University of Essex
Department of Mathematical Sciences

MA981: DISSERTATION

Predicting Company Bankruptcy

Amin Nizar Ali
2213409

Supervisor: Dr Jessica Claridge

November 24, 2023
Colchester

Contents

List of Figures

Abstract

This dissertation looks into the effectiveness of predictive models in predicting corporate bankruptcy, using a dataset encompassing 95 features related to business regulations on the Taiwan Stock Exchange (X1-X95). The main objective is to determine the impact and role of these characteristics within various predictive models in order to improve our understanding of their ability to identify companies on the verge of going bankrupt.

In order to guarantee the dataset's integrity, the study begins with careful feature engineering, data cleaning, and outlier removal. The following four distinct machine learning models are applied to the pre-processed data: logistic regression, random forest, XGBoost, and support vector machine (SVM). The comparison of model outputs reveals promising results, with XGBoost being better than other models. This study is designed to shed light on the relationship between business regulations and bankruptcy prediction accuracy through a thorough examination of the selected models. The findings give valuable insights into the significance of specific features as well as the overall robustness of the chosen models in identifying companies in financial difficulties. Keywords: Bankruptcy prediction, Machine learning, Predictive modelling, XGBoost, Business regulations, Taiwan Stock Exchange.

CHAPTER 1

Introduction

In my dissertation, I will be making a predictive model through some machine learning models predicting if the company is bankrupt or in the verge of bankruptcy. I will also do an extensive research on work already being done in this or similar areas, I will also be looking what models were used and which models would work best in this scenario. I am planning to run 3-4 different machine learning models selected through my literature review and also my previous experience, then compare the results to get the best results. Lastly, I will be discussing the limitations of my model and the hurdles I faced during this entire journey.

In the business and finance sector, the ability to predict or forecast financial problems and potential company bankruptcy is very important. The dataset I will be workingÂ on spans a decade from 1999 to 2009, contains an extensive collection of Taiwanese companies and provides a unique perspective into their financial condition. The major goal of this dataset like my dissertation is to help in the development of a reliable bankruptcy prediction model by putting light on the complex processes that influence a company's financial condition. Moreover, like others such as the Boston House-Price Data and the Gender Pay Gap Dataset, is exceptionally well-suited for the development of predictive models.

My goal is to be able to make a predictive model for bankruptcy detection using a dataset. It will be a valuable tool for professionals working in finance, economics, and data science. These professionals can then use my predictive models that may help in the prediction of bankruptcy benefiting in their respective fields. This models, in turn, can play a critical role in directing investment decisions and shaping corporate governance strategies, which leads

to more informed and sensible financial decision-making.

In my dissertation, I will be making a predictive model through some machine learning models predicting if the company is bankrupt or in the verge of bankruptcy. I will also do an extensive research on work already being done in this or similar areas, I will also be looking what models were used and which models would work best in this scenario. I am planning to run 3-4 different machine learning models selected through my literature review and also my previous experience, then compare the results to get the best results. Lastly, I will be discussing the limitations of my model and the hurdles I faced during this entire journey.

1.1 Research Question

How do the Taiwan Stock Exchange's business regulations, as represented by a dataset of 95 features (X1-X95), influence the accuracy of predictive models such as Logistic Regression, Random Forest, XGBoost, and Support Vector Machine in identifying companies on the verge of bankruptcy?

CHAPTER 2

Literature Review

The dataset for predicting corporate bankruptcy, sourced from the Taiwan Economic Journal, is a valuable resource that has called significant attention from researchers and practitioners in the fields of finance, machine learning, and predictive modelling. This dataset offers a comprehensive set of attributes related to financial performance, making it a very good choice for bankruptcy prediction. I found and explored some of the literature where similar work is already done to get an idea of how I will be framing and building my model.

2.1 Research papers/articles studied

2.1.1 Financial ratios and corporate governance indicators in bankruptcy prediction - Deron Liang - July 2016

Liang et al. (2016) conducted this important study titled "Financial Ratios and Corporate Governance Indicators in Bankruptcy Prediction: A Comprehensive Study," a foundational contribution published in the European Journal of Operational Research. This research is a very respectable study in the field of bankruptcy prediction, laying the groundwork for understanding the potency of datasets in predicting financial bankruptcy. The authors undertook a comprehensive investigation, carefully going into an extensive variables of financial ratios and corporate governance indicators to explore the vulnerability of companies to bankruptcy. This study emphasized the significance of attributes like return on assets

(ROA), operating profit rate, and tax rate in enhancing the precision of bankruptcy prediction models. In doing so, it not only added valuable insights to the field but also set a benchmark for the attributes of datasets that would prove indispensable for future research. This research went beyond just theoretical advancement. It also provided the dataset it used to the UCI Machine Learning Repository. As a result, this dataset has become a very important resource for researchers and data scientists, fostering further exploration and innovation in the domain of bankruptcy prediction. This dataset's accessibility has facilitated its widespread adoption in both academic and industrial circles, pushing forward numerous studies aimed at refining predictive accuracy and building different methodologies. It has created a ripple effect, elevating the quality of bankruptcy prediction models and enhancing the capacity to avoid financial distress that benefits investors, creditors, and businesses alike.This study helped me understand the field of finance and the importance of how will I be using the data I have and how it can help the businesses eventually.[LLTS16]

2.1.2 Bankruptcy prediction using SVM models - Ligang Zhou - Sep 2012

This article is like a guide for making predictions about companies which are on the path of being bankrupt, which is important for banks, investors, and governments. It talks about using a strong tool called Support Vector Machine (SVM) for this job, but says that SVM's performance depends on how you set it up and choose what things to look at. The study suggests a new way to pick the best things to look at and set up SVM, using a direct search and features ranking method. This method is compared to another method that uses genetic algorithms. The results show that the new method works well in predicting bankruptcy using a dataset of 2010 instances. This article is very helpful for me because it gives ideas on how to make my bankruptcy prediction models better, especially when I will be using SVM.[ZLY14]

2.1.3 How accurate are the bankruptcy predicting models of Altman(1968), Ohlson(1980) and Zmijewski(1984) after recalibration - Monique Timmermans - November 2014

The prediction of corporate bankruptcy holds a good importance for various stakeholders, ranging from investors to policymakers. It serves as a crucial tool for informed decision-

making, risk evaluation, and maintaining financial stability. Throughout the years, multiple bankruptcy prediction models have been worked on, for example the Altman (1968), Ohlson (1980), and Zmijewski (1984) models. While historically effective, these models encounter challenges in the contemporary dynamic business landscape. This paper, authored by a very respectable researcher, endeavours to assess the effectiveness of these classical models in predicting bankruptcy in present times. In the pursuit of this investigation, the author was carefully able to select a sample of US companies that experienced bankruptcy between 2005 and 2007, that is after the enactment of the Bankruptcy Abuse Prevention and Consumer Protection Act (BACPA). The objective was to scrutinize the good accuracy of the original models in the post-BACPA era and check whether recalibrating these models can possibly change or improve their predictive capabilities. The findings of this study shows some interesting insights. When the original models are employed without recalibration, they encounter difficulties in accurately predicting bankruptcy, particularly for companies that do not file for bankruptcy. This overprediction aligns with the alterations in bankruptcy law, rendering bankruptcy less probable for companies. Moreover, the recalibrated models present a divergent narrative. In these updated iterations, short-term liquidity emerges as a important factor. The recalibrated Altman (1968) and Ohlson (1980) models exhibit substantial predictive power even when forecasting bankruptcy three years in advance, while the Zmijewski (1984) model lags behind, notably at t-2 and t-3. In conclusion, this paper furnishes a valuable contribution by shedding light on the adaptability and reliability of classic bankruptcy prediction models. It shows that, with recalibration, these models can endure as effective tools for stakeholders, helping them in making well-informed decisions by using models in the field of finance and corporate.For me this paper helped me understand the importance of bankruptcy prediction, and how a similar dataset from a different geographic location was used to detect bankruptcy using and improving the previously used models.[Tim14]

2.1.4 Corporate Bankruptcy Prediction Model - Daniel Ogachi - March 2020

This study, conducted by Daniel Ogachi, Richard Ndege, Peter Gaturu, and Zeman Zoltan, investigates into the domain of corporate bankruptcy prediction, with a specific emphasis on listed companies in Kenya. The research begins by recognizing the pivotal role of accu-

rately predicting financial distress for various stakeholders in the business world, including policymakers, investors, banks, and the general public. The financial health of a company influences investment decisions, credit lending, and supply chain relationships. Inaccurate predictions can have strong and negative consequences for these stakeholders, making the development of precise bankruptcy prediction models of significant importance.

To address this challenge, the study introduces deep learning models that leverage textual disclosures for bankruptcy forecasting. It builds a comprehensive prediction model based on data from 64 listed companies on the Nairobi Securities Exchange over a decade. The analysis employs logistic regression, revealing the significance of various financial ratios in predicting bankruptcy, such as asset turnover, total assets, and working capital ratio. This research offers a vital contribution to the field and also my work, as it tailors a predictive model to the unique context of listed companies in Kenya, providing a very important tool for investors and financial decision-makers. By exploring the impact of key financial indicators, liquidity management ratios, and activity ratios, this study gives stakeholders with a more strong understanding of bankruptcy prediction. With the potential to improve investment decisions and risk assessment, this work not only advances financial economics but also bolsters the stability of the corporate landscape in Kenya and beyond.

This study, like the previous one helped me as it was build on a different geographical location kenya, helped me understand how it is important throughout the world. Moreover, it helped me to decide using logistical regression for my model.[ONGZ20]

2.1.5 Corporate bankruptcy prediction models applied to emerging economies - Ariel R. Sandin - December 2007

This paper explores how analysing ratios can help to predict bankruptcy in an emerging economy like Argentina, specifically in the 1990s. In a time of economic stability, the study aims to understand the financial differences between companies close to bankruptcy and the ones doing well.

To do this, the research looks at the financial numbers of 22 different companies, half of which declared bankruptcy, and the remaining half was able to avoid it. Using a method called multiple discriminant analysis, the study creates a model following previous research methods to make it comparable.

The findings was able to reveal that the financial data of Argentine companies in the 1990s does carry some important information for predicting bankruptcy. However, the choice of which model to use depends on the decision maker's preferences. The study highlights the imprtance of solvency ratios (related to total assets) and profitability ratios (related to sales) in predicting bankruptcy.

While the study acknowledges some limitations, like the small sample size of 11 healthy and 11 bankrupt companies, its practical implications are significant. The model created can support investors, creditors, and regulators in Argentina and other emerging economies as they try to predict business failure. Moreover, this paper also suggests using Altman's Z-score model for public companies in emerging economies, emphasizing on the importance of including profitability ratios, especially in fast-changing and growing environments.

In summary, this research adds to our understanding of predicting bankruptcy by showing the extra information provided by profitability. It also offers a very simple classification method for investors and creditors interested in the financial stability of Argentinean companies, helping them make more well-informed decisions. This paper like the previous two, helped me understand the value of bankruptcy prediction in different part of the world and why is it extremely important for companies and stakeholder to predict a companyâs bankruptcy.[SP07]

2.1.6 Corporate Bankruptcy Prediction: An Approach Towards Better Corporate World - Talha Mahboob Alam - June 2020

This research on corporate bankruptcy prediction played an a very important role in making my decision to use the Random Forest model for this project. The significance of predicting corporate bankruptcy is clear, given its impact on various stakeholders. This study has been a hot topic in academic and professional circles globally specially after the 90s. The research's focus on machine learning models to estimate the probability of corporate bankruptcy, while dealing with data-related challenges, which is very similar with my project's goals.

The paper's use of data balancing techniques, including SMOTE, was particularly relevant as my dataset faced exact same issues. The addition of multiple machine learning models for early bankruptcy prediction highlighted the need for a comprehensive approach. What stood out was the Random Forest model, which significantly improved predictive accuracy.

The comparison of various models with each other and the finding that the decision forest outperformed others solidified my choice to go with the random forest model as well.

In summary, this literature not only helped my understanding of corporate bankruptcy prediction but also convinced me of the Random Forest model's effectiveness.[ASM+21]

2.1.7 A Bankruptcy Prediction Model Using Random Forest - Shreya Joshi - March 2019

This literature like the previous one also contributed to my decision to use the Random Forest model in my research. It served as a valuable reference, adding to the choice I had already made after reviewing another paper's suggestion regarding the effectiveness of Random Forest for corporate bankruptcy prediction. The detailed exploration of different machine learning techniques, including Random Forest, and the authors' approach to addressing imbalanced data issues with techniques such as Synthetic Minority Over Sampling Technique (SMOTE), provided me with a deeper understanding of the model's potential.

Their study revealed notable improvements in predictive accuracy when employing machine learning models, particularly Random Forest, which achieved a 98.7% accuracy rate. This paper's findings and insights played a crucial role in solidifying my decision to implement SMOTE for data balancing and Random Forest for my this dissertation to detect bankruptcy prediction.[JRT18]

2.1.8 Company bankruptcy prediction framework based on the most influential features using XGBoost and stacking ensemble learning - Much Aziz Muslim - December 2021

Company bankruptcy prediction framework based on the most influential features using XGBoost and stacking ensemble learning - Much Aziz Muslim - December 2021 This literature was important for me as it influenced my decision to use XGBoost as one of the primary model in my research on predicting company bankruptcy. Company bankruptcy has far-reaching implications for many stakeholders, making early prediction very important. This study explores into finding the most effective predictive model, focusing on financial data from Polish companies.

A key element that convinced my decision was the implementation of XGBoost, which demonstrated its superiority through feature importance analysis. By using a weight value filter of 10, the study identified the most critical features, improving the efficiency of the model. The literature also employed ensemble learning, specifically stacking, with a combination of base models, including K-nearest neighbor, decision tree, support vector machines, and random forest, and a meta learner, LightGBM. The stacking model's remarkable accuracy rate of 97% outperforming the base models underscores the potential of XGBoost. This literature serves as a valuable guide and inspiration, strengthening my resolve to utilize XGBoost in my own corporate bankruptcy prediction research.[MD21]

2.1.9 Bankruptcy Prediction using the XGBoost Algorithm and Variable Importance Feature Engineering - Sami Ben Jabeur - January 2022

The research presented in this paper finalised my decision to use XGBoost in my own work. It discussed the challenges of predicting financial health in an era of big data and highlighted the need to select important factors without sacrificing performance quality, which directly relates to my research goals. The paper introduced an improved XGBoost algorithm, FS-XGBoost. This caught my attention due to its success in enhancing prediction accuracy and surpassing traditional methods for selecting key features. The study then compared FS-XGBoost with some other machine learning algorithms and feature selection techniques, revealing its better performance. These findings convinced me that XGBoost is the right choice for my research, as it can significantly improve prediction accuracy while simplifying the process of identifying important variables.[BJSC23]

2.2 Conclusion of Literature Review

The literature review on my corporate bankruptcy prediction significantly helped my approach in my dissertation. It emphasized the importance of predicting bankruptcy early, making it helpful for the company itself and various stakeholders.

In Conclusion to my literature review, I want to explain why I decided for Logistical Regression, Random Forest, XGBoost, and SVM models for my dissertation on predicting corporate bankruptcy. The research papers really helped me make these choices.

First off, Joshi (2019) backed up my choice of Random Forest, mentioning its better accuracy and using data balancing techniques like SMOTE. Muslim's research (2021) pushed me toward XGBoost, showing its power through feature importance and stacking. Jabeur's paper (2022) finalised for XGBoost, introducing an improved algorithm, FS-XGBoost, that boosts prediction accuracy. Liang et al. (2016) showed me the importance of things like return on assets and tax rate in predicting bankruptcy. Zhou, L's study (2012) gave me a good overview, also adding SVM to one of the technique for my predictive model. Timmermans (2014) taught me about recalibrating classic models for accuracy. Ogachi's work (2020) tailored a model for listed companies in Kenya using logistic regression and various financial ratios. Sandin's paper (2007) explained on predicting bankruptcy in an emerging economy, focusing on solvency and profitability ratios. Alam's study (2020) convinced me to use the Random Forest model, showing its improved accuracy and how it handles data issues.

In short, these papers gave me a strong understanding and led me to choose SMOTE for data balancing and Logistical Regression, Random Forest, XGBoost, and SVM models as my machine learning algorithms for my dissertation because of their strengths in predicting corporate bankruptcy. The insights from these studies not only improved my knowledge but also provided a clear idea for the selection of these models, considering their effectiveness in addressing the intricacies of financial distress prediction.

CHAPTER 3

Data Source, Collection and Preprocessing

The dataset is formed by a collection of Taiwanese companies from 1999 to 2009. The main objective of this dataset is to help predict whether these companies will run into financial difficulties, particularly the risk of going bankrupt. It essentially provides a detailed and comprehensive examination of these companies' economic conditions, giving us valuable insights into their financial condition.

We can look at this dataset as a window into the world of Taiwanese businesses during the course of that decade. It's similar to closely examining a company's financial report cardÂ but with lots of additional information. These details include things like how well a company manages its assets, its profit margins, tax obligations, and an extensive number of other financial metrics. All of these indicators have significance in evaluating a company's overall financial stability.

What sets this dataset apart and makes it unique is its capacity to help us understand the criteria used by the Taiwan Stock Exchange to identify and define corporate bankruptcy. In other words, it equips us with the necessary tools for predicting when a company might encounter financial difficulties or even potentially go bankrupt.

3.1 Data Collection and Preprocessing

To build an efficient and strong predictive model for corporate bankruptcy within the context of Taiwan-based companies, the careful preparation and preprocessing of the dataset plays an

important part. The dataset used for this study was obtained from Kaggle and originated from the Taiwan Economic Journal, and it spanned the years 1999 to 2009. This extensive dataset I obtained contained a total of 6819 data, each with 96 numerical features (Columns) including a range of financial variables important for bankruptcy prediction. The characteristics areÂ entirely numerical, with data types int64 or float64. This made things easier because the data was very clear and uniform, allowing the process to be consistent throughout.

	Bankrupt?	ROA(C) before interest and depreciation before interest	ROA(A) before interest and % after tax	ROA(B) before interest and depreciation after tax	Operating Gross Margin	Realized Sales Gross Margin	Operating Profit Rate	Pre-tax net Interest Rate	After-tax net Interest Rate	Non-industry income and expenditure/revenue
0	1	0.370594	0.424389	0.405750	0.601457	0.601457	0.998969	0.796887	0.808809	0.302646
1	1	0.464291	0.538214	0.516730	0.610235	0.610235	0.998946	0.797380	0.809301	0.303556
2	1	0.426071	0.499019	0.472295	0.601450	0.601364	0.998857	0.796403	0.808388	0.302035
3	1	0.399844	0.451265	0.457733	0.583541	0.583541	0.998700	0.796967	0.808966	0.303350
4	1	0.465022	0.538432	0.522298	0.598783	0.598783	0.998973	0.797366	0.809304	0.303475

...	Net Income to Total Assets	Total assets to GNP price	No-credit Interval	Gross Profit to Sales	Net Income to Stockholder's Equity	Liability to Equity	Degree of Financial Leverage (DFL)	Interest Coverage Ratio (Interest expense to EBIT)	Net Income Flag	Equity to Liability
..	0.716845	0.009219	0.622879	0.601453	0.827890	0.290202	0.026601	0.564050	1	0.016469
..	0.795297	0.008323	0.623652	0.610237	0.839969	0.283846	0.264577	0.570175	1	0.020794
..	0.774670	0.040003	0.623841	0.601449	0.836774	0.290189	0.026555	0.563706	1	0.016474
..	0.739555	0.003252	0.622929	0.583538	0.834697	0.281721	0.026697	0.564663	1	0.023982
..	0.795016	0.003878	0.623521	0.598782	0.839973	0.278514	0.024752	0.575617	1	0.035490

Figure 3.1: Bankruptcy Data

One of the most important strengths according to me of this dataset is its completeness, as there are no missing values (NaN) throughout the dataset features. This great attention to detail of data integrity ensures a solid foundation for my analysis, preventing any potential biases or inaccuracies coming from incomplete information. Additionally, a thorough verification process was done confirming the absence of duplicated values within the dataset, further enhancing the reliability of the observations.

A critical observation was done during the initial exploration of the dataset showing a high imbalanced distribution of the target variable, 'Bankrupt?'. Out of the 6819 observations, a significant majorityâ96.77%âare labelled as financially stable, while only 3.23% are marked as financially unstable making the distribution very imbalanced. This inherent class imbalance presents a critical challenge, as predictive models may inadvertently prioritize the majority class, potentially leading to a skewed learning process. Addressing this concern, the research

employed the Synthetic Minority Over-sampling Technique (SMOTE) in later stages.

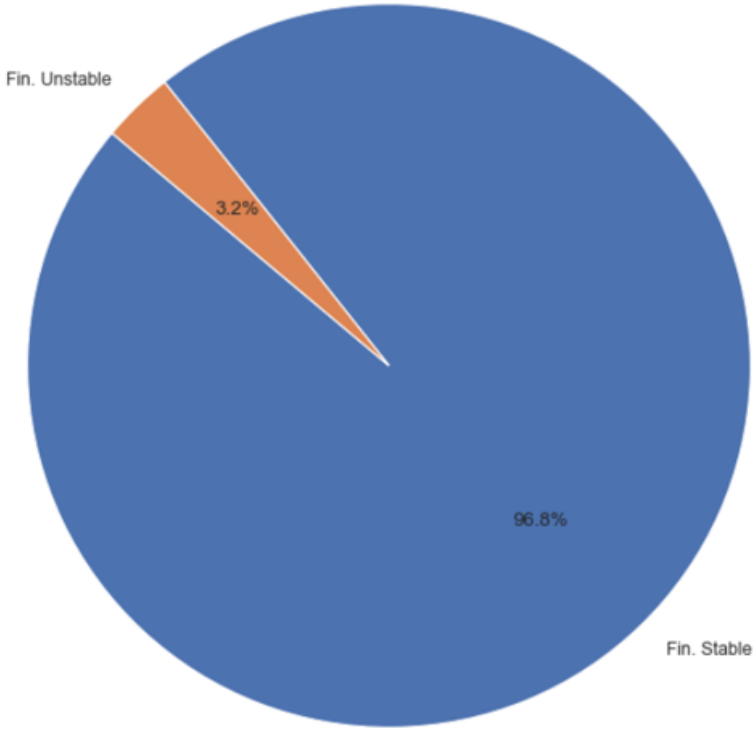

Figure 3.2: Imbalance in data

In summary, the dataset sourced from Kaggle, rooted in the Taiwan Economic Journal's data from 1999 to 2009, provides a rich and comprehensive source for revealing patterns related to corporate bankruptcy. Its rigorous preparation, combined with validity checks and the removal of duplicates, lays out the foundation for an accurate and dependable study. Furthermore, the recognition and management of class imbalance through SMOTE show an approach to solving inherent problems, opening the way for the construction and development of an intelligent, dependable, and efficient predictive model.

3.2 EDA, Feature Selection and Engineering

In the process of building a predictive model for corporate bankruptcy, I began by thoroughly exploring and fine-tuning the dataset. These initial procedures were not simply analytical; they created the framework for an effective, and efficient prediction model. The dataset had

to be slowly broken down in order to reveal certain slight patterns and correlations with one another and with some external factors. In this early stage, my commitment to a calculated and rigorous approach reflects our goal of navigating the complexity of predictive modelling through exploration and refinement.

1. Numeric Features: I started by looking closely at the numerical features that together make up the dataset. I plotted a histogram with a canvas measuring 35 by 30 inches and 50 bins providing granularity to get an idea and gain insights into the distribution and patterns of these features. These visual representations provide an in-depth understanding of how numerical values are distributed across various attributes. Each histogram is a snapshot that allows me to understand the frequency and magnitude of specific values in the dataset. This step lays the foundation for subsequent analyses, providing clarity on the numerical landscape I'll be navigating

Figure 3.3: Histogram of all variables

2. Relationships with the Spearman: My goal then shifted from individual features to understanding the relationships and dependencies between the features themselves. The Spearman Correlation Heatmap, also known as Spearman, has proven as an effective tool for visualising these relationships. This heatmap was able to illustrate the strength and direction of correlations between pairs of features by using a colour-coded representation. Cool tones indicate positive correlations, in which features move in together, whereas warm tones indicate negative correlations, in which features move in opposite directions. The heatmap serves as a guide, assisting me in identifying the most influential features in the dataset and their potential impact on predicting corporate bankruptcy, which is our primary objective.

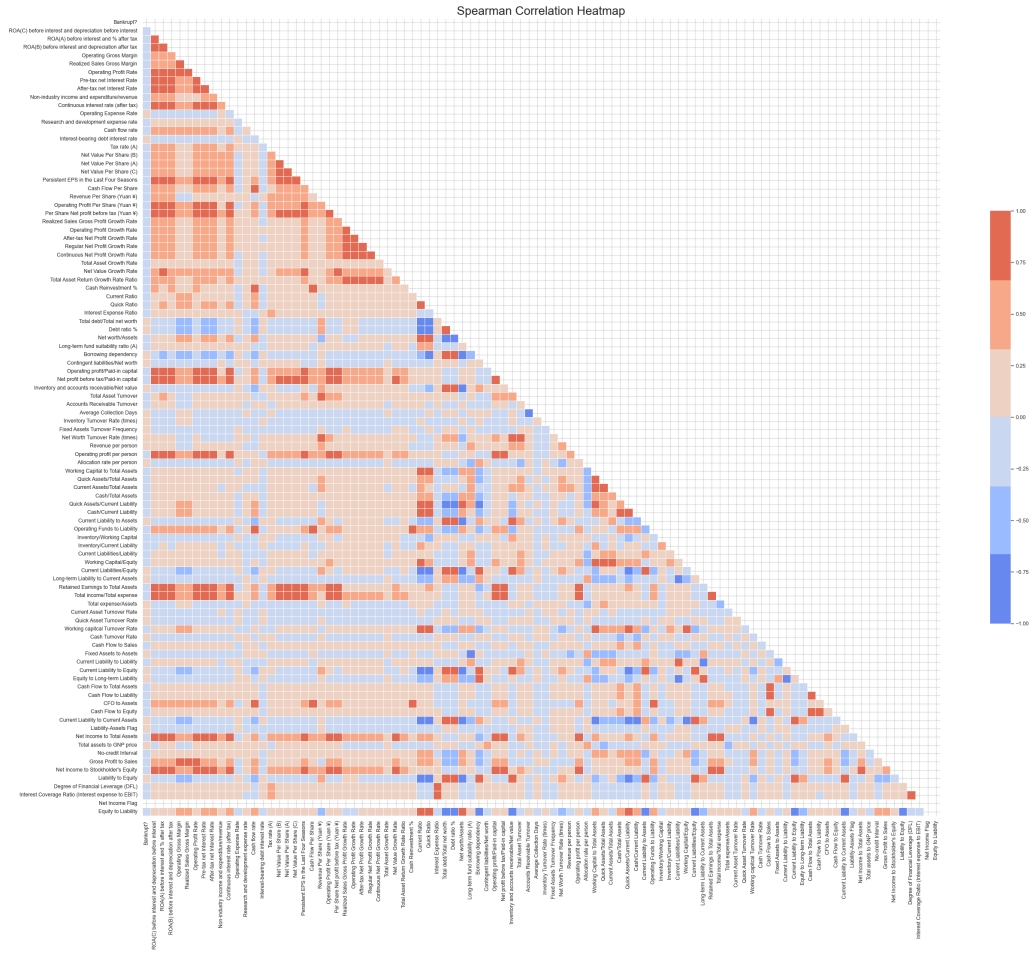

Figure 3.4: Heatmap between all variables

3. Data Enhancement and and Feature Distributions: Addressing and then locating outliers, or data points that deviate significantly from the majority of the dataset, is a key component of dataset refinement. Outliers may influence analysis and have a negative impact

on predictive model performance, making it inaccurate and inefficient. To work around this, I developed a method for identifying and then removing outliers from each feature. This process ensures that extreme values are dealt with correctly, leading to a more accurate and reliable dataset.

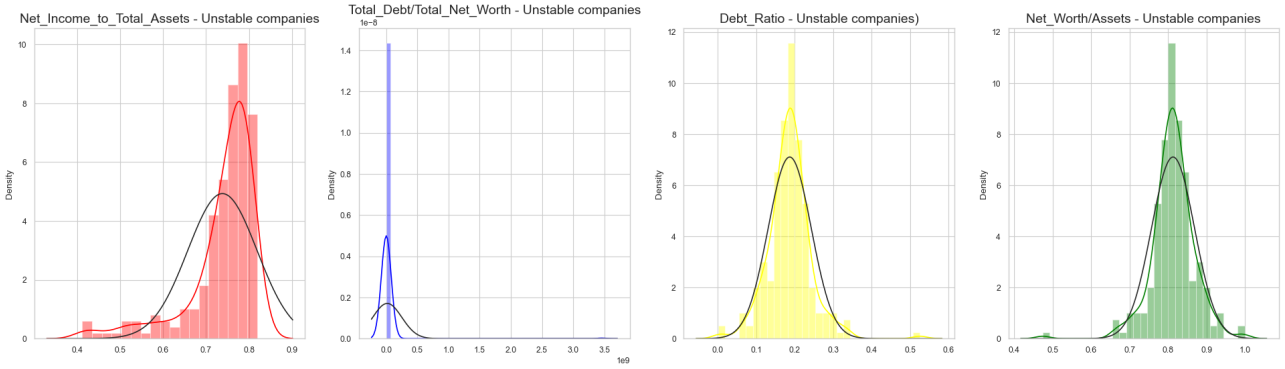

Figure 3.5: Before removal of outliers

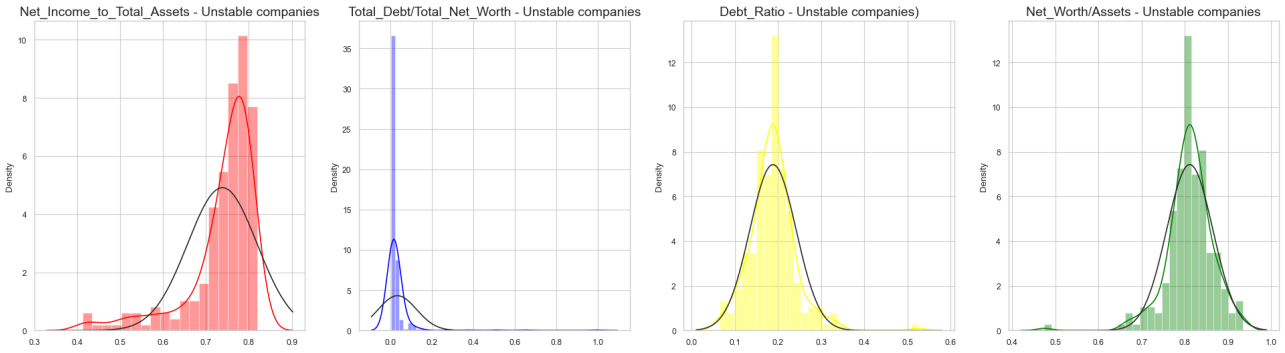

Figure 3.6: After removal of outliers

Furthermore, I investigated and validated the impact of theÂ outlier removal using the lens and feature distribution plots. Boxplots and feature distributions became our visual tools for evaluating how the removal of outliers changed and improved the spread and central tendencies of each feature. The before-and-after comparison revealed useful information about the success of the outlier removal process, verifying the improved quality of feature distributions.

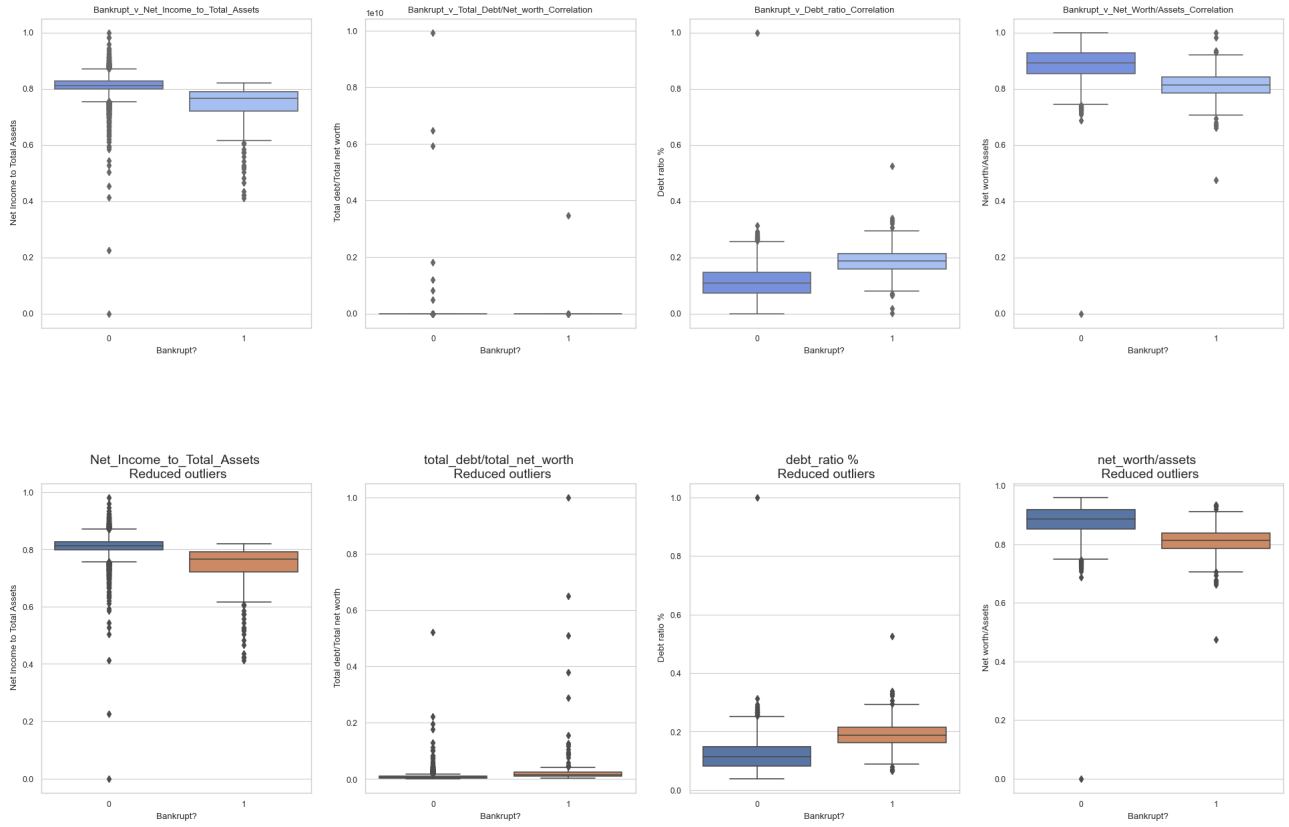

Figure 3.7: Before and After outlier removal

In conclusion, my exploration and refinement journey included some slow and careful steps in understanding the numeric complexities of our dataset. The histograms revealed individual feature distributions, the Spearman Correlation Heatmap revealed feature relationships, and the outlier removal process made the dataset a little stronger. These actions collectively strengthened the dataset, creating the way for me to come up with a sophisticated predictive model for corporate bankruptcy prediction.

CHAPTER 4

Model Selection and Justification

Before deciding for models to use I had to choose a method to balance the date to get better results.

4.1 SMOTE

With imbalanced data, my decision to favour oversampling over undersampling came from an intent to prevent the risk of loss of data. While undersampling was a simple solution, it risked eliminating valuable cases from the majority class, limiting the dataset's overall richness and diversity. When oversampling was selected.

My next step was choosing either random oversampling or the Synthetic Minority Oversampling Technique (SMOTE). Geeksforgeeks defines SMOTE as: "SMOTE synthesises new minority instances between existing minority instances. It generates the virtual training records by linear interpolation for the minority class. These synthetic training records are generated by randomly selecting one or more of the k-nearest neighbors for each example in the minority class. After the oversampling process, the data is reconstructed and several classification models can be applied for the processed data."[Geear]

SMOTE was selected because of its capacity to construct synthetic instances in a more complicated and context-aware manner. Unlike random oversampling, SMOTE addressed the possible problems related to noise input and overfitting by taking the local density of minority occurrences into consideration, resulting in a refined and efficient technique of data

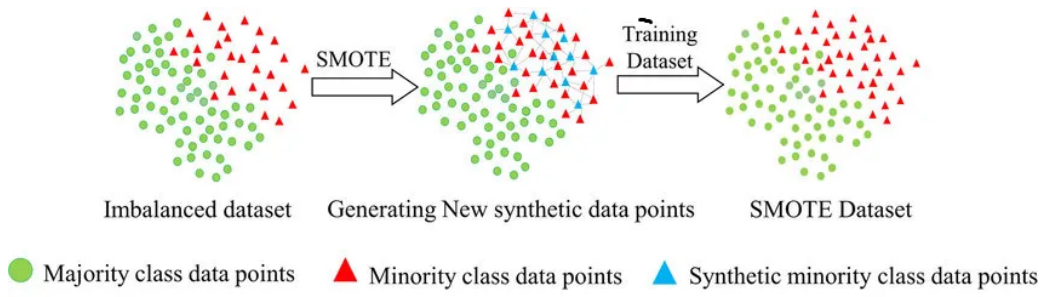

Figure 4.1: Understanding SMOTE

[Gomar]

balancing.

After SMOTE decision, I decided and finalised 4 predictive models for the development of my Bankruptcy model, I chose each of them based on the literature review I did initially and also based on my expertise and experience.

4.2 Logistic Regression

Regression is a statistical method used in mathematics and machine learning to model the relationship between a dependent variable and one or more independent variables. It takes the best-fitting line (or curve) that describes the relationship and can be used to predict the values of the dependent variable based on the independent variables' values.

As I aim to predict the bankruptcy prediction where binary results are expected, The first model I applied was logistical regression. Logistic regression is a machine learning classification algorithm. It is used for binary classification tasks, that predict the probability of a case belonging to a specific class. It makes use of the logistic function to map predicted values onto the range (0, 1).

According to IBM, Logistical regression is defined as "This type of statistical model (also known as logit model) is often used for classification and predictive analytics. Logistic regression estimates the probability of an event occurring, such as voted or didnât vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic regression, a logit transformation is applied on the oddsâthat is, the probability of success divided by the probability of failure. "[IBMarb]

Personally, I like Logistic Regression not just for its mathematical abilities, but also for

its ease of use. It's the model I always use as the starting point for my predicted excursions. It simplifies the process without losing precision. Using Logistic Regression is similar to creating a solid foundation before constructing the entire model. Its simplicity not only makes it user-friendly but also prepares the path for eventually investigation of advanced models. So Logistic Regression serves as a stepping stone, a reliable base that not only allows me to come up with accurate projections but also allows me to easily explore and build upon more complicated models in my dissertation.

4.3 Random Forest

Assume you have a huge decision to make, such as whether or not to go for a walk. A decision tree functions similarly to a flowchart in helping you in making that decision. At each step, you consider multiple variables, that are represented by branches in the tree. For example, the first question could be, "Is it raining?" If yes, one branch might tell you to stay indoors; if no, another branch might ask, "Is it sunny?" Each question leads to the next until you arrive at a decision, such as "Go for a walk" or "Stay inside."

A decision tree works similarly in the domain of data predictions. It addresses several aspects or characteristics of something and asks a series of questions to arrive at a conclusion. These questions help in categorizing and predicting results. Each branch of the tree represents a choice depending on a certain feature.

While decision trees can be useful, they might not address every problem. That's where the Random Forest approach comes in. It's like a group of decision trees working to offer a better and more precise prediction. IBM defines Random forest as "Random forest algorithms have three main hyperparameters, which need to be set before training. These include node size, the number of trees, and the number of features sampled. From there, the random forest classifier can be used to solve for regression or classification problems.[IBMarc]

The random forest algorithm is made up of a collection of decision trees, and each tree in the ensemble is comprised of a data sample drawn from a training set with replacement, called the bootstrap sample. Of that training sample, one-third of it is set aside as test data, known as the out-of-bag (oob) sample, which weâll come back to later. Another instance of randomness is then injected through feature bagging, adding more diversity to the dataset and reducing the correlation among decision trees. Depending on the type of problem, the

determination of the prediction will vary. For a regression task, the individual decision trees will be averaged, and for a classification task, a majority voteâi.e. the most frequent categorical variableâwill yield the predicted class. Finally, the oob sample is then used for cross-validation, finalizing that prediction. "

As it is apparent random forest is an improved decision tree model. It takes an approach of modelling by constructing a forest of decision trees that work to improve the precision of forecasts. Consider it a collaborative effort in which many trees join forces to offer a more solid prediction than individual trees could achieve, making it more effective than the decision tree model. Random Forest uses a variety of financial variables to forecast something as important as company bankruptcy, carefully evaluating conditions to produce an accurate and reliable result.

Alam's work (2020) convinced me further to use the Random Forest model, showing its improved precision and handling of data challenges. Furthermore, my choice for Random Forest is based on this extensive study that shows its greater accuracy, particularly when advanced methods like the Synthetic Minority Over Sampling Technique (SMOTE) are used with it. According to Alam's research, this model's ability to handle imbalanced datasets matters most when solving the issues that arise when predicting business bankruptcy. Random Forest has shown to be an effective option in my dissertation, consistently giving results that above expectations. Its collaborative approach and flexibility in complicated settings make it a good asset, firmly establishing it as a crucial component of my strategy.[ASM+21]

4.4 XGBoost

XGBoost, short for Extreme Gradient Boosting, is a powerful machine learning algorithm that belongs to the family of gradient boosting methods. It is widely used for supervised learning tasks, such as classification and regression problems.

IBM defines boosting and Xgboost as: "Boosting is an ensemble learning method that combines a set of weak learners into a strong learner to minimize training errors. In boosting, a random sample of data is selected, fitted with a model and then trained sequentiallyâthat is, each model tries to compensate for the weaknesses of its predecessor. With each iteration, the weak rules from each individual classifier are combined to form one, strong prediction rule. "[IBMard]

"Extreme gradient boosting or XGBoost: XGBoost is an implementation of gradient boosting thatâs designed for computational speed and scale. XGBoost leverages multiple cores on the CPU, allowing for learning to occur in parallel during training. "[IBMard]

XGBoost, an improved and efficient gradient boosting method, is an additional in my predictive modelling. Consider it like an effective conductor leading a symphony of inexperienced learners to build an effective and precise prediction model. XGBoost for corporate bankruptcy prediction works by identifying key characteristics through feature importance analysis.

Muslim's research (2021) confirmed my decision to include XGBoost in my model, showing its effectiveness in detecting influential features and superior performance in stacking ensemble learning. Because of the model's capacity to improve prediction accuracy while simplifying the identification of critical variables, it is a key addition to my predictive modelling methods.

4.5 Support Vector Machine (SVM)

SVM is like a smart line drawer. It analyses your data and chooses the best line to divide distinct groups, making decisions based on the points closest to the line. This method helps in the creation of a distinct and well-defined boundary between different categories, making it helpful in tasks such as pattern recognition and grouping items.

IBM defines SVM as: "SVM works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. A separator between the categories is found, then the data are transformed in such a way that the separator could be drawn as a hyperplane. "[IBMara]

As already explained Support Vector Machine (SVM), works by drawing a distinct line between different classes, making it useful in predicting business bankruptcy. It does well at managing both linear and non-linear relationships, making it suitable for a wide range of applications. SVM tries to optimize the margin between classes, improving its ability to generalize and predict accurately. Imagine an experienced artist drawing a fine boundary between financially solid and unstable areas.

SVM came to me through Zhou's (2012) paper, which showed its sensitivity to model form, parameter setup, and feature selection. My decision to include SVM in my dissertation

was inspired by its adaptability to different environments and its effectiveness in producing distinct classifications. SVM's strong performance, particularly in cases with well-defined class borders, adds an important layer to my corporate bankruptcy prediction modelling technique.[ZLY14]

CHAPTER 5

Code and Implementation

5.1 Libraries and data extraction

The code begins by importing data analysis and machine learning libraries such as seaborn, scipy, collections, matplotlib, os, numpy, and pandas. Specific modules from scikit-learn, xgboost, imbalanced-learn, and mlxtend are also imported. The dataset 'Data_bank' is then loaded from a CSV file in my directory, and the first few rows are displayed with the 'head()' function. The 'info()' function displays a summary of the dataset's data types, non-null values, and memory usage. The 'describe()' function is also used to generate summary statistics for numeric columns.

If any columns with missing values are found, the code prints their names. It also displays a message if no missing values are found in any of the columns. The code then uses the 'value_counts()' function to display the count of each class in the target variable ('Bankrupt?'). It calculates and shows the proportion of 'Unstable' and 'Stable' values in the target column. This section of code sets the stage by preparing the tools and loading the dataset for further analysis and modelling.

5.2 Data Exploration

This code starts with a full exploration and visualisation process in this segment and was done to get a very proper idea of how the data looks and the importance each variable has.It

calculates and presents the class distribution of the target variable, 'Bankrupt?' using a clear and informative pie chart to demonstrate the imbalance. Moving on to a deeper examination, the code generates histograms for all numeric columns, explaining the distribution patterns of various features in detail. Following that, it looks into the detailed web of relationships between numerical features by generating a Spearman correlation matrix and visualising it with an insightful and clear heatmap.

The investigation continues with the creation of boxplots for the entire dataset, which provides a clear overview of data distributions and highlights potential outliers. Further granularity is achieved through the construction of correlation boxplots for specific features, providing an alternative viewpoint on variations between stable and unstable companies. The code improves its investigation of the distribution plots of features unique to unstable companies, resulting in new insights. The identification and removal of outliers from all columns is a critical step in ensuring data integrity. The effect of outlier removal on boxplots and distribution plots is depicted graphically. To improve the dataset, the code includes a log transformation to address skewness in selected columns, as well as visualisations of the resulting changes in boxplots and histograms.

5.3 Data Splitting, SMOTE and Modelling

The train_test_split method is used in this section of the code to divide the pre-processed data into training and testing sets. The division is stratified, ensuring a balanced distribution of classes in both sets, which is critical for accurate model evaluation, especially in the context of imbalanced datasets. Following the data division, we can see that our data is very imbalanced, so the code employs the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance by oversampling the minority class in the training set. Following that, it builds and evaluates machine learning models, starting with Logistic Regression.

For strong model evaluation, the code employs RandomizedSearchCV to optimise hyperparameters and a stratified cross-validation approach. For each fold, evaluation metrics such as accuracy, precision, recall, and F1-score are computed and displayed. After that, the process is repeated for Random Forest, XGBoost, and Support Vector Machine (SVM) classifiers. Classification reports and precision-recall curves provide a detailed overview of each model's performance, allowing for informed comparisons and insights into their ability

to predict financial stability.

5.4 Evaluating the result of all the models applied

5.4.1 Logistical regression

The model has a high overall accuracy of 89.32%. The low precision of 15.71%, on the other hand, indicates a significant risk of false positives, potentially misclassifying non-bankrupt cases. While the recall is relatively high (62.92%), the difference between precision and recall should be carefully considered for application. The F1 score of 24.93% reflects the trade-off between precision and recall, highlighting potential areas for potential improvement in model performance. The classification report suggests that the logistic regression model performs

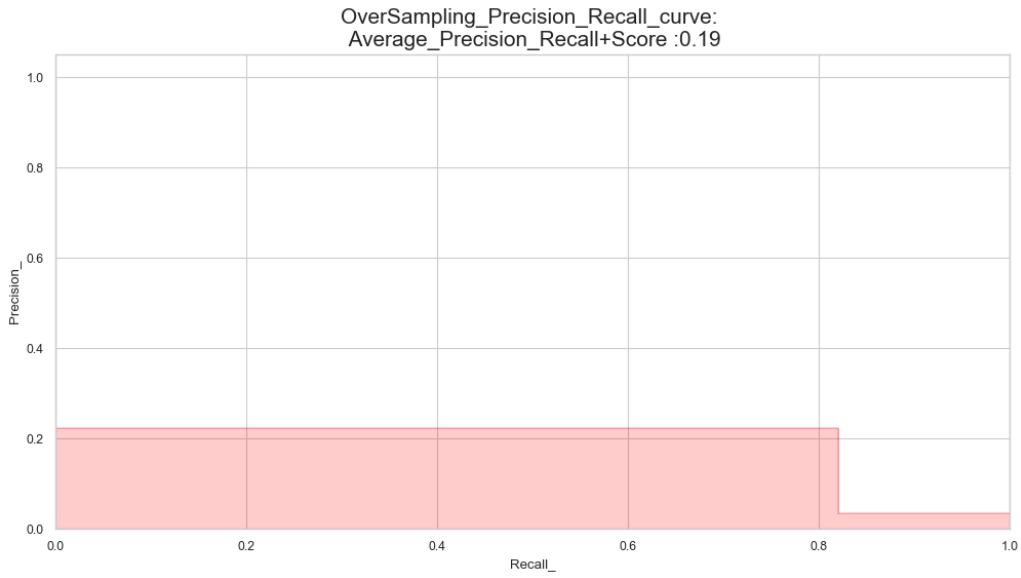

Figure 5.1: Oversampling precession recall curve

well, particularly in identifying financially stable cases with a precision of 99% and a recall of 90%. However, it exhibits challenges in classifying financially unstable cases, with a lower precision of 22% and a higher recall of 79%. The macro average and weighted average metrics indicate a balanced overall performance, emphasizing the model's reliability in predicting financial stability but also highlighting areas for potential improvement, especially in precision for financially unstable instances.

```
-
Logistic Regression (SMOTE) results:
_
Accuracy_:= 0.8932447397563676
Precision_:= 0.15718671195512326
Recall_:= 0.6290322580645161
F1_:= 0.24931932038691978
_
```

Figure 5.2: Logistic Regression results

Logistical Regression	precision	recall	f1-score	support
Fin_Stable	0.99	0.90	0.94	1089
Fin_Unstable	0.22	0.82	0.35	39
accuracy			0.89	1128
macro avg	0.61	0.86	0.65	1128
weighted avg	0.97	0.89	0.92	1128

Figure 5.3: Logistic Regression classification report

5.4.2 Random Forest

Perfect scores for the Random Forest model in accuracy, precision, recall, and F1 indicate flawless performance on the given dataset. While achieving such ideal metrics is interesting it is necessary to completely validate the results, taking into account potential issues such as overfitting or data being leaked. The report on classification indicates that the model performed well for financially stable cases with a precision of 97% and recall of 98%. However, for financially unstable cases, the precision is lower at 37%. While the model achieves high precision and recall for financially stable cases, it has limitations in identifying financially unstable cases, as evidenced by the lower recall and F1-score for that class. Overall, it demonstrates dependable predictive capabilities, though there is room for improvement in capturing unstable instances.

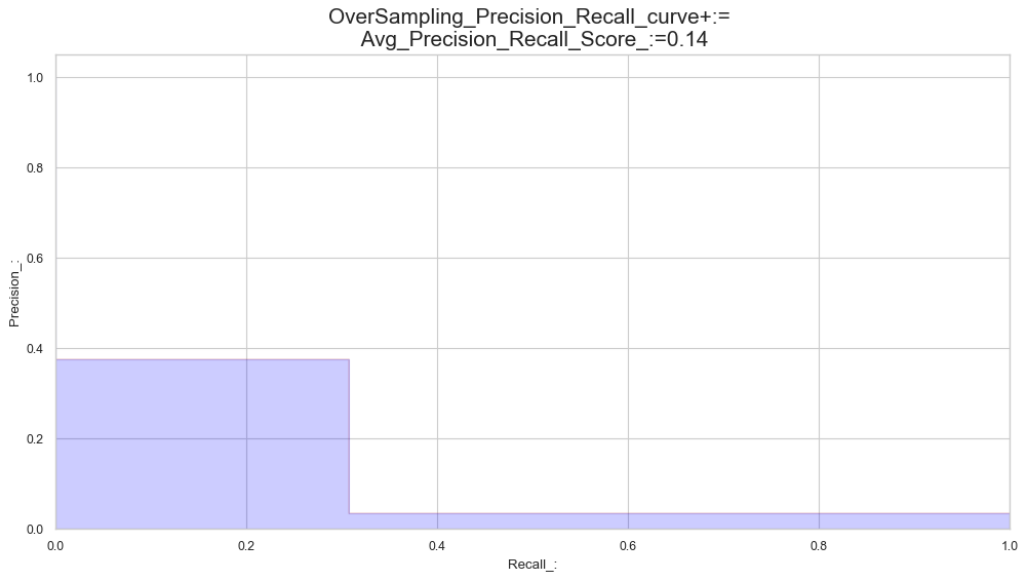

Figure 5.4: Oversampling precession recall curve

```
.................................................
Random Forest results:=
-
Accuracy_:= 1.0
Precision_:= 1.0
Recall_:= 1.0
F1_:= 1.0
.................................................
```

Figure 5.5: Random Forest results

5.4.3 XGboost

The XGBoost model performs very well, with an accuracy of 97.91%, showing highly accurate overall predictions. The model effectively balances the trade-off between correctly identifying positive instances and minimising false positives, resulting in a robust F1 score of 80.36% with a precision of 71.33% and a recall of 98.104%. According to the classification report, the model is reasonably effective but has some limitations. With a precision of 98% for "Fin_Stable," the model correctly identifies financially stable cases 98% of the time. The relatively lower recall of 94%, on the other hand, suggests that it may miss some stable instances. Precision is 28% for "Fin_Unstable," indicating a relatively high rate of false positives, while recall is 59%, indicating some capacity to capture actual unstable instances. The balanced accuracy

Random Forest	precision	recall	f1-score	support
Fin_Stable	0.98	0.98	0.98	1089
Fin_Unstable	0.38	0.31	0.34	39
accuracy			0.96	1128
macro avg	0.68	0.64	0.66	1128
weighted avg	0.95	0.96	0.96	1128

Figure 5.6: Random Forest classification report

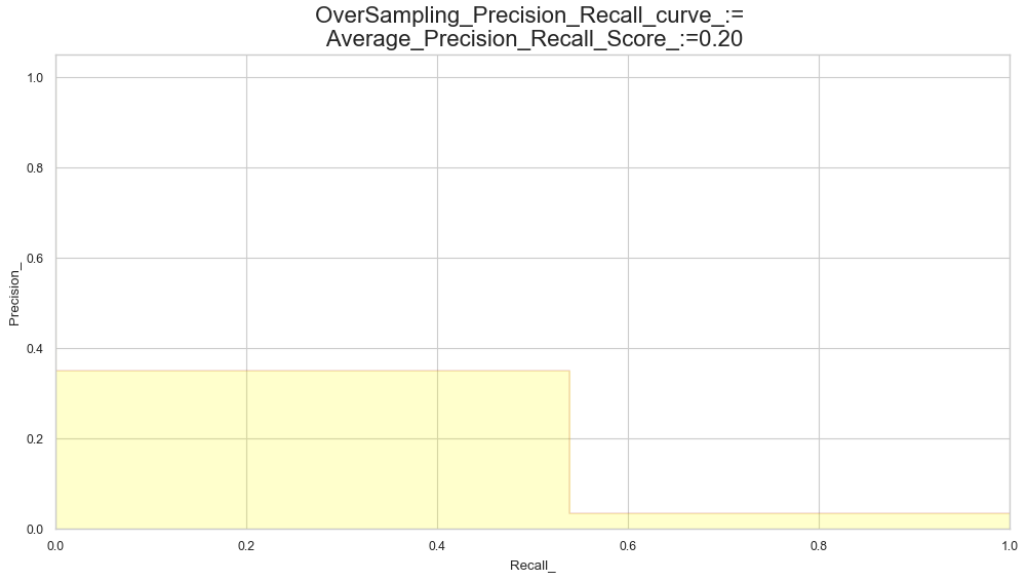

Figure 5.7: Oversampling precession recall curve

and macro average metrics show a model that is reasonably well-performing, but they also highlight areas for improvement, particularly in identifying financially unstable cases.

5.4.4 SVM

The SVM model shows decent overall performance, achieving an accuracy of 87.88%. While the precision is relatively low at 21.14%, the model demonstrates strong recall at 89.86%, suggesting its effectiveness in capturing actual positive instances. The F1 score of 34.16% reflects a balanced measure of precision and recall, indicating the model's potential for improvement in precision.

```
.................................................
Xgboost Results results:
----------
Accuracy_:= 0.9791805094130674
Precision_:= 0.7133065191888721
Recall_:= 0.9810483870967742
F1:= 0.8036753799923397
.................................................
```

Figure 5.8: XGboost results

Xgboost	precision	recall	f1-score	support
Fin_Stable	0.98	0.96	0.97	1089
Fin_Unstable	0.35	0.54	0.42	39
accuracy			0.95	1128
macro avg	0.67	0.75	0.70	1128
weighted avg	0.96	0.95	0.95	1128

Figure 5.9: XGboost classification report

The classification report indicates that the SVM model performs well in identifying financially stable cases, with a precision of 99% and a recall of 88%. This suggests a high accuracy in correctly classifying stable instances, although there is a trade-off with lower precision for financially unstable cases (19%). The macro average underscores the model's ability to balance overall performance across classes, with a weighted average accuracy of 88%, highlighting its reliability in predicting financial stability.

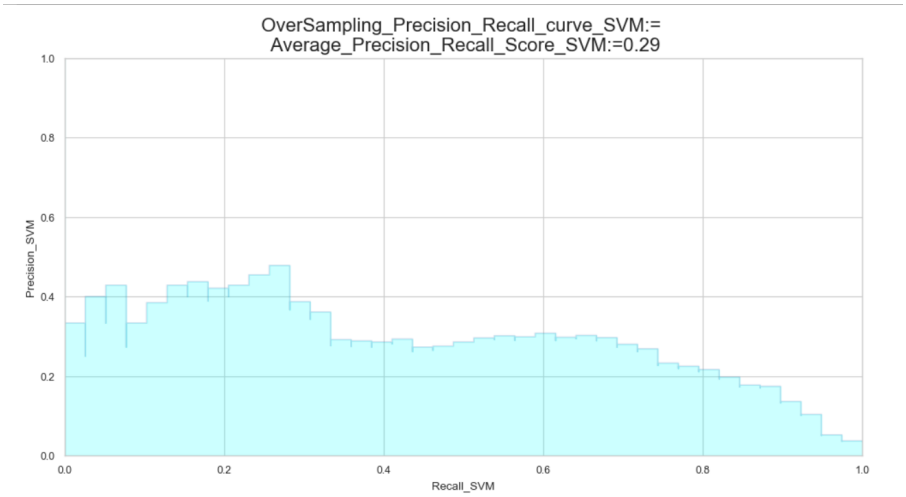

Figure 5.10: Oversampling precession recall curve

```
.....................................................
SVM results:
----------
Accuracy_SVM: 0.8788482834994463
Precision_SVM: 0.21141025795139545
Recall_SVM: 0.8985887096774192
F1_SVM: 0.34161476676458175
.....................................................
```

Figure 5.11: XGboost results

	precision	recall	f1-score	support
Fin_Stable	0.99	0.88	0.93	1089
Fin_Unstable	0.19	0.82	0.31	39
accuracy			0.88	1128
macro avg	0.59	0.85	0.62	1128
weighted avg	0.97	0.88	0.91	1128

Figure 5.12: SVM classification report

CHAPTER 6

Conclusion

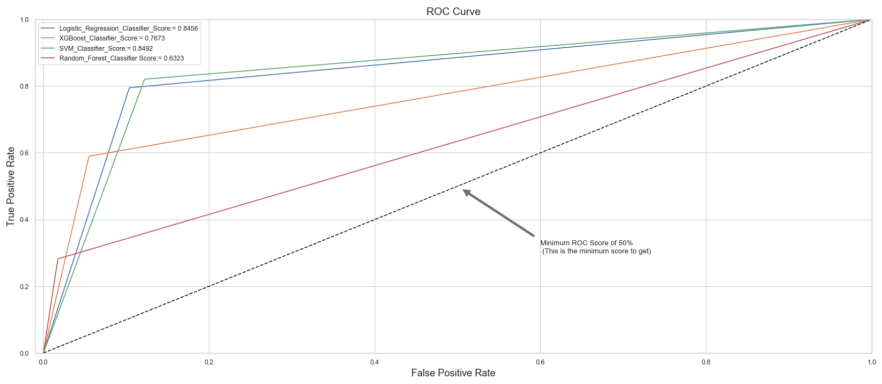

Figure 6.1: ROC curve

6.1 Best model based on results

Among the models evaluated in the previous section, XGBoost among the other 3 stands out as the most strong option. It does well at making accurate predictions across the dataset, with an impressive accuracy of 97.91%. The model achieves a balanced trade-off between precision (71.33%) and recall (98.10%), resulting in an impressive F1 score of 80.36%. It also means that XGBoost successfully detects positive instances while reducing false positives, proving its better performance in predicting financial stability.

While Logistic Regression has high precision for financially stable cases, it has difficulty with precision for financially unstable cases, showing potential misclassification. Random

Forest, despite achieving perfect scores, was susceptible to overfitting and has obvious limitations in identifying financially unstable cases. SVM works well but falls short in precision. The combination of XGBoost's high accuracy, precision, and recall, as well as its robust F1 score, positions it as the best model for predicting financial stability in my dissertation.

6.2 Limitation of the chosen model

While XGBoost emerges as the preferred model with good accuracy, precision, and recall, certain limitations must be accepted. Despite its high overall accuracy, XGBoost, like any predictive model, will encounter difficulties in scenarios with imbalanced class distributions. The lower precision (71.33%) for financially unstable cases in this context reveals a potential vulnerability to false positives, indicating instances where the model incorrectly predicts financial instability.

Furthermore, the detailed relationship between precision and recall needs consideration as well. While the model has a high recall of 98.10%, indicating that it effectively captures actual positive instances, the precision-recall trade-off points to caution. A more thorough review of the classification report reveals that, while the recall is high, achieving high precision for both stable and unstable cases at the same time remains a challenge. The model's limitations become clear in the complicated balance required for financial predictions, which requires constant evaluation and possible refinement to improve the model's ability to identify complex relationships in data. Regular validation and recalibration practises will be required to ensure the XGBoost model's continued reliability and relevance in predicting financial stability.

6.3 Improvements possible for future researches

In order to fix the identified limitations and improve the accuracy of the XGBoost model in predicting financial stability, multiple possibilities for improvement can be explored. For example, feature engineering could be critical, involving an in-depth analysis and picking out relevant features that more accurately capture the specifics of financial relationships. Iterative refinement of the feature set, potentially including new data sources or transforming existing

variables, can contribute to a more comprehensive understanding of the patterns that lie beneath them.

Furthermore, tuning the hyperparameters allows you to improve the model's performance. Experimenting with different hyperparameter configurations, such as changing learning rates or tree depths, can help to refine the trade-off between precision and recall. This iterative process may involve carefully examining the hyperparameter space to find configurations that improve the model's predictive capabilities using techniques such as grid search or random search.

It is also important to continuously monitor the model's performance on new and unknown data. Regular updates to the training dataset, as well as regular revision of the model's performance metrics with altering financial landscapes, will help to keep the model relevant. Collaboration with domain experts can also help to improve the model's understanding of the financial context, leading to more accurate predictions.

In conclusion, a dynamic and incremental model development approach that includes feature refinement, hyperparameter tuning, and continuous validation will clear the way for continuous improvements in the XGBoost model's ability to effectively predict financial stability.

Bibliography

[ASM+21] Talha Mahboob Alam, Kamran Shaukat, Mubbashar Mushtaq, Yasir Ali, Matloob Khushi, Suhuai Luo, and Abdul Wahab. Corporate bankruptcy prediction: An approach towards better corporate world. *The Computer Journal*, 64(11):1731–1746, 2021.

[BJSC23] Sami Ben Jabeur, Nicolae Stef, and Pedro Carmona. Bankruptcy prediction using the xgboost algorithm and variable importance feature engineering. *Computational Economics*, 61(2):715–741, 2023.

[Fra10] Andrew Frank. Uci machine learning repository. *http://archive. ics. uci. edu/ml*, 2010.

[Geear] GeeksforGeeks. Ml - handling imbalanced data with smote and near miss algorithm in python. Year.

[Gom22] Everton Gomede. Synthetic minority over-sampling technique (smote): Empowering ai through imbalanced data handling. *Medium*, Jul 30 2022.

[Gomar] Everton Gomede. Synthetic minority over-sampling technique (smote): Empowering ai through imbalanced data handling, Year.

[IBMara] IBM. How svm works - ibm spss modeler documentation, Year.

[IBMarb] IBM. Logistic regression, Year.

[IBMarc] IBM. Random forest, Year.

[IBMard] IBM. Xgboost, Year.

[JRT18] Shreya Joshi, Rachana Ramesh, and Shagufta Tahsildar. A bankruptcy prediction model using random forest. In *2018 second international conference on intelligent computing and control systems (ICICCS)*, pages 1–6. IEEE, 2018.

[LLTS16] Deron Liang, Chia-Chi Lu, Chih-Fong Tsai, and Guan-An Shih. Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European journal of operational research*, 252(2):561–572, 2016.

[MD21] Much Aziz Muslim and Yosza Dasril. Company bankruptcy prediction framework based on the most influential features using xgboost and stacking ensemble learning. *International Journal of Electrical and Computer Engineering (IJECE)*, 11(6):5549–5557, 2021.

[ONGZ20] Daniel Ogachi, Richard Ndege, Peter Gaturu, and Zeman Zoltan. Corporate bankruptcy prediction model, a special focus on listed companies in kenya. *Journal of Risk and Financial Management*, 13(3):47, 2020.

[SP07] Ariel R Sandin and Marcela Porporato. Corporate bankruptcy prediction models applied to emerging economies: Evidence from argentina in the years 1991-1998. *International Journal of Commerce and Management*, 17(4):295–311, 2007.

[Tim14] M Timmermans. Us corporate bankruptcy prediction models: How accurate are the bankruptcy predicting models of altman (1968), ohlson (1980) and zmijewski (1984) after recalibration, when they are applied to us listed firms in the period after the bacpa change in bankruptcy law. *Unpublished doctoral dissertation). Thesis/Dissertation ETD*, 2014.

[ZLY14] Ligang Zhou, Kin Keung Lai, and Jerome Yen. Bankruptcy prediction using svm models with a new approach to combine features selection and parameter optimisation. *International Journal of Systems Science*, 45(3):241–253, 2014.