Applying Machine Learning in Corporate Default Prediction

Hans Dall-Møller (92218)

&

Oscar Monberg (93273)

Master Thesis

M.Sc. Finance & Investments

Supervisor: Michael Ahm

Copenhagen Business School

May 15th, 2019

No. of pages (characters): 90,35 (205,564)

Abstract

The thesis investigates the degree to which it is possible to apply machine learning in corporate default prediction. Specifically, in order to investigate the overall problem statement, the thesis intends to test and answer three specific research questions.

First, the thesis analyzes whether there is a difference in accuracy between logistic regression and random forest when predicting default. Next, the thesis tests whether the addition of non-firm-specific variables have any effect on model accuracy. Lastly, the thesis investigates whether the driving variables and the precision of the models are conditional on industry when predicting corporate default.

The data used to test the research questions is private company information extracted from the Bureau van Dijk (BVD) database. In addition, the data complies with the following selection criteria. The analysis is conducted on data from 2000-2017, is performed on companies from France, Italy, Portugal and Spain, and fall within the BVD size classification Very Large, Large & Medium and lastly, the data is restricted to manufacturing and wholesale trade defined by SIC-code.

The variables chosen fall within firm-specific and non-firm specific variables. The firm-specific variables consist of financial ratios from the categories; profitability, asset efficiency and solvency. The non-firm-specific variables consist of macroeconomic indicators, a stock market index and commodity prices.

For the first research question, it is found that random forest outperformed logistic regression by 4.7 percentage points indicating that random forest is better at predicting corporate default.

For second research question, it is found that the addition of non-firm-specific variables only minutely increases accuracy for random forest, but has no effect on logistic regression. Conclusively, the addition of non-firm-specific variables does not materially increase accuracy.

For the third research question, it is found that accuracy is highest for the total sample, indicating no gain from splitting samples by industry. To substantiate this claim, it is found that the driving variables for both models tested are greatly similar between industries. In total, this indicates that the driving variables are not conditional on industry.

Abbreviations

BVD	Bureau van Dijk
SMOTE	Synthetic Minority Oversampling Technique
LASSO	Least Absolute Shrinkage and Selection Operator
ROSF	Return on Shareholder Funds
ROCE	Return on Capital Employed
ROTA	Return on Total Assets
PM	Profit Margin
EBITDAM	EBITDA Margin
EBITM	EBIT Margin
CFT	Cash Flow Turnover
NAT	Net Asset Turnover
ST	Stock Turnover
COP	Collection Period
IC	Interest Coverage
CP	Credit Period
CR	Current Ratio
SR	Solvency Ratio
LR	Liquidity Ratio
G	Gearing
AGE	Company Age
GDPG	GDP Growth
IRNC	Interest Rate Nominal Change
INFLNC	Inflation Nominal Change
UNC	Unemployment Nominal Change
STOXX	STOXX 600 Europe % Change
RMPNC	Raw Materials Price Nominal Change
BMNC	Base Metals Nominal Change

List of Figures

List of Tables

Table of Contents

1. Introduction

Corporate default prediction has been researched extensively since the 1960's. The awareness of the topic was established when Edward Altman (1968) published his groundbreaking paper *"Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy"* in which he introduced the Z-score model. In the paper, Altman introduced a simple model that indicated whether a company was in danger of defaulting. Altman found five financial ratios which he suggested contained sufficient information to predict corporate default. Altman was inspired by previous work by William Beaver (1966) who was among the first to use financial ratios as predictors of firm failure. Later, James Ohlson (1980) published a similar paper to Altman in which he introduced the famous O-score as an alternative to Altman's Z-score.

Even though the topic of predicting default has been of interest for a substantial amount of time, the topic has not yet been studied exhaustively. Especially as a result of the financial crisis in 2007-2008, corporate default prediction research is as important as it has ever been, and is an academic area within finance which deserves attention. In addition, technological advances in computing and the subsequent development of non-parametric models has reinvented the way in which default prediction can be analyzed. The fact that the problem can be analyzed using newer methods that can handle enormous amounts of data has been a reason for the continuous interest in the topic.

Overall, there are two main approaches which can be applied when dealing with corporate default prediction: the market-based approach and the machine learning approach. Machine learning models can then be split into parametric and non-parametric models. Parametric statistical models include, among others, logistic regression and discriminant analysis. The commonality between parametric approaches is that they rely on a finite set of parameters and thus a set of assumptions regarding the data. The famous paper by Edward Altman (1968) used discriminant analysis in order to retrieve his Z-score whereas the research by James Ohlson (1980) used logistic regression.

Non-parametric approaches rely on clever algorithms and computing power to see patterns in data. Non-parametric models are not restricted to a finite set of parameters and do thus not rely on any assumptions about the data. This fact is what makes non-parametric algorithms interesting and in theory superior to the classical parametric approaches.

The market-based approach includes the application of option pricing theory such as the KMV-Black-Scholes-Merton model. The market-based approach looks at the problem from an option pricing perspective where the idea is to model the probability of the firm's assets being below a critical default point i.e. the face value of debt.

In this thesis, two different models will be applied; logistic regression, a traditional parametric model and random forest, a non-parametric model. In addition, the explanatory variables used in the analysis are extended to encompass financial ratios as well as non-firm specific variables such as macroeconomic factors, commodity prices and a stock index.

Concretely, the goal with the thesis is to answer three different questions. First, the thesis investigates whether random forest as a representative of non-parametric models is better at predicting corporate default than logistic regression as a traditional parametric method. Second, most of the literature regarding corporate default prediction is concerned with financial ratios. Therefore, the thesis seeks to answer whether non-firm specific variables provide valuable information that will increase the accuracy of the models. Lastly, the thesis explores whether the variables that best explain corporate default vary across industries.

It is believed that the three questions mentioned above are relevant in the field of corporate default prediction. As stated, most research has traditionally focused on financial ratios where only a fraction of the published papers have used other explanatory variables. There might be good reasons for that, but the thesis seeks to further investigate whether the decision to only focus on financial ratios is justified. Furthermore, no research was found regarding the effect of the analyzed variables across industries. In order to explore the most optimal model, is important to understand whether the driving factors of corporate default are conditional on industry. If this is the case, there is a problem in making broad default models which do not distinguish between industries. Lastly, despite being a relatively well researched topic, it is important to shed light on whether non-parametric methods are in fact significantly better at predicting corporate default than traditional parametric models.

In order to answer the three research questions, the thesis is based on data extracted from the Bureau van Dijk (BVD) database. BVD is a Moody's Analytics company. More specifically, the data stems from French, Italian, Spanish and Portuguese private companies in the period 2000-2017. The data consists of 16 financial ratios, one firm-specific variables which is not a ratio, four macroeconomic variables, two commodity variables and a stock index variable. In addition, the data is split between two different industries; manufacturing and wholesale trade. Lastly, the data frequency is annual. In other words, all data in the BVD database stems from annual reports.

1.1 Problem Statement

To which degree is it possible to predict corporate default using machine learning?

1.1.1 Research Questions

1) How does accuracy vary between logistic regression and random forest when predicting corporate default?

2) Does the addition of non-firm-specific variables improve the accuracy of the models?

3) Are the driving variables and the precision of the models conditional on industry?

1.1.2 Hypotheses

H1: It is expected that random forest has superior accuracy compared to logistic regression.

First, in order to substantiate the assertion made above, it is necessary to elaborate on the fundamental difference between parametric and non-parametric models and why non-parametric models should, in theory, provide better results than parametric models in most instances.

Parametric models rely on the premise that a finite set of parameters can adequately describe the data. This in turn means that parametric models have to make assumptions about the data. If the data does not comply with these assumptions, the chosen parametric model will not be built on a reliable foundation which consequently affect the accuracy of the model.

This is also true for logistic regression. Even though logistic regression does not assume any specific distribution of the data, it still draws on several other assumption. First, it is assumed that there is a linear relationship between the predictors and their logit transformed outcomes Stoltzfus (2011). Second, is the assumption of the absence of multicollinearity between the explanatory variables. Multicollinearity means that the predictor variables are not only correlated to the response variable, but also correlated with each other. The assumption of an absence of multicollinearity therefore means that logistic regression assumes that the explanatory variables are not highly correlated with each other, but only correlated to the response variable. Third, it is assumed that there are no influential outliers in the data. If none of the aforementioned assumptions are violated, the logistic regression model will most likely perform well. However, it is often difficult to assemble a dataset consisting of hundreds of thousands of observations across a vast set of explanatory variables and confidently state that none of the assumption are violated to some degree. This is where non-parametric models are useful.

Random forest and other non-parametric models do not rely on a fixed number of parameters. As will be explained in depth later, random forest is based on a clever procedure of combining thousands of randomly built decision trees which are “grown” using binary splitting. Binary splitting does not require any assumptions about the data which in turn makes random forest preferable to logistic regression, in theory. There might be other reasons for choosing logistic regression such as interpretability or computation time, but regarding the accuracy of the two models, random forest should outperform logistic regression if the assumptions behind logistic regression are not met exactly.

To further substantiate the expectations, previous research has found that random forest does in fact yield better results that logistic regression. For instance, Lin & Mcclean (2001), Wagenmans (2017), Barboza, Kimura and Altman (2017) and many others have found that random forest outperforms logistic regression when predicting corporate default.

H2: It is expected that the addition of non-firm-specific variables will improve the accuracy of the models.

It is expected there will be some information gain when including non-firm specific variables, it is, however, not expected that the predictive power of the added variables has a leading effect. This claim will be elaborated now. Due to the fact that non-firm specific variables are dependent on time and therefore not specific to the individual firm, it is not expected that the added variables will constitute a driving factor when predicting individual firm default. In other words, since the samples used are balanced such that the distribution of non-bankrupt firms across years is equal to the bankrupt firms, it is not expected that the non-firm specific variables will have a leading effect on default prediction.

It is, however, expected that the addition of the non-firm specific variables will have some additional predictive power. To understand this notion, it is useful to think in lines of a classification tree. At the top of a classification tree, the most informative variable will be chosen. In other words, the variable which is best a splitting the two classes apart will always be the leading variable. However, variables further down the tree, despite not being the leading variables in explaining default, are also important. For instance, it is possible that companies with a significant negative cash flow to turnover ratio are correlated with lower GDP growth or any other non-firm-specific variable. If this is the case, the non-firm-specific variable will appear further down the tree as a rule for splitting the classes and thus in predicting default. Therefore, it is expected that the addition of non-firm-specific variables will provide some information gain which consequently will result in a higher accuracy of the models.

This expectation is primarily based on intuition as well as an understanding of how the models treat the data, but not on previous research. The reason why the expectation isn't based on research is the fact that conclusion drawn by previous research vary. For instance, Duffie, Saita & Wang (2007) and Carling et al. (2007) report a significant relationship between the state of the economy and default hazard rates of individual firms. However, Koopman et. al (2009) and Koopman, Lucas & Schwaab (2011), argues that macroeconomic indicators alone might either under or over-estimate default risk and that macroeconomic variables could most likely lose their predictive power in conjunction with firm-specific variables.

H3: It is expected that the driving variables and the precision of the models are conditional on industry.

The claim stated above is based on intuition rather than any theoretical justification. It is expected that the importance of the explanatory variables will differ between industries when predicting corporate default due to the nature of the two industries in question. Firstly, companies within the manufacturing industry are defined as firms that use components, parts and/or raw materials to make finished goods. Wholesale firms, on the other hand, do not participate in any type of production, but rather sell large quantities of goods to other parties. Both industries are asset-heavy, where manufacturing is predominantly heavy on long-term assets such as machinery whereas wholesale trade is heavy on short-term assets in the form of inventories. Therefore, it is expected that the general asset efficiency ratios such as net asset turnover and return on total assets are better predictors for manufacturing companies regarding default prediction where stock turnover, collection period, credit period and the current ratio are more telling for wholesale companies. Therefore, from an intuitive point of view, it is expected that the driving factors of default will vary across industries. However, despite these hypothesized differences, it is still expected that general measures of solvency, leverage and cash flows will be important unconditional of industry.

It is noted that the two industries are very broadly defined using the SIC-code classification system. Breaking down the industries more specifically will, everything else equal, yield a more specific class of companies with larger definitive differences between the industries. It is, however, still expected that the driving variables and precision of the models are conditional on industry for the general case.

1.2 Structure

Up to this point, the academic realm of corporate default prediction has been introduced and the problem statement of this thesis has been stated. In addition, three specific questions in conjunction with the hypotheses that show the expectations have been stated. The rest of the thesis will be structured the following way.

First, the thesis will give a comprehensive review of the relevant academic literature that has set the setting for the field of corporate default prediction. In this section, the thesis describes the different methods that one can use to predict default followed by a more specific review of the literature that concentrates on the classification method which will be applied in this thesis. In addition, a presentation of some of the results that previous research has found regarding variable selection and model accuracy will be made.

Second, the methodology section is presented. In this chapter of the thesis, substantial emphasize will be put on how the three different hypotheses are tested. This includes how the datasets will be build, but also a thorough description of different sampling techniques, definitions of how model performance will be measured and how classification works in general. Additionally, a short descriptive overview of the set of explanatory variables that are included in this thesis will be given.

Third, is a description of the data. This section is of significant importance to the rest of the thesis. In this part, the selection process which includes the criteria used to define the dataset from the BVD database will be described. The data description section then moves on to a sample description where a visual presentation of the samples used for the different tests are presented. The sample description will also be used as a way to show and justify some of the changes to the data that is necessary to do. Finally, the data description section moves on to a variable description. Here each variable that was decided to be used in the analysis is thoroughly described. The variable description is followed by descriptive statistics which will show how the variables differ on a descriptive level.

Fourth, is the theoretical background. In this chapter of the thesis, the theory behind the models that are used will be dissected. This part includes the theory behind logistic regression, random forest and synthetic minority oversampling technique (SMOTE). The idea of this part of the thesis is to break down otherwise fairly complicated ideas in order to describe how the models work.

Fifth, after the theoretical background, the results are presented. In this part all the results will be described primarily on a descriptive basis. The idea here is to present the findings in relation to the hypotheses and describe whether the results prove or disprove the hypotheses.

Sixth, closely related to the results is the discussion section. In this section the results are discussed on a deeper level. Furthermore, the interpretation of the results, how the results relate to the academic literature and lastly the limitations of the results will be discussed.

Finally, ideas and recommendations for future research are proposed. This part will summarize future research ideas that can illuminate the limitations of the results, but also ideas that do not relate directly the specific research, but could be interesting to study.

1.3 Scope of the paper

The overall purpose of the thesis is to test to which degree it is possible to predict corporate default using machine learning. In order to investigate this problem statement, the three research questions must be answered.

First, the accuracy of logistic regression as a parametric model and random forest as a non-parametric model when predicting corporate default will be tested. The thesis will not focus on how parametric and non-parametric models compare in general. In other words, the thesis will not compare other models than logistic regression and random forest.

Second, it is tested whether the addition of non-firm-specific variables such as macroeconomic variables improve the accuracy of the models compared to only using firm-specific variables. The conclusion drawn is therefore limited to the set of explanatory variables that was chosen.

Third, it is investigated whether corporate default prediction is conditional on industry and whether the driving factors of default prediction varies across industries. To narrow the scope of the analysis it was decided to use two industries: manufacturing and wholesale trade. Again, the conclusion drawn is limited to these two industries.

Generally, it is important to state that the problem statement, research questions and hypotheses that are presented are questions that are important when investigating and testing the claim that machine learning has its merit in corporate default prediction. Despite the fact that previous academic literature touch upon some of the same issues as this thesis aim to test, it is important to underline that the thesis does not intend to replicate or test the conclusions of any one or more specific previous research papers. Instead, the goal of the thesis is to extend the current research by testing some propositions that haven’t been well documented, but also propositions that are fairly well analyzed.

Lastly, it is important to state that despite the fact that relatively complex algorithms such as logistic regression, random forest and the synthetic minority oversampling technique are used, the thesis will not focus on the proofs behind these models. An in-depth explanation of the intuition behind the ideas is included, but the mathematical proofs are outside the scope of the thesis.

2. Literature Review

The literature review of the relevant contributions to the field of corporate default prediction will be structured the following way. First, the review will cover the initial research within the field and the models which were used at that time, predominantly parametric models. Then a short section of market-based model will be presented followed by recent development and applications of non-parametric models. The last part will focus on the set of variables which have been used to analyze the problem of default prediction.

2.1 Parametric Models

The field of corporate default prediction relates back to the 1930's where FitzPatrick (1932) published a paper in which he investigated 19 pairs of bankrupt & non-bankrupt firms. In the paper, FitzPatrick found that there were persistent differences in the investigated companies' financial ratios at least three years prior to default. Shortly after, Winakor & Smith (1935) presented a similar paper to FitzPatrick, in which they extended the time until default to 10 years. In the paper, Winakor & Smith found that the deterioration of the mean of the financial ratios investigated could be observed 10 years prior to default with the effects becoming more apparent closer to default. Similarly, Merwin (1942) found that the mean ratios could be observed as long as six years prior to default. Despite the fact that corporate default research can be traced back to the 1930's, the significant contributions to the field first appeared in the 1960's by Beaver (1966) & Altman (1968).

Beaver (1966) presented a paper in which he analyzed 30 financial ratios from 158 companies of which 79 had gone bankrupt and 79 were active. The chosen method of analysis was "Dischotomous Classification Test" a univariate approach which was later criticized for its simplicity. To test which financial ratios were better at predicting default, Beaver ranked each financial ratio in ascending order and visually decided a cut-off point which he deemed to be optimal. He iteratively conducted this analysis for each of the 30 financial ratios once at a time. After computing the classification error for each of the 30 cut-off points, he found six of the ratios were better at predicting default. Beaver reported cashflow/total debt, net income/total assets, total debt/total assets, working capital/total assets, current ratio and no-credit interval were best at predicting default.

Altman (1968) criticized the approach of previous research by stating: *"... the adaptation of their results for assessing bankruptcy potential of firms, both theoretical and practically, is questionable"* (Altman, 1968). Specifically, Altman criticized the simplicity of the results and stated that looking at one financial ratio independently is insufficient, prone to faulty interpretation and potentially confusing when explaining default. In order to circumvent this issue, Altman advocated the use of multiple discriminant analysis (MDA). Unlike univariate analysis, the benefit of using MDA is the fact that the method considers the entire predictor space and the interactions between the predictors. In addition, MDA reduces the dimensionality of the data by computing the linear combinations of the predictors which best discriminate between the groups i.e. bankrupt and non-bankrupt companies. According to Altman (1968), the advantages of reducing the dimensionality of the dataset is the fact that the final model only considers the predictors which contain most information. In the paper, Altman investigated 22 financial ratios on a dataset consisting of 66 companies in the period 1946-1965 of which 33 were bankrupt and 33 were non-bankrupt. Altman found that five financial ratios made up the discriminant function:

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 0.999X_5$$

Where X_1 is working capital/total assets, X_2 is retained earnings/total assets, X_3 is EBIT/total assets, X_4 is market value of equity/book value of debt and finally X_5 is sales/total assets. Altman concluded that a Z-score of $Z > 2.99$ should be considered “safe zone”, $1.81 < Z < 2.99$ = “grey zone” and $Z < 1.81$ = “distress zone”. Altman later extended his original research in Altman (1973), Altman & McGough (1974), Altman & Lorris (1976) and Altman, Halderman & Narayanan (1977).

Despite the fact that the framework presented in Altman (1968) is considered as a highly significant paper in the realm of corporate default prediction, several limitations was pointed out by Ohlson (1980). In his paper *“Financial Ratios and the Probablistic Prediction of Bankruptcy”* (1980) inspired by White & Turnbull (1975a, 1975b) and Santomero & Vinso (1977), James Ohlson presented a new probabilistic approach to analyzing corporate default. Even though most bankruptcy studies up until this point in time had been conducted using the MDA approach, Ohlson (1980) criticized the limitations of the method. He mentioned that the MDA approach required that the variance-covariance matrix of the predictors had to be of the same size which in turn meant that there must be an exact equal number of observations belonging to the default and non-default class. In addition, MDA assumes that the predictors are normally distributed which in addition to be problematic in and by itself, does not allow for dummy variables. Lastly, Ohlson (1980) criticized the output of MDA as having little intuitive interpretation. Instead, Ohlson (1980) proposed the use of a probabilistic approach where each observation is given a probability of going bankrupt. In the paper Ohlson presents the famous O-score model:

$$T = -1.32 - 0.407X_1 + 6.03X_2 - 1.43X_3 + 0.0757X_4 - 1.72X_5 - 2.37X_6 - 1.83X_7 + 0.285X_8 - 0.521X_9$$

Where X_1 is the logarithm of total assets/by the GNP price-level index, X_2 total liabilities/total assets, X_3 is working capital/total assets, X_4 is current liabilities/current assets X_5 is a dummy variable that takes the value 1 if the total liabilities exceeds the total assets and zero otherwise, X_6 is net incomes/total assets, X_7 is funds provided by operations/total liabilities X_8 is another dummy variables that takes the value 1 if net income is negative for the last two years and 0 otherwise and lastly X_9 is defined as net incomes at time t minus net income at time t-1 / the absolute value of the net income at time t minus the absolute value of the net income at time t-1.

The three fundamental papers at this point in times was considered to be Beaver (1966), Altman (1968) and Ohlson (1980). However, Zmijewski (1984) criticized all the previous research for having sample bias. Specifically, Zmijewski (1984) states that previous literature used an even number of bankrupt and non-bankrupt companies in their sample despite the fact that only 0.75% of companies in 1934-1984 went bankrupt. Zimjewski argued that "oversampling" the bankrupt firms violates the assumption of random sampling. In addition, Zmijewski argued that previous research suffered from "sample selection bias" which is a result of using a complete data sample selection criterion. In other words, Zimjewski argued that if the probability of going default using complete data is significantly different from not having complete data, then there is a fundamental problem in the model. To circumvent these issues, Zimjewski proposed the use of weighted exogenous sample maximum likelihood (WESML). Despite Zimjewski's critique and proposed solution, the results that took these biases into account did not have any significant statistical difference to the results which did not.

In papers presented by Taffler (1984) and Izan (1984), quadratic discriminant analysis (QDA) were used as an alternative to the MDA proposed by Altman (1968). However, the general conclusion was found that MDA provided better results that QDA did.

2.2 Market-Based Models

A vastly different approach on viewing corporate default was founded by Merton (1973, 1974). Following the development of option pricing theory by Black & Scholes (1972), from which they won the Nobel Prize, Robert Merton provided a link between option pricing, the market value of a firm's assets, the value of its equity and consequently a framework to calculate the probability of default. The intuition provided by Merton was the following; imagine a company financed by equity and zero-coupon bonds (ZCB), then at maturity T of the ZCB, the equity holders can choose to pay the debt holders their promised principal and retain ownership, or they can default and let debt holders take over the assets of the company. This is exactly the same as the equity holders having a call option on the company's assets whereas the debt holders are short a put option on the company's assets plus the face value of debt. This intuition in conjunction with some proposed assumption about the default point made by Kaelhofer, McQuown and Vasicek (KMV), a Moody's subsidiary, together with the newly developed pricing method of Black & Scholes (1972), made it possible to model a stock exchange traded company's probability of default.

Some researchers prefer structural models since they are compatible with the efficient market hypothesis, do not rely on a specific accounting practice as well as not being dependent on sample issues (Agarwal & Taflier, 2008). Miller (2009) show that market-based models outperforms Altman's Z-score in accuracy. This is largely due to the additional information such as asset volatility.

On the other hand, Hillegest, et al (2004) and others question the assumption of the model for example the assumption of normally distributed stock returns is criticized. Another criticism refers to the parameter values which are not directly observable such as asset volatility and expected return on assets which consequently effect the prediction accuracy.

2.3 Non-Parametric Models

Recently, more advanced models that do not rely on a finite set of parameters have been used extensively for classification problems. Despite the fact that the theory behind some of the algorithms were developed centuries ago, the computing power of today has allowed some of these computationally heavy algorithms to be deployed effectively. Among these attractive non-parametric models are support vector machines (SVM) developed by Vapnik & Lerner (1963), Vapnik & Chervonenkis (1964, 1974). SVM is an extension of the theory of maximal margin classifier which utilizes clever "kernel" tricks that allow for non-linear decision boundaries. SVM has proved to provide impressive results for bankruptcy classification in several studies such as Van Gestel et. al (2003) and Shin et. al (2005). However, despite the high accuracy found when using the SVM algorithm, the relationship between the predictors and the response variable is very difficult to interpret.

Another non-parametric model which has been used extensively is neural networks (NN). Odom & Sharda (1990) and Tam & Kiang (1992) were some of the first to apply NN for default prediction. Both studies found that NN outperformed other models such as logistic regression, k-nearest neighbors and decision trees. However, Pompe & Feelders, 1997) found that NN does not outperform traditional parametric models such as MDA.

Lastly, another highly appreciated non-parametric model is random forest (RF). RF is an extension to classical decision trees. RF was developed by Breiman (2001, 2001) who was inspired by Amit & German (1997) and further developed by Cutler et. al (2011). The RF algorithm utilizes the idea behind decision trees by building thousands of separate decision trees and letting the majority vote determine each observations class. RF has also been used in several studies and proven to yield good results while still allowing for interpretation. For instance, Lin & Mcclean (2001), Wagenmans (2017), Altman (2017) found that random forest outperform other classification models.

2.4 Variables

Financial ratios or accounting data has been used since the 1960's to estimate corporate default with over 185 ratios with significance in predicting default. Wang (2011) lists all variables that are significant in more than 4 papers for predicting default and this list consists of 27 different ratios across five categories: "profitability, liquidity, operational efficiency, capital structure and firm size". As it is evident, the vast majority of previous research on corporate default prediction has focused on accounting formation in terms of financial ratios. The founding fathers of corporate default prediction, Beaver (1966), Altman (1968) and Ohlson (1980) all used financial ratios. However, more recent studies in corporate default prediction has tried to incorporate other variables into the analysis such market variables, industry variables, macroeconomic indicators and payment behavioral data.

Keasey & Watson (1991) suggest that the incorporation of market variables is important. There is evidence that market variables in the form of stock returns, stock volatility, market to book ratios and earning per share ratios have significant correlation to bankruptcy Wang (2011). In addition, Chava & Jarrow (2004) suggest that industry effects are important in default prediction.

Regarding the effect of macroeconomic variables, Duffie, Saita & Wang (2007) report a significant relationship between the state of the economy and default hazard rates of individual firms. Carling et al. (2007) also report a significant relationship between the macroeconomy and individual firm default. On the other hand, Koopman et. al (2009) and Koopman, Lucas & Schwaab (2011), argue that macroeconomic indicators alone might either under or over-estimate default risk and that macroeconomic variables could most likely lose their predictive power in conjunction with firm-specific variables.

3. Methodology

The following sections are dedicated to explaining the research design which will be applied in order to answer the research questions and prove or disprove the hypotheses. Specifically, the choice of models and the rationale behind the choice, the general analytical approach behind classification, sampling techniques, how model performance is measured, the specific setup which tests the hypothesis and lastly the packages used in R to conduct the analyses are explained.

3.1 Models

The problem of corporate default prediction can be tackled using three different overall approaches. One approach is using market-based models such as Black-Scholes-Merton's option pricing theory. The other two approaches fall within the supervised learning realm or more specifically, classification. In this thesis a focus on classification is chosen. There are, however, numerous different classification algorithm which can be applied. Classification algorithms can be divided into two different categories; parametric and non-parametric classifiers. Parametric classifiers are defined as models which take a finite number of parameters. Such models could be logistic regression or linear discriminant analysis. Non-parametric models, on the other hand, are not limited to a finite number of parameters. Classification models which are non-parametric are for instance k-nearest neighbors, support vector machines, neural network and random forest. In this thesis the problem is analyzed using logistic regression as a parametric classifier and random forest as a non-parametric classifier.

The rationale behind choosing a model from each category is to test whether a non-parametric model outperforms a parametric model which, in theory, would be the case especially if the assumption behind the parametric model is violated. The rationale behind choosing logistic regression as the parametric model is that logistic regression has some desirable features. Logistic regression assigns a probability for each observation belonging to the bankrupt class. This feature is useful for default prediction. The rational for choosing random forest as a parametric model is based on the fact that random forest in previous applications yield high accuracy while still maintaining an output which can be interpreted relatively easily. Other classification algorithms such as neural networks and support vector machines, which have also been tested to yield good results, are much more difficult, if not impossible, to interpret.

3.2 Classification

The general approach for classification unconditional of which classification algorithm is applied is the following. First, the dataset is divided into two different sub samples, a training set and a testing set. Next, the classification model is trained on the training set and then the trained model is tested on the testing set. If the accuracy of the model is determined by how accurate the model is on the training data, the model will most likely provide unrealistically good results. The reason for this is that the model can "allow" to overfit the data. Therefore, the trained model is tested on a dataset which it "doesn't know". By doing this, the modeler ensures that the model generated is not only suited for one specific dataset, the training data. When testing a model on a dataset which the original model hasn't seen before, the true performance of the model can be evaluated.

However, before building a model on a training set, the following steps should be done. First, before dividing the sample into a training and testing set, the classification models must have an approximately equal distribution of the two classes, i.e. non-bankrupt and bankrupt observations, in the training set. This issue is a major challenge for classification in general since most classification problems have highly skewed datasets where one class is significantly overrepresented. The samples used in this thesis is no different. Due to the nature of reality, the samples used have significantly less bankruptcy observations compared to non-bankrupt observations. There are three general approaches to tackle this problem which will be discussed in section 3.3.

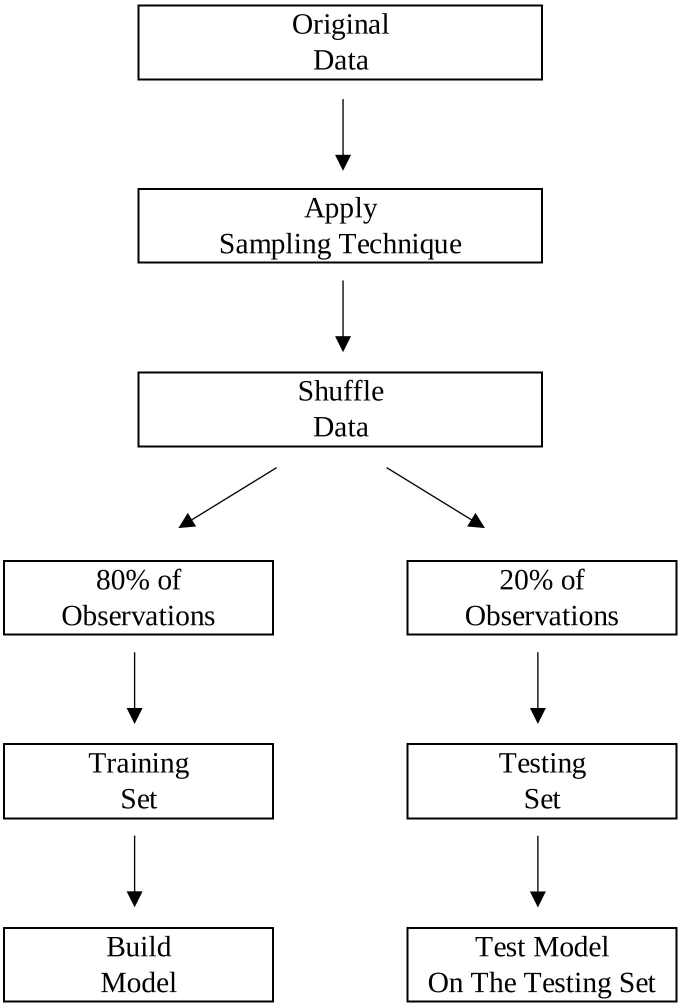
Next, when the original data has been treated by one of the sampling techniques, the data should be randomized. In order to eliminate any structural bias in the sample, the sample is randomly shuffled to mix the order of the observations in the sample such that the allocated observations to the training and testing set doesn't suffer from any bias from the original sample. Randomizing the order of the sample is of outmost importance. If, for instance, the original sample is ranked by a given variable such as "year" or "country", and split into a training and testing set without shuffling the data, the model will be built on a training set which is not representative of the testing set. The result of not shuffling the data before splitting it into a training and testing set will be a highly biased model with poor and misleading results.

After balancing the sample such that there is an equal distribution of the two classes and shuffling the observations, the sample is split into the aforementioned training set consisting of 80% of the data and the remaining 20% is allocated to the testing set. The 80/20 split is the most commonly used split, but there is no scientifically right way to split the data in the training and testing set.

Lastly, after the aforementioned steps have been done, the classification algorithms can be applied and a model generated. Afterwards, the modeler has the option to prune the parameters of the models in order to generate a general model which perform well on all sorts of datasets. When pruning the parameters of the model it is important that the model is pruned on the testing data and not the training data in order to increase the accuracy of the model. If a model's parameters are pruned to yield a higher accuracy on the training data, the model will be highly overfit and specific to the training data.

The general approach is summarized in figure 1.

Figure 1: Classification Procedure



3.3 Sampling

The three general approaches to deal with a skewed sample are oversampling by replication, synthetic minority oversampling technique (SMOTE) and undersampling. Oversampling by replication is simply to copy the observations belonging to the minority class, i.e. bankrupt companies, until there is an approximate equal number of observations in both classes. For instance, if the training sample consists of 100 bankrupt observations and 1,000 non-bankrupt observations, the 100 bankrupt observations are copied 10 times in order to have an equal number of observations between the two classes. Oversampling by replication is highly prone to overfitting. This approach will not be adopted.

The other oversampling approach is synthetic minority oversampling technique (SMOTE). The theory behind this approach will be elaborated in depth later, but the general idea is to synthetically generate minority class observations using a k-nearest neighbor approach. Simply put, SMOTE evaluates which k number of observations belonging to the minority class is closest to the chosen observation and then generates an observation within that space. Here k is a parameter which can be chosen by the modeler. If k is 1, SMOTE finds the closest minority class and generates a minority observation in the space between the chosen observation and its nearest minority class neighbor. If k is 5, SMOTE generates a minority class within the space of the five nearest neighbors of the chosen minority class. SMOTE then loops through the minority observations and generates synthetic observations in order to have an approximately equal number of observations between the two classes. Following the same example as above, if there are 1,000 non-bankrupt observations and 100 bankrupt observations, SMOTE generates 900 additional bankrupt observations such that there is 1,000 of each class. This approach will be adopted

The last method is undersampling. The idea is simple, choose a random subsample of the majority class such that there is an approximately even number of observations within each class. For instance, if the sample consists of 1,000 non-bankrupt observations and 100 bankrupt observations, 100 random observations of the 1,000 non-bankrupt cases are chosen and the rest discarded. The obvious issue with undersampling is that potential informative information is excluded. This approach will also be adopted.

3.4 Model Performance

Throughout the thesis, a confusion matrix is applied as the metric for evaluating model performance. A confusion matrix indicates the number of observations that are true positives, false positives, true negatives and false negatives. An illustration of the idea behind a confusion matrix is seen in figure 2.

By adding the upper left quadrant in figure 2 with the lower right quadrant and dividing that with the total sum of all quadrants the Overall Accuracy (OA) of the model is found. By dividing the upper left quadrant by the sum of the two upper quadrants, the accuracy of the Non-Bankrupt predictions is found. This measure is usually denoted “specificity” however, this will be called Active Accuracy (AA) for the sake of intuition. In a similar fashion, by dividing the lower right quadrant with the sum of the two lower quadrants, the accuracy of the Bankrupt predictions is found, which is normally denoted “sensitivity”, but will be denoted as Bankrupt Accuracy (BA).

Figure 2: Confusion Matrix

	Predicted Class	
Actual Class	# of True Non-Bankrupt	# of False Bankrupt
	# of False Non-Bankrupt	# of True Bankrupt

3.5 Hypothesis Testing

First of all, three overall samples each containing the explanatory variables seen in table 2 are used. The first sample consists of observations from manufacturing companies, the second for wholesale trade companies and the last is a combined sample which does not distinguish between industries.

Figure 3: Samples

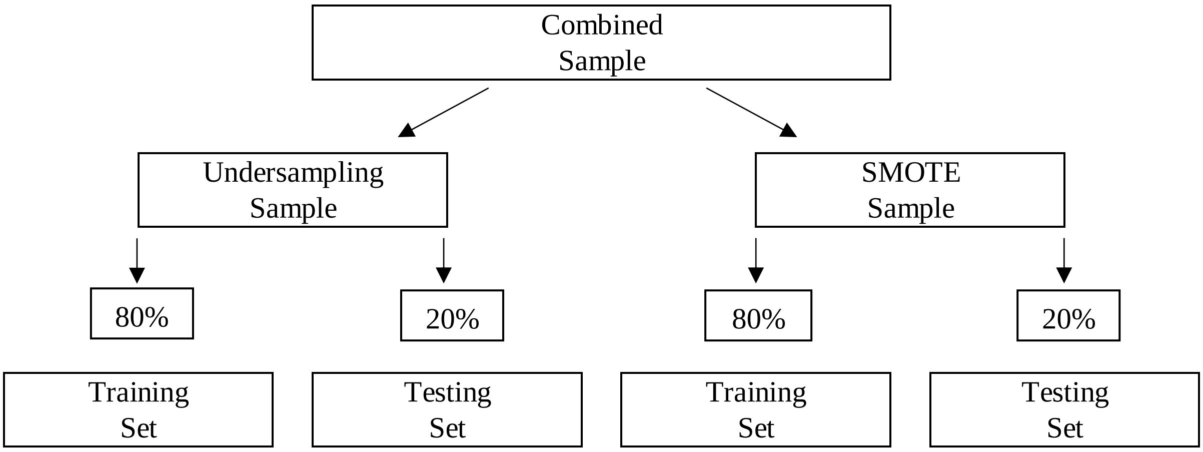
Manufacturing Sample + Wholesale Trade Sample = Combined Sample

In order to test Hypothesis 1, where the performance between logistic regression and random forest is evaluated, the combined sample will be used. For Hypothesis 1, it is not important whether the data stems from one industry or the other, but instead the focus is on model performance. In addition, the analysis will be conducted with the two mentioned sampling techniques, undersampling and SMOTE.

More specifically, two samples are used. A combined dataset is created using undersampling and a combined dataset is created where the training set has been treated by the SMOTE algorithm. Both the undersampling sample and the SMOTE sample are divided into a training set containing 80% of the observations and 20% in the testing set. For the SMOTE case, the 80% distributed to the training sample will undergo the SMOTE algorithm.

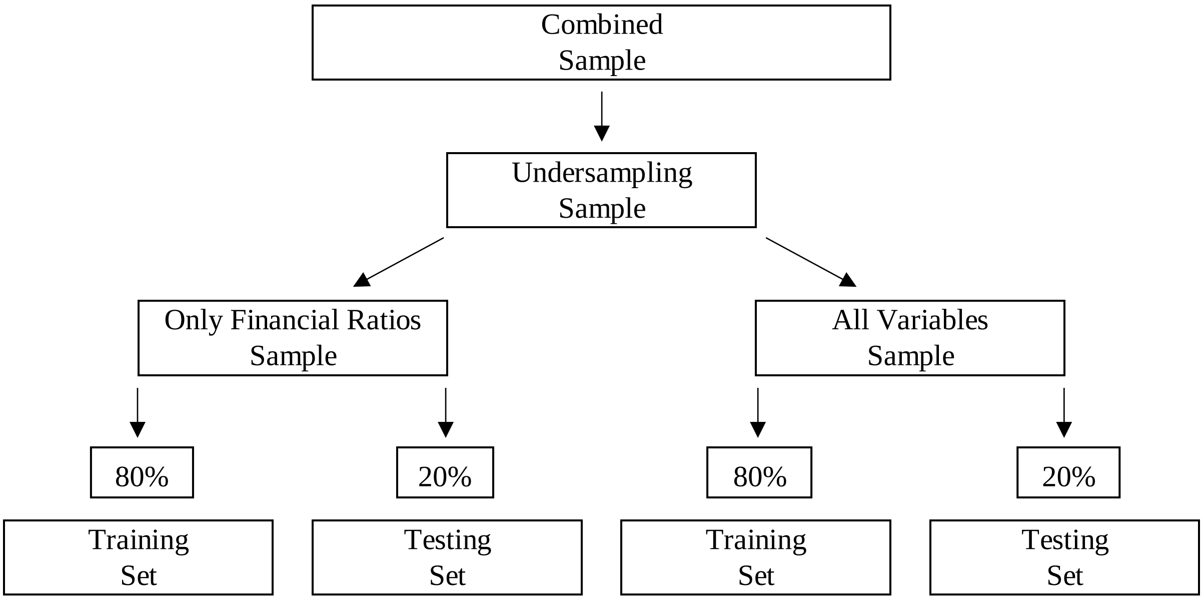
In summary, two samples exist, an undersampling sample which is split 80/20 into a training and testing set and a SMOTE sample which is also split 80/20. Next, a logistic regression and random forest model are tested on each sample, i.e. the combined undersampling sample and the SMOTE sample. The accuracy of the two classification approaches are tested on the respective test sets.

Figure 4: Testing Hypothesis 1



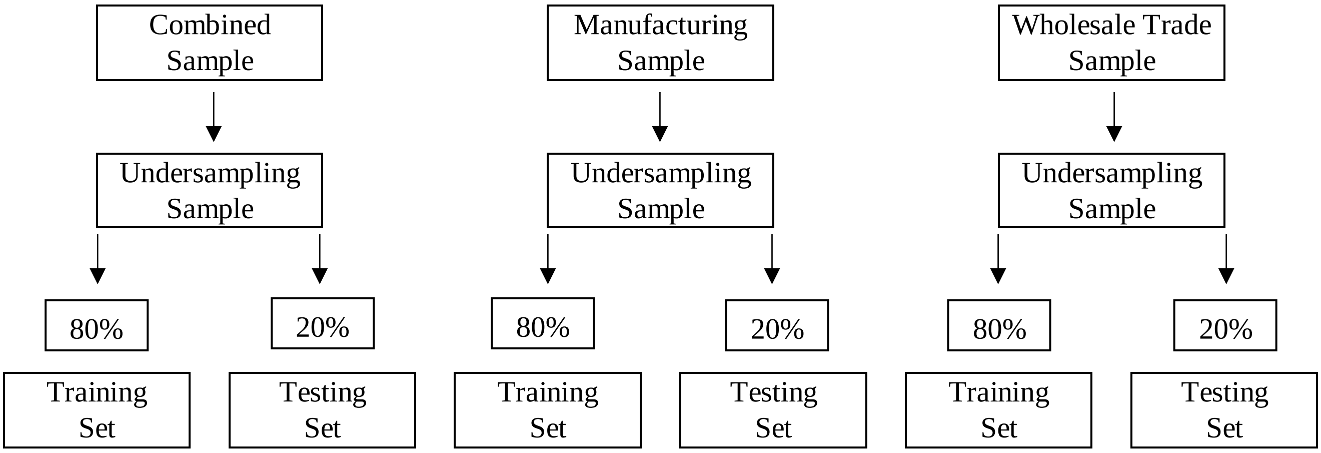
In Hypothesis 2, the focus is on testing whether the addition of non-firm-specific variables add any predictive power to the analysis. To test Hypothesis 2, the following will be done. First, the combined sample is used as the base. The combined sample is then split into two samples where one only contains financial ratios and another containing all explanatory variables. Next, conditional on the results from Hypothesis 1, either both sampling procedures will be adopted, i.e. undersampling and SMOTE or if one of the sampling procedures prove to outperform the other, the one that outperforms will be used. Afterwards, the same procedure as for Hypothesis 1 will be followed. A logistic regression and random forest model are trained on the combined dataset which has all explanatory variables and on the sample with only financial ratios. In the figure below, it is assumed that only undersampling is used for this test.

Figure 5: Testing Hypothesis 2



To investigate Hypothesis 3, the combined sample, the manufacturing sample and wholesale trade sample will be used. In order to test whether the accuracy of default prediction varies across industries as well as investigating which variables are most important in predicting default, the three samples are used. In addition, as in Hypothesis 2, either undersampling or SMOTE or both will be used, conditional on the results from testing Hypothesis 1. The illustration is assuming that only undersampling is used.

Figure 6: Testing Hypothesis 3



3.6 Analysis

All analysis will be conducted in the programming language "R". More specifically, for the Random forest analysis, the package "randomForest" written by Andy Liaw and Matthew Wiener (Liaw & Wiener, 2018) based on research by (Breiman, 2001) and (Cutler, Cutler, & Stevens, 2011) will be used. Logistic regression will be conducted using the standard "Generalized Linear Models (glm)" function written by the R core team based on research by (Dobson, 1990), (Hastie & Pregibon, 1992), (McCulagh & Nelder, 1989) and (Venables & Ripley, 2002). In addition, the LASSO regularization procedure will be conducted using "the Lasso and Elastic-Net Regularized Generalized Linear Models (glmnet)" package written by (Friedman, Hastie, Tibshirani, Simon, Narasimhan & Qian, 2018). Finally, the Synthetic Minority Sampling Technique (SMOTE) will be applied using the "smotefamily" package in R written by Siriseriwan Wacharasak (Siriseriwan, 2018) based on research by (Bunkhumpornpat, Sinapiromsaran, & Lursinsap, 2009).

4. Data Description

In this section of the thesis, the rationale behind the selection criteria of the data extracted from the BVD data base is explained. Furthermore, an in-depth explanation of the explanatory variables is provided.

4.1 Selection Criteria

In order to select the data, each observation must comply with the following criteria;

- The company's financial data stems from 2000-2017
- The company is from France, Spain, Portugal or Italy
- The company operates in either Manufacturing or Wholesale Trade
- The company falls into BVD's classification of Very Large, Large or Medium company size
- The company's financial data stems from unconsolidated financial statements
- The company's Legal status for non-defaulted firms is:
  - "Active"
- The company's Legal status for defaulted firms is:
  - "Active (default of payments)"
  - "Active (insolvency proceedings)"
  - "Bankrupt"
  - "Dissolved (bankruptcy)"
- A defaulted company's latest financial statement is used, others are excluded
- The company must have no missing values

First, in order to limit the analysis to present time as well as to minimize the time changing bias of financial ratios, the time-horizon is limited to 2000-2017. It is commonly believed that financial ratio trends are time dependent and in order to mitigate this bias, narrowing the time-horizon is justified.

Second, to get a sufficient data foundation, a focus on France, Spain, Portugal and Italy is chosen. This decision is based on research conducted by Liapis et. al (2013). The paper investigates the economic similarities of EU countries (Netherlands, UK, Luxembourg, Germany, Finland, Sweden, Austria, Denmark, Belgium, Ireland, Italy, Spain, France, Greece and Portugal) using cluster analysis. The paper analyzes the differences between these aforementioned countries based on the banking sector, tax structure, wages, GDP, current account of balance of payments and government debt and deficit. The conclusion is that the countries in the analysis can be divided into two primary clusters. The first cluster consists of southern European countries i.e. France, Spain, Portugal, Italy and Greece whereas the second cluster consists of the rest of the mentioned countries. Based on the research, the thesis is focused on the southern European countries with the exception of Greece. Since there is limited data available on the BVD data base for Greece, Greece is excluded from the sample.

Third, in order to answer the third research question, a focus on two industries namely manufacturing and wholesale trade is decided. The reason for this decision is based on two factors. One, these two industries are substantially different which is hypothesized to be evident when analyzing which explanatory variables the classification models pick to predict default. Second, both industries have a decent amount of data which is important when conducting classification analyses. The two industries have been classified using the Standard Industrial Classification (SIC). Companies with a SIC code within the 2000-3999 range are classified as manufacturing companies, whereas firms within the 5000-5199 range are classified as belonging to wholesale trade.

Fourth, in order to retrieve a sufficient number of bankrupted firms in the data set, the search criteria are extended to include companies from the very large, large and medium sized segment. This classification excludes firms with operating revenue below 1 million EUR, total assets below 2 million EUR and number of employees below 15.

Fifth, it is expected that unconsolidated financial reports give a more accurate picture of the primary operations and thus the "health" of a company compared to consolidated financial statements which combines financial data from subsidiaries. It is thus decided that all companies with consolidated financial statements are excluded.

Sixth, BVD has 15 different types of "legal statuses" and it is decided that active companies are categorized by having the "active" status and bankrupt companies by having either "Active (default of payments)", "Active (insolvency proceedings)", "Bankrupt" or "Dissolved (bankruptcy)". In table 1, all the categories used in the BVD database are presented. In the classification column of table 1, it is indicated how the legal status classifications are treated in the thesis.

Table 1: Legal Status

Legal Status from BVD	Classification
Active	Active
Active (branch)	Disregarded
Active (default of payments)	Bankrupt
Active (dormant)	Disregarded
Active (insolvency proceedings)	Bankrupt
Bankruptcy	Bankrupt
Dissolved	Disregarded
Dissolved (bankruptcy)	Bankrupt
Dissolved (demerger)	Disregarded
Dissolved (liquidation)	Disregarded
Dissolved (merger or take-over)	Disregarded
In liquidation	Disregarded
Inactive (branch)	Disregarded
Inactive (no precision)	Disregarded
Unknown	Disregarded

Seventh, in order to predict why some companies default and others do not, the latest available financial statement of the defaulted companies are retained and all older statements excluded.

Lastly, the sample is narrowed by excluding all companies with missing values.

4.2 Variable Selection

In the corporate default prediction literature, a vast number of different variables have been analyzed in order to predict default. Most research has been conducted based on financial ratios whereas other explanatory variables such as macroeconomic factors, market variables and commodity prices appear less frequently. In this thesis, 17 firm specific factors of which 16 are financial ratios, four macroeconomic factors, two commodity price indicators and one stock index are used. It is hypothesized that it is necessary to include other variables than financial ratios in order to find the driving factors of corporate default. Each variable is depicted in table 2. The variables will be explained thoroughly in section 4.3.

Table 2: Explanatory Variables

Variable Name	Abbreviation	Category
Financial Ratios		
Return on Shareholder Funds	ROSF	Profitability
Return on Capital Employed	ROCE	Profitability
Return on Total Assets	ROTA	Profitability
Profit Margin	PM	Profitability
EBITDA Margin	EBITDAM	Profitability
EBIT Margin	EBITM	Profitability
Cash Flow Turnover	CFT	Profitability
Net Asset Turnover	NAT	Asset Efficiency
Stock Turnover	ST	Asset Efficiency
Collection Period	CP	Asset Efficiency
Interest Coverage	IC	Solvency
Credit Period	CP	Solvency
Current Ratio	CR	Solvency
Solvency Ratio	SR	Solvency
Liquidty Ratio	LR	Solvency
Gearing	G	Solvency
Other Firm Specific		
Company Age	AGE	Firm Specific
Macroeconomic		
GDP Growth	GDPG	Macro
Interest Rate Nominal Change	IRNC	Macro
Inflation Nominal Change	INFLNC	Macro
Unemployment Nominal Change	UNC	Macro
Stock Index		
STOXX 600 Europe % Change	STOXX	Index
Commodities		
Raw Materials Price Nominal Change	RMPNC	Commodity
Base Metals Nominal Change	BMNC	Commodity

4.3 Explanatory Variables

4.3.1 Financial Ratios

In this section, the 16 financial ratios that constitute the explanatory variables with which the analyses are conducted in order to predict corporate default are explained. As seen in table 2, the 16 financial ratios can be broadly categorized into three overall categories i.e. profitability, asset efficiency and solvency. First, it is explained what these three categories depict in general and thereafter each financial ratio within these three categories are explained in detail. Since these financial ratios lay the foundation of the analysis, it is important clarify what they tell about a company's financial health.

Profitability Ratios

There are numerous different profitability ratios, but the common purpose of them all is to give an indication of a company's ability generate profits. In the analysis, the profitability measures that depict how efficiently a company can generate profits based on assets, employed capital, equity and cash flows are in focus. Even though these profitability measures undoubtedly are correlated, they all have an important interpretation which is fundamental when assessing their relationship to credit risk. In the following section, each of the profitability ratios that are deployed as an explanatory variable in the analyses will be defined.

$$Return\ on\ Shareholder\ Funds\ = \frac{\text{Profit Before Tax}}{\text{Shareholders Funds}} \tag{4.1}$$

In 4.1, shareholder funds is defined as total equity, i.e. capital + other shareholder funds. Return on shareholder funds measures the profitability of the company relative to equity and is thus an investor-oriented profitability measure. All else equal, it is expected that return on shareholder funds is inversely correlated with default risk. The reason is that equity holders are the last to being paid, and the measure thus indirectly tells the company's ability to pay its financial liabilities. A low return on shareholder funds indicates that the company is barely able to pay its financial obligations. A persistent low or even negative return on shareholder funds is a warning sign since it indicates that the primary operations is inadequate and outside capital is necessary for the firm to pay its financial obligations.

$$Return\ on\ Capital\ Employed = \frac{(\text{Profit Before Tax } + \text{ Interest Paid})}{(\text{Shareholders Funds } + \text{ Non Current Liabilities})} \tag{4.2}$$

In 4.2, Non-current liabilities is defined as all long term financial debts + other long term liabilities and provisions. Return on capital employed (ROCE) measures how efficiently a company can generate profits relative to its capital employed. All else equal, it is expected that return on capital employed is inversely correlated with default risk and specifically in asset intense industries.

$$Return\ on\ Total\ Assets = \frac{\text{Profit Before Tax}}{\text{Total Assets}} \tag{4.3}$$

Return on total assets (ROTA) measures how efficiently a company uses its assets to generate profits. Relative to the other profitability measures ROCE and ROTA are particularly important measures in asset intense industries. All else equal, it is expected that return on total assets is inversely correlated with default risk.

$$Profit\ Margin = \frac{\text{Profit Before Tax}}{\text{Operating Revenue}} \tag{4.4}$$

The profit margin measures profit before tax as fraction of operating revenue and is thus measuring the percentage of sales that result in profits. The profit margin is an important measure in terms of credit rating since the measure fundamentally tells how much profit is made from a unit of sales. All else equal, it is expected that the profit margin is inversely correlated with default risk.

$$EBITDA\ Margin = \frac{\text{EBITDA}}{\text{Operating Revenue}} \tag{4.5}$$

The EBITDA margin measures a company’s operating profitability by excluding taxes, depreciation and amortization from the earnings and dividing by operating revenue. The EBITDA margin is thus not a suitable measure for comparing profitability across industries that differs significantly regarding asset intensity due to the substantial differences in depreciation. All else equal, it is expected that the EBITDA margin is inversely correlated with default risk, but it is acknowledged that the explanatory power of the EBITDA margin relative to default risk most likely differs significantly across industries due to the various differences in amortization/depreciation and cost-recognition.

$$EBIT\ Margin = \frac{\text{EBIT}}{\text{Operating Revenue}} \tag{4.6}$$

The EBIT margin is similar to the EBITDA margin except that the EBIT margin incorporates depreciation and amortization into the calculation. The EBIT margin gives a more accurate view of a company's operating profitability since depreciation and amortization is inevitably a part of a company's operations. The EBIT margin is still expected to vary substantially across industries due to the differences in tax and interest burden as well as differences in inherent capability to produce profit. All else equal, it is expected that the EBIT margin is inversely correlated with default risk.

$$Cash\ Flow\ Turnover = \frac{Cash\ Flow}{Operating\ Revenue} \tag{4.7}$$

Cash flow turnover measures the fraction of cash flow that the company generates from its operating revenue. The measure is an indication of how effective a company is at producing cash flow from its operations. The measure is important from a credit risk perspective as a high ratio indicates a company's ability to generate cash flow that will cover the financial liabilities. Therefore, all else equal, it is expected that the cash flow turnover is inversely correlated with default risk.

Asset Efficiency Ratios

Financial ratios within the asset efficiency category measures how efficient a company is a managing their assets in order to generate sales, but also how efficient a company is at collecting payments and managing inventories. Asset efficiency ratios are expected to be especially important for industries which have high average inventories and significant trade receivables. An industry where assets efficiency ratios are expected to have substantial explanatory power regarding default risk is industries such as wholesale- and retail trade.

$$Net\ Asset\ Turnover = \frac{Operating\ Revenue}{\text{Shareholders Funds} + \text{Non Current Liabilities}} \tag{4.8}$$

Net asset turnover measures the ability to generate sales from the assets. The measure indicates of how efficient the company utilizes its assets to generate revenue. All else equal, it is expected that net asset turnover is inversely correlated with default risk.

$$Stock\ Turnover = \frac{Operating\ Revenue}{Stocks} \tag{4.9}$$

In 4.9, stock is defined as total inventories i.e. raw materials + goods in progress and finished goods. The stock turnover shows how many times a company has sold and replaced its inventory over a given period. The measure is an indication of how efficient the company is at selling its inventory. A low stock turnover may imply weak sales, possible overstocking or both. All else equal, it is expected that the stock turnover is inversely correlated with default risk.

$$Collection\,Period = \frac{Debtors}{Operating\ Revenue} * 360 \tag{4.10}$$

In 4.10, debtors are defined as trade receivables from clients and customers only. The collection period measures the average amount of days it takes to receive payment from a sale. All else equal, it is expected that stock turnover is positively correlated with default risk. If a company has a low collection period measured in days, the company is efficient as receiving payments and thus lowering the risk of a counterparty defaulting on its obligations. On the other hand, if the collection period is high, the company is inefficient at managing its trades receivable which has a negative effect on cash flows and also an increased probability of not receiving a promised payment.

Solvency Ratios

Solvency ratios depict a company's ability to cover its financial obligations in the short and long run. Solvency ratios are therefore highly indicative of a company's financial health and important for credit evaluations. Leverage ratios are included under solvency ratios since they depict the same, namely the mix of financing of the company.

$$Interest\ Coverage = \frac{Operating\ Profit}{Interest\ Paid} \tag{4.11}$$

The interest coverage ratio measures how many times the operating profit can cover interest expenses. The interest coverage is highly relevant in credit ratings since it is ultimately the ability of a company to cover its financial obligations that determine whether they default or not. All else equal, it is expected that the interest coverage ratio is inversely correlated with default risk.

$$Credit\,period = \frac{Creditors}{Operating\ Revenue} * 360 \tag{4.12}$$

In 4.12, creditors are defined as debt to suppliers and contractors. The credit period indicates how many days on average it takes a company to pay its short-term financial obligations. A high credit period relative to industry peers might indicate that a company uses its suppliers' money to fund operations cheaply. All else equal, it is expected that the credit period is positively correlated with default risk.

$$Current\ ratio = \frac{Current\ Assets}{Current\ Liabilities} \tag{4.13}$$

The current ratio shows the relationship between short-term assets such as accounts receivable and inventory against short-term liabilities i.e. accounts payable and short-term loans. A low current ratio indicates that the company has trouble financing its short-term financial obligations with their short-term assets. This is a warning sign. On the other hand, a too high current ratio might indicate that the company is inadequately managing their short-term assets. However, all else equal, it is expected that the current ratio is inversely correlated with default risk.

$$Solvency\ Ratio = \frac{Shareholder\ Funds}{Total\ Assets} \tag{4.14}$$

The solvency ratio measures the fraction of assets which is funded using equity. A high solvency ratio indicates that the company is mostly financed by investors. On the contrary, a low solvency ratio means that the company is highly debt financed. Since debtholders require interest and are senior to equity holders, a low solvency ratio can potentially be a warning sign. Therefore, all else equal, it is expected that the solvency ratio is inversely correlated with default risk.

$$Liquidity\ Ratio = \frac{Current\ Assets - Stocks}{Currenct\ Liabilities} \tag{4.15}$$

The liquidity ratio indicates the ability of the company to pay its liabilities using liquid current assets. A liquidity ratio of 1 indicates that the company can finance its short-term liabilities by selling their liquid assets. The measure is therefore important in a credit rating sense. If the liquidity ratio is below 1, the company might not able to pay its short-term financial obligations. On the other hand, if the liquidity ratio is well above 1, the company can easily liquidate current assets to pays it current liabilities. Thus, all else equal, it is expected that the liquidity ratio is inversely correlated with default risk.

$$Gearing = \frac{Non\ Current\ Liabilities + Loans}{Shareholder\ Funds} \tag{4.16}$$

Gearing is a measure of the financial leverage of the company. Gearing simply shows the relationship between debt and equity in a company. It is expected that gearing is positively correlated with default risk.

4.3.2 Macroeconomic Indicators

In this section, the four macroeconomic indicators of the analysis are explained. In H2 it is expected that non-firm-specific variables, including macroeconomic indicators, will improve the precision of the models. The motivation behind the hypothesis is based on the assumption that the worse general economic conditions are, the higher the probability of default amongst individual companies. Whether this causality is due to higher cost of capital via interest rates, lower output due to weakened demand, higher costs due to rising input prices, or some other factor is not specified. The assumptions are, however, explicitly stated in relation to the effect of each individual indicator.

GDP Growth rate

The GDP growth rate is a measure of the growth rate of the economy. Excluding government spending and trade, when the gross domestic product of a country is growing either investment or consumption is growing. Both of which are positive signs of the economy. Therefore, it is expected that an inverse relation between GDP and default probability exists.

Interest rate

The interest rate is the cost of borrowing money in the market. While the nominal interest rate can be difficult to interpret, the change in interest rates has implications in the economy that leads to the believe that a rising interest rate is positively correlated with the probability of default. First, the central banks typically raise rates during the end of the economic cycle to prepare for a slowdown. Therefore, raising rates tend to forecast economic slowdowns in which case the probability of default is expected to rise.

Second, there exists an inverse relation between interest rates and asset prices. The lower the interest rate the higher the asset prices since a low yield implies a higher price. Therefore, when rates are rising and asset prices are falling and moving towards the default point, a higher probability of default is expected. This fits with the framework of the market-based models such as the Black-Scholes-Merton model described in 2.2. In order to capture the latest development of interest rates, the nominal change in the yield on government bonds over the latest year is included as a variable. The variable is denoted IRNC for *interest rate nominal change* as in table 2.

$$IRNC_{it} = Yield_{it} - Yield_{it-1} \tag{4.17}$$

The IRNC is calculated as the annual difference of the respective countries' monetary-related interest rates. In 4.17, *i* captures France, Spain, Italy and Portugal respectively and t goes from 2000-2017.

Inflation

Inflation is a measure of rise in the prices of goods and services. Higher inflation means higher prices while radical inflation indicates something deeply troubling in the economy. It is expected that a rising inflation will be positively correlated with the probability of default for individual companies due to the generally worse state of the economy during high inflation. The caveat of inflation is the fact that neither very high or very low inflation is regarded as productive for the economy and therefore individual companies. This means that the direction of the correlation may differ from the expectations. In order to capture the latest development of inflation the nominal change in inflation over the latest year is included. The variable is denoted INC for *inflation nominal change* as in table 2.

$$INC_{it} = Inflation_{it} - Inflation_{it-1} \tag{4.18}$$

The INC is calculated as the annual difference of the respective countries' inflation. In 4.18, *i* captures France, Spain, Italy and Portugal respectively and *t* goes from 2000-2017.

Unemployment rate

The unemployment rate is a measure of the fraction of able people who are currently unemployed. In general, the unemployment rate is falling when the economy is growing since companies need to hire additional staff while the unemployment rate is rising when companies are struggling and laying off staff. Therefore, it is expected that unemployment is positively correlated with higher default probability. In order to capture the effect of the latest development of unemployment the nominal change in unemployment over the latest year is included. The variable is denoted UNC for *unemployment nominal change* as in table 2

$$UNC_{it} = Unemployment_{it} - Unemployment_{it-1} \tag{4.19}$$

The UNC is calculated as the annual difference of the respective countries' unemployment rate. In 4.19, *i* captures France, Spain, Italy and Portugal respectively and *t* goes from 2000-2017.

4.3.3 Other Variables

In this section, the remainder of the variables in the analysis will be explained. This includes two commodity price indices and a stock index. While commodity prices explain the prices of inputs for some companies, especially manufacturing, the stoxx index is a measure of the capitalization of European companies.

Raw Materials Index

Raw materials are primary commodities such as timber, chemicals, fuels and plastics that are used to produce a variety of other goods. Manufacturing companies are especially sensitive to changes in raw materials prices since they rely heavily on raw materials for inputs. The raw materials index is in real prices, and therefore corrected for inflation, which means that the changes in index value is a result only of the real economic underlying causes and not a monetary effect. It is expected that a higher raw materials index is positively correlated with the probability of default. In order to capture the effect of the latest development of the Raw Materials Index the nominal change in the Raw Materials Index over the latest year is included. The variable is denoted RMPNC for *raw materials price nominal change* as in table 2.

$$RMPNC_t \;=\; Price_t - Price_{t-1} \tag{4.20}$$

The RMPNC is calculated as the annual difference of the prices in the raw materials index. In 4.20, *t* goes from 2000-2017.

Base Metals Index

Base metals are the industrial metals copper, lead, nickel and zinc. Like the case with raw materials, manufacturing companies are sensitive to base metal prices due to cost pressure. The Base Metals Index is in real prices, and therefore corrected for inflation, and therefore the changes in index value is a result only of the real economic underlying causes and not a monetary effect. It is expected that a higher Base Metals Index is positively correlated with the probability of default. In order to capture the effect of the latest development of the Base Metals Index, the nominal change in the Base Metals Index over the latest year is included. The variable is denoted BMNC for *base metal nominal change* as in table 2.

$$BMNC_t \;=\; Price_t - Price_{t-1} \tag{4.21}$$

The BMNC is calculated as the annual difference of the prices in the base metals index. In 4.21, *t* goes from 2000-2017.

Stoxx 600 Europe Index

The Stoxx 600 Europe Index is a combination of 600 large, middle size and small companies covering 17 European countries and 90% of the market capitalization of the European stock market. This makes the Stoxx a proxy for the market value of European business. When the market value is high it is expected that the probability of default is low and therefore, Stoxx is expected to be inversely correlated with probability of default.

$$STOXX_t = \frac{Price_t}{Price_{t-1}} - 1 \tag{4.22}$$

In 4.22, STOXX is calculated as the percentage change of the price of the index where *t* goes from 2000-2017.

4.4 Data Source

4.4.1 Firm-Specific Data

Due to the empirical nature of the thesis and the vastness of data required to conduct the analyses, the choice of the data source is a critical element. In order to retrieve sufficient data, it was decided to use Bureau van Dijk (BVD) to extract all relevant firm specific financial information. BVD is a Moody's analytics company which specializes in compiling financial information of private companies (Dijk, 2019).

4.4.2 Macroeconomic Data

All Macroeconomic data is extracted from the International Monetary Fund's (IMF) data bank (IMF, 2019). More specifically, GDP figures are retrieved from IMF's "GDP and Components" library, Inflation is extracted from IMF's "Prices, Production and Labor" library, Interest Rates are retrieved from the "Interest Rates" library and finally unemployment figures are found as a separate dataset in IMF's data bank.

4.4.3 Stock Index

The Euro 600 Stoxx prices are extracted from the Wall Street Journal's "historical prices" subsection (WSJ, 2019).

4.4.4 Commodities

The commodity data used has been extracted from the World Bank's Commodity Markets website (World Bank, 2019). The World Bank published a commodity price update February 4th 2019, which is used. The report states the price development of a vast collection of stand-alone commodities as well as commodity indices over time. The price development of the "Raw Materials" and "Base Metals" indices as explanatory variables was used in the analyses.

4.5 Validity of Sources

As mentioned, the firm specific data is extracted from the BVD data base, which is a Moody's analytics company. The BVD database is regarded as being highly valid for several reasons. First, since the company operates under Moody's Corporation's analytics division, it is expected that the data has significant validity as Moody's is a well-known and respected credit rating institution. Moody's provides investors and other stakeholders with continuously updated credit ratings, research and financial intelligence and is therefore an essential part of the transparency of the global capital markets. Second, BVD states they carefully capture and enrich the data before making it available by appending and standardizing it. In addition, BVD states that they have information of around 300 million companies around the world and capture data from more than 160 separate providers and hundreds of BVD's own sources (Dijk, 2019).

The Macroeconomic data is extracted from the International Monetary Fund, the commodity prices from the World Bank and the Euro 600 STOXX prices from the Wall Street Journal. All of the aforementioned data sources are regarded as highly valid.

5. Descriptive Statistics

5.1 Sample Description

Throughout this section, the samples are visualized. The output obtained from BVD consists of reporting basis, the annual report year, company name, standardized industrial classification (SIC code). The reporting basis indicates whether the financial data is from unconsolidated or consolidated financial statements. The annual report year indicates which reporting year the data is obtained from. The company and country name variables are self-explanatory and the SIC code indicates which industry each company operates in. The legal status variable depicts the status of each company. After narrowing the sample by the criteria mentioned in 4.1, the primary sample ends up with the following size;

Table 3: Primary Sample | Legal Status

	Manufacturing	Wholesale Trade
Legal Status	#	#
Active	829,090	694,427
Active (default of payments)	815	269
Active (insolvency proceedings)	1,134	579
Bankruptcy	1,267	976
Dissolved (bankruptcy)	712	488

Table 3 depicts the number of observations belonging to the different legal status categories for the two industries. The "Active" companies show the number of non-defaulted companies. As it can be seen, 829,090 and 694,427 non-defaulted companies belong the manufacturing and wholesale trade respectively. The four remaining legal status categories are lumped together as the overall class of "bankrupt".

Table 4: Primary Sample | Year

	Manufacturing				Wholesale Trade			
	Active		Bankrupt		Active		Bankrupt	
Year	#	%	#	%	#	%	#	%
2000	221	0.0%	0	0.0%	209	0.0%	0	0.0%
2001	262	0.0%	1	0.0%	252	0.0%	0	0.0%
2002	384	0.0%	0	0.0%	414	0.1%	0	0.0%
2003	1,169	0.1%	5	0.1%	1,096	0.2%	1	0.0%
2004	1,782	0.2%	3	0.1%	1,780	0.3%	4	0.2%
2005	3,512	0.4%	13	0.3%	3,751	0.5%	10	0.4%
2006	9,242	1.1%	41	1.0%	9,655	1.4%	26	1.1%
2007	66,488	8.0%	101	2.6%	54,658	7.9%	41	1.8%
2008	76,159	9.2%	152	3.9%	61,724	8.9%	55	2.4%
2009	77,374	9.3%	161	4.1%	63,668	9.2%	64	2.8%
2010	78,927	9.5%	216	5.5%	64,853	9.3%	99	4.3%
2011	81,126	9.8%	339	8.6%	67,344	9.7%	183	7.9%
2012	82,638	10.0%	392	10.0%	69,026	9.9%	211	9.1%
2013	84,174	10.2%	526	13.4%	71,339	10.3%	330	14.3%
2014	83,362	10.1%	716	18.2%	70,860	10.2%	577	25.0%
2015	87,795	10.6%	717	18.3%	74,593	10.7%	457	19.8%
2016	84,517	10.2%	480	12.2%	71,182	10.3%	239	10.3%
2017	9,956	1.2%	65	1.7%	8,019	1.2%	15	0.6%
2018	2	0.0%	0	0.0%	4	0.0%	0	0.0%
Total	829,090	100%	3,928	100%	694,427	100%	2,312	100%

Table 4 displays the distribution of the sample across years. This distribution is going to be important when dealing with macroeconomic variables. As it can be seen from table 4, the sample is imbalanced concerning the percentage of observation across years given the companies are active or bankrupt. If the analyses are conducted on an imbalanced dataset using variables that are not firm-specific, but are specific to time, the classification algorithms will see "synthetic" patterns in the data that is merely a result of the imbalanced data set and not true patterns that can predict default. Therefore, the sample must be rebalanced such that the active companies follow the same distribution as the bankrupt observations.

Table 5: Rebalanced Sample | Year

	Manufacturing				Wholesale Trade			
	Active		Bankrupt		Active		Bankrupt	
Year	#	%	#	%	#	%	#	%
2000	0	0.0%	0	0.0%	0	0.0%	0	0.0%
2001	116	0.0%	1	0.0%	0	0.0%	0	0.0%
2002	0	0.0%	0	0.0%	0	0.0%	0	0.0%
2003	580	0.1%	5	0.1%	120	0.0%	1	0.0%
2004	348	0.1%	3	0.1%	481	0.2%	4	0.2%
2005	1,509	0.3%	13	0.3%	1,201	0.4%	10	0.4%
2006	4,760	1.0%	41	1.0%	3,124	1.1%	26	1.1%
2007	11,725	2.6%	101	2.6%	4,926	1.8%	41	1.8%
2008	17,646	3.9%	152	3.9%	6,608	2.4%	55	2.4%
2009	18,690	4.1%	161	4.1%	7,689	2.8%	64	2.8%
2010	25,075	5.5%	216	5.5%	11,894	4.3%	99	4.3%
2011	39,354	8.6%	339	8.6%	21,986	7.9%	183	7.9%
2012	45,507	10.0%	392	10.0%	25,350	9.1%	211	9.1%
2013	61,063	13.4%	526	13.4%	39,647	14.3%	330	14.3%
2014	83,120	18.2%	716	18.2%	69,323	25.0%	577	25.0%
2015	83,236	18.3%	717	18.3%	54,905	19.8%	457	19.8%
2016	55,723	12.2%	480	12.2%	28,714	10.3%	239	10.3%
2017	7,546	1.7%	65	1.7%	1,802	0.6%	15	0.6%
2018	0	0.0%	0	0.0%	0	0.0%	0	0.0%
Total	455,998	100%	3,928	100%	277,770	100%	2,312	100%

In table 5, the rebalanced dataset can be seen. One could argue that there could be a potential issue with excluding otherwise informative data. However, since the analyses will be biased and yield invalid results it is a necessity. In addition, given the abundance of data, more than sufficient records of active data is available. As it will be seen later, the amount of data is actually going to cause problems, but this issue is a common problem in classification which will be discussed in-depth later.

5.2 Undersampling & SMOTE

As discussed in the methodology section, to utilize classification algorithms, the dataset has to be evenly split between the two classes. Simply put, an approximately equal amount of bankrupt and non-bankrupt companies is needed in the sample. This is a substantial issue in most classification problems as the dataset often is significantly skewed towards one class. The data used in this thesis is no exception. Specifically, in this dataset, the distribution between active and bankrupt firms is 99.149% and 99.175% of active firms and thus 0.854% and 0.825% of the total sample belonging to the bankrupt class between manufacturing and wholesale trade respectively. There are three general ways to deal with such issue as explained. The three methods are undersampling, oversampling by replication and synthetic oversampling. As stated in 3.3, oversampling by replication is disregarded, but the other two methods will be used. Table 6 and 7 will display the samples using the two methods.

5.2.1 Undersampling

Since the primary sample is balanced correctly, a random sample can be drawn without concern about the distribution of the randomized sample. The randomized sample should follow the distribution of the primary sample due to the law of large numbers when the number of observations is very high. In the sample shown in table 6, there is 3,928 and 2,312 of bankruptcy observations between manufacturing and wholesale trade respectively. It can be argued that the number of observations is not high enough to ensure that the randomly downsizing of the active class follow the exact distribution of the primary sample. It is, however, assumed that it does which is also confirmed in the appendices. Conclusively, table 6 shows the manufacturing and wholesale trade samples used when undersampling is applied.

Table 6: Undersample

	Manufacturing	Wholesale Trade
Classification	#	#
Active	4,019	2,305
Bankrupt	3,928	2,312

5.2.2 SMOTE

An issue with undersampling is the fact that a significant amount of observations is disregarded and therefore potential information is lost. A way to avoid this problem is to oversample by replicating the minority class x amount of times to have as many observations of that class as the other. However, oversampling by replication will most likely resulted in an overfit model. To avoid this issue, it is possible to synthetically over-sample by using the k-nearest neighbor algorithm. The theory behind SMOTE will be discussed later. The issue with SMOTE in this case is that the total sample size becomes too big to handle. For instance, if the training set consists of 1,500 bankruptcy observations and 200,000 active observations, SMOTE will synthetically create 198,500 new bankruptcy observations which yields a training sample of 400,000 observations. In order to handle this amount of data using classification algorithms, a specialized computer with extensive Random Access Memory (RAM) is needed. In order to navigate around this issue, the primary sample must be narrowed down by a factor of 18 in the manufacturing case, and by a factor of 11 in the wholesale trade case. Table 7 shows the sample used for the classification analyses where the SMOTE function is applied.

Table 7: SMOTE Sample

	Manufacturing	Wholesale Trade
Classification	#	#
Active	25,097	25,301
Bankrupt	3,928	2,312

5.3 Descriptive Statistics

In the following section, the descriptive statistics of the explanatory variables are presented. In table 8, the mean and standard deviation for the active and bankrupt companies from the manufacturing and wholesale trade samples are shown for the total sample. Since it is assumed that the undersampling and SMOTE samples follow the same distribution, the mean and standard deviation should follow the primary sample which in turn makes table 8 representative. This is confirmed in the appendices. Following table 8, a brief elaboration of the descriptive statistics is given. The purpose of the elaboration is to describe how the mean of the variables, primarily the financial variables, looks in relation to the expectations presented in 4.3.

Table 8: Descriptive Statistics

	Manufacturing				Wholesale Trade			
	Active (N=455,998)		Bankrupt (N=3,928)		Active (N=277,770)		Bankrupt (N=2,312)	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Financial Ratios								
Profitability								
ROSF	0.14	0.43	-0.23	1.07	0.17	0.43	-0.13	1.09
ROCE	0.13	0.26	-0.01	0.59	0.17	0.32	0.07	0.61
ROTA	0.05	0.09	-0.02	0.11	0.05	0.08	0.00	0.12
PM	0.03	0.09	-0.06	0.16	0.03	0.06	-0.02	0.12
EBITDAM	0.08	0.08	0.01	0.13	0.05	0.06	0.01	0.10
EBITM	0.04	0.08	-0.03	0.14	0.03	0.06	0.00	0.10
CFT	0.06	0.07	-0.01	0.14	0.03	0.05	-0.01	0.10
Asset Efficiency								
NAT	3.97	9.27	6.57	18.76	8.27	19.74	13.13	25.66
ST	27.04	73.66	18.02	57.88	29.78	82.10	32.62	93.87
COP	94.78	71.24	115.04	110.04	76.91	70.02	95.26	106.20
Solvency								
IC	34.72	105.16	8.46	69.89	36.35	106.81	17.99	88.21
CP	60.36	53.78	87.97	74.33	60.05	58.48	84.08	92.49
CR	1.92	2.06	1.25	0.98	1.95	2.48	1.52	2.76
SR	0.37	0.22	0.20	0.15	0.35	0.22	0.20	0.17
LR	1.40	1.62	0.82	0.70	1.31	1.88	1.18	2.37
G	1.31	1.76	2.96	2.62	1.18	1.71	2.68	2.60
Firm Specific								
AGE	27.93	15.71	26.93	16.59	24.22	13.84	21.42	14.04
Macroeconomic Factors								
GDPG	0.01	0.02	0.00	0.02	0.01	0.02	0.01	0.02
IRNC	-0.01	0.01	-0.01	0.01	-0.01	0.01	-0.01	0.01
INFLNC	0.00	0.01	0.00	0.01	0.00	0.01	0.00	0.01
UNC	0.00	0.01	0.00	0.01	0.00	0.02	0.00	0.01
Stock Index								
STOXX	0.05	0.14	0.05	0.14	0.06	0.12	0.06	0.12
Commodities								
RMPNC	-0.97	7.52	-0.97	7.52	-1.25	7.09	-1.25	7.10
BMNC	-3.15	11.06	-3.15	11.06	-3.00	9.99	-3.00	9.99

The primary rebalanced sample consists of 730,008 companies of which 449,926 are from the manufacturing subsample and 280.082 from the wholesale trade sample. Since bankruptcy is a relatively rare event the samples are skewed toward companies not affected of a default event. As such, 3,928 manufacturing companies defaulted in the sample period while 455,998 companies did not. 2,312 wholesale trade companies defaulted while 277,770 did not.

For the manufacturing subsample the mean of any profitability ratio is positive for the active companies while negative for all except one ratio for the bankrupt companies. This indicates that the expectation of an inverse relationship between the profitability ratios and the probability of default is true. This is also highly intuitive since negative profitability will erode the equity of any company if allowed to continue long enough. Regarding the asset efficiency ratios, it is found that the mean of net asset turnover is higher for bankrupt companies compared to the active companies indicating a positive relationship between net asset turnover and default for manufacturing companies. This contradicts the expectation. On the other hand, Stock turnover is positive and collection period is negative, for active companies which is in line with expectation. For the solvency ratios, the interest coverage ratio, current ratio, solvency ratio and liquidity ratio are all greater for the active companies compared to the bankrupt companies. Gearing and credit period are negative for the active companies compared to the bankrupt companies. The direction of the relationship of all the solvency ratios confirms the expectations.

For the wholesale trade subsample, the mean of the profitability ratios is higher for active than bankrupt companies for any profitability ratio. This indicates a similar relation as for the manufacturing sample and is in line with expectation. For the asset efficiency ratios, net asset turnover is lower for active companies as was the case for the manufacturing sample. In addition, the stock turnover is lower for active companies which is contrary to the manufacturing subsample. The collection period, however, is higher for bankrupt companies than active companies in the wholesale trade sample in a similar manner to the manufacturing sample. The interest coverage ratio, current ratio, solvency ratio, liquidity ratio, credit period and finally gearing follow the same pattern as for the manufacturing sample. This confirms the expectation as stated.

The age of the companies is generally higher for active companies compared to bankrupt companies, however, the difference is relatively small. The mean values of macro variables and market variables are very similar across active and bankrupt companies which should be the case since the sample has been balanced across years.

6. Theoretical Background

6.1 Logistic Regression

Logistic regression was developed in 1958 by the statistician David Cox. Logistic regression is useful when modeling the probability of a binary outcome. In relation to default prediction, logistic regression can model the probability of corporate default and non-default (Tibshirani et al, 2017).

Figure 7: Linear Regression

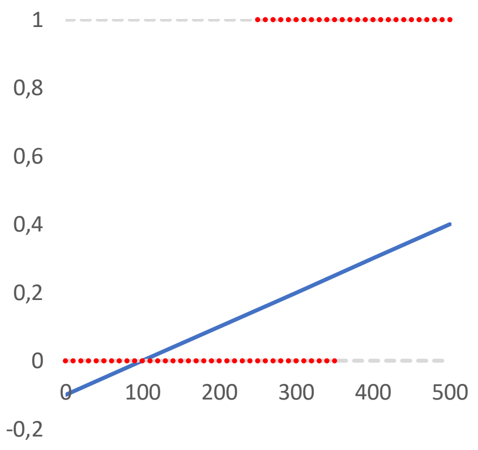


Figure 8: Logistic Regression

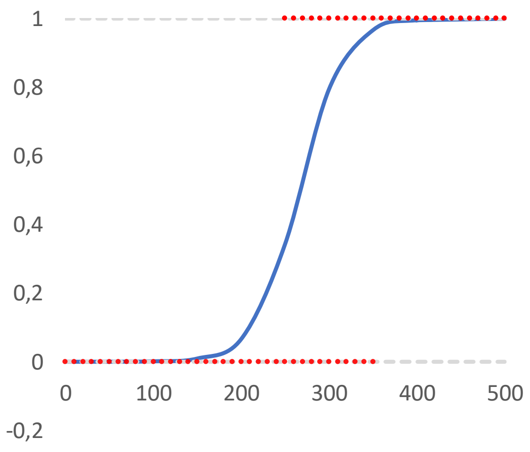


Figure 7 shows the probability of default using linear regression and figure 8 shows the probability of default using logistic regression. Using linear regression some probabilities are negative whereas all probabilities for logistic regression are between 0 and 1.

We can write the probability of default as

$$\text{Probabilty (default } = \text{ yes} \,|\, \text{Input variable)}$$

For logistic regression the value of the probability is between 0 and 1. So for any value of the input variable a default prediction can be made.

6.1.1 Maximum Likelihood Function

The goal of maximum likelihood is to find the optimal way to fit a distribution to the data. In order to calculate the maximum likelihood function, the likelihood of observing each individual observation is calculated. Then the likelihoods are multiplied together which gives a line through the data which tells the likelihood of observing the data. Next, the line is shifted and the likelihood is calculated again. The curve with the maximum value for the likelihood is the one that maximizes the likelihood of observing the data. This is the best fitted line.

6.1.2 The Logistic Model

Since linear models predicts negative probabilities for values close to zero when fit to a binary response, P(X) must be modelled using functions that give outputs between 0 and 1 for all values of X (Tibshirani et al., 2017). This can be done using the logistic function

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \qquad (6.1)$$

The right hand of the formula above shows the fit of the logistic regression model. For low input variables probabilities are estimated close to but never below zero while high input variables predict probabilities close to but not above 1. The logistic function will always produce an s-shaped curve of the form shown in figure 8 and thus always give sensible predictions.

It is possible to manipulate formula 6.1 and find

$$\frac{P(X)}{1 - P(X)} = e^{\beta_0+\beta_1 X} \qquad (6.2)$$

The left-hand side of 6.2 is the odds. The odds can take any value between 0 and infinity. Low values of the odds indicate a low probability of default while high values indicate a high probability of default. An odds of 1/4 indicates that that 20% of people will default since $\frac{0,2}{1-0,2}$ =1/4.

By taking the logarithm of both sides of 6.2 the following is found

$$\log\left(\frac{P(X)}{1 - P(X)}\right) = \beta_0 + \beta_1 X \qquad (6.3)$$

The left-hand side of 6.3 is called logit. A logistic regression model has a logit that is linear in X.

6.1.3 Estimating Regression Coefficients

Since the coefficients are unknown, they must be estimated on training data. The method of maximum likelihood is used to fit the model due to its better statistical properties than non-linear least squares (Tibshirani et al., 2017). β_0 and β_1 are estimated such that the predicted probability of default corresponds as close as possible to the individuals observed state of default. Despite the fact that the s-shaped function is associated with logistic regression, the coefficients are presented in terms of the log odds or logit function.

Table 9: Logistic Regression Coefficients

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.34874	0.04982	27.07	2.00E-16	***
X15	-5.1062	0.16471	-31	2.00E-16	***

***' = 0.001, '**' = 0.01, "*" = 0.1, '.' = 0.1

Table 9 shows the output of a logistic regression. By computing the standard error, the accuracy of the coefficient estimates can be found and the z-value may be measured. A large z-statistic value indicates evidence against the null hypothesis. The null hypothesis states that β_1=0 which implies that default is not dependent on the input variable. When the P-value is small the null hypothesis can be rejected which means that default is dependent on the input variable.

6.1.4 Predicting Default

After estimating the coefficients, it is simple to compute the probability of default given the input variables. By plugging in the input variable into 6.1 the probability default may be estimated.

6.1.5 Multiple Logistic Regression

Now, prediction of binary responses using multiple predictors will be discussed. It is possible to extend from simple to multiple logistic regression such that formula 6.3 can be extended to

$$\log\left(\frac{\mathrm{P(X)}}{1-\mathrm{P(X)}}\right) = \beta_0 + \beta_1 \mathrm{X}_1 + \cdots + \beta_p \mathrm{X}_p \tag{6.4}$$

Where X=(X_1,…, X_p) are p predictors. Equation 6.4 can be rewritten as

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}} \tag{6.5}$$

Similarly to simple logistic regression, the maximum likelihood method is used to estimate β_0, β_1,…, β_p

Table 10: Multiple Logistic Regression Coefficients

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	0.5292104	0.1300923	4.068	4.74E-05	***
X1	-0.0455181	0.0848365	-0.537	5.92E-01	
X2	0.0777748	0.1385859	0.561	5.75E-01	
X3	-2.3086361	0.7650097	-3.018	2.55E-03	**
X4	0.1367805	0.746409	0.183	8.55E-01	
X5	-9.298257	1.6125614	-5.766	8.11E-09	***
X6	-3.6354698	1.0704279	-3.396	6.83E-04	***
X7	7.0066538	1.8983673	3.691	2.23E-04	***
X8	0.0029473	0.002627	1.122	2.62E-01	
X9	0.0014703	0.0003712	3.961	7.48E-05	***
X10	-0.0008311	0.0004554	-1.825	6.80E-02	.
X11	0.0013283	0.000416	3.193	1.41E-03	**
X12	0.0019621	0.0006112	3.21	1.33E-03	**
X13	0.0655901	0.045201	1.451	1.47E-01	
X14	-0.254875	0.0721659	-3.532	4.13E-04	***
X15	-3.2418419	0.2706744	-11.977	2.00E-16	***
X16	0.1058862	0.0185204	5.717	1.08E-08	***
X17	-0.0019132	0.0019149	-0.999	3.18E-01	
X18	9.1643363	2.9194364	3.139	1.70E-03	**
X19	-5.5711844	4.2641657	-1.307	1.91E-01	
X20	6.1836254	5.2467102	1.179	2.39E-01	
X21	1.1588774	3.7272425	0.311	7.56E-01	
X22	0.254656	0.3690484	0.69	4.90E-01	
X23	-0.0007404	0.0062114	-0.119	9.05E-01	
X24	-0.0043581	0.0047588	-0.916	3.60E-01	

***' = 0.001, '**' = 0.01, "*" = 0.1, '.' = 0.1

Table 10 shows the output of logistic regression using multiple input variables. Analogously to simple logistic regression by computing the standard error, the accuracy may be measured. A large z-statistic value indicates evidence against the null hypothesis that β_p=0 which implies that default is not dependent on this input variable. When the P-value is small the null hypothesis can be rejected which means that default is dependent on that specific input variable.

6.1.6 The Lasso

The *least absolute shrinkage and selection operator (Lasso)* is one way to perform variable selection. While other methods such as ridge regression will generate models that use all input variables as predictors, the lasso is a way to include only a subset of the most important variables by reducing insignificant variables to a value of zero. By doing so, it is possible to overcome problems of model interpretation using other methods such as ridge regression (Tibshirani et al., 2017).

More specifically, the lasso coefficient β_λ^L minimizes the quantity

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j = RSS + \lambda\sum_{j=1}^{p}\beta_j \tag{6.6}$$

For sufficiently large values of λ, lasso forces some of the coefficients estimates to be exactly equal to zero.

Table 11: LASSO Regularization

(Intercept)	0.6342301		
X1	.	X13	.
X1	.	X14	-0.0312133
X3	-2.3293221	X15	-3.176993
X4	-0.8233796	X16	0.072926
X5	-2.8356174	X17	.
X6	.	X18	0.2054826
X7	.	X19	.
X8	.	X20	.
X9	.	X21	.
X10	.	X22	.
X11	.	X23	.
X12	0.001514	X24	.

Table 11 shows the output of lasso regularization. As shown, only 8 variables in addition to the intercept are included which means that 16 variables have been discarded. In general, the lasso lists the variable and tests their significance. If the value is beneath a certain threshold then the variable is excluded. This process is repeated until all variables are significant. An interesting result of the lasso is that some variables that are not significant initially will become significant. This is due to assumptions of logistic regression that there is no multicollinearity between the input variables (Stoltzfus, 2011). When some of the variables are excluded that indeed are heavily correlated, then the remainder will become significant.

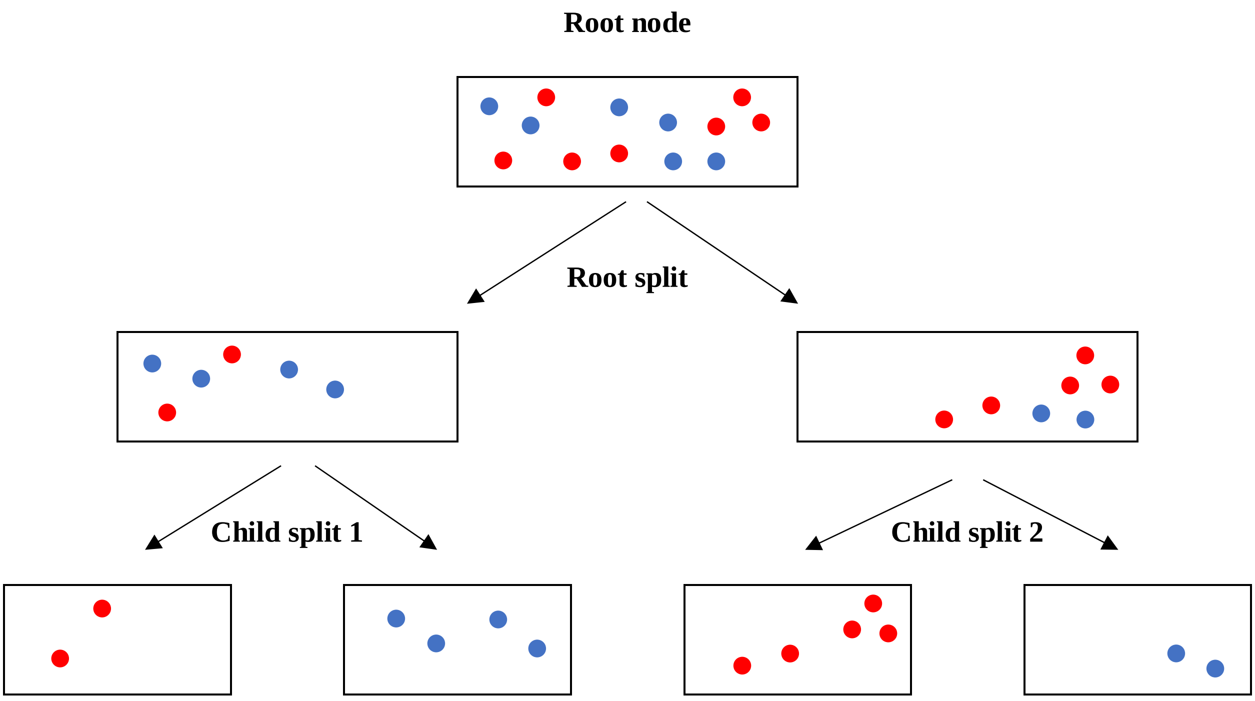
6.2 Random Forest

To understand how Random forest works, it is necessary to go through binary recursive partiotioning trees, or simply; classification trees. Since Random forest is an extension of classification trees, it is a precondition to understanding Random forest that the idea behind classification trees is established.

6.2.1 Classification Trees

To begin with, it proves helpful to illustrate how classification trees work by showing a figure of the general idea behind the model. As it can be seen in figure 9, the root node depicts the entire predictor space and each blue circle represent one class, for example the "active" companies and the red cricles represent the other class i.e. "bankrupt" companies.

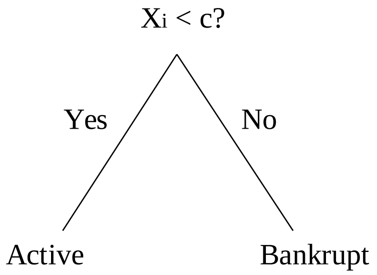
Figure 9: Classification Tree Illustration



The goal of classification trees is to split the data such that each new split is more "pure" than the previous split. The splitting criterion used will be elaborated later, but for now, "pure" means that each split contains a higher percentage of one class compared to before the split. As it is illustrated in figure 9, the root node is split into two root splits. It is evident that each of these splits are more "pure" than the root node since the root note contains 6 and 7 equivalent to 46.2% and 53.8% of each class whereas after the split, the left root split has 4 (66.7%) of the blue class and 2 (33.3%) of the red class and the right split root split has 2 (28.6%) of the blue class and 5 (71.4%) of the red class. This process of splitting the data continues until some stopping criteria is met. Such criteria could be that the data should not be split further once there is less that x observations in a node. The last split in a classification tree is called the terminal node, so each of the four last boxes are terminal nodes.

When a classification tree has been built based on a training sample, each observation in an independent test sample goes through the tree and depending on the rules of the tree each observation will end in a terminal node and receive a classification i.e. either blue = "Active" or red = "Bankrupt". An example is seen in figure 10. Here X_i is a predictor variable and c is the split-point. If X_i is smaller than c, the observation will be classified as "Active" and if it is larger than c, it will take the other path and be classified as "Bankrupt".

Figure 10: Splitting



6.2.2 Splitting Criterion

The basic principals of classification trees have now been established, but in order to fully understand how the algorithm decides whether to split a variable at a given point, the splitting criterion used must be described. There are several different splitting criteria, but one of the most common is the gini index.

$$Gini\ Index = \sum_{k=1}^{k} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{6.7}$$

In 6.7, k is the number of classes, $\hat{p}$ is the proportion of observations in the mth split that belong to class k. The gini index measures the amount of information there is in each split. A Gini Index of 0.5 indicates that there is no valuable information whereas a low gini index suggests that there is indeed valuable information i.e. a split with predominately one class. Figure 11 illustrates how the gini index is computed in practice using the example from figure 9.

Figure 11: Gini Index

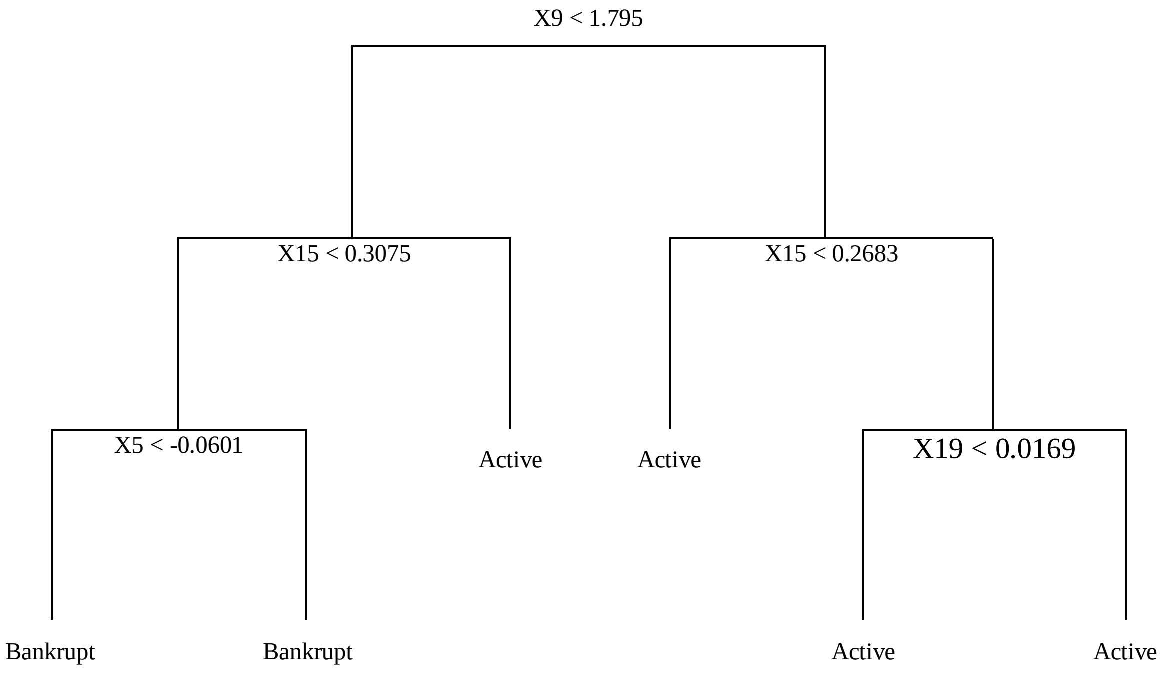
	Root Node	
Class	Blue	Red
# Obs	6	7
Proportion (p)	46%	54%
p(1-p)	0.249	0.249
Gini Index	**0.497**	

	Root split 1		Root split 2	
Class	Blue	Red	Blue	Red
# Obs	4	2	2	5
Proportion (p)	67%	33%	29%	71%
p(1-p)	22%	22%	20%	20%
Gini Index	0.444		0.408	
Weighted Gini Index	**0.425**			

As it can be seen in figure 11, the gini index of the root node is 0.497 since the sum of each gini index for the blue and red class is 0.249. Below the root node, is the gini index of the each of the root splits i.e. 0,44 and 0,408. The final gini index of the root split is then the weighted average of the left and right root split i.e. (0,44*6+0,408*7)/13 = 0.425. Since 0.425 is lower than 0.497, the split has a “gain” in information. A gini index of 0.5 suggests that there is no valuable information, but as the gini index decreases, the more information is to be gained from the split. The goal of the classification algorithm is then to compute all possible splits and evaluate their respective gini index. The split with the lowest gini index is then chosen. This process is continued until the stopping criteria is met and the tree is finalized.

Figure 12 is an illustration of a “real” generated classification tree from the R package “tree”. The tree has been modified to fit the format of the thesis. The tree is based on real data, but the variables are generically named since the results are secondary here.

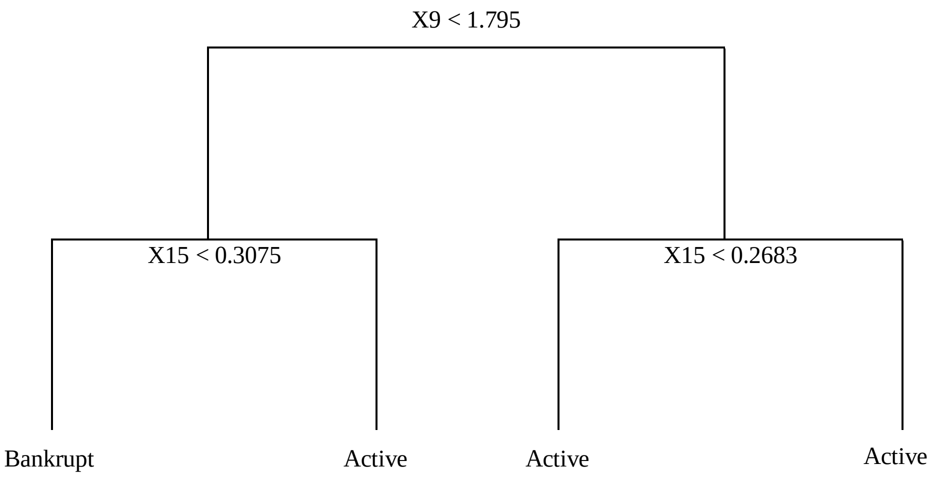
Figure 12: Classification Tree



6.2.3 Pruning

As it can be seen from figure 12, some of the splits at the bottom of the tree seems to be redundant. Specifically, the splits at X5 and X19 do not give any additional information since both the splits yield the same results i.e. “Bankrupt” for X5 and “Active” for X19 no matter what the value of the observation is. To deal with this problem, it is possible to “prune” the tree or in other words shrink the tree to eliminate splits that do not yield additional information, but only increase the risk of overfitting the training data. Without going into too much detail, the pruning is based minimizing a parameter in the cost-complexity function which investigates the effect of collapsing splits in the tree in relation to the “goodness of fit” to the data. If a split can be collapsed without exceeding some “goodness of fit” boundary, the split is taken away and the tree compressed.

Figure 13: Pruned Classification Tree



6.2.4 Classification Tree Issues

Even though the classification tree algorithm is a clever way of analyzing data while still being easy to interpret, there is a major issue than one has to be aware of. Classification trees are prone to having high variance which result in instability. Small changes in the data can lead to a dramatic change in the outcome of the splits being picked due to the hierarchical nature of the splitting process. In other words, an error in the top of the tree will funnel down to every split below. This claim is substantiated in table 12.

Table 12: Instability of Classification Trees

Seed	1		2		3		4	
	Variable	Value	Variable	Value	Variable	Value	Variable	Value
Root Node	X9	<1.865	X9	<1.795	X9	<1.735	X15	<0.265
Left Split	X1	<-0.205	X1	<-0.205	X1	<-0.159	X9	<1.085
Right Split	X15	<0.268	X15	<0.268	X15	<0.268	X9	<1.555
Left Child	X15	<0.317	X15	<0.351	X15	<0.351	X15	<0.085
Right Child	X19	<0.012	X19	<0.017	X19	<0.017	X19	<0.012

In table 12, four different classification trees are built using the same data set. The only difference is that the data is randomly shuffled with different seeds before splitting it into a training and test set. As it can be seen, for seed 1, 2 and 3 the same variables are chosen, but with different cut-off points for the variables. Using seed = 4 a completely different tree is grown using another set of variables than the previous trees. Table 12 is a good example of the instability of classification trees and how small changes in the data can have a substantial effect on the subsequent tree built and ultimately the result.

One way to avoid this problem is to build a "forest" of trees i.e. thousands of classification trees where each tree is generated using a random subset of the training data and a random subset of the explanatory variables. This is exactly the idea behind random forest, which will be elaborated next.

6.2.5 Random Forest

As briefly mentioned above, Random forest is an extension to the traditional classification tree. There are three components of the Random forest that need to be elaborated which is "bagging", "decorrelated trees" and "variable importance".

6.2.6 Bagging

Bagging is a powerful procedure which significantly eliminates the high variance which classification trees are prone of. First, however, it might prove useful to explain why bagging works in the context of a regression tree. A regression tree is similar to classification trees except the tree predicts a continuous variable instead of a discrete i.e. "Active" or "Bankrupt".

Given a set of n independent observations $Z_i ... Z_n$ where each observation has a variance of σ^2, the variance of the mean $\bar{Z}$ is $\frac{\sigma 2}{n}$. This result indicates that averaging a set of observation will reduce the total variance (James, Witten, Hastie, & Tibshirani, 2013). Therefore, in order to reduce the variance and increase the predictive power, it is useful to take numerous training samples and build a regression tree on each of the training samples and averaging the predictions. However, this method is not applicable in real life since multiple training sets are not available. This is where the bagging procedure steps in. Instead of averaging the predictions generated using multiple regression trees built on sperate training sets, it is possible to randomly pick multiple B sub-samples from one training set. After, the regression tree algorithm is trained on each of the B randomly picked sub-samples from the training set in order to get $\hat{f}^{*1}(x), \hat{f}^{*2}(x) \dots \hat{f}^{*b}(x)$ predictions and finally averaging them;

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x) \tag{6.8}$$

Translating this result into a classification tree context, instead of averaging the predictions, the easiest way is to have each observation in the training set moved through each of the classification trees to receive a class. The class of a given observation predicted by each classification tree is then stored. Finally, the majority vote decides whether the observation should be classified as “Active” or “Bankrupt”. In other words, if a 1,000 classification trees are generated, an observation will receive 1,000 predictions, one for each classification tree. It is then the most commonly occurring class among the 1,000 predictions that decide the final prediction of the observation. This procedure is done of each observation in the training sample.

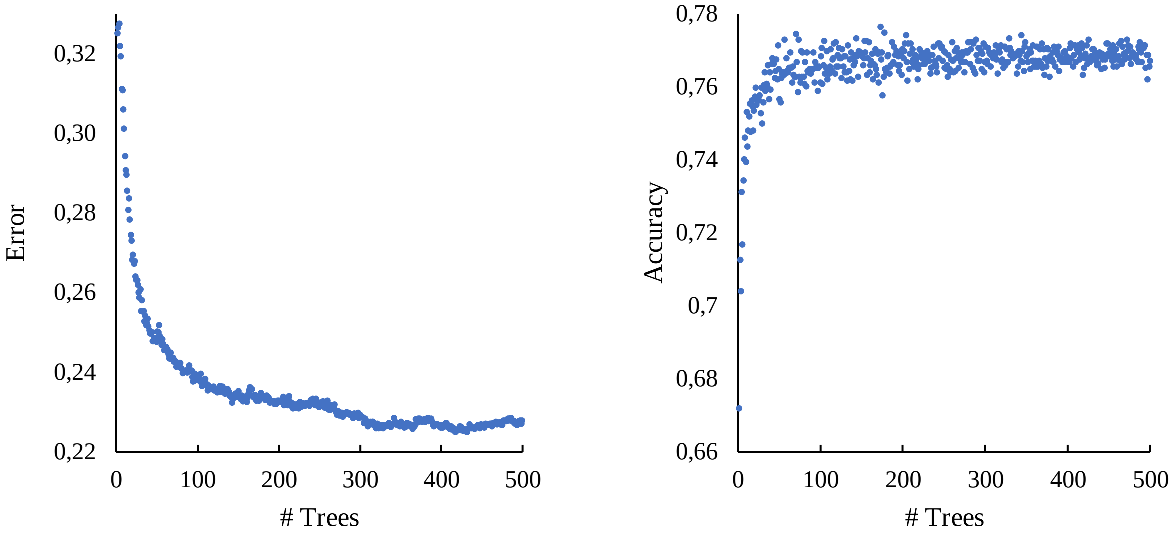
6.2.7 Decorrelated Trees

The powerful idea behind bagging has now been explained, but there is another trick which is used in the random forest algorithm that further reduces variance. Imagine a training set with *m* explanatory variables, the chance is that there is one of these variables which is better at prediction default than the rest of the variables. By only using the bagging approach, each of the 1,000 generated trees will have this strong variable at the top of the tree for each of the 1,000 trees. It might sound reasonable that the variable that predicts default best is at the top each time, but the problem is that each of the 1,000 generated trees will be highly correlated (James, Witten, Hastie, & Tibshirani, 2013). The issue is that averaging the variance of highly correlated trees will not reduce the total variance as much as desired. Again, the averaging procedure relates to regression trees, but the same logic is applied for the usefulness of the majority vote used in classification trees.

Thus, instead of simply generating 1,000 trees using all explanatory variables available, the random forest algorithm chooses a random sample of the *m* predictors at each split of the tree. By doing this, each tree will be inherently different with different predictors at each split of the 1,000 trees. This might prove a problem if only 5 trees are generated, but as the number of trees increase, the variance decreases and the predictive power increases as seen in figure 14.

In figure 14, it is evident that as the number of trees generated increase, the error of the random forest model decreases. In connection, when looking at the accuracy, it is evident that the accuracy of the predictions increases as the number of trees increase.

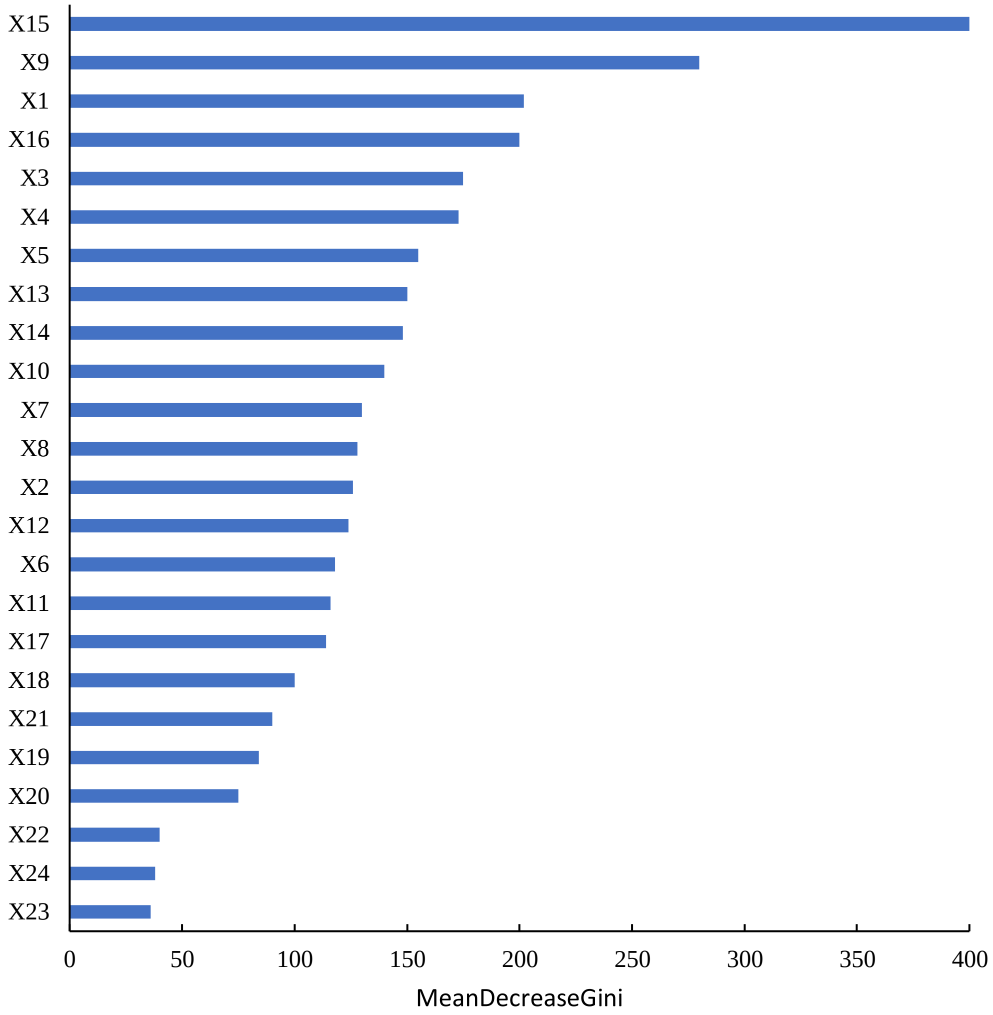
Figure 14: Error & Accuracy As A Function Of # Trees



6.2.8 Variable Importance

Unlike the traditional classification tree where it is possible to see which variables are at each split as well as the cut-off point, random forest does not provide such an overview. Due to the fact that random forest is based on many trees, it is not possible to see an "average" tree. Instead the decrease in the gini index for each variable at every split across all trees is averaged. The variable importance thus indicates which predictors increase the "purity" of the data the most and can therefore be used in order to evaluate which variables are in fact better at prediction default in this case. A plot of the variable importance of the first 15 variables for the example used through this chapter can be seen in figure 15

Figure 15: Variable Importance



In figure 15, it is evident that variable X15 is best at predicting default since it decreases the gini index the most and thus the variable that provides most purity of the nodes when used for splitting the data. This logic is continued and finally it is thus evident that the 5 most important variables for predicting default is variable X15, X9, X1, X16 and X3. It is important to mention that the total value of gini is arbitrary and cannot be used to compare different models using different datasets. This is due to the fact that the mean decrease in gini will increase as the number of explanatory variables increase. In other words, a mean decrease in gini of 300 cannot be compared to a value of 200 found using another dataset. The relative gini between variables in a plot such as figure 15 is however not arbitrary.

6.3 SMOTE

The purpose of this sections is similar to that of the previous, namely to give an intuitive understanding of how synthetic minority oversampling technique (SMOTE) works. Since class imbalance is arguably one of the most important challenges to consider before applying any classification model, it is important to understand how they work. As explained in the methodology section, the three general sampling methods are undersampling, minority oversampling by replication and SMOTE.

As mentioned, oversampling by replication will not be used and undersampling is fairly straightforward. However, SMOTE is more complex and requires some explanation. In its essence, the SMOTE algorithm generates synthetic samples of the minority class by applying the K nearest neighbor (K-NN) approach. Illustrations will be used to explain how this works.

Figure 16: Class Imbalance

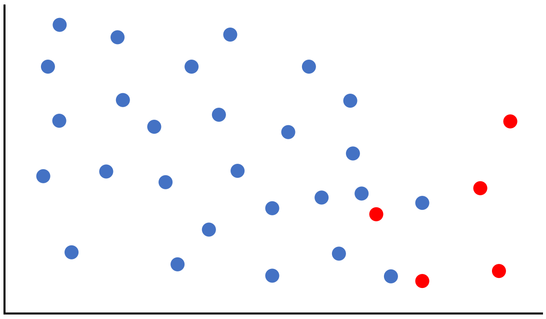
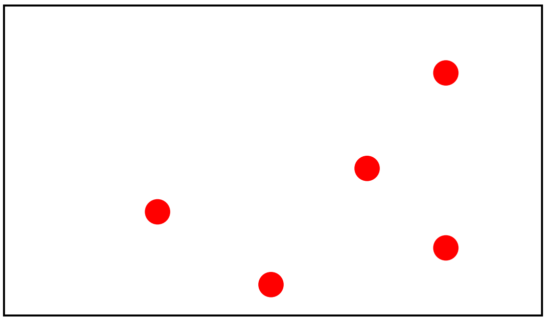


Figure 16 depicts a simple sample distribution of a training sample in 2-dimensional space where the two colors represent each class. As before, blue and red represent “Active” and “Bankrupt” companies respectively. As seen in figure 16, the two classes are highly imbalanced, where the “Active” observations outweigh the “Bankrupt” significantly. The SMOTE algorithm seeks to generate more observations of the minority class such that there is approximately an equal distribution of both classes. The purpose of oversampling the minority class is to give the classification models more data in the training sample in order to hopefully train a classification model more accurately. Generally, classification algorithms will face difficulties predicting the minority class if the minority class is significantly underrepresented.

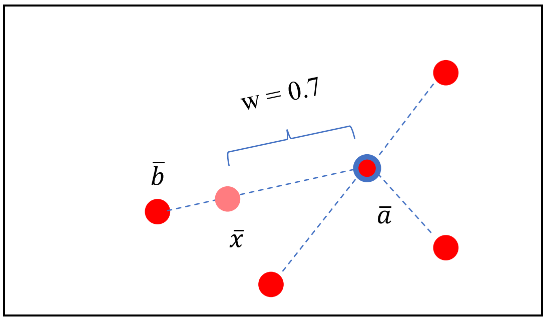
If we zoom in on the minority class and forget about the majority class, the following picture emerges;

Figure 17: Zoomed-in



The SMOTE algorithm consists of three steps. First, SMOTE randomly chooses one of the minority observations and computes the linear distance between the chosen observation and all its neighbors. Then, SMOTE ranks the distances and finds the "k" nearest neighbors. For instance, if the modeler chooses k = 4, all of the minority classes above will be found since there are only 5 observations in total. After this, depending on the amount of synthetic observations needed, SMOTE chooses one of the k-nearest neighbors and places a synthetic observation in the line segment connecting the original randomly chosen observation and the random chosen k-nearest neighbor by multiplying the linear distance with a value between 0-1. Figure 18 illustrates this process.

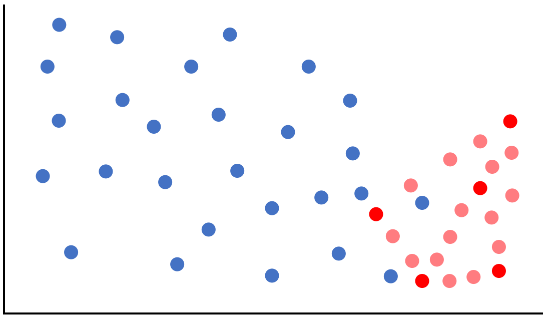
Figure 18: SMOTE k=4



In figure 18, $\bar{a}$ is the initial randomly chosen observation, its k=4 nearest neighbors are all of the 4 surrounding observations and the blue line segments are the distance between $\bar{a}$ and the 4 nearest neighbors. Of the 4 nearest neighbors, $\bar{b}$ is randomly chosen and a synthetic observation $\bar{x}$ is generated by multiplying the distance between $\bar{a}$ and $\bar{b}$, with a random number between 0 and 1, which is 0.7 in this example. Concretely, $\bar{x}$ is generated by randomly interpolating the two samples such that $\bar{x} = \vec{a} + w(\vec{b} - \vec{a})$ where w is a randomly chosen weight in

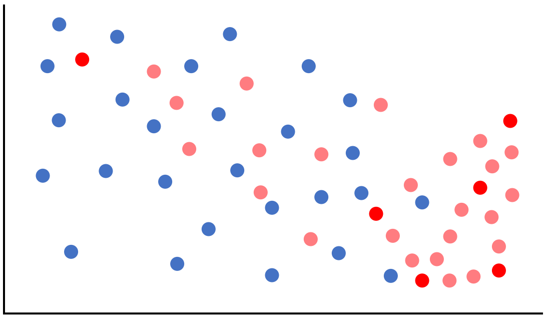
[0,1] (Last, Douzas, & Bacao, 2017). This process is then iterated such that a new randomly chosen observation becomes $\bar{a}$ with a related $\bar{b}$ and finally a new synthetic $\bar{x}$. The final result can be seen in figure 19 where the pink circles represent the synthetically generated observations;

Figure 19: After SMOTE



As one might imagine, one of the major issues with the SMOTE algorithm is outliers. If, for instance, there is an outlier deep into the majority class space, the SMOTE algorithm will generate synthetic observations which will distort purpose of oversampling the minority class. An illustration of the implications of outliers can be seen in figure 20.

Figure 20: SMOTE With Outlier



Even though the SMOTE algorithm is powerful in classification problems and enables a classification model such as Random forest to have “more” data to train the model, there are potential issues with SMOTE as seen in figure 20.

7. Results

This section is devoted to testing the three hypotheses stated in 1.1.2. The section will be divided into three parts, each of which are dedicated to answering the three hypotheses. The goal is to present detailed results which will lay the foundation of material to discuss and interpret in relation to answering the hypotheses.

The first part will depict some detailed sensitivity analyses of the performance of random forest and logistic regression in relation to the accuracy of predicting default. The analyses will be conducted using both undersampling and synthetic minority oversampling technique (SMOTE).

The second part will dig slightly deeper than a pure model accuracy comparison. This part will shed light on whether the addition of non-firm-specific variables will result in higher accuracy using both random forest and logistic regression compared to financial ratios alone.

The last and final part of the analysis will focus on whether the accuracy of default prediction is conditional on industry using both random forest and logistic regression. In addition, the final part will look at which variables that are found to be the most informative when predicting default for both random forest and logistic regression and whether the variables chosen are different on an industry level.

7.1 Hypothesis 1

To test this hypothesis, the results of the accuracy are reported on two different dimensions, a model and sampling dimension. First, the results using random forest and logistic regression using the undersampling technique are presented. Next, the same type of output using SMOTE as the sampling method is shown.

To give a justified assessment of the performance of random forest and logistic regression, it is important to test and compare the accuracy of each model using two vastly different types of sampling techniques.

Lastly, all results presented in this section are based on the combined sample. In other words, since the goal is to address model performance, it is not important to divide the analysis into the two samples; wholesale trade and manufacturing, instead, the combined sample is used as described in 3.6.

7.1.1 Undersampling

Table 13: Random Forest Accuracy With Undersampling

	# of Trees (n)								
	200			500			1000		
# of splits (m)	AA	BA	OA	AA	BA	OA	AA	BA	OA
4	0.762	0.788	0.775	0.753	0.789	0.771	0.753	0.791	0.772
5	0.749	0.788	0.768	0.755	0.789	0.771	0.750	0.794	0.771
6	0.753	0.791	0.771	0.748	0.796	0.771	0.750	0.796	0.773
7	0.750	0.791	0.770	0.750	0.796	0.773	0.750	0.793	0.771
8	0.742	0.794	0.767	0.754	0.791	0.772	0.751	0.793	0.771

Table 13 shows a sensitivity analysis of the random forest accuracy across number of trees and number of variables included in each tree. As mentioned in 3.4, AA (Active Accuracy) stands for the accuracy of predicting firms that are active and predicted to still be active, BA (Bankrupt Accuracy) is the accuracy of the model in relation to predicting firms that did in fact go bankrupt and lastly, OA (Overall Accuracy) is the accuracy of the model as a whole.

As it can be seen from the table 13, the accuracy for AA, BA and OA increases across number of trees, which is expected. The accuracy across number of variables per tree is relatively stable with a slight increase in accuracy up until 6 variables per tree and a decrease in accuracy past that point. Looking at this exact point which yields the best accuracy, an AA of 75%, BA of 79,6% and OA of 77,3% is found. This result indicates that the random forest algorithm, using the undersampling technique, 1,000 trees and 6 variables per tree, predicts default with roughly 80% accuracy.

As mentioned briefly, an important observation is that the accuracy of random forest is stable and only marginally improving at 200 trees as well as being highly stable across number of splits per tree. This observation indicates that random forest is a robust technique which does not require the modeler to worry too much about whether the parameters should be fine-tuned. In addition, the difference in value between AA and BA is relatively small which is preferred when building classification models. In this case, when using undersampling with a roughly equal number of observations of each class in both the training and testing set, the only possible way to get a high OA is when both AA and BA are high. This is, however, not true in other cases where the testing sample is imbalanced.

Table 14: Logistic Regression Accuracy With Undersampling

	Full Model			LASSO Regularization		
Cut-off	AA	BA	OA	AA	BA	OA
0.1	0.071	0.992	0.540	0.749	0.685	0.716
0.2	0.268	0.965	0.622	0.790	0.644	0.715
0.3	0.428	0.916	0.677	0.817	0.610	0.712
0.4	0.576	0.836	0.702	0.857	0.556	0.712
0.5	0.734	0.717	0.726	0.886	0.513	0.706
0.6	0.857	0.556	0.713	0.905	0.466	0.694
0.7	0.928	0.399	0.658	0.915	0.427	0.667
0.8	0.975	0.238	0.600	0.930	0.391	0.656
0.9	0.991	0.117	0.546	0.941	0.350	0.640

Table 14 indicates the accuracy of the logistic regression model using the full model i.e. with all variables included as well as the LASSO regularization accuracy where only the most informative variables are kept. In addition, table 14 reports the sensitivity of the two models' accuracy across cut-off points. Since the logistic regression predicts a probability for each observation going bankrupt, the cut-off points state at which probability the model should classify and observation as bankrupt. For instance, 0.1 or 10% means that if an observation is predicted to have a probability of default of 10%, all observations >=10% will be classified as bankrupt and observations <10% as not defaulting. This sensitivity analysis might seem redundant as the AA accuracy will, by intuition, be much higher when the cut-off point is high with a corresponding low BA. Conversely, the BA accuracy will be high when the cut-off point is low etc. However, the reason why it is important to test this is the fact that the aforementioned intuition is not always that straightforward.

As seen, the AA and BA varies significantly across cut-off point for the full logistic model which confirms the intuition, but the same is not equally true for the LASSO regularization. First, however, zooming into each model separately, the cut-off point for the full logistic regression model reaches its maximum OA at 0.5 or 50% with a corresponding OA of 72.6%, an AA of 73.4% and BA of 71.7%. This result confirms the intuition, since a cut-off point of 0.5 will classify observations <50% as not defaulted and >=50% as defaulted. On the other hand, by looking at the LASSO regularization, this result is quite different. When shrinking the model to only consider the most significant variables, a cut-off point of 10% yields the best Overall Accuracy of 71.6%. The reason for why this might be the case is that when shrinking the model to obtain a simpler model, some information is kept out. This substantiates the claims that a sensitivity analysis is important even when the intuition suggests a simple picture.

There is therefore a tradeoff between the full logistic regression model and the simpler LASSO model. The full model is more complicated, but yields higher accuracy, where the LASSO model is simpler and easier to apply and interpret, but with a slight decrease in accuracy.

7.1.2 SMOTE

Table 15: Random Forest Accuracy With SMOTE (k=5)

	# of Trees (n)								
	200			500			1000		
# of splits (m)	AA	BA	OA	AA	BA	OA	AA	BA	OA
4	0.948	0.434	0.890	0.947	0.435	0.889	0.937	0.463	0.885
5	0.949	0.433	0.891	0.947	0.431	0.889	0.937	0.461	0.885
6	0.946	0.431	0.888	0.948	0.428	0.890	0.940	0.454	0.886
7	0.891	0.429	0.949	0.949	0.432	0.891	0.941	0.454	0.887
8	0.949	0.424	0.890	0.949	0.424	0.890	0.942	0.448	0.887

The logic of table 15 is similar to table 13, but using the SMOTE sampling technique. As it can be seen, the Overall Accuracy using SMOTE is higher than using undersampling. However, drawing a conclusion based on that result alone will be highly misleading. Since the test sample using SMOTE is much larger than undersampling, there will be a significant imbalance between active and bankrupt companies which ultimately results in a high Overall Accuracy when AA is high.

Interestingly, the BA is much lower across number of trees and variables per tree than using undersampling, approximately 45% against 79%. This is in direct contrast to what is expected due to the fact that the purpose of the SMOTE is exactly to synthetical increase the number of observations of the minority class (Bankrupt observations) in the training sample in order to build a better model. What is found, however, is that applying SMOTE does in fact not increase the accuracy of bankrupt prediction, but rather decreases it. The reason for this is not exactly clear, but most likely due to outliers/noise in the data which the SMOTE algorithm will further enhance. This issue was described in 6.3. Therefore, in conclusion, even though the Overall Accuracy is higher using SMOTE than undersampling, SMOTE is not found to be superior to undersampling, but rather inferior due the fact that the high OA comes at the expense of a low BA.

Table 16: Logistic Regression Accuracy With SMOTE (k=5)

	Full Model			LASSO Regularization		
Cut-off	AA	BA	OA	AA	BA	OA
0.1	0.156	0.974	0.245	0.837	0.550	0.805
0.2	0.356	0.924	0.418	0.862	0.509	0.824
0.3	0.535	0.852	0.570	0.881	0.470	0.836
0.4	0.692	0.748	0.698	0.899	0.438	0.848
0.5	0.812	0.588	0.787	0.913	0.400	0.857
0.6	0.899	0.439	0.849	0.928	0.369	0.867
0.7	0.951	0.310	0.881	0.940	0.345	0.875
0.8	0.981	0.186	0.894	0.950	0.316	0.881
0.9	0.993	0.110	0.896	0.958	0.286	0.885

Similarly, table 16 is built around the same logic as table 14, but applying SMOTE instead of undersampling. The general pattern for logistic regression is similar to what was evident from random forest using undersampling compared to SMOTE. Looking at the results of the full logistic regression model at a cut-off of 0.5 and above, a higher OA is found compared to using undersampling, but at the expense of BA. For instance, the OA of the full model at 0.5 is 78.7% using SMOTE compared to 72.6% using undersampling, but with a Bankruptcy Accuracy of 71.7% against 58.8%. The LASSO regularization results are more even across cut-off points for the SMOTE, but still at the expense of a low BA for the exact same reason as random forest i.e. the testing sample is imbalanced with a significant overrepresentation of active companies. Therefore, even though the Overall Accuracy is high, the model is in fact worse than applying undersampling.

7.1.3 Partial Conclusion

From the results presented above, it can be concluded, firstly, that applying the SMOTE sampling technique for the data does not provide any improvement in the models for neither random forest nor logistic regression. Instead, SMOTE increases the noise of the data which disrupts the purpose. The SMOTE sampling procedure will therefore not be applied for any of the coming analyses.

Regarding the actual hypothesis of which model is better at predicting bankruptcy, it can be concluded that random forest does perform better than logistic regression. Even though it can be argued that random forest only performed marginally better than logistic regression, a 4.7 percentage point difference is substantial. Therefore, it can be concluded that the non-parametric strength that random forest should have in theory, is evident which confirms hypothesis 1.

On the other hand, random forest is computationally significantly more complex than logistic regression which in turn means that running the random forest algorithm is time consuming. It is thus useful to know that logistic regression, despite the decrease in accuracy, can be applied with relatively satisfying results using a fraction of the time. The reason why this conclusion is important should be seen in a business application perspective. Applying machine learning to analyzing classification problems in real-time is done at an increasing frequency and knowing that "simpler" models performs reasonably well at a lower computation complexity is important.

7.2 Hypothesis 2

The thesis proceeds with the analysis by addressing the second hypothesis by investigating whether the addition of non-firm-specific variables are in fact useful for the prediction of bankruptcy. The following results are presented in a similar way to hypothesis 1. In addition, the results are still computed on the combined sample and using undersampling only. The results are shown using random forest and logistic regression in order to be more confident in the conclusion on whether the addition of non-firm-specific variables increase the accuracy or not.

Table 17: Random Forest Accuracy (Firm-Specific and All Variables)

	Firm-specific			Firm-specific + non-firm-specific		
# of splits (m)	AA	BA	OA	AA	BA	OA
4	0.755	0.736	0.745	0.753	0.791	0.772
5	0.753	0.740	0.746	0.750	0.794	0.771
6	0.752	0.734	0.743	0.750	0.796	0.773
7	0.750	0.736	0.743	0.750	0.793	0.771
8	0.750	0.732	0.741	0.751	0.793	0.771

Table 17 presents the results using 1,000 trees (n=1,000) as well as a sensitivity analysis across splits per tree (m= 4 to 8). The left-hand side of the table shows the accuracy of random forest using financial ratios only. The accuracy, i.e. AA, BA and OA, is very consistent across variables per split with an AA of approximately 75.2%, BA of 73.5% and OA of 74.3%. On the right-hand side is the accuracy of the Random forest algorithm using the whole predictor space i.e. financial ratios and macroeconomic variables. The AA is around 75%, BA of 79.5% and OA of 77.2%. The results indicate that the addition of non-firm-specific variables is in fact useful for the prediction of bankruptcy.

Table 18: Logistic Regression Accuracy (Firm-Specific and All Variables)

	Firm-specific			Firm-specific + non-firm-specific		
Cut-off	AA	BA	OA	AA	BA	OA
0.1	0.070	0.985	0.519	0.071	0.992	0.540
0.2	0.259	0.951	0.598	0.268	0.965	0.622
0.3	0.421	0.893	0.652	0.428	0.916	0.677
0.4	0.592	0.825	0.709	0.576	0.836	0.702
0.5	0.734	0.725	0.730	0.734	0.717	0.726
0.6	0.853	0.580	0.716	0.857	0.556	0.713
0.7	0.930	0.388	0.665	0.928	0.398	0.658
0.8	0.981	0.191	0.594	0.975	0.238	0.600
0.9	0.993	0.095	0.553	0.991	0.117	0.546

Looking at table 18, where the same analysis is conducted using logistic regression (full model), it is found that the peak Overall Accuracy is found at a cut-off point of 0.5. Interestingly, however, this indicates that there is no difference between the accuracy when performing the analysis on financial ratios alone and the full predictor space.

Logistic regression is therefore not able to extract any additional valuable information from the addition of the non-firm-specific variables. As found in 7.1, random forest performs better than logistic regression. In addition, the results from 7.2 indicate that random forest, unlike logistic regression, is in fact also able to extract information from less important variables with the result of a higher accuracy of 3 percentage points for OA and 4.2 percentage points for BA.

7.2.1 Partial Conclusion

We have shown that the addition of non-firm-specific variables increase the accuracy of default prediction using random forest. The same conclusion can’t be drawn from the logistic regression where the addition of non-firm-specific variables do not seem to have an effect. Furthermore, it can be concluded that the increase in accuracy for random forest is moderate. It is clear, that the variables that contain the most valuable information regarding bankruptcy prediction is to be found in the financial ratios. This result is not surprising, as the analysis is conducted on a firm level. It is, however, interesting that the addition of non-firm-specific variables is not impacting the accuracy for logistic regression. Despite the fact that a moderate increase in accuracy is found for random forest when including non-firm-specific variables, but not for logistic regression, leads to the conclusion that non-firm-specific variables do not always have an effect.

7.3 Hypothesis 3

To test this hypothesis, the accuracy of both models is reported on both the wholesale and manufacturing subsamples as well as the combined sample. The sample is split into accuracy of active companies (AA) which corresponds to companies that have not defaulted and accuracy of bankrupt companies (BA) as well as the overall accuracy of the model. First, each subsample will be addressed for both models and at last an analysis of the variables that drive corporate default will follow.

Table 19: Random Forest Accuracy (All Samples)

	Combined			Wholesale			Manufacturing		
# of splits (m)	AA	BA	OA	AA	BA	OA	AA	BA	OA
4	0.753	0.791	0.772	0.773	0.767	0.770	0.754	0.761	0.757
5	0.750	0.794	0.771	0.757	0.761	0.759	0.759	0.766	0.763
6	0.750	0.796	0.773	0.767	0.763	0.765	0.755	0.771	0.763
7	0.750	0.793	0.771	0.774	0.761	0.768	0.750	0.769	0.760
8	0.751	0.793	0.771	0.762	0.767	0.765	0.754	0.766	0.760

Table 19 shows the accuracy of the prediction using random forest with undersampling and n=1000 on the combined data and the two subsamples using various numbers of splits.

It is found that the overall accuracy of the combined sample is at its highest of 0.773 using 6 splits while it is noted that the variation in overall accuracy of the combined sample is very low with the lowest accuracy being 0.771 for both 5, 7, and 8 splits. The highest accuracy of bankruptcy prediction is also at 6 numbers of splits with an accuracy of 0.796 while the highest accuracy of a non-default event is at 4 splits at 0.753.

Moving onto the subsamples, in the wholesale subsample, it is found that the highest overall accuracy is at 7 splits with a value of 0.768. The highest accuracy of bankruptcy prediction is found at the extremes of 4 and 8 splits both with a value of 0.767. In the manufacturing sample, the highest overall accuracy is found at both 5 and 6 splits both with a value of 0.763. The highest accuracy of bankruptcy prediction is found at 6 splits with a value of 0.771 whereas the highest accuracy of the active data is found 5 splits with a value of 0.759. There is found very little difference in overall accuracy with a maximum value of 0.773 for the combined sample and 0.768 and 0.763 for wholesale and manufacturing respectively which actually means that the combined sample has higher overall accuracy than splitting the data, however, with a very minor difference.

In conclusion, there is found very little difference in accuracy when subsampling the manufacturing and wholesale industry both internally and relative to the combined sample as a whole when using random forest. This leads to the conclusion that no superior insight is gained from a standpoint of accuracy from subdividing the industries of wholesale trade and manufacturing in predicting default using random forest.

Table 20: Logistic Regression Accuracy (All Samples)

	Combined			Wholesale			Manufacturing		
Cut-off	AA	BA	OA	AA	BA	OA	AA	BA	OA
0.1	0.071	0.992	0.540	0.038	0.998	0.505	0.152	0.973	0.556
0.2	0.268	0.965	0.622	0.160	0.967	0.552	0.316	0.945	0.625
0.3	0.428	0.916	0.677	0.345	0.917	0.623	0.464	0.907	0.682
0.4	0.576	0.836	0.702	0.531	0.817	0.679	0.598	0.834	0.713
0.5	0.734	0.717	0.726	0.710	0.690	0.700	0.706	0.738	0.721
0.6	0.857	0.556	0.713	0.843	0.529	0.681	0.837	0.605	0.724
0.7	0.928	0.398	0.658	0.915	0.359	0.645	0.928	0.440	0.688
0.8	0.975	0.238	0.600	0.963	0.204	0.594	0.980	0.257	0.624
0.9	0.991	0.117	0.546	0.991	0.083	0.550	0.995	0.132	0.570

Table 20 shows the accuracy of the prediction using logistic regression with undersampling on the combined data and the two subsamples using various cut-off points.

It is found that the highest accuracy of the combined data is found at a cut-off point of 0.5 yielding 0.726 in overall accuracy. As expected, the accuracy of the bankruptcy prediction is falling with the cut-off value while the accuracy of the non-bankrupt prediction is rising with the cut-off value as a clear result of lumping more or less observations into either category. This is the case for both the combined sample as well as both subsamples.

Moving onto the subsamples, the highest accuracy of the wholesale subsample is found at a cut-off point of 0.5 with a value of 0.700. The highest accuracy of the manufacturing sample is found at a cut-off point of 0.6 with a value of 0.724. It is found that the overall accuracy is highest for the combined sample relative to both subsamples. It is worth noting that the accuracy of bankruptcy prediction also is highest for the combined sample when using the cut-off point that yields the highest overall accuracy for each individual sample. When doing so it is found that the combined sample has an accuracy of 0.717, the wholesale sample has an accuracy of 0.690, and the manufacturing sample has an accuracy of 0.605 when predicting default using the cut-off point that yields the highest total accuracy.

Thus, it is found that no additional accuracy is gained from subdividing the wholesale and manufacturing industries when predicting corporate default. In fact, the combined sample had higher overall accuracy as well as higher accuracy for predicting default at this cut-off point, although the difference is considered minor. This leads to the conclusion that no superior insight is gained from a standpoint of accuracy from subdividing the industries of whole sale and manufacturing in predicting default using logistic regression.

7.3.1 Variable Importance

In this section, the variables that drive corporate default will be presented. First, the random forest model will be addressed. The variable importance will be listed for the combined sample as well as for both the wholesale trade and manufacturing subsamples. The variables are ranked by their importance measured by the mean decrease in gini. Second, the logistic regression model will be addressed. The significant variables will be listed with and without lasso regularization for the combined sample as well as the wholesale trade and manufacturing subsamples. An attempt will be made to compare the important variables across industries for both models in order to determine whether the variables that drive corporate default are similar or different. This will show whether the preliminary conclusions of 7.3 are supported.

Random Forest - Combined Sample

Figure: 21 Variable Importance - Combined Sample - n=1,000, m=6

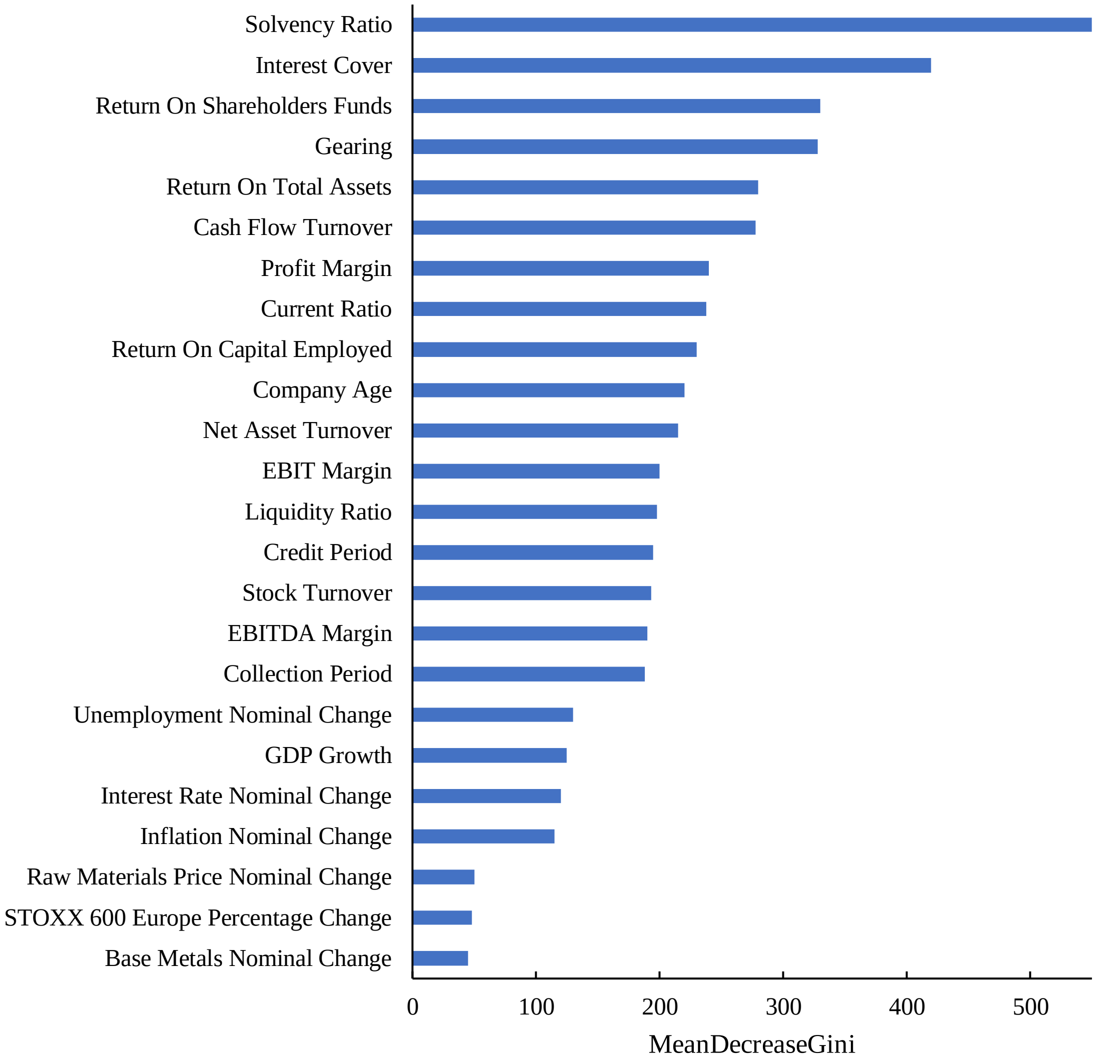


Figure 21 shows variable importance using random forest on the combined sample. The variables are ranked by the measure mean decrease in gini (MDG). As described in section 6.2.2, the gini index is a measure of the purity of the nodes in a classification tree. A decrease in gini from one split to another indicates that the chosen variable and the connected rule purifies the split. Therefore, MDG measures which variables decrease the gini index most across all trees in random forest. The actual number on the x-axis in the plot is arbitrary in the sense that it will change depending on the number of explanatory variables in the sample. One can therefore not say that a MDG of 500 is good. The relative difference between the MDG of the variables is, however, not arbitrary. As seen from figure 21 the most important variable as measured using MDG is the solvency ratio followed by the interest coverage ratio, return on shareholders' funds, gearing, return on total assets, and cash flow turnover. It is noteworthy that the solvency ratio which is the most important variable and gearing which is the fourth most important variable both are measures of roughly the same thing – the capital structure of the company. Whereas solvency ratio measures the fraction of the total assets financed by equity, gearing measures the total debt as a fraction of the equity. It is also worth noting that two other relatively similar ratios are included amongst the top 6 most important variables – return on shareholders' funds and return on total assets. Both profitability measures but are differentiated by the amount of leverage and the cost of debt.

Random Forest - Wholesale Trade Sample

Figure 22: Variable Importance - Wholesale Trade Sample - n=1,000, m=6

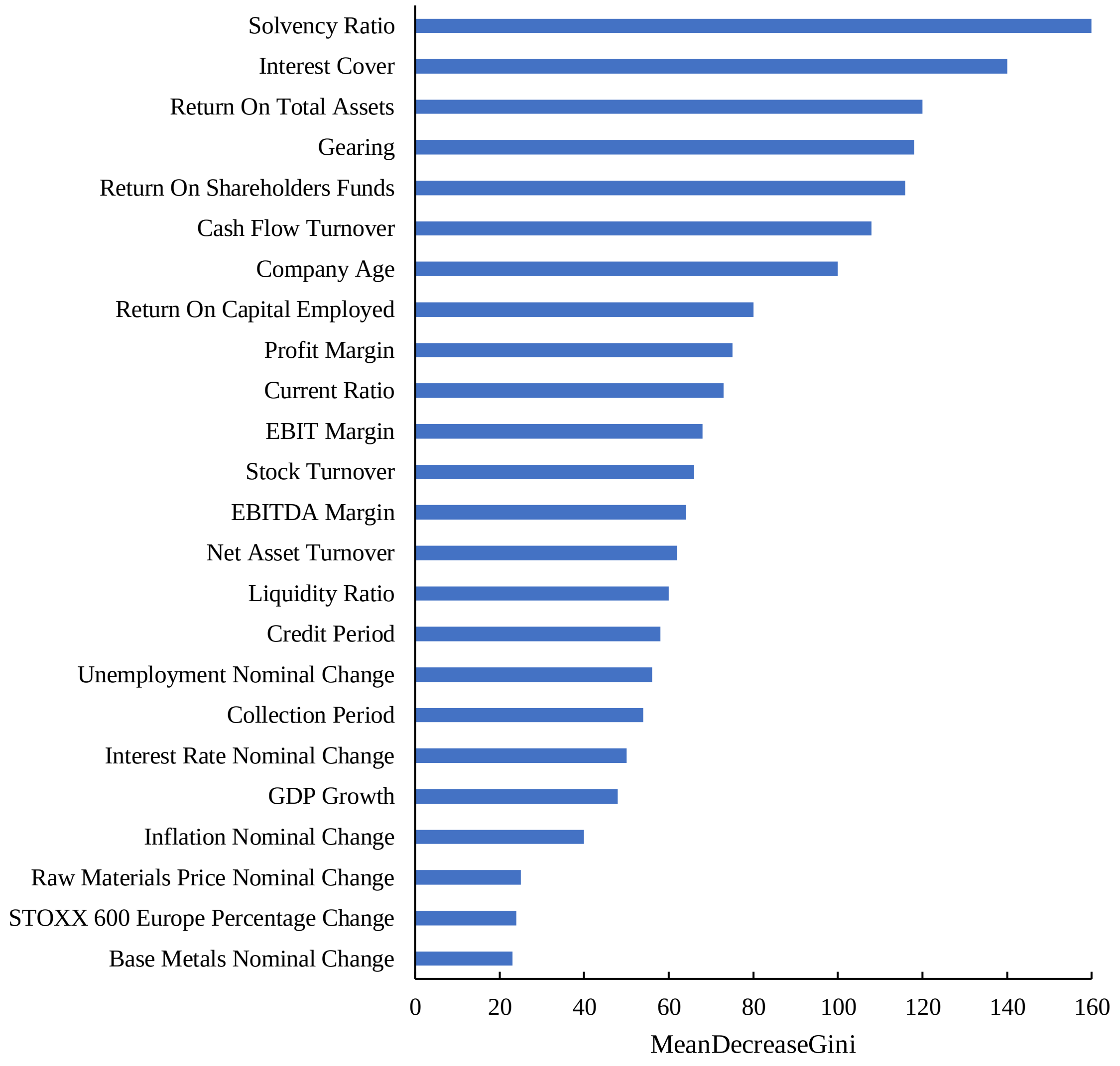


Figure 22 shows variable importance using random forest on the wholesale trade sample. As seen from figure 22 the most important variable as measured using MDG is the solvency ratio followed by the interest coverage ratio, return on total assets, gearing, return on shareholders' funds, and cash flow turnover. In this regard, the wholesale sample is not much different the combined sample with the only difference being that return on total assets is 5th and not 3rd and vice versa for return on shareholders' funds. In other words, the exact same 6 variables are the most important when measured by MDG.

Random Forest – Manufacturing Sample

Figure 23: Variable Importance - Manufacturing Sample - n=1,000, m=6

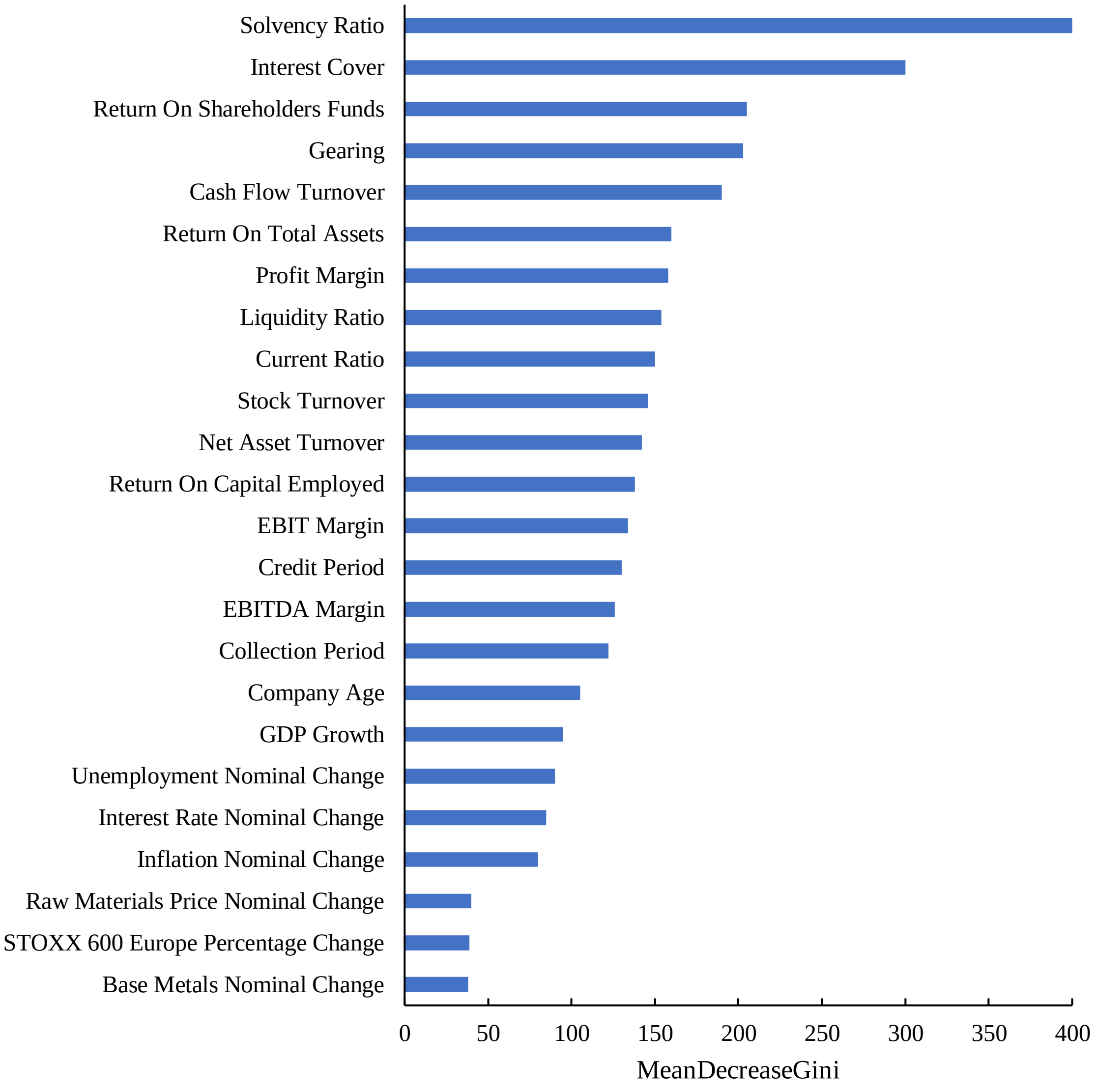


Figure 23 shows variable importance using random forest on the manufacturing sample. As seen from figure 23 the most important variable as measured MDG is the solvency ratio followed by the interest coverage ratio, return on shareholders' funds, gearing, cash flow turnover, and return on total assets. Again, the same six ratios are the most important as measured by MDG. For the manufacturing sample, cash flow turnover has gained relative importance, while amongst the two profitability measures return on shareholders fund is again the most important.

Thus, it is concluded that the most important variables listed with relative importance include solvency ratios, profitability ratios and leverage ratios. There is no difference in the variables included for the various subsamples and even only minor difference in sequence between the variables. Therefore, it is concluded that when using random forest on the subsamples manufacturing and wholesale trade, default prediction is not conditional on industry.

Logistic Regression - Combined Sample

Table 21: Logistic Regression Coefficients (Combined Sample)

	Full Logistic Regression Model					**LASSO**
	Estimate	Std. Error	z value	Pr(>|z|)		
(Intercept)	2.2E-01	9.9E-02	2.2E+00	2.8E-02	*	0.522
Return_on_shareholders_funds	-2.0E-01	7.0E-02	-2.9E+00	4.3E-03	**	-0.081
Return_on_capital_employed	5.6E-02	1.1E-01	5.0E-01	6.2E-01		.
Return_on_total_assets	2.4E-01	5.4E-01	4.5E-01	6.5E-01		.
Cash_flow_Turnover	-8.5E-03	6.7E-01	-1.3E-02	9.9E-01		-1.386
Profit_margin	-1.4E+01	1.5E+00	-9.4E+00	2.0E-16	***	-4.538
EBITDA_Margin	-3.1E+00	9.5E-01	-3.3E+00	1.1E-03	**	.
EBIT_Margin	1.0E+01	1.7E+00	6.0E+00	1.6E-09	***	.
Net_assets_turnover	3.8E-03	1.8E-03	2.1E+00	3.2E-02	*	.
Interest_cover	1.0E-03	2.8E-04	3.5E+00	4.1E-04	***	.
Stock_turnover	2.1E-05	2.9E-04	7.2E-02	9.4E-01		.
Collection_period	5.9E-04	3.4E-04	1.8E+00	7.8E-02	.	0.000
Credit_period	2.0E-03	4.4E-04	4.6E+00	4.1E-06	***	0.001
Current_ratio	8.0E-02	3.9E-02	2.1E+00	3.9E-02	*	.
Liquidity_ratio	-5.0E-02	4.5E-02	-1.1E+00	2.8E-01		.
Solvency_ratio	-3.0E+00	2.0E-01	-1.5E+01	2.0E-16	***	-2.960
Gearing	1.2E-01	1.4E-02	8.5E+00	2.0E-16	***	0.095
Company_Age	-6.0E-03	1.6E-03	-3.8E+00	1.7E-04	***	-0.001
GDP_Growth	1.5E+01	2.3E+00	6.4E+00	1.7E-10	***	1.581
Interest_Rate_Nominal_Change	-4.4E+00	3.4E+00	-1.3E+00	2.0E-01		.
Inflation_Nominal_Change	4.8E+00	4.1E+00	1.2E+00	2.4E-01		.

Unemployment_Nominal_Change	1.1E+01	3.0E+00	3.7E+00	2.5E-04	***	.
STOXX_600_Europe_Percentage_Change	5.0E-01	2.9E-01	1.7E+00	9.0E-02	.	.
Raw_Materials_Price_Nominal_Change	6.3E-03	5.0E-03	1.3E+00	2.0E-01		.
Base_Metals_Nominal_Change	-8.7E-03	3.7E-03	-2.4E+00	1.9E-02	*	.

***' = 0.001, '**' = 0.01, '*' = 0.1, '.' = 0.1

Table 21 shows the coefficients of logistic regression on the combined sample. As shown in table 21, 14 variables are significant initially before using LASSO regularization. After LASSO regularization 9 variables are significant. They include return of shareholders' funds, cash flow turnover, profit margin, collection period, credit period, solvency ratio, gearing, company age and GDP growth. Since collection period, credit period and company age all have estimates equal to or close to zero they are less relevant. Cash flow turnover, profit margin and solvency ratio have the largest negative values which means that they are the most negatively correlated with default. On the other hand, GDP growth was positively correlated with corporate default indicating that the higher the growth in the country of domicile, the larger the probability of default. The basis for this relation is unknown. One reason could be the fact that one country within the sample had higher GDP growth in the estimation period while also having higher rates of default for unrelated reasons. Thus, it is not expected that the relation is causal but a data issue. Also, gearing was positively correlated with default indicating that higher leverage is correlated with corporate default.

Logistic Regression - Wholesale Trade Sample

Table 22 shows the coefficients of logistic regression on the wholesale trade subsample. As shown in table 22, 12 variables are significant initially before using LASSO regularization. After LASSO regularization 12 variables are significant. Since 5 variables have values close to zero, they are less relevant. The most relevant include return on shareholders' funds, return on total assets, cash flow turnover, profit margin, solvency ratio, gearing, and company age. Again, cash flow turnover, profit margin and solvency ratio have the largest negative values indicating that these three variables are the most negatively correlated with corporate default among wholesale trade companies. For the wholesale trade subsample, return on total assets is positively correlated with default which strikes against expectations. This indicates that the more profitable the assets are the higher the probability of default, which is difficult to explain. Gearing was also positively correlated with default indicating that higher leverage increases the probability of corporate default which is in line with expectations.

Table 22: Logistic Regression Coefficients (Wholesale Trade Sample)

	Full Logistic Regression Model					LASSO
	Estimate	Std. Error	z value	Pr(>|z|)		
(Intercept)	-2.1E-02	1.7E-01	-1.3E-01	9.0E-01		0.427
Return_on_shareholders_funds	-4.3E-01	1.3E-01	-3.4E+00	6.6E-04	***	-0.080
Return_on_capital_employed	1.4E-01	1.8E-01	7.9E-01	4.3E-01		.
Return_on_total_assets	4.2E+00	7.7E-01	5.4E+00	5.7E-08	***	0.116
Cash_flow_Turnover	-5.4E+00	2.0E+00	-2.8E+00	5.6E-03	**	-2.566
Profit_margin	-2.1E+01	3.4E+00	-6.2E+00	6.6E-10	***	-3.890
EBITDA_Margin	1.3E+00	3.0E+00	4.3E-01	6.7E-01		.
EBIT_Margin	1.3E+01	4.4E+00	3.0E+00	2.7E-03	**	.
Net_assets_turnover	4.9E-03	2.3E-03	2.1E+00	3.8E-02	*	0.002
Interest_cover	-8.6E-05	5.0E-04	-1.7E-01	8.6E-01		.
Stock_turnover	5.2E-04	4.3E-04	1.2E+00	2.2E-01		.
Collection_period	3.6E-04	5.9E-04	6.1E-01	5.4E-01		0.000
Credit_period	2.0E-03	6.8E-04	2.9E+00	3.9E-03	**	0.001
Current_ratio	9.6E-02	6.9E-02	1.4E+00	1.6E-01		0.033
Liquidity_ratio	-2.9E-02	8.0E-02	-3.6E-01	7.2E-01		.
Solvency_ratio	-2.4E+00	3.3E-01	-7.5E+00	8.0E-14	***	-2.059
Gearing	1.5E-01	2.4E-02	6.2E+00	5.4E-10	***	0.143
Company_Age	-1.5E-02	3.0E-03	-4.8E+00	1.3E-06	***	-0.011
GDP_Growth	1.5E+01	3.8E+00	4.0E+00	5.7E-05	***	.
Interest_Rate_Nominal_Change	-1.0E+01	5.9E+00	-1.7E+00	8.9E-02	.	.
Inflation_Nominal_Change	4.8E+00	6.8E+00	7.1E-01	4.8E-01		.
Unemployment_Nominal_Change	2.0E+01	5.1E+00	3.8E+00	1.2E-04	***	.
STOXX_600_Europe_Percentage_Change	5.1E-01	5.0E-01	1.0E+00	3.1E-01		.
Raw_Materials_Price_Nominal_Change	1.4E-02	8.6E-03	1.7E+00	9.6E-02	.	.
Base_Metals_Nominal_Change	-6.7E-03	6.3E-03	-1.1E+00	2.8E-01		0.001

***' = 0.001, '**' = 0.01, '*' = 0.1, '.' = 0.1

Logistic Regression – Manufacturing Sample

Table 23: Logistic Regression Coefficients (Manufacturing Sample)

	Full Logistic Regression Model					**LASSO**
	Estimate	Std. Error	z value	Pr(>|z|)		
(Intercept)	5.3E-01	1.3E-01	4.0E+00	5.5E-05	***	0.672
Return_on_shareholders_funds	-1.9E-01	9.7E-02	-1.9E+00	5.2E-02	.	.
Return_on_capital_employed	2.6E-01	1.4E-01	1.8E+00	6.4E-02	.	.
Return_on_total_assets	-3.1E+00	7.8E-01	-4.0E+00	7.5E-05	***	-2.867
Cash_flow_Turnover	-1.6E-01	7.6E-01	-2.0E-01	8.4E-01		-0.992
Profit_margin	-1.1E+01	1.7E+00	-6.6E+00	3.1E-11	***	-2.454
EBITDA_Margin	-3.5E+00	1.1E+00	-3.2E+00	1.2E-03	**	.
EBIT_Margin	9.6E+00	2.0E+00	4.8E+00	1.2E-06	***	.
Net_assets_turnover	9.2E-04	2.4E-03	3.8E-01	7.1E-01		.
Interest_cover	1.8E-03	3.8E-04	4.8E+00	1.9E-06	***	.
Stock_turnover	-6.0E-04	4.6E-04	-1.3E+00	2.0E-01		.
Collection_period	1.3E-03	4.2E-04	3.0E+00	2.5E-03	**	.
Credit_period	2.0E-03	6.1E-04	3.3E+00	1.2E-03	**	0.001
Current_ratio	7.0E-02	4.9E-02	1.4E+00	1.6E-01		.
Liquidity_ratio	-2.5E-01	7.8E-02	-3.2E+00	1.5E-03	**	-0.005
Solvency_ratio	-3.4E+00	2.8E-01	-1.2E+01	2.0E-16	***	-3.237
Gearing	9.6E-02	1.9E-02	5.1E+00	3.4E-07	***	0.067
Company_Age	-2.1E-03	1.9E-03	-1.1E+00	2.9E-01		.
GDP_Growth	1.2E+01	2.9E+00	4.0E+00	5.4E-05	***	.
Interest_Rate_Nominal_Change	-4.7E+00	4.4E+00	-1.1E+00	2.9E-01		.
Inflation_Nominal_Change	8.4E+00	5.3E+00	1.6E+00	1.1E-01		.
Unemployment_Nominal_Change	2.9E+00	3.8E+00	7.6E-01	4.5E-01		.
STOXX_600_Europe_Percentage_Change	2.9E-01	3.7E-01	7.9E-01	4.3E-01		.
Raw_Materials_Price_Nominal_Change	-8.3E-03	6.3E-03	-1.3E+00	1.9E-01		.
Base_Metals_Nominal_Change	-4.0E-03	4.8E-03	-8.3E-01	4.0E-01		.

***' = 0.001, '**' = 0.01, '*' = 0.1, '.' = 0.1

Table 23 shows the coefficients of logistic regression on the manufacturing sub sample. As shown in table 23, 12 variables are significant initially before using LASSO regularization. After LASSO regularization 8 variables are significant of which four have high negative values indicating negative correlation with default. Again, cash flow turnover, profit margin and solvency ratio have high negative values indicating a negative relation with default. Unlike for the wholesale trade sample, return on total assets has a high negative value indicating a strong negative relation to corporate default.

In conclusion, the most important variables when predicting default using logistic regression include cash flow turnover, profit margin and solvency ratio. This is common among all three samples. There is however, found differences conditional on industry where return on total assets is highly negatively correlated with default among manufacturing companies, positively correlated with default among wholesale trade companies and insignificant for the combined sample.

7.3.2 Partial Conclusion

The general differences between the industries are small in term of the accuracy of the models both when using random forest and logistic regression. The variables that are important between the models are almost identical across industries. This is especially true for random forest. For logistic regression the important variables are also mostly the same between industries, but with slight variation. On the other hand, the important variables found are different when comparing random forest and logistic regression. For instance, interest coverage is not found to be significant for logistic regression after applying the LASSO. In general, it is found that default prediction is not conditional on industry in terms of accuracy. In addition, when looking at random forest and logistic regression in isolation, the important variables are nearly identical indicating that corporate default is not conditional on industry.

8. Discussion

The results from section 7 have primarily been presented on a descriptive basis with limited focus on why the results are interesting in a broader perspective. In this chapter of the thesis, the emphasize and focus of attention is on the bigger picture of the relevance of the results. Concretely, the interpretation of the results, how the results relate to the literature and the limitations of the results are discussed.

For instance, what is the implication of the fact that random forest is better at predicting default than logistic regression? What is the significance of the fact that non-firm-specific variables seem to be redundant in some cases? What is the implication of the fact that the important drivers of corporate default seem to be unconditional on industry? How does all these results coincide with the academic literature? Is it possible to broadly conclude anything from the results, or are there limitations to the results which require further research? These are some of the questions that this chapter of the thesis will focus on.

8.1 Hypothesis 1

8.1.1 Interpretation

The results found from investigating the first hypothesis is that random forest is more accurate that logistic regression when predicting corporate default. In addition, it is found that applying SMOTE actually decreased the BA of both models compared to undersampling. In order to address the implications of these results, it is useful to first discuss the advantages and disadvantages of both models. By doing this, it is possible to relate the implications of the results to appropriate model application. In other words, in which cases are either model more appropriate given the results presented in the previous section?

Logistic regression is a parametric model which has some desirable properties that are different from random forest. First of all, logistic regression analyzes a problem using a finite set of parameters which makes it possible to provide an actual equation of the relationship between each explanatory variable and the impact on the probability that an observation belongs to a class. The fact that it is possible to investigate the exact relationship between variables and the probability makes interpretation easy. In some problems, the interpretation of the variables is more important than gaining a few percentage points of accuracy using sophisticated classification algorithms which are difficult to interpret. However, there is always a tradeoff between accuracy and interpretability which is highly specific to the problem. The fact the accuracy of logistic regression is not substantially lower than for random forest is a useful result. This means that it is indeed possible to apply logistic regression for problems where the interpretation of the variables is important without losing a substantial amount of accuracy. On the other hand, for problems where accuracy is highly important, random forest is preferred.

Apart from the fact that logistic regression gives a clear equation of the relationship between the variables and the probability of an observation belonging to a class, logistic regression provides other statistical output which is otherwise lost when using random forest. The statistical output of logistic regression contains the significance of each explanatory variable as well as standard errors and confidence intervals. These statistical properties provide an in-depth understanding of the variables and how they interact which is yet another important factor when interpretability of the variables is more important than the accuracy.

In addition to the raw statistical output from the standard logistic regression, it is possible to utilize several regularization extensions such as the LASSO. The LASSO, as seen, is a way to simplify the model which in addition to making the model easy to apply also gives some important insights regarding the relationship between the explanatory variables and the probability assigned to an observation. As it has been shown, the results from a full logistic regression model can be significantly different to a regularized model. When there is a significant difference, it indicates the effect of having correlated variables in the data set and how this affects the explanatory variables' impact to the outcome. This property is highly desired when interpretability is important.

Even though logistic regression is preferred when the interpretation of the variables is more important than gaining a few percentage points in accuracy, there are also drawbacks which one need to be aware of. Logistic regression relies on several assumptions such a no multicollinearity between the variables and a linear relationship between the log odds and the explanatory variables. These assumptions can in some cases be violated especially in a data-heavy world where a model is fed with hundreds if not thousands of variables. Despite the fact that it is possible to check whether these assumptions are violated or not, it is most likely impractical for problems of substantial size. Furthermore, logistic regression is not built to tackle categorical variables well.

The issues stated above can be considerable when looking at the problem from a business point of view. For instance, a bank is inarguably interested in investigating their credit exposure by looking at the probability of bankruptcy of their lenders. In such cases, the bank wants to use a model which is able to handle all sorts of data types and relationships between the variables in order to predict the risk of bankruptcy of their lenders as accurately as possible. In these cases, "black box" classification algorithms where the interpretation of the variables is less important than the accuracy is preferred. This is where the implication of the results regarding random forest is important.

In addition to the fact that random forest was more accurate than logistic regression, the fact that the algorithm can handle all sorts of data types is desired for "black box" purposes. For "black box" problems where the variable importance is less of a concern, random forest is preferred to logistic regression. Even though random forest is significantly more complex than logistic regression and thus requires more computation time, the fact that random forest has proven to be robust and able to handle large amounts of data of all sorts makes it a perfect candidate for specific situations.

Apart from looking at the results regarding the performance of logistic regression and random forest in isolation, the surprising result of the SMOTE must be addressed. It is evident from the results that applying SMOTE as a sampling procedure, despite its theoretically desirable properties, did not increase the accuracy of the bankruptcy prediction of either model, but in fact it decreased them. This result was surprising since the fundamental idea behind SMOTE is to exactly circumvent the problem of having to discard valuable information due to the natural imbalance of the classes with the goal to have more data to train the models. The general implication of this result might seem straight forward; do not use SMOTE. However, making such general conclusion would be unsophisticated. Even though it is found that SMOTE did not add any valuable information in the problem other than increasing the complexity substantially, the issue is potentially specific to the data used in the thesis. This will be discussed further in 8.1.3.

In short, the overall implication of the result when testing the difference in accuracy between random forest and logistic regression is that logistic regression can be used for cases where the variable understanding is important without losing a considerable amount of accuracy. However, in cases where the variable importance is not important, but the accuracy of the model is, random forest is preferred. Furthermore, it can be concluded that applying SMOTE should be done with caution as there is no guarantee that the performance of the models will improve or suffer from a sample which has been treated by the SMOTE algorithm.

8.1.2 Academic Literature

As touched upon in the literature review, an extensive amount of research has been conducted on predicting corporate defaults. However, narrowing the scope of the overall literature to the relevant papers which has investigated problems which closely resemble the ones raised in this thesis, it is found that the results are mostly in line with previous research. For instance, Lin & Mcclean (2001) analyzed corporate default prediction for 1133 UK companies between 1980-1999 using financial ratios and found that decision tree-based models outperform logistic regression by 2.6 percentage points. Wagenmans (2017) investigated the difference in accuracy between random forest and logistic regression in bankruptcy prediction for 97,671 companies. Wagenmans performed his analysis on payment behavioral data which is a different set of explanatory variables compared the ones used in this thesis, but he found similar results. Wagenmans found that random forest outperforms logistic regression for different time until default horizons by an average of approximately 2 percentage points. Barboza, Kimura and Altman (2017) found similar results from their research of 449 bankrupt and 13,300 solvent US & Canadian companies using financial ratio data. They found that random forest outperforms logistic regression by around 9 percentage points. On the contrary to previous research, Figini, Savona and Vezzoli (2016) find that logistic regression outperforms random forest marginally when predicting corporate default on 742 German SME companies using financial data.

Apart from previous research focusing on the accuracy between random forest and logistic regression, research taking into account the sampling technique and how the results vary across these different types of sampling methods is limited for default prediction. There has been plenty of research on the effects of SMOTE in retail loan predictions, fraud detections and other applications, but few investigate the effect of SMOTE in corporate default prediction. One paper, however, investigates the effect of different sampling techniques for random forest and other machine learning algorithms on corporate default prediction for polish firms. Almayyan and Almayyan (2018) presented a paper on default prediction of polish firms using a vast collection of financial ratios. The paper investigated the accuracy of different machine learning models including random forest for different forecasting periods with and without SMOTE. From the paper, it was evident that the results suffered from undergoing the SMOTE algorithm for every forecasting period. The paper did, however, not specify where the decrease in accuracy stemmed from. In other words, whether it was due to a decrease in bankruptcy prediction, active prediction or both. The result, however, resembles the surprising finding that SMOTE can potentially be a harmful technique with respect to accuracy.

Overall, the mentioned literature shows results which are equivalent to the results found in this thesis. The general consensus is that non-parametric models such as random forest outperforms logistic regression as a parametric model when predicting corporate default. In addition, however scarce, the literature around the application of SMOTE as a sampling technique in bankruptcy prediction finds similar results to the ones found in this thesis.

8.1.3 Limitations

Even though the results seem to be confirming previous literature, it is important to be aware of the limitations of the research conducted in this thesis in order to not confidently state conclusions which might not be appropriate. Since all the research presented previously is conducted on different samples, from different periods, different countries and using a different set of explanatory variables, the conclusions that have been reached might not be perfectly equivalent to the ones found in this thesis. Therefore, it is important to mention the potential limitations of the research conducted in this thesis in relation to what is possible to conclude. This section will focus on this.

Since this thesis is only focusing on the performance of random forest and logistic regression and that the results show that random forest outperforms logistic regression, it cannot be stated whether this conclusion applies on a general level between parametric and non-parametric models. In other words, despite the fact that random forest is a non-parametric model and logistic regression is a parametric mode, it cannot be concluded that non-parametric models in general outperforms traditional parametric models. The underlying approach of the vast amount of non-parametric models suggest that such conclusion would be naive. In relation, parametric models can be highly suited for some problems where the data obey the underlying assumptions made from the various standard parametric models. The model choice and the expected performance of parametric and non-parametric models is highly problem specific, and it is important to emphasize that when performing classification analyses, the understanding of the data and how the variables interact has to be the first point of attention before choosing the model. In connection, as mentioned, the purpose of the analysis might also vary where for some problems variable importance is more important than accuracy.

In addition, as previous research has shown, it cannot be stated that random forest will always outperform logistic regression. There is, however, overwhelming evidence that this is generally the case, but as stated, each research paper in based on a different set of variables, samples etc. Furthermore, it has been the case in some research that logistic regression is in fact outperforming random forest even though this result seems to be rare.

Regarding the results using the SMOTE sampling technique, the lack of performance could be specific to the data used in this thesis. In other words, the data may contain outliers which, as shown in the 6.3, can further enhance outliers in the data. In other cases, the problem of outliers might not be an issue and in such cases, the SMOTE sampling technique will most likely prove useful. In addition, the first generation of SMOTE was used in this thesis. There have been several extensions to the algorithm which tries to circumvent the issue of outliers. For instance, borderline-SMOTE, Density-based SMOTE, Relocating Safe-level SMOTE, and Safe-level SMOTE are all extensions to the original SMOTE algorithm which tries to deal with the problem of outliers. The four mentioned extensions have different approaches, but the fundamental idea is to detect outliers and based on those detections, generate synthetic minority observations in areas where the algorithm believes the observations should be put to be more representative.

The reason for mentioning these extensions is that it cannot be ruled out that the idea behind SMOTE is useful for corporate default prediction in general. Some of the extensions might potentially circumvent the issue and provide great results. It can, however, be said that the general SMOTE algorithm is not useful for the data used in this thesis. As concluded in the 8.1.1, it must be emphasized that using the SMOTE algorithm and its extensions should be done with caution.

8.2 Hypothesis 2

8.2.1 Interpretation

When looking at the results found when testing the second hypothesis, it was evident that the financial ratios contained the overwhelming predictive power and that non-firm-specific variables added little to no information. For random forest, the non-firm-specific explanatory variables did yield a slight increase in accuracy, but for logistic regression, there was no effect. Despite the fact that these results seem relatively easy to comprehend, the implications of the results found in 7.2, and why this result might have occurred will be discussed.

By looking at the implications of the results from an intuitive point of view, the results could indicate that non-firm-specific variables such as macroeconomic and commodity price variables are already incorporated in the financial ratios. It is sensible that variables which are outside the control of a company will be summarized indirectly through the financial ratios. For instance, if GDP growth is decreasing or simply low, demand for products will, everything else equal, be low relative to a world where GDP growth is high. This effect will show up in the financials of companies especially in profitability ratios. Therefore, GDP growth in itself might not explain anything since the true effect on a firm-level is incorporated in the profitability ratios. Similarly, if the general prices of commodities are increasing, manufacturing companies, which buys commodities to build products, could be hurt. However, some companies will be better at mitigating this effect and therefore the commodity price index may not yield any additional information since the real effect is seen in the financial ratios of each company. The bottom line is that it is reasonable to think that non-firm-specific variables in and of itself might be telling for the number of bankruptcies in an economy, but at a firm-specific level where firm-specific variables are included, this effect is better portrayed in the financial ratios which ultimately makes the non-firm-specific variables unnecessary.

On the other hand, it was found that the addition of non-firm-specific variables did increase accuracy for random forest. This indicates that random forest as a non-parametric model is able to extract information for non-firm-specific data that logistic regression was not able to. This leads to the believe that non-firm-specific variables might not be completely redundant, but further research is needed to determine this claim conclusively. Since logistic regression is unable to extract additional information, but random forest is, it is unclear, at this point, whether a true relationship actually exists.

Next, by looking at the implications of the result from an academic standpoint, the result could indicate that future research efforts should not only be put into expanding the predictors space towards non-firm-specific variables, but also focus should be pointed towards a deeper investigation of firm-specific financial data. In the research conducted in this thesis 16 different financial variables were used, but the list could be expanded substantially. For instance, it could be interesting to not only look at the nominal values of the ratios, but instead the changes in ratios over time. There could potentially be significant information to be gained by looking at how ratios for companies develop over a given time horizon instead of looking at ratios from a static point of view. Furthermore, the predictor space could be extended to incorporate other firm-specific data such as payment data, managerial data or other untraditional firm-specific variables. This will be discussed further in section 10.2.

By looking at the implications of the result from a business point of view, the result indicates that businesses, for instance banks, which find the problem of predicting bankruptcy important, do not need to worry about adding non-firm specific variables to their classification analysis. Instead, businesses should focus on predicting bankruptcy using firm-specific variables that they might already have in their systems. This can potentially be an important implication since it can be difficult to extend the predictor space with variables that have to be found from outside data sources.

In general, it can be concluded from the results that financial ratios are the predominate driver of bankruptcy and that additional information gained from non-firm-specific variables is minimal. To substantiate this claim, the effect of adding non-firm-specific variables yield no information using logistic regression while additional accuracy was found using random forest. The implications of this result from an academic point of view is that there should be more focus on firm-specific variables and less on non-firm-specific variables. From a business point of view, the implication of the result is that businesses, that find the problem of predicting default interesting, should not broaden their predictor space with non-firm-specific data. Lastly, the reason why the financial ratios are the overwhelming driver of default is most likely due to the fact that outside information such as GDP growth, inflation, commodity prices etc. are already incorporated in the financial ratios.

8.2.2 Academic Literature

The vast majority of previous research on corporate default prediction has focused on accounting information in terms of financial ratios. The founding fathers of corporate default prediction, Beaver (1966), Altman (1968) and Ohlson (1980) all used financial ratios. Since then, corporate default prediction has evolved to not only include financial ratios, but also market variables, industry variables and macroeconomic indicators. Despite the fact that other variables than financial ratios have been tested, there is no clear conclusion whether they yield any additional information.

There is evidence that market variables in the form of stock returns, stock volatility, market to book ratios and earning per share ratios have significant correlation to bankruptcy. However, since the research is conducted on private companies, market variables are not obtainable and thus redundant to the discussion.

Regarding the effect of macroeconomic variables, Duffie, Saita & Wang (2007) report a significant relationship between the state of the economy and default hazard rates of individual firms. Carling et al. (2007) also report a significant relationship between the macroeconomy and individual firm default. However, both research papers had a vastly different approach. Both papers analyzed the problem using econometric analysis rather than a classification approach as the one used in this thesis. On the other hand, Koopman et. al (2009) and Koopman, Lucas & Schwaab (2011), argues that macroeconomic indicators alone might either under or over-estimate default risk and that macroeconomic variables could most likely lose their predictive power in conjunction with firm-specific variables. Again, the mentioned research is conducted using a different approach than this thesis, but the conclusion found by Koopman et. al (2009) is in line with what is found in this thesis i.e. that macroeconomic variables do not yield a significant amount of predictive power at a firm-level when analyzed alongside firm-specific variables.

The general conclusion of the result found in relation to academic research is that the problem is difficult to give a clear-cut answer on. The result indicates that non-firm specific variables are mostly redundant and that the overwhelming predictive power of corporate default is found from financial ratios. However, the data is perhaps not favoring non-firm-specific variables in the first place. This will be discussed in 8.2.3.

8.2.3 Limitations

As mentioned in section 4, the data consists of financial ratios from annual data. The consequence of this is that non-firm-specific variables are also annual. This is a significant shortfall of the analysis. Optimally, the analysis should have been conducted on quarterly data. Doing this would allow for a more variation within each explanatory variable. When balancing the data such that there is a proportional number of observations belonging to the years from which the financial data is from, the variation in the non-firm specific variables becomes limited. However, as explained in section 5.1, balancing the data is necessary. Having a dataset which is more detailed regarding the frequency of data i.e. quarterly data, would potentially give the non-firm-specific variables a better chance of yielding some additional information. It is therefore important to state that the result found when testing hypothesis 2, should be seen in the light of the data. Using annual data, non-firm-specific variables yield a minimal impact on accuracy, but it cannot be ruled out that non-firm-specific data could have an effect on a different dataset using quarterly observations.

In addition, it cannot be stated that non-firm-specific data in general does not yield any additional information even using annual data. It might be the case that the set of explanatory variables was not chosen optimally.

8.3 Hypothesis 3

In this section the results presented in section 7.3 will be discussed. First, an interpretation of the results will be presented followed by a discussion of the implications of the results. Then the limitations will be discussed.

8.3.1 Interpretation

First, the results from earlier sections will be presented and their meaning will be discussed.

Table 24: Best Overall Accuracy

	Combined	Wholesale	Manufacturing
Random forest	0.773	0.768	0.763
Logistic regression	0.726	0.700	0.724

Table 24 shows the best overall accuracy from each subsample using both models. As shown above, the accuracy of the random forest model on the combined sample is higher than for either subsample alone, however only to a minor degree. This means that the model performs better when applied on the combined sample than for either subsample alone. This is contrary to the expectations that specific variables are indicative of default conditional on industry which should result in higher accuracy in the specified subsamples due to higher degree of discrimination. This was not the case. The accuracy of the logistic regression model on the combined sample is higher than for each sub-sample alone in a similar manner to random forest. Again, this means that the model performs better when applied on the combined sample than for either subsample alone. Thus, this indicates that hypothesis 3 was incorrect since no superior insight is gained in accuracy of the default prediction from subdividing the industries of wholesale trade and manufacturing in southern Europe. Now, the variables involved in each model are addressed.

Table 25: Variable Importance - Random forest

Variable Rank	Combined	Wholesale	Manufacturing
1	SR	SR	SR
2	IC	IC	IC
3	ROSF	ROTA	ROSF
4	G	G	G
5	ROTA	ROSF	CFT
6	CFT	CFT	ROTA

Table 25 shows the six most important variables for the random forest model ranked by their mean decrease in gini. As shown in table 25, the exact same six variables yield the largest mean decrease in gini for the three samples. In other words, the variables that are important when predicting corporate default using random forest is found to be identical. This supports the earlier conclusion on accuracy and shows that the driving factors when predicting default using random forest is not conditional on industry.

Table 26: Variable Importance - LASSO

Significant*(+/-)	Combined	Wholesale	Manufacturing
*	ROSF (-)	ROSF (-)	ROTA (-)
*	CTF (-)	ROTA (+)	CFT (-)
*	PM (-)	CFT (-)	PM (-)
*	SR (-)	PM (-)	SR (-)
*	G (+)	SR (-)	G (+)
*	GDPG (+)	G (+)	

Table 26 shows the significant variables when applying logistic regression with lasso regularization. +(-) means that the variable is positively (negatively) correlated with corporate default. As seen in table 26 there are definite similarities in some of the variables across the subsamples. Cash flow turnover, profit margin and solvency ratio are all negatively correlated with corporate default for all subsamples while gearing is positively correlated with corporate default for all subsamples. This was in line with expectation. Some other variables like GDP growth which is positively associated with corporate default for the combined sample and return on total assets which is positively associated with corporate default for the wholesale trade sample were not in line with expectation. These variables will be thoroughly discussed in a later section.

In general, the variables that are important when predicting corporate default using logistic regression is found to be similar across the samples with certain exceptions. This supports the earlier conclusion on accuracy and shows that the driving factors when predicting default using logistic regression is not conditional on industry.

Hypothesis 3 states the expectations: "*It is expected that the driving variables and the precision of the models are conditional on industry*". This is based on the idea that companies of a given industry should share certain similar characteristics. These characteristics should be visible in the accounting statements and thus in the financial ratios based on accounting data. Thus, it was expected that the accounting information that is important when predicting corporate default should differ across industries. This was not found. On the contrary, the results indicate that the driving factors in predicting corporate default is not conditional on industry. This conclusion is based on the fact that the models performs better (higher accuracy) when no subdivision of industries is done. In conclusion it was found that when a model is trained only on data from companies within the same industry, the model performs worse.

It is also found that the exact same six variables are the most important when using the random forest model while very similar variables where significant when using logistic regression. Again, this goes against the expectations stated in hypothesis 3. If hypothesis 3 in fact was true, differences in the types of variables that are important should differ materially between industries. This is not the case. In conclusion, no material difference between the driving factors were found when subdividing the industries.

8.3.2 Academic Literature

Wang (2011) summarizes the ratios that are significant for default prediction by listing 27 accounting variables and grouping the variables in five categories: profitability, operational efficiency, liquidity, capital structure and firm size. The specific accounting variables must be significant in more than four different studies to qualify for the list. Wang (2011) also lists the expected sign for the variable indicating a positive or negative relationship with corporate default.

The variables found to be most important in this thesis by application of random forest are solvency ratio, interest coverage ratio, return on shareholders funds, gearing, return on total assets and cash flow turnover. Of these six variables, the solvency ratio, interest coverage ratio and return on shareholders fund were not found to be significant in more than four studies (Wang, 2011). It is noted that very few ratios including common equity or shareholders funds were included as significant in more than four previous studies which explains why solvency ratio and return on shareholders funds are not included since they are equity measures. Gearing is the only equity measure found to be significant in more than four studies. The findings in this thesis, however, shows that the equity measures in fact are important relative to other measures included in the analysis.

The interest coverage ratio was found to be the second most important variable for all subsamples of random forest but is not found to be significant in more than four studies. Intuitively, the interest coverage ratio should be important for default as it demonstrates the ability of the company to service its interest payments, which is crucial for the survival of the company. This is supported by the findings in this thesis, but not in the broader literature. It is noted that several liquidity measures using only balance sheet items such as current assets and liabilities are found to be significant in other studies such as the current ratio and the quick ratio. These types of ratios were not found to be important relative to others in this thesis.

The variables found to be most important in this thesis by application of logistic regression include cash flow turnover, profit margin, solvency ratio, and gearing. Of these four, profit margin and solvency ratio were not found to be significant in more than 4 other studies. Relatively similar measures for profit margin and capital structure, however, are included such as operating income / sales, total liabilities / total assets and gearing. It is expected that multicollinearity between similar types of ratios exist which would explain why some profitability and capital structure ratios are found to be significant in more than 4 studies while others are not. In this study, both profit margin, EBIT margin, and EBITDA margin were initially significant in the application of logistic regression before lasso regularization. After lasso regularization only profit margin is significant. This is expected to be the case since all three variables basically explain the same thing about the company: the margin of profitability including different cost types.

Generally, the types of ratios found to be important in this thesis are very similar to the types of ratios found in other studies specifically profitability and solvency ratios. One difference between the findings of the broader literature and this thesis is the relative importance of equity and asset measures of profitability and leverage. Whereas this thesis finds great importance of equity measures, generally asset measures are found to be significant in the broader literature.

The overall results found both in terms of variables and precision indicated that the driving factors in predicting default were not conditional on industry. This result was unexpected and leads to the rejecting of hypothesis 3, specifically for the manufacturing and wholesale trade industries in southern Europe. Only two industries were tested which means that no conclusion can be made as to whether some driving factors are specific to other industries, more narrowly defined groups or specific geographies. On the other hand, no results were found that indicates such a relationship.

Two significant variables specifically in the application of logistic regression were found that were highly unexpected. For the combined sample, GDP growth was found to be positively correlated with corporate default. This indicates that the more the economy is growing the higher the probability of default should be, all else equal. This is found to be counter intuitive since higher growth in the economy usually is associated with better business environment. It is hypothesized by the authors that the reason for this unexpected result is the fact that the data used for the analysis is unbalanced in terms of relative rates of bankruptcies between countries. If one country has higher rates of bankruptcy and higher GDP growth in the training data, the model will infer a relationship that is real historically but without any causal relation.

For the wholesale trade sample return on total assets was found to be positively associated with corporate default. This indicates that higher returns on assets should increase the risk of corporate default which is highly counter intuitive. When comparing with other similar variables such as return on shareholders funds and profit margin that are negatively correlated with default for the same sample, it is found that no other variable points in the same direction. One could argue that despite the fact that the specific results goes against intuition, the model may still perform very well since variables such as profit margin, cash flow turnover and return on shareholders funds also are included and are highly negatively correlated with default, balancing the total effect of the profitability measures on the model.

An argument could be made that size of the samples used could influence the accuracy of the model. Therefore, combining two subsamples should automatically increase the accuracy of the model exclusively because of the increase in the size of the combined sample. This would explain the results found on accuracy of both models where the combined sample consistently was more accurate in predicting corporate default. Since the same variables were consistently shown to predict default across the two industries analyzed no indication was made that the driving factors are conditional on industry. This, however, is only shown for a small subset of the total industries. Therefore, further analysis should be conducted to show whether this holds true for any subset of industry. No prior research found has demonstrated this hypothesis to a full extend. Such an analysis would also show more clearly the effect on precision on combining all industries compared to the subsample alone which ties back to the argument at the start of this section.

For now, however, the analysis conducted showed that precision is gained from combining the samples and that default drivers are not specific to industries such as concluded earlier.

The general implication of the results is that a general model has superiority over a model on a more narrowly defined subsample. More research on other industries should be done in order to fully determine whether all industries share the general default drivers as the wholesale trade sample and manufacturing sample was shown to do.

The findings of this thesis are similar to those of earlier research in some of the types of ratios found to be important. Here, profitability- and solvency ratios are found to be the most important types of default drivers while other studies also find operational efficiency, liquidity and firm size to be significant drivers (Wang, 2011).

It was found that equity measures of profitability were the most important default drivers in this thesis. This is not generally found in the literature since only one equity measure, gearing, is significant in more than four studies (Wang, 2011).

This thesis adds to earlier research by exploring the industry specificity of default drivers for corporations. More thorough industry analysis must be done in order to fully determine whether any such specific drivers exist. Types of businesses that are truly specific and unlike others in an accounting sense are expected to be the most likely to have specific default drivers. Real estate management and consulting are examples of types of business that are expected to have significantly different accounting statements simply due to the nature of the business.

8.3.3 Limitations

The generalizability of the results is limited by the number of industries analyzed. Since the analysis was restricted to the wholesale trade and manufacturing industries the results and resulting conclusions are limited to these industries alone. Moreover, the question of the specificity of default drivers can only be answered for the industries analyzed and therefore no generalization can be made yet for other industries and geographies.

Restricting the analysis to two samples had other consequences for the result. The fact that the variables of the combined sample were similar to the two industry samples is no coincidence since both subsamples are substantial parts of the combined sample. Therefore, it should be expected that combining only two samples would yield similar significant variables as either sample alone due to fact that each sample is such a material part of the combination. Were many different subsamples analyzed, each subsample would only be a fraction of the combined sample and therefore any dominant variables in the subsamples would average out in the combination. If the important variables still are similar this is irrelevant. This is another reason to conduct further research on more industries in order to gain the full picture.

9. Conclusion

This thesis tests the degree to which it is possible to predict corporate default using machine learning. To test the overall problem statement, three specific hypotheses are tested. It is tested how accuracy vary between logistic regression and random forest. Then, the effect on accuracy by the addition of non-firm-specific variables is tested. Finally, it is tested whether the driving variables and the precision of the models when predicting corporate default are conditional on industry.

In order to test the first hypothesis, a logistic regression model with and without LASSO and a random forest model was performed on the combined sample. Both models are tested using undersampling and SMOTE techniques. It is found that random forest outperformed logistic regression by 4.7 percentage points using undersampling indicating that random forest is superior to logistic regression for corporate default prediction. In addition, it is found that applying the SMOTE algorithm distorted the results for bankruptcy prediction for both random forest and logistic regression which indicates that undersampling is the preferred sampling method for accuracy in corporate default prediction.

To test the second hypothesis, a logistic regression and random forest is performed on the combined sample using financials ratios versus using financial ratios and non-firm-specific variables such as macroeconomic and market variables. With random forest it is found that the addition of non-firm-specific variables slightly increases accuracy. For logistic regression no difference is found. Conclusively, no significant effect on accuracy is found overall.

In order to test the third hypothesis a logistic regression and random forest is performed on the combined sample as well as the manufacturing sample and wholesale trade sample. It found that accuracy is highest for the combined sample, indicating no gain from splitting samples by industry for the industries tested. To substantiate this claim, it is found that the driving variables for both models tested are greatly similar between industries. In total, this indicates that the driving variables are not conditional on industry.

Conclusively it is shown that random forest outperforms logistic regression in terms of accuracy which confirms hypothesis 1. It is shown that the addition of non-firm-specific variables does not materially improve accuracy and thus hypothesis 2 is rejected. Lastly, it is found that the precision and driving variables are not conditional on industry and therefore hypothesis 3 is rejected.

10. Further Research

10.1 Hypothesis 1

The analysis was conducted using logistic regression as a parametric model and random forest as a non-parametric model. It would be interesting to see how other models perform such as linear- and multiple discriminant analysis as parametric models and K-nearest neighbors, neural networks and support vector machines using different kernels e.g. linear, polynomial and radial basis function. Expanding the set of models used to analyze corporate default prediction would enable one to conclude whether there is in fact a difference between parametric and non-parametric models regarding accuracy. Comparing the accuracy between a vast set of different classification models would also be interesting regarding the underlying approaches of the models. For instance, if there is little difference between the accuracy of the non-parametric models, it indicates that the problem can be analyzed sufficiently using different approaches to the data. On the contrary, if there is a significant difference, it would be interesting to investigate why some non-parametric models are in fact better than the others.

Similarly, it would be interesting to see whether extensions to the traditional SMOTE algorithm e.g. borderline-SMOTE, Density-based SMOTE, Safe-level SMOTE and Relocating Safe-level SMOTE would yield different results than what was seen with the traditional SMOTE procedure. It might be the case that some of the extensions are able to dissect the outliers and synthetically create minority observations that improve the accuracy of the models dramatically. It could also be the case that other classification algorithms such as neural networks and support vector machines, which have been praised for their accuracy, are able to better use the increased, but noisy dataset which SMOTE generates. All these questions could be interesting to investigate in future research.

10.2 Hypothesis 2

Regarding the second hypothesis, there are some interesting prospects for future research that can either substantiate the conclusions that was found in this thesis or bring other views to life. Firstly, a comprehensive analysis using different classification methods should be employed to conclude whether other models are able to extract information from non-firm-specific variables. If this is the case, a larger focus on non-firm-specific variables can be justified.

In addition, it is suggested that future corporate default prediction research using the classification approach should be conducted on quarterly financial data. Having a deeper time dimension allows for several interesting extensions to the research problem. Firstly, the incorporation non-firm-specific data will become more sophisticated which can potentially reveal some other conclusions than found in this thesis. Secondly, looking at the financial variables, it is suggested that future research should not only be based on the absolute ratios, but changes in ratios as well. There is potentially substantial information in the change of ratios over time. The inclusion of the change in financial ratios over time can be done with annual data, but should preferably be done with quarterly data since it will give a more sophisticated view of the actual change.

Lastly, it is suggested that future research should investigate how the accuracy of corporate default prediction is conditional on time until default. For instance, how much does the accuracy decrease when predicting default one year before the event happens compared to two, three, four or five years? The analysis could also be brought down to a quarterly level which perhaps is more relevant. In other words, instead of predicting default on annual intervals, an investigation of the accuracy of default on a quarterly basis could be interesting.

10.3 Hypothesis 3

The analysis done to answer hypothesis 3 was constrained by the number of industries included. In order to fully answer hypothesis 3 further research is needed. Further research should be based on fully exploring whether the results found in this thesis holds true for all industry segments. By doing so, it is possible to fully determine whether a generalized model always performs better and whether some variables are uniformly important or conditional on industry.

The analysis should cover at least two points that would ensure an exhaustive analysis.

First, there must be a large enough fraction of industries included to fully cover all the different types of companies. Optimally all major industries and/or branches of business should be included. The goal is to ensure that all different types of businesses are covered such that all possible dependent variables can be found. More data gives more confidence in the conclusion whatever way the data points. Before a final conclusion on industry specificity can be made, all incumbent industries must be analyzed.

Second, the industries should generally be defined as narrowly as possible in order to get as homogenous a subsample as possible. This should increase the likelihood that specific industry-dependent variables are significant due to higher specificity of business and higher uniformity within the sample.

11. Bibliography

Agarwal, V., & Taffler, R. J. (2007). Twenty-five years of the Taffler z-score model: does it really have predictive ability. *Accounting and Business Rsearch*, 285-300.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 589-609.

Altman, E. I. (1973). Predicting Railroad Bankruptices in America. *The Bell Journal of Economics and Management Science*, 184-211.

Altman, E. I., & Loris, B. (1974). A Financial Early Warning System For Over-The-Counter Broker-Dealers. *The Journal of Finance*.

Altman, E. I., & McGough, T. (1974). Evaluation of a Company as a Going Concern. *Journal of Accountancy*.

Altman, E. I., Haldeman, R. G., & Narayanan, P. (1977). Zeta Analysis: A new model to identify bankruptcy risk corporations. *Journal of Banking and Finance*, 29-54.

Amayyan, W., & Almayyan, H. (2018). Bankruptcy Prediction using Random Forest and Particle Swarm Optimization. *Journal of Convergence Information Technology*.

Bank, T. W. (2019, February 5). *Commodity Markets*. Retrieved from The World Bank: http://www.worldbank.org/en/research/commodity-markets

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems With Applications*, 405-417.

Beaver, H. W. (1966). Financial Ratios as Predictors of Failure. *Journal of Accounting Research*, 71-111.

Bentley, J. L. (1975). Multidimensional Binary Search Trees Used for Associative Searching. *ACM*, 509-517.

Breiman, L. (2001). Random Forests. *Machine Learning*, 5-32.

Brijs, T., Gilbert, S., Vanhoof, K., & Wets, G. (2018). Using Shopping Baskets to Cluster Supermarket Shoppers.

Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2009). Safe-level synthetic minority oversampling technique for handling the class imbalanced problem. *Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, 475-482.

Carling, K., Jacobsen, T., Linde, L., & Roszbach, K. (2007). Corporate credit risk modeling and the macroeconomy. *Journal of Banking and Finance*, 845-868.

Chava, S., & Jarrow, R. A. (2008). Bankruptcy Prediction with Industry Effects. In S. Chava, & R. A. Jarrow, *Financial Derivatives Pricing* (pp. 517-549). World Scientific Publishing.

Cook, R. D. (1977). Detection of Influential Observation in Linear Regression. *Technometrics*, 15-18.

Cover, T., & Hart, P. (1967). Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, 21-27.

Cutler, A., Cutler, D. R., & Stevens, R. J. (2011). Random Forests. *Machine Learning*.

Dijk, B. V. (2019, February 4). *bvdinfo*. Retrieved from Bureau Van Dijk Web Site: https://www.bvdinfo.com/en-gb/our-products/data/international/orbis

Dobson, A. J. (1990). *An Introduction to Generalized Linear Models.* London: Chapman and Hall.

Duffie, D., Saita, L., & Wang, K. (2007). Multi-period corporate default prediction with stochastic covariates. *Journal of Financial Economics*, 635-665.

Figini, S., Savona, R., & Vezzoli, M. (2016). Corporate Default Prediction Model Averaging: A Normative Linear Pooling Approach. *Intelligent Systems In Accounting, Finance And Management*, 6-20.

FitzPatrick, J. P. (1932). A Comparison of Ratios of Successful Industrial Enterprises with Those of Failed Firms. *Certified Public Accountant*, 598-605; 656-662; 727-731.

Hastie, T. J., & Pregibon, D. (1992). Generalized linear models. In J. M. Chabers, & T. J. Hastie, *Statistical Models in S.* Springer - Wadsworth & Brooks/Cole Mathematics Series.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.

Hillegest, S. A., Keatin, E. K., Cram, D. P., & Lundstedt, K. G. (2004). Assessing the probability of bankruptcy. *Review of Accounting Studies*, 5-34.

IMF. (2019, February 5). *Data*. Retrieved from Internatioanl Monetary Fund: https://www.imf.org/en/Data

Izan, H. Y. (1984). Corporate distress in Australia. *Journal of Banking and Finance*, 303-320.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with applications in R.* New York: Springer.

Keasey, K., & Watson, R. (1991). Financial Distress Prediction Models: A Review of Their Usefulness. *British Journal of Management*, 89-102.

Koopman, S. J., Kraussl, R., Lucas, A., & Monteiro, A. B. (2009). Credit cycles and macro fundamentals. *Journal of Empirical Finance*, 42-54.

Koopman, S. J., Lucas, A., & Schwaab, B. (2011). Modeling frailty-correlated defaults using many macroeconomic covariates. *Journal of Econometrics*, 312-325.

Last, F., Douzas, G., & Bacao, F. (2017). Oversampling for Imbalanced Learning Based on K-Means and SMOTE.

Liapis, K., Rovolis, A., Galanos, C., & Thalassinos, E. (2013). The Cluseters of Economic Similarities between EU Countries: A View Under Recent Financial and Debt Crisis. *European Research Studies*, 41-66.

Liaw, A., & Wiener, M. (2018, 3). randomForest. *Cran R Project*. Retrieved from Cran r project.

Lin, Y. F., & Mclean, L. (2001). A data mining approach to the prediction of corporate failure. *Knowledge-Based Systems*, 189-195.

McCulagh, P., & Nelder, J. A. (1989). *Generalized Linear Models.* London: Champan and Hall.

Merwin, L. C. (1942). Financing Small Corporations in Five Manufacturing Industries, 1926-1936. *National Bureau of Economic Research*.

Ohlson, A. J. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 109-131.

Siriseriwan, W. (2018, February 22). *smotefamily*. Retrieved from Cran R Project: https://cran.r-project.org/web/packages/smotefamily/smotefamily.pdf

Stoltzfus, C. J. (2011). Logistic Regression: A Brief Primer. *Academic Emergency Medicine*, 1099-1103.

Taffler, R. (1983). The assessment of company solvency and performance using a statiscal model. *Account and Business Research*, 295-308.

Taffler, R. J. (1984). Empirical models for the monitoring of Uk corporations. *Jorunal of Banking and Finance*, 199-227.

Tam, K. Y., & Kiang, M. Y. (1992). Managerial applications of neural networks: the case of bank failure prediction. *Management Science*, 926-947.

Van Gestel, T., Baesens, B., Suykens, J., Espinoza, M., Baestaens, D. E., Vanthienen, J., & De Moor, B. (2003). Bankruptcy prediciton with least squares support vector machine classifier. *Computational Intelligence for Financial Engineering*, 1-8.

Vapnik, V., & Chervonenkis, A. (1964). A note on one class of perceptrons. *Automation and Remote Control*.

Vapnik, V., & Lerner, A. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, 774-280.

Venables, W. B., & Ripley, B. D. (2002). *Modern Applied Statistics with S.* New York: Springer.

Wagenmans, F. (2017). Machine Learning in Bankruptcy Prediction.

Wang, Y. (2011). Corporate Default Prediction: Models, Drivers and Measurements.

White, R. W., & Turnbull, M. (1975a). The Probability of Bankruptcy: American Railroads.

White, R. W., & Turnbull, M. (1975b). The Probability of Bankruptcy for American Industrial Firms.

Winakor, A., & Smith, F. R. (1935). Changes in Financial Structure of Unsuccessful Industrial Companies. *Bureau of Business Research*.

WSJ. (2019, February 22). *Historical Prices*. Retrieved from The Wall Street Journal: https://quotes.wsj.com/index/XX/SXXP/historical-prices

Zimijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Account Research*, 59-86.

12. Appendices

12.1 Appendix - Descriptive Statistics (Undersampling)

Descriptive Statistics (Undersampling Sample)

	Manufacturing				Wholesale Trade			
	Active (N=4,019)		Bankrupt (N=3,928)		Active (N=2,305)		Bankrupt (N=2,312)	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Financial Ratios								
ROSF	0.14	0.39	-0.23	1.07	0.17	0.45	-0.13	1.09
ROCE	0.13	0.28	-0.01	0.59	0.17	0.35	0.07	0.61
ROTA	0.05	0.09	-0.02	0.11	0.05	0.08	0.00	0.12
PM	0.03	0.08	-0.06	0.16	0.03	0.06	-0.02	0.12
EBITDAM	0.08	0.08	0.01	0.13	0.05	0.06	0.01	0.10
EBITM	0.04	0.08	-0.03	0.14	0.03	0.05	0.00	0.10
CFT	0.06	0.08	-0.01	0.14	0.03	0.05	-0.01	0.10
NAT	4.16	9.59	6.57	18.76	8.22	18.23	13.13	25.66
ST	28.18	79.00	18.02	57.88	32.79	87.97	32.62	93.87
COP	93.92	71.96	115.04	110.04	78.21	70.31	95.26	106.20
IC	33.24	106.55	8.46	69.89	33.81	97.69	17.99	88.21
CP	59.45	49.01	87.97	74.33	60.37	61.88	84.08	92.49
CR	1.90	1.85	1.25	0.98	1.87	1.76	1.52	2.76
SR	0.37	0.21	0.20	0.15	0.34	0.22	0.20	0.17
LR	1.37	1.43	0.82	0.70	1.26	1.42	1.18	2.37
G	1.31	1.74	2.96	2.62	1.21	1.73	2.68	2.60
Firm Specific								
AGE	29.18	15.68	26.93	16.59	24.14	13.67	21.42	14.04
Macroeconomic Factors								
GDPG	0.01	0.02	0.00	0.02	0.01	0.02	0.01	0.02
IRNC	-0.01	0.01	-0.01	0.01	-0.01	0.01	-0.01	0.01
INFLNC	0.00	0.01	0.00	0.01	0.00	0.01	0.00	0.01
UNC	0.00	0.01	0.00	0.01	0.00	0.02	0.00	0.01
Stock Index								
STOXX	0.05	0.14	0.05	0.14	0.05	0.12	0.06	0.12
Commodities								
RMPNC	-1.07	7.58	-0.97	7.52	-1.31	7.12	-1.25	7.10
BMNC	-3.30	10.99	-3.15	11.06	-3.19	9.94	-3.00	9.99

12.2 Appendix - Descriptive Statistics (SMOTE)

Descriptive Statistics (SMOTE Sample)

	Manufacturing				Wholesale Trade			
	Active (N=25,097)		Bankrupt (N=3,928)		Active (N=25,301)		Bankrupt (N=2,312)	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Financial Ratios								
ROSF	0.15	0.40	-0.23	1.07	0.17	0.44	-0.13	1.09
ROCE	0.13	0.24	-0.01	0.59	0.17	0.35	0.07	0.61
ROTA	0.05	0.09	-0.02	0.11	0.05	0.08	0.00	0.12
PM	0.04	0.09	-0.06	0.16	0.03	0.06	-0.02	0.12
EBITDAM	0.08	0.08	0.01	0.13	0.05	0.06	0.01	0.10
EBITM	0.04	0.08	-0.03	0.14	0.03	0.06	0.00	0.10
CFT	0.06	0.07	-0.01	0.14	0.03	0.05	-0.01	0.10
NAT	3.85	7.25	6.57	18.76	8.67	23.36	13.13	25.66
ST	27.06	73.79	18.02	57.88	30.12	83.48	32.62	93.87
COP	95.10	71.76	115.04	110.04	76.27	69.24	95.26	106.20
IC	35.38	106.10	8.46	69.89	37.30	108.12	17.99	88.21
CP	60.12	52.99	87.97	74.33	59.77	59.13	84.08	92.49
CR	1.94	2.18	1.25	0.98	1.95	2.48	1.52	2.76
SR	0.38	0.22	0.20	0.15	0.35	0.22	0.20	0.17
LR	1.41	1.74	0.82	0.70	1.32	1.88	1.18	2.37
G	1.30	1.75	2.96	2.62	1.19	1.71	2.68	2.60
Firm Specific								
AGE	30.04	15.68	26.93	16.59	26.06	13.80	21.42	14.04
Macroeconomic Factors								
GDPG	0.01	0.02	0.00	0.02	0.01	0.02	0.01	0.02
IRNC	-0.01	0.01	-0.01	0.01	-0.01	0.01	-0.01	0.01
INFLNC	0.00	0.01	0.00	0.01	0.00	0.01	0.00	0.01
UNC	0.00	0.01	0.00	0.01	0.00	0.02	0.00	0.01
Stock Index								
STOXX	0.05	0.14	0.05	0.14	0.06	0.12	0.06	0.12
Commodities								
RMPNC	-1.06	7.53	-0.97	7.52	-1.29	7.14	-1.25	7.10
BMNC	-3.20	11.02	-3.15	11.06	-2.98	9.99	-3.00	9.99

12.3 Appendix - R Code

```r
#Data

# Load Data (Combined Undersampling Sample With All Variables)

mydata<-read.table(file.choose(),header=T)


# Shuffle Data

set.seed=1

mydatashuffle<-mydata[sample(nrow(mydata)),]


# Divide Data Into Training And Testing

traindata<-mydatashuffle[1:10051,]

testdata<-mydatashuffle[10052:nrow(mydatashuffle),]


# Classification Tree

# Install Necessary Package

install.packages("tree",dependencies=TRUE)

library(tree)


#Build Classification Tree

Classificationtree<-tree(Bankruptcy~.,traindata)

plot(Classificationtree)

text(Classificationtree)


#Deviance of Tree

Prunedtree<-cv.tree(Classificationtree,FUN=prune.tree)

plot(Prunedtree)


#Pruned Tree

mybestsize <- Prunedtree $size[which(Prunedtree $dev==min(Prunedtree $dev))]

Prunedtree <- prune.tree(Prunedtree,best=mybestsize)

plot(Prunedtree)

text(Prunedtree)
```

```
# Random Forest
# Install Necessary Package
install.packages("randomForest",dependencies=TRUE)
library(randomForest)

# Run Random Forest Algorithm (n=1000, m=6)
randomforestmodel <- randomForest(Bankruptcy~.,traindata,ntree=1000,mtry=6,importance=TRUE)
randomforestprediction <- predict(randomforestmodel, testdata[,-25], type='class')
randomforesttable <- table(testdata[,25], randomforestprediction)
AAaccuracyRF<-( randomforesttable [1,1])/sum(randomforesttable [1,])
BAaccuracyRF<-( randomforesttable [2,2])/sum(randomforesttable [2,])
OAaccuracyRF<-sum(diag(randomforesttable))/sum(randomforesttable)

# Variable Importance
varImpPlot(randomforestmodel,sort=TRUE,n.var=24,type=2,main=NULL)

# Error & Accuracy as a Function of # Trees
Acc<-vector("numeric",500)

for(x in 2:500){

  err.acc.model <- randomForest(Bankruptcy~.,traindata,ntree=x,mtry=6,importance=FALSE)
  err.acc.predict <- predict(err.acc.model, testdata[,-25], type='class')
  err.acc.table <- table(testdata[,25],err.acc.predict)
  accuracy<-sum(diag(err.acc.table))/sum(err.acc.table)
  error<-err.acc.model$err.rate[,1]
   Acc[x]<-accuracy
}
```

```
# Logistic Regression
logisticregressionmodel<-glm(Bankruptcy~.,data=traindata, family="binomial")

# Testing Logistic Regression Model With Cut-off = 0.5
logisticprobability<-predict(logisticregressionmodel,newdata=testdata[,-25],type="response")
logisticprediction<-rep("No",2513)
logisticprediction [logisticprediction>0.5]="Yes"
logistictable<-table(testdata[,25], logisticprediction)
AAaccuracyLR<-( logistictable [1,1])/sum(logistictable [1,])
BAaccuracyLR<-( logistictable [2,2])/sum(logistictable [2,])
OAaccuracyLR<-sum(diag(logistictable))/sum(logistictable)


# LASSO Regularization
# Install Necessary Packages
install.packages("glmnet",dependencies=TRUE)
library(glmnet)
install.packages("dplyr",dependencies=TRUE)
library(dplyr)

# Run LASSO Regularization
x<-as.matrix(traindata[,-25])
y<-as.matrix(traindata[,25])

cv.lasso<-cv.glmnet(x,y,family="binomial",alpha=1)
coef(cv.lasso,cv.lasso$lambda.1se)
lasso.model<-glmnet(x,y,family="binomial",alpha=1,lambda=cv.lasso$lambda.1se)

# Test LASSO Model With Cut-off = 0.5
x.test<-as.matrix(testdata[,-25])

lasso.probability<-lasso.model %>% predict(newx=x.test)
lasso.predict1<-rep("No",2513)
lasso.predict1[lasso.probability>0.5]="Yes"
lasso.table<-table(testdata[,25],lasso.predict1)
AA.accuracy.lasso<-(lasso.table[1,1])/sum(lasso.table[1,])
BA.accuracy.lasso<-(lasso.table[2,2])/sum(lasso.table[2,])
OA.accuracy.lasso<-sum(diag(lasso.table))/sum(lasso.table)

# SMOTE
# Install Necessary Package
install.packages("smotefamily",dependencies=TRUE)
library(smotefamily)

#Load Data (Combined SMOTE Sample With All Variables)
mydata1<-read.table(file.choose(),header=T)
set.seed=1
mydatashuffle<-mydata1[sample(nrow(mydata1)),]

traindata<-mydatashuffle[1:45310,]
testdata<-mydatashuffle[45311:nrow(mydatashuffle),]

# Run SMOTE
traindata1<-SMOTE(traindata[,-25],traindata[,25],K=5,dup_size=5)

# Assign The Dataframe Of The SMOTE Output To A Variable
traindata<-traindata1$data

# Convert Class To Factor
traindata$class<-as.factor(traindata$class)
colnames(traindata)[which(names(traindata) == "class")] <- "Bankruptcy"
```