FINDINGS FORMAT TO EXAMINATION

NOTIFICATION NO: 586/2025

NOTIFICATION DATE: 30/09/2025

NAME OF SCHOLAR: MOHD RAAGIB SHAKEEL

NAME OF SUPERVISOR: DR. TAUFEEQUE AHMAD SIDDIQUI

NAME OF CO-SUPERVISOR: DR. SHAHZAD ALAM

NAME OF DEPARTMENT: DEPARTMENT OF MANAGEMENT STUDIES

TOPIC OF RESEARCH: CORPORATE DEFAULT RISK PREDICTION USING MACHINE LEARNING TECHNIQUES

FINDINGS

1. For the Indian Bankruptcy Dataset (IBD), KNN was the optimal imputation technique for accuracy (0.9618), recall (0.98), F1-score (0.966) and AUC (0.9929), while Mean gave the best precision (0.99). For the Polish Bankruptcy Dataset (PBD), KNN was the optimal imputation technique for accuracy and precision across all five years, and for recall in every year except Year 4 (Mean); for AUC, KNN was best in Years 1, 2 and 5, EM in Year 3 and Mean in Year 4. Thus, in most cases KNN was the optimal imputation technique.

2. For the IBD, XGBoost with KNN achieved the highest accuracy, recall, F1-score and AUC, while XGBoost with Mean achieved the highest precision. For the PBD, XGBoost with KNN was the best combination for accuracy, recall and AUC in most years, while the Stack (RF+GBT+AdaBoost) and Stack 1 (XGB+GBT+RF) ensembles gave the best precision or accuracy in some years.

3. All evaluation metrics are important, but in the context of bankruptcy prediction their significance, in descending order, is AUC score, recall, F1-score, precision and accuracy.

4. The results of the IBD and PBD datasets are consistent, with XGBoost combined with KNN being the leading combination across both datasets, and only minor variations across certain metrics and years.

5. On comparison with previous studies, the XGBoost and stacked-ensemble models with KNN, Mean or EM imputation match or outperform the accuracy, precision, recall, F1-score and AUC reported in prior work.

SUMMARY OF ABSTRACT

Predicting corporate default risk has always been a crucial topic in banking and finance, since the failure of a company causes massive losses not only to market participants such as investors, creditors and financial institutions but to the whole economy. This study predicts corporate default risk using machine learning techniques, taking the data of legally defaulted (bankrupt) companies and using their financial ratios to predict bankruptcy, which can help creditors, investors, policymakers and financial institutions avoid significant losses or prepare for early intervention. In the Indian context, the Insolvency and Bankruptcy Code (IBC), 2016 provides the framework for corporate insolvency, establishing the Insolvency and Bankruptcy Board of India (IBBI) and defining the Corporate Insolvency Resolution Process (CIRP) and liquidation under which defaulting firms are identified. Traditional bankruptcy prediction has relied on statistical models such as discriminant analysis, binary response (logit and probit) models and hazard models. Although interpretable and easy to implement, these models often assume linearity and normality, struggle to capture nonlinear patterns and cope poorly with imbalanced data. Machine learning algorithms provide greater accuracy and adaptability, handle large datasets with many features, deal better with imbalanced data and are more resilient to multicollinearity. This study therefore develops and evaluates a machine learning framework for corporate default risk prediction, focusing on the handling of missing values, class imbalance and the choice of evaluation metrics.

The study uses two datasets. The first is a newly constructed Indian Bankruptcy Dataset (IBD), which solely examines bankrupt Indian companies that officially defaulted by filing under Section 7 of the IBC, 2016, and are undergoing CIRP or are in liquidation, using data from the IBBI for the period 2016-2022. It consists of 2,445 observations, of which 410 are bankrupt and 2,035 non-bankrupts, described by 29 financial ratios. As this dataset is newly constructed, no previous study has used it. The second is the Polish Bankruptcy Dataset (PBD) from the EMIS database, which spans five years with a total of 43,405 cases and contains 64 financial ratios; it is used to evaluate the consistency and robustness of the findings obtained from the IBD.

Both datasets contain missing values. Three imputation techniques - Mean, K-Nearest Neighbour (KNN) and Expectation-Maximization (EM) - were applied separately, and the Synthetic Minority Over-sampling Technique (SMOTE) was used to handle class imbalance. Eleven machine learning classifiers were then applied, including Logistic Regression, Decision

Tree, Support Vector Machine, K-Nearest Neighbour, AdaBoost, Gradient Boosting Tree, Random Forest, XGBoost, Stack (RF+GBT+AdaBoost) and Stack 1 (XGB+GBT+RF). Performance was assessed using seven evaluation metrics: accuracy, precision, recall, F1-score, confusion matrix, AUC score and the AUROC curve.

The results show that KNN was the optimal imputation technique in most cases, and XGBoost combined with KNN was the best classifier-imputation combination. For the IBD, XGBoost with KNN achieved an accuracy of 0.9618, recall of 0.98, F1-score of 0.966 and AUC of 0.9929, while XGBoost with Mean gave the highest precision of 0.99. For the PBD, XGBoost with KNN was the leading combination across the five years, with the stacked ensembles best for precision or accuracy in some years; accuracy ranged from 0.9965 to 0.9998, recall reached 1.00 in every year, and AUC ranged from 0.9997 to 0.9999. Among the metrics, AUC and recall are the most important for bankruptcy prediction, as failing to identify an actual bankruptcy (false negative) is more costly than a false positive. The results of the IBD and PBD are consistent, and the proposed models match or outperform the results of previous studies.

The study carries both policy and managerial implications. For policy, the Early Warning System developed by the Ministry of Corporate Affairs, along with SEBI, can use the XGBoost model with KNN imputation and SMOTE, and recall-driven evaluation, to enhance the early detection of financial distress and support financial stability. For management, banks, investment firms, creditors, suppliers and corporate risk managers can use the proposed models in credit assessment, lending and portfolio decisions and early intervention. The main limitations are the available period of Indian bankruptcy data and the difficulty of capturing sudden shocks such as pandemics or economic crises. Future research may explore additional imputation and feature selection techniques, alternative sampling methods such as ADASYN, Borderline-SMOTE and NearMiss, the integration of qualitative factors such as corporate governance and management quality, and the use of explainable AI and stress-testing.