



**QUEEN'S
UNIVERSITY
BELFAST**

**QUEEN'S
BUSINESS
SCHOOL**

**Financial Ratios as Predictors of Company Bankruptcy: A Predictive Model
Approach**

Research Report

Name: Muhammad Muneeb Ullah Ansari

Student ID: 40426685

Word Count: 8,249

**Research Report submitted in part fulfilment of the degree of Master of
Science in Business Analytics**

September 2024

Queen's Business School

Candidate Declaration

Declaration

This is to certify that:

- i. The portfolio comprises only my original work;
- ii. AI technologies (e.g. chat GTP) have not been used in the writing of the portfolio dissertation.
- iii. No portion of the work referred to in the dissertation has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.



[Candidate's Signature]

Muhammad Muneeb Ullah Ansari

Printed Name

27 - 08 - 2024

Date

Acknowledgments

I would like to express my deepest appreciation to my supervisor, Reza Yousefi Zenouz, whose invaluable guidance and steadfast patience have profoundly influenced this dissertation. His expertise has been instrumental in refining this work, and I consider myself fortunate to have had the chance to learn under her mentorship.

Additionally, I extend my sincere gratitude to Queen's Business School for cultivating an environment that encourages academic excellence and inquiry. I am especially thankful to our program director, Dr. Byron Graham, for his unwavering encouragement throughout this academic journey.

Finally, my heartfelt thanks go to my family. Their unwavering support and belief in my abilities have been the foundation upon which this dissertation is built.

Abstract

The following study aims to create a machine learning model for predicting corporate bankruptcy. The dataset used is from the UCI Machine Learning Repository containing 95 financial ratios for 6,819 companies in Taiwan.

The study starts by explaining Bankruptcy followed by a literature review on previous quantitative approaches for bankruptcy, causes of bankruptcy and implications on a bankrupt firms.

For the methodology, we accurately and step-by-step follow CRISP-DM (Cross-Industry Standard Process for Data Mining) research framework and explain our approach considering it. For Feature Selection, we deploy a Decision Tree. We perform Isolation Forest for Outlier Analysis. As our data had class imbalance, we also tested 4 sampling methods which were Random Oversampling, Random Under-sampling, SMOTE (Synthetic Minority Oversampling Technique) and ADASYN (Adaptive Synthetic Oversampling). We tested 5 predictive models which included Random Forests, Logistic Regression, MDA (Multiple Discriminant Analysis), SVM (Support Vector Machines) and KNN (K-Nearest Neighbor). We performed Cross-Validation 10 folds 3 times to reduce overfitting. After diving deeper into the best performing model to extract more insights, we also applied IML (Integrated Machine Learning) methods to understand the complex relationships between variables and the outcome.

Our study finds out that out of five algorithms chosen to create the model, SVM gives the highest accuracy of 95.45%. The best sampling method in our study was concluded to be SMOTE. The financial ratios most indicative of bankruptcy were found out to be Net Income to Total Assets, Equity to Liability, Net Income to Stockholder Equity and Total Debt to Total Net Worth.

We also extracted some interesting insights about the data. The first insight is that bankrupt and non-bankrupt companies are not very distinguishable from one another, revealed by deep diving into the best performing model. The second insight is that changes in financial ratios are correlated to bankruptcy but they do not directly cause it, revealed by IML methods.

The study compares its findings to those present in the literature to discover that findings in this study are consistent with literature. The study concludes by highlighting some limitations, scope for future work and how can this study be of use to different stakeholders.

Keywords

Bankruptcy, Company Distress, Machine Learning, Feature Selection, Sampling, Interpretable Machine Learning.

Table of Contents

Declaration	1
Acknowledgments	2
Abstract	3
Keywords	4
1- Introduction	8
1.1- Overview	8
1.2- Scope	9
2- Literature Review	10
2.1- Review of Quantitative Approaches in Bankruptcy Prediction	10
2.2- Causes of Bankruptcy	12
2.3- Implications Bankrupt Firms	13
3- Methodology	15
3.1- Business Understanding	16
3.2- Data Understanding	16
3.3- Data Preparation	17
3.3.1- Feature Selection	17
3.3.2- Outlier Removal	17
3.4- Modeling	18
3.4.1- Sampling Methods	18

3.4.1.1- SMOTE (Synthetic Minority Oversampling Technique).....	18
3.4.1.2- ADASYN (Adaptive Synthetic Sampling)	19
3.4.1.3- Random Oversampling.....	19
3.4.1.4- Random Under-sampling	19
3.4.2- Predictive Modeling.....	19
3.4.2.1- MDA (Multiple Discriminant Analysis)	20
3.4.2.2- SVM (Support Vector Machines)	20
3.4.2.3- KNN (K-Nearest Neighbor)	20
3.4.2.4- LR (Logistic Regression)	20
3.4.2.5- RF (Random Forest).....	21
3.4.3- Cross-Validation	21
3.5- Model Evaluation Metrics	21
3.5.1- TNR (True Negative Rate).....	22
3.5.2- NPV (Negative Predicted Value)	22
3.5.3- F2 Score.....	23
3.5.4- ROC (Receiver Operating Characteristics) Curve	23
3.5.5- Interpretable Machine Learning.....	24
3.5.5.1- PDP (Partial Dependency Plots)	24
3.5.5.2- ALE (Accumulated Local Effects).....	24
4- Findings	25

4.1- Exploratory Visualizations.....	25
4.2- Correlation Analysis	28
4.3- Outlier Analysis using Isolation Forest.....	32
4.4- Feature Selection Using Decision Trees	33
4.5- Testing Sampling Methods	34
4.6- Testing Predictive Models	39
4.7- Further Analyzing Best Model.....	43
4.8- Interpretable Machine Learning Using PDP and ALE	48
5- Discussions	69
6- Conclusions and Recommendations	71
References	74
Appendix.....	82
Dissertation Checklist	82

1- Introduction

1.1- Overview

The definition of Bankruptcy is such that it varies greatly depending on what literature you refer.

A study by Zhao, Quenniche, and De Smedt (2024) summarized bankruptcy definition into 3 categories: Bankruptcy / Legal Definitions, Financial Distress Definitions, Hybrid Definitions.

Bankruptcy / Legal Definitions: *“A company is liquidated, reorganized, or ruled by court decision as bankrupt.”*

Financial Distress Definitions: *“A company has negative net income for two consecutive years.”*

Hybrid Definitions: *“A Company is classified as financially distressed whenever it meets the following two conditions: (1) the firm is inactive, has merged, is suspended, dissolved, or undergone liquidation (either voluntary or by court order), gone bankrupt or equivalent; (2) its net income is negative for three consecutive years.”*

The Global Financial Crisis of 2008 has led to a rise in bankruptcy cases across the globe and has sparked a new area of research focused on predicting this occurrence not only at the national level but also on a global scale, identifying the shared characteristics of companies that have been impacted (Alaminos, Del Castillo, and Fernández, 2016). Bankruptcy can occur due to many reasons of which Financial mismanagement, deteriorating market conditions, regulatory changes, and unforeseen external shocks are included (Kirkos, 2015). Bankruptcy is also of a contagious nature such that firms whose business operations depend on bankrupt company's activities may follow in bankruptcy (Doumpos and Zopounidis, 1999). The primary reason companies want to avoid bankruptcy is because its legal and financial costs are prohibitive,

meaning so large that companies cannot afford it (Weiss, 1996). As noted by Papan and Spyridou (2020), while bankruptcy has many definitions, there is no universally accepted definition. Contributive of this, there is no universally accepted bankruptcy prediction model throughout literature (Shi and Li, 2019). The earliest study of modern bankruptcy can be traced back to Fitzpatrick (1932) where he predicted bankruptcy using 20 company financial accounting ratios.

1.2- Scope

This study aims to assess the accuracy of machine learning techniques in predicting corporate bankruptcy and to identify the key factors influencing bankruptcy risk. To achieve this, five different machine learning models will be developed: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), and Multiple Discriminant Analysis (MDA). The research will involve cleaning and preprocessing the dataset to eliminate irrelevant information and address any anomalies. The performance of these models will be evaluated using various statistical metrics such as Negative Predicted Value ($TN / (TN + FP)$), Specificity (TN Rate), F2 score, and AUC. To enhance the transparency and understanding of the results, the study will employ Partial Dependency Plots (PDP) and Accumulated Local Effects (ALE) for analyzing feature interactions.

The findings from this research can serve as a crucial foundation for drafting new financial regulations or refining existing ones related to corporate bankruptcy. By providing precise predictions of bankruptcy risks, policymakers can make decisions that are data-driven and evidence-based, ensuring that interventions are targeted and effective (Claessens and Yurtoglu, 2013). Furthermore, accurate predictions can assist in aligning corporate governance practices

with broader economic stability goals, particularly those related to financial health and sustainability of businesses (Laeven and Valencia, 2018).

2- Literature Review

2.1- Review of Quantitative Approaches in Bankruptcy Prediction

The earliest study of modern bankruptcy can be traced back to Fitzpatrick (1932) where he predicted bankruptcy using 20 company financial accounting ratios. He compared failed and successful firm's ratios and found that the successful companies showed good ratios while the failed firms had subpar ratios when compared with industry standards and trends. He concluded that two significant ratios were Net Worth to Debt and Net Profits to Net Worth. He also focused on asserting less importance on the Quick Ratio and Current Ratio for firms with long-term liabilities.

Beaver (1966) took this approach further by gauging individual ratios' predictive abilities in classifying bankrupt and non-bankrupt firms. He used 30 ratios for 158 firms, half of which were bankrupt. The companies consisted of 38 different industries. He concluded that Net Income to Total Debt constituted of the greatest predictive power at 92% accuracy which was followed by Net Income to Sales at 91%. He also found that 3 ratios had the same predictive power at 90% accuracy which were Cash Flow to Total Debit, Cash Flow to Total Assets and Net Income to Net Worth.

Both previous mentioned studies are univariate studies. Univariate studies are a type of statistical analysis that focuses on examining and describing a single variable within a dataset. These studies aim to understand the basic properties and distribution of that one variable without considering any relationships with other variables. For a deeper understanding of univariate studies, please refer to the work by Cleff and Cleff (2014).

The study by Altman (1968) was the first multivariate study for bankruptcy prediction. He used the "Z-score" model which indicated the likelihood of bankruptcy if a firm's Z-score fell within a specific range. He created the five-factor model specifically for bankruptcy for manufacturing firms. Initially the model gave a 95% predictive accuracy for bankruptcy one year before the firm failed. However accuracy significantly declined to 72%, 48%, 29% and 36% for two, three, four and five years, respectively.

A multivariate study involves the simultaneous analysis of more than two variables to understand the relationships between them and how they jointly influence an outcome. This type of study is particularly useful in research fields where complex interactions between variables are examined. For deeper understanding of multivariate studies, please refer to the work by Tabachnick, Fidell, and Ullman (2013).

The study by Ouenniche and Tone (2017) was one of the earliest where machine learning was implemented. It introduced an out-of-sample evaluation framework for assessing the performance of Data Envelopment Analysis (DEA) models. Their framework was designed to address the issue of overfitting in DEA models by validating their predictive power on data not used during the model development process by dividing the data into training and test sets. This approach provided more robust and reliable predictions by ensuring that the model's performance is not overly tailored to the training data. The results of this study included a 0% Type 2 error and 100% sensitivity.

2.2- Causes of Bankruptcy

This literature review will focus on causes of bankruptcy that are not mostly related to financial ratios and how other factors influence bankruptcy.

Some studies focus on a single factor to prove that they standalone impact bankruptcy. In a study by Baum and Mezias (1992), they explore how the density and proximity of competing hotels influenced the likelihood of organizational failure of another hotel. They concluded that hotels were more likely to fail when surrounded by a high concentration of competitors in proximity, highlighting the significance of localized competition in determining organizational outcomes within a specific geographic area.

A study by Greening and Johnson (1996) focuses on management and strategic decisions during times of organizational crisis. They research whether actions and strategies implemented by managers significantly impact the outcomes of an organization in a critical situation. Through an analysis of various case studies and data, they conclude that both managerial decisions and strategic approaches play crucial roles in navigating crises. Hence it is suggested that poor management and ill-considered strategies increase the likelihood of a crisis and decrease chances of recovery.

Similarly a study done by Swaminathan (1996) examines how conditions present at the time of an organization's founding influence its likelihood of bankruptcy over time. The author proposes a "trial-by-fire" model which suggests that organizations built under hardships develop resistance and hence reduce their risk or mortality. The research analyzes survival rates of organizations and concludes that firms built under hostile environments have lower probability of bankruptcy in the long run.

Some studies focus on the interaction of certain factors that might cause bankruptcy. The study by Back (2005) focuses on three main categories of predictors which include Previous Payment Behavior, Management Background Variables and Financial Ratios. By focusing on these factors, the study aimed to identify patterns and indicators that could signal a financial distress. The findings concluded that past payment behaviors such as missed or delayed payments is a strong distress indicator. Additionally certain management characteristics like Experience, Education, Tenure and key financial ratios like Liquidity Ratios, Leverage Ratios, Profitability Ratios and Cash Flow Ratios play a key role and explaining bankruptcy.

Another study done by Bradley and Rubach (2002) explores the role of trade credit in the financial health of a small businesses. Trade credit is referred to accumulate where suppliers allow businesses to purchase goods or services on credit. The report concludes that trade credit is often essential for operations of a small business however it can lead to financial instability if not managed properly. Some of the issues created because of poor trade credit management include overextension of credit, late payments and disrupted cash flow.

2.3- Implications Bankrupt Firms

When a firm enters financial distress, it can either go through bankruptcy reorganization where a formal court order is passed when company declares bankruptcy or settle a restructuring agreement with creditors outside of court (Donaldson et, al. 2020). An out of court agreement usually leads to coordination problems amongst creditors (Morris and Shin, 2004). Hence when bankruptcy reorganization process starts and the firm declares bankruptcy, courts are required to give a legal order to sell or distribute assets of the insolvent firm so that loan terms, bank recovery rates and leverage ratios can be maintained (Acharya, Rangarajan, and Kose, 2008). After such a crisis, firms report that keeping employees becomes a concern as they start to rapidly leave the firm

(Berk, Stanton, and Zechner, 2010). Additionally it becomes very difficult for a firm to protect its reputation because of which it must convince clients, trade creditors and suppliers to carry on doing business so that the crisis does not escalate (Epaulard and Zapha, 2022). For the case of smaller firms, lenders require that owners give personal collateral on the loan, hence owner's personal assets get liquidated because of the firm's bankruptcy (Berkowitz and White, 2004). This essentially means that limited liability does not apply on such debts. In countries that might lack bankruptcy frameworks, a single bankruptcy could lead the owner to have a lifetime of debt and barring them from doing business again (Armour and Cumming, 2008). Going bankrupt does not only cause loss of capital owned by the firm but additional costs are incurred. In a study by Branch (2002), costs of a bankrupt firms are classified as Real costs (borne directly by the bankrupt firm), Real costs (borne directly by the claimants), Losses to the bankrupt firm that are offset by gains to other entities, Real costs (borne by parties other than the bankrupt firm and/or its claimants).

3- Methodology

The following is a flowchart of the major technical tasks carried out.

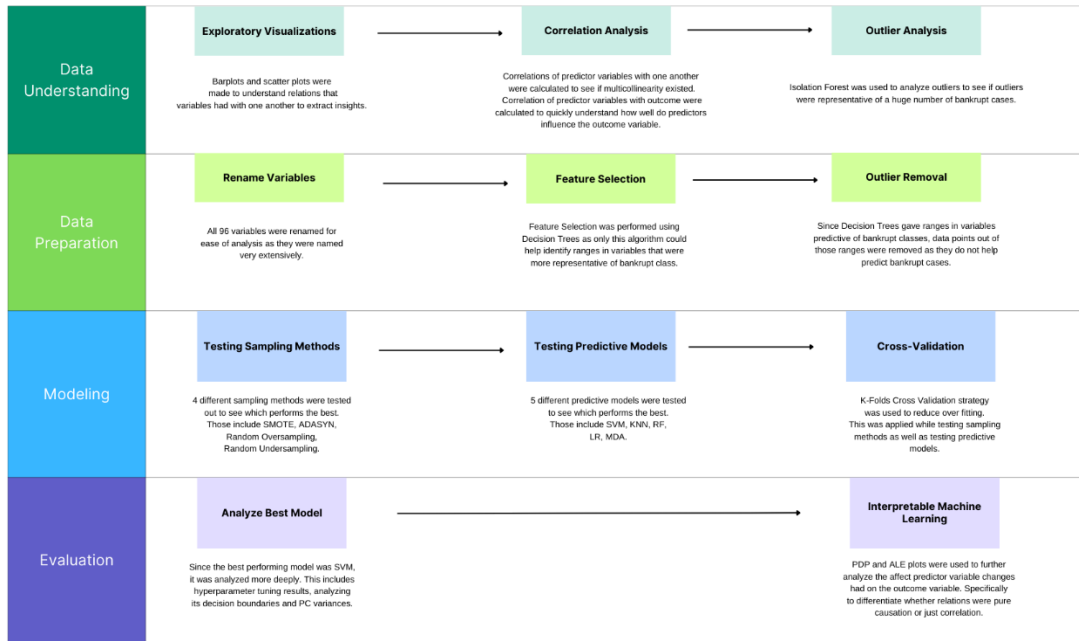


Figure 1: Technical Flowchart

For this study, the Cross-Industry Standard Process for Data Mining (CRISP-DM) research framework will be adopted. This is done to ensure that the study is easily replicable while being comprehensive in an orderly fashion. The following figure displays the six essential stages of CRISP-DM.

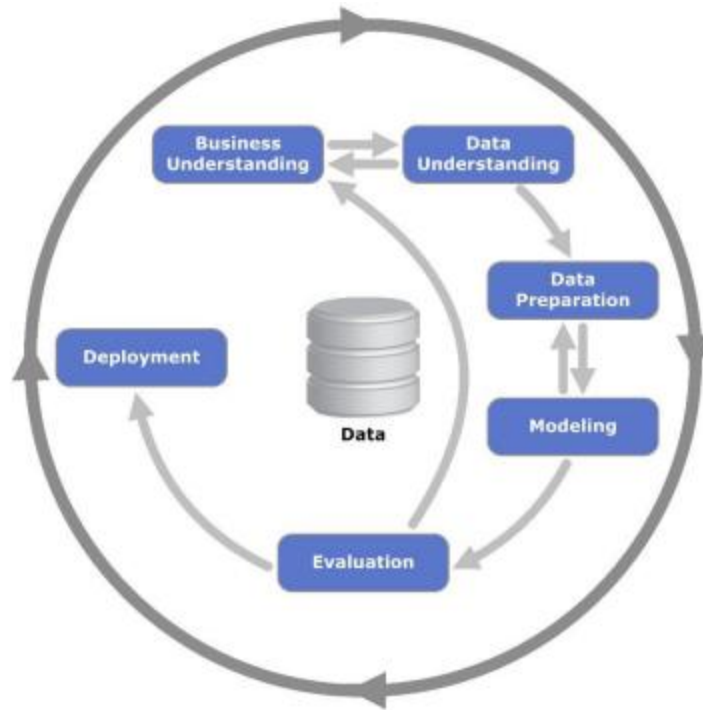


Figure 2: CRISP-DM Model

(Source: Data Science Process Alliance (2018))

3.1- Business Understanding

This is crucial business context developed in the “Literature Review” section in this report.

3.2- Data Understanding

The data used is from the UCI Machine Learning Repository about Financial Ratios for 6,819 Taiwanese companies as rows and 96 variables (UCI Machine Learning Repository, 2020). The target variable is binary indicating whether a company goes bankrupt or not and the input variables are continuous numeric values. The data was originally collected from the Taiwan Economic Journal for the years 1999 to 2009. Company bankruptcy was defined based on the business regulations of the Taiwan Stock Exchange. For more details on the dataset and metainformation, refer the citation (UCI Machine Learning Repository, 2020).

3.3- Data Preparation

Our dataset was generally very clean. It did not have any missing values that needed to be treated. However since each column or financial ratio was very extensively named because of its nature, we replaced 95 names of the variables with letters which spanned from A to CQ. For a list of what variables was replaced by what letters, please refer to the “Appendix” of the Technical Report.

The data preparation step of CRISP-DM is further divide into two: Feature Selection and Outlier Removal.

3.3.1- Feature Selection

Since our analysis could not revolve around all 95 variables, we had to perform a feature selection technique to reduce the number of variables. For that reason we decided to use Decision Trees as a feature selection algorithm. This approach had also been taken in studies by Gayatri et al. (2010), as well as Wang and Li (2008). For further details, refer to the “Feature Selection” heading in the technical report.

3.3.2- Outlier Removal

For dealing with outliers present in the data, we used the same decision tree model created for feature selection purposes. This approach is very similar to that taken by John (1995) where he also mentions that advanced machine learning algorithms are not as effective at addressing the outlier problem compared to the manual approaches by certain statistical methods like decision trees.

We also aided our outlier analysis with an Isolation Forest Algorithm. The Isolation Forest algorithm is a machine learning technique used for anomaly detection, particularly effective in

identifying outliers in large datasets. Unlike traditional clustering algorithms, Isolation Forest works by isolating observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature. For more details on the algorithm, refer the work by Liu, Ting, and Zhou (2008).

For more details on the whole process, refer to the “Outlier Removal” section in the technical report.

3.4- Modeling

Since we found out that there is a class imbalance in our dataset for the target variable, we will apply certain Sampling Methods to resolve that issue.

We will use certain Predictive Models to determine which works best in our study.

We also use Cross-Validation to reduce overfitting in the models.

Hence the modeling section in the CRISM-DM will be further dividing into Sampling Methods, Predictive Modeling and Cross-Validation.

3.4.1- Sampling Methods

The following are all the sampling methods that we tested throughout our study. For more details on rationale for selection of each sampling method, please refer the “Sampling Methods” heading in the technical report.

3.4.1.1- SMOTE (Synthetic Minority Oversampling Technique)

SMOTE works by generating synthetic instances of the minority class to balance the class distribution. This is achieved by randomly selecting a data point from the minority class and finding its k-nearest neighbors. For more details, refer to the work by Chawla et al. (2002).

3.4.1.2- ADASYN (Adaptive Synthetic Sampling)

Like SMOTE, ADASYN generates synthetic data points to balance the distribution between majority and minority classes. However, ADASYN goes a step further by adaptively generating more synthetic data for the minority class instances that are harder to learn, i.e., those that are closer to the decision boundary. For more details, refer to the work by He et al. (2008).

3.4.1.3- Random Oversampling

The method works by randomly duplicating instances from the minority class until the class distribution is balanced. While simple, ROS can effectively improve model performance by providing the classifier with more opportunities to learn from the minority class. For more details, refer to the work by Dummond and Holte (2003).

3.4.1.4- Random Under-sampling

The method works by randomly removing samples from the majority class until the class distribution is balanced. This approach helps prevent the model from being biased towards the majority class, leading to more accurate predictions for the minority class. For more details, refer to the work by Kubat and Matwin (1997).

3.4.2- Predictive Modeling

The following are the predictive models that we tested throughout our study. For more details on the rationale for selecting each model, please refer the “Predictive Modeling” heading in the technical report.

3.4.2.1- MDA (Multiple Discriminant Analysis)

Multivariable Discriminant Analysis (MDA) is a statistical method used for classifying observations into predefined groups based on multiple predictor variables. The technique identifies a linear combination of these variables that best separates the classes, creating a discriminant function that maximizes the variance between the groups while minimizing the variance within each group. For more details, refer to the work by Hair (2009).

3.4.2.2- SVM (Support Vector Machines)

SVM works by finding the hyperplane that best separates the data into different classes, maximizing the margin between the nearest data points of the classes, known as support vectors. This method is particularly effective in high-dimensional spaces and is robust against overfitting. For more information, refer to the work by Pedregosa et al. (2011).

3.4.2.3- KNN (K-Nearest Neighbor)

KNN operates on the principle that similar data points are likely to be near each other in the feature space. When classifying a new data point, the algorithm identifies the 'k' closest points (neighbors) in the training dataset and assigns the most common class among these neighbors to the new point. For more details, refer to the work by Zhang (2016).

3.4.2.4- LR (Logistic Regression)

Logistic Regression is a fundamental statistical method used for binary classification tasks, where the goal is to model the probability that a given input belongs to one of two possible classes. Logistic regression uses the logistic function to map predicted values to a probability between 0 and 1. For more details, refer to the work by Hosmer, Lemeshow, and Sturdivant (2013).

3.4.2.5- RF (Random Forest)

Random Forest (RF) is an ensemble learning method that is widely used for both classification and regression tasks. It works by constructing a multitude of decision trees during training and outputting the class that is the mode of the classes (for classification). The key advantage of RF is its ability to improve predictive accuracy by reducing overfitting, which is a common issue in single decision trees. For more details, refer to the work by Breiman (2001).

3.4.3- Cross-Validation

K-fold cross-validation is a robust and widely adopted technique in machine learning for evaluating the performance and generalization ability of models. The process involves dividing the dataset into 'k' equal-sized subsets, or folds. The model is trained on 'k-1' folds and validated on the remaining fold. This process is repeated 'k' times, with each fold acting as the validation set once. The average of the results from these iterations provides a more accurate estimate of the model's performance on unseen data. For more details, refer to the study by Raschka and Mirjalili (2019).

In our study, we had cross-validated our models 10 folds 3 times. So we received a total of 30 results for each model from which the best model was selected. We not only applied cross-validation to our Predictive Models, but we applied cross-validation during evaluating Sampling Methods as well.

3.5- Model Evaluation Metrics

We will use certain Evaluation Metrics that determine which model works best. Generally, classification problems have fewer evaluation metrics as compared to regression problems (Moreau and Wassermann, 2024).

Since we were only interested in the predictive accuracy of one class of the target variable, we did not focus on evaluation metrics that indicate the overall predictive accuracy of the model. In our study, the bankrupt class was the negative class which was denoted by 1 or X1. Hence the following were the metrics used to evaluate each model.

We will also use certain Interpretable Machine Learning Methods to explain changes in variables with respect to target variable more deeply.

3.5.1- TNR (True Negative Rate)

The True Negative Rate (TNR), also known as specificity, is a critical evaluation metric in binary classification tasks in machine learning. It measures the proportion of actual negatives that are correctly identified by the model. In other words, TNR is the ratio of true negatives (TN) to the sum of true negatives and false positives (FP).

$$TNR = \frac{\text{True Negatives (TN)}}{\text{True Negatives (TN)} + \text{False Positives (FP)}}$$

For more details, refer to the study by Xu, Zhang, and Miao (2020).

3.5.2- NPV (Negative Predicted Value)

It measures the proportion of true negatives among all instances that were predicted as negative by the model. In other words, NPV indicates how well a model can correctly identify negative cases, thereby assessing its reliability in predicting the absence of a condition or class. NPV is particularly valuable in scenarios where false negatives need to be minimized, such as bankruptcy cases, where missing a negative case could have severe consequences.

$$NPV = \frac{True\ Negatives\ (TN)}{True\ Negatives\ (TN) + False\ Negatives\ (FN)}$$

For more details, refer to the study by Xu, Zhang, and Miao (2020).

3.5.3- F2 Score

As opposed to F1, the F2 score gives more weight to recall, which makes it suitable for applications where false negatives are more critical than false positives, which is the case in our study. This metric is calculated using a formula that adjusts the balance between precision and recall, emphasizing recall twice as much as precision.

$$F2 = (1 + 2^2) * \frac{Precision * Recall}{4 * Precision + Recall}$$

Or:

$$F2 = \frac{5 * TP}{5 * TP + 4 * FP + FN}$$

For more details, refer to the study by Sokolova and Lapalme (2009).

3.5.4- ROC (Receiver Operating Characteristics) Curve

The Receiver Operating Characteristic (ROC) curve is a graphical representation used to evaluate the performance of a binary classification model. It plots the true positive rate (sensitivity) against the false positive rate (1-specificity) across different threshold values. The ROC curve provides insight into the trade-off between sensitivity and specificity. AUC (Area

Under Curve) will also be utilized which numerically depicts how well the tradeoff between the measures plotted on the ROC Curve is balanced. For more details, refer the study by Fawcett (2006).

The above explanation for ROC Curve is better suited if our interested class is the positive class. Since the bankruptcy class in our case is negative, hence we would need to alter out ROC curve to fit our needs. We would have it plot true negative rate (TNR or Specificity) against the false negative rate (1-Sensitivity or FNR) so that the results of the ROC curve are more impactful for us.

3.5.5- Interpretable Machine Learning

We are going to use two Interpretable Machine Learning methods which include PDP (Partial Dependency Plots) and ALE (Accumulated Local Effects). For more details about the rationale of using these methods, refer to the “Interpretable Machine Learning” heading in the technical report.

3.5.5.1- PDP (Partial Dependency Plots)

PDPs help understand the relationships between the target variable and one or more features in a model. PDPs show the marginal effect of a feature on the predicted outcome, while averaging out the effects of all other features. This makes PDPs useful for understanding the impact of a single feature in isolation, particularly in complex, non-linear models like gradient boosting machines or random forests. For more details, refer the study by Friedman (2001).

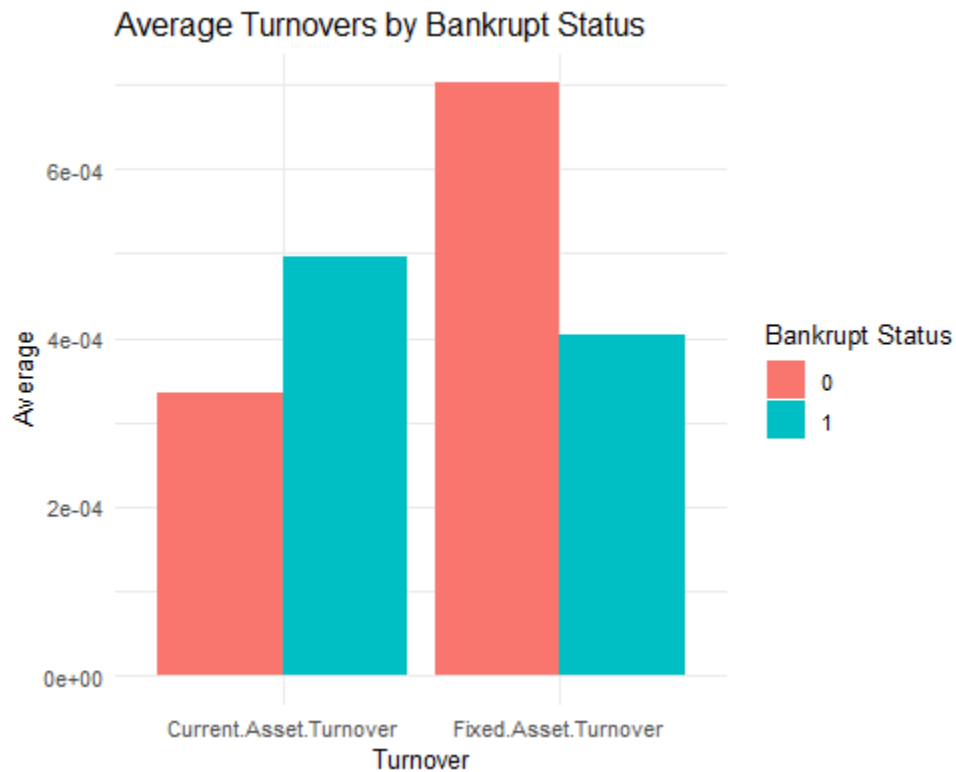
3.5.5.2- ALE (Accumulated Local Effects)

Accumulated Local Effects (ALE) plots are an advanced method for interpreting machine learning models, particularly useful in cases where Partial Dependence Plots (PDPs) might be

misleading due to interactions between features. Unlike PDPs, ALE plots focus on the local effects of features by computing changes in the prediction as the feature varies, while keeping other features fixed. This makes ALE plots more accurate for capturing the true effect of a feature, especially in models where features interact. For more details, refer to the work of Apley and Zhu (2020).

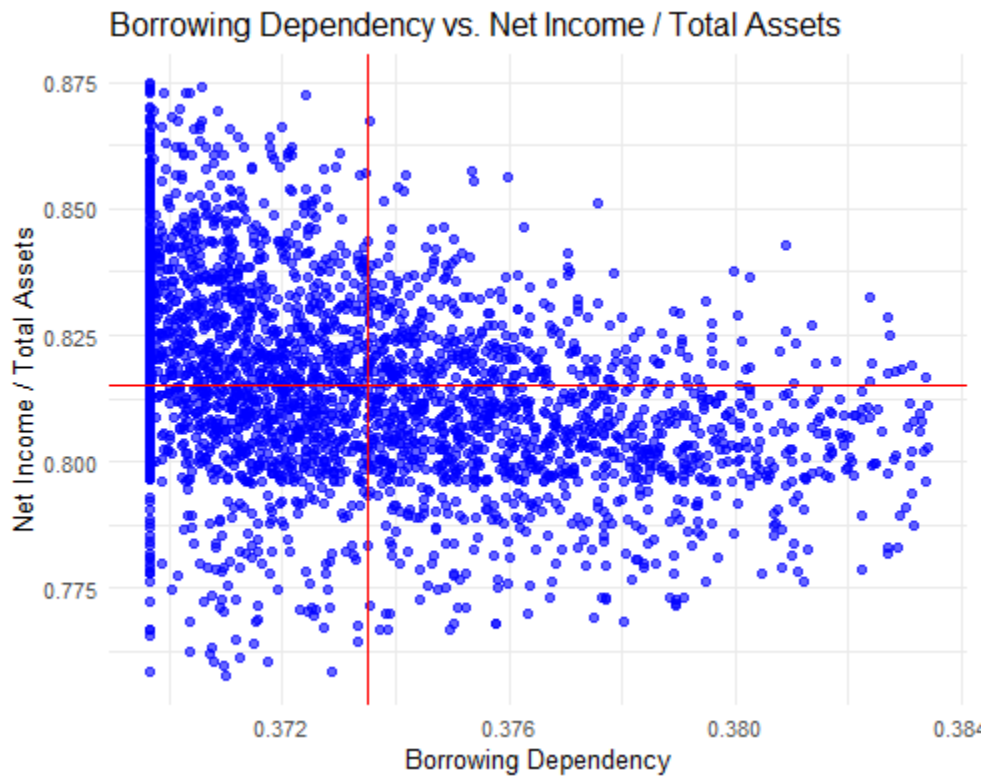
4- Findings

4.1- Exploratory Visualizations



The bar chart here is very clearly indicating how bankrupt companies tend to focus more on short term productivities compare to long term goals and strategies. This could imply that bankrupt companies aggressively use their current assets to generate revenue, possibly indicating financial stress or inefficiency in managing current assets. Stronger companies less likely to bankrupt tend

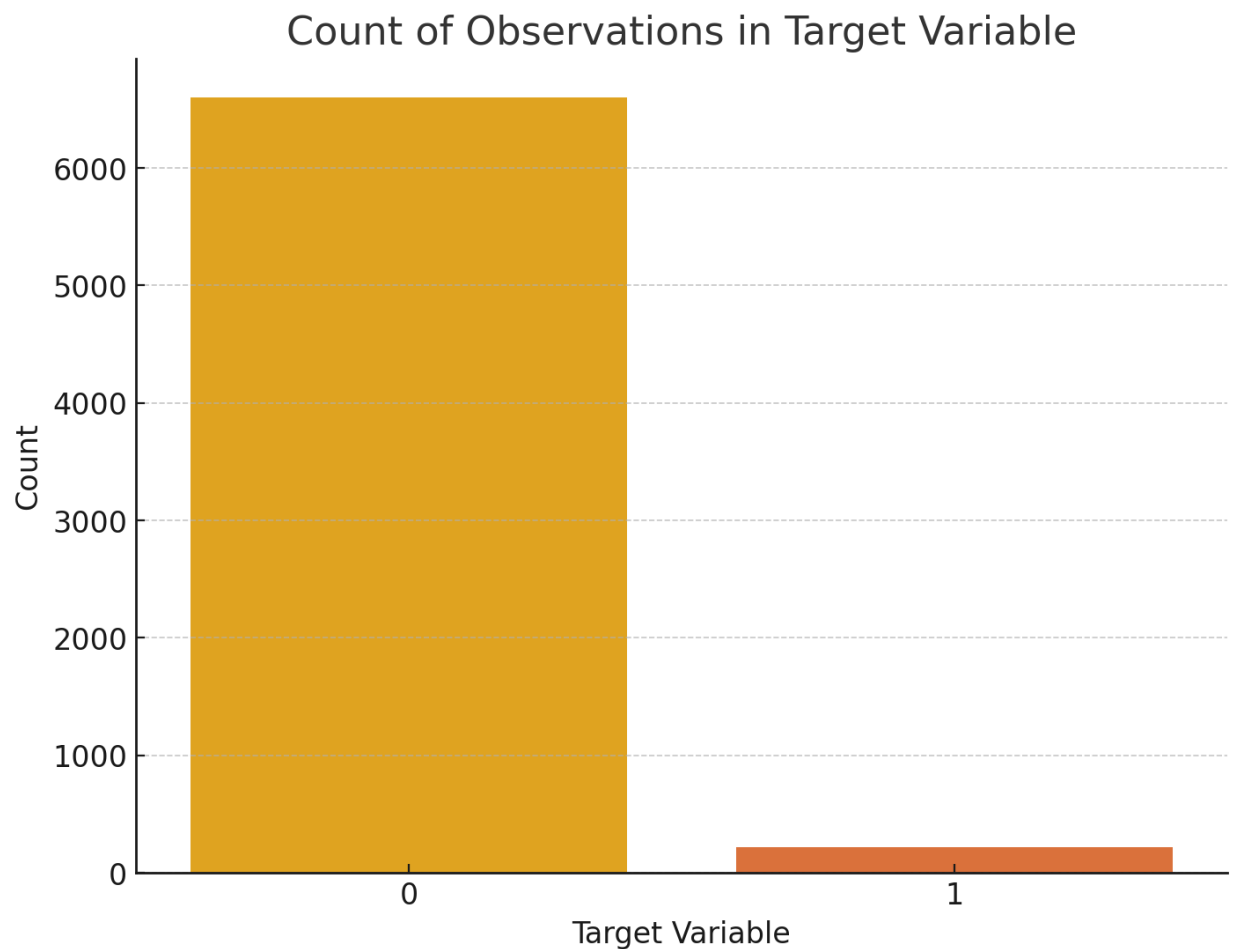
to have a much higher fixed asset turnover and a lower current asset turnover which might be indicative of financial stability. Note that the difference between the turnovers for non-bankrupt companies is high compared to bankrupt companies.



The red lines cutting both axis represent the average values. For a Taiwanese company in general, 37.3% percent of their assets will be liabilities which indicates that Taiwan might not have a culture of depending too much on debt. We see an interesting “fawning in” shape of the scatter plot. This implies that as liabilities take up a greater portion of a company’s assets, its range of relative net income becomes smaller. Generally there are no companies that are high in debt and high in net income. We also see that a greater portion of companies have a higher net income to total assets ratio irrespective of borrowing dependency specially at lower levels of borrowing dependencies.

We first found out that there is a strong class imbalance in our dataset. “0” represents not bankrupt and “1” represents bankrupt.

```
> summary(as.factor(df$bankrupt))  
    0     1  
6599  220
```



Being 220 observations for “1” (Bankrupt Cases) and 6599 observations for “0” (Non-bankrupt Cases), Bankrupt cases only constitute 3.2% of the whole observations and non-bankrupt cases

constitute of 96.8% of the dataset. This makes it evident at this stage that sampling techniques will be required to get reliable modeling results.

4.2- Correlation Analysis

Since we have a huge number of variables, it is very difficult to show correlation pairs for variables and display that in a matrix. Hence we created a distribution of correlations between variables and plotted it on a boxplot.

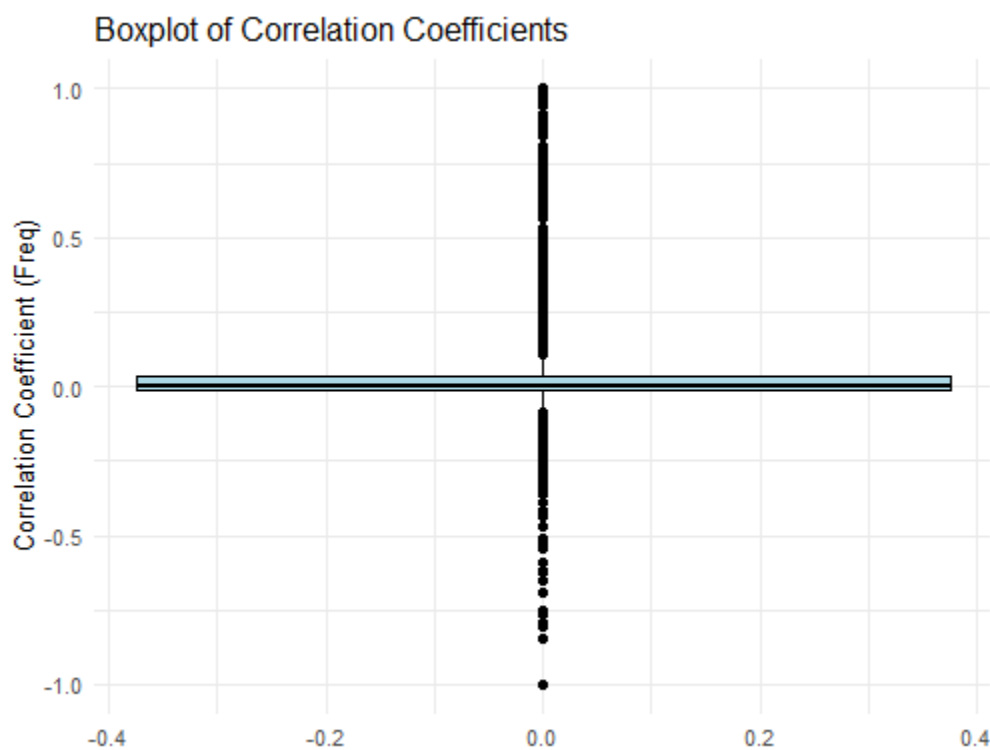


Figure 3: Boxplot of Predictor Variable Pair Correlations

To supplement this graph, we extracted summary statistics of the distribution of correlations.

```
> summary(correlation_pairs$Freq)
      Min.   1st Qu.   Median     Mean   3rd Qu.     Max.    NA's
-1.00000 -0.01190  0.00155  0.03710  0.03590  1.00000   188
```

It is very evident that we have a lot of correlation pairs with very high correlations as seen in the boxplot. We can see that even though the average or majority of the variables are not highly correlated as the average correlation is 0.03710, there is still an unignorable number of high correlations seen on the boxplot. Hence moving forward we would need to remove variables with high correlations to prevent the issue of multicollinearity in the models.

Similarly boxplot of correlations of predictor variables with target variables were created. This is also supported by summary statistics.

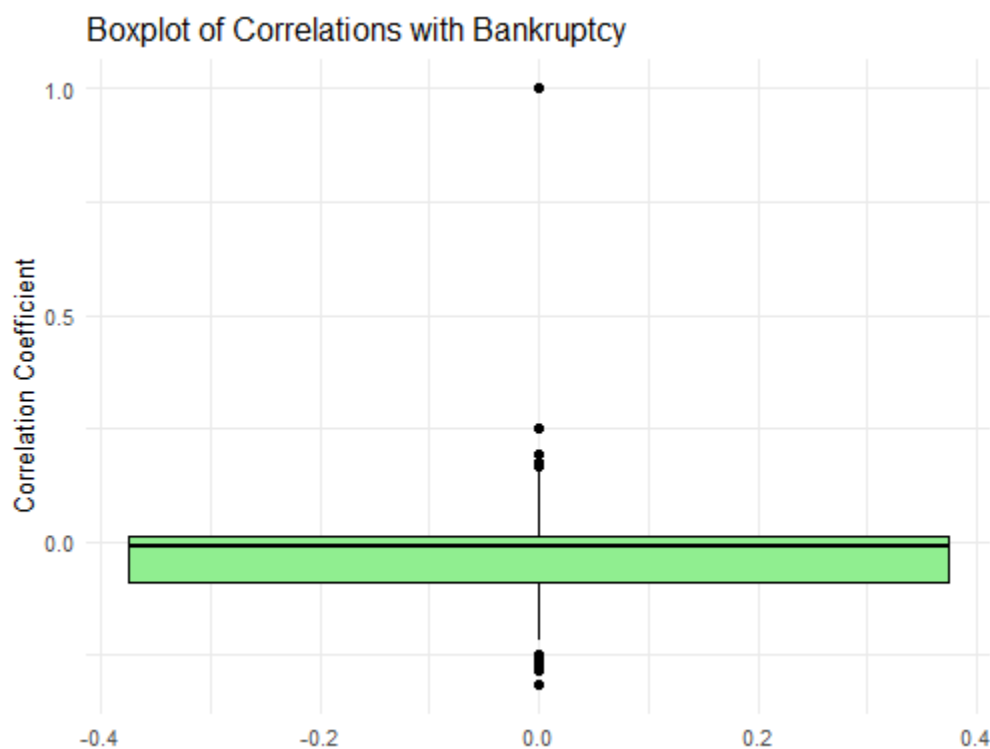


Figure 4: Boxplot of Correlations of Predictors with Bankruptcy

```
> summary(bankruptcy_correlations$Correlation)
      Min.   1st Qu.   Median     Mean   3rd Qu.    Max.     NA's 
-0.315457 -0.089612 -0.008857 -0.021360  0.012121  1.000000      1
```

On the other hand, correlations with the outcome variable for each of the predictor variables are not strong either as displayed by this boxplot of correlations. Summary statistics with a low average correlation and table of correlations also show the same. The point with 1.0 correlation coefficient is bankruptcy and can be ignored.

We created the following table to see what variables are most correlated with the outcome.

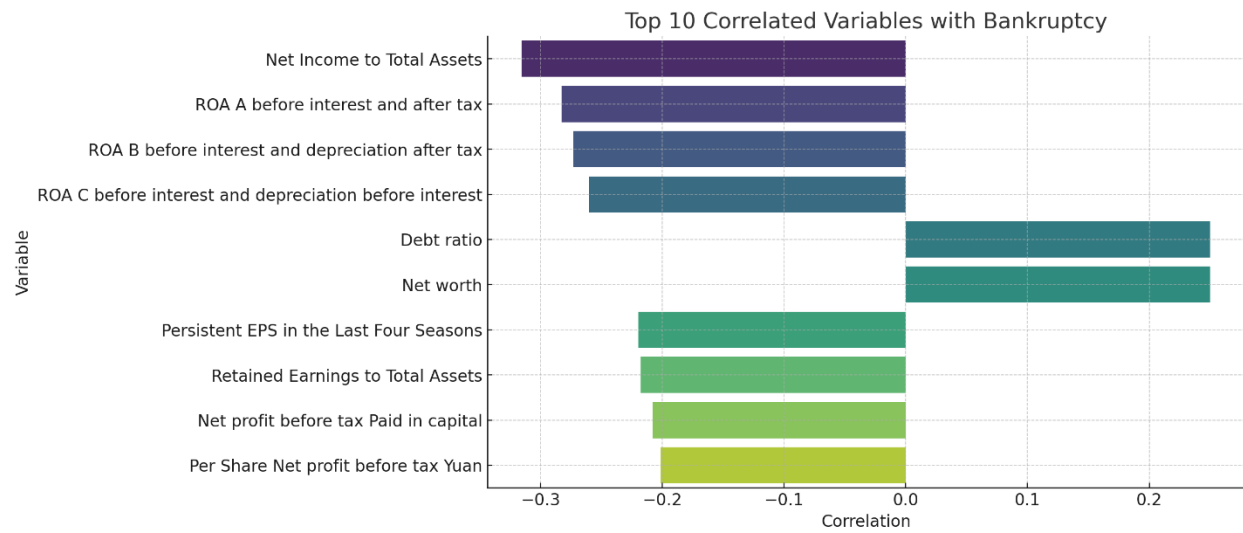
```
> print(bankruptcy_correlations)
```

	Variable	Correlation
1	Bankrupt.	1.0000000000
2	Net.Income.to.Total.Assets	-0.3154569716
3	ROA.A..before.interest.and...after.tax	-0.2829405849
4	ROA.B..before.interest.and.depreciation.after.tax	-0.2730513179
5	ROA.C..before.interest.and.depreciation.before.interest	-0.2608065575
6	Debt.ratio..	0.2501609621
7	Net.worth.Assets	-0.2501609621
8	Persistent.EPS.in.the.Last.Four.Seasons	-0.2195596812
9	Retained.Earnings.to.Total.Assets	-0.2177787800
10	Net.profit.before.tax.Paid.in.capital	-0.2078565200
11	Per.Share.Net.profit.before.tax..Yuan...	-0.2013948345
12	Current.Liability.to.Assets	0.1944944359
13	Working.Capital.to.Total.Assets	-0.1930833758
14	Net.Income.to.Stockholder.s.Equity	-0.1809869884

We can see that the highest correlation with the outcome is for the variable

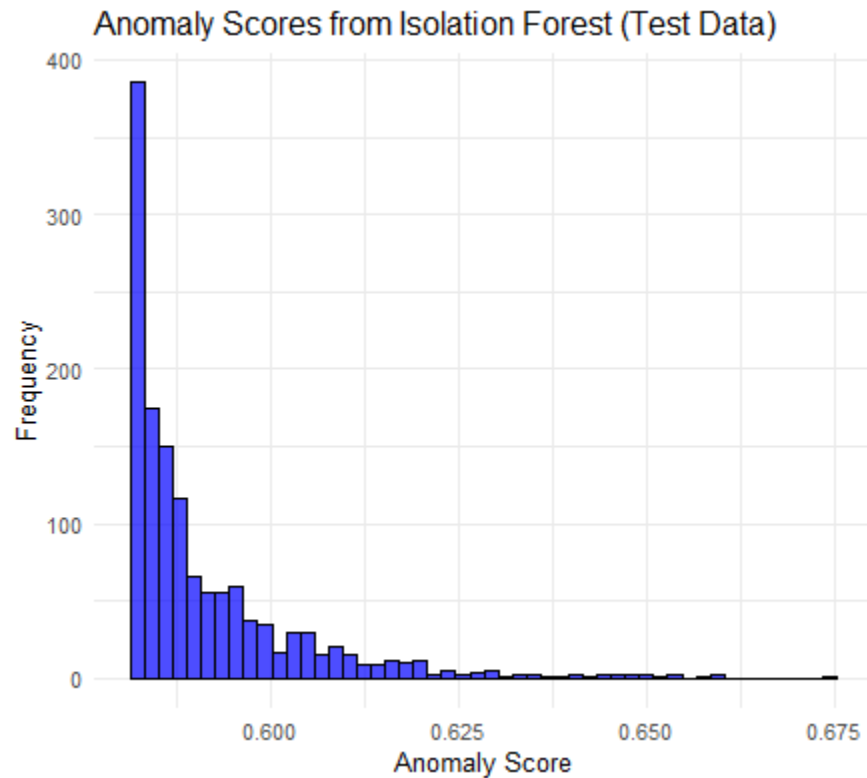
“Net.Income.to.Total.Assets” at -0.315 which is not a strong correlation.

The following is a table for the top 10 variables highly correlated with the outcome variable.

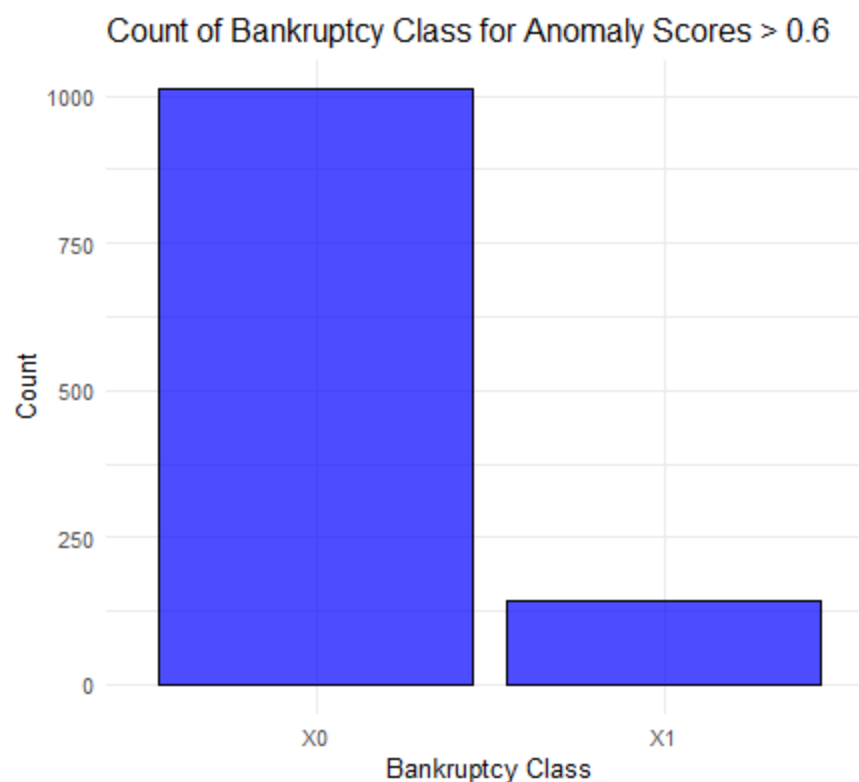


4.3- Outlier Analysis using Isolation Forest

Outliers in this study hold a great importance as anomalies in the data could be representative of bankrupt cases. Hence performing Isolation Forest gave the following results.



High anomaly scores with lower frequencies represent outliers. We cannot omit a very high percentage of our data after classifying them as outliers because that will cause loss of valuable information. However we can focus on data with anomaly score greater than 0.6 as that constitutes of only 15% of our data.

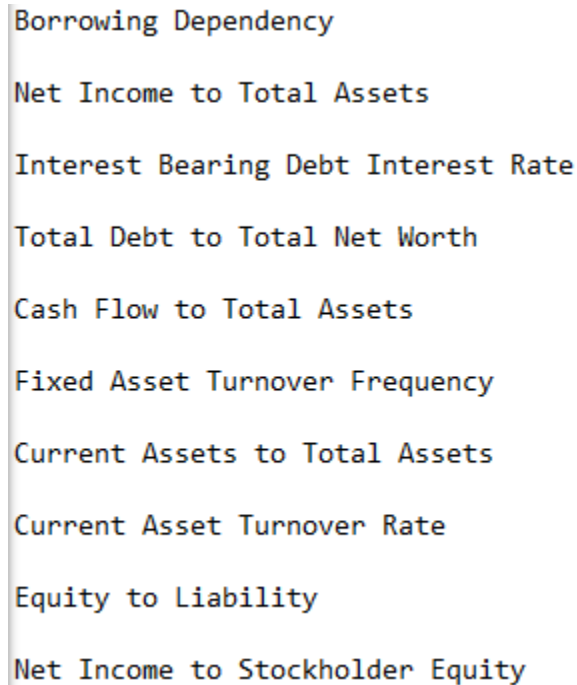


Here X1 represents bankruptcy and X0 represents non-bankruptcy. In the total dataset we have 220 bankrupt cases. This plot shows that 125 of those bankrupt cases are in outliers, which is over 50% of all the bankrupt cases. So despite having a lot of outliers in the dataset, we cannot remove them as it will remove a lot of valuable information about bankruptcy.

4.4- Feature Selection Using Decision Trees

After feature selection technique was applied to extract most important variables (refer to “Feature Selection” heading in the technical report), there were 18 variables extracted. These are the financial ratios highly predictive of bankruptcy according to the feature selection algorithm. After all variables were manually observed and removed, the following were the final 10 variables selected. Refer to the “Feature Selection” heading in the technical report for more details. Note that for the selected method of feature selection, there is no numerical way to

quantify significance of each variable to the outcome. Further discussion on variables is present in the “Discussions” heading in this report.

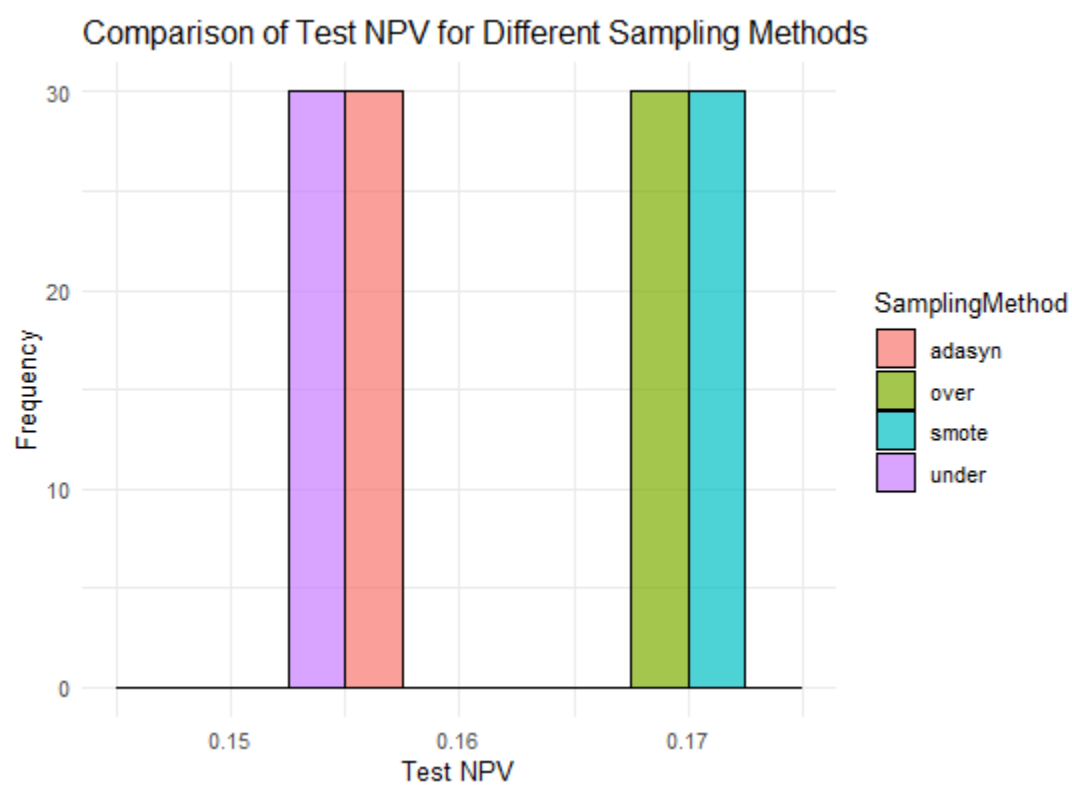
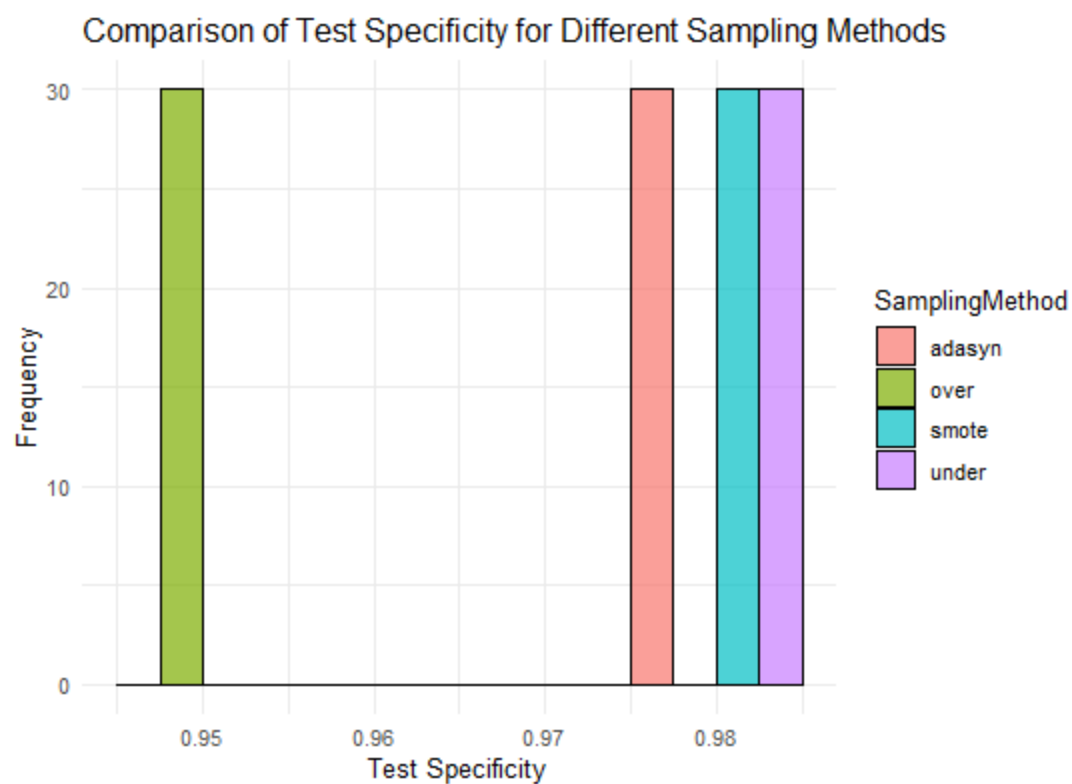


- Borrowing Dependency
- Net Income to Total Assets
- Interest Bearing Debt Interest Rate
- Total Debt to Total Net Worth
- Cash Flow to Total Assets
- Fixed Asset Turnover Frequency
- Current Assets to Total Assets
- Current Asset Turnover Rate
- Equity to Liability
- Net Income to Stockholder Equity

Figure 3: List of Variables Selected for Modeling

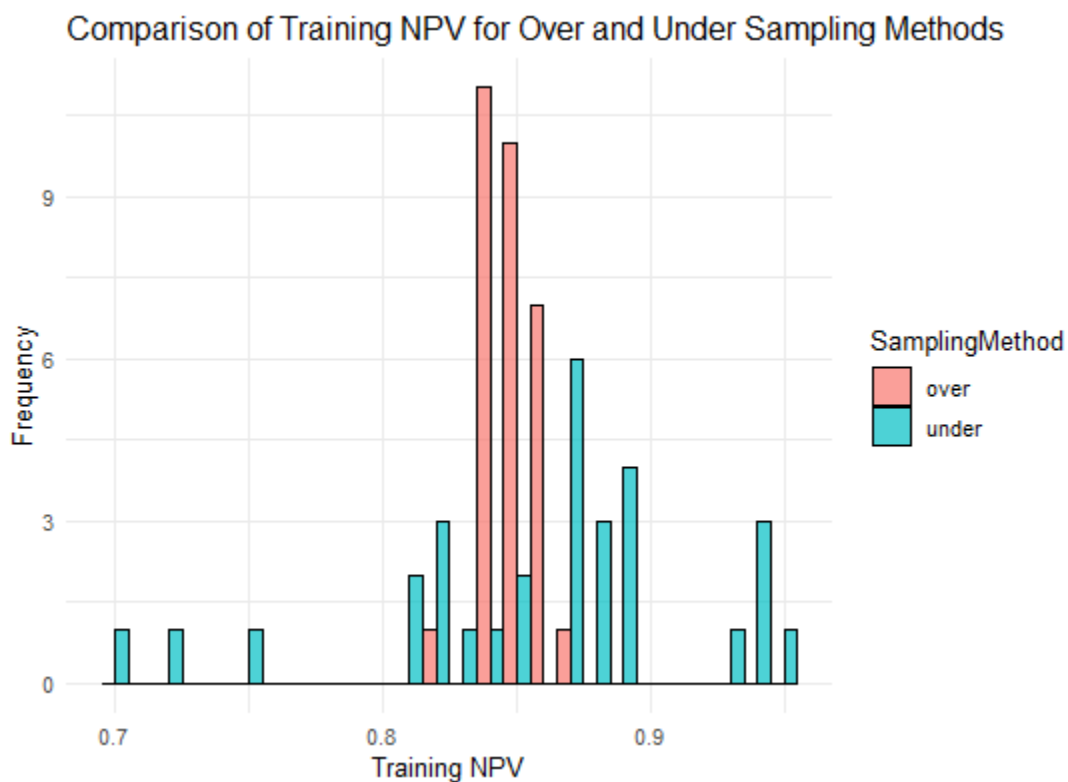
4.5- Testing Sampling Methods

The following are test accuracies for different sampling methods. “Frequency” on the y-axis being 30 implies that for all 30 occurrences of cross-validation (10 folds, 3 times giving a total of 30 results), test accuracies were the same.

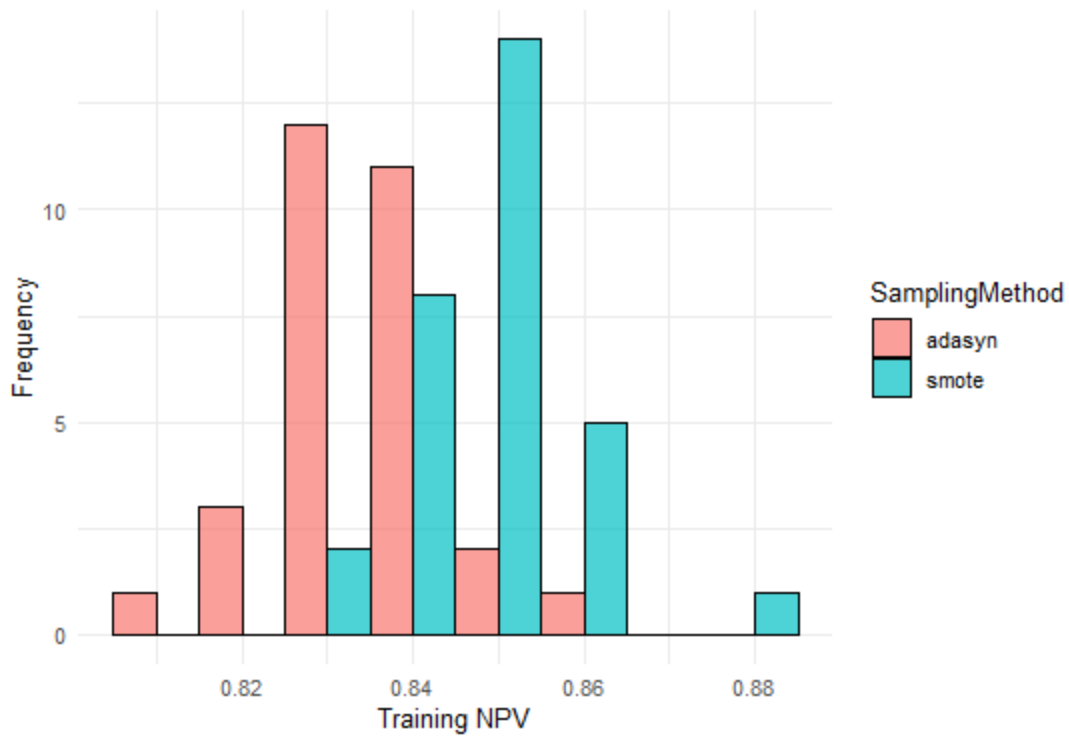


For each of the sampling techniques, the specificity of the model is high which represents the model's ability to correctly identify the total cases of bankruptcy. However NPV of each model is not as high since classes are not as separable in the data. This represents that models trained are always going to be more biased towards predicting the negative class (bankruptcy) and in return give a high False Positive error.

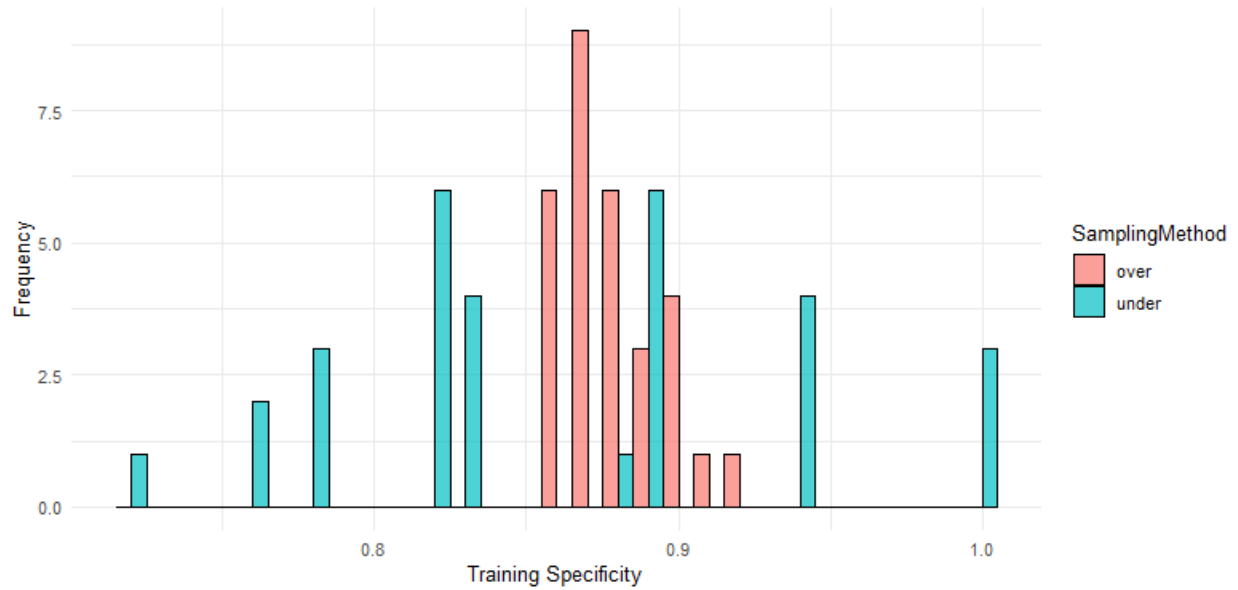
We also have results for training accuracies to put into perspective how well the model learned patterns in the dataset.

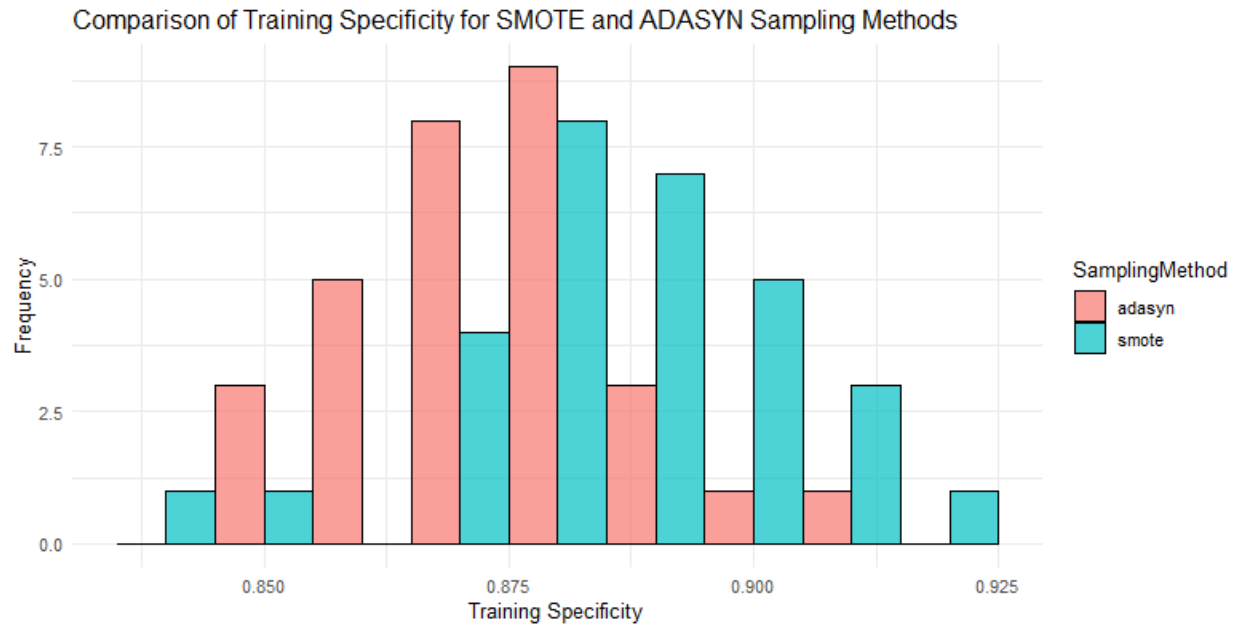


Comparison of Training NPV for SMOTE and ADASYN Sampling Method



Comparison of Training Specificity for Over and Under Sampling Methods





In these plots, since all 30 results of cross-validation did not give the same accuracy, we see a distribution of accuracies in training accuracies.

Every sampling technique is getting a high accuracy for both training specificity and training NPV. Since only testing NPV is low for each model, this represents that the model is getting overfitted only in terms of NPV and not specificity despite cross-validating results. This is again due to the classes in the dataset being difficult to separate.

We will choose SMOTE as our sampling technique because we are getting the highest NPV with the second highest specificity. Every model is getting a high specificity so the deciding factor is highest NPV because that makes our model least overfitted than other models.

4.6- Testing Predictive Models

The following are relevant evaluation metrics for each of the models.

	MDA	SVM	KNN	LR	RF
Specificity (test)	0.8409	0.9545	0.3864	0.9545	0.5682
Specificity (train)	0.767	0.8636	0.7784	0.875	1
NPV (test)	0.1588	0.1787	0.1771	0.175	0.2941
NPV (train)	0.1487	0.1617	0.3948	0.1578	1
F2 Score	0.876522	0.8797814	0.9475776	0.8759181	0.9586323
Area Under The Curve (AUC)	0.9360052	0.957871	0.7056224	0.9549418	0.9314219

Figure 6: Evaluation Metrics for Predictive Models

Interesting to see that Non-parametric models like KNN and RF have very low-test specificity, indicating that they are not good at identifying bankrupt cases. LR is performing well in terms of test specificity, it means there may be simple relations in the dataset. We can see that RF is getting overfitted on the training set as we get a perfect accuracy on them but poor accuracy on test set. Plus it is the only model giving a high test neg pred value. But since its test specificity is very poor, having high test negative predictive value is considered not as important. Increasing number of trees to 1000 for RF caused no difference in results.

KNN (0.9476) and RF (0.9586) have the highest F2 scores, indicating that these models are better at minimizing false negatives compared to the others. But again as they have lowest test specificity of 0.3864 for KNN and 0.5682 for RF, their ability to minimize false negatives is completely overshadowed as they are incompetent at detecting negative cases or bankrupt companies. KNN also has the lowest AUC of 0.705 which again suggests that it is not effective at distinguishing between positive and negative classes.

SVM (0.9579) and LR (0.9549) have the highest AUC values, indicating strong discriminatory power between the classes and suggesting that these models are generally well-calibrated.

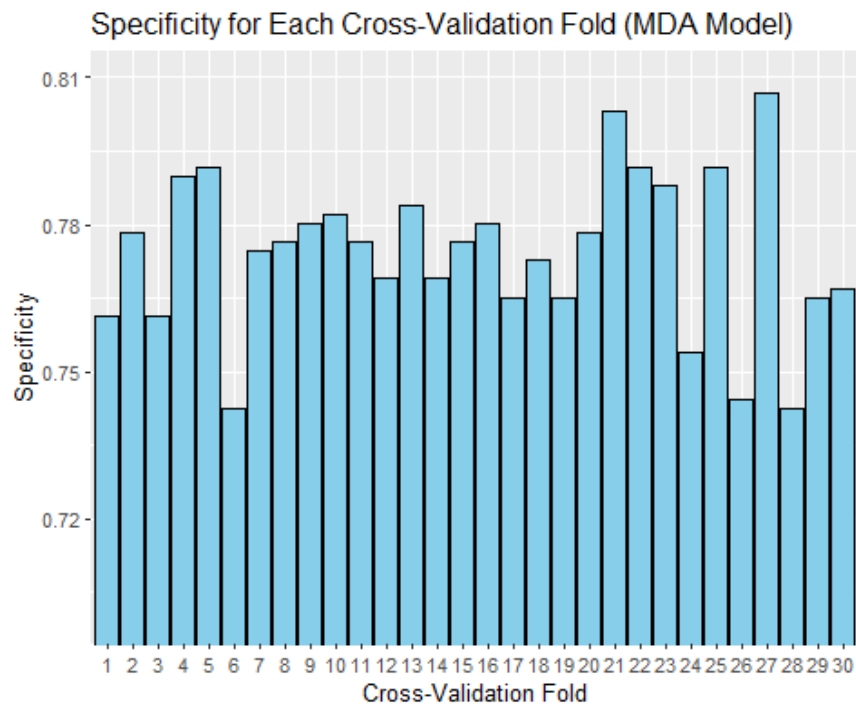
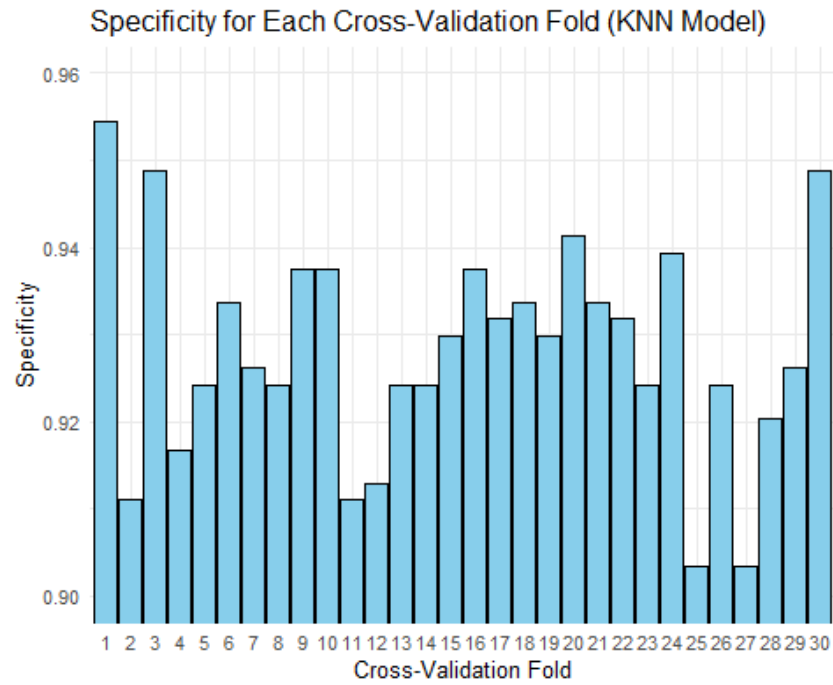
SVM and LR are also the two models with the highest test specificity indicating the highest competence at detecting bankrupt cases. The test specificity for SVM and LR is 0.9545. Even though RF has the highest test NPV, it is not impactful as the model has very poor test specificity. Hence the next best models with highest test NPV are also SVM and LR. SVM has a test NPV of 0.1787 and LR of 0.175.

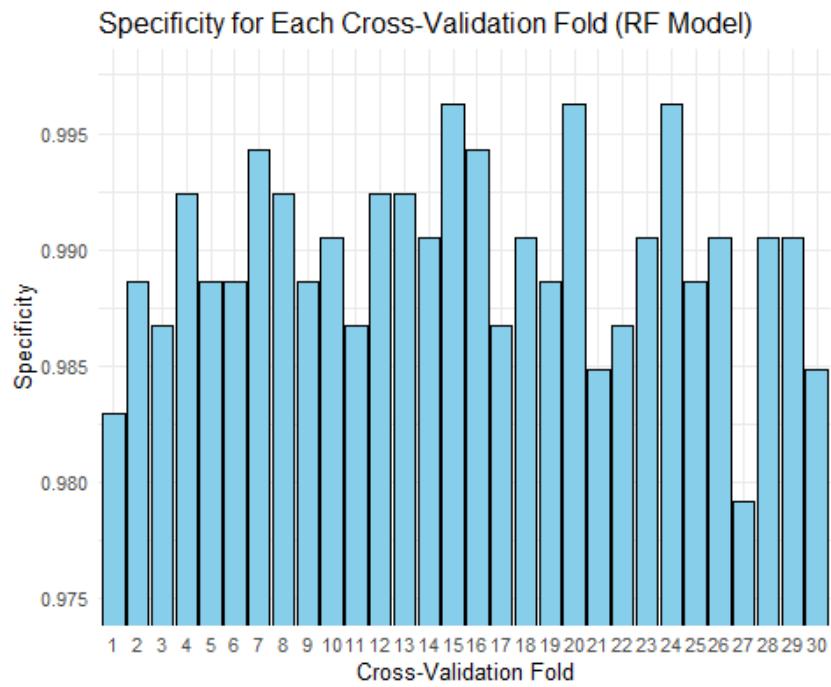
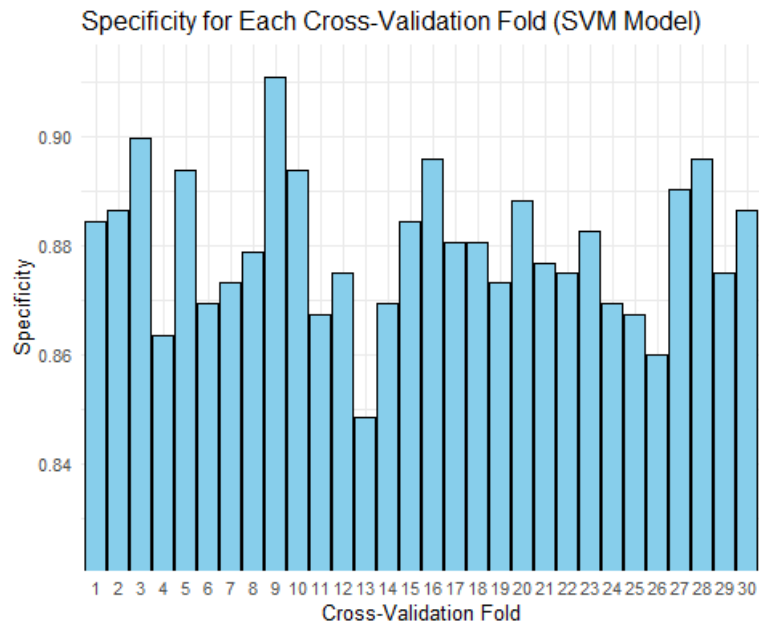
So moving forward we are choosing SVM as our model as it offers the highest test NPV of 0.1787 and test specificity of 0.9545. The reason why we did not choose LR is because SVM is only performing marginally better than LR on test NPV.

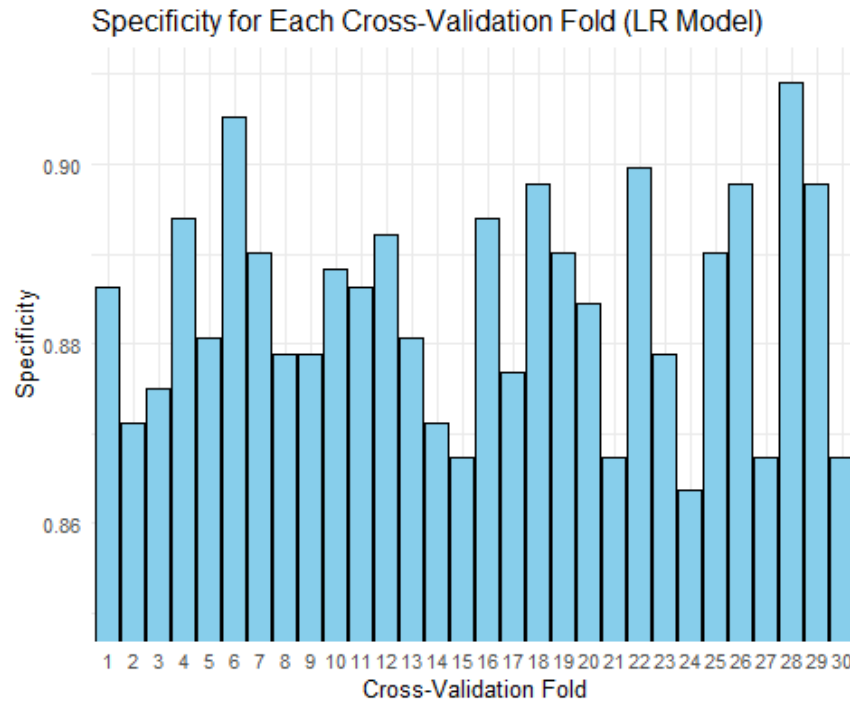
The following is the confusion matrix for SVM corresponding to its accuracies shown in the table.

Prediction	Reference	
	X0	X1
X0	1126	2
X1	193	42

The following are distributions of cross-validation results for each of the models. Results in the table above for training accuracies are averages of the cross-validated accuracies.







4.7- Further Analyzing Best Model

Since we chose SVM as our best model, we will further analyze it. The following are the coefficients and intercept of the model.

```
> print(svm_coefficients)
          AN          CH          N          AJ          CB          AW          BD          BS          CQ          CL
[1,] -0.2850222 -2.548604 -0.2707409  2.187711 -0.3162389  0.1666704 -0.1555692 -0.1112216 -1.60929  2.39801

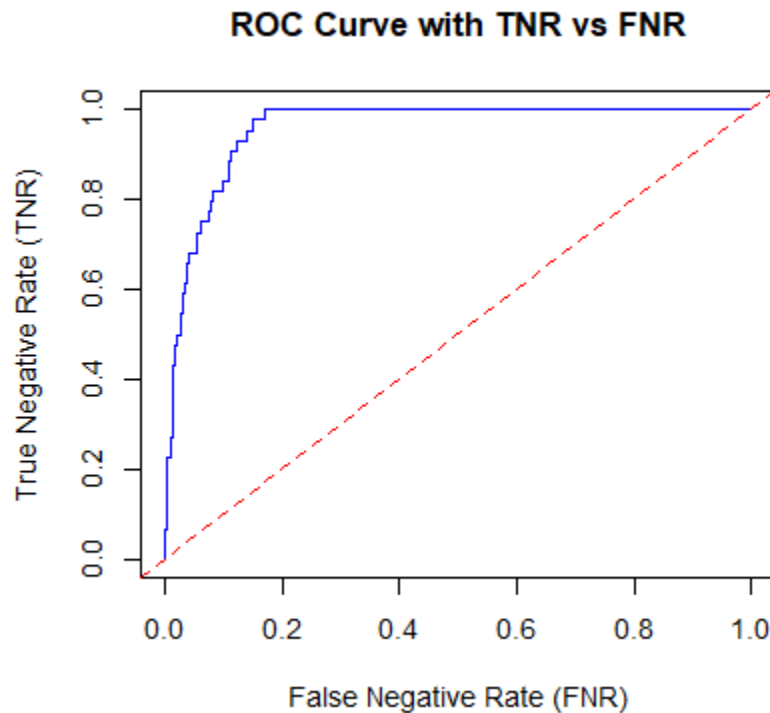
> print(svm_intercept)
[1] 0.0621981
```

If a feature has a negative coefficient, an increase in this feature will decrease the value of the linear combination, moving the prediction towards the negative class (X1: bankrupt). The magnitude (absolute value) of the coefficient indicates the strength of the feature's influence on the decision boundary. Hence features that are heavily contributing towards bankruptcy by their increase are Net Income to Total Assets (CH) and Equity to Liability (CQ). This is interesting as generally

higher net income and equity are indicative of a healthy business. Features that are heavily contributing towards non-bankruptcy are Net Income to Stockholder Equity (CL) and Total Debt to Total Net Worth (AJ). Total Debt to Total Net Worth is also interesting as generally greater debt is an indicator of financial distress.

The intercept in an SVM model with a linear kernel (often referred to as the bias term) is a value that shifts the decision boundary away from the origin in the feature space. An intercept of 0.0621981 means that the decision boundary is slightly shifted in favor of predicting the non-bankrupt class (X0). However, since the intercept is quite small, it has a minimal impact compared to the feature coefficients.

The following is the ROC curve for SVM.



Traditionally ROC Curves have True Positive Rate on the Y-axis and False Positive Rate on the X-axis. Since we are more concerned with True Negative Rate as Bankruptcy is our negative class, we will alter the ROC curve to show True Negative Rate on the Y-axis and False Negative Rate on the X-axis.

The ROC curve indicates that the model is performing well at identifying bankruptcy (the negative class). The TNR is high, meaning that the model correctly identifies most bankrupt companies, and the FNR is low, meaning that it rarely misclassifies bankrupt companies as non-bankrupt. Based on this curve, the model seems reliable for distinguishing bankrupt from non-bankrupt companies, which is crucial if the primary concern is to avoid missing any bankrupt companies.

The following are results for hyperparameter tuning using cross-validation.

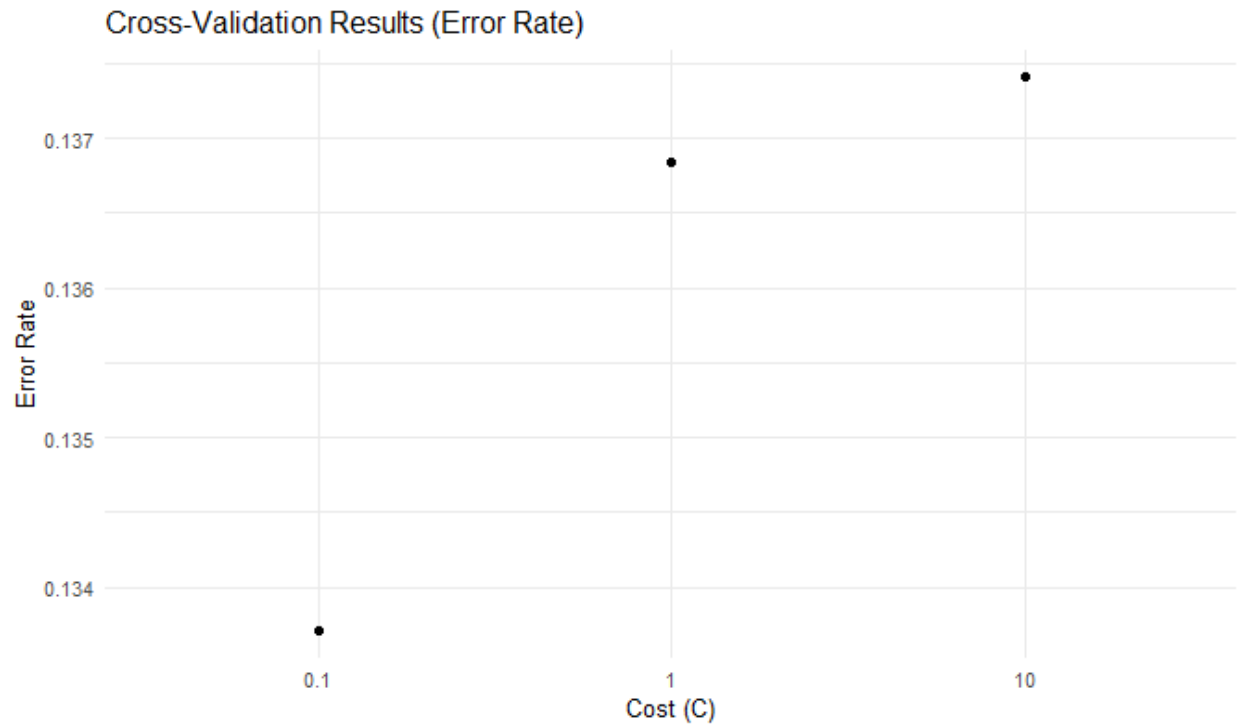
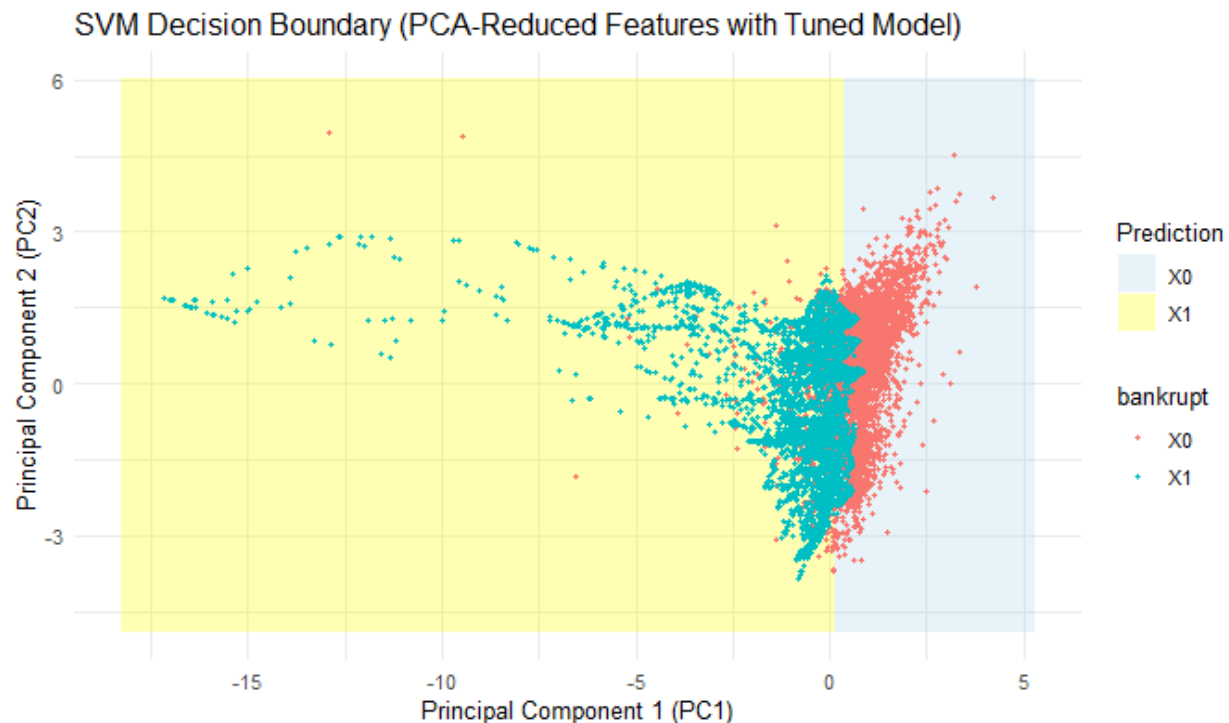


Figure 7: Cross-Validation Results for Cost Tuning for SVM

The cost parameter C controls the trade-off between maximizing the margin and minimizing the classification error in SVM. A smaller C encourages a wider margin but allows for more classification errors, while a larger C aims to classify all training examples correctly by potentially allowing a narrower margin. The error rate on the y-axis represents the proportion of incorrectly classified instances during cross-validation. The cost value of 0.1 has the lowest error rate, around 0.134. This suggests that the SVM model performs best with a lower cost value, balancing the trade-off between margin width and classification accuracy effectively.

The following is the decision boundary of the SVM model.



The decision boundary created by the SVM model divides the PCA space into two regions. The decision boundary is vertical, meaning that the first principal component (PC1) plays a significant role in the classification. Most of the non-bankrupt points (represented in red) lie within the light blue area, indicating that they are correctly classified by the model. However, there is a significant overlap in the central region where many of the points labeled "X1" (bankrupt, in teal) are also located within the light blue area, meaning they are misclassified as non-bankrupt. The yellow region on the left contains points classified as "X1" (bankrupt). However, some points labeled as "X0" (non-bankrupt) are within this region, suggesting they are misclassified as bankrupt by the model. There are also some points in the blue region on the right that are labeled as "X1" but are predicted as "X0," indicating misclassification of bankrupt cases. The plot indicates that while the SVM model generally separates the classes well, especially in

the extremities (far left and far right of the plot), there is significant overlap in the center, leading to misclassifications.

The following are values for variance explained for each of the Principal Components created for the decision boundary of the SVM model.

```
> print(pca_variance_table)
  PCA  Variance_Explained  Cumulative_Variance_Explained
1  PC1          0.22482645                0.2248264
2  PC2          0.15465257                0.3794790
3  PC3          0.11436696                0.4938460
4  PC4          0.10259737                0.5964434
5  PC5          0.09995073                0.6963941
6  PC6          0.08797808                0.7843722
7  PC7          0.07763947                0.8620116
8  PC8          0.06198851                0.9240001
9  PC9          0.05084860                0.9748487
10 PC10         0.02515127                1.0000000
```

Figure 8: Variance Explained by Principal Components for Decision Boundaries

This table shows that two Principal Components are not sufficient to explain the variation in the data. It will take 7 Principal components to explain at least 80% variation in the data.

4.8- Interpretable Machine Learning Using PDP and ALE

Our focus now is to compare PDP with ALE plots for each variable. If both plots are similar, that means changes in probability caused by change of the variable is only because of the change of variable (causation). If plots are not the same, that means changes in probability are not occurring because of change in variable only and there may be other variables that maybe causing those changes (correlation). Those other variables are most likely removed during the feature selection stage.

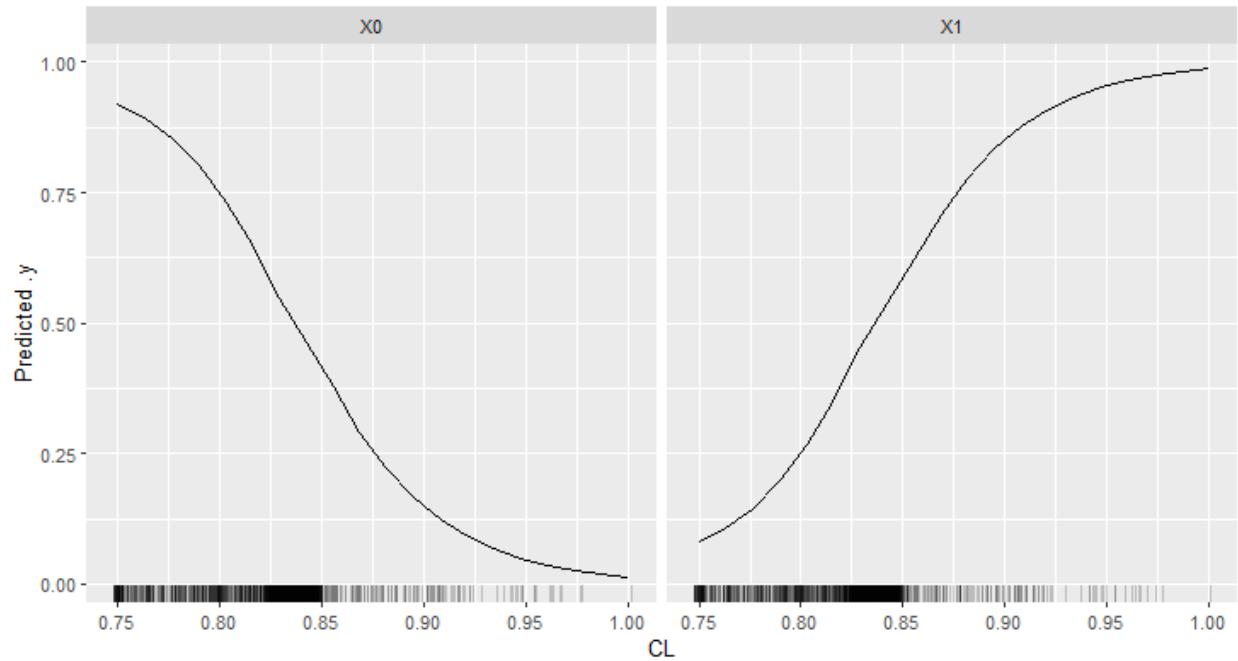


Figure 9: Net Income to Stockholder Equity PDP

This is one of the highest contributing variables. For CL (Net Income to Stockholder Equity) most of the values are less than 0.85 which is giving a lesser probability for bankruptcy. There are fewer companies with values higher than 0.85 and for them the probability for bankruptcy is much higher. For companies with CL of 0.85, the probability of bankruptcy is approximately 63%. For a CL of 0.9, the probability of bankruptcy is 88%. The highest probability of bankruptcy is 100% at a value of 1.00 for CL.

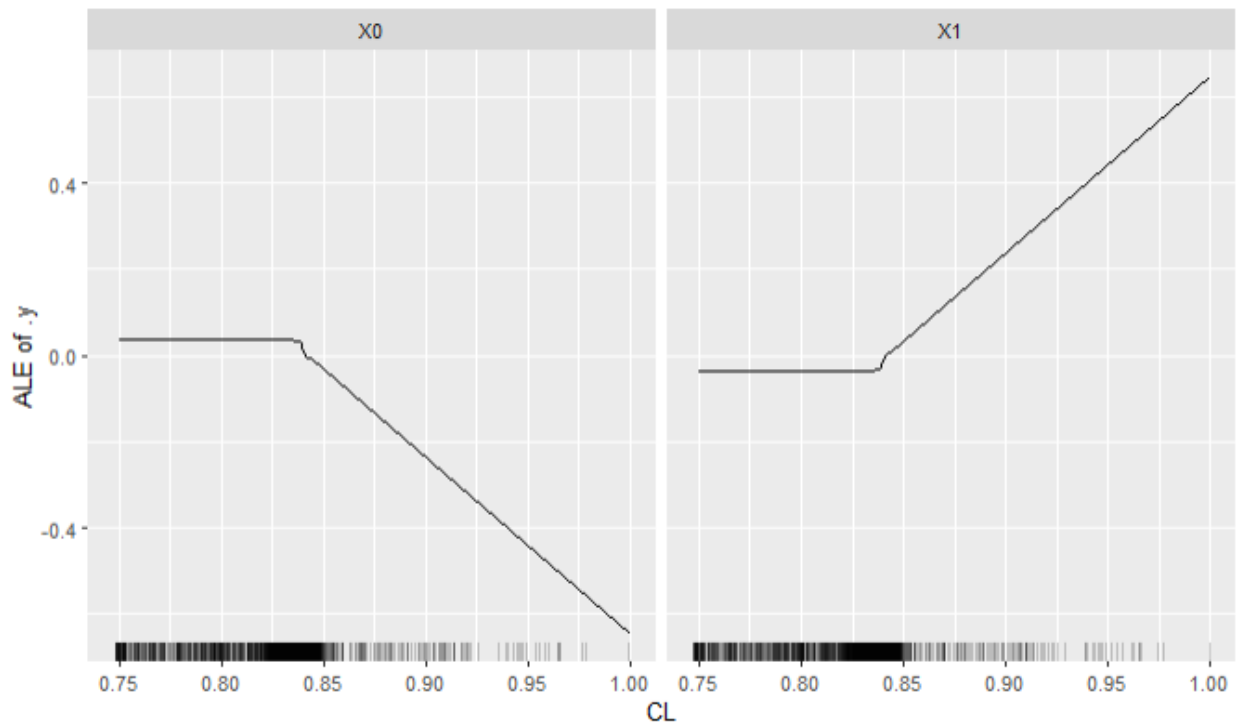


Figure 10: Net Income to Stakeholder Equity ALE

Changes in CL (Net Income to Stakeholder Equity) do play a role in predicting bankruptcy. However both plots follow a same general trend but they are not exactly similar. That means even though solely CL changes are causing bankruptcy changes (causation), there are other variables also correlated to CL causing the changes of bankruptcy.

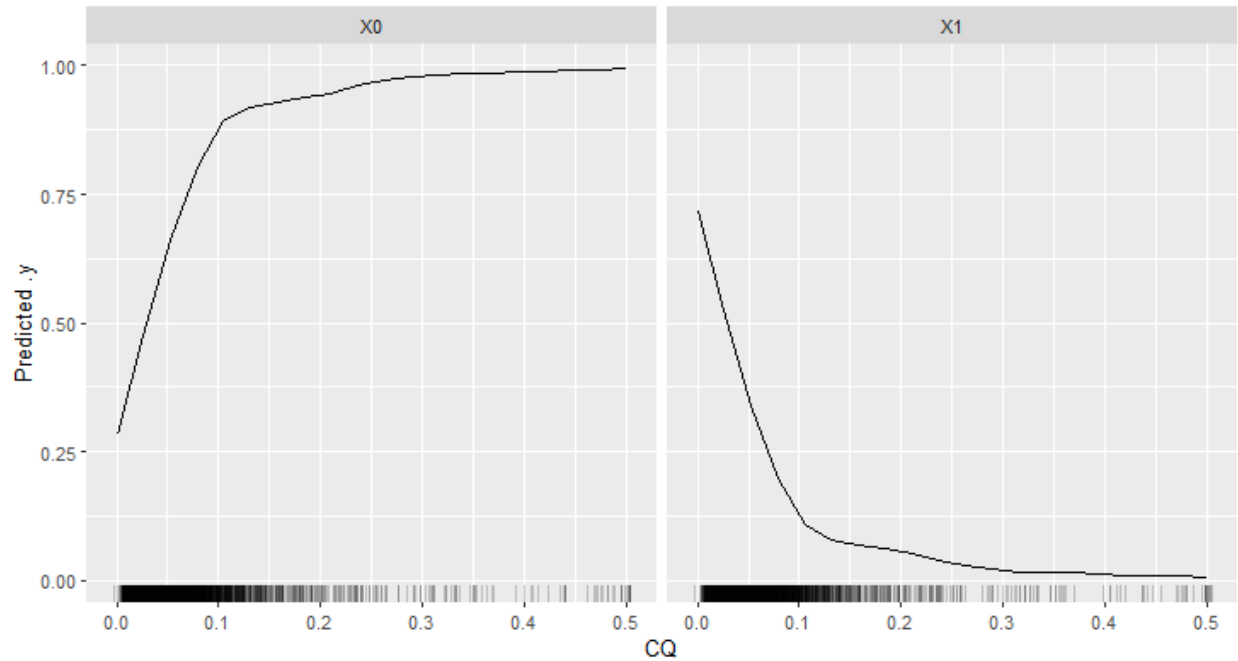


Figure 11: Equity to Liability PDP

CQ (Equity to Liability) is another highly contributing variable. The highest probability of bankruptcy of 75% is when Q is 0. There is a steep drop in bankruptcy probability as the value of CQ increases and soon experiencing a decreased rate of decrease as CQ becomes greater than 0.1.

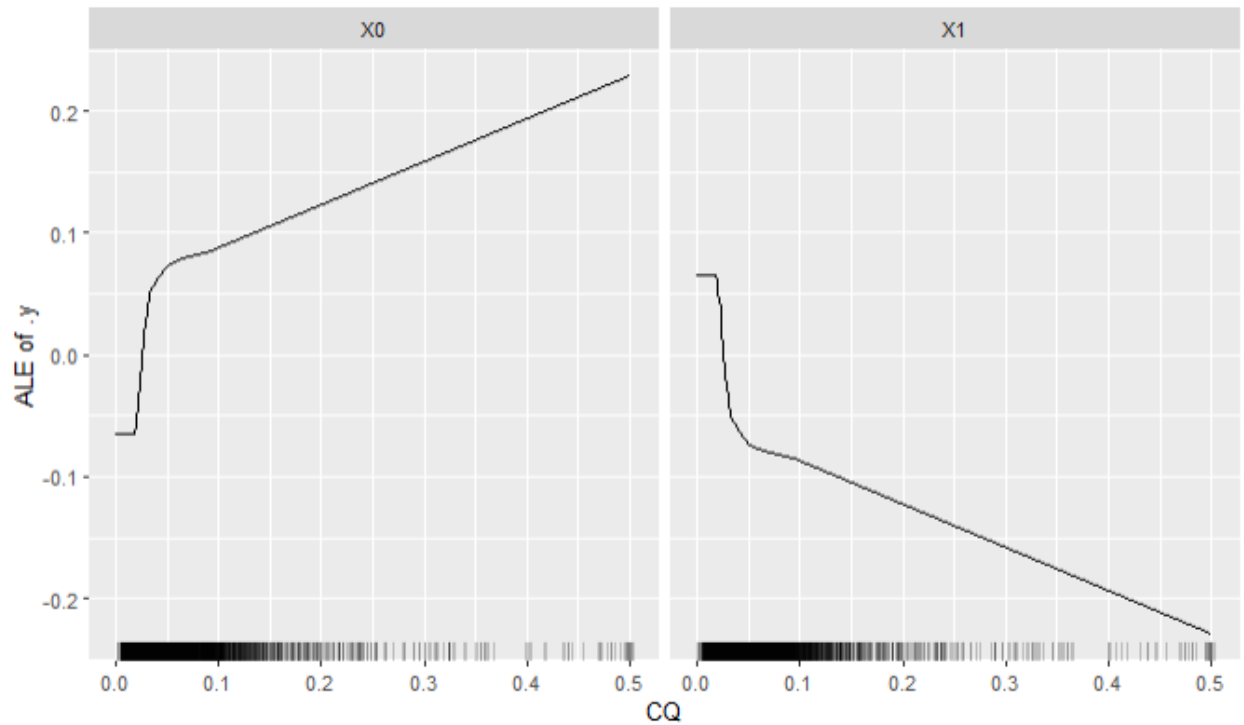


Figure 12: Equity to Liability ALE

Changes in bankruptcy by change in CQ (Equity to Liability) are because of CQ (causation) and possibly because of other correlated variables as both plots are not exactly similar but causing the same general trend.

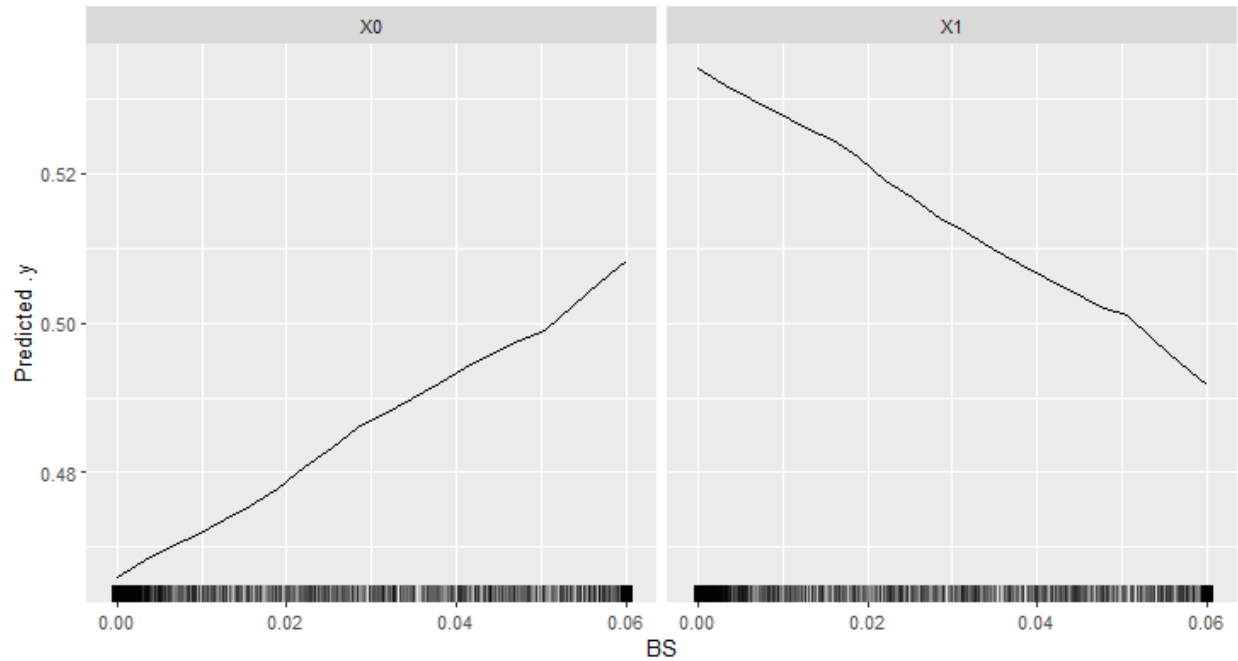


Figure 13: Current Asset Turnover Rate PDP

BS (Current Asset Turnover Rate) is a variable that has values evenly distributed throughout its range of 0 to 0.06. The probability of bankruptcy is almost linear to BS with highest probability of 52% at the value of 0. Hence this variable is not as useful to predict bankruptcy.

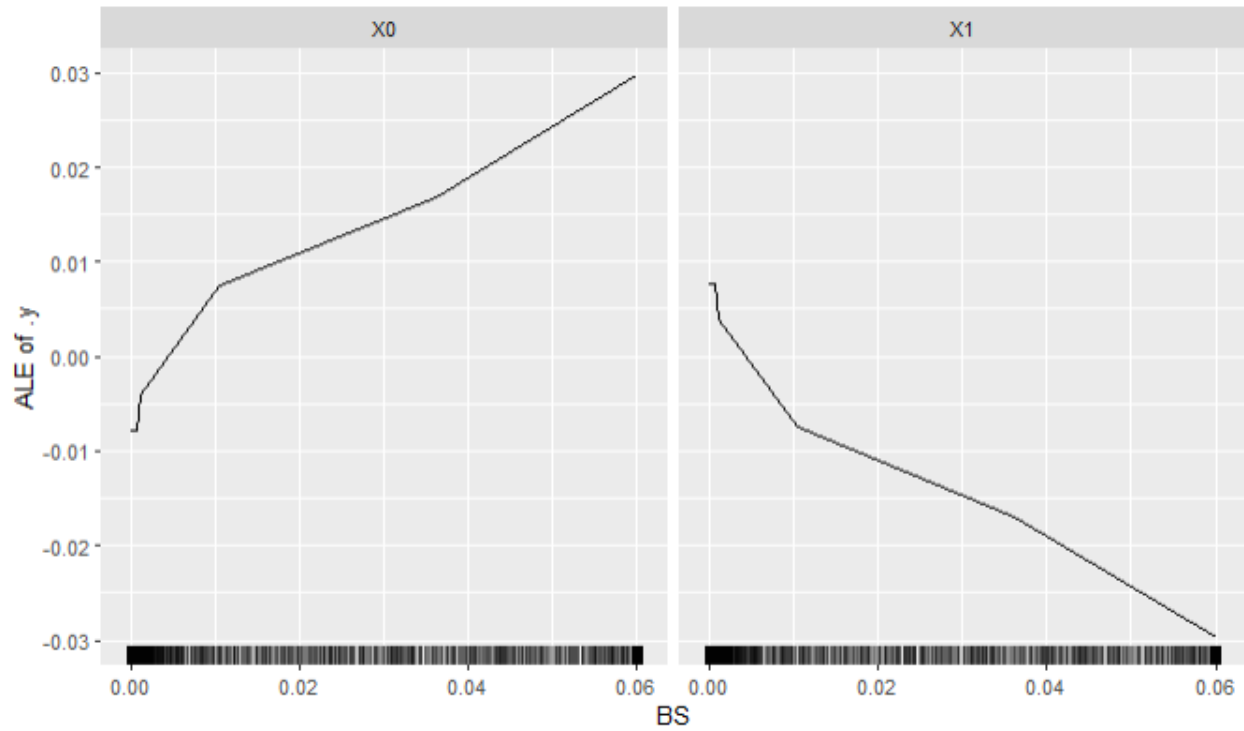


Figure 14: Current Asset Turnover Rate ALE

Both plots are the same for BS (Current Asset Turnover Rate), change in bankruptcy by change in BS are being caused by BS (causation) and other possible correlated variables.

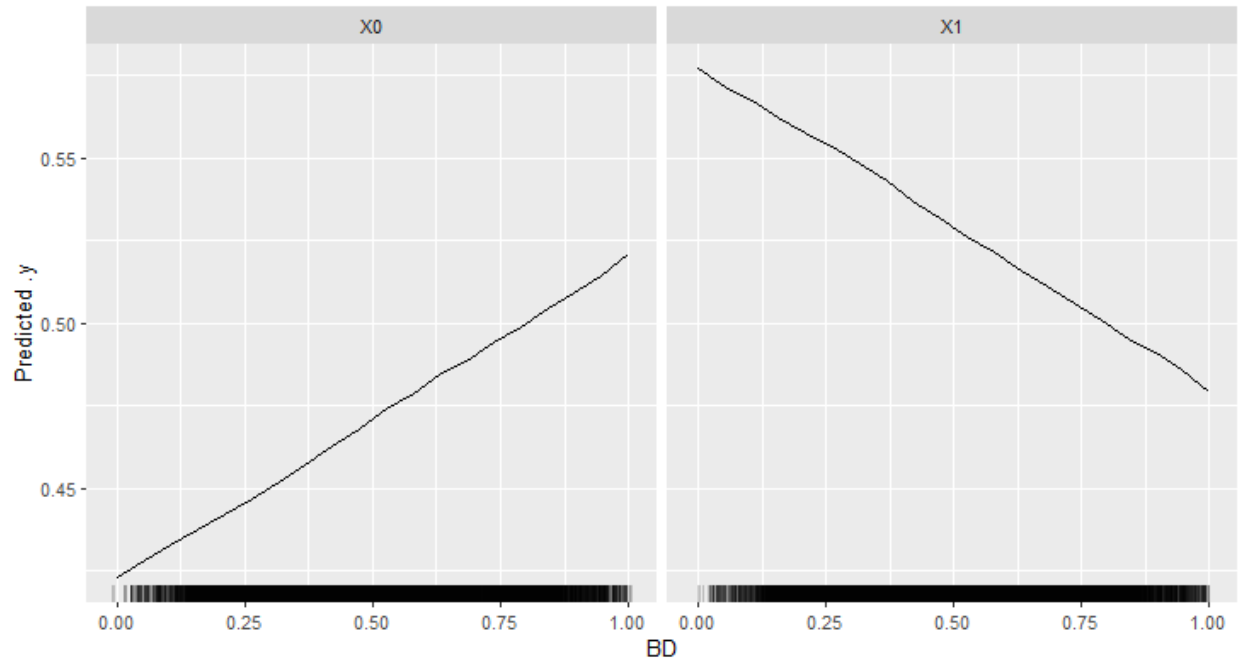


Figure 15: Current Asset to Total Asset PDP

BD (Current Asset to Total Assets) is another variable with an even distribution however a low maximum probability of bankruptcy of 58% at BD value of 0. Probability of bankruptcy is linearly related to this variable.

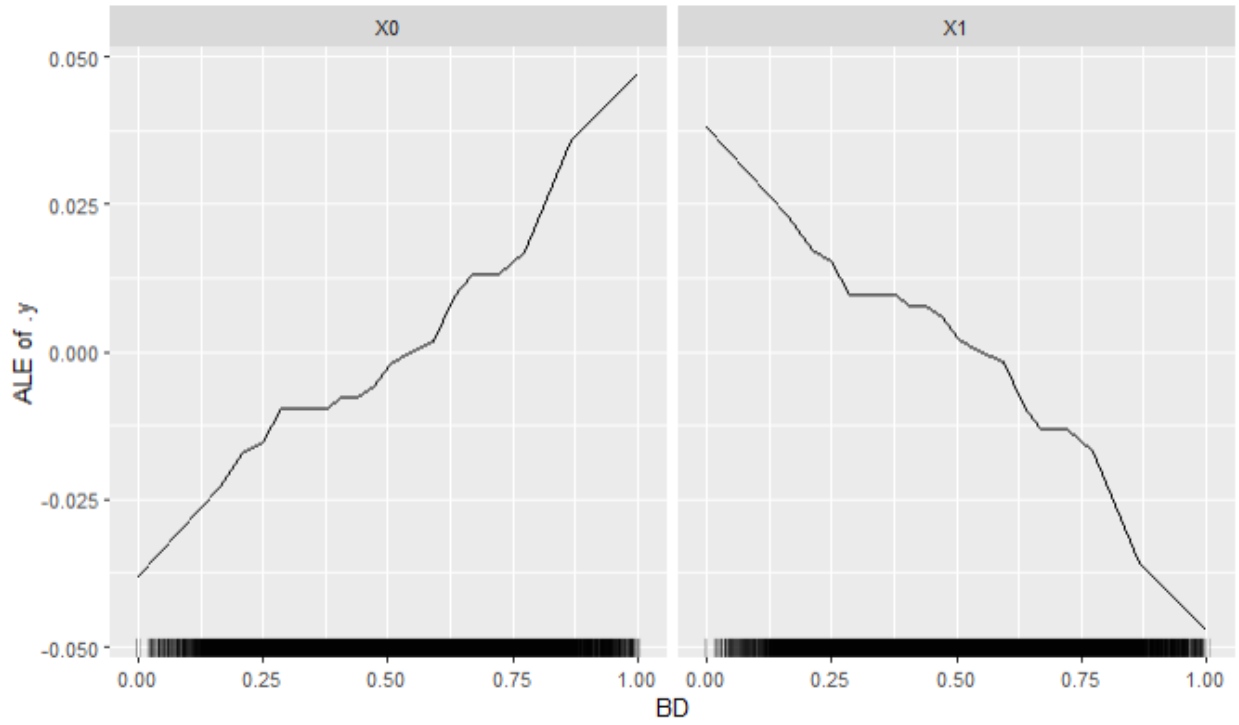


Figure 16: Current Assets to Total Assets ALE

Similar thing can be same for BD (Current Assets to Total Assets) in terms of plot similarity.

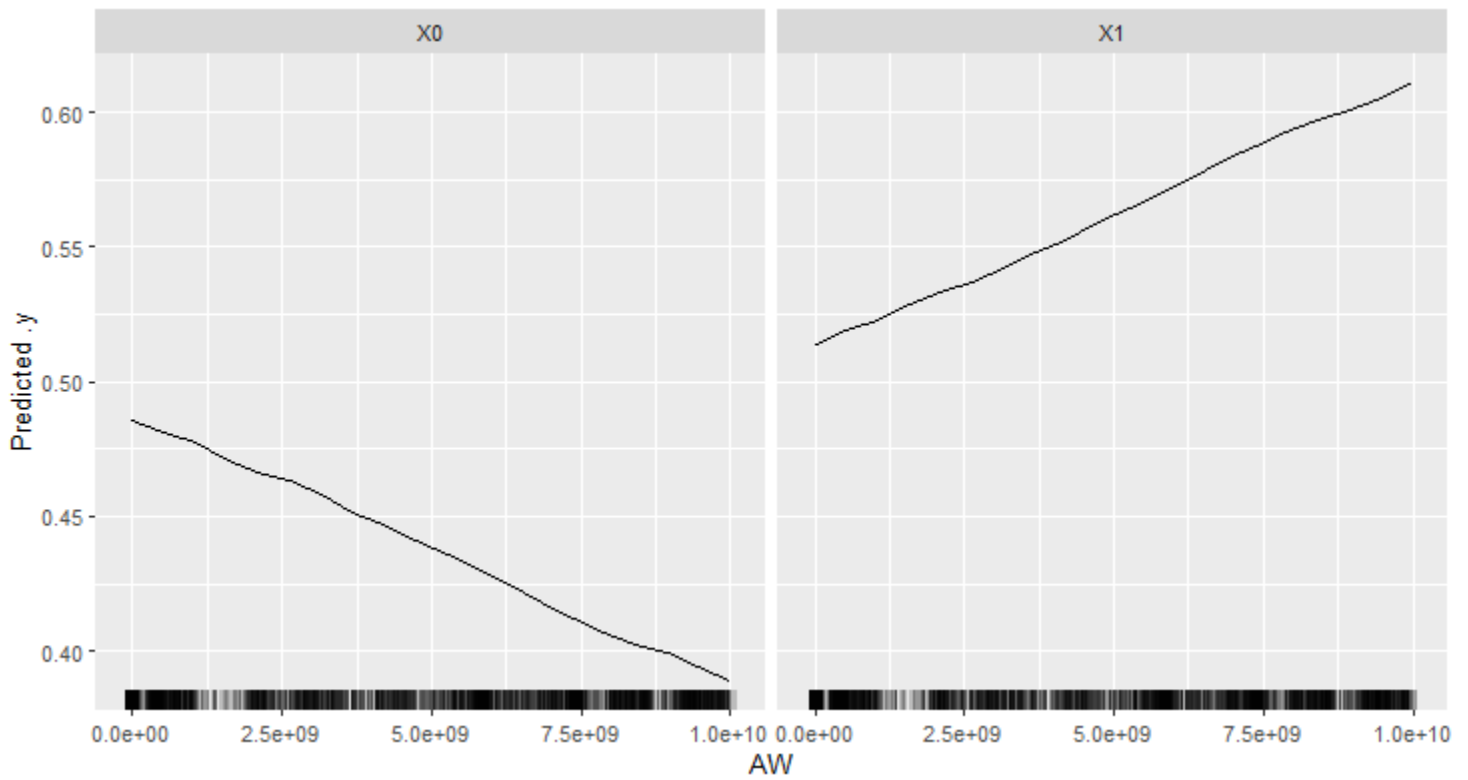


Figure 17: Fixed Asset Turnover Frequency PDP

AW (Fixed Asset Turnover Frequency) consists of an even distribution throughout its range and probability of bankruptcy being linear to it. At a maximum value of 1.0e10, probability of bankruptcy is 60%.

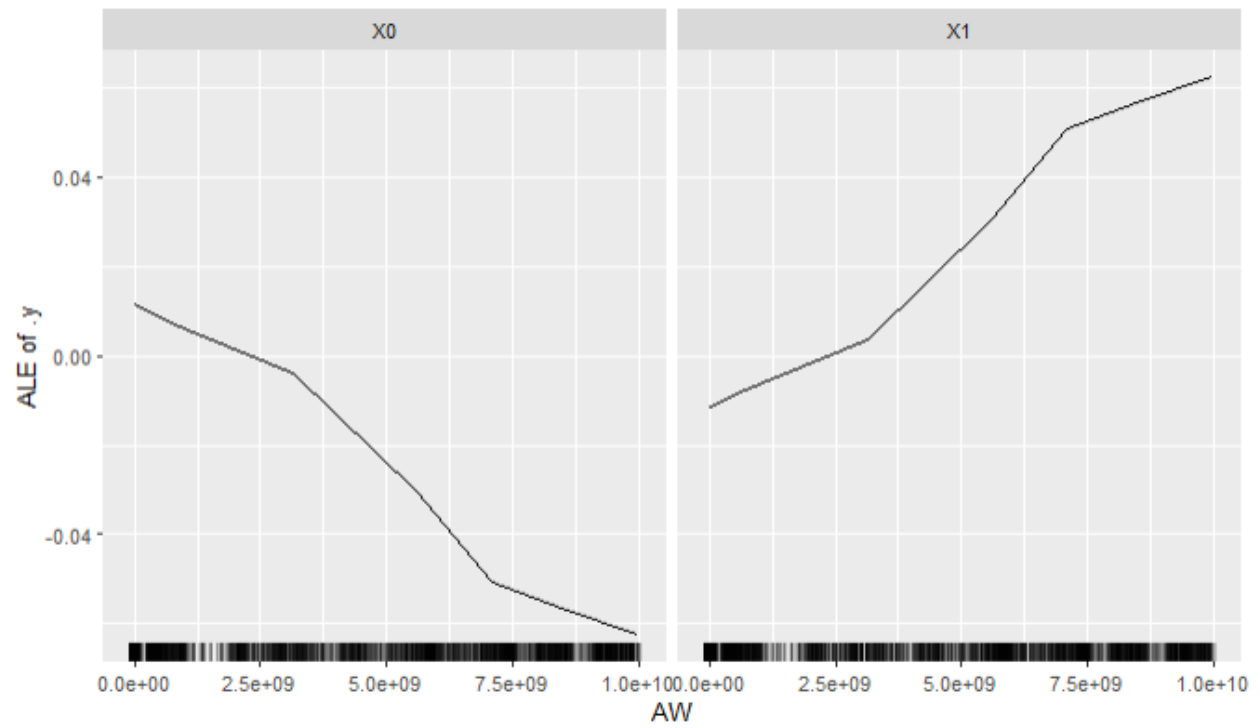


Figure 18: Fixed Asset Turnover Frequency ALE

Same is the case for AW (Fixed Asset Turnover Frequency) for plot similarity.

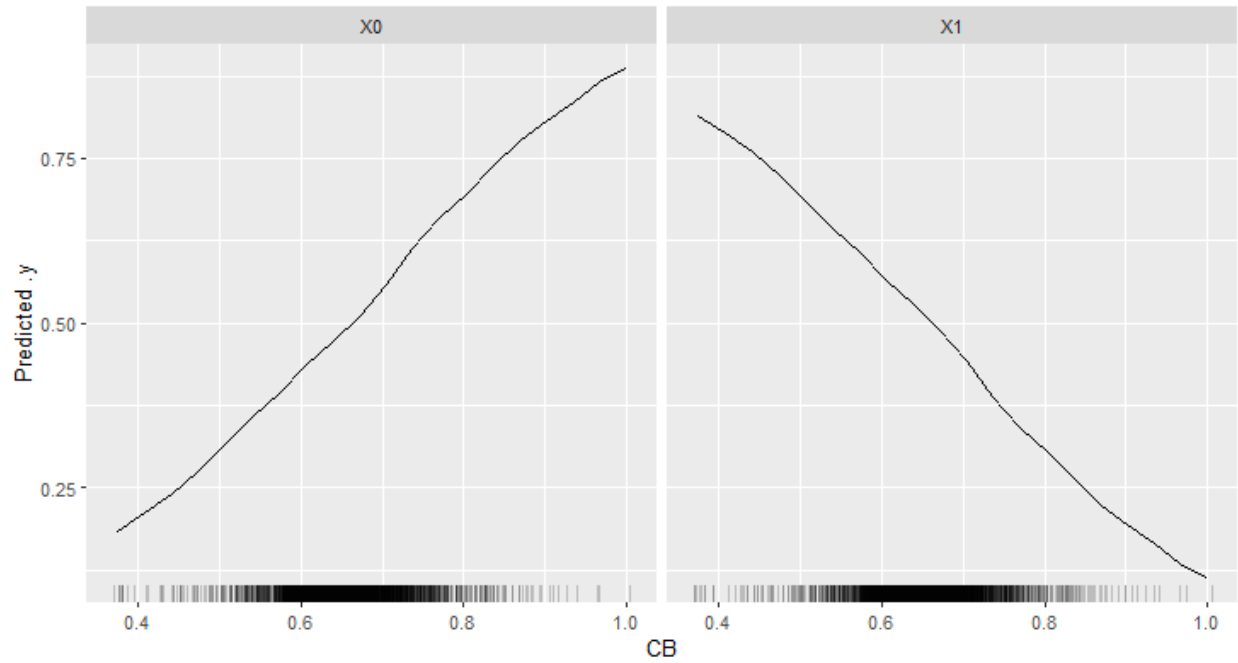


Figure 19: Cashflow to Total Assets PDP

CB (Cashflow to Total Assets) has a maximum probability of bankruptcy of 80% at a value of 0.4. The probability of bankruptcy is also linearly related.

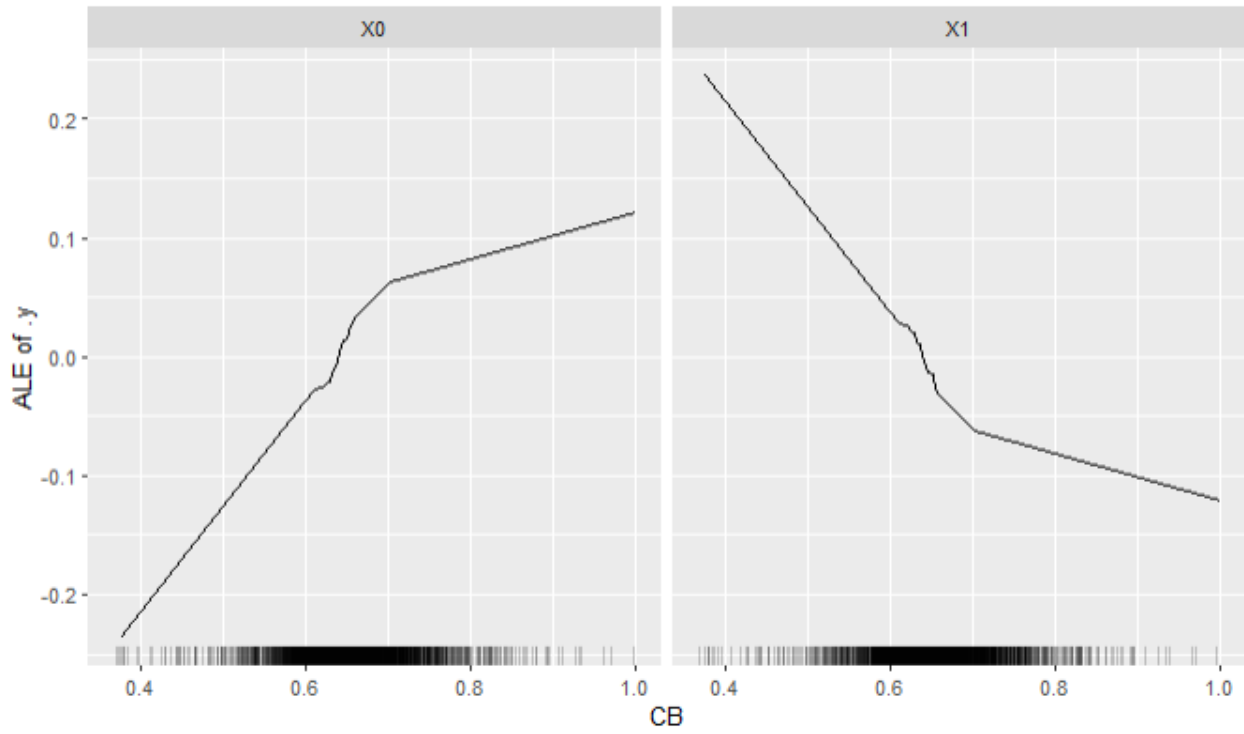


Figure 20: Cash Flow to Total Assets ALE

Same for CB (Cash Flow to Total Assets) for plot similarity.

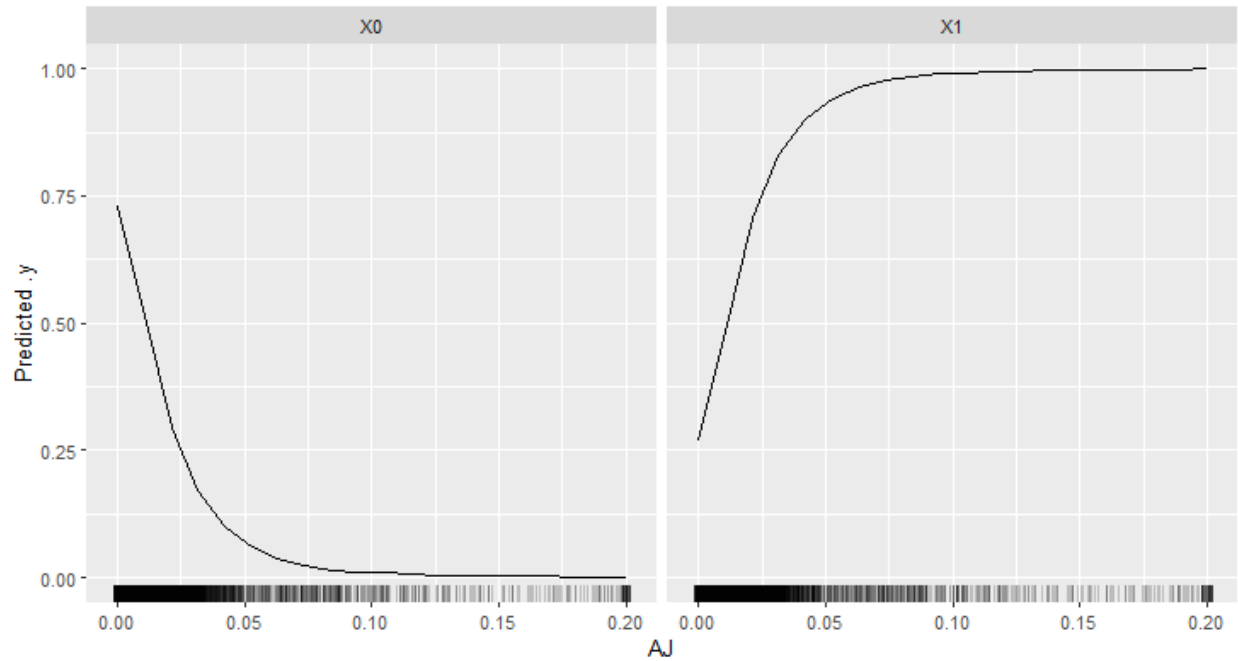


Figure 21: Total Debt to Total Net Worth PDP

AJ (Total Debt to Total Net Worth) is another highly contributing variable. The highest probability of bankruptcy is 100% at a value of 0.2. Considering that the probability of bankruptcy sharply increases till a value of 0.05 for AJ after which it starts to become constant, the probability of bankruptcy at 0.05 is 88%.

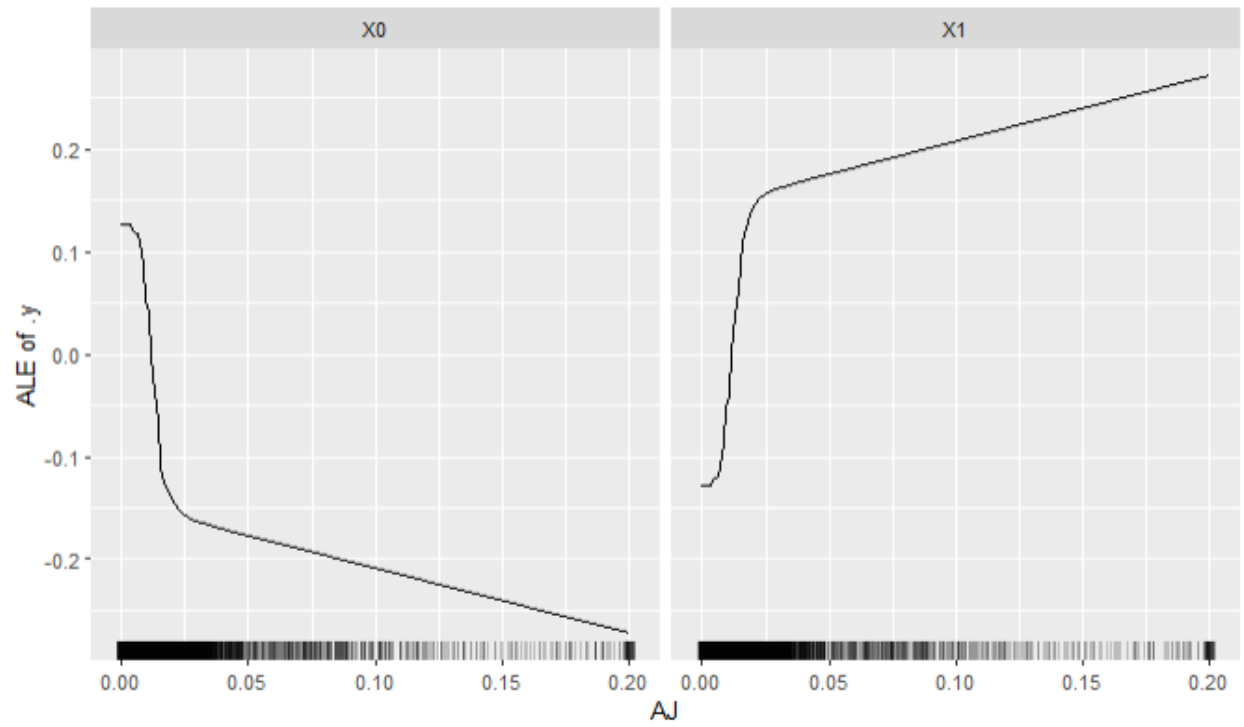


Figure 22: Total Debt to Total Net Worth ALE

Same for AJ (Total Debt to Total Net Worth) for plot similarity.

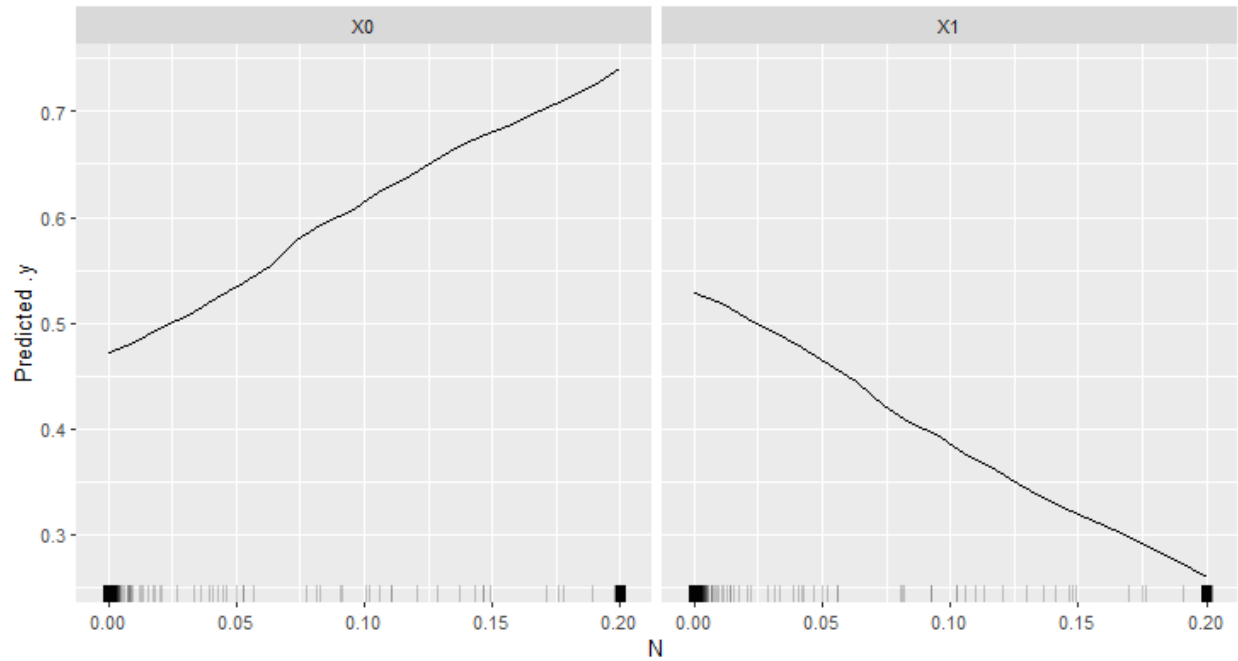


Figure 23: Interest Bearing Debt Interest Rate PDP

N (Interest Bearing Debt Interest Rate) has a maximum probability of bankruptcy of 53% and it only has values concentrated at 0 and 0.2 which are its max and minimum values.

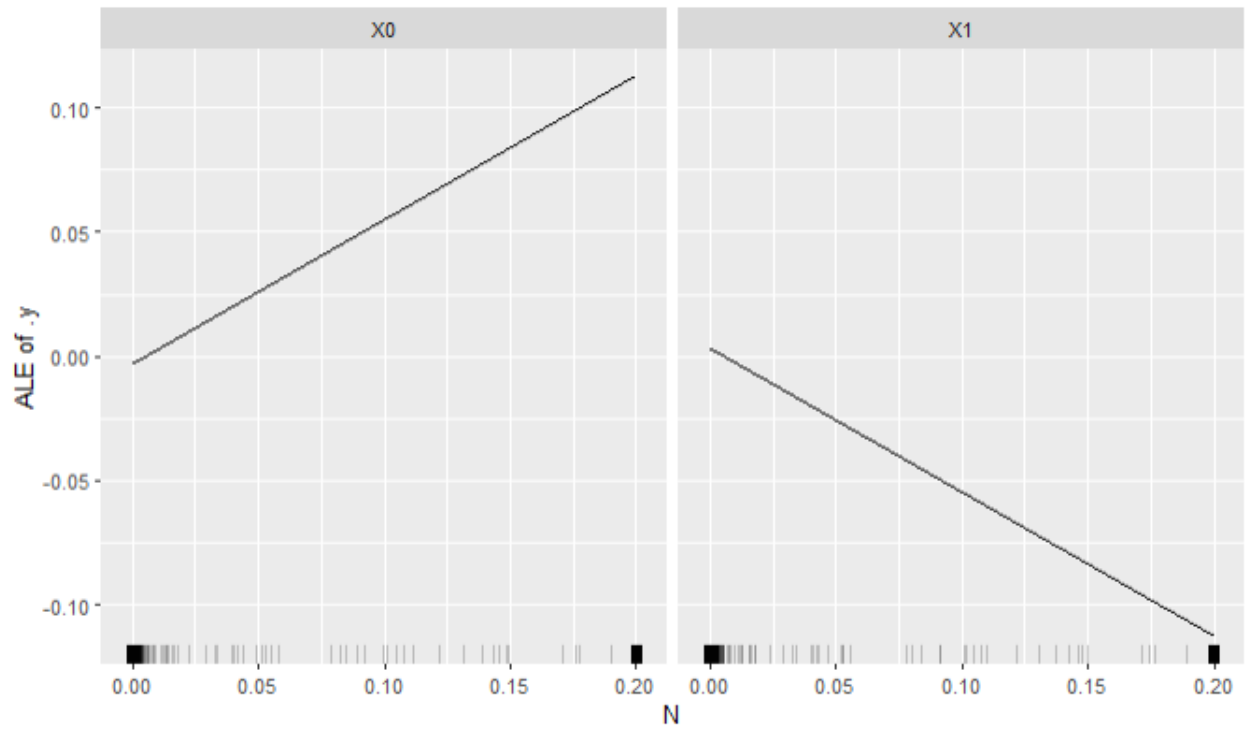


Figure 24: Interest Bearing Debt Interest Rate ALE

Both plots similar for N (Interest Bearing Debt Interest Rate). Hence N and Bankruptcy are causation. Other variables playing negligible role.

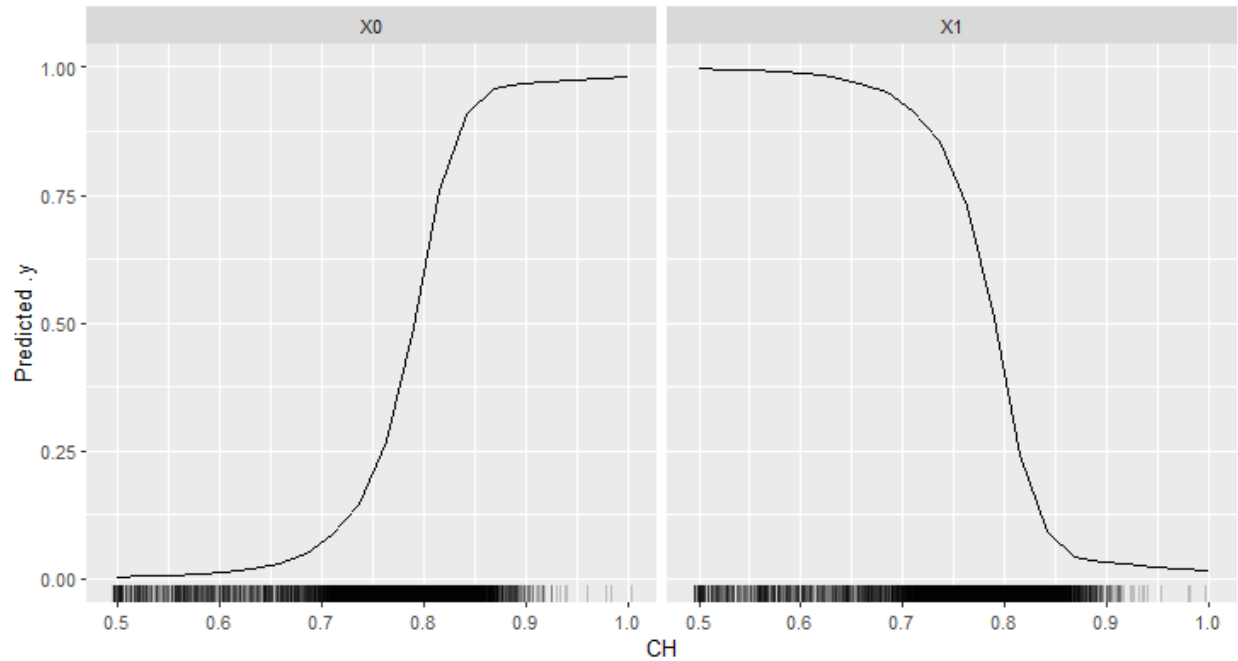


Figure 25: Net Income to Total Assets PDP

CH (Net Income to Total Assets) is another highly contributing variable. It has the highest probability of bankruptcy at 100% at a value of 0.5. An increase in CH is causing the probability of bankruptcy to decrease.

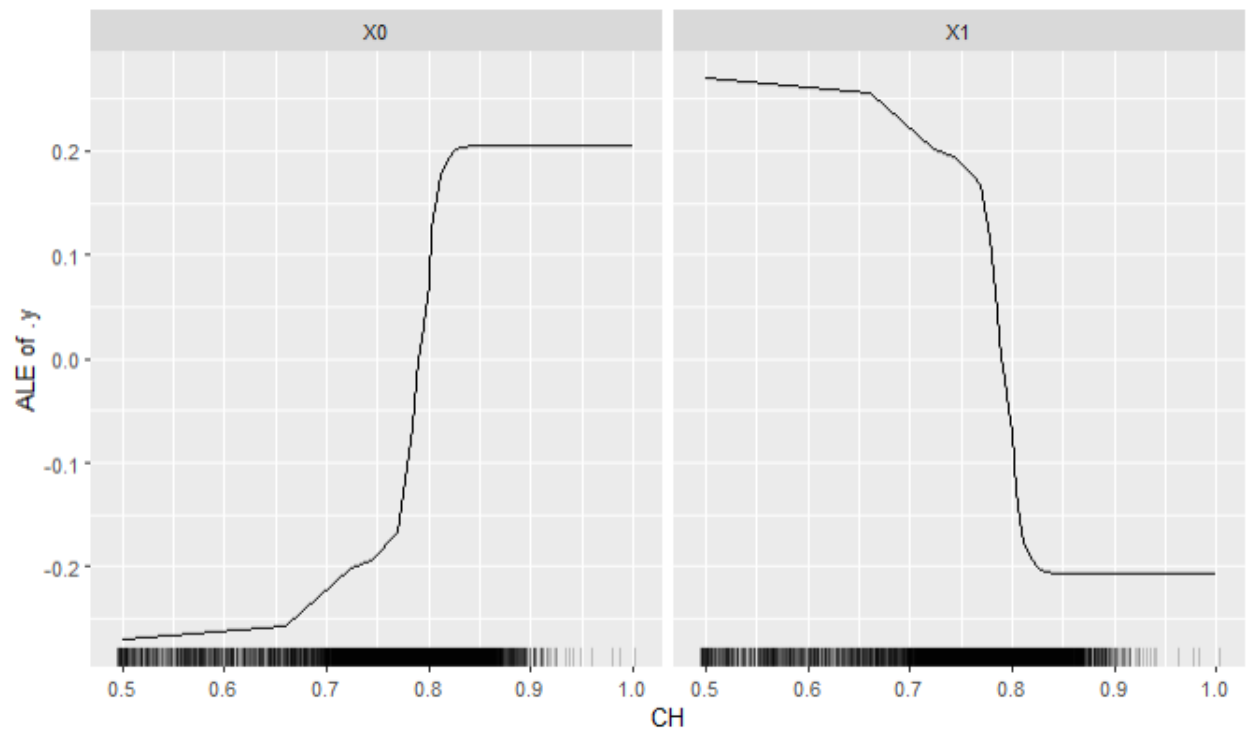


Figure 26: Net Income to Total Assets ALE

CH (Net Income to Total Assets) somewhat follows the same trend. Hence CH and Bankruptcy are causation. But the plots are not identical, this means that other variables do play a slight role.

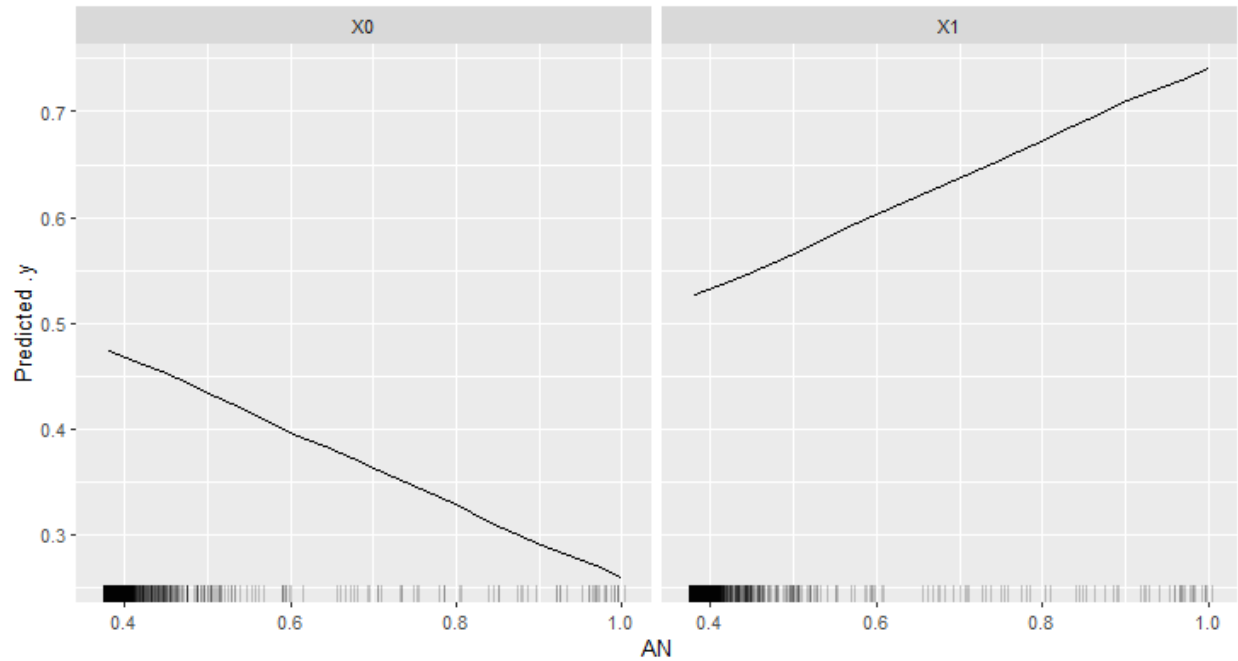


Figure 27: Borrowing Dependency PDP

AN (Borrowing Dependency) is also linearly related to probability of bankruptcy however most of the values of AN are cluttered towards the bottom end of the range. The maximum probability of bankruptcy is 54% at a value of 0.4.

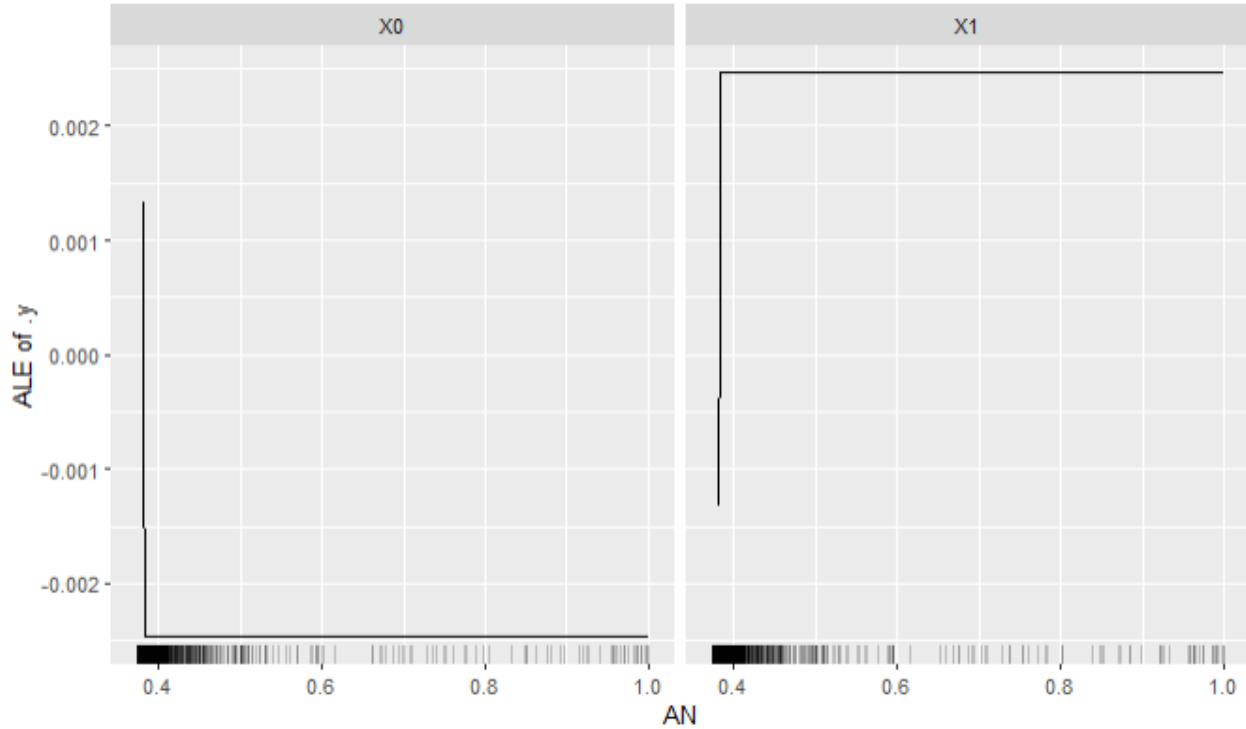


Figure 28: Borrowing Dependency ALE

Completely different plots for AN (Borrowing Dependency). This means that changes of bankruptcy probability by AN are only correlated and not causation. Solely changes in AN are causing no change in bankruptcy probability.

In conclusion, only Interest-Bearing Debt Interest Rate is the variable that is purely causing changes in bankruptcy only because of its changes (causation). Net Income to Total Assets changes to bankruptcy are mostly being caused by itself (causation and a little correlation). Rest of the variables are not directly causing those changes in bankruptcy by their change (causation and correlation). Those include AJ, CB, AW, BD, BS, CQ, CL. The only variable that is not causing any change to bankruptcy by its change is Borrowing Dependency (correlation), other correlated variables to it are causing those changes.

5- Discussions

To reiterate the main findings:

- The SVM model with SMOTE as the sampling technique performed the best where we got an accuracy (specificity) of 95.45%.
- Net Income to Total Assets, Equity to Liability, Net Income to Stockholder Equity and Total Debt to Total Net Worth are strongest predictors of corporate bankruptcy according to our research.
- IML methods revealed that mostly variables predict bankruptcy as they are correlated to it, not that they directly cause it (correlation vs causation).
- SVM Decision boundary revealed that bankrupt companies are not easily distinguishable and hence Principal Components are not able to explain a lot of variance.

For the study done by Mahembe (2024), SMOTE had also been the highest performing sampling technique. The second-best technique was WNN (Weighted Nearest Neighbor) but it only performed well for the Random Forest model.

Alam et al. (2021) performed a similar study. When compared with under sampling methods, SMOTE was also the better performing. However in his study, the best performing predictive algorithm was Decision Forest at 99% accuracy while SVM gave 92% accuracy. His study also summarized results of previous studies where different countries or datasets had different best performing models and a very wide range of best predictive accuracies.

In a study by Liang et al. (2016), multiple feature selection techniques and multiple machine learning algorithms were tested out. It was found that irrespective of which feature selection technique was used, SVM model performed the best. Liang performed this study on a Taiwan

Bankruptcy Dataset which gave an 82% accuracy. This is also consistent with the findings of our study where SVM performs the best however our study gave a higher accuracy.

Another study done on Taiwan bankruptcy dataset was by Tsai (2014) which gave a highest accuracy of 86% while using a Decision Tree composed of 80-100 classifiers using the boosting method along with the boosting method. He also tested out the SVM model like our study but it did not perform as well.

Barboza, Kimura and Altman (2017) performed their study and found that random forest was the highest performing model at 87% accuracy where logistic regression was second at 69% accuracy. His study also summarizes results of previous studies which again highlights that different datasets have different best performing models with a wide range of best accuracies.

Hansradz (2024) performed a similar study where the importance of Profitability Ratios, Liquidity Ratios (Cash Flow Ratios) and Leverage Ratios was highlighted for the prediction of bankruptcy. Kronová et al. (2024) also concluded that Profitability Ratios and Leverage ratios are the most indicative of bankruptcy. Two important variables concluded in our study are also Profitability Ratios which include Net Income to Total Assets and Net Income to Stockholder Equity. The other two important ratios concluded in our study that fall under Leverage Ratios are Equity to Liability and Total Debt to Total Net Worth. However our study did not highlight any Liquidity Ratios (Cash Flow Ratios). For more details on what financial ratios fall under what kinds of ratios, please refer to the study by Barnes (1987).

Bottai, Crosato and Liberati (2024) also found out Profitability Ratios to be very indicative of corporate bankruptcy. However his study also highlighted the importance of Activity Ratios for predicting bankruptcy in SMEs. Kaleem et al. (2024) similarly concluded Activity Ratios to be

important along with Profitability, Liquidity and Cashflow Ratios. Our study however indicated no Activity Ratio and Cash Flow ratio to be very predictive of corporate bankruptcy.

6- Conclusions and Recommendations

The following were some of the limitations in our study and possible scope for any future work.

- 1- Data was very noisy which was highlighted by the decision boundary plot and poor variation explained by principal components. This resulted in bankruptcy classes to be very overlapped and difficult to distinguish by the model. This is potentially the reason why negative predicted value for every model was low.
- 2- We only tried out one feature selection method. But we tested 4 sampling techniques and 5 models. Future studies can have more than one feature selection techniques tested as well.
- 3- We got results of models with high specificity but low negative predicted value. Even though the algorithm is useful for application in business contexts, it will raise false bankruptcy alarms. Future studies could focus increasing negative predicted value further.
- 4- Most bankruptcy predictions have a timeline which consider predicting bankruptcy a certain period before it occurs. Our dataset had no such time data hence such an analysis could not be done.

The value provided by our study is that it confirms the importance of Profitability Ratios and Leverage Ratios for bankruptcy while creating a model that is applicable in business contexts for predicting bankruptcy.

Financial institutions, such as banks and lending companies, can use such a bankruptcy prediction model to evaluate the creditworthiness of their clients. By predicting the likelihood of bankruptcy, lenders can make informed decisions on whether to extend credit, adjust interest rates, or require collateral. This reduces the risk of loan defaults, ensuring a more stable and profitable lending portfolio. For more details on how financial institutions can use such models, refer to the work by Altman and Saunders (1997).

Investors and portfolio managers can rely on these models to gauge the financial health of companies in which they plan to invest. Identifying companies at risk of bankruptcy enables investors to avoid potential losses and strategically allocate their resources. For example, they might opt for more stable investments or engage in risk management strategies like diversification or short selling. For more details on how investors and portfolio managers rely on such models, refer to the work by Jones, Johnstone, and Wilson (2017).

Regulatory bodies can apply these models to monitor the financial health of companies within specific industries, ensuring compliance with financial stability requirements. This helps prevent systemic risks that could lead to broader economic instability. By identifying and addressing issues early, regulators can maintain the health of the financial system and protect consumers and investors. For more details on how regulatory bodies can use such a model, refer to the work by Altman and Rijken (2004).

Businesses can also use bankruptcy prediction models to assess the financial stability of their suppliers or partners. This assessment can companies avoid disruptions in their supply chain caused by the sudden bankruptcy of a key supplier, ensuring continuity in operations and reducing the risk of operational setbacks. For more details on how businesses can use such a model logistically, refer to the work by Hoffman, Schiele and Krabbendam (2013).

Companies themselves can use bankruptcy prediction models to monitor their financial health and manage risks. Early detection of financial distress allows firms to take proactive measures, such as restructuring, managing debts, or securing additional funding, potentially preventing bankruptcy. This proactive approach strengthens the company's overall financial strategy and long-term viability. For more details on how companies can use bankruptcy prediction models to monitor their financial health and manage risks, refer to the work by Dimitras, Zanakis, and Zopounidis (1996).

References

- Fitzpatrick, P.J., 1932. A comparison of the ratios of successful industrial enterprises with those of failed companies.
- Beaver, W.H., 1966. Financial ratios as predictors of failure. *Journal of accounting research*, pp.71-111.
- Altman, E.I., 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4), pp.589-609.
- Barnes, P., 1987. The analysis and use of financial ratios: A review article. *Journal of Business Finance & Accounting*, 14(4).
- Baum, J.A. and Mezias, S.J., 1992. Localized competition and organizational failure in the Manhattan hotel industry, 1898-1990. *Administrative science quarterly*, pp.580-604.
- John, G.H., 1995, August. Robust Decision Trees: Removing Outliers from Databases. In *KDD* (Vol. 95, pp. 174-179).
- Weiss, L.A., 1996. Bankruptcy resolution: direct costs and violation of priority of claims. *Corporate Bankruptcy: Economic and Legal Perspectives*, 260, p.263.
- Greening, D.W. and Johnson, R.A., 1996. Do managers and strategies matter? A study in crisis. *Journal of management Studies*, 33(1), pp.25-51.
- Swaminathan, A., 1996. Environmental conditions at founding and organizational mortality: A trial-by-fire model. *Academy of management journal*, 39(5), pp.1350-1377.

- Dimitras, A.I., Zanakis, S.H. and Zopounidis, C., 1996. A survey of business failures with an emphasis on prediction methods and industrial applications. *European journal of operational research*, 90(3), pp.487-513.
- Kubat, M. and Matwin, S., 1997, July. Addressing the curse of imbalanced training sets: one-sided selection. In *Icml* (Vol. 97, No. 1, p. 179).
- Altman, E.I. and Saunders, A., 1997. Credit risk measurement: Developments over the last 20 years. *Journal of banking & finance*, 21(11-12), pp.1721-1742.
- Doumpos, M. and Zopounidis, C., 1999. A multicriteria discrimination method for the prediction of financial distress: The case of Greece. *Multinational Finance Journal*, 3(2), pp.71-101.
- Breiman, L., 2001. Random forests. *Machine learning*, 45, pp.5-32.
- Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp.1189-1232.
- Branch, B., 2002. The costs of bankruptcy: A review. *International Review of Financial Analysis*, 11(1), pp.39-57.
- Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, pp.321-357.
- Bradley, D.B. and Rubach, M.J., 2002. Trade Credit and Small Business: A Cause of Business Failures. Technical report, Small Business Advancement National Center, University of Central Arkansas.

Drummond, C. and Holte, R.C., 2003, August. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II* (Vol. 11, No. 1–8).

Berkowitz, J. and White, M.J., 2004. Bankruptcy and small firms' access to credit. *RAND Journal of Economics*, pp.69-84.

Altman, E.I. and Rijken, H.A., 2004. How rating agencies achieve rating stability. *Journal of Banking & Finance*, 28(11), pp.2679-2714.

Morris, S. and Shin, H.S., 2004. Coordination risk and the price of debt. *European Economic Review*, 48(1), pp.133-153.

Back, P., 2005. Explaining financial difficulties based on previous payment behavior, management background variables and financial ratios. *European Accounting Review*, 14(4), pp.839-868.

Fawcett, T., 2006. An introduction to ROC analysis. *Pattern recognition letters*, 27(8), pp.861-874.

Armour, J. and Cumming, D., 2008. Bankruptcy law and entrepreneurship. *American law and economics review*, 10(2), pp.303-350.

Hotz, N. (2018) 'What is CRISP DM?', Data Science Process Alliance, 10 September. Available at: <https://www.datascience-pm.com/crisp-dm-2/> (Accessed: 26 August 2024).

He, H., Bai, Y., Garcia, E.A. and Li, S., 2008, June. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence) (pp. 1322-1328). Ieee.

Wang, Y.Y. and Li, J., 2008. Feature-selection ability of the decision-tree algorithm and the impact of feature-selection/extraction on decision-tree results based on hyperspectral data.

International Journal of Remote Sensing, 29(10), pp.2993-3010.

Liu, F.T., Ting, K.M. and Zhou, Z.H., 2008, December. Isolation forest. In *2008 eighth ieee international conference on data mining* (pp. 413-422). IEEE.

Hair, J.F., 2009. Multivariate data analysis.

Sokolova, M. and Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Information processing & management, 45(4), pp.427-437.

Berk, J.B., Stanton, R. and Zechner, J., 2010. Human capital, bankruptcy, and capital structure. *The Journal of Finance*, 65(3), pp.891-926.

Gayatri, N., Nickolas, S., Reddy, A.V., Reddy, S. and Nickolas, A., 2010, October. Feature selection using decision tree induction in class level metrics dataset for software defect predictions. In *Proceedings of the world congress on engineering and computer science* (Vol. 1, pp. 124-129).

Acharya, V.V., Sundaram, R.K. and John, K., 2011. Cross-country variations in capital structures: The role of bankruptcy codes. *Journal of Financial Intermediation*, 20(1), pp.25-54.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12, pp.2825-2830.

Claessens, S. and Yurtoglu, B.B., 2013. Corporate governance in emerging markets: A survey. *Emerging markets review*, 15, pp.1-33.

Tabachnick, B.G., Fidell, L.S. and Ullman, J.B., 2013. Using multivariate statistics (Vol. 6, pp. 497-516). Boston, MA: pearson.

Hoffmann, P., Schiele, H. and Krabbendam, K., 2013. Uncertainty, supply risk management and their impact on performance. *Journal of purchasing and supply management*, 19(3), pp.199-211.

Hosmer Jr, D.W., Lemeshow, S. and Sturdivant, R.X., 2013. *Applied logistic regression*. John Wiley & Sons.

Cleff, T. and Cleff, T., 2014. Univariate data analysis. Exploratory Data Analysis in Business and Economics: An Introduction Using SPSS, Stata, and Excel, pp.23-60.

Tsai, C.F., Hsu, Y.F. and Yen, D.C., 2014. A comparative study of classifier ensembles for bankruptcy prediction. *Applied Soft Computing*, 24, pp.977-984.

Kirkos, E., 2015. Assessing methodologies for intelligent bankruptcy prediction. *Artificial Intelligence Review*, 43, pp.83-123.

Alaminos, D., Del Castillo, A. and Fernández, M.Á., 2016. A global model for bankruptcy prediction. *PloS one*, 11(11), p.e0166693.

Liang, D., Lu, C.C., Tsai, C.F. and Shih, G.A., 2016. Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European journal of operational research*, 252(2), pp.561-572.

Zhang, Z., 2016. Introduction to machine learning: k-nearest neighbors. *Annals of translational medicine*, 4(11).

Ouenniche, J. and Tone, K., 2017. An out-of-sample evaluation framework for DEA with application in bankruptcy prediction. *Annals of Operations Research*, 254, pp.235-250.

- Jones, S., Johnstone, D. and Wilson, R., 2017. Predicting corporate bankruptcy: An evaluation of alternative statistical frameworks. *Journal of Business Finance & Accounting*, 44(1-2), pp.3-34.
- Barboza, F., Kimura, H. and Altman, E., 2017. Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, 83, pp.405-417.
- Laeven, M.L. and Valencia, M.F., 2018. Systemic banking crises revisited. International Monetary Fund.
- Shi, Y. and Li, X., 2019. An overview of bankruptcy prediction models for corporate firms: A systematic literature review. *Intangible Capital*, 15(2), pp.114-127.
- Raschka, S. and Mirjalili, V., 2019. Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2. Packt publishing ltd.
- Papana, A. and Spyridou, A., 2020. Bankruptcy prediction: the case of the Greek market. *Forecasting*, 2(4), pp.505-525.
- Xu, J., Zhang, Y. and Miao, D., 2020. Three-way confusion matrix for classification: A measure driven view. *Information sciences*, 507, pp.772-794.
- Apley, D.W. and Zhu, J., 2020. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4), pp.1059-1086.
- UCI Machine Learning Repository, 2020. Taiwanese Bankruptcy Prediction. Available at: <https://doi.org/10.24432/C5004D>
- Donaldson, J.R., Morrison, E.R., Piacentino, G. and Yu, X., 2020. Restructuring vs. bankruptcy. *Columbia Law and Economics Working Paper*, (630).

Alam, T.M., Shaukat, K., Mushtaq, M., Ali, Y., Khushi, M., Luo, S. and Wahab, A., 2021.

Corporate bankruptcy prediction: An approach towards better corporate world. *The Computer Journal*, 64(11), pp.1731-1746.

Epaulard, A. and Zapha, C., 2022. Bankruptcy costs and the design of preventive restructuring procedures. *Journal of Economic Behavior & Organization*, 196, pp.229-250.

Moreau, T. and Wassermann, D., 2024. Regression and classification. In *Medical Image Analysis* (pp. 57-84). Academic Press.

Mahembe, W., 2024. Use of machine learning in bankruptcy prediction with highly imbalanced datasets: The impact of sampling methods.

Hansradz, M.H., 2024. A Comparative Study between Logistic Regression and Artificial Neural Network Models for Bankruptcy Prediction on Companies Listed on Bombay Stock Exchange in the Indian Market.

Kronová, J., Trebuňa, P., Pekarčíková, M. and Fil'o, M., 2024. Logit Model for Prediction of Financial Health in Automotive Industry. *International Journal of Industrial Engineering: Theory, Applications and Practice*, 31(4).

Bottai, C., Crosato, L. and Liberati, C., 2024. Prediction of SMEs Bankruptcy at the Industry Level with Balance Sheets and Website Indicators. In *INTERNATIONAL CONFERENCE OF ADVANCED RESEARCH METHODS AND ANALYTICS* (pp. 235-241). Editorial Universitat Politècnica de València.

Kaleem, M., Jusoh, H., Raza, H., Sadiq, M. and bin Hamzah, A.H., 2024. A machine learning approach to predict bankruptcy in Chinese companies with ESG integration. *Pakistan Journal of Commerce and Social Sciences (PJCSS)*, 18(2), pp.335-357.

Zhao, J., Ouenniche, J. and De Smedt, J., 2024. Survey, classification and critical analysis of the literature on corporate bankruptcy and financial distress prediction. *Machine Learning with Applications*, p.100527.

Appendix

Dissertation Checklist

Name: Muhammad Muneeb Ullah Ansari

Date Submitted: 27-08-2024

Signature (Digital): 

I confirm that my dissertation contains the following prescribed elements:

- ✓ My dissertation portfolio meets the style requirements set out in the MSc Business Analytics Portfolio Dissertation Handbook including a word count on the front page of each element.
- ✓ I have reviewed the Turnitin similarity report prior to submission.

- ✓ My dissertation title captures succinctly the focus of my dissertation
- ✓ My title page is formatted as prescribed in the MSc Business Analytics Portfolio Dissertation Handbook
- ✓ The abstract provides a clear and succinct overview of my study
- ✓ Each element contains a Table of Contents, and List of Figures and Tables (where appropriate)
- ✓ My dissertation contains a statement of acknowledgement (optional)

- ✓ The Introduction section of the research report, at a minimum, covers each of the following issues:
 - Background to/context of the project
 - Research question(s), aim(s) and objectives
 - Why the project is necessary/important
 - A summary of the Methodology
 - Outline of the key findings
 - Overview of chapter structure of the remainder of dissertation

- ✓ The Background section of the research report, at a minimum, covers each of the following issues:
 - Synthesises the key technical literature relating to the topic
 - Synthesises the key theoretical literature relating to the topic

- ✓ The methodology section of the research report:
 - Details the procedures adopted in carrying out the project (e.g. the data source/acquisition, data processing, procedures for maximising rigour and robustness, methods of data analysis etc). Contains ethical considerations and decisions. This section should not duplicate the technical report, which focuses more on the detailed technical choices and steps.
- ✓ The findings section of the research report reports the results in detail and provides possible explanations for the various findings
- ✓ The discussion section of the research report makes appropriate linkages between the findings and the literature reviewed.
- ✓ The conclusions section of the research report includes:
 - Conclusions about each research question and/or hypothesis
 - General conclusions about the research problem
 - Implications for theory, for policy and/or management practice
 - Limitations of the research
 - Suggestions for practice and future research
- ✓ The technical report, log book, and reflective discussion have each been included.
- ✓ The reference list is in alphabetical order and follows the Harvard system
- ✓ I have signed and dated the Candidate Declaration