Master Thesis report on

Efficiency of LightGBM technique in Bankrupty Prediction using Polish dataset

Submitted in partial fulfillment of the

requirements of the degree of

Master of Science

in

Big Data and Business Analytics

by

Mr. Shidharth Bammani

(Matriculation Number 11011885)

Under the guidance of

Dr. Frank Schulz

Department of Information and Media Design

and

Under the guidance of

Prof. Swati Chandna

Department of Information and Media Design

DEPARTMENT OF INFORMATION AND MEDIA DESIGN

SRH HOCHSCHULE HEIDELBERG

28-November-2020

Declaration of Authorship

I, Shidharth Bammani, declare that this thesis titled, " Efficiency of LightGBM technique in Bankruptcy Prediction data using Polish Data " and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

_____

Date:

_____

Abstract

by

Financial Insolvency is the major issue faced by companies worldwide, as many factors lead to the financial Insolvency of a company. To predict the bankruptcy earlier than it happens will eventually give the companies adequate time to react and work on the factors leading to it. The study on bankruptcy prediction has always been an interesting topic to the creditors and investors in evaluating a firm's likelihood of becoming bankrupt. The predictive model aims to alert the company of financial distress by combining various econometric parameters known as financial ratios. However, the financial data are becoming more sophisticated, and the traditional methods may not deal with such data or may perform poorly. At the same time, new classification and statistical techniques are emerging for the last few decades. In this master thesis, we did a comparative research analysis of traditional statistical models such as Gaussian Naïve Bayes, Logistic Regression, Random forest, Decision tree and ensemble learning method XGBoost with respect to LightGBM model for bankruptcy prediction. We have chosen the Polish company data set from the UCLA-LoPucki Bankruptcy Research Database (BRD). The data holds financial ratios in order to reflect higher-order statistics. The research's main purpose is to prove that ensemble learning methods work well as classifiers for bankruptcy data. Light Gbm provides much higher accuracy and optimizes time consumption than the existing classifiers, as the observation proves. ...

Acknowledgements

I would like to thank SRH Hochschule for giving me the chance to do my Master's thesis. I am grateful to Dr.Frank Schulz for his guidance and valuable insights leading to successful completion and submission of this dissertation.

I would like to convey to Dr. Frank Schulz my heartfelt appreciation for constructive input and guidance helped me find the path for this thesis. i would also like to thank Prof. Swati Chandna for helping me and providing me with time, support and valuable feedback that I neeeded the most.

Last but not the least i would like to thank my parents, my sister, my brother-in-law and my friends Sachin, Taha and Abhinav for their constant support and encouragement which helped me in completion of this dissertation. ...

Contents

List of Figures

List of Tables

For/Dedicated to/To my…

Chapter 1

Introduction

Bankruptcy is a term we use when the company is unable to repay its debts to the creditors, it is a legal status imposed on the company by a court order. No company or a person would like to entitle this status and would avoid becoming financially insolvent, hence enter of the term bankruptcy prediction. Bankruptcy prediction is art of predicting the bankruptcy before it happens and it can be helpful for not just the company which is trying to survive a financial turmoil but also to various other entities. Bankruptcy prediction has been a vast area of interest for management, creditors and investors as well. While performing the prediction we assess the underlying problems that cause bankruptcy and alert the company by providing distress signals. To identify the cause, a lot of research has been done in the field of finance domain by introducing financial ratios or the economic indicators in the study. The financial ratios are comparative magnitude that are derived from the companies financial statement or also known as the balance sheet. The financial ratios is a common word in the accounting domain and financial analyst have been using these ratios to understand a company's strengths and weaknesses. Bankruptcy prediction can be traced back to it early days in 1932 by FitzPatrick where around 20 firms where analysed to identify the bankrupt companies, although the research was not conducted in the way we generally perform with statistical analysis but careful understanding of the ratios and the trends[24]. We notice the importance of the finantial ratios and the study takes a step further by introducing Z scores, a mutiple discriminant analysis of financial data by [1]. Later in 1980 Ohlson introduced O'scores the application of logit regression in the bankruptcy prediction domain[24]. Having the early introduction of statistical methods lead to more research in this domain and eventually leads to introduction of

machine learning algorithms. The modern approach to prediction of bankruptcy is an important topics for the researchers and financial investors and creditors.
Further in bankruptcy prediction more sophisticated and advanced models where introduced. There has been a lot of literature available in this domain and we can refer the research of [11] where the author had addressed most of the literature done in the field of bankruptcy till 2009. Most of the issues that are addressed in bankruptcy prediction is the modelling technique. Modelling of bankruptcy prediction is an important part of the research, misclassifying the firms we cost the management tremendously and hence introducing new bankruptcy models that have high accuracy and performance is crucial. Hence it is justifiable to have the main attention on the modelling techniques. After the financial crisis in 2008-2009, more of the focus was yet again actualized on bankruptcy prediction and implementation of newer models to the domain. In the year 2019 a lot of companies were filled for bankruptcy due to COVID 19. Hence, bankruptcy prediction is a trending topic and it is interesting to identify the potential of the new models. Causes of bankruptcy depends on many factors, the modelling of new machine learning technique is one way to tackle the problem but predicting bankruptcy may have several other reasons. Financial analyst and researchers are interested in finding the underlying problems of bankruptcy, feature selection is another way in alerting the company about the distress signals. For selection of feaures, one must have in depth knowledge of the finance domain to highlight the financial indicators. Along with the feature selection the type of failure should also be addressed in the current situations where the company might have different groups to classify instead of being either bankrupt or non bankrupt. We can have multiple state of the financial standings of a company, some of the states are liquidation, insolvency, vulnerability etc.

1.1 Motivation

Bankruptcy prediction has always been a part of active research work, it is always in trend to come up with better prediction models, financial indicators, and other factors that lead to bankruptcy. In the current scenario, a lot of companies filed for bankruptcy due to COVID 19 pandemic. The potential of new models to identify the bankruptcy

factors can be good research to have in the bankruptcy prediction domain. The LightGBM model has been recently released by Microsoft in 2018, and not much of the research work is available. LightGBM is known for its faster training speed and accuracy in Kaggle competition. Hence, knowing how this ensemble learning technique works on the bankruptcy data can be interesting.

1.2 Problem Definition

The problem definition of our research can be summarized in the below statement:

What are the conclusion that can be drawn from the implementation of the newly introduced ensemble learning technique LightGBM in corporate bankruptcy prediction?

1.2.1 Objectives

The objectives of the thesis consist of three parts and are given below.

The first objective of our research is to build classification models for predicting bankruptcy, our key focus is to compare the performance of the LightGBM model with the traditional methods (Logistic Regression, Gaussian Naive Bayes, Decision tree) and ensemble learning methods (Random Forest, XGBoost) on balanced data with default parameters for boosted algorithms. We have to keep on updating our models while predicting bankruptcy data as new financial ratios and economic indicators are introduced in the domain and aim for achieving better results.

Our second objective is to evaluate the performance on highly skewed bankruptcy data of two significant gradients boosted machines XGBoost which has won many Kaggle competitions and comparing with LightGBM which has less research work done.

The third objective of our thesis is to evaluate the LightGBM model with a gradient one-sided sampling technique. Gradient boosting framework with GOSS implementation to the best of our knowledge has not been used in the bankruptcy data prediction and we will evaluate the model. In real-world bankruptcy prediction, our model needs to be robust and faster in predicting the outcome and most importantly it needs to be precise despite having highly skewed data.

1.2.2 Research Questions

We have stated the problem definition and the objective of this thesis, we will summarize our objective and form research questions mentioned below.

- How well the LightGBM performs in terms of bankruptcy prediction comparing with the reference models?
- How well the LightGBM model performs in comparison to XGBoost with data containing missing values and class imbalance?
- Can we improve the corporate bankruptcy prediction by implementing the LightGBM GOSS technique to the highly skewed Polish bankruptcy dataset?

1.3 Contribution

In this thesis we provide a comparative analysis of bankruptcy prediction models, we implement the LightGBM technique to the Polish bankruptcy dataset and evaluate the performance of the model and check the efficiency of the model in comparison to traditional statistical models like Logistic Regression, Gaussian Naive Bayes, Decision trees, Random Forest and gradient boosting technique XGBoost. Best to our knowledge LightGBM is not been used in the research of bankruptcy prediction as it is recently introduced by Microsoft in 2018. However, LightGBM has been used in other studies for classification problems. The model is insensitive to class imbalance and missing values in the data. To improve the prediction result we have used advanced hyperparameter tuning to achieve the best results. The model is evaluated by using the AUC ROC curve and with optimum threshold values. We test our model on the Polish bankruptcy dataset.

1.4 Overview of the section

The thesis structure is divided into 7 sections are described in the following order. Section 2 introduces the literature done in the field of bankruptcy prediction and the

related works to our proposed model. Section 3 describes the methodology and different modelling approaches taken. Section 4 explains the architecture of our experimental setup. Section 5 provides with the implementation of our models. Section 6 consists of the evaluation of our models and the comparison results. Section 7 is a discussion of our proposed model and Section 8 is for conclusion and future scope.

Chapter 2

Literature Survey

In this section, we will be stating some of the existing state of art present in the machine learning world that have been in practice for bankruptcy prediction. In the first segment, we present the empirical bankruptcy prediction models and techniques. In the second segment of this section, we present the specialized models and techniques related to bankruptcy prediction and review the customized models and techniques that will help us understand the related works and outlines the difference from our proposed model.

2.1 Background

In this section we will be presenting the literature in the domain of bankruptcy prediction and explain the research papers in attempt to show case the empirical methods and techniques that help us to understand the existing work.

Edward Altman[1], proposed a novel approach to financial ratio analysis, the paper asses the quality of traditional ratio analysis as an analytical technique and proposes a multiple statistical discriminant method where new financial ratios are introduced which are used for investigating the performance of corporate companies. In this paper, the shortcomings of the traditional ratios used for bankruptcy prediction are discussed, and further, a discriminant method is developed and its compatibility with traditional ratios. The financial ratios are introduced to hold the potential to be predictors of bankruptcy. The discriminant functions transform the financial ratio values to a single discriminant score which is known as the Z score. Although the model

performed well in predicting the bankrupt company as it relies on the data directly from each company and is only good if the data is provided by companies.

Back, Barbro and Laitinen, Teija and Sere, Kaisa (1996)[5], have evaluated a different technique for feature selection of neural network that predicts the failure of the company one year before and gives best results. The financial predictors are selected by three different techniques and each has a different understanding of the relation among the variables and the chances are higher of finding the right predictive variable to be selected. The techniques used are Linear Discriminant Analysis and the variables are selected based on the linear combination, Logit analysis is based on logistical cumulative function and the third technique is Genetic algorithms which are based on mechanics of natural selection and natural genetics. Among the three techniques used for variable selection, the genetic algorithms have outperformed the other two. The study shows that we can have the same predictors for machine learning models as that of statistical methods and obtain better results although the models predict 1 year before actual failure but do not predict well for 2 or 3 years before failure.

Edward Altman (2000)[2], proposed a revised Z score model, denoting as Z''. In this paper, Altman is presenting the need to rethink the earlier published Z score models in 1968 and 1977. By the end of 1999, there was an assumption among the academicians that the ratio analysis to be eliminated as an analytical technique in assessing the performance measure of corporate companies. The analytical and practical value of financial ratios are highlighted, and the characteristics of business failures are examined to quantify the effective predictor and indicator variables. The utility of ratio analysis is discussed and tries to bridge the gap between traditional ratio analysis and more rigorous statistical analysis techniques. The model was applied to manufacturing companies and that it displayed good results, outperformed the other bankruptcy strategical analysis, and showed no change when applied the same to retail firms. The Zeta models show better accuracy when compared to the old Z score model and the application is the same as the previous employed models.

Peña, Tonatiuh & Martinez-Jaramillo, Serafin & Abudu, Bolanle. (2009)[20], made a comparison of classification performance of statistical analysis Altman Z score and machine learning algorithms such as Logistic Regression, Least-squares Support Vector Machine and Gaussian classifiers (Bayesian Fisher Discriminant and Warped GP's)

concerning the bankruptcy prediction problem. In this paper, the new techniques of Gaussian processes are used for binary classification and show that standard approaches for classification are based on parametric models and Gaussian processes that are non-parametric perform well in comparison to well-established techniques like Altman Z score and Logistic Regression. However, this technique is not suitable for complex problems where data is to be separated between classes.

Philippe du Jardin (2009)[11], has provided extensive research of almost all the empirical work that has been conducted on bankruptcy prediction and sketched a well-structured informative paper. The study shows us the focus of research that has been done on different issues, we find that most study has been done over modeling techniques and their selection, second, we see that study has been done on finding the right predictors for the models, the third issue that has been addressed is the type of failure that can be predicted for example whether a company is bankrupt or not bankrupt, different states of financial health and so on, the fourth issue is the sample size of data and fifth issue that is rarely addressed in the analysis of uncommon variables. The study also tells a lot about the reasons for the company's failure and the evaluation criterion.

Fang-Mei Tseng, Yi-Chung Hub (2010)[21], applied four forecasting techniques, Logit model, Quadratic interval logit model, backpropagation MLP and Radial basis function network (RBFN) to the bankruptcy data of companies in England, the data holds less observation. The study shows that radial basis function network produced the best result in predicting the bankrupt and non-bankrupt companies followed by backpropagation MLP, quadratic interval logit model including (including defuzzy) were similar to each other and Logit model performance was the least. It was interesting to know that RBFN and MLP are similar to each other, the difference is that in RBFN each hidden node has a radial basis function, and these models perform well with fewer data.

Miche, Yoan and Séverin, Eric and Lendasse, Amaury (2011)[26], conducted a classification of bankruptcy data that have a high amount of missing data by using a novel technique in bankruptcy prediction. The research states the use of ENN i.e. ensemble of nearest neighbors which is similar to KNN but the difference is we K is not searched for assigned but the ensemble of nearest neighbor learners are used and are combined

linearly. Euclidean distance metric is used to take care of missing data. There are four datasets on which the classification task is performed and zero to fifty percent of missing data. The model is robust when ensembled and the result are not deteriorating and provides good results.

Chaudhuri, Arindam (2013)[9], investigated different approach of building a predictive model and the impact of having choice of threshold, data sampling and the business cycle on bankruptcy prediction. The research states use of both parametric and non-parametric techniques for prediction and have implemented four bankruptcy models namely Hazard, Mixed Logit, Bayesian Networks and Rough Bayesian networks. The study shows that the cut off points, sampling strategy can affect the bankruptcy model prediction and given optimal cut-off point the chances of misclassification is lowered. The study provides insight in choosing the condition of the bankruptcy model. The result shows that Bayesian networks performed poorly in comparison to the other models.

Zhou, Ligang, and Lai, Kin Keung (2016) [28], proposed an approach in dealing with missing data in bankruptcy prediction. AdaBoost machine learning is implemented using a gradient boosting framework and different imputation techniques. The results show that the evaluation of such a model combined with the imputation technique has good prediction results and does not affect the classification. The model is also trained with different samples to train them well. Neural network is selected as the base learner for boosting technique and imputation methods Knn, mean and global closest fit is used, and the data is of firms in the USA.

Nanxi Wang (2017)[23], proposed three new machine learning models Autoencoders, Neural networks with dropouts, and Support Vector Machine application on bankruptcy data. The results are compared to already established methods such as Robust Logistic Regression, Inductive learning algorithms, and Genetic algorithm and the machine learning models show high accuracy. The newly adopted models show improvement in the aspect of Overfitting, and large features spaces are taken care of.

Flavio Barboza, Herbert Kimura, Edward Altman. (2017)[6], made a research on computational methods and compared the results with the traditional statistical techniques and proves the superiority of computational models over them. The comparison is made between statistical models Logistic Regression, Discriminant Analysis,

and early machine learning models namely Neural Network, Support Vector Machine, Random Forest, Bagging, and Boosting in predicting bankruptcy one year before the event. In the study, Altman Z score variables are used as the predictive variables along with six new financial indicators. The result shows substantial improvement in the applied machine learning techniques. The research says that the machine learning models show 10 percent more accuracy than the traditional statistical methods, and the ensemble techniques Bagging and Boosting outperforms other models and the result for the random forest is like it and the ANN, Logistic Regression and Multiple Discriminant Analysis have low predictive accuracy. The study displays that credit risk is associated with bankruptcy prediction and that Altman and Ohlson's models are still relevant to bankruptcy prediction as they provide a simple and consistent framework along with predictive power.

Le, Tuong and Son, Le and Vo, Minh and Lee, Mi and Baik, Sung (2018)[14], proposed a novel technique to solve the class imbalance in bankruptcy data prediction. The class imbalance is handled by using clustering technique, CBoost which stands for Cluster based Boosting algorithm and does classification task. The centroid of the clusters are taken as the weights for training the consecutive base learners. The research also implements IHT i.e. Instance Hardness Threshold to remove the noise in the data. The study states that CBoost outperforms other techniques such as gradient Boosting method where the class imbalance is taken care by oversampling techniques such as SMOTEEN.

Tomasz Korol. (2019)[12], implemented a dynamic approach to bankruptcy prediction model by using four different methods to build forecasting models. The methods used are fuzzy sets, two artificial neural network methods, and decision trees methods. The objective of the paper is to find the relevance in prediction if there is a change in indicators and it states that the change of values in indicators do not immediately indicate the company is in distress or not. The research shows the influence of the dynamic approach using four different techniques and changes in the effectiveness of the models. The dynamic approach to financial ratios let us know more about the economic conditions of the company. The fuzzy set model proved to be the best among others in terms of effectiveness and predictive abilities. The limitation is access to data being limited and requires organized data for evaluation.

Yuri Zalenkov (2020)[27], proposed a novel approach to bankruptcy data prediction by introducing Survival Analysis Technique. The research states that a lot of studies have been done on bankruptcy prediction and most of them are based on classification and they let us know about the posterior possibility of bankrupt companies and the time is not specified as in when the company may become insolvent. Yuri came up with the idea of using a technique that is new to financial prediction, the SA technique has been used prior in the medical and technical science field. This technique extracts useful information from the given data and estimates the impact of the feature variables. In this study, a comparison of survival analysis techniques and classification techniques is also done, and the SA techniques perform well.

2.2 Related works

In this section, we will explain the research work related to our proposed model. Zięba, Maciej and Tomczak, Sebastian and Tomczak, Jakub (2016)[29], proposed a novel technique in bankruptcy prediction by implementing an ensemble of boosted tree learners, taking gradient boosting framework, and using Extreme gradient boosting technique. The data is of bankrupt companies of Poland and the author has coined the word synthetic features for the predictive variables. The research shows that the extreme gradient boosting technique works well with the dataset that has data imbalance. We see that the financial predictive data holds high variance for a small sample of data and ensemble learners like neural networks lead to poor performance whereas regularized boosted trees have proved to have higher accuracy results. The study has done a comparison of XGB with other binary classification methods that are not commonly used and the result of the XBG classifier is proved to be superior to other reference methods. Wyrobek, Joanna, and Kluza, Krzysztof. (2019)[25], compared Gradient Boosting technique and two traditional statistical methods i.e. Linear Discriminant Analysis and Logit function in the field of bankruptcy prediction and state two research hypotheses that GBM is better at predicting failure and it performs well if the data is raw and normalized from financial statement instead of financial ratios. The research proves the hypothesis and shows that LDA and Logit function does not perform well if the data has high variance whereas boosted trees are excellent learners

when the results are combined, also logit functions are prone to outliers. The study says that due to strong pre pruning and the trees are shallow and require less memory and the higher the number of trees, the better is the model but the additional trees are only added if the previous learner is improved. GBM is sensitive to parameter tuning but if set precisely then better the accuracy results. Gradient boosting decision trees tend to have a short-term nature and the models need to be given new data from time to time. Meng, Qi (2018)[16], proposed two novel techniques using Gradient boosting decision trees framework as base, introducing Gradient-based one-sided sampling (GOSS) and Exclusive Feature Bundling (EFB). These two techniques when incorporated with gradient boosting technique tends to outperform well known boosting techniques like Extreme Gradient Boost and pGBRT in terms of memory optimization and faster training speed, hence it is called as LightGBM. GOSS estimates accurately the information gain and plays an important role in it even if the data is small, whereas EFB reduces the number of features by bundling the features which are mutually exclusive. The newly proposed technique makes the standard GBM 20 times faster in training process.

Chapter 3

Methodology

For a data mining project the two most common approaches that are followed by industry specialist are CRISP-DM (Cross Industry Process for Data Mining) and KDD (Knowledge Discovery in Databases). There is another popular approach known as SEMMA and is developed by SAS institute and the acronym stands for Sample, Explore, Modify, Model, Assess and refers to the way of conducting Data Mining projects. CRISP-DM and SEMMA both incorporates the steps of KDD [3]. CRISP-DM appears to be more complete than SEMMA and we will be using it to explain the process flow. The architecture and the process flow is shown in the fig 4.1

3.1 Business Understanding

The first objective before starting any project is to gather the complete objective of the project. The primary aim of the research is to recognize the earlier signs of bankruptcy and alert the company or the individual about the early financial standings of the company. This can be achieved by analyzing the financial indicators of the company. So our primary objective is to build a bankruptcy prediction model that will help the creditors, managers, etc, regarding the current standing of the company and alert them about the potential risk of bankruptcy soon. As observed in the literature survey, the number of globally bankrupt companies is less compared to the nonbankrupt companies, and that many factors may lead to bankruptcy. To identify the bankrupt companies our model needs to be robust and identify the bankruptcy distress with more accuracy based on the training data which tends to be less. Secondly, there can be many factors which may lead to bankruptcy and to identify those factors we need to have an

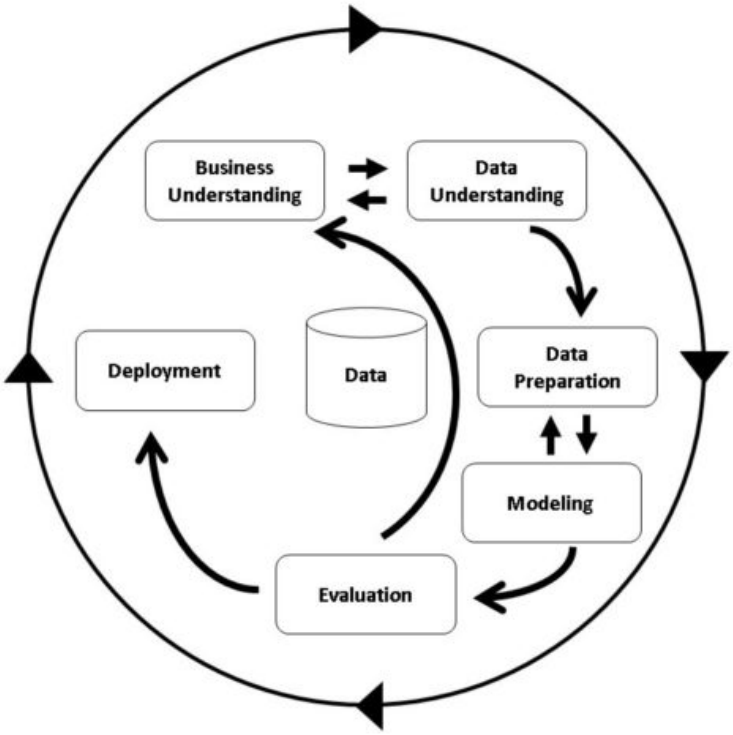

FIGURE 3.1: CRISP-DM Methodology

understanding of the economic indicators and the financial records as they high variance with relatively small data samples[29] so building a model that efficiently utilizes such data and provides accurate prediction results. The research would help us take further in the field of bankruptcy prediction and will be beneficial for the industry experts.

3.2 Data Understanding

For our next step in building a better prediction model the understanding of the data is crucial and is the second step in the CRISP-DM process. To build a reliable model we need to understand its component and the data that is required for its working. We can find data that is available on multiple sources although we cannot say that it is reliable at the first glance, the data obtained from the source might be unethical as well. So extraction of data from a few sources which are reliable and ethical requires

a considerable amount of time. For our research work, we require the financial ratios of companies that belong to a particular region for a certain period of time which includes bankrupt and nonbankrupt cases. The data description and data source are described below. We have taken the data-set of Polish companies and is hosted by UCI Machine Learning Repository, it is a huge repository and is freely available for research and learning purposes. The reason for choosing this data set is because Poland is the 6th largest economy in the European Union. During the year 2004, a major number of manufacturing companies went bankrupt in Poland, and the bankrupt companies were analyzed in the period of 2000-2012. The data is archived from Emerging Markets Information Service database (EMIS), it holds more than 540 publication of the companies located globally including Polish Dataset. The service provider has publications based on financial information, macroeconomics and companies news, industrial reports, etc. The data is suitable for our research as it holds economic indicators as attributes (synthetic features) [29] and has around 10,000 records of the companies that are evaluated from the period 2007 to 2013 in 5 different time frames. Based on the collected it is divided into 5 classification cases which depend on the forecasting. We have chosen the data for the 2nd year of forecasting and the class label indicates the bankruptcy status after 4 years.The below table gives us a summary of the data.

3.2.1 Data Exploration

We have selected the dataset and have initial understanding of the variables and data type. In this step we will explore and understand the data and the best way to convey is by visualization. The visual representation of data will help us to comprehend the structure of the data, the sparsity and correlation within the dataset, missing values, data imbalance and the way the values are scattered. Below we have given below the findings of the Polish bankruptcy dataset.

Fig 3.3 shows the variables of different data type. The variables are mostly in object and float format and need to converted in to float and maintain uniformity in the data preparation.

ID	Description	ID	Description
X1	net profit / total assets	X33	operating expenses / short-term liabilities
X2	total liabilities / total assets	X34	operating expenses / total liabilities
X3	working capital / total assets	X35	profit on sales / total assets
X4	current assets / short-term liabilities	X36	total sales / total assets
X5	[(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365,	X37	(current assets - inventories) / long-term liabilities
X6	retained earnings / total assets	X38	constant capital / total assets
X7	EBIT / total assets	X39	profit on sales / sales
X8	book value of equity / total liabilities	X40	(current assets - inventory - receivables) / short-term liabilities
X9	sales / total assets	X41	total liabilities / ((profit on operating activities + depreciation) * (12/365))
X10	equity / total assets	X42	profit on operating activities / sales
X11	(gross profit + extraordinary items + financial expenses) / total assets	X43	rotation receivables + inventory turnover in days
X12	gross profit / short-term liabilities	X44	(receivables * 365) / sales
X13	(gross profit + depreciation) / sales	X45	net profit / inventory
X14	(gross profit + interest) / total assets	X46	(current assets - inventory) / short-term liabilities
X15	(total liabilities * 365) / (gross profit + depreciation)	X47	(inventory * 365) / cost of products sold
X16	(gross profit + depreciation) / total liabilities	X48	EBITDA (profit on operating activities - depreciation) / total assets
X17	total assets / total liabilities	X49	EBITDA (profit on operating activities - depreciation) / sales
X18	gross profit / total assets	X50	current assets / total liabilities
X19	gross profit / sales	X51	short-term liabilities / total assets
X20	(inventory * 365) / sales	X52	(short-term liabilities * 365) / cost of products sold)
X21	sales (n) / sales (n-1)	X53	equity / fixed assets
X22	profit on operating activities / total assets	X54	constant capital / fixed assets
X23	net profit / sales	X55	working capital
X24	gross profit (in 3 years) / total assets	X56	(sales - cost of products sold) / sales
X25	(equity - share capital) / total assets	X57	(current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation)
X26	(net profit + depreciation) / total liabilities	X58	total costs /total sales
X27	profit on operating activities / financial expenses	X59	long-term liabilities / equity
X28	working capital / fixed assets	X60	sales / inventory
X29	logarithm of total assets	X61	sales / receivables
X30	(total liabilities - cash) / sales	X62	(short-term liabilities *365) / sales
X31	(gross profit + interest) / sales	X63	sales / short-term liabilities
X32	(current liabilities * 365) / cost of products sold	X64	sales / fixed assets

FIGURE 3.2: Data Description

```
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 3.3: Data type of variables

The below fig3.4 bar chart of missing values in the attributes, we see that attribute 37 attribute 21 has more number of missing values.

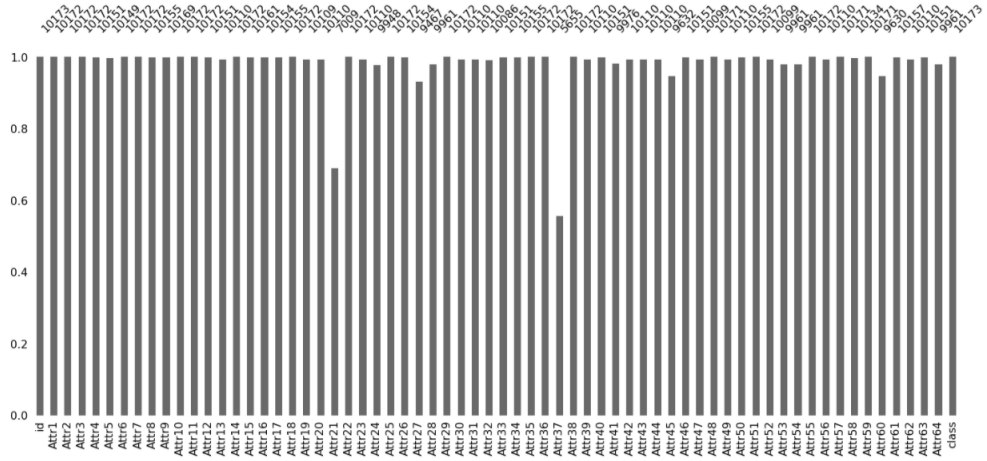

FIGURE 3.4: Missing values Bar chart

3.3 Data Pre-processing and Transformation

Now that we have an understanding of the data and we proceed to our next step i.e.data preparation. Preparation of the data is an essential part of the data mining process, after interpreting the given data the necessary steps that need to be taken for modeling is determined in this step. The financial datasets often hold a large amount

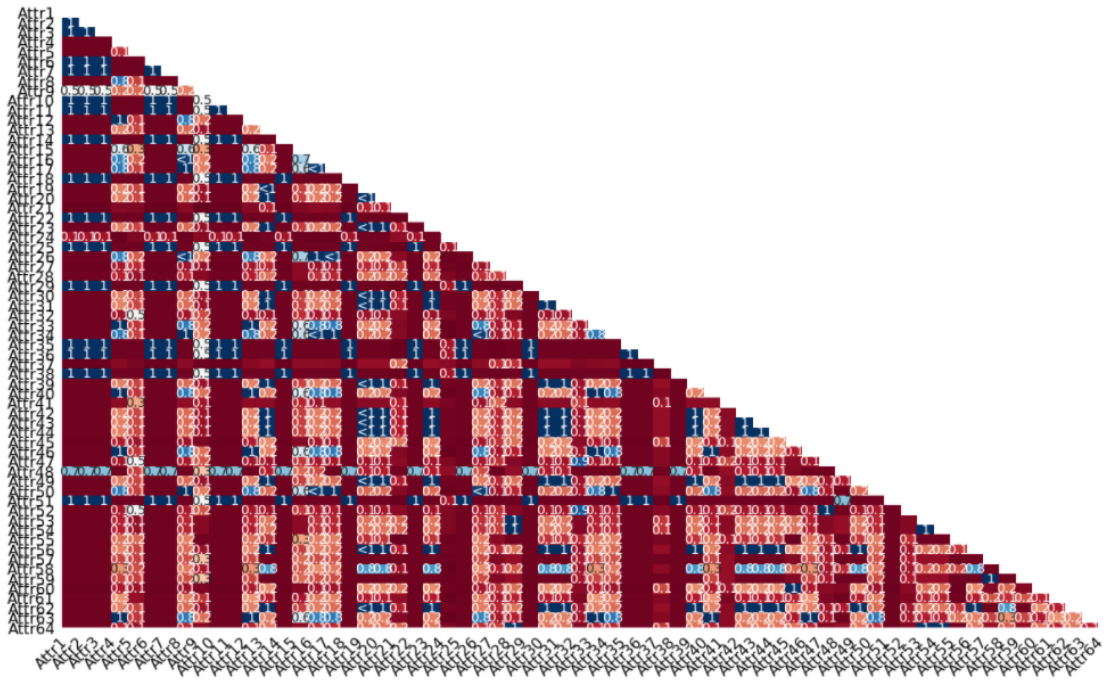

FIGURE 3.5: Heatmap for Missing values

of data and Big data tends to have noise in form of special characters, missing values, blank spaces, different data formats, etc, and have effects on the performance of the model. It is important to handle the noise in the preparation stage. While handling such problems in the data we must be careful in order to avoid affecting the essence of the data terribly as this results in poor predictions. The selection of variables is also determined in this process, as to selecting the predictive indicators for our bankruptcy data. The variables that have a high impact on the outcome can be said to have better prediction outcomes. We will try to keep the records of the data intact as more the data we have, the better prediction can be expected. The steps taken for data preparation are mentioned below.

- Converting file type: The earlier file format extracted from the UCI Machine Learning repository is of type ARFF format. We will convert the file to CSV file format for our model.

- Creating Data frames: We will store the data in pandas data frames for better performing operations.

Dataset	Number of Instances	Bankrupt instances	Non Bankrupt instances	Percentage of minority class samples
2nd year	10173	400	9773	3.93%

TABLE 3.1: Data summary

Dataset Characteristics	Multivariate
Number of Features	64
Has missing values?	Yes
Associated task	Classification

TABLE 3.2: Dataset summary

- Changing the data type of the variables: After observing the data types in the previous process we will convert all numerical values to float data type.

3.3.1 Handling Missing values

The presence of blank values can be a problem for a few of our models and we will try to solve this by using the imputation method. To deal with it we fill the blank values with NaN placeholders in the data frames. We do not want to remove the records with missing values but we do not want to lose data points that may have valuable information. There are different approaches to deal with missing data, in this study, we will be using MICE (Multivariate Imputation by Chained Equation) and it is a principled technique for imputations and offers better improvements than single imputation techniques[4]. Since our data has multivariate characteristics this technique works well with multiple imputation procedures. "MICE is particular multiple imputation techniques [8]. We have missing data are Missing at Random (MAR) and to handle this MICE is used as it performs well with data that are MAR[4].

3.3.2 Handling Data Imbalance

The class imbalance is an issue when it comes to binary classification problems [10]. The meaning of having class imbalance is having more number of instances for a certain class. We have major class imbalance in our data as shown in the below table.

The class imbalance will lean our model in predicting certain classes more often than the actual value and may affect the model's outcome[reference from analyicsvidya] and we may say the model is biased and inaccurate for the developed model. To overcome class imbalance we can either perform oversampling or undersampling. Since we focus on utilizing most of our data, we will be performing oversampling. Oversampling can be performed with different techniques. In this study, we will be using the SMOTE technique to handle class imbalance. Smote creates synthetic minority class samples and performs better for high dimensional data[7], in our case the economic indicators are high dimensional.

3.3.3 K-fold Cross validation

The data is cleaned and prepared for our classification models, to provide the data to our models we will use K-fold cross validation technique to split our training data and test data. The data is split into 'K' number of groups and to begin with the procedure of K-fold, the data is shuffled at random and split into K- groups, and in each iteration for each unique group is taken as the test data and the remaining is taken as the training data. This will help to understand how the model is performing on the unseen data. K-fold CV is a popular method and results in less biased and estimates the model's skill to predict the outcome[15].

3.4 Modelling Approach

In this section we will discuss about the different models that are to be implemented. This process is important and very crucial, we will be implementing our models on the pre-processed data and evaluate them. The details of the algorithms are mentioned below.

3.4.1 Gaussian Naïve Bayes

Gaussian Naïve Bayes is a variant of Naïve Bayes, it follows Gaussian normal distribution. Naïve Bayes algorithm is a machine learning algorithm for classification. It is

used for spam filtration, text classification etc. The occurrence of the features are independent hence the name 'naive' . Bayes is the base for the Naïve Bayes Algorithm. Bayes theorem is given as:

- P(A | B) = P(B | A) P(A) / P(B)
- Where, P(A | B) is probability of occurrence of event A given the event B is true.
- P(A) and P(B) are probabilities of occurrence of event A and event B.
- P(B | A) is probability of the occurrence of event B given event A is true.

In a classification problem with features and classes say 'C', the main aim of the Naïve Bayes algorithm is to calculate the conditional probability of an object with a feature vector that belong to the class 'C'. In Gaussian Naïve Bayes classifier, we assume that feature values associated with each class is distributed according to a normal distribution [13]. The likelihood of the features assumed to be:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

3.4.2 Logistic Regression

Logistic regression is widely used to classify the data into two classes. Regression analysis is done when the target/dependent variable is categorical or binary. Like all the regression analysis, logistic regression is a predictive analysis. It gives us the relationship between one dependent variable which is categorical and one or more independent variables which are nominal or ordinal. Logistic regression is used in many binary class prediction example such as spam detection where we can determine whether the mail is a spam or not. Logistic regression is also used in the field of financial areas including transaction fraud detection, financial stress prediction etc. In logistic regression the dependent variable 'y' can be of range 0 to 1. The function used at the core of the model is sigmoid function also called as logistic function.

$$S(x) = \frac{1}{1 + e^- x}$$

Where 'e' is the base of natural logarithm and 'x' is the actual numerical value. Logistic regression is similar to linear regression, the difference is the output value being modeled is a binary value (0/1) rather than numerical value.

3.4.3 Decision Tree

A decision tree is a supervised machine learning algorithm and can be seen as classifier which divides the data in form of partition or groups. A tree has many analogies in life and is influenced in wide areas of machine learning, covering both classification and regression problems.[19] Decision tree can be built on several algorithms, some of them are ID3, CART and C4.5 etc. Below mentioned decision tree diagram indicates tree consisting of multiple nodes and forms a rooted tree, and is directed by the node called as the Root node. The root node does not have any incoming edges. The nodes with outgoing edges are decision nodes and the rest of the nodes are leaves. The leaf node can be a predicted value or probability value. While building a decision tree we specify the condition on the decision node in order to classify the data into subsets and make decisions to split the decision node into leaves nodes or terminal node. The algorithms like ID3, CART, and C4.5 decides the splitting of the nodes based on information gain, Gini index, and gain ratio respectively. So different algorithms use different metrics to split the data. The structure of the tree help us to form a series of conditions applied on the decision node and help us find the hidden data in the future.

$$Gini = 1 - \sum(P_i)2$$

for i=1 to number of classes

We are using CART algorithm, it can do classification and regression problems and sci-kit learn python library uses the same library used in the CART. As we discussed different metrics for making decision points, CART uses Gini impurity and below in the formulae mentioned.

3.4.4 Random Forest

Random forest is one of the most popular and powerful supervised machine learning algorithm, it can perform both regression and classification problems [18]. As the name suggests, this algorithm uses a number of decision trees to create a forest. The robustness of the forest depends on the number of trees. Generally, more trees means a more robust forest. Similarly, in a random forest classifier, the accuracy of the results is proportionally related to the number of trees. More trees will give better accuracy in the results. The advantages Random forest takes care of the missing values in the data and maintains the accuracy. It does not overfit the model and can handle large amount of data with high dimensionality. Random forest is utilized in various sectors including banking, medical and stock market. In Banking it can be used to classify the loyal customers and fraud customers of a bank. In Medical sector it can be used to identify the correct combination of components to validate the medicine and can also help in identifying disease by analyzing the patient's records. In stock market it is used for identifying the behavior of the stock and calculate the profit and loss of a particular stock. Pseudocode:

- Assume N number of cases in training set. Where the sample of the N cases is selected at random but with replacement.
- If there are X features, a number x < X is specified at the node, to split the nodes using the best split on these x. The value of x is constant while building the forest.
- Each tree is grown to the full extent without pruning.
- Predict new data by aggregating the prediction of n trees (in classification: majority voting among the trees).

3.4.5 XGBoost

XGboost is a ensemble boosting machine learning technique which uses gradient boosting as its framework. It is known for having better performance and accuracy and can be used for classification, regression, ranking and user defined prediction problems.

XGBoost works very well with the small-medium structured data. XGBoost when applied to the financial data especially in bankruptcy prediction, the econometric features show high variance for small sample sets. It indicates that most of the values of some indicators are hidden but there are some companies that are described by relatively high/small values of those features. As an outcome the neural network and logistic regression when used as base learners for gradient based models, perform poorly and make incorrect prediction. Whereas in ensemble learning the order of the feature values is taken rather than values itself [29]. This ensemble boosted learning algorithms has received a lot of appreciation in the Kaggle competition [29]. XGBoost is a combination of software and hardware optimization technique that yields high output. XGBoost improves upon the base GBM framework through systems optimization and algorithmic enhancements.

3.4.6 Light GBM

LightGBM is another powerful ensemble boosting technique developed by Microsoft and uses gradient boosting framework [23]. LightGBM is claimed to be much more faster and efficient than xgboost though both use the same underlying algorithm that is gradient boosting decision trees(GBDT) they introduce different methods to improve the training and efficiency. The key difference between GBDT and LightGBM is that GBDT grows the decision tree level-wise whereas LightGBM grows the trees leaf- wise and splitting the data. The leaf-wise splitting helps to reduce most of the loss as boosting algorithms reduces the loss in each iteration [25]. The leaf-wise training can construct trees similar to the level-wise but it is not vice versa. The leaf wise training has higher chances of overfitting but can be taken care of by tuning the parameters. Light GBM uses two novel techniques GOSS(Gradient One Side Sampling) and EFB (Exclusive Feature Bundling) which makes it light and efficient. We will be implementing the GOSS method in one of our case studies. Since the gradient boosting learns with computation of information gain, GOSS is very effective in finding the information gain with fewer data samples [23], it takes all the high gradient data and does random sampling over the data with small gradient. LightGBM is not used in

bankruptcy data prediction and hence we will evaluate the model based on the variables provided to it and see the results.

Chapter 4

Design Specification

In the figure[4.1] the architecture shows the flow of our research. We begin with gathering the data then preparing the data with pre-processing steps such as handling missing values with MICE imputation technique and dealing with class imbalance by oversampling the data with SMOTE technique. Further splitting the data into train and test and preparing K-folds. Next step is to train our different classifier models with the training data and predict the results based on test data set and evaluating the accuracy of our models.

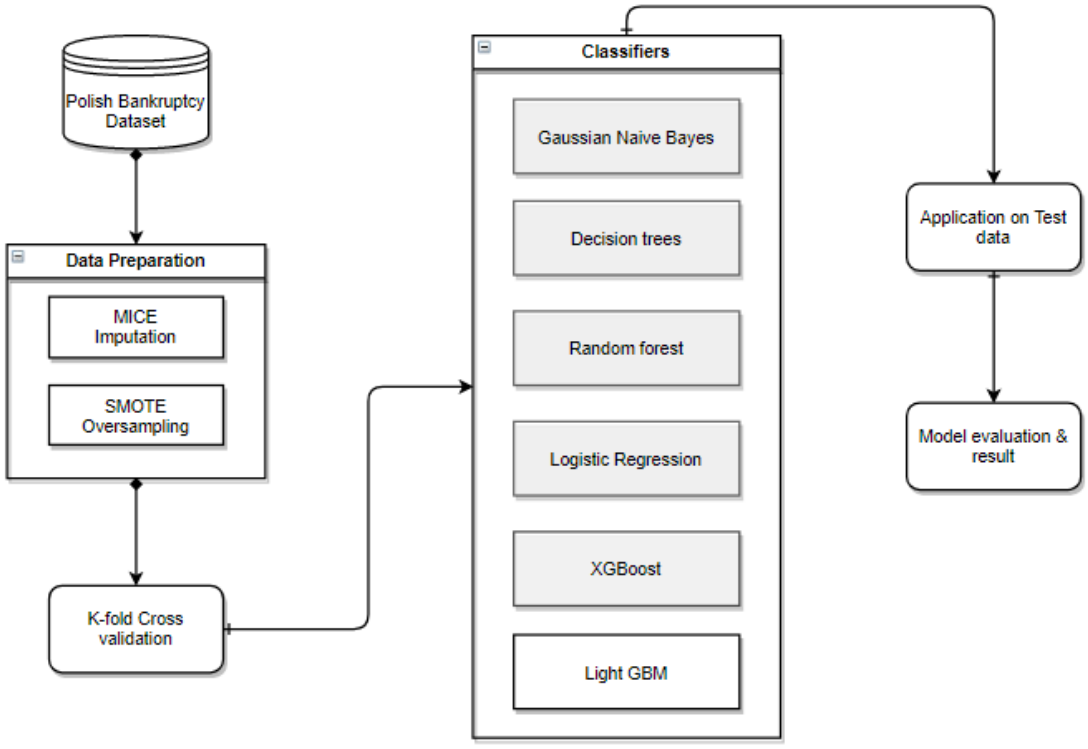

FIGURE 4.1: Architecture

Chapter 5

Implementation

In this section we will explain the implementation of our models proposed for bankruptcy prediction in detail and the techniques used for data preparation. Also, it describes the different tasks that are performed using LightGBM model and its implementation with the primary data. The entire implementation was executed in Python language version 3.7.4 and carried out using Jupyter Notebook as Integrated Development Environment(IDE). We chose to execute our research work in Python language as it easy for implementation and the provides with extensible nature and huge amount of libraries that makes the implementation process hassle free. Python language has a huge community and there are many packages available to implement pre-processing steps and model training.

5.1 Implementation of the models with pre-processed data

Now, we will explain the step wise implementation of our proposed models for classification prediction of bankruptcy data. We begin with the data sourced for our research, the data consist of 5 years forecasting period financial data, we will using this data and have an understanding of the data. The data file is in ARFF file format and using Python's pandas library we will load and store the data in data frames. After looking the data frames we see the structure of the data, it consists of 65 columns in which 64 of them are financial ratios and the last column is categorical which contains binary values. For '0' it indicates that the instances of a particular company are not bankrupt and '1' is for bankrupt instances. This type of data is suitable for binary classification. The 64 attributes are our feature variables and the last column is our target

variable. We further start the pre-processing of the data by looking at the data type of the variables. The attributes are in integer format and are changed to float format and the class variable to integer format. We visualize the data structure and figure out the missing values present in the data and the imbalance of class variable instances. We notice that we have many instances with missing values, first we drop the instances with missing values and perform MICE imputation technique to manage the data loss. Now, we check the class imbalance by taking a count of the class variable and notice the datasets are high imbalanced and for training our model with both the instances with adequate sample of data, we perform oversampling using SMOTE technique. The imputed data frames are now taken and oversampling is performed and generate even number of over sampled data for both the classes. Next step we prepare to split our data for training and testing our proposed models. For splitting the data, we use K fold cross validation technique and split the data into K-folds of training and test data set. We set the k value to 'n' and n is the size of the dataset and gives opportunity for every test set to be a hold out set[15]. Further we applied our resampled data set on 6 different classification models, and the models implemented are Gaussian Naive Bayes Logistic Regression, Decision Tree, Random forest, Extreme Gradient Boosting and Light Gradient Boosting. All of the models are using default parameters as a binary classifier and are available in Python's sklearn library. To evaluate each of the models we calculate the accuracy by using metrics like true positive, false negative, false positive and true negatives and append the mean of each iteration of the K-folds to these metrics. The model is evaluated based on accuracy, precision and recall values for each of the years.

5.2 Implementation of LightGBM classifier and XGBoost with raw data

In this section we will be describing the steps taken to implement the XGB and LightGBM clasifier with the raw dataset. The data which we recieved originally is taken in to consideration, we will be applying the models to the 2nd year of the forecasting period to do a comparision of both the models. The dataset is in ARFF format and

is converted to CSV format as per our model requirement. The csv data is stored in the pandas data frames and the data type of the variables are changed to float format in the similar way as the previous implementation. The missingv values present in the dataset are replaced with NaN as place holder. As both the models can be implemented on the data with missing values. We split the data in test and train using sklearn's train & test split package. We set the features and the target variables and apply both the models on the train and test data. For evaluation of the models we use AUC and ROC curve, we measure the classifiers ability to distinguish the class labels. We also calculate the accuracy of the models by calculating the difference between actual and the predicted values.

5.3 Implementation of Light GBM model with GOSS and Tuning of hyper parameter

In this part we will be explaning the implementation of Light GBM with gradient one sided sampling(GOSS). As the data is high dimensional and GOSS excels in such classification problems. We begin the primary data obtained from the source we organise the data and store them in pandas data frames, then we convert the data types and prepare the data set for training and testing. We begin with importing the lightgbm library in python, splitting the dataset in to training and validation sets and wrap them in Light GBM's Datasets. Next we provide the core parameters for the training of the gradient boosting machine, we refer the Light GBM documentation github for hyper parameter tuning [17].

Below are the list of some the core-parameters we applied :

- boosting_type: 'goss',
  GBM type: Based on the method for gradient boosting machine to be applied.
  Example: gradient boosted decision tree, rf (random forest), dart, goss.

- objective: 'binary',
  We specify the optimization object depending on the classification problem.
  Example: binary, regression, multiple class variables.

- 'learning_rate': 0.05,
  This parameter controls the gradient descent learning or shrinkage rate,and the step size.

- num_leaves : 31,
  number of trees we want to build, the best number of leaves to have is 42 for better training.

- nthread : 4,
  number of threads to use for LightGBM, best set to number of actual cores.

- metric: auc
  It is a metric to calculate during validation: area under curve (auc).

Now we can train the Gradient One-sided Sampling using LightGBM. We have wraped the training set in a function and train the GOSS. We will plot the training and the validation results, and to evaluate the model we plot AUC-ROC graph and calculate the AUC score for our model. Few of the advance parameters are mentioned below that help us to improve the auc score of our GOSS based gradient boosting machine.

- max_depth: 5
  we specify the depth of the tree, shallow trees avoid over fitting of the model.

- min_child_samples: 21
  We can specify the number of samples at a leaf node.

- min_split_gain: 0
  to perform split with the minimal loss gain

- lambda_l1: 0.5
  L1 regularization

- lambda_l2:
  L2 regularization

- max_bin: 700
  The number of bins to create are specified and larger bins improves the accuracy

Note we are still using the raw data and testing the model's accuracy. The Light GBM model is highly sensitive to hyper parameter tuning and must be carefully set to get better prediction results.

Chapter 6

Evaluation

The main aim of our research is to evaluate the best model for bankruptcy prediction, and we use binary classification to solve this problem. To say which classification model performed best or had better results than other classifiers is to compare the evaluation metrics of these models. We are evaluating our models based on metrics such as accuracy, precision, sensitivity, specificity, auc_score and AUC-ROC curve. These metrics are good for evaluating a binary classification models as stated in the blog article [22]. The metrics are derived from the confusion matrix values that holds true positive, false negative, true negative and false positive and are used for the computation of the specified metrics. We can interpret the meaning of these metrics in terms of bankruptcy prediction. The Accuracy metrics tells us about the overall classification of correct predictions. The Precision metrics can be interpreted as to how often the model correctly predicts bankrupt companies. AUC score metrics explains how well the model distinguishes the binary class and AUC-ROC graph plot explains the performance of the classifier on the graph using a curve. The AUC ROC curve are plot using True Positive Rate(TPR) against False Positive Rate(FPR) at various threshold values. Sensitivity (true positive) tells us how often the classifier correctly predicts the outcome when the company is bankrupt and Specificity(true negative) tells how often the classifier predicts the outcome correctly when the company is non-bankrupt. We have used cross validation so each dataset a confusion matrix is constructed with the mean value so that the models are unbiased. The calculation of metrics is done by using below given equations.

$$Accuracy = \frac{TP + TF}{TP + TF + FP + FN}$$

Models	Accuracy	Precision	Sensitivity	Specificity
Gaussian Naïve Bayes	50%	50%	98%	3%
Logistic Regression	44%	44%	43%	44%
Random Forest	92%	91%	93%	90%
Decision tress	91%	90%	93%	89%
XGBoost	97.4%	98%	97%	97%
LightGBM	98.6%	98%	98%	98%

TABLE 6.1: Overall Performance of the models

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

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

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{TN + FP}$$

6.1 Case study 1: Comparison of overall models with Imputation and Balanced data

We have evaluated the result of all the base models that are built on the balanced data and with imputed data and are given in the table below. The table[6.1] displays the evaluation metrics and the models accuracy. We can infer from the table[6.1] that Logistic regression and Gaussian Naive Bayes show poor performance having the lowest accuracy although Gaussian Naive Bayes has high recall(sensitivity) value but has very low specificity. The Random Forest classifier and the Decision trees model have the almost same accuracy with 92% 91% respectively. The LightGBM shows superiority in overall performance and is a slight edge over the accuracy of the XGBoost model.

Model	AUC score train	AUC score test	Training time
LightGBM	1.0	0.94	0.82 sec
XGBoost	1.0	0.92	5.35 sec

TABLE 6.2: Performance of the XGBoost and LightGBM models

6.2 Case study 2: Comparision of XGB and LightGBM classifiers with Unbalanced data and Missing values

In the second case study, we have evaluated the model performance using the AUC-ROC metrics curve, AUC test, and train score and training time of XGBoost and LightGBM. Both the models are trained using data with a feature missing values and highly imbalanced class data. We display the AUC-ROC curve for both the models in the fig[7.3] and fig[6.1]. We have compared the models in the table[**??**] and interpret that LightGBM has a higher AUC score and its training time is very less compared to the XGBoost model. Both the model predicts the training data correctly for all the instances whereas when the models are tested with unknown data they have AUC score of 94% for Light GBM and 92% for XGBoost. In the fig[7.3] the curve seems to dip at True positive rate between 0.8 to 1.0 and that AUC-ROC curve for LightGBM fig[6.1] the curve looks good covering the maximum area under the curve.

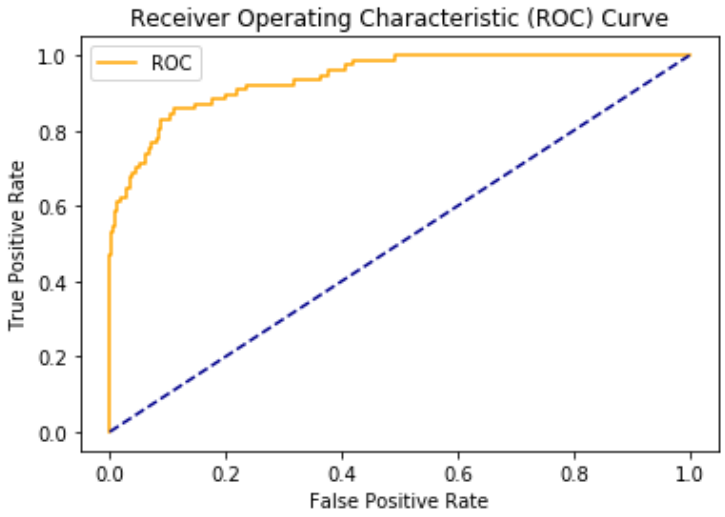

FIGURE 6.1: AUC-ROC curve for LightGBM model

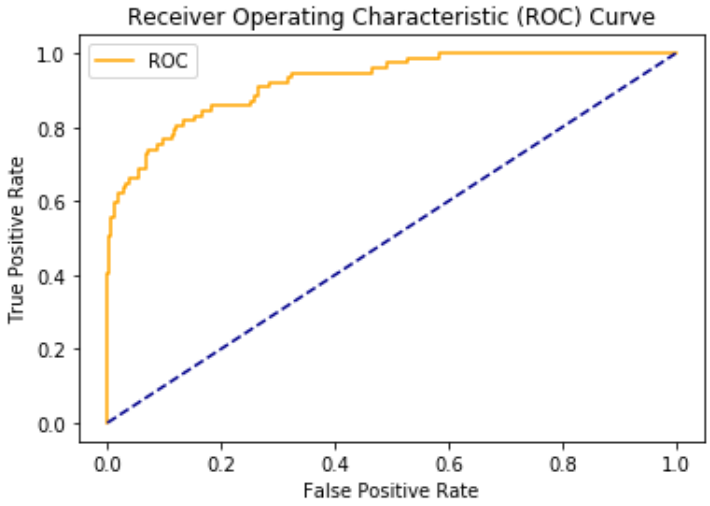

FIGURE 6.2: AUC-ROC curve for XGBoost model

6.3 Case study 3: Model Performance with LightGBM: GOSS and tuned parameters

To evaluate the performance of the LightGBM GOSS model we are using the AUC ROC curve, AUC score, and Accuracy metrics. The LightGBM model is trained with raw data that we obtained from the data source, the data has missing values and is highly imbalance, we evaluate our model as to how the LightGBM model performs with the GOSS technique. The model has good accuracy of 97% and the AUC score is around 0.927, the fig[6.3] displays the AUC ROC curve for our model. Although the best accuracy was obtained using advanced hyper parameter tuning. The model is very sensitive with the hyper tuning and displays changes in performance with the slightest change in the parameter. This is the best accuracy we could reach along with the hyper tuning of our LGB model. In the fig[6.4] we see the bar chart of attributes that have high impact on the LGB model's preformance.

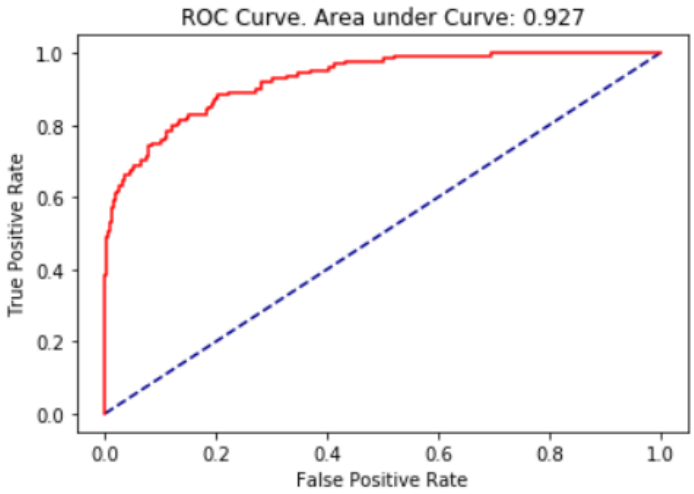

FIGURE 6.3: AUC Model3

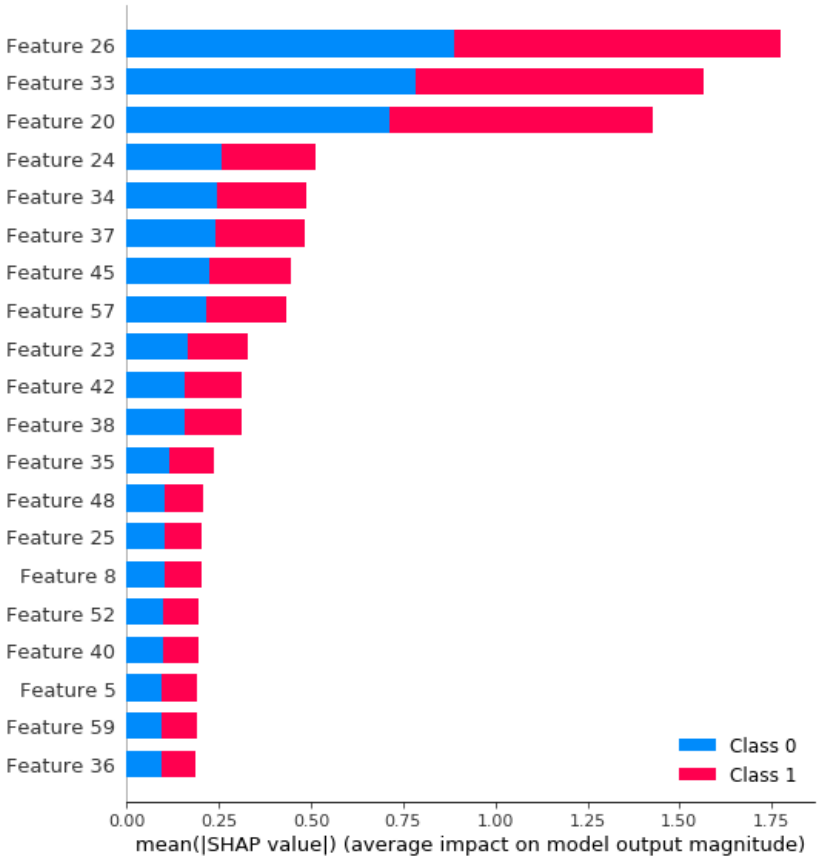

FIGURE 6.4: Features with high impact on the LigtGBM model

Chapter 7

Discussion

The Accuracy metric of the models tells us how accurately the model classifies the instances accurately, the overall performance of the classification model. The Sensitivity metrics helps us understand how accurately our model predicts the bankrupt instances correctly. This metric is important as we do not want our model to miss out on indicating that the company is bankrupt, and understanding this metric will help to alert the concerned person. Specificity is another important metric, it tells how correctly our model predicts the non-bankrupt companies, the classification model's accuracy can also be determined by looking at the Sensitivity and Specificity metric and comparing the values with each other. In the table[6.1] we see for Gaussian Naive Bayes classifier the model poorly detects the company which is nonbankrupt and classifies almost all the instances as bankrupt, we can say that the model overfits a single class. The overall performance of traditional models Logistic regression and Gaussian Naive Bayes is very low. The Decision trees and Randomly forest performed better than traditional models, but while predicting the bankruptcy of the company we need to have more efficient models. We can see the Specificity metric for the Decision tree and Random forest is less than recall metrics, both the model tend to distinguish bankrupt class label more often than nonbankrupt class. However, if we look at the results for XGBoost and LightGBM they are almost the same with a 1 percent difference in the accuracy as shown in the table[6.1].

The table[6.1] tells us the comparison results of models with default parameters and are trained on the data which is balanced and complete. We are interested in the working and performance of Light GBM on bankruptcy data and the best ensemble learning algorithm is that is widely used is XGBoost, so we made a comparison of

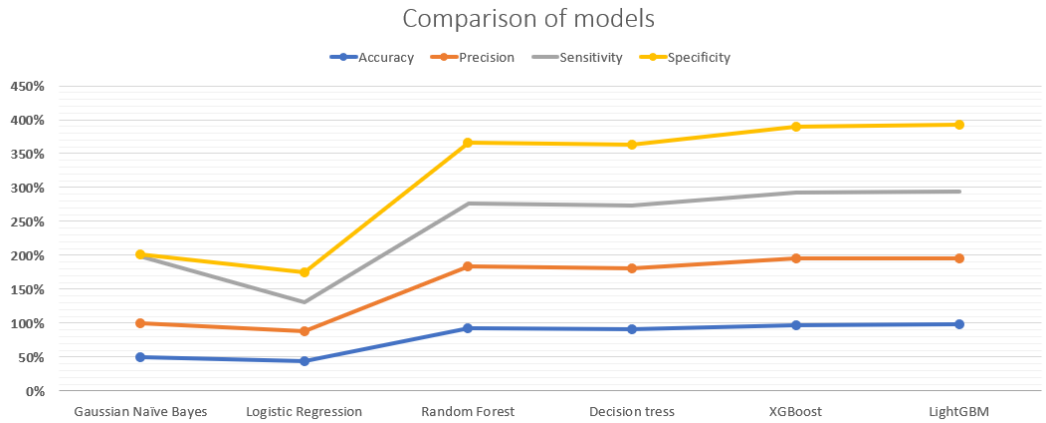

FIGURE 7.1: Comparison of overall models

both, and the results are displayed in the table[6.2]. As we had discussed the importance of AUC ROC, we evaluated both the models, and the results show that the LGB classifier is provided good AUC score than XGBoost. Also, is if we note the training time the model is faster than XGBoost[23]. The data we used was relevantly small but in a broader scenario when the financial big data is used for prediction, the LightGBM model will be ready for execution of the unknown data in no time. The AUC ROC curve is plotted using Sensitivity and False Positive rate, it is a great efficiency metrics for data that is highly skewed sample data. The fig[6.1] displays the AUC ROC curve of the LightGBM model.

Now we discussed the two case studies results, in the last case study the Accuracy and AUC ROC is used as evaluation metrics, we used gradient one-sided sampled boosted tree. Even with the small sampled dataset the model can achieve an accuracy of 96 percent and it even contains missing values which the Gradient boosting framework takes care of in parameter tuning, secondly, the data is highly imbalanced, with these factors taking under consideration the LightGBM model GOSS technique provides significant results. Although to achieve this accuracy we had to tune the hyperparameters, the method is sensitive to parameter tuning.

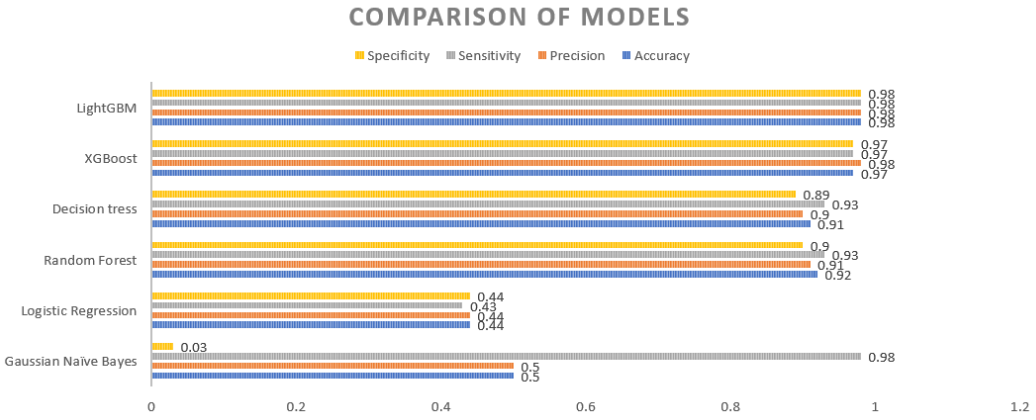

FIGURE 7.2: Bar chart Comparison of overall models

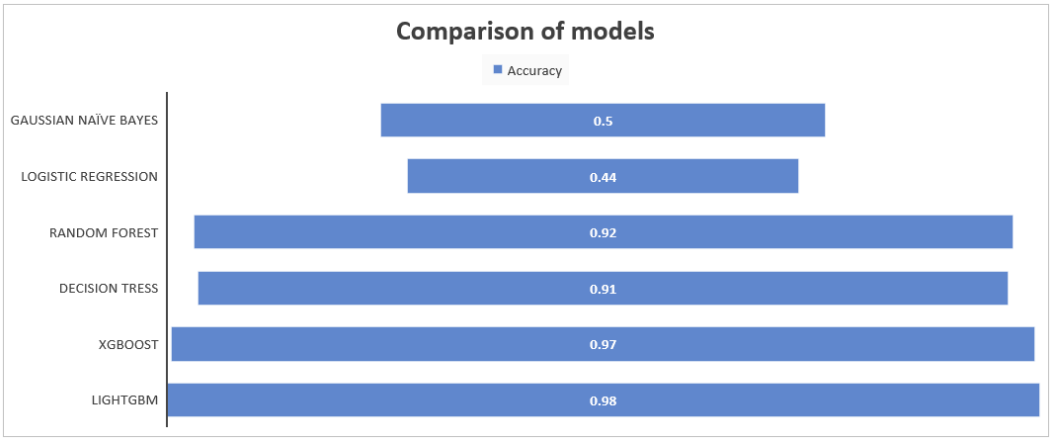

FIGURE 7.3: Funnel chart of overall models accuracy

Chapter 8

Conclusion & Future Scope

In this section, we will discuss the work done in the thesis so far. We have done a comprehensive analysis of 6 different models namely Logistic Regression, Gaussian Naive Bayes, Decision trees, Random forest, Extreme Gradient Boosting, and Light Gradient Boosting. A comparison of all the models' performance is made and the newly introduced LightGBM model performs better than the traditional models. The performance of the LightGBM model is evaluated with two types of data inputs i.e. one with balanced data and comparison of the results with the other reference models and the second with raw imbalanced data with missing values and the model's performance on both the dataset outperforms the rest of the models. A close comparison of the XGBoost and LigtGBM model is done and the LightGBM gives similar results even with the highly skewed data. Also in this work, a LightGBM model was built on the GOSS technique to see how well the model performs with highly imbalanced data with a small sample size, and the maximum accuracy is achieved by tuning the hyperparameters of the model. The training time of LightGBM proves to be much faster and the classification of the class labels is done much accurately. The traditional statistical models Logistic Regression and Gaussian Naive Bayes display poor performance although to be fair the experimental setup of these models was not to their optimal level. The results for the Decision tree and Random forest are similar and better results can be expected with parametric implementation. These models need the data to balanced and without missing values to perform. Whereas gradient boosting models handle missing values and correctly classifies the class labels. We notice the attributes (net profit+depreciation)/total liabilities, operating expenses/short-term liabilities, (inventory*365)/sales seem to have a high impact on the LightGBM

model and that these attributes can be counted as key performing indicators in the bankruptcy prediction models. Further, we can say that the data used in the experimental setup can be improved, the dataset holds lots of missing values and has a major class imbalance, in future the implementation of our LightGBM model may bring out some more insights and well accuracy as a bankruptcy prediction model. The models also are sensitive to hyperparameters and further work can be done in finding the right parametric values to achieve even better results. Next, we can also check the performance of the LightGBM model compared to other models that have been very well implemented to their optimal performance and do fair modeling of the reference models.

Bibliography

[1] Edward Altman. "I., 1968, Financial ratios, discriminant analysis and the prediction of corporate bankruptcy". In: *Journal of Finance* 23.4 (1968), pp. 589–609.

[2] Edward I Altman. "Predicting financial distress of companies: revisiting the Z-score and ZETA® models". In: *Handbook of research methods and applications in empirical finance*. Edward Elgar Publishing, 2013.

[3] Ana Azevedo and Manuel Santos. "KDD, semma and CRISP-DM: A parallel overview". In: Jan. 2008, pp. 182–185.

[4] Melissa Azur et al. "Multiple Imputation by Chained Equations: What is it and how does it work?" In: *International journal of methods in psychiatric research* 20 (Mar. 2011), pp. 40–9. DOI: 10.1002/mpr.329.

[5] Barbro Back, Teija Laitinen, and Kaisa Sere. "Neural networks and genetic algorithms for bankruptcy predictions". In: *Expert Systems with Applications* 11 (Dec. 1996), pp. 407–413. DOI: 10.1016/S0957-4174(96)00055-3.

[6] Flavio Barboza, Herbert Kimura, and Edward Altman. "Machine Learning Models and Bankruptcy Prediction". In: *Expert Systems with Applications* 83 (Apr. 2017). DOI: 10.1016/j.eswa.2017.04.006.

[7] Rok Blagus and Lara Lusa. "SMOTE for High-Dimensional Class-Imbalanced Data". In: *BMC bioinformatics* 14 (Mar. 2013), p. 106. DOI: 10.1186/1471-2105-14-106.

[8] Stef Buuren and Catharina Groothuis-Oudshoorn. "MICE: Multivariate Imputation by Chained Equations in R". In: *Journal of Statistical Software* 45 (Dec. 2011). DOI: 10.18637/jss.v045.i03.

[9] Arindam Chaudhuri. "Bankruptcy Prediction Using Bayesian, Hazard, Mixed Logit and Rough Bayesian Models: A Comparative Analysis". In: *Computer and Information Science* 6 (Apr. 2013). DOI: 10.5539/cis.v6n2p103.

[10] S.M.A. Elrahman and A. Abraham. "A review of class imbalance problem". In: *J. Netw. Innov. Comput.* 1 (Jan. 2013), pp. 332–340.

[11] Philippe du Jardin. "Bankruptcy prediction models: How to choose the most relevant variables?" In: *Bankers, Markets Investors* (Jan. 2009), 39–46.

[12] Tomasz Korol. "Dynamic Bankruptcy Prediction Models for European Enterprises". In: *Journal of Risk and Financial Management* 12 (Dec. 2019), p. 185. DOI: 10.3390/jrfm12040185.

[13] Nikhil Kumar. *Gaussian Naive Bayes Classifier*. https://www.geeksforgeeks.org/naive-bayes-classifiers/.

[14] Tuong Le et al. "A Cluster-Based Boosting Algorithm for Bankruptcy Prediction in a Highly Imbalanced Dataset". In: *Symmetry* 10 (July 2018), p. 250. DOI: 10.3390/sym10070250.

[15] Machine Learning Mastery. *K FOLD*. https://machinelearningmastery.com/k-fold-cross-validation.

[16] Qi Meng. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". In: Apr. 2018.

[17] LightGBM Microsoft. *LGBM documentation*. https://lightgbm.readthedocs.io/en/latest.

[18] Avinash Navlani. *Random Forest*. https://www.datacamp.com/community/tutorials/random-forests-classifier-python.

[19] Harsh Patel and Purvi Prajapati. "Study and Analysis of Decision Tree Based Classification Algorithms". In: *International Journal of Computer Sciences and Engineering* 6 (Oct. 2018), pp. 74–78. DOI: 10.26438/ijcse/v6i10.7478.

[20] Tonatiuh Peña, Serafin Martinez-Jaramillo, and Bolanle Abudu. "Bankruptcy Prediction: A Comparison of Some Statistical and Machine Learning Techniques". In: (Jan. 2011), pp. 109–131. DOI: 10.1007/978-3-642-16943-4_6.

[21] F.M. Tseng and Hu Yi-Chung. "Comparing four bankruptcy prediction models: Logit, quadratic interval logit, neural and fuzzy neural networks". In: *Expert Systems with Applications* 37 (Mar. 2010), pp. 1846–1853. DOI: 10.1016/j.eswa.2009.07.081.

[22] Analytics Vidhya. *AUC ROC Analytics vidya*. https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning.

[23] Nanxi Wang. "Bankruptcy Prediction Using Machine Learning". In: *Journal of Mathematical Finance* 07 (Jan. 2017), pp. 908–918. DOI: 10.4236/jmf.2017.74049.

[24] Wikipedia. *Bankruptcy prediction*. https://en.wikipedia.org/wiki/Bankruptcy_prediction.

[25] Joanna Wyrobek and Krzysztof Kluza. "Efficiency of Gradient Boosting Decision Trees Technique in Polish Companies' Bankruptcy Prediction: Part III". In: Jan. 2019, pp. 24–35. ISBN: 978-3-319-99992-0. DOI: 10.1007/978-3-319-99993-7_3.

[26] Qi Yu et al. "Bankruptcy Prediction with Missing Data". English. In: *SDM ICDM The International Conference on Data Mining (ICDM) ICDM*. VK: airc. 2011, pp. 279–285.

[27] Yuri Zelenkov. "Bankruptcy Prediction Using Survival Analysis Technique". In: July 2020. DOI: 10.1109/CBI49978.2020.10071.

[28] Ligang Zhou and Kin Keung Lai. "AdaBoost models for corporate bankruptcy prediction with missing data". In: *Computational Economics* 50.1 (2017), pp. 69–94.

[29] Maciej Zięba, Sebastian Tomczak, and Jakub Tomczak. "Ensemble Boosted Trees with Synthetic Features Generation in Application to Bankruptcy Prediction". In: *Expert Systems with Applications* 58 (Apr. 2016). DOI: 10.1016/j.eswa.2016.04.001.