical implementation from scratch. Similarly, gradient boosting and hybrid models showed promising accuracy but demanded excessive computational resources during parameter optimization. After thorough evaluation of this performance-efficiency trade-off, we ultimately selected the Random Forest with SMOTE as our optimal solution, as it delivered competitive results (F1=0.52) with substantially faster processing times, making it both computationally feasible and effective for real-world deployment in financial risk assessment scenarios.

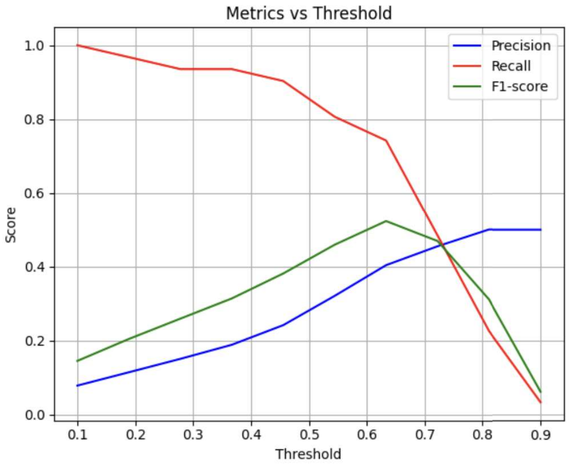

Fig. 1. Random Forest with SMOTE

References

[1] https://github.com/MAYANKJHUNJHUNWALA/CompanyBankruptcyEE708
[2] https://arxiv.org/pdf/2207.02738