Research Journal of Computer Systems and Engineering (RJCSE)
Volume 4 Issue 1 (2023) | Pages: 30 – 38 | e-ISSN: 2230-8571; p-ISSN: 2230-8563
https://doi.org/10.52710/rjcse.60

Prediction of Corporate Bankruptcy using Machine Learning

Received: 09 March 2023; Revised: 12 May 2023; Accepted: 02 June 2023

Aakansha Vyas
Associate Consultant,
Performio, Melbourne
aakansha.vyas@performio.co

Mr. Prasad B. Chaudhari
Department of Artificial Intelligence & Data Science,
Vishwakarma Institute of Information Technology, Pune - India
prasad.chaudhari@viit.ac.in

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 SMOTEENN hybrid resampling technique on Polish Bankruptcy dataset. We next apply four classifiers to the altered data: Random Forest, Decision Tree, KNN, and AdaBoost, and assess the effectiveness of each model. As a result of our suggested approach, the Random Forest classifier had the highest accuracy of 89%, while the AdaBoost model as a whole outperformed it with a Recall of 73% and a Geometric Mean of 80%.

Keywords

Corporate bankruptcy, deep learning, SMOTTEN, AdaBoost, SVM

1. Introduction

One prevalent and significant source of fear among creditors, investors, employees, and management is bankruptcy. This phrase describes a corporation that lacks an operational source of funding to run its operations and is unable to pay back the obligations it owes creditors. Being in this precarious situation—where the company has reached a standstill and left its suppliers and customers high and dry—is not easy. These days, there are more competitors and unpredictability in the global market, which might push big businesses towards bankruptcy. One recent example is the British travel operator Thomas Cook, which abruptly filed for bankruptcy, throwing away millions of dollars' worth of investor funds and forcing 21,000 employees out of employment. As such, it is hard to overstate the harm done in terms of monetary loss. In addition to having a detrimental effect on a nation's economy, corporate bankruptcy spreads recessions [1]. A company's ability to make economic decisions is greatly impacted by its ability to predict bankruptcy. This is because all businesses, no matter how big or little, have an impact on investors, community members, industry participants, and policy officials. Because the corporate sector's performance has a significant impact on a nation's economy, it might occasionally be crucial for creditors to verify financial data and estimate the likelihood of bankruptcy. These kinds of choices have the potential to impact both the growth of an economy and the development of a corporation.

The recent global financial crisis that many nations have experienced is proof that the need for a quick and reliable bankruptcy model has arisen. The two primary methods used worldwide to forecast the likelihood of a company filing for bankruptcy are: Structural approach, which uses data mining techniques to forecast the chance of default by closely examining the firm's characteristics and interest rates. Authors in [2] have thoroughly examined statistical methods; also, [3] have investigated intelligent ways in addition to statistical techniques to estimate the bankruptcy rate. In this subject, various methodological techniques utilising generalised linear models and multidimensional analysis have already been investigated. However, the fundamental reason why linear models are insufficiently trustworthy and ineffective at determining the relationship between economic indicators is the increasing amount of data available. During the past few decades, artificial intelligence and machine learning have been used to predict company insolvency.

Both market-based and accounting-based variables in numerical format have been used in the majority of research in this area. It is commonly known that a company's bankruptcy status depends on a wide range of characteristics rather than just a few, making it challenging to anticipate with any degree of accuracy. Sales, purchases, assets, liabilities, EBIT, trade data, earnings per share, and so on are examples of numerical qualities. Even if it is understood how important it is to predict bankruptcy, in order to increase prediction performance and reduce computing time and cost, the appropriate qualities now need to be used. Second, there are a tonne of businesses that are not bankrupt, but there are also very few companies that file for bankruptcy—a small number of businesses that have the power to upend entire industries and the economy. It is difficult to deal with this kind of class imbalance since the model would be skewed in favour of the majority class, or non-bankrupt companies, if there was no adequate strategy in place and it was not carried out. These two conundrums must be properly addressed in order to create a prediction model that works.

Therefore, in comparison to state-of-the-art methods, can we increase the corporate bankruptcy prediction performance utilising a new combination of a feature extraction technique and a resampling strategy?

Using a collection of algorithms, we examine each variable's link to the dependent variable in the feature selection technique. We then pick and retain only those variables that have a strong association and are valuable for prediction. On the other hand, resampling techniques are employed to rectify dataset imbalances.

2. 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.

The majority of research focuses on publicly traded corporations, but a sizable portion of small and medium-sized businesses receive little to no attention. Since they make up a large portion of the Slovak economy, these enterprises have been identified in [7] as a part of an intriguing study. The selected dataset was divided into four sectors: manufacturing, retail, agricultural, and construction. It had 21 financial ratios. Three one-class classification techniques were used: least-squares anomaly detection, one-class support vector machines, and isolation forests [8]. According on the assessment year, the model based on one-class LSAD obtained prediction scores ranging from 76% to 91%.

In [9] authors suggest using the cluster-based boosting method C-boost in conjunction with the Instance Hardness threshold (IHT), which is typically employed to eliminate noisy instances. The imbalance within the classes is addressed using the C-boost algorithm. This scenario uses a Korean bankruptcy dataset, and a C-boost prediction model is constructed following resampling. The findings demonstrated that the suggested framework beat several of the current techniques, including the GM-Boost algorithm and a method that employed the SMOTEENN as the oversampling approach, with an AUC of 87%.

In addition, authors in [10] investigated the topic of bankruptcy and employed the Random Forest algorithm to forecast the bankruptcy rate. This study emphasises the significance of having the proper machine learning tools, an acceptable dataset for training the model, and other considerations like feature selection/imbalance concerns. After extracting the financial records for each of the 50 companies—50 bankrupt and the other 50 not—a genetic algorithm was used to identify the most significant aspects. In addition, the model is constructed using an ensemble of decision trees, or random forests. The results indicate that while the model was somewhat successful in making predictions, it is unreliable owing to the little amount of data and requires more research.

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.

In a study by authors in [10], the authors investigated the use of an ensemble-based model, such as Random forest, for predicting Turkish company bankruptcy, much like in earlier studies. There were twenty financial ratios in the sample. The main benefit of employing this method is that it can handle categorised, binary, and numerical data without the need for translation. The model's accuracy, which was the evaluation metric utilised, was 94% with all features and 96% with six features.

Building on earlier research, in [11] implemented a number of classifiers, including support vector machines, decision trees, and neural networks, using the KDD technique.

Accuracy, recall, and precision were the evaluation measures that were employed. The neural network-based model outperformed SVM and DT in terms of prediction accuracy, scoring 78%, according to the results. The authors experimented with several train-test splits, and interestingly, the outcomes differed greatly, demonstrating that the ratio of train-test split also affects prediction accuracy. Authors in [12] put out 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.

Authors in [12] conducted a thorough analysis of several popular data mining methods used to the bankruptcy prediction problem. This study took into account a variety of methodologies, including machine learning and statistical approaches. Additionally, meta-heuristic optimization-based machine learning techniques for enhancing prediction accuracy have been covered. Accuracy, sensitivity, specificity, and precision were the evaluation parameters taken into consideration. With an accuracy of 95%, specificity of 95%, and precision of 94%, SVM-PSO demonstrated the highest metrics performance among the data mining techniques utilising the Apache Mahout tool.

In a study published in [13], 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.

Authors in [14] conducted a similar analysis using financial ratios as attributes. The features that have the greatest influence on the bankruptcy prediction problem were chosen by the authors using the Random Forest model and the Genetic algorithm feature selection technique. In order to categorise the financial indicators as influential or non-influential, the non-linear relationship between them must first be analysed. Five models were considered to forecast bankruptcy, in order to determine the most influential features. Bootstrap Aggregation has been applied to decision trees in order to overcome the issue of large variation. The study's shortcoming is that an extremely small dataset—just 14 businesses—was selected using an 80:20 train-test split.

Although the model can accurately predict bankruptcy in certain situations, there is no guarantee of a high prediction rate because of the small size of the dataset used. Authors in [15] conducted a comprehensive analysis of all the methods that have been used to date in this field and released their review paper lately. The main takeaways from this were, firstly, that there is increasing interest in the subject of bankruptcy prediction, particularly in the wake of the 2008 global financial crisis. Second, because no eminent scholars have collaborated in this field in the previous few decades, there is very little co-authorship in this field. The author emphasises the value of cooperation and how influencer cooperation may advance this field's study [16]. Thirdly, Logistic Regression and Neural Network models are the most commonly utilised in this field. However, because of recent developments in the fields of computer science and artificial intelligence, one can see the application of creative solutions to this issue.

A study in [17] [18] and colleagues did a literature review with the aim of reviewing prior research on the prediction of bankruptcy in Poland. The objective is to determine how far research has come in forecasting bankruptcy in Poland and how this compares to worldwide trends. The most important findings are that this topic has been studied across a wide range of industries, including logistics, meat, manufacturing, trade, transportation, construction, farms, etc. The research started in the 1990s, which was a little later than studies done on the United States region. This was due to the fact that the academics' initial interest in the subject only developed following the first bankruptcies that occurred after 1990.

Altman designed one of the most extensively utilised and well-liked models. However, there is still disagreement among academics regarding the value and effectiveness of applying the instruments and methods covered in this study. Therefore, the author recommends that future research concentrate on the significance of such instruments and techniques that could help to enhance the prediction performance. 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. 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.

3. Methodologies Used

Getting the researcher's entire goal in order is the first priority before starting any project. The main goal of this research is to forecast, 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 investors/creditors and that can warn them about any impending bankruptcy threat to a company.

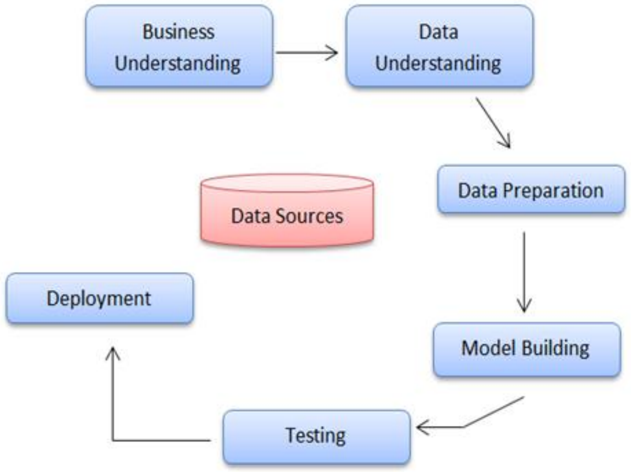

Figure 1: Workflow of the Methodology

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 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.

B Data Pre-Processing

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-colinearity: 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.

C 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.

Table 1: Attribute Features

Feature	Description
Attr21	Sales
Attr23	Profit
Attr25	Expense
Attr27	Profit acquired on sales
Attr29	Liabilities
Attr49	Costs

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. Authors in [20] evaluated the applicability of this technique in the field of bioinformatics, and found that it is rarely utilised to predict corporate bankruptcy. 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.

D Data Balance

Based on their ability to improve node purity, tree-based strategies used in random forests are ranked. This is known as mean decrease in impurity, or gini impurity. The maximum reduction in node purity occurs near the top of the tree, whereas the minimum decline occurs near the base. Thus, a subset of important features is generated in this fashion by first identifying a certain node and then trimming away from it. Either over-sampling the minority class, under-sampling the majority class, or using a hybrid strategy that combines the two is known as a sampling-based technique. 5. The present investigation will employ the infrequently employed hybrid approach, SMOTEENN, to address the issue of class imbalance. This method works well for the problem since we have a very unbalanced dataset and need to expand the minority class while, at the same time, under-sampling the majority class to some degree because there are a lot of cases in the majority class.

E Algorithms Used

This step in the machine learning process is seen to be crucial and important. The suggested models are to be put into 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.

- KNN: The KNN algorithm operates under the presumption that nearby instances of comparable cases exist. Using distance calculation methods, the most widely used of which is the Euclidean distance, it determines the distance between the instances. If the selected K value can both lower the amount of errors and preserve the algorithm's capacity to produce correct predictions, then it is the right choice. We will assess this technique on our dataset because we haven't seen it used much in the literature.
- 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 below. 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 hybrid SMOTEENN approach. The final step involves feeding the processed data to four distinct classifiers and assessing each one's performance using the testing data.

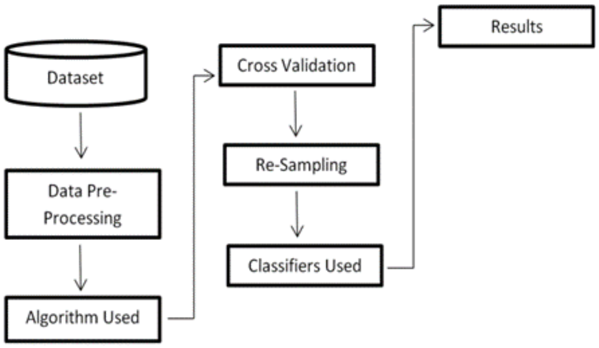

Figure 2: Architectural Flow.

4. 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 SelectFromModel 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 SMOTEENN approach to balance the classes in order to prevent over-fitting and create a dependable model. We employed a hybrid strategy, which can oversample the minority class and under-sample the majority class, because of the significant imbalance. For resampling, the SMOTEENN 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 AdaBoost, K Nearest Neighbours, Decision Tree, and Random Forest. These models can be found in the Python sk-learn library as several packages.

5. Results

Case Study 1:

After selecting features, each classifier's base model is constructed using the unbalanced data. Table 2 shows the performance of each classifier individually. The data indicates that while the classifiers' accuracy is high, their recall is low, indicating that they are unable to accurately forecast cases of bankruptcy. This demonstrates how the substantial class imbalance in our model has caused bias and overfitting. KNN had the lowest recall of 14% whereas decision tree did the best, outperforming the other classifiers with a recall of 56% although still being extremely poor. Since all classifiers essentially equally score on the other two metrics—specificity and accuracy—the decision tree's general manager (GM) scores the highest at 73% and the KNN's lowest at 36%. We further implemented SMOTEENN on the classifiers, maintaining the same feature set, and assessed their performance in order to address the low recall issue and enhance performance.

Case Study 2:

We have now run the identical experiment on the resampled dataset using SMOTEENN. Table 3 makes it evident that the model has improved, as evidenced by the acquired recall values. With a recall value of 73%, the AdaBoost classifier performed better than the others. This is not a bad number considering that the programme can accurately forecast 73% of all bankrupt instances. Additionally, KNN had the lowest recall of 60%, which is still greater than the maximum number from the last study. Additionally, the GM values have marginally improved as all models' specificity has somewhat decreased. This indicates that the model has improved its ability to anticipate when a company would go bankrupt. With a GMean of 80%, AdaBoost performed the best among the classifiers, while KNN got the lowest GM value, at 67%. We will also talk about how SMOTEENN affects certain classifiers for a given dataset and collection of features.

Table 2: Results

Model	Accuracy	Sensitivity	Specificity	Mean
Random forest	90%	70%	89%	75%
Decision tree	87%	64%	89%	76%
KNN	80%	62%	81%	68%
AdaBoost	89%	74%	90%	81%

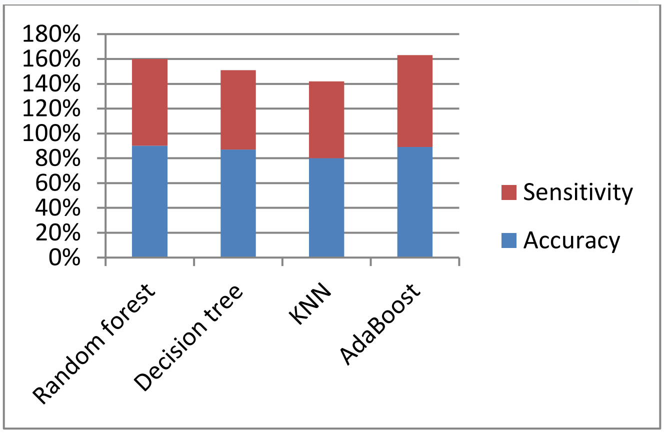

Figure 3: Recall

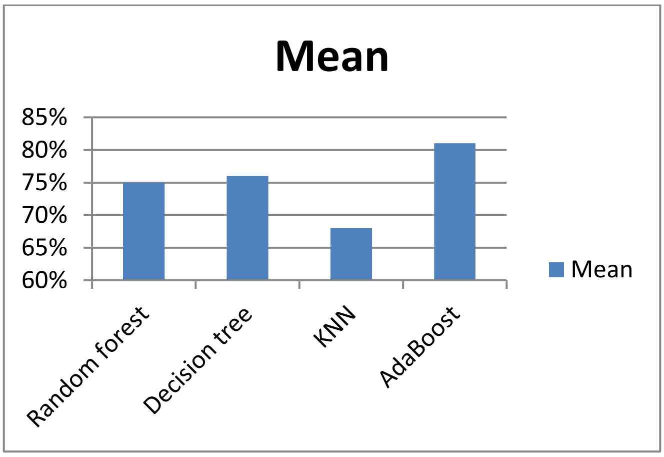

Figure 4: G-Mean

6. Conclusions and Future Works

As evidenced by the earlier research, there has been a lot of interest in the field of bankruptcy prediction over the past few decades, and various strategies have been used in an effort to achieve the best possible forecast performance. As previously mentioned, the difficult aspect is identifying the best financial characteristics that primarily contribute to a company's bankruptcy. An additional challenge in this research is the issue of class imbalance. In this study, we constructed four models using a new mix of a resampling technique and a feature selection technique. In terms of prediction accuracy, our methods of Random Forest feature selection and SMOTEENN with Random Forest classifier outperform the AdaBoost classifier, and comparable methods with AdaBoost classifier perform better than other classifiers in terms of Geometric Mean and Recall values. These methods can be investigated further using a different set of financial attributes and other classifiers. We restricted our investigation in this study to bankruptcy data from Poland. One can investigate bankruptcy data from any other region in the future. Additionally, KNN classifier shown a notable increase in performance while using SMOTEENN. This combination can be investigated further in additional classification issues of a similar nature.

References

[1] Alrasheed, D., Che, D. and Stroudsburg, E. (2017). Improving Bankruptcy Prediction Using Oversampling and Feature Selection Techniques, pp. 440–446

[2] Ayyadevara, V. K. and Ayyadevara, V. K. (2018). Random Forest, Pro Machine Learning Algorithms (Iciccs): 105–116

[3] Balcaen, S. and Ooghe, H. (2006). 35 years of studies on business failure : an overview of the classic statistical methodologies and their related problems, 38: 63–93

[4] Barboza, F., Kimura, H. and Altman, E. (2017). Machine learning models and bankruptcy prediction, Expert Systems with Applications 83: 405–417. URL: http://dx.doi.org/10.1016/j.eswa.2017.04.006

[5] Behr, A. and Weinblat, J. (2017). Default prediction using balance-sheet data : a comparison of models, 18(5): 523–540

[6] Bernanke, B. B. E. N. S. (2015). Bankruptcy , Liquidity , and Recession, 71(2): 155–159.

[7] Cunningham, P. and Delany, S. J. (2014). k-Nearest neighbour classifiers k -Nearest Neighbour Classifiers, (April 2007)

[8] Devi, S. S. and Radhika, Y. (2018). A Survey on Machine Learning and Statistical Techniques in Bankruptcy Prediction, 8(2)

[9] Ding, Y. (2008). Forecasting financial condition of Chinese listed companies based on support vector machine, 34: 3081–3089

[10] Elrahman, S. M. A. and Abraham, A. (2013). A Review of Class Imbalance Problem, 1: 332–340

[11] Kotsiantis, S. B., Kanellopoulos, D. and Pintelas, P. E. (2006). Data Preprocessing for Supervised Leaning, 1(1): 111–117

[12] Kumar, P. R. and Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review, 180: 1–28

[13] Landgrebe, I. (1991). A Survey of Decision Wee Classifier Methodology, 21(3)

[14] Le, T. (2018). SS symmetry Oversampling Techniques for Bankruptcy Prediction : Novel Features from a Transaction Dataset

[15] Le, T., Vo, M. T., Vo, B., Lee, M. Y. and Baik, S. W. (2019). A Hybrid Approach Using Oversampling Technique and Cost-Sensitive Learning for Bankruptcy Prediction, 2019

[16] Linnga, O. (2018). Bankruptcy prediction based on financial ratios using Jordan Recurrent Neural Networks : a case study in Polish companies Bankruptcy prediction based on financial ratios using Jordan Recurrent Neural Networks : a case study in Polish companies

[17] Liu, H. and Motoda, H. (1998). Feature Extraction, Construction and Selection: A Data Mining Perspective, Kluwer Academic Publishers, Norwell, MA, USA

[18] Monard, M. C. (2017). A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data, 6(1): 20–29.