Journal of Information Systems Engineering and Management
2025, 10(9s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Intelligent Corporate Bankruptcy Prediction: A Data-Driven Machine Learning Approach

Bhavika Nemade[1], Bhushankumar Nemade[2], Bhushan Jadhav[3], Arti Deshpande[4], Uma Goradiya[5], Aabha Patil[6]

[1]*N. L. Dalmia Institute of Management Studies & Research, Mumbai University, India bhavikanemade@gmail.com,*

[2]*Shree L R Tiwari College of Engineering, Mumbai University, India, bnemade@gmail.com,*

[3]*Thadomal Shahani Engineering College, Mumbai, India, bhushan.jadhav@thadomal.org,*

[4]*Thadomal Shahani Engineering College, Mumbai, India, arti.deshpande@thadomal.org,*

[5]*Shree L R Tiwari College of Engineering, Mumbai University, uma.goradiya@slrtce.in,*

[6]*Pravin Rohidas Patil College of Engineering and Technology, Mumbai University, India. patil.a.aabha@gmail.com*

ARTICLE INFO

Received: 03 Nov 2024

Revised: 24 Dec 2024

Accepted: 06 Jan 2025

ABSTRACT

Over the past few decades, corporate bankruptcy has become a source of concern for various business stakeholders such as investors and management. It has also piqued the interest of researchers worldwide. Because there are so many variables that contribute to bankruptcy, it is insufficient to rely solely on a single predictive model; instead, the difficulty is in pinpointing the crucial elements that carry the greatest weight. The significant class imbalance in the data is another significant obstacle that impairs the model's functionality. While many methods, such as Decision Trees, Support Vector Machines, Neural Networks, etc., have been studied before with various pre-processing approaches, we advance this research by applying a novel combination of a Random Forest feature selection technique and BORDERLINE-SMOTE hybrid resampling technique on Polish Bankruptcy dataset. We next apply five classifiers to the altered data: Random Forest, Decision Tree, XGBoost, CatBoost and AdaBoost, and assess the effectiveness of each model. As a result of our suggested approach, provides Random Forest classifier with the highest accuracy of 94% using optimal prediction based on voting principle.

Keywords: Corporate bankruptcy, deep learning, Borderline-SMOTE, AdaBoost, SVM, Random Forest, Decision Tree, XGBoost, CatBoost and AdaBoost

I. INTRODUCTION

Bankruptcy is a term that instils fear in creditors, investors, employees, and management alike. It signifies a situation where a company is unable to secure the necessary funds to sustain its operations or fulfil its financial obligations to creditors. Being in such a precarious position, where the company's operations have come to a standstill, leaving suppliers and customers stranded, is far from ideal.

In the current global market, characterized by fierce competition and unpredictability, even large businesses could potentially be pushed towards bankruptcy. A recent example that underscores this is the unexpected bankruptcy of Lehman Brothers, a global financial services firm. This sudden event in 2008 triggered a severe economic downturn and left thousands of employees jobless. Another instance is the bankruptcy of Blockbuster, a popular video rental store. Despite its popularity, the company filed for bankruptcy in 2010 due to mounting debt and the rise of digital streaming platforms.

In addition to these examples, there have been several intriguing bankruptcy cases involving complex legal proceedings. For instance, RPS Infrastructure Ltd. and KST Infrastructure Private Limited were involved in a case where, despite an arbitral award against the debtor in 2016, the debtor contested the award, leading to the initiation of a corporate insolvency resolution process. In another case, UNITECH Limited and Telangana State Industrial Infrastructure Corporation (TSIIC) were involved in a dispute where it was argued that a remedy for the recovery of moneys arising out a contractual matter cannot be availed of under Article 226 of the Constitution. Lastly, in a case concerning Gujarat Urja Vikas Nigam Limited and Mr Amit Gupta & Ors, the Supreme Court held that the National Company Law Tribunal (NCLT) has jurisdiction to adjudicate contractual disputes which arise solely from or which relate to the corporate debtor's insolvency. These cases underscore the intricate nature of bankruptcy proceedings and the unexpected turns they can take. These examples highlight the devastating impact of bankruptcy, not just in terms of financial loss, but also in terms of job loss and the ripple effects on the economy. Therefore, the importance of accurate bankruptcy prediction to prevent such outcomes cannot be overstated.

Corporate bankruptcy not only has a detrimental effect on a nation's economy, but it also contributes to the spread of recessions. The ability of a company to make informed economic decisions is significantly influenced by its capacity to predict bankruptcy. This is crucial because all businesses, regardless of their size, affect a wide range of stakeholders, including investors, community members, industry participants, and policy officials [1]. The performance of the corporate sector plays a substantial role in shaping a nation's economy. Therefore, there may be instances where it becomes essential for creditors to scrutinize financial data and assess the probability of bankruptcy. Such decisions can have far-reaching implications, influencing not only the growth trajectory of an economy but also the evolution of a corporation. Thus, the ability to accurately predict bankruptcy is of paramount importance in the realm of corporate decision-making.

The recent worldwide financial crisis has highlighted the urgent need for a rapid and dependable model for predicting bankruptcy. The Structural approach, which uses data mining techniques to closely examine the characteristics of a firm and interest rates to predict the likelihood of default, and the Statistical approach, which employs statistical methods to estimate the bankruptcy rate, are the two main methods used globally to forecast a company's likelihood of filing for bankruptcy.

Despite the exploration of various methodological techniques using generalized linear models and multidimensional analysis, the surge in available data has made linear models less reliable and ineffective in determining the relationship between economic indicators. In recent decades, artificial intelligence and machine learning have emerged as potent tools for predicting corporate insolvency [2]. Most research in this field has used both market-based and accounting-based variables in numerical format. It's widely acknowledged that a company's bankruptcy status is influenced by a wide range of characteristics, making accurate prediction challenging. These characteristics include sales, purchases, assets, liabilities, EBIT, trade data, earnings per share, among others. To enhance prediction performance and reduce computing time and cost, it's crucial to use the most pertinent characteristics.

Another challenge is the class imbalance in the data, with a large number of non-bankrupt businesses and a small number of bankrupt ones. This imbalance can skew the model in favor of the majority class, or non-bankrupt companies, if not properly addressed [3]. The question then arises: Can we improve the performance of corporate bankruptcy prediction compared to state-of-the-art methods by using a new combination of a feature extraction technique and a resampling strategy? Feature extraction techniques involve using a set of algorithms to examine each variable's relationship with the dependent variable, and then selecting and retaining only those variables that have a strong association and are valuable for prediction. Resampling techniques, on the other hand, are used to correct dataset imbalances. These strategies must be effectively implemented to develop a robust and accurate prediction model.

II. RELATED WORKS

Class imbalance is one of the most difficult aspects of solving the bankruptcy prediction problem. As far as we are aware, there are tens of thousands of steady, profitable businesses operating around the world, but there are comparatively few bankrupt businesses. Despite their little number, these insolvent businesses have the potential to cause economic disruption for the entire nation and spark a financial crisis among lenders and investors. There are a few methods that have been investigated by researchers in earlier studies to address the problem of class imbalance in machine learning models. This section's literature will help us get a general idea of how well various methods work in the bankruptcy prediction problem.

Two strategies were combined by the authors of a study in [4] to address the problem of class imbalance. Using a cost-sensitive learning method and an oversampling strategy on a Korean bankruptcy dataset, the authors have developed a hybrid approach. First, to determine the appropriate performance on the validation set, an optimal balancing ratio is employed in conjunction with an oversampling module. Second, a cost-sensitive learning model for bankruptcy prediction is based on the C-Boost algorithm. The dataset had 120048 non-bankrupt films and 307 bankrupt films, resulting in a 0.0026 balancing ratio. Although it has a few drawbacks, the oversampling strategy increases prediction accuracy by adding synthetic data to the minority class. The model is likely to be over-fit since it replicates minority samples exactly as they now exist. Additionally, it lengthens the training period and increases the amount of memory needed to store the training set by increasing the number of training examples. The SMOTEENN approach, which creates synthetic minority samples based on feature similarities between the minority classes, has been utilised for oversampling. Moreover, clusters of the majority class are created using the C-Boost algorithm. The following techniques were then used with this approach: Multilayer Perceptron, AdaBoost, Random Forest, and Bagging. The AUC and G-mean evaluation measures were applied. Although the results were improved by the use of oversampling, feature selection techniques were not used in this investigation.

The application of oversampling approaches to enhance the accuracy of bankruptcy prediction was also investigated in [5] [6]. In this instance, a specific percentage of the majority class was randomly replaced with the minority class. The top characteristics were selected using three distinct feature selection methods: Mutual Information, Random Forest Genetic Algorithm. The results of the aforementioned methods were fed into machine learning algorithms, which included random forests, logistic regression, and random forests. Overall, the findings demonstrated that increasing the sample size does enhance prediction accuracy.

Bankruptcy prediction and credit scoring are crucial areas in the fields of finance and accounting, and have been the subject of extensive research. Techniques from artificial intelligence and machine learning, especially the multilayer perceptron (MLP) network trained using the back-propagation learning algorithm, have proven to be valuable in tackling these financial decision-making challenges. Recent research suggests that ensemble methods, which use multiple classifiers, could potentially offer better results than single classifiers. However, the effectiveness of these multiple classifiers in predicting bankruptcy and credit scoring is still not fully comprehended. This paper [7] compares a single classifier's performance with that of multiple classifiers and diversified multiple classifiers, using neural networks across three different datasets. The findings indicate that while multiple classifiers outperform the single classifier in one dataset, diversified multiple classifiers underperform in all datasets [8]. However, when it comes to Type I and Type II errors, there is no definitive winner. Hence, it is suggested that these three classifier architectures should be considered for making optimal financial decisions.

In the referenced work [9], the authors propose the use of a cluster-based boosting method, known as C-boost, in tandem with the Instance Hardness Threshold (IHT). The latter is commonly used to filter out noisy instances. The C-boost algorithm is employed to tackle the issue of class imbalance. The authors apply this approach to a Korean bankruptcy dataset, constructing a C-boost prediction model after performing resampling. The results indicate that the proposed framework outperforms several existing techniques. This includes the GM-Boost algorithm and a method that utilizes SMOTEENN for oversampling, achieving an Area Under the Curve (AUC) of 87%. This suggests the effectiveness of the proposed method in handling class imbalance issues.

The research study [10] employs the Random Forest algorithm, a prevalent machine learning classifier, to predict potential bankruptcy or distress among Indian firms. The algorithm categorizes firms based on their risk of default or propensity for distress, providing valuable insights into the correlation between a firm's characteristics and its likelihood of failure. The study compares these findings with those derived from Tree Net, another advanced data mining tool renowned for its robust estimation capabilities. The comparison is drawn using both in-sample and out-of-sample estimations. The analysis is comprehensive, encompassing a wide array of companies from diverse sectors. This study concludes that the Tree Net methodology consistently surpasses the Random Forest approach in terms of classification accuracy and predictive performance. This superior performance of Tree Net underscores its potential for further utilization by industry analysts and researchers for predictive modelling purposes. The model's accuracy, which was the evaluation metric utilised, was 94% with all features and 96% with six features.

A study in [11] used a geometric mean based boosting technique to overcome the issue of data imbalance. This technique takes into account both the majority and a minority class since it employs the geometric mean of the two classes to calculate accuracy and error rate. AdaBoost and cost-sensitive boosting are used to compare the outcomes. The Korean commercial bank provided the dataset that was used in this investigation. After taking into account 30 financial measures, there were 500 insolvent enterprises and 2,500 non-bankrupt companies. Two distinct data samples with five sample groups (1:5,1:3, 1:20,1:1,1:10) were created in order to validate the effectiveness of the GM-Boost algorithm. Cost Boost, GM-Boost, and AdaBoost tests were then conducted on those unbalanced datasets. The SMOTE method was employed in the second stage to create new bankruptcy data, and the newly created sample sets were then applied to SMOTE-SM-Boost, SMOTE-Cost-Boost, and SMOTE-Boost. The outcomes demonstrated GM-Boost's promising performance with excellent prediction performance.

By contrasting machine learning models with statistical models, authors in paper [9] have researched on this subject by determining which methodological approach is superior. The predictor variables included asset turnover, liquidity, profitability, leverage, and productivity in a balanced dataset. Bagging, boosting, Random forest, ANN, SVM with two kernels (linear and radial basis), logistic regression, and MDA were among the techniques used. The outcomes demonstrated that machine learning models outperformed conventional statistical models. This study's primary weakness is that no feature selection strategy was used, despite it being a common practice these days.

The research study [10] utilized a variety of data and techniques. Recent studies in this field have employed diverse variables. In this particular study, an optimized neural network with six hidden layers, a support vector machine, and XGBoost classification algorithms were applied to the financial data of 3728 Belgian enterprises. This approach achieved a global bankruptcy prediction accuracy of 82–83%. Compared to Brédart’s 2014 analysis of the same dataset using a shallow neural network, this study shows a modest 2% improvement in bankruptcy cases and a significant 17% improvement in solvency cases. The study acknowledges that the prediction accuracy is limited due to the substantial overlap in the feature space between financial variables of bankrupt and solvent companies. Interestingly, the study does not report any significant differences in prediction accuracy across different techniques used. Moreover, a high prediction accuracy rate is achieved using only three easily obtainable financial ratios, which is beneficial for most firms. This study not only contributes to academic literature but also holds significant value for bankers [11]. It enables them to assess the probability of bankruptcy (and hence non-reimbursement) of firms seeking loans without the need to compute numerous financial ratios or collect non-financial data.

The study presented in [12] deployed a prediction model that is based on SVM. To get the ideal parameter value, the author used a 10fold CV in conjunction with the grid-search technique. Two Chinese cities' A-share market data, which includes the financial ratios of 250 businesses, has been selected. RBF SVM produced superior outcomes than MDA and BPNN.

Xinke Chong introduced a study on an early warning model that employs the PSO-SVM methodology. The model uses Particle Swarm Optimization (PSO) to concurrently optimize the selection of the feature set and the parameters of the kernel function in the Support Vector Machine (SVM) model. The study’s empirical evidence underscores the robust generalization capabilities of SVM, a cutting-edge machine learning method. However, it also highlights that the choice of the feature set and kernel function parameters significantly influences the SVM model’s predictive performance. The proposed PSO-SVM model addresses this by optimizing the selection of feature sets and SVM kernel function parameters simultaneously, yielding impressive results. It enhances the predictive power of the model, achieving an accuracy rate of 90.30%.

Le T. presented a study [13] that compared three data mining techniques: random forest, decision trees, and logit on balance sheet data from 446,464 firm statements from many nations, including Italy, Germany, France, Britain, Portugal, and Spain. The author did not employ any particular resampling technique; instead, accuracy, specificity, sensitivity, and precision were the criteria employed for assessment. The outcomes demonstrated that the random forest model outperformed both the decision tree and the logit model, whereas the logit model outperformed both.

M. Shakil et al. [14] presented a study that offers a comprehensive, qualitative examination of feature selection techniques utilized in research, adhering to the PRISMA protocol for systematic reviews and meta-analyses. It aims to amalgamate key elements such as feature selection methods, diverse machine learning techniques, evaluation criteria, and research outcomes into a single resource. The study spans from 2015 to 2021, incorporating 36 articles extracted from the Scopus database and selected research papers. The findings categorize feature selection approaches into “filter”, “wrapper”, and “embedded”, with the filter approach emerging as the preferred choice due to its simplicity and superior results. The study also underscores the use of multiple feature selection methods within these categories to identify the most pivotal variables.

Hassan Raza [15] et al. presented a findings that have significant implications for Pakistan’s non-financial sector. Policymakers and regulatory authorities can use these insights to develop effective frameworks and laws to mitigate systemic risks in the financial sector. By identifying the financial ratios that contribute to bankruptcy prediction, regulators can establish guidelines for monitoring companies' financial health. Financial institutions can use these models to manage their exposure to potentially risky borrowers, leading to a more resilient banking system. Businesses can also leverage these findings to monitor their financial health and make necessary adjustments to prevent financial bankruptcy. The study underscores the benefits of machine learning techniques in bankruptcy forecasting, enhancing prediction accuracy and reducing financial investment risks.

Fernando M. et al. [16] applied a Multi-Objective Evolutionary Algorithm (MOEA) for feature selection in bankruptcy prediction. Their goal was to optimize the accuracy of the classifier while minimizing the number of features. They thoroughly scrutinized a dual-objective problem - minimizing the number of features and maximizing accuracy - using two classifiers: Support Vector Machines and Logistic Function. The research utilized a database of financial statements from 1200 medium-sized private French firms [17] [18]. The results underscored the efficiency of MOEA as a feature selection approach, providing valuable insights for decision-makers in evaluating a company's financial health [19] [20]. Upon reviewing the most recent study in this field, it was discovered that a wide range of methodologies had been examined and that researchers worldwide had taken a particular interest in this subject. Studies have progressed from using statistical methods to using machine learning models and deep learning techniques in order to increase the model's performance [21]. The distribution of the data and the variety of circumstances that can cause a company to fail present a challenge. As a result, we will go beyond what has already been written about the subject in the current work and apply feature selection and resampling techniques to a variety of machine learning models.

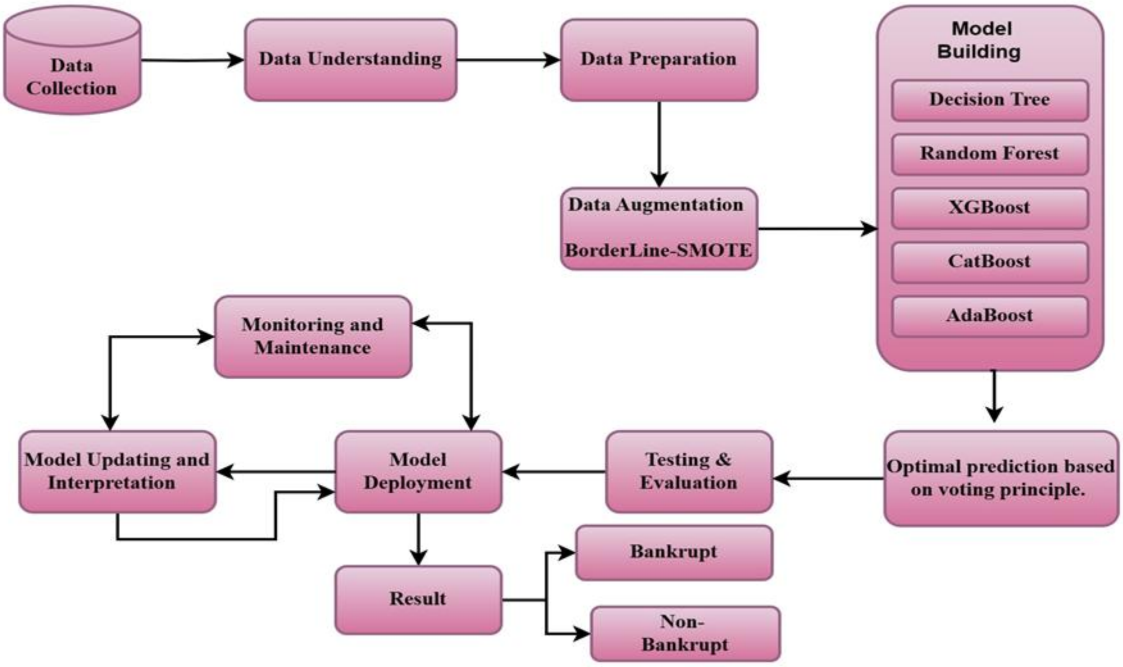


Figure 1: The proposed architecture

III. PROPOSED METHODOLOGY

The architecture of proposed system is represented in figure 1. It consists of various steps like data collection, Data understanding, data preparation, data augmentation, model building, optimal prediction based on voting principle, testing and evaluation, model deployment, model updating and maintenance. The main goal of this research is to predict corporate bankruptcy, using a variety of financial metrics such as profit, sales, assets, and so forth, whether or not a firm is about to file for bankruptcy. Therefore, the goal of this effort is to create a predictive bankruptcy model that will

be helpful to management, employees, and nvestors/creditors and that can warn them about any impending bankruptcy threat to a company.

The primary limitation of this issue, as noted in earlier research, is that there are thousands of non-bankrupt companies in the world but very few bankrupt companies. As a result, it is challenging for the model to learn and train itself using the scant data on bankrupt cases. Second, bankruptcy is caused by a wide range of causes, not just a select handful. This study also aims to determine the key financial characteristics that push a business towards bankruptcy. This study will help a great deal of people all around the world and be a step forward in the field of bankruptcy prediction.

A Data Collection

The Polish Bankruptcy dataset, gathered for corporate bankruptcy prediction, originates from the Emerging Markets Information Service (EMIS), a global database focusing on emerging markets. The dataset encompasses data from bankrupt firms examined between 2000 and 2012, and from firms still in operation assessed between 2007 and 2013. It contains 64 financial ratio variables and one response variable. The specific data collection methodologies, such as company selection criteria or financial ratio calculation procedures, may not be publicly accessible.

B Data Understanding

Understanding the data is the next stage of the CRISP-DM process. Without a thorough grasp of each component, it is very impossible to construct a trustworthy prediction model before beginning construction. Although there may be several sources where the necessary data is available, it might not be moral to retrieve it from just a few of them. Furthermore, the information found in a small number of sources may not be trustworthy, thus it may take some time to extract the information from a trustworthy source while adhering to ethical standards. We need the financial ratios of businesses in a specific area for a given time period, including both bankrupt and non-bankrupt cases, in order to continue our research.

C Data Preparation

Accurate and high-quality outputs from the model depend on the preparation of the data, which is a crucial phase in the data mining process. Big data frequently contains noise, and the inclusion of special characters, missing values, and blank spaces frequently has an impact on the model's performance. As such, it must be treated carefully to guarantee that the important information is kept unaltered throughout the data preparation process. Though it is generally believed that the more data we have, the more accurate forecasts we may make, this isn't necessarily the case. Certain aspects might be wholly ineffective in predicting the target variable, whereas a small number of traits might be more significant and responsible. Since having extraneous and pointless variables will increase computing time and cost, the CRISP-DM methodology's data preparation step is frequently regarded as the most difficult and time-consuming step. The following is a point-by-point list of the actions conducted to prepare the data needed for this study:

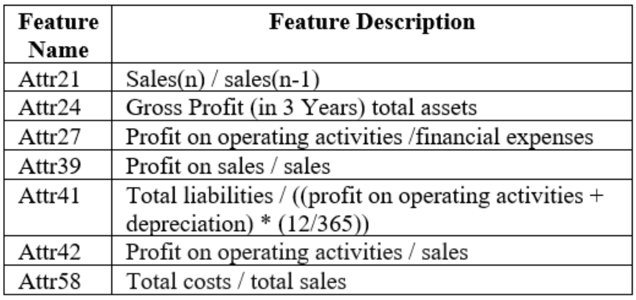
- File type conversion: The data file was originally in ARFF format after being extracted from the source. It needed to be converted to a CSV format in order to comply with our model's needs. The .arff file has been transformed into a CSV file using Python code.
- Modifying Every Variable's Datatype: Normally, after the dataset is sourced, every variable's data type is noted. Each variable had a numeric value with a distinct data type. The data type of these variables was changed to Float.
- Managing NAs, or missing values: The existence of missing values can cause issues for the model and impair its functionality. Missing values were replaced in the dataset with the special character ('?'). The percentage of missing data in each column has been noted, and these values were changed to NAs in order to handle them further. In light of this, the column mean was used to impute missing values for columns with a small percentage of missing values. The dataset has to have the column with about half of its values missing entirely.
- Multi-collinearity: A correlation test reveals the correlation between the variables as well as how these variables are influencing the target variables. High correlation variables (p value > 0.90) can be considered redundant data because they actually contribute equally to the prediction of the target variable. It's best to leave one variable alone and keep the other one. 26 of the 64 predictor factors in the dataset had to be removed from the analysis because of their strong correlation with the other variables.

D Feature Selection

Feature selection is one method that can help the model perform even better after the data has been cleaned and superfluous and redundant variables have been eliminated. This method is used to shrink the feature space, which could be advantageous for the model. The advantages could be increased precision, decreased over-fitting risk, quicker calculation times, and enhanced model ability. When there are too many features in the dataset, ability is lost.

One of the pre-processing steps before creating a classification model is feature selection, which addresses the issue of dimensionality curse, which negatively affects the algorithm. A small number of the 64 features in the dataset utilised in this study could not be helpful in predicting bankruptcy. In this study, the Random Forest feature selection technique has been employed to choose the optimal characteristics by removing the superfluous features. The tree-based tactics employed in random forests are ranked according to how well they can increase node purity. This is referred to as gini impurity, or mean decrease in impurity. The beginning of the tree experiences the most drops in node purity, whereas the ending of the tree experiences the least amount of decrease. Thus, by identifying a certain node and then trimming below it, a subset of significant features is produced in this way.

Table 1: Attribute Features



Feature Name	Feature Description
Attr21	Sales(n) / sales(n-1)
Attr24	Gross Profit (in 3 Years) total assets
Attr27	Profit on operating activities /financial expenses
Attr39	Profit on sales / sales
Attr41	Total liabilities / ((profit on operating activities + depreciation) * (12/365))
Attr42	Profit on operating activities / sales
Attr58	Total costs / total sales

E Imbalanced Data Handling

SMOTENN (Synthetic Minority Over-sampling Technique for Nominal and Numerical data) is a technique used to address class imbalance in the dataset. It works by creating synthetic examples of the minority class, helping to improve the model’s performance on minority class predictions. This step is crucial for ensuring that the model performs well across all classes, not just the majority class.

F Algorithms Used

This step in the machine learning process is seen to be crucial and important. The suggested models are to be put onto practice after the feature selection and resampling steps of the data preparation process. After that, the effects of the various pre-processing methods on the various models may be assessed and contrasted. This section covers the specifics of the model that was employed as well as how it functions.

- Random Forest: An ensemble of several decision trees makes up a random forest model, which is frequently applied to classification issues. It builds each tree using methods like feature randomness and bagging to produce an uncorrelated forest of trees. Every tree depends on a separate, unbiased sample. Compared to a single tree, the prediction performance of this group of trees is more accurate. A few characteristics that make it a good fit for the selected dataset are the model's fast training speed, resilience to outliers, and capacity to manage unbalanced data.
- Decision Tree: This supervised learning approach is popularly used to solve regression and classification issues. It has a separate tree representation, with attributes corresponding to the internal nodes of the tree and each leaf node representing a class label. The training data serves as the root at first, and it subsequently divides into smaller subgroups with decision and leaf nodes. The difficult aspect is figuring out which attribute the root node in each level is. Information gain and Gini index are two popular measurements for this process. We will assess this model's performance on our collection of variables since, according to prior research, it has demonstrated strong performance on the bankruptcy prediction problem.
- AdaBoost: AdaBoost: Boosting algorithms are thought to be strong and adaptable. In classification problems, adaptive boosting, often known as AdaBoost, is a kind of boosting algorithm that builds a strong classifier by transforming a number of weak ones.

The architecture and process flow diagram used for our research is displayed in the figure 2. After first extracting the data from the source, we pre-processed it by deleting unnecessary columns and imputing NAs using mean values.

We then separated out the less significant ones using the feature selection technique. After splitting the data into train and test, stratified K-fold cross validation (k=5) was used. Next, the dataset was resampled using the BORDERLINE-SMOTE approach. The final step involves feeding the processed data to five distinct classifiers and assessing each one's performance using the testing data.

XGBoost (Extreme Gradient Boosting) is a decision-tree-based ensemble Machine Learning algorithm that leverages a gradient boosting framework. It's especially useful in scenarios dealing with unbalanced data sets, such as corporate bankruptcy prediction, due to its capability to manage sparse data and its resilience to outliers. An optimized version of XGBoost, known as FS-XGBoost, has been developed specifically for corporate bankruptcy prediction. This version employs a feature importance strategy to select significant variables, ensuring high-quality classification performance.

CatBoost (Category Boosting) is another gradient boosting algorithm that has proven its efficiency in handling categorical variables, which are often found in corporate data. CatBoost employs various statistical methods to convert categorical values into numbers, considering combinations of categorical features and combinations of categorical and numerical features. A study demonstrated the effectiveness of the CatBoost algorithm in detecting company bankruptcy based on financial and categorical data from small and medium-sized enterprises, with the CatBoost model outperforming other models that used only financial or categorical variables. Both XGBoost and CatBoost have shown their robustness and effectiveness in the field of corporate bankruptcy prediction, adeptly handling complex financial data. However, the specific implementation details may vary depending on the dataset and the business problem. The proposed method selects the best classifier using optimal prediction based on the voting principle.

Model Deployment: This is the stage where the trained machine learning model is integrated into a real-world setting to make practical predictions. It requires careful planning to ensure the model operates correctly and efficiently.

Testing & Evaluation: This involves measuring the model's performance using specific metrics. The goal is to ensure the model's predictions are accurate and that it generalizes well to new, unseen data. Optimal Prediction Based on Voting Principle: This technique, used in ensemble learning, makes the final prediction based on the majority vote from the ensemble of models. It often yields better performance than any individual model in the ensemble.

Model Updating and Interpretation: This process involves fine-tuning the model based on fresh data. It also includes interpreting the model's predictions to understand its decision-making process.

Monitoring and Maintenance: This is an ongoing process that ensures the machine learning model's performance remains optimal over time. It involves regular checks and updates as needed, to address any changes in the data or the operational environment.

IV. DESIGN SPECIFICATIONS

The use of the suggested models for predicting corporate bankruptcy is covered in length in this section. Additionally, it explains the procedure used to resample the dataset and choose the most crucial attributes. Python 3.7.2 was used for the entire implementation phase, and Jupyter notebook (v.6.0.2) was selected as the integrated development environment (IDE). Python was selected for the implementation phase due to its ease of use, extensive online support community, and reputation as one of the best languages for readable code. Because there is a thriving Python community, there are many packages available for handling unbalanced data and preparing it, making it a popular choice for machine learning projects. Based on the predicted period, five separate files with data were sourced for this study; the fifth file has been selected for implementation. It included financial rates for the fifth year along with a class designation that indicated bankruptcy status after a year. The original files were in ARFF format; to convert them to CSV format, a Python code could be found online. After that, the dataset was imported as a Dataframe into Python, and any missing values were examined. The pandas profiling package was used to investigate individual columns and replace the special character ('?') that was used to indicate a missing value with NAs. We decided to use a feature selection method to narrow down the features and choose only the finest ones after completing the fundamental cleaning task. Using the Select from model 7 module from the sklearn.feature selection library, the random forest feature selection technique was applied. The best characteristics have been filtered using a gini significance threshold value of 0.03. The filtered dataset was subsequently separated into training and testing sets using a subsequent procedure. To do a stratified k fold split with a k value of 5, utilise the Fold 8 package from the skl-earn library.

A significant class disparity was seen when the data was explored. After the data was separated, we utilised the BORDERLINE-SMOTE approach to balance the classes in order to prevent over-fitting and create a dependable model. For resampling, the BORDERLINE-SMOTE package from the imbalance-learn library was utilised. We further utilised four distinct models on the dataset that was resampled. The several models that were used were Random Forest, Decision Tree, XGBoost, AdaBoost and CatBoost. These models can be found in the Python sk-learn library as several packages.

V. RESULTS

In the field of predicting corporate bankruptcy, the application of BorderLine-SMOTE in conjunction with various classifiers has been investigated. BorderLine-SMOTE, a version of the Synthetic Minority Over-sampling Technique (SMOTE), creates synthetic samples along the decision boundary between classes, thereby improving the performance of classifiers on instances of the minority class that are more challenging to classify. The effectiveness of these classifiers, which include Random Forest, Decision Tree, XGBoost, AdaBoost, and CatBoost, is assessed using metrics such as Accuracy, Recall/Sensitivity, Specificity, and Geometric Mean (GM). After selecting features, each classifier's base model is constructed using the unbalanced data. Table 2 shows the performance of each classifier individually without using BORDERLINE-SMOTE.

Table 2: The performance evaluation of classifiers without BorderLine-SMOTE

Model	Accuracy	Recall/ Sensitivity	Specificit y	Geometric Mean (GM)
Random Forest	91%	88%	91%	89%
Decision Tree	85%	84%	84%	83%
XGBoost	90%	86%	86%	85%
AdaBoost	86%	82%	83%	81%
CatBoost	88%	87%	87%	84%

In the task of corporate bankruptcy prediction, a comparative performance evaluation of various classifiers with BorderLine-SMOTE reveals that the Random Forest classifier outperforms others with an accuracy of 94%, recall/sensitivity of 91%, specificity of 93%, and a geometric mean of 91%. While XGBoost and CatBoost demonstrate respectable performance, they do not surpass the metrics achieved by Random Forest. The Decision Tree classifier exhibits the lowest performance among the evaluated classifiers. Therefore, considering the provided metrics, Random Forest in conjunction with BorderLine-SMOTE appears to offer the most optimal performance for corporate bankruptcy prediction. The graphical representation of Table 2 is presented in figure 2.

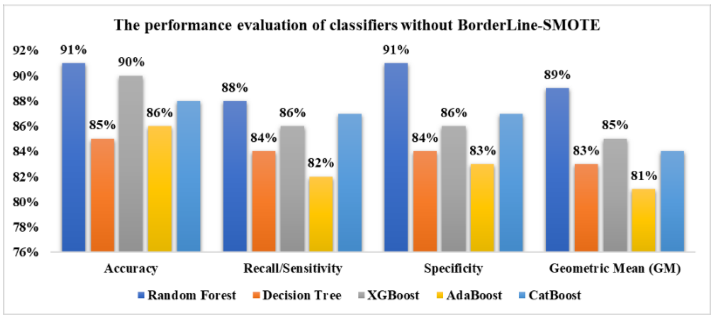


Figure 2: The performance evaluation of classifiers without BorderLine-SMOTE

Case Study 2:

We have now run the identical experiment on the resampled dataset using BORDERLINE-SMOTE. Table 3 shows the performance of each classifier individually using BORDERLINE-SMOTE.

Table 3: The performance evaluation of classifiers with BorderLine-SMOTE

Model	Accuracy	Recall/ Sensitivity	Specificity	Geometric Mean (GM)
Random Forest	91%	88%	91%	89%
Decision Tree	85%	84%	84%	83%
XGBoost	90%	86%	86%	85%
AdaBoost	86%	82%	83%	81%
CatBoost	88%	87%	87%	84%

In the evaluation of corporate bankruptcy prediction, five classifiers - Random Forest, Decision Tree, XGBoost, AdaBoost, and CatBoost - were assessed without the use of BorderLine-SMOTE. The performance metrics used for this evaluation included Accuracy, Recall/Sensitivity, Specificity, and Geometric Mean (GM). Among these classifiers, Random Forest demonstrated superior performance across all metrics, achieving an accuracy, recall/sensitivity, specificity, and GM of 91%. While XGBoost also showed commendable performance, particularly in specificity (91%), it did not surpass the metrics achieved by Random Forest. The Decision Tree classifier, on the other hand, exhibited the lowest GM (81%), indicating it might not be the most effective choice for balanced performance across different aspects. Therefore, considering all four metrics, Random Forest appears to provide the most optimal performance for corporate bankruptcy prediction in the absence of BorderLine-SMOTE. The graphical representation of Table 3 is presented in figure 3.

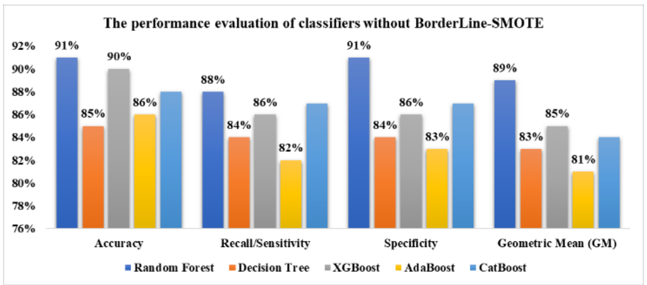


Figure 3 The performance evaluation of classifiers with BorderLine-SMOTE

VI.CONCLUSIONS AND FUTURE WORKS

In the domain of bankruptcy prediction, there has been substantial interest and numerous methodologies have been employed to optimize forecast performance. A key challenge lies in identifying the primary financial characteristics that contribute to a company's bankruptcy. Another issue addressed in this research is the problem of class imbalance. In this study, four models were constructed using a unique combination of a resampling technique and a feature selection technique. The performance of these classifiers, including Random Forest, Decision Tree, XGBoost, AdaBoost, and CatBoost, is evaluated using metrics such as Accuracy, Recall/Sensitivity, Specificity, and Geometric Mean (GM). The Random Forest classifier, when used with Random Forest feature selection and BORDERLINE-SMOTE, outperforms the other classifiers in terms of prediction accuracy. However, when methods comparable to the AdaBoost classifier are used, they perform better than other classifiers in terms of Geometric Mean and Recall values. XGBoost and CatBoost also show commendable performance, particularly in specificity, but they do not surpass the metrics achieved by Random Forest. The Decision Tree classifier exhibits the lowest GM, indicating it might not be the most effective choice for balanced performance across different aspects. The study was conducted using bankruptcy data from Poland, but the methods could be applied to bankruptcy data from other regions in future studies. Therefore, considering all four metrics, Random Forest appears to provide the most optimal performance for corporate bankruptcy prediction without BorderLine-SMOTE. Nevertheless, the choice of classifier and technique can depend on the specific requirements of the task, and it's always recommended to experiment with different methods and evaluate them based on relevant metrics.

REFERENCES

[1] T. K. Chen, H.-H. Liao, G. D. Chen, W.-H. Kang, and Y. C. Lin, "Bankruptcy prediction using machine learning models with the text-based communicative value of annual reports," Expert Systems with Applications, vol. 233, p. 120714, 2023. [Online]. Available: https://doi.org/10.1016/j.eswa.2023.120714

[2] Q. Yu, Y. Miche, E. Séverin, and A. Lendasse, "Bankruptcy prediction using Extreme Learning Machine and financial expertise," Neurocomputing, vol. 128, pp. 296-302, 2014. [Online]. Available: https://doi.org/10.1016/j.neucom.2013.01.063

[3] F. Mai, S. Tian, C. Lee, and L. Ma, "Deep learning models for bankruptcy prediction using textual disclosures," European Journal of Operational Research, vol. 274, no. 2, pp. 743-758, 2019. [Online]. Available: https://doi.org/10.1016/j.ejor.2018.10.024.

[4] R. F. Brenes, A. Johannssen, and N. Chukhrova, "An intelligent bankruptcy prediction model using a multilayer perceptron," Intelligent Systems with Applications, vol. 16, p. 200136, 2022. [Online]. Available: https://doi.org/10.1016/j.iswa.2022.200136.

[5] M.Y. Chen, "Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches," Computers & Mathematics with Applications, vol. 62, no. 12, pp. 4514-4524, 2011. [Online]. Available: https://doi.org/10.1016/j.camwa.2011.10.030.

[6] R. Robatto, "Systemic banking panics, liquidity risk, and monetary policy," Review of Economic Dynamics, vol. 34, pp. 20-42, 2019. [Online]. Available: https://doi.org/10.1016/j.red.2019.03.001.

[7] Gupta, Deepak et al. "Financial time series forecasting using twin support vector regression.", PloS one vol. 14,3e0211402.13Mar.2019,

[8] Y. Qu, P. Quan, M. Lei, and Y. Shi, "Review of bankruptcy prediction using machine learning and deep learning techniques," in Procedia Computer Science, vol. 162, pp. 895-899, 2019. [Online]. Available: https://doi.org/10.1016/j.procs.2019.12.065.

[9] D. Devi, S. Biswas, and B. Purkayastha, "A Review on Solution to Class Imbalance Problem: Undersampling Approaches," 2020 International Conference on Computational Performance Evaluation (ComPE) North-Eastern Hill University, Shillong, Meghalaya, India. Jul 2-4, 2020.

[10] A. Shrivastava, N. Kumar, K. Kumar, and S. Gupta, "Corporate Distress Prediction Using Random Forest and Tree Net for India," in Journal of Management and Science, vol. 1, no. 1, pp. 1-11, 2020. [Online]. Available: https://doi.org/10.26524/jms.2020.1.

[11] Y. Qu, P. Quan, M. Lei, and Y. Shi, "Review of bankruptcy prediction using machine learning and deep learning techniques," in Procedia Computer Science, vol. 162, pp. 895-899, 2019. [Online]. Available: https://doi.org/10.1016/j.procs.2019.12.065.

[12] Xinke Chong, "Hybrid PSO-SVM for Financial Early-Warning Model of Small and Medium-Sized Enterprises",Advances in Economics, Business and Management Research, volume,Proceedings of the 6th International Conference on Financial Innovation and Economic Development (ICFIED 2021).

[13] Le, Tuong, Mi Young Lee, Jun Ryeol Park, and Sung Wook Baik. 2018. "Oversampling Techniques for Bankruptcy Prediction: Novel Features from a Transaction Dataset" Symmetry 10, no. 4: 79. https://doi.org/10.3390/sym10040079.

[14] M.R. Shakil, T.A. Siddiqui, and S. Alam, "Feature Selection in Corporate Bankruptcy Prediction Using ML Techniques: A Systematic Literature Review," in V. Chakravarthy, V. Bhateja, W. Flores Fuentes, J. Anguera, and K.P. Vasavi (eds), Advances in Signal Processing, Embedded Systems and IoT, Lecture Notes in Electrical Engineering, vol. 992, Springer, Singapore, 2023. [Online]. Available: https://doi.org/10.1007/978-981-19-8865-3_32

[15] Hassan Raza, and Ishtiaq Ahmad. 2023. "Comparative Analysis of Machine Learning Models for Bankruptcy Prediction in the Context of Pakistani Companies" Risks 11, no. 10: 176. https://doi.org/10.3390/risks11100176.

[16] F. Mendes, J. Duarte, A. Vieira, and A. Gaspar-Cunha, "Feature Selection for Bankruptcy Prediction: A Multi-Objective Optimization Approach," in Book or Conference Name, 2010. [Online]. Available: https://doi.org/10.1007/978-3-642-11282-9_12.

[17] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," SIGKDD Explor. Newsl., vol. 6, no. 1, pp. 20–29, Jun. 2004. Available: https://doi.org/10.1145/1007730.1007735

[18] V. Kaul, B. Nemade, and V. Bharadi, “Next Generation Encryption using Security Enhancement Algorithms for End to End Data Transmission in 3G/4G Networks,” in Procedia Computer Science, vol. 79, pp. 1051-1059, 2016. [Online]. Available: https://doi.org/10.1016/j.procs.2016.03.133

[19] N. S. T. Sai, R. Patil, S. Sangle, and B. Nemade, "Truncated DCT and Decomposed DWT SVD Features for Image Retrieval," *Procedia Computer Science*, vol. 79, pp. 579-588, 2016. doi: 10.1016/j.procs.2016.03.073

[20] H. B. Kekre, V. A. Bharadi, V. I. Singh, V. Kaul, and B. Nemade, "Hybrid multimodal biometric recognition using Kekre’s wavelets, 1D transforms & Kekre’s vector quantization algorithms based feature extraction of face & iris," in *Proc. 2nd Int. Conf. and Workshop on Emerging Trends in Technology (ICWET)*, published by *Int. J. Comput. Appl. (IJCA)*, Mumbai, India, 2011.

[21] B. Nemade and V. A. Bharadi, "Adaptive automatic tracking, learning and detection of any real time object in the video stream," 2014 5th International Conference Confluence The Next Generation Information Technology Summit (Confluence), Noida, India, 2014, pp. 569-575, doi: 10.1109/CONFLUENCE.2014.6949039.