National College of Ireland

Corporate Bankruptcy Prediction using Machine Learning Techniques

MSc Research Project
Data Analytics Jan 2019-20

Shantanu Deshpande
Student ID: x18125514

School of Computing
National College of Ireland

Supervisor: Dr. Vladimir Milosavljevic

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Shantanu Deshpande
Student ID:	x18125514
Programme:	MSc. Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Dr. Vladimir Milosavljevic
Submission Due Date:	12/12/2019
Project Title:	Corporate Bankruptcy Prediction using Machine Learning Techniques
Word Count:	7657
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard speci ed in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	28th January 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not su cient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator o ce must be placed into the assignment box located outside the o ce.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Corporate Bankruptcy Prediction using Machine Learning Techniques

Shantanu Deshpande
x18125514

Abstract

Corporate bankruptcy has been a cause of concern for business stakeholders including the management, investors etc. since last few decades and has been of interest among the researchers worldwide. It is not enough to rely on a single predictive model due to the vast number of factors responsible for bankruptcy and thus the challenge lies in identifying only the key factors that are more responsible. Another major hurdle is the high class imbalance within the data which hinders the model's performance. Though there are numerous techniques that have been previously explored which include Decision Trees, SVM, Neural Networks etc. with di erent pre-processing strategies, we further take this research forward by using a novel combination of a Random Forest feature selection technique along with SMOTEENN hybrid resampling technique on Polish Bankruptcy dataset. Further we implement four classi ers, Random Forest, Decision Tree, KNN and AdaBoost on the transformed data and evaluate each model's performance. The results show that under our proposed strategy, Random forest classi er showed highest accuracy of 89% whereas overall AdaBoost model performed better with Recall 73% and Geometric Mean as 80%.

1 Introduction

Among the creditors, investors, employees and management there is one common and major cause of concern i.e. Bankruptcy. It is a term used to signify a company that has no operational source of money to operate the business and is in no position to repay the debts it owes to its creditors. It is a di cult situation to be in as the company comes down to a stand-still state and leaves the employees and suppliers/customers in a high & dry state. Increased competitors and uncertainty in the global economy these days can be a cause to drive large organizations towards bankruptcy. The recent example being that of the British travel company, Thomas Cook which suddenly declared bankruptcy and left 21,000 people out of work and million dollars of investor money down the drains. Thus, it is impossible to overstate the damage caused in terms of the nancial loss. Corporate bankruptcy propagates recession and also negatively impacts the economy of a country (Bernanke; 2015). Prediction of bankruptcy is of high importance in terms of the ability of economic decision-making as all types of companies, be it large or small, concerns investors, local people within the community, industry participants and thus it in uences the policy makers as well as the global economy. The economy of a country is largely dependent on the performance of the corporate sector, hence at times it becomes very important for a creditor to validate the nancial indicators and thereby predict the

probability of bankruptcy. Such type of decisions can not only a ect the development of a company but also the development of a country's economy.

The global nancial crisis faced by several countries worldwide in the recent years is an indication that a timely and credible bankruptcy model is required. The two main approaches followed globally to predict the possibility of a company's bankruptcy are: Structural approach wherein the attributes of the rm and also the interest rates are carefully examined and thus the probability of default is predicted with the use of data mining methods. The statistical methods have been explored in detail by Balcaen and Ooghe (2006) and in addition to this Kumar and Ravi (2007) have explored intelligent techniques as well in conjunction with statistical techniques to predict the rate of bankruptcy. Different methodological approaches using multidimensional analysis and generalized linear models have already been explored in this eld. But mainly because of the growing abundance of data, linear models are not reliable enough and also not e ective in terms of identifying the relationship among the economic indicators. In the last couple of decades we have seen the use of machine learning and arti cial intelligence for determining the corporate bankruptcy prediction.

Most of the studies in this eld have made use of accounting-based variables as well as market-based variables that are in numeric format. It is well understood that a company's bankruptcy status is not dependent on a few factors but on a diverse set of attributes which makes it di cult to accurately predict the odds. The numerical attributes may consist of the sales, purchases, assets, liabilities, EBIT, trading data, earning per share etc. Despite knowing the importance of predicting bankruptcy, utilizing the right attributes to improve the prediction performance and decrease computational time and cost is required at present. Secondly, the presence of non-bankrupt companies is tremendous and on the other hand only a handful of companies go bankrupt although these handful of companies are enough to disrupt several industries and the economy as a whole. Dealing with this kind of class imbalance is a challenging part as without proper plan and execution, the model would be biased towards the majority class i.e. non-bankrupt companies. In order to build an e ective prediction model it is important to tackle these two dilemmas appropriately.

Hence, the research question for this study will be - *Can we improve the corporate bankruptcy prediction performance using a novel combination of a feature extraction technique and a resampling technique as compared to the state-of-the art techniques?*

In this study, we have focused on the pre-processing and data transformation phase which is infact one of the most important phases of a data mining process. The aim is to use an appropriate feature selection technique in conjunction with a resampling technique and subsequently the transformed data will then be fed to few machine learning models and have their performances compared with that in the previous literature. In feature selection technique, we study the relationship of each variable with the dependent variable using a set of algorithms and identify and keep only those that show strong relationship and are useful for prediction. Whereas resampling techniques are used in order to balance an imbalanced dataset.

2 Related Work

One of the main challenging part of dealing with the bankruptcy prediction problem is the problem of class imbalance. As we know the fact that there are tens and thousands

of non-bankrupt companies worldwide that may be stable and performing well but on the other hand there are only a handful of bankrupt companies. Even though their count is small, these bankrupt companies can disrupt the country as a whole in terms of the economic loss and also create nancial crisis among the investors and lenders. To address the issue of class imbalance in machine learning models, there exists several techniques that have also been explored by the researchers in the previous studies. The literature in this section will give us a general understanding of how well these techniques are suitable in the bankruptcy prediction problem.

2.1 Dealing with Class Imbalance

In a study conducted by Le et al. (2019) the authors have used two techniques in conjunction to address the issue of class imbalance. The authors have developed a hybrid approach using cost-sensitive learning and oversampling technique on the Korean bankruptcy dataset. Firstly, oversampling module is used with an optimal balancing ratio to gure an ideal performance on the validation set. Secondly, CBoost algorithm is used as a cost-sensitive learning model for bankruptcy prediction. There were 307 bankrupted rms and 120048 non-bankrupt ms in the dataset thus making a balancing ratio of 0.0026. Oversampling technique adds synthetic data to the minority class thus improving prediction accuracy although it comes with few disadvantages. As it makes exact copies of existing minority samples, the model is likely to be over t. Also it increases the number of training examples and thereby increasing the training time along with the amount of memory required for holding the training set. For oversampling, SMOTE-ENN technique has been used that generates synthetic minority samples that are based on the feature similarities between the minority class. Further, CBoost algorithm is used to make clusters of the majority class. This approach was then applied on following methods: Bagging, AdaBoost, Random Forest, Multilayer Perceptron. The evaluation metrics used were AUC and Gmean. The use of oversampling did improve the results however this study lacked the use of feature selection methods.

Alrasheed et al. (2017) also explored the use of oversampling techniques to imrove bankruptcy prediction accuracy. The minority class in this case was randomly duplicated to a certain percent of the majority class. Three di erent feature selection techniques were used for chosing the best features, Mutual Information, Random Forest Genetic algortithm. The outcome of above techniques were fed to machine learning algorithms like Neural Network, decision trees, logistic regression, K-nearest neighbour, support vector machine and random forest and these models were evaluated based on ROC AUC, F1 score, Precision and Recall. Overall the results showed that oversampling does improve the prediction performance.

In another study by Le (2018) ve oversampling techniques have been studied on the bankruptcy prediction problem namely, SMOTE, Adaptive synthetic sampling, Borderline-SMOTE, SMOTE+TOMEK and SMOTE+ENN. Korean dataset containing rms nancial ratios was chosen for the study and the following models were implemented after oversampling, Random Forest, Decision Tree, Multi-layer Perceptron and SVM. The evaluation metric used was AUC and results showed that SMOTE+ENN on Random Forest model achieved the best AUC of 84%.

Most of the studies are focssed on companies that are listed on stock market but there's a large chunk of small-and-medium scale companies that are not much studied . In an interesting study by Zori et al. (2019), the authors have tagetted such companies of Slovak

region as they represent a signi cant part of an economy. The chosen dataset had 21 nancial ratios and been bifurcated based on their sector- manufacture, retail, agriculture, construction. Three one-class classi cation methods were adopted: an isolation forest, a one-class SVM and a least-squares approach to anomally detection. The results showed that model based on one-class LSAD achieved prediction score between 76% to 91% based on the evaluation year. Resampling techniques and cost based learning for dealing with imbalanced dataset has not been used in this study and the author mentioned them as future research areas. Also, the use of feature selection techniques has been strongly put forward by the author as an interesting area for future work.

Further in a study performed by Faris et al. (2019) the problem of highly imbalanced class distribution is tackled by using hybrid approach that is a combination of SMOTE and ensemble methods. Not only resampling techniques but also the authors have used ve di erent feature selection techniques to gure out the most important attributes for bankruptcy prediction on a dataset that consisted of Spanish companies . Two variations of SMOTE, ADASYN Borderline-SMOTE has been used and the feature selection techniques used were, Correlation feature selection (CFS), ReliefF, Information Gain (IG), Classi er attribute evaluator (CAE) and lastly Correlation attribute evaluator. The above preprocessed dataset was then thouroughly used to build model using di erent algorithms like KNN, decision tree, Naive bayes, arti cial neural network and ensemble methods like Boosting, bagging, Random forest, Rotation forest, DECORATE. The evaluation metrics used were Accuracy, GMean and AUC. Results showed that SMOTE with AdaBoost ensemble achieved highest accuracy of 98% but with a type2 error of 45% and Gmean 0.73. The author suggests to explore cost-sensitive learning approachs in the future. Also data from few other nancial sources can be explored in the future.

The study conducted by Le (2018) proposes the use of cluster based boosting algorithm known Cboost along with Instance Hardness threshold (IHT) that is normally used for removing the noise instances. The Cboost algorithm is used for dealing with the imbalance within the classes. A Korean bankruptcy dataset is used in this case and Cboost prediction model is built after resampling. The results showed that the proposed framework achieved an AUC of 87% and outperformed some of the existing methods such as GMBoost algortithm (Kim et al.; 2014) and also a method that used the SMOTEENN (Le; 2018) as the oversampling method.

Further the issue of bankruptcy has been studied by Veganzones and Severin (2018) where the authors have used Random Forest algorithm to predict the rate of bankruptcy. The importance of having an appropriate dataset for training the model, the right machine learning tools and other factors like feature selection/imbalance issues are highlighted in this study. The nancial reports have been individually extracted for 50 bankrupt and 50 non-bankrupt companies and Genetic algorithm has been used for selecting the most important features. Further an ensemble of Decision trees i.e. Random forest is used for building the model and the results showed that the model was successful to some extent regarding the predictions however due to the limited nature of data the model is not reliable and need to be studied further.

A Geometric mean based boosting algorithm has been used in a study performed by Kim et al. (2014) to resolve the data imbalance problem. This algorithm considers both the minority and majority class as during the calculation of accuracy and error rate it uses the geometric mean of the two classes. The results are compared with cost-sensitive boosting and AdaBoost. The dataset used in this study was gathered from Korean commercial bank. The number of bankrupt companies were 500 whereas non-bankrupt

were 2500 with 30 nancial ratios considered. To verify the performance of GMBoost algorithm, two di erent data samples were constructed with ve sample groups (1:5,1:3, 1:20,1:1,1:10) and then experiments using CostBoost, GMBoost and AdaBoost have been performed on those imbalanced datasets. For the second stage, in order to generate new bankruptcy data, SMOTE algorithm has been used and the newly generated sample sets are applied on SMOTE-SMBoost, SMOTE-CostBoost and SMOTE-Boost. The results showed promising performance by GMBoost with high prediction performance.

2.2 Supervised Machine Learning techniques for bankruptcy prediction

The study conducted by Barboza et al. (2017) takes the research on this topic to a next level by comparing machine learning models with the statistical models and thereby assess which methodological approach is better. A balanced dataset was chosen and the predictor variables were liquidity, pro tability, leverage, productivity and asset turnover. The techniques that were implemented included bagging, boosting, Random forest, ANN, SVM with two kernels (linear and radial basis), logistic regression and MDA. Results showed that traditional statistical models underperformed than machine learning models. A major limitation of this study is that there was no use of any feature selection technique which is a standard approach nowadays.

Similar to the previous studies, in a study by Rustam and Saragih (2018) the authors have also explored the use of ensemble based model ie Random forest for predicting bankruptcy for Turkish companies. The dataset consisted of 20 nancial ratios. The primary advantage of using this technique is that it is able to handle numerical variables, binary variables and categorical variables without any requirement of scaling. The evaluation metric used was accuracy and the model achieved accuracy of 94% with all features and 96% with 6 features.

Extending the previous studies, Geng et al. (2014) have used the KDD methodology to implement several classi ers namely, neural network, decision tree and support vector machine. The dataset used for this study contained 31 nancial indicators dated 3-4 years prior to the companies receiving bankruptcy notice. Numerous models were built based on statistical probabilistic theory, namely, CR tree, DT, NN, C5.0, logit , bayes probability, SVM DA. Models built on SVM, C5.0, NN outperformed other techniques and were further used for data mining process. The evaluation metrics used were Accuracy, Recall and Precision. Results showed that the model built on neural network showed higher prediction accuracy than SVM and DT i.e. 78%. The authors used various train-test splits and notably the results varied signi cantly thus proving that prediction accuracy is also dependent on the ratio of train-test split.

A prediction model that is based on SVM has been proposed by Ding (2008). The author has made use of grid-search technique along with 10fold CV for identifying best parameter value. A-share market data of two Chinese cities has been chosen containing the nancial ratios of 250 rms. The results of RBF SVM were better than MDA and BPNN.

A detailed survey of some of the common data mining techniques used on bankruptcy prediction problem have been performed by Devi and Radhika (2018). The techniques considered in this study ranged from statistical methods to machine learning techniques. Also machine learning algorithms based on meta-heuristic optimization that are used to improve prediction accuracy have also been discussed. The evaluation metrics considered

were accuracy, sensitivity, speci city and precision. Apache Mahout tool has been used for the data mining process and from this it was observed that SVM-PSO achieved the highest metrics performance with accuracy of 95%, speci city of 95% and precision of 94%.

In a study performed by Behr and Weinblat (2017) the authors have performed a comparison of three data mining techniques, logit, decision trees & random forest on balance sheet data of several countries, namely, Italy, Germany, France, Britain, Portugal and Spain covering 446,464 rm statements. No speci c resampling technique was performed by the author and the metrics used for evaluation were Accuracy, speci city, sensitivity and the precision. The results showed that logit model outperformed the decision tree model and the random forest model outperformed both the decision tree & the logit model.

With the use of nancial ratios as attributes, a similar study has been carried out by Ayyadevara and Ayyadevara (2018). The authors have applied Random forest model and used a feature selection technique called Genetic algorithm for chosing the features that most in uence the bankruptcy prediction problem. The process includes analysing the non-linear relationship among the nancial indicators and further to this classifying them as in uential ratio or non-in uential ratio. To identify the most in uencing features, ve bankruptcy prediction models were taken into consideration. To address the problem of high variance in decision tree, Bootstrap Aggregation has been used. The drawback of this study is that a very limited dataset of only 14 companies companies was chosen with 80:20 train-test split. Though the model is successful in predicting bankruptcy in some cases, however due to the limited nature of the chosen dataset, there is no guarentee of a strong prediction rate.

In a review paper published recently by Shi and Li (2019), the authors have done a thorough study of all the existing techniques that have been deployed in this area. The key observations from this were, rstly, the topic of bankruptcy prediction is of growing interest especially after the global nancial crisis in 2008. Secondly, there is hardly any co-authorship in this area as no top researchers worked in collaboration in last few decades. The author highlights the importance of collaboration and how it can improve the research in this eld if in uencers collaborate. Thirdly, the most frequently used models in this area are Neural Network and Logistic regression (Logit). Although in the recent years one can observe the use of innovative methods to tackle this problem due to the advancements in the eld of computer science and arti cial intelligence.

2.3 Existing studies on corporate bankruptcy prediction in Poland

A study conducted recently by Linnga (2018) have used similar dataset as used by Naidu and Govinda (2018) i.e. bankrupt/non-bankrupt status and nancial statement of Polish companies. The authors have used a type of Arti cial neural network (ANN) known as Jordan Recurrent Neural Network on the nancial ratios. The activation function used is a logistic sigmoid function whereas that for the output signal is linear function. The model achieved the best prediction performance when trained with 5 neurons in the hidden layer and the accuracy in this case was 81%.

The study by Naidu and Govinda (2018) also shows the use of bankruptcy data of Polish companies which consists of 10008 non-bankrupt and 495 banrupt companies and applied arti cial neural network and random forest. These techniques have also been used

by Geng et al. (2014). The activation function used in the arti cial neural network model was sigmoid and bootstrap aggregating technique was used for training the model. From the feature size of 64, only three features were selected: EBIT, liquidity and solvency. Random Forest technique was also implemented however it showed a high degree of error, 5%.

In a literature study conducted by Prusak (2019) the authors attempt to review the previous studies on bankruptcy prediction in Poland. The aim is to understand the level of advancement in the research conducted to predict bankruptcy in Poland and where does this stand in terms of the global trends. the key observations have been that a wide number of sectors like logistics, meat, manufacturing, trade, transport, construction, farms etc. have been explored on this topic and particularly the research started during the period of 1990's a bit of delayed as compared to the studies conducted on United States region. This was because the initial interest among the researchers arose only after the initial bankruptcies that started taking place after 1990. One of the most popular and widely used model was developed by Altman. Although there is still doubt among researchers on the impact and usefulness of using the tools and techniques that are described in this study. Hence as a part of future work, the author suggests to focus on the importance on such tools and methods that might be bene cial in improving the predictive performance.

After having gone through the recent literature work in this area, it has been found that numerous techniques have been studied on this topic and this topic has been of speci c interest to the researchers globally. The studies evolved from statistical methods to the recent use of machine learning models and deep learning techniques to improve performance of the model however the challenging part is the multitude of factors that can lead a company to bankruptcy and the distribution of the data. Thus, in the current work, we will explore the topic beyond existing literature and use feature selection technique along with resampling technique on various machine learning models.

3 Methodology

The two most common approaches for planning a data mining project that are followed by industry experts are CRISP-DM (Cross Industry Process for Data Mining) and KDD (Knowledge Discovery in Databases). Another modelling approach known as SEMMA is mostly used by SAS enterprise miner. CRISP-DM covers complete modelling process in six steps whereas KDD approach involves nine steps & also based on Sha que and Campus (2014) CRISP-DM is more complete than SEMMA hence in this study, we will be using CRISP-DM as our research methodology.

The architecture and the process ow is shown in gure 1

3.1 Business Understanding

The foremost objective before beginning any project is to gather the complete objective of the project. The primary aim of this research is to predict the possibility whether a company is on the verge of bankruptcy or not based on various nancial indicators of the company like the pro t, sales, assets etc. Thus the objective of this work is to build a predictive bankruptcy model that will be useful for investors/creditors, management and employees and which can alert them about any potential bankruptcy threat over a

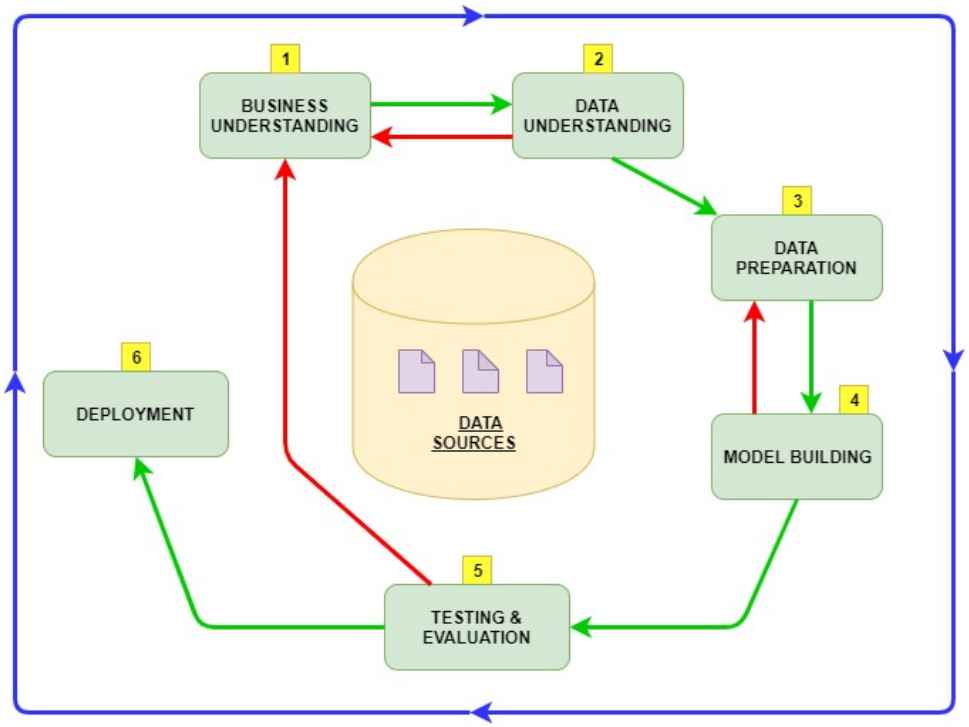

Figure 1: CRISP-DM Methodology

company in the near future. As observed in the previous literature work, the main constraint surrounding this problem is that there are thousands of non-bankrupt companies globally and only a handful of bankrupt companies hence it is di cult for the model to learn & train itself based on the limited information about the bankrupt cases.

Secondly, there are multitude of factors that are responsible for bankruptcy and not just few. Another objective of this study is to identify the most important nancial attributes that drive a company towards bankruptcy. This research would be a positive step ahead in the eld of bankruptcy prediction and will be bene cial for a wide number of people globally.

3.2 Data Understanding

The next phase in the CRISP-DM process is understanding the data. Before building a prediction model having a proper understanding of its components is critical without which it is almost possible to build a reliable model. The required data may be available across multiple sources however fetching data from few of them may be unethical. Also the data available on few sources may be unreliable therefore extraction of data from a reliable source and that too ethically requires considerable amount of time.

For proceeding with this research, we require the nancial ratios of companies of a particular region over a certain period of time which includes bankrupt and non-bankrupt cases. Thus, the data source and dataset that has been used is described below:

UCI Machine Learning Repository: The chosen dataset consists of Polish companies and is hosted by UCI Machine Learning Repository for free. The reason for choosing Polish companies is because Poland is the sixth largest economy in the European Union

and as per McKinsey's [1] report in 2015, Poland will become Europe's new growth engine by 2025. The dataset consists of 5910 instances i.e. nancial statements within which 410 instances represented the bankrupt companies and 5500 represented the non-bankrupt companies. Depending on the forecasting period, ve di erent classi cation cases have been distinguished. We will be using the data of 5th year i.e. it contains the nancial statements from the fth year of the forecasting period and the relevant class label that indicates the status of bankruptcy after one year. There are in all 64 numerical attributes which are ratios derived from the net pro t, liabilities, working capital, EBIT etc. The complete dataset can be downloaded from URL[2]

3.2.1 Data Exploration

One of the quickest method to explore and understand the data is with the help of visualizations. The results of exploring data visually can be powerful in terms of comprehending the structure of the data, the way the values are scattered, and the presence of any correlation within the dataset. Therefore, the ndings of Polish bankruptcy dataset are as below:

Identifying the datatype of all the variables [Fig 2]:

Most of the variables are in object datatype and need to be converted to oat datatype as part of data pre-processing.

```
Attr1      object
Attr2      object
Attr3      object
Attr4      object
Attr5      object
           ...
Attr61     object
Attr62    float64
Attr63     object
Attr64     object
class       int64
Length: 65, dtype: object
```

Figure 2: Data type of variables

Checking the presence of null values in the dataset [Fig. 3]:

From the gure, it is observed that only a single column has more than 40% missing values whereas the remaining columns have very few missing values which can be replaced by the mean value of the column.

[1] https://www.mckinsey.com/~/media/mckinsey/business%20functions/economic%20studies%20temp/our%20insights/how%20poland%20can%20become%20a%20european%20growth%20engine\/poland%202025_full_report.ashx

[2] https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data

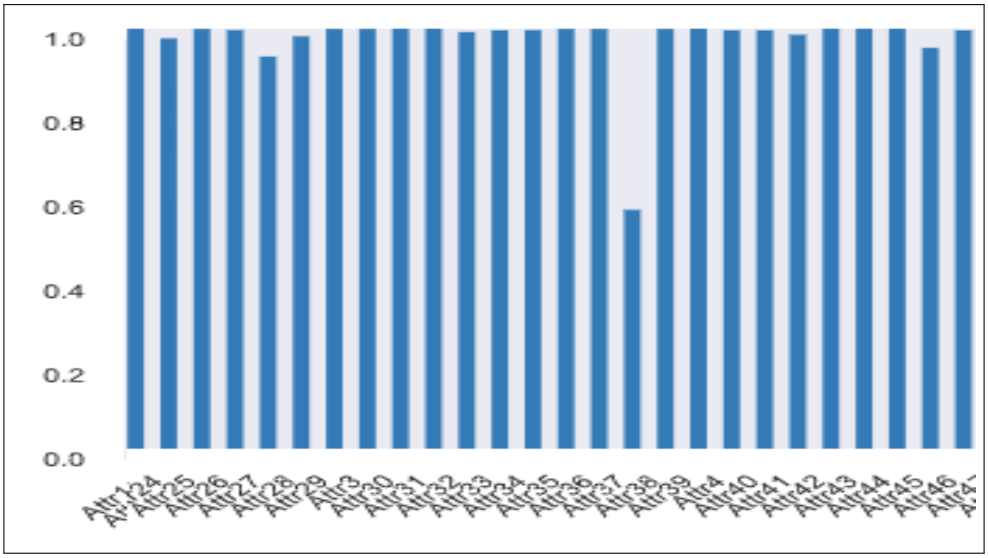

Figure 3: Presence of null values

Plotting bar-graph to explore the distribution of bankrupt and non-bankrupt cases [Fig. 4]:

A high level of class imbalance can be observed as we have 5500 non-bankrupt rms and only 410 bankrupt rms in the dataset. This issue needs to be addressed before we implement our models.

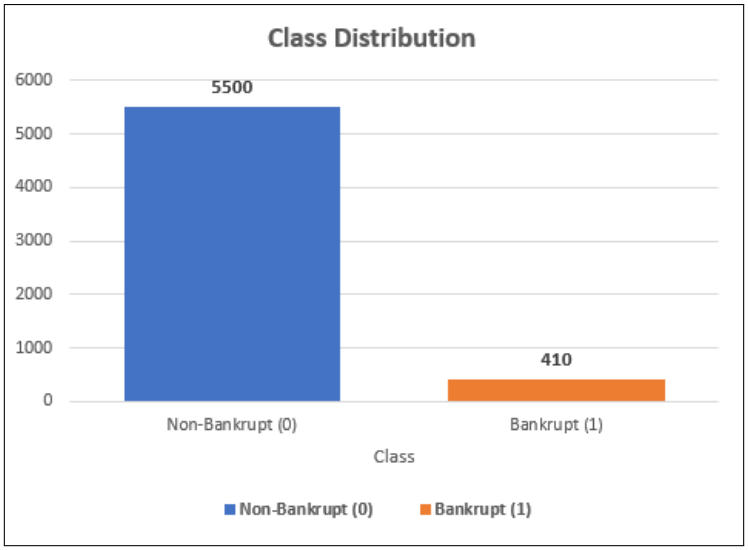

Figure 4: Class Distribution

The correlation within the variables is observed using below gure [Fig. 5]:

This graph shows the multicollinearity and correlation between our dependant variable and the independent variables. The 'class' variable in this case is our dependent variable which denotes the bankruptcy status of the company.

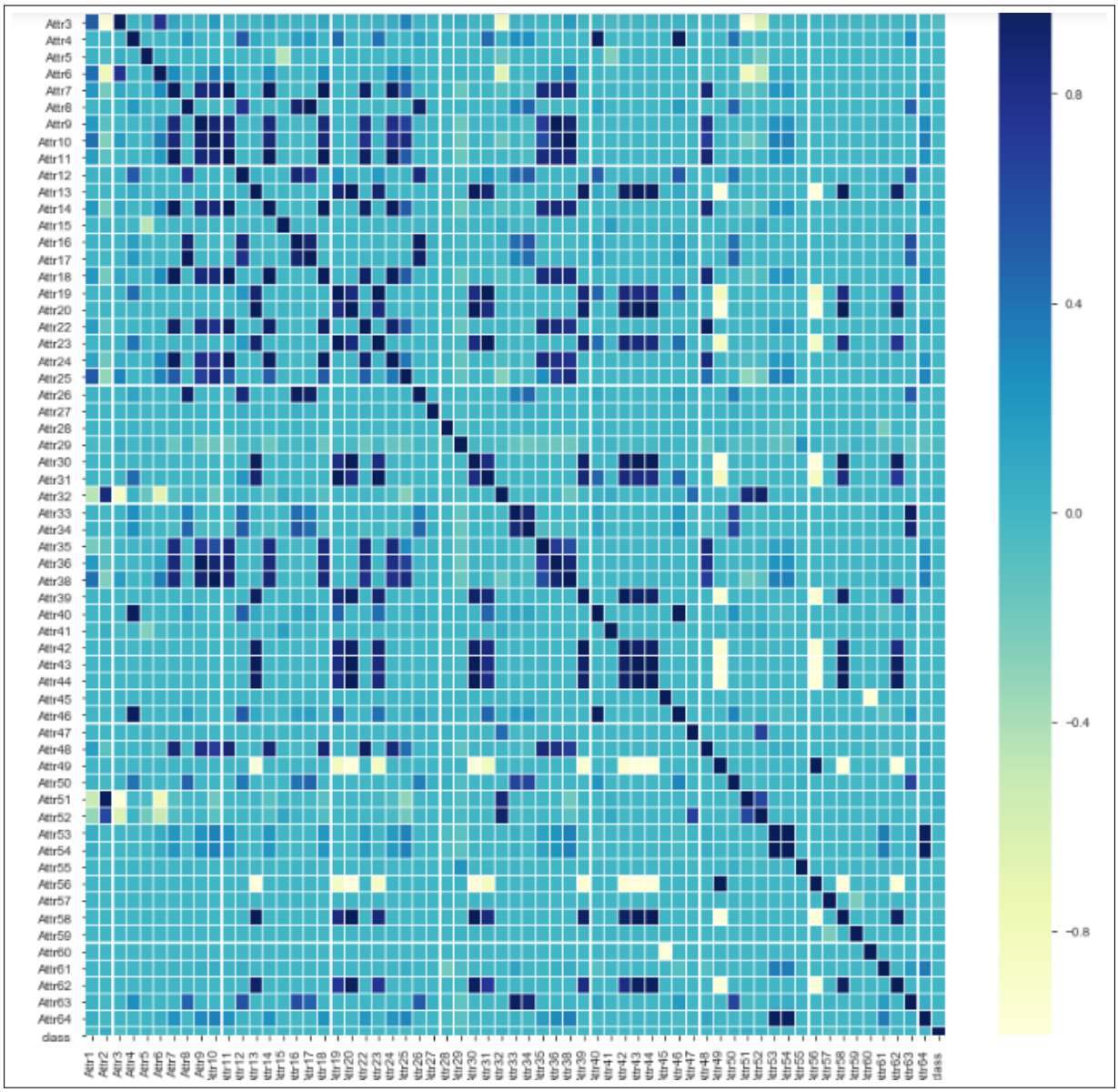

Figure 5: Correlation Matrix

3.3 Data Pre-processing & Transformation

Preparation of the data is a critical step in the process of data mining and also important in order to achieve accurate and quality results from the model. Big data often does contain noise and the presence of special characters, missing values, blank spaces often does a ect the performance of the model and should be handled carefully so as to ensure that the crucial information is preserved and not tampered during the data preparation phase. Although it is assumed that the more data we have, the better predictions can be expected however this is not always the case. There may be a limited subset of features that are more responsible and important for predicting the target variable and some of the variables may be completely useless (Kotsiantis et al.; 2006) . Having redundant and useless variables will increase the computational time and cost hence the data preprocessing phase is often considered as the most challenging and time consuming phase in the CRISP-DM methodology. The steps undertaken to prepare the data required for this study is listed point-wise below:

Converting the le type : Initially after extracting the data le from the source, it was in an ARFF format. To meet with our model requirements, it had to be converted to a CSV format. A Python code has been used to convert the .ar le to a CSV le.

Changing the Datatype of all variables : Normally once the dataset is sourced, the datatypes of all the variables are observed. The values in all the variables were numeric with di erent datatype. These variables were converted to Float datatype.

Handling the NA's or missing values : Presence of missing values are troublesome for the model and can hinder models performance. In the place of missing values, the dataset had special character ('?'). To further deal with these values it were converted to NA's and the percentage of missing values in each column has been observed. Based on this, for columns with a small percent of missing values, the missing values were imputed by the mean of the column. The column that had almost half it's values missing had to be completely removed from the dataset.

Multicolinearity : A correlation test tells us how these variables are a ecting the target variables and also the correlation within the variables. The variables that are highly correlated with each other (p value greater than 0.90) can be said as redundant data that is infact equally contributing in predicting the target variable. It is advisable to remove one of the variable and keep the other one. Out of 64 predictor variables, the dataset contained 26 variables that were highly correlated with the other variables hence had to be dropped from the study.

3.3.1 Feature Selection

Post the process of cleaning the data and removing unnecessary and redundant variables, one of the technique that can further improve the model's performance is known as feature selection. This technique is used for reducing the feature space which may turn out bene cial for the model. The bene ts may include improvement in the accuracy, reducing the risk of over tting, faster computational time and better explain-ability of the model. Explainability is lost when we have too many features in the dataset (Liu and Motoda; 1998).

Feature Name	Feature Description
Attr21	sales (n) / sales (n-1)
Attr24	gross pro t (in 3 years) / total assets
Attr27	pro t on operating activities / nancial expenses
Attr39	pro t on sales / sales
Attr41	total liabilities / ((pro t on operating activities + depreciation) * (12/365))
Attr42	pro t on operating activities / sales
Attr58	total costs /total sales

Table 1: Various Features Selected

Feature selection is one of the preprocessing step that is used prior to building a classi cation model and it tackles the curse of dimensionality problem that has a negative impact on the algorithm. The dataset that has been used in this study consisted of 64 features and few of those may not be useful for predicting bankruptcy. To eliminate the unnecessary features, Random Forest feature selection technique has been used in this study for selecting the best features. In Bioinformatics eld, the usefulness of this technique is studied by Qi (2016) and this technique has been rarely used for predicting corporate bankruptcy. In random forest, the tree-based strategies that are used ranks on the basis of how well they are able to improve the node purity. [3] This is known as mean decrease in impurity or otherwise called as gini impurity. The greatest decrease in the purity of nodes takes place at the start of the tree, whereas the least decrease happens at the end of the tree. Thus, in this a subset of important features is created by nding a particular node and then pruning below it.

The important features that will be used for further preprocessing and model building process are mentioned in Table 1.

3.3.2 Tackling Imbalanced Data

The presence of uneven class distribution is one of the common problems in classi cation modelling (Elrahman and Abraham; 2013). This means that there is a stark di erence in the number of observations of one class than that of the other class. This is a challenging problem in the eld of machine learning because most of the algorithms are designed to perform best when both the classes are equally balanced. In the case of class imbalance, the predictive model that would be developed may turn out to be biased and inaccurate[4]. This may increase the possibility of misclassi cation of the minority class.

A Sampling based approach consists of either oversampling of minority class, undersampling of majority class or Hybrid, which is a combination of both. [5] In this study we will be using a rarely used hybrid approach, SMOTEENN, for dealing with the problem of class imbalance (Monard; 2017). This approach is suitable for this problem as we are dealing with a highly imbalanced dataset where it is necessary to increase the minority class and on the other hand, as there are many instances in the majority class, it is ne to undersample the majority class to some extent.

Synthetic Minority Oversampling Technique (SMOTE) synthetically creates new minority instances between the real minority class and that is done on the basis of nearest

[3] https://chrisalbon.com/machine_learning/trees_and_forests/feature_selection_using_random_forest/
[4] https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classi cation-problem/
[5] http://www.chioka.in/class-imbalance-problem/

neighbours found using Euclidean distance between the data points. ENN separates the instances where the class label is di erent from that of the two of its nearest neighbours. The instances are then removed by ENN method based on the prediction made by KNN. Only those instances are removed whose prediction made is di erent from the majority class.

The distribution of instances in both the classes before and after SMOTEENN on training data can be seen in Figure 6

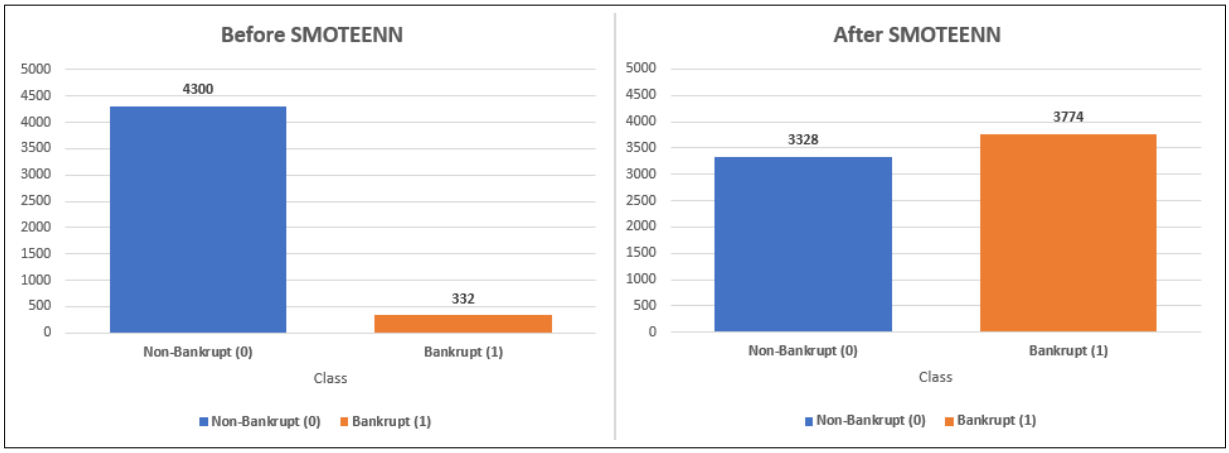

Figure 6: Distribution of instances after resampling

3.4 Modelling Approach

This part is considered to be critical and essential in the machine learning process. Further to the process of data preparation that included feature selection and resampling, the proposed models are to be implemented. The impact of the preprocessing techniques on di erent models can then be evaluated and compared. The details about the model used and the working of the model is discussed in this section.

Random Forest:
A random forest model is an ensemble of many decision trees and is often used in classi cation problems. It uses techniques such as feature randomness and bagging for building each tree such that a uncorrelated forest of tree is obtained (Khoshgoftaar; 2007). Each of the tree relies on an independent random sample. The prediction performance of this collection of trees is more accurate than the individual tree. Few of the factors that make it suitable choice for the chosen dataset include the quick training speed of the model, being robust to outliers and the ability to handle unbalanced data.

Decision Tree:
This is a widely used supervised learning algorithm for solving classi cation and regression problems (Landgrebe; 1991). It has an individual tree representation wherein each leaf node represents a class label and attributes corresponds to internal node of the tree. It starts with the training data as the root and then breaks down into smaller subsets having decision nodes and leaf nodes. For identifying the attribute for root node in each and every level is the challenging part and the

two attribute selection measures that are commonly used are Information gain and Gini index. Based on the learnings from previous literature, this algorithm has performed well on the bankruptcy prediction problem hence we will evaluate the performance of this model on our set of variables.

K Nearest Neighbors:
KNN algorithm works on the assumption that similar instances exist in close proximity. It calculates the distance between the instances using distance calculation formulas, the most popular being Euclidean distance (Cunningham and Delany; 2014). The chosen K value should be considered right if it is able to reduce number of errors and along with that maintain algorithms ability of making accurate predictions. In the literature work, we have not seen a wide use of this technique and will therefore evaluate this technique on our dataset.

AdaBoost:
Boosting algorithms are considered to be exible and powerful. Adaptive Boosting or AdaBoost is a type of boosting algorithm used in classi cation problems that constructs a strong classi er by converting a set of weak ones (Freund et al.; 1999). The classi cation equation for Adaboost can be represented as -

$$F(x) = sign(\sum_{m=1}^{M} \ _m f_m(x)) \tag{1}$$

where mth weak classi er is denoted by f_m and $_m$ as its corresponding weight. Thus from the formula it is interpretable that it is a weighted combination of several week learners. Adaboost is one of the rst successful boosting algorithm for binary classi cation problems and it boosts the performance of decision trees. As it is not widely used in the bankruptcy prediction problem, its performance will be evaluated on the set of variables used in this study.

4 Design Specification

Figure 7 shows the architecture/process ow diagram followed for our research. Initially we gathered the data from the source followed by preprocessing steps that included imputing NA's by mean values, removing redundant columns. Further we used feature selection technique to identify the most important and eliminated the rest. Then strati ed K-fold cross validation (k=5) has been applied to split data into train and test followed by a hybrid resampling technique, SMOTEENN to resample the dataset. Lastly the processed data is feeded to four di erent classi ers and their performance is evaluated on the testing data.

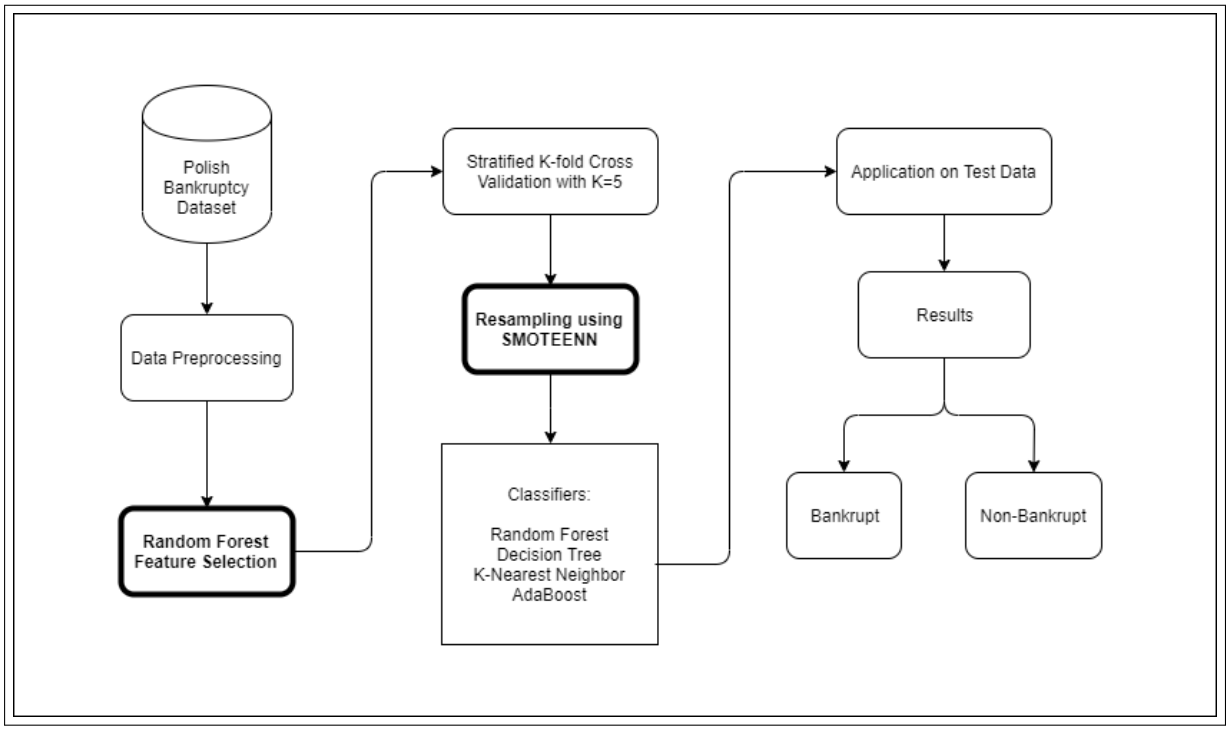

Figure 7: Architecture Diagram

5 Implementation

This section describes the implementation of the proposed models for corporate bankruptcy prediction in detail. Also it describes the process undertaken for selecting most important features and resampling the dataset. The complete implementation part was carried out in Python language v.3.7.2 and Jupyter notebook (v.6.0.2) has been chosen as the Integrated Development Environment (IDE). Python has been chosen for implementation part because it is easy to use, has a wide online support forum and also is considered one of the top languages in terms of code readability. Due to the presence of an active python community, there are plenty of packages available for data preprocessing and handling imbalanced data and thus has been a popular choice for machine learning projects.

The data sourced for this study consisted of ve individual les based on the forecasting period out of which the fth le has been chosen for implementation. It contained nancial rates from 5th year and corresponding class label indicating status of bankruptcy after 1 year. The sourced les were in ARFF format and a python code was available online[6] which was used to convert the le into CSV format. The dataset was then imported into Python as Dataframe and checked for missing values. The special character ('?') in place of missing value was replaced with NA's and individual columns were explored using *pandas_pro ling* package. After the basic cleaning activity, we chose a feature selection method to reduce the features and select only the best ones. Random forest feature selection technique was used and this was done using the *SelectFromModel*[7] package from the *sklearn.feature_selection* library. A threshold value of 0.03 for the gini

[6]https://github.com/haloboy777/ar tocsv/blob/master/ar ToCsv.py

[7]https://scikit-learn.org/stable/modules/generated/sklearn.feature$_s$*election:SelectF romM odel:html*

importance has been used to lter best features. The ltered dataset has been then used for further process of splitting into training and testing set. *Strati edKFold*[8] package frorm sklearn library is used for performing strati ed k fold split with k value of 5.

During data exploration, a high class imbalance was observed. To avoid over tting and build a reliable model, we used SMOTEENN technique after the data was split, to balance the classes. Due to high imbalance, we used a hybrid approach that can oversample the minority class and undersample the majority class. *SMOTEENN* package from the *imbalance-learn* library was used for resampling. Further we applied four di erent models on the resampled dataset. The various models that were implemented Random Forest, Decision tree, K Nearest neighbours and AdaBoost. These models are available in the form of di erent packages in sklearn library in Python. For each model, during the iteration process that takes place number of times based on value of K, the metrics like true positives, true negatives, false positives and false negatives were appended in a list and then the mean of these values were treated as the model outcome and was used to calculate the individual metrics. The evaluation metrics used in this study were accuracy, speci city, sensitivity and geometric mean of speci city & sensitivity. For each model, these metrics were calculated and compared by plotting bar plots. It was found out that model built using Random forest achieved highest accuracy and KNN achieved the lowest accuracy. Further a detailed evaluation and comparison of the results is carried in section 6.

6 Evaluation

The main purpose of our study is to evaluate the performance of our model and assess whether the techniques used are suitable for this problem. The performance of the models are compared using metrics like accuracy, recall, speci city and geometric mean of recall & speci city. These metrics are chosen as they are suitable for binary classi cation problems such as ours and are commonly used in previous studies like the one conducted by Faris et al. (2019). These metrics can be derived from the values gathered from the confusion matrix which contain true positives, true negatives, false positives and false negatives. The signi cance of these values in terms of bankruptcy context is as follows- TP's are the companies that are actually bankrupt and model predicted as bankrupt; TN's are the companies that are actually non-bankrupt and model predicted as non-bankrupt; FP are the companies that are actually non-bankrupt and model predicted as bankrupt and FN's are companies that are actually bankrupt and model predicted as non-bankrupt. As we have used cross validation, for each model ve confusion matrix were generated. For further calculations and to build an unbiased model, mean value of the confusion matrix was obtained for each model and the metrics were calculated as per below equations-

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Recall{=}Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

[8] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html

Model	Accuracy	Recall/Sensitivity	Specificity	Geometric Mean (GM)
Random Forest	95%	38%	99%	62%
Decision Tree	93%	56%	96%	73%
KNN	92%	14%	98%	36%
AdaBoost	95%	49%	98%	70%

Table 2: Performance of Models before Resampling

$$GM = \sqrt{Sensitivity{:}Specificity} \tag{5}$$

The GM aggregates both the measures, speci city & sensitivity. This measure is suitable for imbalanced datasets as the primary goal of a classi er is to improve sensitivity while maintaining the speci city (Tharwat; 2018).

6.1 Case Study 1: Model Performance on Unbalanced Data

A base model for each classi er is built on the unbalanced data after feature selection. The performance of individual classi er can be observed in the Table 2. From the table we can infer that the accuracy of the classi ers are high however the recall is poor which shows that it is failing to predict the bankrupt cases. This shows that our model is biased and largely over t due to a high class imbalance. Decision tree outperformed the other classi ers with recall of 56% although still being very poor whereas KNN had the lowest recall of 14%. The other two metrics, speci city and accuracy is almost same for all classi ers therefore GM of Decision tree is highest at 73% and KNN is lowest at 36%. To overcome the issue of low recall and improve performance, we further applied SMOTEENN on the classi ers while keeping the same feature set and evaluated their performance.

6.2 Case Study 2: Model Performance on Balanced Data

Now, we have performed the same experiment on the dataset that has been resampled using SMOTEENN. From Table 3, we can clearly observe that the model has now improved as seen from the obtained recall values. The AdaBoost classi er outperformed others in terms of recall with value of 73%. which means that the model is able to correctly predict 73% of the total bankrupt cases which is not a bad score. Also, KNN showed lowest recall of 60% which is still higher than the highest value in previous experiment. Also there's slight improvement in GM values as speci city has decreased slightly for all models. This shows that the model is now better in terms of predicting a bankrupt company. Among the classi ers, AdaBoost showed the best performance with GMean of 80% and KNN had the lowest GM value of 67%.

Further we will discuss about the e ect of SMOTEENN on individual classi ers for chosen set of features and dataset.

Model	Accuracy	Recall/Sensitivity	Specificity	Geometric Mean (GM)
Random Forest	89%	66%	90%	77%
Decision Tree	86%	65%	87%	75%
KNN	74%	60%	75%	67%
AdaBoost	88%	73%	89%	80%

Table 3: Performance of Models after Resampling

7 Discussion

The Recall metric will tell us how accurately our model has predicted a bankrupt company as bankrupt. This is an important metric for this type of classi cation problems as we don't want a bankrupt company to be mis-classi ed as non-bankrupt. Another metric, speci city will tell us how accurately model has predicted a non-bankrupt company as non-bankrupt. Although this metric is also important, more importance is given to Recall as on a broader scenario, having a model that will classify a bankrupt company as non-bankrupt will make it less reliable and may attract huge losses for investors. Hence, the aim is to improve the recall and keep the model trustworthy.

After thorough observation of the results from Table 2 and Table 3 we can state that a signi cant improvement is visible in the sensitivity/recall and Gmean values of the classi ers after the use of SMOTEENN resampling technique for handling the issue of data imbalance on the chosen set of features. Through this technique we successfully oversampled the minority class, bankrupt cases, and undersampled the majority class i.e. non-bankrupt cases and it is a rmed that by doing this there is an improvement in the recall and Gmean performance.

The e ect of SMOTEENN on di erent classi ers varied signi cantly out of which the performance of KNN improved drastically. The individual performance is compared in Figure 8. Initially, the model gave a recall of 14% and after resampling it had a recall of 60%. Even Random forest model showed signi cant improvement in the recall value after resampling the data. As seen from Figure 9, AdaBoost achieved the highest Gmean value of 80% while having Acccuracy of 88% and Random Forest outperformed all the other classi ers in terms of accuracy at 89%. After going through previous studies, it was observed that these techniques, with variation in feature selection strategy and resampling technique, have been used very sparsely in this problem area and AdaBoost and Random Forest have achieved signi cant results in the studies conducted by Faris et al. (2019) and Le (2018). There isn't a direct improvement in the performance of our model than the previously mentioned studies however few metrics like Recall of AdaBoost is promising and worth exploring further. Our study suggests that based on all the metrics studied, AdaBoost is a better classi er than the other models considered in this study and the improvement in the performance of KNN classi er is notable.

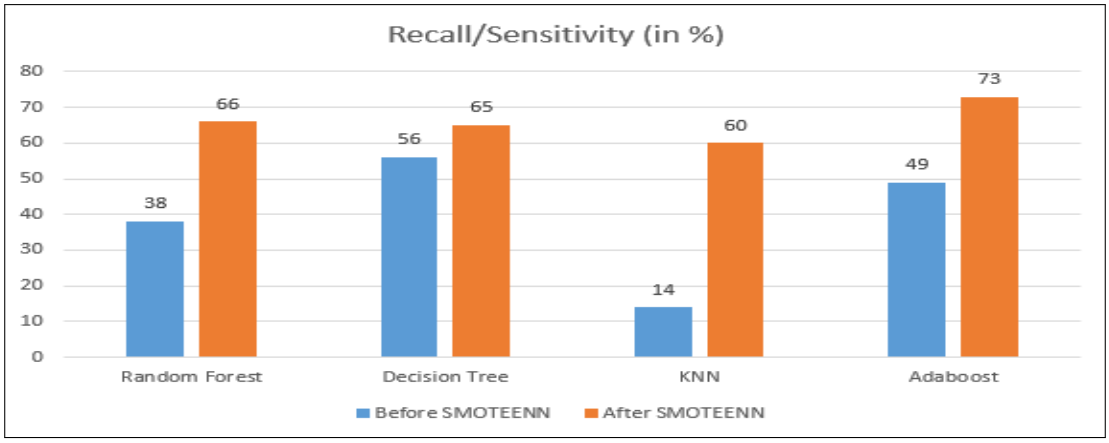

Figure 8: Recall values of di erent Classi ers

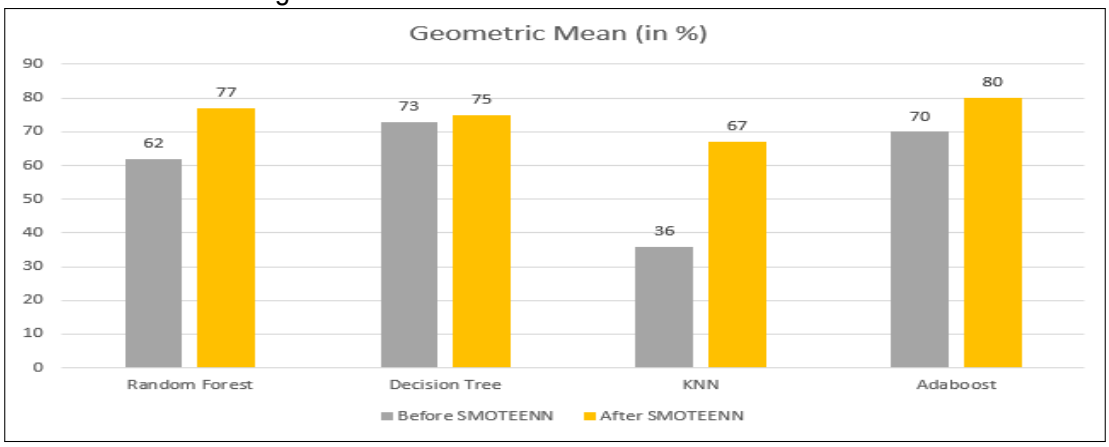

Figure 9: Gmean values of di erent Classi ers

8 Conclusion & Future Work

As seen from the previous literature, the topic of bankruptcy prediction has been of interest in last few decades and numerous techniques have been implemented in the quest for achieving optimum prediction performance. The challenging part, as discussed, lies in selecting the best nancial attributes that are majorly responsible for a company's bankruptcy. Another hurdle in this study is the problem of imbalance in the classes. In this research, we have used a novel combination of a feature selection technique and a resampling technique and implemented four models on the same. Our techniques, Random Forest feature selection and SMOTEENN with Random Forest classi er is a better model followed by AdaBoost classi er in terms of the prediction accuracy and the similar techniques with AdaBoost classi er outperforms other classi ers in terms of Geometric Mean and Recall values.

These techniques can further be explored on other classi ers along with di erent set of nancial attributes. In this study, we limited our research to Polish bankruptcy data. In future, one may explore bankruptcy data of any other region. Further, KNN classi er showed impressive performance improvement with SMOTEENN. This combination can further be explored in other similar type of classi cation problems.

Acknowledgement

I would like to extend my sincere gratitude to my supervisor Dr. Vladimir Milosavljevic for providing constant feedback and valuable suggestions throughout the implementation phase of this research project. I would also like to thank my family & friends for their constant support and encouragement without which this research would have been impossible.

References

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

Ayyadevara, V. K. and Ayyadevara, V. K. (2018). Random Forest, *Pro Machine Learning Algorithms* (Iciccs): 105{116.

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.

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*

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

Bernanke, B. B. E. N. S. (2015). Bankruptcy , Liquidity , and Recession, **71**(2): 155{159.

Cunningham, P. and Delany, S. J. (2014). k-Nearest neighbour classi ers k -Nearest Neighbour Classi ers, (April 2007).

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

Ding, Y. (2008). Forecasting nancial condition of Chinese listed companies based on support vector machine, **34**: 3081{3089.

Elrahman, S. M. A. and Abraham, A. (2013). A Review of Class Imbalance Problem, **1**: 332{340.

Faris, H., Abukhurma, R., Almanaseer, W., Saadeh, M., Mora, A. M., Castillo, P. A. and Aljarah, I. (2019). Improving nancial bankruptcy prediction in a highly imbalanced class distribution using oversampling and ensemble learning : a case from the Spanish market, *Progress in Arti cial Intelligence* .
URL: *https://doi.org/10.1007/s13748-019-00197-9*

Freund, Y., Schapire, R. E., Avenue, P. and Park, F. (1999). A Short Introduction to Boosting, **14**(5): 771{780.

Geng, R., Bose, I. and Chen, X. (2014). *Prediction of nancial distress: An empirical study of listed chinese companies using data mining*, Elsevier B.V.
URL: *http://dx.doi.org/10.1016/j.ejor.2014.08.016*

Khoshgoftaar, T. M. (2007). An Empirical Study of Learning from Imbalanced Data Using Random Forest, pp. 310{317.

Kim, M.-j., Kang, D.-k. and Bae, H. (2014). Expert Systems with Applications Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction, *EXPERT SYSTEMS WITH APPLICATIONS* (September).
URL: *http://dx.doi.org/10.1016/j.eswa.2014.08.025*

Kotsiantis, S. B., Kanellopoulos, D. and Pintelas, P. E. (2006). Data Preprocessing for Supervised Leaning, **1**(1): 111{117.

Kumar, P. R. and Ravi, V. (2007). Bankruptcy prediction in banks and rms via statistical and intelligent techniques { A review, **180**: 1{28.

Landgrebe, I. (1991). A Survey of Decision Wee Classi er Methodology, **21**(3).

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

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**.

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

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

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

Naidu, G. P. and Govinda, K. (2018). Bankruptcy prediction using neural networks, *2018 2nd International Conference on Inventive Systems and Control (ICISC)* (Icisc): 248{251.

Prusak, F. I. (2019). Corporate bankruptcy prediction in Poland, **15**(1): 10{20.

Qi, Y. (2016). Random Forest for Bioinformatics, pp. 307{323.

Rustam, Z. and Saragih, G. S. (2018). Predicting Bank Financial Failures using Random Forest, *2018 International Workshop on Big Data and Information Security (IWBIS)* pp. 81{86.

Sha que, U. and Campus, L. (2014). A Comparative Study of Data Mining Process Models (KDD , CRISP-DM and A Comparative Study of Data Mining Process Models (KDD , CRISP-DM and SEMMA), (November).

Shi, Y. and Li, X. (2019). An overview of bankruptcy prediction models for corporate rms : A systematic literature review, **15**(2): 114{127.

Tharwat, A. (2018). Classi cation Assessment Methods, *Applied Computing and Informatics* .
URL: *https://doi.org/10.1016/j.aci.2018.08.003*

Veganzones, D. and Severin, E. (2018). PT, *Decision Support Systems* p. #pagerange#.
URL: *https://doi.org/10.1016/j.dss.2018.06.011*

Zori, M., Gnip, P., Drotar, P. and Gazda, V. (2019). Bankruptcy prediction for small- and medium-sized companies using severely imbalanced datasets, (February).