Predicting Financial Distress Using Machine Learning Models: The Case of Banking Sector in Ethiopia

Belay Getachew
HiLCoE School of Computer Science and Technology, Addis Ababa, Ethiopia
belay.acca@gmail.com

Mesfin Kifle
HiLCoE School of Computer Science and Technology, Addis Ababa, Ethiopia
Kiflemestir95@gmail.com

ABSTRACT

This research delves into the application of machine learning models to predict financial distress within the Ethiopian commercial banking sector. The specific inquiries guiding the study's objective to build an accurate predictive model for bank financial distress. The questions seek to determine how machine learning models can be developed in the Ethiopian banking sector. The study leverages the well-established CAMEL framework, which examines Capital Adequacy, Asset Quality, Management, Earnings, and Liquidity, as well as the Altman Z-score measurements-to label distress banks, to develop a comprehensive predictive model. This study evaluates several machine learning algorithms—including Random Forest and Neural Networks—using thirty years of financial data from 18 Ethiopian commercial banks to identify distress patterns. The researchers have employed optimization techniques such as SMOTE oversampling, hyperparameter tuning, and ensemble methods to refine the models' ability to anticipate financial distress before it materializes.

The findings demonstrate that the Random Forest model- addressed dataset imbalance & overfitting problem by applying SMOTE oversampling combined with class weighting and evaluated through stratified cross- validation demonstrated exceptional performance and achieves an overall accuracy of 97.70% and a Kappa statistic of 0.9458, significantly improving the identification of distressed banks. Followed by another ensemble method - REP (Reduced Error Pruning)-decision trees ensemble method like Bagging exhibit strong predictive capabilities, with an accuracy of 96.07%. The study emphasises the importance of data pre-processing, feature selection, and dimensionality reduction in building robust models. By integrating machine learning with the CAMEL framework, the research provides a data-driven approach to financial distress prediction, offering valuable insights for regulators, bank management, and investors. The study concludes with recommendations for implementing machine learning-based early warning systems, improving data quality, and fostering collaboration among stakeholders to enhance the stability and resilience of the Ethiopian banking sector.

Keywords: Financial *Distress Prediction, Machine Learning, National Bank of Ethiopia, Ethiopian Commercial Banks*

1. Introduction

The challenges faced by banking supervisors is increasing due to the growth in the number and size of banks, emphasizing the need for proactive monitoring of financial health [1, 7]. Traditional methods have had limited success, leading to the adoption of machine learning (ML) techniques, which have shown strong performance in predicting financial distress [10, 21, 15]. Financial distress is a critical condition indicating ineffective management or failure to resolve problems—not necessarily bankruptcy—and early warning models are essential for its prevention. [8]. The National Bank of Ethiopia (NBE) is transitioning to ML due to the limitations of traditional statistical methods in a complex environment. The research focuses on commercial banks within Ethiopia's formal banking sector, which includes the private and public, noting the sector's significance to the national economy and its projected growth [6].

The National Bank of Ethiopia's (NBE) priority to protect depositors by ensuring bank robustness through thorough health and stability assessments using early warning systems and statistical models [4]. Despite existing frameworks and risk monitoring, some Ethiopian banks have faced distress [5], indicating a need for improved monitoring and the aggregation of comprehensive data into a single measure of financial strength. The increasing number and size of banks in Ethiopia pose new stability challenges, motivating this research to improve financial distress prediction. The study aims to identify the most accurate model to serve as an early warning system and is considered a pioneering effort in applying machine learning to data from Ethiopian commercial banks.

Financial distress for commercial banks as the inability to meet financial obligations due to factors like non-performing loans, low capital, or poor risk management [23], leading to significant negative consequences including reduced lending, loss of depositor confidence, and potential bankruptcy [9]. While the Ethiopian banking system has experienced distress, outright failures have been avoided, and the NBE currently relies on subjective expert judgment and the CAMEL rating system (Capital Adequacy, Asset Quality, Management, Earnings, Liquidity) to assess bank health. This current method is time-consuming and potentially inconsistent, underscoring the need for accurate, reliable, and efficient machine learning models for financial distress prediction, despite challenges such as identifying effective variables and dealing with imbalanced datasets [13].

The study is guided by specific inquiries aimed at developing an accurate predictive model for bank financial distress. The questions seek to determine how machine learning models can be developed for this purpose in the Ethiopian banking sector. They also investigate the utility of dimensionality reduction and feature extraction techniques, asking which methods are more effective and if new insights can be gained from transformed features [13]. A central question focuses on identifying the most effective features from the CAMEL evaluation framework to be included in the machine learning models for

predicting financial distress in Ethiopian commercial banks [19].

The overall aim of the research is to develop a predictive model by investigating machine learning models capable of predicting financial distress within the Ethiopian Banking sector. This general objective involves identifying the most influential features [13, 19], managing imbalanced datasets [12], and optimizing feature selection relevant to financial distress [13]. The specific objectives further detail these goals, including identifying key financial indicators contributing to distress that provide insights for stakeholders [19], comparing the effectiveness of various supervised classification models to propose suitable prediction models for commercial banks [16], and developing a prediction model utilizing financial variables, specifically ratios, from these banks [11].

2. Related work

Research on financial distress prediction (FDP) has evolved significantly, moving from traditional statistical methods to advanced machine learning techniques [10, 21, 15]. Early approaches like Altman's Z-score and logistic regression laid the foundation, but recent studies increasingly leverage machine learning algorithms such as decision trees, neural networks, support vector machines, random forests, gradient boosting and ensembles [10,16]. These advanced methods are favored for their ability to capture complex, non-linear relationships in financial data, offering improved prediction accuracy compared to their traditional counterparts [10, 15].

Studies in the global context showcase this methodological shift and application across various financial sectors and regions. Research on UK banks utilized novel supervisory data, finding Random Forest models superior for early warning systems and employing interpretability techniques [14]. Investigations in Vietnam focused on listed firms, using ML to gauge distress probability and identify key internal determinants, reporting high accuracy rates with models like Altman and Zmijewski [2]. Further work on UAE firms compared various ML classifiers and emphasized the benefit of ensemble techniques like Majority Voting, Random Forest, and AdaBoost for enhancing prediction accuracy using financial data.

Other global research reinforced the superiority of machine learning in specific contexts. A study on US banks [3] comparing traditional methods with ML on a large dataset found Artificial Neural Networks and k-nearest Neighbor algorithms to be the most precise for forecasting bank failures [15]. Similarly, a review of literature on Taiwanese companies highlighted XGBoost as the most accurate predictor among supervised models and noted the enhanced precision of hybrid models like DBN-SVM, demonstrating the continued exploration of advanced and combined ML approaches [16].

In the African context, research adapts these successful ML techniques to unique regional challenges like economic volatility and data limitations [17, 18]. Studies here commonly employ logistic regression, decision trees, neural networks, support vector machines, and ensemble methods [17, 18]. For instance, research on Egyptian SMEs found

that integrating financial variables with firm characteristics significantly improved distress prediction accuracy, with Artificial Neural Networks (MLPs) outperforming other tested statistical methods, although macroeconomic indicators had no significant impact on NN performance in that study [17].

Further research in the African context, such as a study on Tunisian firms, specifically explored the use of Neural Networks for financial distress prediction [18]. This work examined different model architectures and time frames, identifying key financial ratios like short-term debt, capital structure, sales growth, and liability ratios as significant factors enhancing predictive performance, and selecting the best model based on its ability to maximize accuracy given specific information structures [18].

Moving to the Ethiopian context, the application of ML for financial distress prediction is still emerging, reflecting the region's unique economic landscape. Existing studies have largely focused on identifying determinants of financial health and distress using more traditional statistical analyses across sectors like commercial banks [19], insurance companies [20, 22, 24], and microfinance institutions [21]. Findings indicate the importance of internal factors like profitability, liquidity, solvency, and asset quality, alongside some macroeconomic variables [19, 24]. The full potential of ML in Ethiopia remains largely untapped, underscoring the need for more comprehensive data collection and tailored ML model development for enhanced risk assessment and financial stability.

The related work establishes that research in Financial Distress Prediction (FDP) has fundamentally shifted from traditional statistical methods to advanced machine learning (ML) techniques, which consistently demonstrate superior predictive performance due to their ability to capture complex, non-linear relationships in financial data. Lessons drawn from global contexts indicate the efficacy of various algorithms, including Random Forests for early warning systems in UK banks, the precision of Artificial Neural Networks (ANNs) and k-nearest Neighbor algorithms for forecasting US bank failures, and the enhanced accuracy provided by ensemble methods like Majority Voting, AdaBoost, and hybrid models like DBN-SVM. Moreover, studies highlight the importance of integrating financial variables with firm characteristics and optimizing model architecture based on specific factors like profitability, liquidity, and liability ratios to significantly improve prediction accuracy across regions like Egypt and Tunisia. The Researcher should bother conducting this research because, despite this established global success, the application of ML for FDP remains largely untapped and emerging within the Ethiopian context. Although prior Ethiopian studies using traditional statistical analyses have identified key internal financial factors influencing distress, there is a critical need to move beyond these methods to develop tailored ML models and engage in more comprehensive data collection. This effort is essential to adapt successful global techniques to Ethiopia's unique economic landscape, thereby enabling enhanced risk assessment and supporting overall financial stability.

3. Methodology

3.1. Research Design, Method & Rational

3.1.1 Research Design & Method

The research utilizes design science approach by starting quantitative task to analyze financial data from commercial banks in Ethiopia. This involves collecting numerical data, such as balance sheets and income statements, which are suitable for statistical and computational analysis. The primary justification for this approach is the aim to develop a predictive model that can be broadly applied across the banking sector. Machine learning models are employed to analyze this quantitative data. To implement this design, a comprehensive dataset spanning a maximum of 33 years from Ethiopian commercial banks is collected. A correlational design is then applied, using machine learning techniques, to identify relationships between various financial ratios and the occurrence of financial distress. The process involves training these models on a portion of the historical data and validating their ability to predict financial distress on a separate test set. The effectiveness of the predictive models is objectively evaluated using statistical measures. Key performance metrics such as accuracy, precision, recall, and the area under the ROC curve are used to assess how well the machine learning models predict financial distress. These measurable criteria are crucial for establishing the credibility and practical value of the research findings.

3.1.2. Methodological Rationale

The methodological rationale for this study centers on integrating the well-established CAMEL framework with machine learning to create a robust predictive model specifically for the Ethiopian banking sector. By analyzing five core dimensions—Capital Adequacy, Asset Quality, Management, Earnings, and Liquidity—alongside Altman Z-score measurements to accurately label distressed banks, the research ensures that the models are grounded in proven financial indicators. This comprehensive approach is applied to a significant dataset encompassing 18 Ethiopian commercial banks over nearly three decades (1990–2022), allowing for a rigorous evaluation of various algorithms, including Logistic Regression, Artificial Neural Networks, and Random Forest, to identify the most effective predictive tool.

To ensure the models are both reliable and actionable, the methodology prioritizes advanced optimization and data preprocessing techniques to handle the complexities of financial data. the study specifically addresses dataset imbalance and overfitting by employing SMOTE oversampling and class weighting, which are critical for refining the models' ability to anticipate distress before it materializes. These refinements, validated through stratified cross-validation, enabled the Random Forest model to achieve an exceptional 97.70% accuracy rate. By focusing on feature selection and dimensionality reduction, the research provides a sophisticated, data-driven framework designed to serve

as an early warning system for regulators, bank management, and investors seeking to enhance sector stability.

3.2. Data, Model & Business Understanding

3.2.1. Data Understanding

The methodology employed in this research begins with the collection of historical financial data for 18 commercial banks in Ethiopia over a period of up to 33 years (1990 to 2022). This data, which includes balance sheets, income statements, and cash flow statements, is gathered from National Bank of Ethiopia Repository. Following collection, rigorous data pre-processing techniques are applied to clean and transform the data, addressing issues such as missing values, inconsistencies, and errors. The process also involves checking for outliers and imbalanced datasets. Subsequently, feature selection is carried out to identify the most relevant variables for predicting financial distress for respective model. This involves exploring the data through statistical analysis and domain knowledge, utilizing techniques such as correlation analysis, feature importance ranking, dimensionality reduction methods, Principal Component Analysis (PCA), and Correlation-based Feature Selection (CFS). The predictor variables are derived from financial ratios based on the CAMEL framework (Capital, Asset Quality, Management, Earnings, and Liquidity), while the predicted variable is the financial distress status, determined using the Altman Z-score model. The Altman Z-score is calculated using specific financial ratios and classifies banks into 'safe', 'grey', or 'distress' zones based on the resulting score. Data analysis revealed a dataset imbalance between distressed and non-distressed banks, with most observations falling into the "Not Distress" category.

3.2.2. Model Understanding

With the data prepared, the study utilizes various machine learning algorithms for training predictive models. The algorithms specifically mentioned include Logistic Regression, Decision Trees, Artificial Neural Networks (ANN), K-Nearest Neighbor (KNN), Support Vector Machines (SVM), and Random Forest (RF). Model training involves splitting the dataset into training (66%) and testing (34%) sets Techniques such as cross-validation and hyperparameter tuning are employed during training to assess effectiveness and generalization capabilities. Parameter optimization techniques like Grid Search, genetic algorithms, simulated annealing, and gradient descent are also supported and used in this process. The performance of the trained models is evaluated using a variety of metrics. These include accuracy, precision, recall, F1 score, ROC curve, and AUC score. The confusion matrix is also used to provide a detailed breakdown of true positives, true negatives, false positives, and false negatives.

In developing our model specifications, we examined an extended set of variables that follow under the classification categories of CAMEL (i.e. Capital, Asset Quality, Management, Earnings, and Liquidity). Specifically, the independent variables tested are

the following (Table 1):

We have considered a comprehensive set of variables categorized under CAMEL (Capital, Asset Quality, Management, Earnings, and Liquidity) to assess the financial distress and risk profile of banks. The independent variables tested within each category provide a detailed insight into key aspects such as Capital Adequacy, Asset Quality, Management Efficiency, Earnings, and Liquidity.

There are various models that are used to measure the financial distress condition of firms those are, debt service coverage (DSC), Springate S score model, Olson model, standard poor's model, and Altman Z-score model to measure the financial distress condition[24].

Table 1: List of Predictor Variables

Category	Code	Indicator Name	Formula
Capital Adequacy (C)	X_1	Total risk-based Capital Adequacy Ratio (CAR)	Total Capital / Risk-Weighted Assets
	X_2	Debt to Equity Ratio	Total Debt / Total Equity
	X_3	Equity Capital to Total Assets	Total Equity / Total Assets
	X_4	Tier 1 risk-based Capital Ratio	Tier 1 Capital / Risk-Weighted Assets
	X_5	Tier 2 risk-based Capital Ratio	Tier 2 Capital / Risk-Weighted Assets
	X_6	Core capital (leverage) ratio	Tier 1 Capital / Total Assets
Asset Quality (A)	X_7	Loan Loss Provision to Total Loans Ratio	Loan Loss Provisions / Total Loans
	X_8	Loan Loss Provision to Total Asset Ratio	Loan Loss Provisions / Total Assets
	X_9	Loan Loss Provision to Net Income Ratio	Loan Loss Provisions / Net Income
	X_10	Noncurrent assets to total assets ratio	Noncurrent Assets / Total Assets
	X_11	Noncurrent assets to total loan ratio	Noncurrent Assets / Total Loans
	X_12	Noninterest income to total income	Noninterest Income / Total Income
Management (M)	X_13	Management Efficiency Ratio	Total Income / Total Assets
	X_14	Efficiency Ratio	Interest Expenses / Interest Income
	X_15	Operating Expense Ratio	Interest Expenses / Total Income
	X_16	Noninterest income to average assets	Noninterest Income / Average Assets
	X_17	Noninterest expense to average assets	Noninterest Expense / Average Assets
	X_18	Net Operating Income to Total Assets	Net Operating Income / Total Assets
Earnings (E)	X_19	Return on Assets (ROA)	Net Income / Average Total Assets
	X_20	Return on Equity (ROE)	Net Income / Average Shareholders' Equity
	X_21	Net Interest Margin (NIM)	(Interest Income - Interest Expenses) / Average Earning Assets
	X_22	Loan Loss Provision to Net Income Ratio	Loan Loss Provisions / Net Income
	X_23	Retained Earnings to Average Equity	Retained Earnings / Average Equity
	X_24	Interest Income to Total Income Ratio	Interest Income / Total Income
Liquidity (L)	X_25	Net Loan to Total Assets	(Net Loans - Loan Loss Provisions) / Total Assets
	X_26	Net Loan to Total Deposit Ratio	Net Loans / Total Deposits
	X_27	Net Loan to Core Deposit Ratio	Net Loans / Core Deposits
	X_28	Liquid Asset to Total Asset Ratio	Liquid Assets / Total Assets
	X_29	Liquid Asset to Demand Deposit Ratio	Liquid Assets / Demand Deposits
	X_30	Liquid Asset to Total Deposit Ratio	Liquid Assets / Total Deposits

This research uses the Altman Z-Score model, developed by Edward Altman, which is a widely used formula to predict the likelihood of bankruptcy or financial distress in a Bank. The Z-Score is calculated using multiple financial ratios such as working capital, profitability, leverage, liquidity, and book value of equity. A Z-Score below a certain threshold indicates a higher risk of financial distress

The following Equation for financial distress or possibility of bankruptcy of the

emerging market financial industry has been analysed.

Z-Score Distress model: $Z=3.25 + 6.56X1 + 3.26X2 + 6.72X3 + 1.05X4$

Where:

X_1: Working Capital/Total Assets (Weight: 6.56): This ratio measures the proportion of a Bank's assets that are financed by its working capital. A higher value indicates a stronger financial position.

X_2: Retained Earnings/Total Assets (Weight: 3.26): This ratio assesses the proportion of a Bank's assets that are financed by its retained earnings. A higher value suggests a more stable financial condition.

X_3: Earnings Before Interest and Taxes (EBIT)/Total Assets (Weight: 6.72): This ratio evaluates the profitability of a Bank relative to its total assets. A higher value indicates better earnings generation.

X_4: Book Value of Equity/Total Liabilities (Weight: 1.05): This ratio compares the market value of a Bank's equity to its total liabilities. A higher value suggests a stronger market position.

By plugging in the values for each ratio, we can obtain the Altman Z-Score. The resulting score provides an indication of a Banks's financial stability and the likelihood of distress. Generally, a higher Z-Score indicates a lower distress risk, while a lower Z-Score suggests a higher risk. The following status will be applicable for the values of Z-score

Table 2: Zones of Discriminations

Range	Description	Classification
If Z is greater than 2.99	The Bank is financially sound and is considered to be in the safe zone, with a low probability of financial distress	Not Distress Bank's
If Z is between 1.81 and 2.99	The Bank is considered to be in the grey zone, with a moderate probability of face financial distress in the near future	Distress Bank's
If Z is less than 1.81	The Bank is considered to be in the distress zone, with a high probability of face financial distress in the near future.	

Source: https://fastercapital.com/content/Altman-Z-Score--How-to-Predict-Your-Comp any-s-Bankruptcy-Risk.html access on March 22 2024

The entire dataset of 331 observations was categorized into two groups, distress and non-distress, using computation results (z-scores). This classification was done to prepare the data for further machine learning analysis. To address the imbalance in the dataset, techniques such as oversampling or under sampling has been applied to ensure that both distress and non-distress categories are adequately represented in the training data for the machine learning model.

3.2.3. Business Understanding

Based on Altman Z-score analysis (2018-2022) as per Table-3, most of the top five

Ethiopian banks showed financial stability, with Z-scores consistently above the distress threshold. While AIB and HB experienced brief distress early in the period but recovered, CBE remained stable overall despite a negative retained earnings ratio potentially impacting long-term sustainability. The data generally indicates a safe financial position and improving Z-scores over time, suggesting effective management.

Table 3: Financial Distress Measurement (Altman Z-score)

Bank	Year	X1_Working Capital to Total Asset	x2_Retained Earning to Total Asset	x3_EBIT to Total Assets	x4_Market Value E to Liablity	Z_score	Distress Zone
AIB	2018	0.3655899	0.0294158	0.0403386	0.1331842	2.905084	Distress
	2019	0.3117509	0.035289	0.0516059	0.1482993	2.662634	Distress
	2020	0.3081701	0.0330627	0.0580274	0.1548984	5.9319676	Not Distress
	2021	0.2741224	0.0292913	0.0531949	0.1435105	5.6518881	Not Distress
	2022	0.248984	0.0276286	0.058097	0.1289778	5.4992434	Not Distress
BOA	2018	0.3438682	0.0113075	0.0342086	0.1530539	5.9332259	Not Distress
	2019	0.3088874	0.0124923	0.0378234	0.1441415	5.7225481	Not Distress
	2020	0.2361483	0.0073752	0.0291762	0.1108581	5.1356408	Not Distress
	2021	0.1845526	0.0104724	0.0390494	0.090857	4.852617	Not Distress
	2022	0.1762833	0.01432	0.0471818	0.1051219	4.8805416	Not Distress
CBE	2018	0.6468256	0.0000141	0.0314466	0.0896274	4.5486061	Not Distress
	2019	0.6744908	-0.00029	0.0394233	0.0758318	4.7692079	Not Distress
	2020	0.6626263	0.000179	0.0355432	0.0647496	4.6536657	Not Distress
	2021	0.675763	-0.00144	0.0372804	0.0574154	4.7438157	Not Distress
	2022	0.7114404	0.000001	0.0714884	0.0662304	5.2169927	Not Distress
DB	2018	0.3981222	0.0364749	0.0304822	0.1483012	3.0911462	Not Distress
	2019	0.3429171	0.0113082	0.0265066	0.1386784	5.8601375	Not Distress
	2020	0.3118289	0.0116657	0.0276944	0.1386993	5.6653684	Not Distress
	2021	0.260513	0.0122622	0.029327	0.1197217	5.3217251	Not Distress
	2022	0.257078	0.0108442	0.0378174	0.1399369	5.37285	Not Distress
HB	2018	0.3886565	0.0131442	0.0359815	0.1177892	2.9579105	Distress
	2019	0.3174811	0.0101598	0.0284603	0.1210795	5.6841837	Not Distress
	2020	0.2749926	0.0140221	0.0322798	0.1422083	5.4659024	Not Distress
	2021	0.2062532	0.0134309	0.0355943	0.1361444	5.0289509	Not Distress
	2022	0.1739454	0.0129591	0.0507281	0.1204644	4.9007088	Not Distress
NIB	2018	0.3816733	0.0124425	0.0341592	0.1450286	2.9261687	Distress
	2019	0.3224108	0.0145417	0.0351151	0.1505146	5.8064344	Not Distress
	2020	0.2888269	0.0172708	0.0366848	0.1577811	5.6131993	Not Distress
	2021	0.2648258	0.0140803	0.036112	0.1508357	5.4342091	Not Distress
	2022	0.2553814	0.0145682	0.0383534	0.1520451	5.3901767	Not Distress

On the otherhand, the Distress Status Distribution plot (figure-1) clearly shows the count of banks in financial distress compared to those that are stable. It reveals that the majority of observations, 270 banks, fall under the "Not Distress" category (financially stable), while a smaller number, 61 banks, are labeled as "Distress". This distribution highlights a significant dataset imbalance, which is important to consider for predicting financial distress. The researcher notes the need to adjust the dataset to address this imbalance, possibly using techniques such as oversampling or under-sampling, to prevent the machine learning model from favoring the majority class.

Separately, the Bank Distribution plot illustrates how often each specific bank appears in the dataset. It shows an uneven representation of banks, with some like CBE and AIB appearing more frequently than others like CBB and ZB. This variation is noted as important because banks with fewer data points might not be adequately represented in the analysis, potentially affecting the accuracy of the conclusions drawn about them. These plots collectively help in understanding the dataset's composition, with the identified imbalance in distress status and uneven bank representation posing challenges for the machine learning process. Addressing these issues is deemed crucial for ensuring accurate and meaningful findings and guides the data preparation and model building steps.

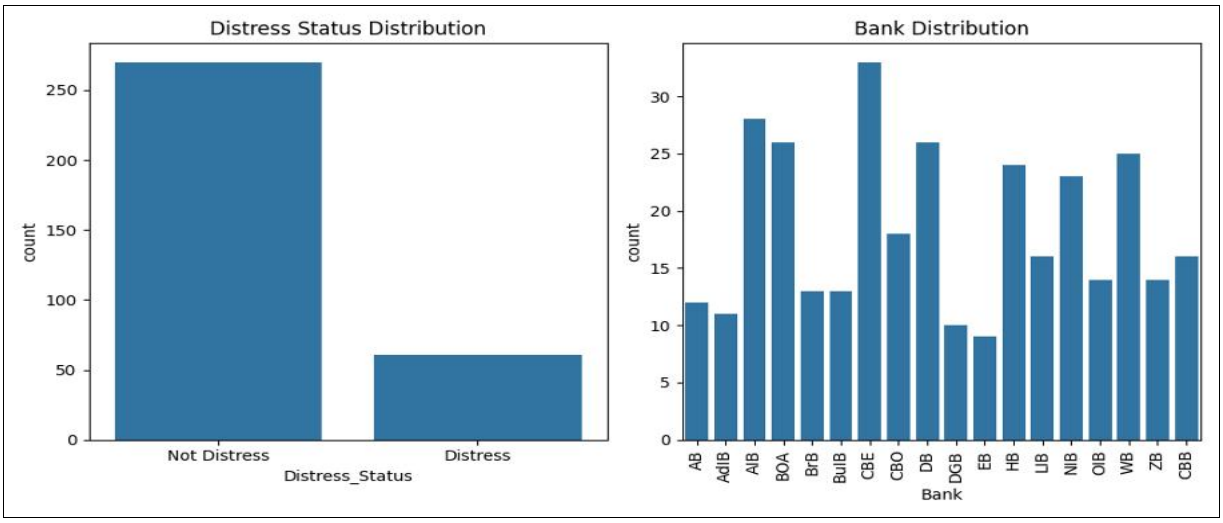

Figure 1: Predictor Variables Measurement

4. Experimentation

4.1. Logistic Regression

Baseline: Established a baseline using Logistic Regression as per table 4 with default settings. It achieved an Overall Accuracy of 89.1% (10-Fold CV), 85% (Test Split), and 96.4% (Training Set). The Test Split Kappa Statistic was 0.545. While overall accuracy was decent, the significant difference between training and test/CV performance (e.g., 96.4% Train vs 85% Test Accuracy) and lower performance on the minority 'Distress' class (Test TP Rate 0.577, F-Measure 0.638) highlighted potential overfitting and class imbalance issues.

Cost Sensitive Learning (CSL): The Cost Sensitive Learning approach, which incorporates a cost matrix to handle class imbalance, was applied. However, this approach yielded results identical to the baseline across all metrics and evaluation methods (e.g., Test Accuracy 85%, Test Distress TP Rate 0.577, Test Kappa 0.545), indicating it minimized misclassification costs but did not alter the predictive performance.

SMOTE: Introduced SMOTE oversampling to improve minority class recall. Similar to CSL, the SMOTE experiment also produced performance metrics identical to both the baseline and CSL models (e.g., Test Accuracy 85%, Test Distress TP Rate 0.577, Test Kappa 0.545), suggesting it did not improve predictive ability in this application.

Hyperparameter Tuning: Focused on Hyperparameter Tuning (using SMOTE-processed data and grid search). This approach demonstrated a clear improvement, particularly in generalization to unseen data. Test Accuracy rose to 89.4%, and the Test Kappa Statistic increased significantly to 0.708. Error metrics on the Test Set decreased (e.g., Test RMSE 0.299, Test RAE 34%). Performance on the minority 'Distress' class on the Test Split saw substantial gains: TP Rate (Recall) increased to 0.808, Precision to 0.75,

and F-Measure to 0.778. Test ROC Area for 'Distress' reached 0.923 and PRC Area 0.807. While Training Accuracy remained high at 96.4%, Hyperparameter Tuning was effective in improving both generalization and minority class prediction compared to the initial methods.

Combined Approach: Explored a Combined Approach by integrating SMOTE, cost matrix application, and hyperparameter tuning. The results for this combined strategy were identical to those achieved solely by the Hyperparameter Tuning experiment. Test Accuracy remained 89.4%, Test Kappa 0.708, and the Test Distress class performance mirrored hyperparameter Experiment with TP Rate 0.808, Precision 0.75, and F-Measure 0.778. While this achieved the best performance observed among all experiments, significantly outperforming the baseline (85% Test Accuracy), CSL, and SMOTE methods, the identical metrics to hyperparameter Experiment, indicate that adding the cost matrix and combining with SMOTE (as implemented here) did not yield further predictive performance improvements beyond those already achieved by hyperparameter tuning alone. Addressing the high Training Accuracy of 96.4% compared to Test Accuracy remains important for ensuring robustness on truly unseen data.

Table 4: Experiment on Logistic Regression

4.2. Decision Tree

Decision Trees are a popular and intuitive machine learning method used for classification and regression tasks. They work by recursively splitting the data into subsets based on feature values, creating a tree-like structure of decisions.

Baseline: The goal of this experiment as per Table -5 was to establish a baseline

performance for the REPTree classifier using its default settings with 10-fold cross-validation1. The classifier options were set to M=2, V=0.001, L=-1, N=3, and I=1.01. The REP Tree classifier achieved a strong overall performance in predicting bank financial distress with a 92.75% accuracy and a Kappa statistic of 0.7759. The resulting decision tree was remarkably simple, consisting of a single split based on the feature L_X_30, indicating its significant predictive power. While the model performed well overall, it showed class-specific nuances, with high precision (0.977) and recall (0.933) for "Not Distress" but slightly lower precision (0.753) and F-measure (0.821) for the "Distress" class, suggesting potential class imbalance and room for improvement in recall for distressed instances. The reliance on L_X_30 underscores its crucial role, but the simple tree structure suggests potential for improved robustness and complexity handling, with suggestions for further analysis of L_X_30, feature engineering, and exploring more advanced methods or parameter tuning.

Parameter Sweeps: This category involved a series of parameter sweeps to meticulously adjust key parameters such as Minimum Instances (M), Minimum Variance (V), Maximum Tree Depth (L), Minimum Instances per Split (N), and Batch Size (I) using auto-weka, along with exploring different pre-processing and evaluation strategies to pinpoint optimal configurations. The experiments specifically evaluated REPTree's performance with 10-fold cross-validation and two different train/test splits (66/34 and 70/30). A key finding was the remarkably stable performance of the REP Tree classifier across all tested configurations and evaluation methods. Performance metrics like the F-measure consistently hovered between 0.857 and 0.966, and the weighted True Positive Rate showed a narrow range of variation, indicating a robust model not overly sensitive to the training/testing split. Overall performance metrics like "Correctly Classified (%)" (94.56%), "Incorrectly Classified (%)" (5.44%), and Kappa Statistic (0.8236) remained identical across experiments varying data partitioning and evaluation methods. The use of the Best First attribute selection algorithm also did not significantly change the model's performance. Training times were very short across all these experiments.

Ensemble Methods (Bagging): Within the exploration of ensemble learning methods like bagging and boosting, the Bagging classifier demonstrated consistently strong performance when evaluated using the provided metrics. The best results were seen in both the 66/34 and the 70/30 splits, achieving an accuracy of 96.07%, a Kappa score of 0.8651, and an areaUnderROC of 0.99, indicating the model's robustness in classifying both positive and negative instances and high effectiveness in both setups. Class-specific performance also showed consistent precision, recall, and f-measure across the split experiments. The attribute handling strategy included using Greedy Stepwise and CfsSubsetEval in some configurations, aligning with the Random Subspace method concept. Interestingly, the Bagging model achieved the same 96.07% accuracy with and without attribute selection, suggesting the original feature space was adequate and feature

selection, while potentially beneficial, was not critical for this model on this dataset. Training times were negligible for the Bagging model, especially for the data splits.

Ensemble Methods (Random Subspace): Also explored as part of the ensemble methods, the Random Subspace classifier exhibited varied overall performance across tested configurations, with accuracy ranging from 92.45% to 96.07%. The highest accuracy (96.07%) was achieved with a 70/30 split, while the 10-fold cross-validation setup resulted in the lowest accuracy (92.45%). Despite the variation in overall accuracy, the class-specific ROC Area values were consistently high across all setups (0.982 for cross-validation, 0.968 for 66/34 split, and 0.99 for 70/30 split), indicating a strong ability to discriminate between the two classes. Class-specific precision values were also very high. The attribute handling strategy varied, employing Greedy Stepwise in some experiments or disabling attribute selection in others. The model achieved high results even without attribute selection, suggesting the original feature space is highly effective, with attribute selection providing further refinement. Training time was negligibly small, except for the cross-validation experiment, which was significantly costlier.

Table 5: Experiment on Decision Tree

	Baseline	Parameter Sensitive Analysis								Ensemble Methods				
	10 Fold Cross Validation	10 Fold Cross Validation		Percentage Split (66/34)			Percentage Split (70/30)			10 Fold Cross Validation	Percentage Split (66/34)		Percentage Split (70/30)	
Classifier & Metric	Baseline	REPTree (fMeasure)	REPTree (weightedTruePositive Rate)	REPTree (fMeasure)	REPTree (weightedTruePositiveRate)	REPTree (errorRate)	REPTree (fMeasure)	REPTree (errorRate)	REPTree (fMeasure)	REPTree (weightedTruePositiveRate)	RandomSubSpace (areaUnderROC)	RandomSubSpace (Precision)	Bagging (areaUnderROC)	RandomSubSpace (Precision)	RandomSubSpace (Precision)
Tried Configurations		287	270	254	267	259	292	267	292	255	236	258	214	258	214
Attribute Search		null	weka.attributeSelection.BestFirst	null	weka.attributeSelection.BestFirst	weka.attributeSelection.BestFirst	null	weka.attributeSelection.BestFirst	null	weka.attributeSelection.BestFirst	weka.attributeSelection.GreedyStepwise	null	weka.attributeSelection.GreedyStepwise	null	weka.attributeSelection.GreedyStepwise
Attribute Search Args		[]	[-D, 2, -N, 3]	[]	[-D, 2, -N, 3]	[-D, 2, -N, 3]	[]	[-D, 2, -N, 3]	[]	[-D, 2, -N, 3]	[-R]	[]	[-C, -R]	[]	[-C, -R]
Attribute Eval Args		[]	[-M]	[]	[-M]	[-M]	[]	[-M]	[]	[-M]	[]	[]	[-L]	[]	[-L]
Estimated Error Rate		n/a	n/a	n/a	n/a	0.05438	n/a	0.05438	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Training Time (sec)		0	0	0	0.01	0.003	0	0.016	0	0	0.218	0.039	0.109	0.04	0.172
Overall Performance															
Correctly Classified (%)	92.7492	94.56%	94.56%	94.56%	94.56%	94.56%	94.56%	94.56%	94.56%	94.56%	92.45%	94.86%	96.07%	94.86%	96.07%
Incorrectly Classified (%)	7.2508	5.44%	5.44%	5.44%	5.44%	5.44%	5.44%	5.44%	5.44%	5.44%	7.55%	5.14%	3.93%	5.14%	3.93%
Kappa Statistic	0.7759	0.8236	0.8236	0.8236	0.8236	0.8236	0.8236	0.8236	0.8236	0.8236	0.7224	0.8164	0.8651	0.8164	86.51%
Mean Absolute Error	0.112	0.0964	0.0964	0.0964	0.0964	0.0964	0.0964	0.0964	0.0964	0.0964	0.1577	0.1613	0.0981	0.1613	0.0981
Root Mean Squared Error	0.2448	0.2195	0.2195	0.2195	0.2195	0.2195	0.2195	0.2195	0.2195	0.2195	0.2378	0.2354	0.1907	0.2354	0.1907
Relative Absolute Error (%)	37.0976	31.94%	31.94%	31.94%	31.94%	31.94%	31.94%	31.94%	31.94%	31.94%	52.26%	53.43%	32.49%	53.43%	0.32486
Root Relative Squared Error (%)	63.1351	56.62%	56.62%	56.62%	56.62%	56.62%	56.62%	56.62%	56.62%	56.62%	61.34%	60.72%	49.17%	60.72%	49.17%
Class -Specific Performance (Not Distress)															
Confusion Matrix															
a b <-- classified as															
** **| a = Not Distress	252 18	259 11	259 11	259 11	259 11	259 11	259 11	259 11	259 11	259 11	265 5	267 3	266 4	267 3	`266 4
** **| b = Distress	6 55	7 54	7 54	7 54	7 54	7 54	7 54	7 54	7 54	7 54	20 41	14 47	9 52	14 47	9 52
TPR	0.933	0.959	0.959	0.959	0.959	0.959	0.959	0.959	0.959	0.959	0.981	0.989	0.985	0.989	0.985
FPR	0.098	0.115	0.115	0.115	0.115	0.115	0.115	0.115	0.115	0.115	0.328	0.230	0.148	0.230	0.148
Precision	0.977	0.974	0.974	0.974	0.974	0.974	0.974	0.974	0.974	0.974	0.930	0.950	0.967	0.950	0.967
Recall	0.933	0.959	0.959	0.959	0.959	0.959	0.959	0.959	0.959	0.959	0.981	0.989	0.985	0.989	0.985
F-Measure	0.955	0.966	0.966	0.966	0.966	0.966	0.966	0.966	0.966	0.966	0.955	0.969	0.976	0.969	0.976
ROC Area	0.909	0.922	0.922	0.922	0.922	0.922	0.922	0.922	0.922	0.922	0.982	0.968	0.990	0.968	0.99
PRC Area	0.967	0.967	0.967	0.967	0.967	0.967	0.967	0.967	0.967	0.967	0.996	0.988	0.998	0.988	0.998
Class -Specific Performance (Distress)															
TPR	0.902	0.885	0.885	0.885	0.885	0.885	0.885	0.885	0.885	0.885	0.672	0.770	0.852	0.770	0.852
FPR	0.067	0.041	0.041	0.041	0.041	0.041	0.041	0.041	0.041	0.041	0.019	0.011	0.015	0.011	0.015
Precision	0.753	0.831	0.831	0.831	0.831	0.831	0.831	0.831	0.831	0.831	0.891	0.940	0.929	0.940	0.929
Recall	0.902	0.885	0.885	0.885	0.885	0.885	0.885	0.885	0.885	0.885	0.672	0.770	0.852	0.770	0.852
F-Measure	0.821	0.857	0.857	0.857	0.857	0.857	0.857	0.857	0.857	0.857	0.766	0.847	0.889	0.847	0.889
ROC Area	0.909	0.922	0.922	0.922	0.922	0.922	0.922	0.922	0.922	0.922	0.982	0.968	0.990	0.968	0.99
PRC Area	0.673	0.757	0.757	0.757	0.757	0.757	0.757	0.757	0.757	0.757	0.911	0.907	0.956	0.907	0.956

4.3. Artificial Neural Network (ANN)

Artificial Neural Networks are advanced computational models inspired by the structure and function of the human brain. They consist of layers of interconnected nodes (neurons) that work together to recognize complex patterns and relationships within data. ANNs are particularly powerful for handling non-linear problems and large datasets, making them widely used in tasks such as classification, prediction, and pattern recognition.

Table 6: Experiment on Artificial Neural Network (ANN)

Varying Number of Hidden Layers: An experiment explored the impact of the number of hidden layers on MLP performance for classifying banks into "Not Distress" or "Distress" using 31 financial attributes. An optimal configuration identified utilized three hidden layers with dynamic sizing. Trained with backpropagation (learning rate 0.3, momentum 0.2, 500 max epochs) and evaluated using 10-fold stratified cross-validation, this model achieved a good overall accuracy of 91.5% and a Kappa of 0.722. However, the recall for the "Distress" class was lower at 0.787, with 13 distressed instances misclassified as "Not Distress", indicating room for improvement particularly in predicting the minority class despite reasonable overall success.

Varying Neurons in ONE Hidden Layer: This experiment examined the effect of the number of neurons in a single hidden layer on MLP performance. After trying various numbers, a model with a single hidden layer containing 30 neurons was identified as optimal. Using backpropagation (learning rate 0.3, momentum 0.2, 500 max epochs with early stopping) and 10-fold stratified cross-validation on 31 financial attributes (minus the initial three), this model achieved a slightly lower overall accuracy of 90.6% compared to the three-layer architecture, with a Kappa of 0.687. The recall for the "Distress" class was

worse, with 16 distressed instances misclassified as "Not Distress", highlighting a notable false negative problem for the distress class and a performance trade-off based on this network structure.

Varying Learning Rate and Momentum: Experiments explored different combinations of learning rates and momentum values for training MLPs. An optimal model emerged with two hidden layers, each containing 5 neurons, trained with a learning rate of 0.2 and momentum of 0.5 using backpropagation for up to 500 epochs with early stopping. This model was evaluated on a single 66%/34% train/test split rather than cross-validation. It achieved a classification accuracy of 89.4% on the test split with a Kappa of 0.708. Similar to other architectures, the "Distress" class showed a lower recall rate of 0.808, with 5 distressed instances incorrectly classified as "Not Distress" on the test set, indicating challenges in correctly identifying distressed banks despite a competitive accuracy level.

Number of Training Epochs: This experiment investigated the impact of different training epoch numbers while keeping learning rate (0.3), momentum (0.2), and automatically determined hidden layer size constant. An optimal setting utilized an MLP with a single automatically determined hidden layer size, trained with backpropagation for a maximum and optimal 50 epochs with early stopping. Evaluated on a single 66%/34% train/test split, this model achieved an overall accuracy of 91.2% and a Kappa statistic of 0.757. Crucially, the recall for the "Distress" class was 0.846, with only 4 distressed instances misclassified as "Not Distress" on the test set, demonstrating a better ability to correctly identify distressed banks compared to the two-layer model and highlighting the importance of the number of epochs.

Impact of Feature Selection: This analysis focused on refining the dataset by identifying and potentially removing irrelevant features before training the model. Using information gain and selection frequency across cross-validation folds, features L_X_30, L_X_28, and A_X_12 were consistently identified as the most informative predictors for financial distress. Other features like C_X_1 showed some importance, while a substantial number demonstrated no predictive value (zero information gain and never selected). The findings suggest that prioritizing the most informative features (especially from 'Liquidity' and 'Asset Quality' categories) while excluding irrelevant ones is likely to improve model accuracy, robustness, and generalization performance.

Handling Imbalanced Data: This experiment explored the influence of data pre-processing techniques, specifically SMOTE and a combination of SMOTE and Resampling, to address class imbalance common in financial distress prediction. Using an MLP with an automatically determined single hidden layer and fixed learning rate/momentum, two models were trained on a dataset of 30 financial attributes and evaluated on a single train/test split (66%/34%). Both models achieved similar, high overall accuracies (94.4%) and Kappa statistics (around 0.88) on the test split. The second model using both SMOTE and Resampling showed only marginally better performance,

particularly slightly higher recall for the "Distress" class compared to the model using only SMOTE., suggesting potential benefits from combining these techniques for improving the prediction of the minority class.

4.4. K-Nearest Neighbor Algorithm Experiments

The K-Nearest Neighbor algorithm is a simple, instance-based learning method used for classification and regression. It classifies data points based on the majority class of their closest neighbors in the feature space. KNN is easy to implement, requires no prior training, and is effective in scenarios where the decision boundary is irregular or unknown.

Baseline Performance Without Scaling: This initial experiment as per table-8 evaluated the k-NN algorithm without any feature scaling or normalization, using an optimal k-value of 3 and 10-fold stratified cross-validation. The model achieved an overall accuracy of 86.10%, correctly classifying 285 out of 331 instances, with a Kappa statistic of 0.462. However, the model exhibited a clear performance imbalance, struggling to correctly classify instances of the minority "Distress" class, misclassifying 34 out of 61 "Distress" instances as "Not Distress". This suggested that feature scales might hinder performance and that scaling could be beneficial.

Table 7: Experiment on K-Nearest Neighbor Algorithm Experiments

Baseline Performance with Scaling: This experiment established a baseline for k-NN performance after applying Min-Max scaling, using Weka's IBk implementation and an optimal k=7 determined by cross-validation. Evaluating with 10-fold stratified cross-validation, the model achieved an overall accuracy of 88.52% and a Kappa statistic of 0.5131. While overall accuracy improved, there was still a clear disparity in per-class performance; the "Not Distress" class had a very high Recall (0.993), while the "Distress" class had a much lower Recall (0.410), with the model misclassifying 36 out of 61 "Distress" cases.

Parameter Tuning and Standardization Impact: This experiment combined Min-

Max normalization and Z-score standardization sequentially with parameter tuning using cross-validation for optimal k selection (k=3). Using 10-fold stratified cross-validation, the model achieved an overall accuracy of 86.10% and a Kappa statistic of 0.462, which is similar to the baseline without scaling. The per-class performance remained imbalanced, with high Recall for "Not Distress" (0.956) but significantly lower Recall for "Distress" (0.443), indicating difficulty identifying distress cases and a bias towards the majority class.

Parameter Tuning and Standardization (Z-score): This experiment specifically evaluated k-NN performance after applying Z-score standardization with cross-validation for k selection (k=3). The results, assessed using 10-fold stratified cross-validation, were identical to Experiment 3, showing an overall accuracy of 86.10% and a Kappa statistic of 0.462. The detailed accuracy by class again highlighted the imbalance, with high Recall for "Not Distress" (0.956) and a lower Recall for "Distress" (0.443), signifying limitations in accurately classifying the minority class and a consistent bias towards the majority class.

Impact of Feature Selection: This experiment assessed the impact of feature selection (using CfsSubsetEval and Best First search) on k-NN performance in conjunction with Min-Max scaling. The process selected a subset of 5 attributes out of the original 31, and the k-NN algorithm was trained on this reduced set with min-max scaling and cross-validation selecting k=7. The model achieved an overall accuracy of 88.52% and similar per-class performance metrics (Recall of 0.993 for "Not Distress" and 0.410 for "Distress"), showing similar performance levels to Experiment 2 which had the full input set. While feature selection didn't improve performance, it reduced model complexity and could offer computational efficiency benefits, although the class imbalance issue persisted.

4.5. Support Vector Machines

Support Vector Machines are powerful supervised learning models used for classification and regression tasks. They work by finding the optimal hyperplane that best separates data points of different classes in a high-dimensional space. SVMs are especially effective in cases with clear margin separation and can handle both linear and non-linear classification through the use of kernel functions.

Establish Baseline SVM Performance: The results of this baseline linear SVM model show a significant class imbalance. While the model performs well in identifying instances of 'Not Distress', it fails to do so for 'Distress' with a very low recall, suggesting the model is biased towards the majority class. The relatively low Kappa score reflects the models inability to identify the instances that belong to the minority class. Therefore, the model's ability to generalize to unseen instances is questionable. This highlights the need for improvement in future experiments. Several avenues can be explored: trying different kernels (like polynomial or RBF) that can model a more complex decision boundary could help, feature selection can be explored so that the model is more focused in some data instances, and the researcher could try to balance the dataset. In our case, the model's

inability to correctly identify instances of 'Distress' suggests this may be a very hard problem to solve. The goal now for the rest of the experiments is to improve on this baseline and to try and overcome the class imbalance in the dataset.

Table 8: Experiment on Support Vector Machines

Metric	Experiment 1 (Baseline)	Experiment 2 (Polynomial Kernel)	Experiment 3 (RBF Kernel)	Experiment 4 (Tuned Polynomial)	Experiment 5 (Feature Selection)	Experiment 6 (Standardized Data)
Overall Performance						
Correctly Classified (%)	82.78	89.73	81.57	86.1	89.73	90.03
Incorrectly Classified (%)	17.22	10.27	18.43	13.9	10.27	9.97
Kappa Statistic	0.12	0.614	0	0.359	0.614	0.628
Mean Absolute Error	0.172	0.103	0.184	0.139	0.103	0.0997
Root Mean Squared Error	0.415	0.321	0.429	0.373	0.321	0.3157
Relative Absolute Error (%)	57.02	34.01	61.03	46.02	34.01	33.01
Root Relative Squared Error (%)	107.02	82.66	110.72	96.14	82.66	81.43
Class-Specific Performance (Not Distress)						
TPR	0.996	0.97	1	0.996	0.97	0.97
FPR	0.918	0.426	1	0.738	0.426	0.41
Precision	0.828	0.91	0.816	0.857	0.91	0.913
Recall	0.996	0.97	1	0.996	0.97	0.97
F-Measure	0.904	0.939	0.899	0.921	0.939	0.941
ROC Area	0.539	0.772	0.5	0.629	0.772	0.78
PRC Area	0.828	0.907	0.816	0.857	0.907	0.91
Class-Specific Performance (Distress)						
TPR	0.082	0.574	0	0.262	0.574	0.59
FPR	0.004	0.03	0	0.004	0.03	0.03
Precision	0.833	0.814	?	0.941	0.814	0.818
Recall	0.082	0.574	0	0.262	0.574	0.59
F-Measure	0.149	0.673	?	0.41	0.673	0.686
ROC Area	0.539	0.772	0.5	0.629	0.772	0.78
PRC Area	0.237	0.546	0.184	0.383	0.546	0.558

Impact of Polynomial Kernel: The application of a polynomial kernel and hyperparameter tuning has provided significant improvements to the classification performance compared to the baseline model with a linear kernel. The major improvement is in the classification of the minority ‘Distress’ class. While the polynomial kernel has addressed some of the limitations of the linear model, it is clear the class imbalance still leads to reduced performance in the minority class. The model's performance on the ‘Not Distress’ class is much higher than the ‘Distress’ class, and has improved over the baseline results, so the model is more capable of classifying both classes. This indicates that other methods can be used to further improve the models performance on the ‘Distress’ class. The next step will be to explore other methods, such as trying to further improve on hyperparameter tuning with other types of kernels, feature selection, or balancing the dataset.

Impact of RBF Kernel: The results of the RBF kernel SVM, especially the severe class imbalance, indicate that the chosen gamma value and the RBF kernel settings, have completely failed at this specific classification task. The model performed worse than the baseline and polynomial kernels and has a Kappa score of 0 and cannot classify the minority class, making this model the worst performing so far. Although it has high recall in the majority class, it has completely misclassified the minority class. This may be due to a poor value of gamma, the selected range of the grid search might have been suboptimal, the selected number of iterations might be insufficient, and that there is not enough variation in the data set to train the RBF kernel properly. Compared to the linear kernel (Experiment 1), which at least had some performance in both classes, and the polynomial kernel (Experiment 2) which improved the performance, the RBF model has a much worse

performance overall. The next steps should involve a careful examination of the range of the grid search for gamma. However, it may be more prudent to discard this approach in favor of a better model. It is clear that further tuning of the RBF kernel or different preprocessing techniques may be required to get this model to perform better.

Hyperparameter Tuning for Best Kernel: The fine-tuning process of the polynomial kernel using a grid search resulted in a model that has lower overall performance than the original setup of the polynomial kernel from experiment 2. This indicates that more fine-tuning of the parameters was not beneficial, and suggests that the original polynomial settings from experiment 2 were already optimized. While the performance of the majority class is very high, it has come at the cost of the performance of the minority class. This shows how using more complex hyperparameter tuning can sometimes lead to issues and lower performance, because of the inherent class imbalance in the dataset, and the inability of the chosen model to generalize. Although this experiment has given us very useful results as it shows that further tuning is not useful, this also shows us that other types of methods should be explored, such as feature selection and balancing the data, to overcome this problem. The results of this model should be compared against all the other experiments.

Feature Selection with SVM: The results of using a polynomial kernel after feature selection showed no change in performance. The metrics are all the exact same as when feature selection was not performed. In this specific case the selected attributes that were used for training the model have had no impact, and implies that the filtered data set had similar classification properties with the original data set. This is very useful information, as it shows us that our feature selection method has not improved the model, nor has it damaged the model performance. The data set can be reduced to include only the selected attributes, and would not impact the model's performance, this is very useful if the data set becomes large and computation becomes an issue. This experiment shows us that for this data set there isn't a great need for feature selection, as the polynomial model already provides great performance.

Impact of Data Standardization: The results indicate that standardizing the data prior to training the polynomial kernel SVM has improved the classification performance. This can be seen across all aspects of the output including overall accuracy, kappa, MAE, RMSE, ROC, PRC and the results from the confusion matrix. Although these results show a relatively small improvement in performance, standardizing the data seems to have slightly helped the model, which is a useful insight. The standardizing of the dataset can lead to better model performance if the data features are not on a similar scale. From these experiments it is clear that the polynomial kernel with the appropriate hyperparameters and standardizing the data is the best solution. This indicates that standardizing the dataset should be done in any future use cases.

4.6. Random Forest model

Random Forest is an ensemble learning technique that combines multiple decision trees to improve classification and regression accuracy. By building numerous trees on random subsets of the data and features, it reduces overfitting and enhances generalization. This approach is known for its robustness, ability to handle large datasets, and effectiveness in dealing with complex, non-linear relationships.

Tuning Number of Trees: This experiment as per table-10 focused on optimizing the number of trees in the Random Forest model to balance complexity and stability and find the point where performance plateaus. Using 30 trees (numTrees = 30), the baseline model achieved an overall accuracy of 94.86% and a Kappa statistic of 0.821, suggesting substantial agreement. However, due to class imbalance, performance differed significantly between classes. While the "Not Distress" class had high recall (98.1%) and precision (95.7%), the "Distress" class showed a lower recall of 80.3%, meaning 19.7% of actual distress cases were missed. This was further reflected in 12 false negatives and 5 false positives in the confusion matrix. The initial default parameters related to randomly chosen attributes and minimum variance were used and likely needed further optimization.

Table 9:Experment on Random Forest model

Metric	Tuning Number of Trees	Impact of Tree Max Depth	Feature Subset Size	Using Class Weights to Handle Imbalance
Overall Performance				
Correctly Classified (%)	94.864	94.864	94.864	97.7041
Incorrectly Classified (%)	5.136	5.136	5.136	2.2959
Kappa Statistic	0.8212	0.8212	0.8212	0.9458
Mean Absolute Error	0.1079	0.1079	0.1079	0.0689
Root Mean Squared Error	0.2134	0.2134	0.2134	0.1594
Relative Absolute Error (%)	35.7292	35.7292	35.7292	16.0592
Root Relative Squared Error (%	55.0286	55.0286	55.0286	34.433
Class-Specific (Not Distress)				
TPR (Recall)	0.981	0.981	0.981	0.993
FPR	0.197	0.197	0.197	0.057
Precision	0.957	0.957	0.957	0.975
F-Measure	0.969	0.969	0.969	0.983
ROC Area	0.977	0.977	0.977	0.993
PRC Area	0.994	0.994	0.994	0.997
Class-Specific (Distress)				
TPR (Recall)	0.803	0.803	0.803	0.943
FPR	0.019	0.019	0.019	0.007
Precision	0.907	0.907	0.907	0.983
F-Measure	0.852	0.852	0.852	0.962
ROC Area	0.977	0.977	0.977	0.993
PRC Area	0.911	0.911	0.911	0.976

Impact of Tree Max Depth: This experiment investigated how limiting the maximum depth of individual trees influences the model's ability to identify financial distress, aiming to balance underfitting and overfitting. The model was configured with a depth parameter set to 10. Evaluated with stratified cross-validation, it achieved an overall accuracy of 94.86% and a Kappa statistic of 0.8212, similar to the previous experiment. Class-specific metrics revealed a clear disparity: "Not Distress" had strong results (Recall 0.981, Precision 0.957), while the "Distress" class had a lower Recall of 0.803 and Precision of 0.907. The confusion matrix showed 5 false positives and a concerning 12 false negatives. The source notes that while a depth of 10 allows for reasonably complex learning, the

performance limitations suggest that simply increasing depth is not enough, and addressing the 12 false negatives, class imbalance techniques, and potentially feature engineering are crucial for improving "Distress" class identification.

Feature Subset Size: This experiment tested the influence of randomly selecting 5 features at each node split with the depth still set at 10, aiming to reduce variance and avoid overfitting. Trained with 30 trees and a maximum depth of 10, this model also demonstrated a strong overall performance, achieving an accuracy of 94.86% and a Kappa statistic of 0.8212 using stratified 10-fold cross-validation. The detailed accuracy by class showed excellent performance for the "Not Distress" class (Recall 0.981, Precision 0.957) and commendable performance for the "Distress" class (Recall 0.803, Precision 0.907). The confusion matrix revealed 5 false positives and 12 false negatives, which is noted as demonstrably lower than in prior results. While overall performance is promising, continued efforts on the "Distress" class to minimize false negatives are recommended, including fine-tuning parameters, exploring feature engineering, and addressing class imbalance.

Using Class Weights to Handle Imbalance: This experiment employed SMOTE oversampling to balance the dataset and improve the detection rate. Evaluated with stratified cross-validation, this model achieved exceptional performance. The overall accuracy soared to 97.70%, indicating a substantial improvement, and the Kappa statistic reached 0.9458, suggesting outstanding agreement beyond chance. Detailed accuracy by class revealed impressive gains: "Not Distress" had near-perfect performance (Recall 0.993, Precision 0.975), and more importantly, the "Distress" class showed remarkable improvement with a high Recall of 0.943 and Precision of 0.983, along with a significantly improved Precision-Recall Curve Area. The confusion matrix illustrated the enhancement, with only 2 false positives and a substantial reduction to only 7 false negatives, representing a major leap in minimizing overlooked distress cases. This model, leveraging SMOTE, achieves highly accurate performance, particularly in identifying financial distress, and its significantly reduced false negatives reinforce its practical utility.

5. Discussion

The research explores the application and evaluation of various machine learning algorithms—including Logistic Regression (LR), Decision Trees (REPTree), Artificial Neural Networks (ANNs), K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Random Forest—for financial distress prediction within the Ethiopian banking sector. The primary goal was to create reliable prediction models tailored to this unique context, moving beyond the limitations of traditional static models. The study meticulously defined its variables, utilizing the CAMEL technique for predictor variables and Altman Z-score measurements to define the predicted variable (financial distress indicated by a Z-Score below 2.99). A rigorous methodology was followed, encompassing thorough data preparation, which included cleaning, preprocessing, and feature selection, crucial for enhancing the relevance and applicability of the data for model development. The

performance of each algorithm was rigorously evaluated based on metrics such as accuracy, precision, and F-Measurement.

Logistic Regression (LR) served as a valuable baseline model due to its inherent simplicity and interpretability. While prior studies demonstrated moderate accuracy for LR, this research achieved a notable average accuracy of 96.4% on the training set, significantly surpassing previously reported results (e.g., 86.23% by). LR's limitations, specifically its reliance on linear decision boundaries and sensitivity to imbalanced datasets, were acknowledged. However, the study found that optimizing LR using feature selection techniques improved its accuracy by 3–4%, suggesting that LR can remain a reliable tool when combined with appropriate preprocessing. Similarly, REPTree, a decision tree algorithm, demonstrated competitive accuracy, achieving an average training accuracy of 96.07%, which was higher than the 89.67% average reported in prior studies. Although Decision Trees are prone to overfitting, a challenge addressed in this research through pruning techniques, their interpretability and flexibility in handling mixed data types make them a popular choice. The performance of REPTree was further enhanced when integrated into ensemble methods.

The more complex, non-linear algorithms, Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs), also showed strong predictive capabilities, though they introduced challenges related to interpretability. ANNs, specifically, achieved a high average accuracy of 94.41% on datasets processed with SMOTE (Synthetic Minority Oversampling Technique), notably higher than the 88.31% average from previous research. ANNs excel at learning intricate patterns and scaling to high-dimensional data, but they require extensive experimentation and careful tuning of hyperparameters to achieve optimal performance. SVMs, prized for their ability to find optimal hyperplanes in high-dimensional spaces, achieved an average accuracy of 90.03%, outperforming the prior average of 83.44%. Like ANNs, SVMs required careful hyperparameter tuning and employed techniques like grid search and cross-validation to improve robustness. Despite the high accuracy of both ANNs and SVMs, their "black-box" nature limited their usefulness in scenarios where model explainability is critical.

K-Nearest Neighbors (KNN) presented a balance of capability and limitation, achieving an average accuracy of 88.52%. KNN is flexible enough to handle non-linear patterns, but the research observed high computational cost for large datasets, limiting its scalability. This challenge, along with sensitivity to imbalanced datasets, was partially addressed through the application of feature selection techniques. However, the resulting relatively low Sensitivity/Recall (70.15%) indicated that further advanced sampling techniques might be necessary to correctly identify distressed cases. In contrast, the Random Forest (RF) model emerged as the most robust and highest-performing algorithm. By employing SMOTE and class weights, RF achieved an exceptional average accuracy of 97.70%, substantially surpassing the 91.54% average from previous studies. Random Forest is advantageous because it is less prone to overfitting than single decision trees and provides

insights into feature importance, which enhances its overall interpretability and applicable for decision-makers.

The superior performance of the Random Forest model (97.70% accuracy) observed in this study aligns with recent developments in financial distress literature, particularly regarding the dominance of machine learning algorithms and the necessity of addressing data imbalance. Consistent with our findings on the robustness of ensemble methods, Barboza and Altman [25] demonstrated that Random Forest models significantly outperform traditional logistic regression when predicting distress in emerging markets, validating the shift towards non-linear algorithmic approaches. Furthermore, the effectiveness of employing SMOTE to mitigate class imbalance in this research mirrors the conclusions of Kitova et al. [26], who established that integrating advanced oversampling techniques with machine learning classifiers substantially enhances predictive precision in imbalanced financial datasets.

In general, the comprehensive evaluation of machine learning techniques underscored that success in predicting financial distress is dependent not on identifying a single "best" algorithm, but rather on understanding the trade-offs and diligently combining strategies like hyperparameter tuning, careful data preprocessing, and addressing class imbalance. While various models like LR and REPTree benefited from feature selection and ensemble methods, Random Forests and ensemble methods proved generally most robust for high predictive accuracy and handling imbalanced data. The optimization efforts confirmed that when enhanced with SMOTE and class weights, the Random Forest model stands out for its high accuracy and superior recall in identifying distressed cases. Consequently, the researcher strongly recommends prioritizing the implementation of the Random Forest model with SMOTE as a robust Machine Learning-Based Early Warning System for financial distress prediction within the Ethiopian banking sector.

6. Conclusion & Future Work

This research predicted financial distress in the Ethiopian commercial banking sector by leveraging the CAMEL framework for predictor variables and a modified Altman Z-score to label distress, using data from 18 banks spanning 1990-2022. The study explored various machine learning algorithms (Logistic Regression, Decision Trees, ANNs, k-NN, SVMs, and Random Forest), employing optimization techniques like SMOTE oversampling, hyperparameter tuning, and ensembles, and emphasized the importance of data preprocessing and feature selection. Findings showed that the Random Forest model, specifically optimized with SMOTE oversampling combined with class weighting and evaluated via stratified cross-validation, achieved the highest performance (97.70% accuracy, 0.9458 Kappa), significantly improving the identification of distressed banks by effectively addressing imbalance and overfitting. Bagging was the second-best performer (96.07% accuracy). Recommendations include implementing the optimized Random Forest model as a machine learning-based Early Warning System, improving data quality, and fostering collaboration among stakeholders to enhance sector stability.

More investigation is required to delve into employing more sophisticated machine-learning methods like deep learning and ensemble techniques for predicting financial distress in the Ethiopian banking industry. Additionally, research should focus on incorporating macroeconomic factors and qualitative data into the models to improve their predictive power.

Reference

[1] O. Rehn, 'Olli Rehn: Monetary policy and bank supervision in Europe after the last financial and sovereign debt crisis and challenges for the future', 2018.

[2] K. L. Tran, H. A. Le, T. H. Nguyen, and D. T. Nguyen, 'Explainable Machine Learning for Financial Distress Prediction: Evidence from Vietnam', Data, vol. 7, no. 11, p. 160, 2022, doi: 10.3390/data7110160.

[3] F. S. Shie, M.-Y. Chen, and Y.-S. Liu, 'Prediction of corporate financial distress: an application of the America banking industry', Neural Comput. Appl., vol. 21, no. 7, pp. 1687–1696, Oct. 2012, doi: 10.1007/s00521-011-0765-5.

[4] C. Tarkocin and M. Donduran, 'Constructing Early Warning Indicators for Banks Using Machine Learning Models', North Am. J. Econ. Finance, 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:263654031

[5] E. Abdu, 'Financial distress situation of financial sectors in Ethiopia: A review paper', Cogent Econ. Finance, vol. 10, Feb. 2022, doi: 10.1080/23322039.2021.1996020.

[6] M. Abate and R. Kaur, 'BANKING SECTOR IN ETHIOPIA: ORIGIN AND PRESENT STATE', EPH - Int. J. Bus. Manag. Sci., vol. 9, pp. 1–13, Apr. 2023, doi: 10.53555/eijbmsv9i2.134.

[7] M. Summer, 'Bank Supervision and Resolution: National and International Challenges', Financ. Stab. Rep., pp. 107–111, 2011.

[8] R. B. Whitaker, 'The early stages of financial distress', J. Econ. Finance, vol. 23, no. 2, pp. 123–132, Jun. 1999, doi: 10.1007/BF02745946.

[9] T. C. Opler and S. Titman, 'Financial Distress and Corporate Performance', J. Finance, vol. 49, no. 3, pp. 1015–1040, Jul. 1994, doi: 10.1111/j.1540-62611994.tb00086.

[10] F. Barboza, H. Kimura, and E. Altman, 'Machine learning models and bankruptcy prediction', Expert Syst. Appl., vol. 83, pp. 405–417, Oct. 2017, doi: 10.1016/j.eswa.2017.04.006.

[11] P. Ravi Kumar and V. Ravi, 'Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review', Eur. J. Oper. Res., vol. 180, no. 1, pp. 1–28, Jul. 2007, doi: 10.1016/j.ejor.2006.08.043.

[12] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, 'SMOTE: Synthetic Minority Over-Sampling Technique', J Artif Intell Res JAIR, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[13] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, 'Gene Selection for Cancer Classification Using Support Vector Machines', Mach. Learn., vol. 46, pp. 389–

422, Jan. 2002, doi: 10.1023/A:1012487302797.

[14] J. Suss and H. Treitel, 'Predicting Bank Distress in the UK with Machine Learning', SSRN Electron. J., 2019, doi: 10.2139/ssrn.3465753.

[15] H. Le and J.-L. Viviani, 'Predicting bank failure: an improvement by implementing Machine learning approach on classical financial ratios', Res. Int. Bus. Finance, vol. 44, Jul. 2017, doi: 10.1016/j.ribaf.2017.07.104.

[16] Y.-P. Huang and M.-F. Yen, 'A new perspective of performance comparison among machine learning algorithms for financial distress prediction', Appl. Soft Comput., vol. 83, p. 105663, Aug. 2019, doi: 10.1016/j.asoc.2019.105663.

[17] D. T. M. Shahwan and M. A. Fadel, 'Machine Learning Models and Financial Distress Prediction of Small and Medium- Sized Firms: Evidence from Egypt'.

[18] F. Abid, A. Zouari, and A. F. Zouari, 'Financial distress prediction using neural networks: The Tunisian firms experience', no. 2, pp. 399–406.

[19] K. Meher and H. Getaneh, 'Impact of determinants of the financial distress on financial sustainability of Ethiopian commercial banks', Banks Bank Syst., vol. 14, pp. 187–201, Oct. 2019, doi: 10.21511/bbs.14(3).2019.16.

[20] E. Zelie and F. Wassie, 'Examining the Financial Distress Condition and Its Determinant Factors: A Study on Selected Insurance Companies in Ethiopia', World J. Educ. Humanit., vol. 1, p. p64, Aug. 2019, doi: 10.22158/wjeh.v1n1p64.

[21] H. A. Meressa, 'Evaluating Financial Distress Condition of Micro Finance Institutions in Ethiopia Using Altman's Revised Z-Score Model', Int. J. Account. Res., vol. 3, no. 4, pp. 35–42, Jan. 2018, doi: 10.12816/0044418.

[22] Y. Isayas, 'Financial distress and its determinants: Evidence from insurance companies in Ethiopia', Cogent Bus. Manag., vol. 8, p. 1951110, Jan. 2021, doi: 10.1080/23311975.2021.1951110.

[23] F. Betz, S. Oprică, T. A. Peltonen, and P. Sarlin, 'Predicting distress in European banks', J. Bank. Finance, vol. 45, pp. 225–241, 2014, doi: https://doi.org/10.1016/j.jbankfin.2013.11.041.

[24] T. Negash, G. Tesfaye, and O. Erana, 'Determinants of financial distress: evidence from insurance companies in Ethiopia', J. Innov. Entrep., vol. 13, Feb. 2024, doi: 10.1186/s13731-024-00369-5.

[25] F. Barboza and E. I. Altman, 'Predicting financial distress in Latin American companies: A comparative analysis of logistic regression and random forest models', *The North American Journal of Economics and Finance*, vol. 72, p. 102158, 2024, doi: 10.1016/j.najef.2024.102158.

[26] K. Kitova, B. Toleva, and I. Ivanov, 'Improving Financial Distress Prediction through Clustered SMOTE for Imbalanced Data', *Advances in Artificial Intelligence and Machine Learning*, vol. 5, no. 2, pp. 3663–3680, Apr. 2025