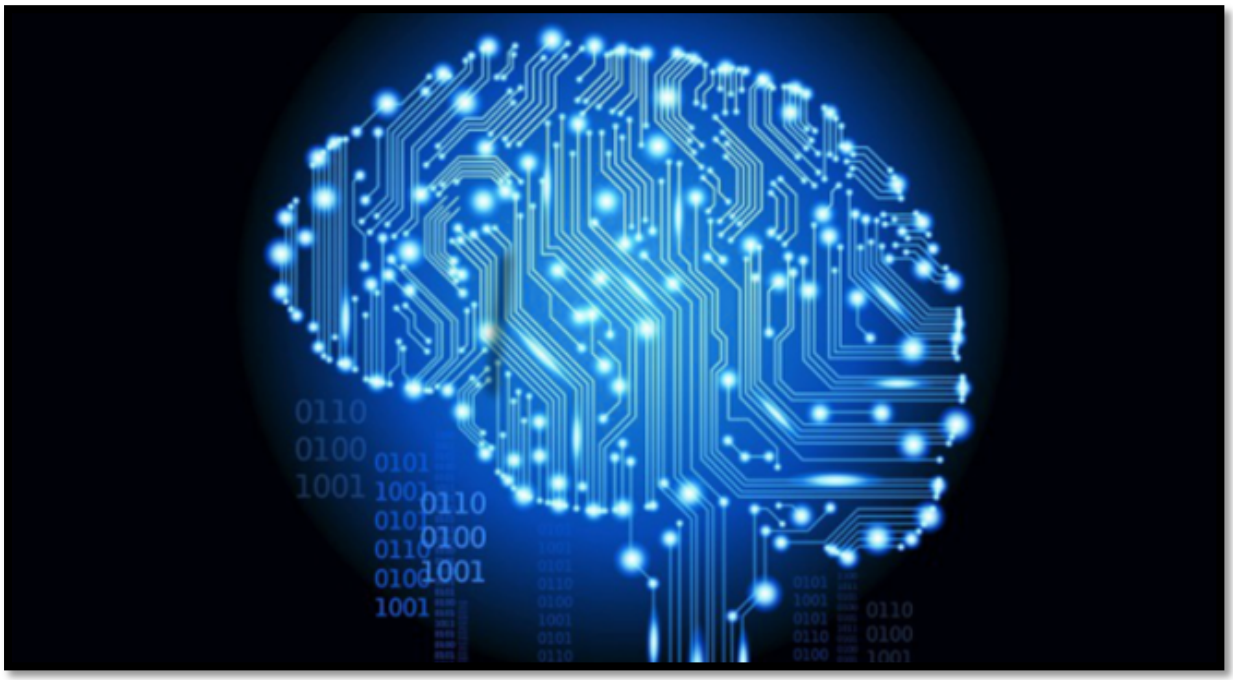
Master Thesis

DEFAULT PREDICTION

With the use of Machine Learning



Foudsigelse af Konkurs ved brug af Machine Learning

Supervisor: Jens Dick-Nielsen

Characters (including spacing): 237,864
Total pages: 104.6
15th of May 2020
Copenhagen Business School

Andreas Kjøller-Hansen
Student number: 102945
Cand.merc. in Finance and Accounting

Sara Skovhøj Jensen
Student number: 102932
Cand.merc. in Accounting, Strategy and Control

Abstract

This thesis investigates the classification of default and non-default on companies from the USA over the time period 1987-2015. The data is split according to two time horizons and whether market variables are included or not. This results in four data sets. The classification is done with the use of five different machine learning methods, logistic regression, neural network, linear SVM RBF SVM, and random forest. The models are evaluated by the accuracy and the distribution of type 1 and type 2 errors, and the ROC curve and its AUC measure. When only taking the accuracy and the distribution of the error types into account, the best methods when predicting default on data including accounting and market variables are neural network and linear SVM, whereas the best method on the data sets only including accounting variables is random forest. When the AUC measure and the ROC curve is taken into account, random forest is the best to predict default at all tested data sets. Overall the conclusion is that random forest, in general, is the most appropriate method when it comes to the empirical results on the data sets used in this thesis. The thesis also investigates variable selection with the use of logistic regression and random forest, and it concludes that the two methods are conflicting since random forest states some variables as least important variables, while logistic regression includes these in its models.

Finally, the results of the thesis are transferred to non-listed Danish firms with a focus on the capital requirement of the credit lender. There are two approaches to calculate the capital requirement, the IRB and the standardized approach. The larger credit institutions in Denmark primarily use the IRB approach, which uses the credit risk model of the credit lender to calculate values for PD, LGD, and EAD, and the approach benefits from setting lower capital requirements. There are other benefits of having a more precise credit risk model since it will imply the calculation of provision being more accurate and the evaluation of potential customers being more trustworthy and fair. The last part shows that the empirical results of the thesis are in accordance with other results from previous default studies.

Table of Contents

ABSTRACT **2**

1. INTRODUCTION **5**

1.1. MOTIVATION 6

1.2. RESEARCH QUESTION 6

1.3. LITERATURE REVIEW 7

1.3.1. Introduction *7*

1.3.2. Mathematical Methods for Predicting Default *7*

1.3.3. New Approaches for Predicting Default *9*

1.3.4. Conclusion and Thesis Contribution *10*

1.4. DELIMITATION 11

2. METHODOLOGY **13**

2.1. APPLIED THEORY OF SCIENCE METHOD 13

2.2. THE METHODOLOGY OF THE PAPER 14

2.2.1. Data Collection *14*

2.2.2. Quality Assessment *22*

2.3. THE STRUCTURE OF THE PAPER 23

3. RISK MANAGEMENT **24**

3.1. REGULATION IN RISK MANAGEMENT 24

3.1.1. Credit Risk *25*

3.2. PREDICTING DEFAULT 26

3.2.1. Default Prediction Methods *26*

3.2.2. Machine Learning *30*

4. EMPIRICAL RESULTS **42**

4.1. ONE YEAR PRIOR TO DEFAULT INCLUDING MARKET AND ACCOUNTING VARIABLES 42

4.1.1. Logistic Regression *43*

4.1.2. Neural Network *46*

4.1.3. Support Vector Machine *47*

4.1.4. Random Forest *50*

4.1.5. The Machine Learning Methods Compared to One Another *53*

4.2. FIVE YEARS PRIOR TO DEFAULT INCLUDING MARKET AND ACCOUNTING VARIABLES 55

4.2.1. Logistic Regression *56*

4.2.2. Neural Network *59*

4.2.3. Support Vector Machine *60*

4.2.4. Random Forest *62*

4.2.5. The Machine Learning Methods Compared to One Another *65*

4.3. One Year Prior to Default Including ONLY Accounting Variables 67
4.3.1. Logistic Regression 68
4.3.2. Neural Network 70
4.3.3. Support Vector Machine 72
4.3.4. Random Forest 74
4.3.5. The Machine Learning Methods Compared to One Another 77
4.4. Five Years Prior to Default Including ONLY Accounting Variables 79
4.4.1. Logistic Regression 80
4.4.2. Neural Network 82
4.4.3. Support Vector Machine 83
4.4.4. Random Forest 85
4.4.5. The Machine Learning Methods Compared to One Another 88
4.5. Sub Conclusion 90
5. COMPARING THE MACHINE LEARNING MODELS 92
5.1. Comparing the Models Between the Data sets 92
5.1.1. Accuracy and Distribution of the Error Types 92
5.1.2. Best Model in terms of ROC and AUC 94
5.1.3. Discussion of the Difference between Accuracy and AUC 95
5.2. Variable Selection 98
5.2.1. All Selected Variables 99
5.2.2. Further Variable Selection 101
5.2.3. Industry Level 103
5.3. Including or Excluding Market Variables 104
5.4. Sub Conclusion 104
6. DANISH MARKET FOR CREDIT LENDING 106
6.1. Power and Calibration for Credit Risk Models 106
6.1.1. Capital Requirement and the IRB Approach 107
6.1.2. Internal Advantages of a Better Credit Risk Model 109
6.2. Results of the Thesis into the Perspective of the Litterature 110
6.3. Sub Conclusion 111
7. CONCLUSION 112
8. BIBLIOGRAPHY 115
9. APPENDIX 119

1. Introduction

Financial institutions are aware that the risk and return are in most cases, closely linked to one another. To get an acceptable return, the institution must obtain some type of risk. Since the financial crises in 2007-2009, risk management has got further attention from governments, regulators, and the financial institutions themselves. For credit institutions, the largest corporate risk exposure is the credit risk which means they must measure the creditworthiness of the borrowing firm. This is the purpose of default prediction for credit lenders to measure and calculate the credit risk. The studies within default prediction have been many over the years. It started with univariate analyses where the study of Beaver (1966) was the most widely known. Different multivariate analysis, such as Altman's (1968) Z-score and Ohlson's (1980) O-score followed. There is a great variation in default prediction processes from how many and which factors should be considered to which methods should be used to develop the model. These multivariate analyses are still developing, and since the ability to use personal computers to build models arises, the use of machine learning for predicting default has increased (Gissel, Giacomino, & Akers, 2007).

The credit lending institutions need to follow different rule sets and regulations for the benefit of the customers and to make sure they keep being solvent. One of them is the financial rules known as the Basel accords by the Basel Committee on Banking Supervision (BCBS). By replacing Basel I with Basel II, the credit lenders got the opportunity to develop their own models by the IRB approach and to use it to measure the credit risk. Therefore, this leads to an increased importance of developing default prediction models for credit lenders themselves. In 2013, some years after the Basel II was implemented, the larger Danish credit institutions used the IRB approach to calculate the credit risk for more than 80% of their loans to other firms (Sørensen, 2013).

This paper contributes to the studies regarding default prediction by using five different machine learning models, logistic regression, neural network, linear support vector machine, RBF support vector machine, and random forest. These methods create models on four different data sets considering two different time horizons and whether market variables are included or not. These models are analysed and discussed with a focus on the accuracy, the ability to separate correctly between the classes, and with the focus of variables selected. It is found that with accuracy as the evaluation measure, the neural network and linear SVM performs best when predicting default on data that includes market variables, and random forest performs best on data that exclude market variables. With the other evaluation measure, AUC, random forest is at all times the best method to separate correctly between the classes. To establish how the theoretical perspective of the paper can be used in practice, the assessment of the Danish market for credit lending is investigated regarding how the result can play a role for the credit

lender. First, it is found that the credit risk model is used to measure credit risk under the IRB approach. The IRB approach is preferred for all larger credit institutions since it lowers the capital requirement compared to the standardized approach. Second, it is found how a more precise credit risk model can be an advantage for the credit lender within the areas of estimating provision and evaluating a potential customer.

1.1. Motivation

In the field of finance, we have been taught different ways to determine credit risk, hence default prediction. The methods we were taught, such as the Z-score and O-score, are relatively old. Thus, it would be interesting to investigate this field to see if we would be able to challenge these methods. After an elective in data science, where we got to work with different machine learning models in R, our interest in building default prediction models rose. This combination of the elective and the main course on our studies then created a motivation to build default prediction models using different machine learning methods.

1.2. Research Question

In line with the motivation for default prediction, this thesis addresses the following research question:

Which default prediction method among those tested will be the most appropriate to use evaluated by the accuracy and the ability to separate between classes, and how can the credit lender benefit from a more precise credit risk model?

To answer this research question, the following sub-questions are to be answered:

- How can the machine learning models be trained to be more precise in predicting default?
- Which model has the highest accuracy and AUC, and what do these measures indicate? Which of them would be preferred when deciding the most appropriate default prediction model?
- How do some models use variable selection to determine which variables are most important?
- Do market variables add any predictive power in the models?
- How do credit lenders calculate credit risk, and what is the difference between the standardized and the IRB approach?
- How does credit lenders calculate their capital requirement, and what role does the credit risk model play?
- How can the credit risk model help to calculate provision and evaluate a potential customer?

1.3. Literature Review

1.3.1. Introduction

Through time many studies have focused on the problem of default prediction. Altman (1968), with his Z-score, was one of the first to establish a method of predicting bankruptcy with a traditional statistical method. He was followed by, among others, Ohlson (1980) and his O-score. In the 90s and onwards the technology, hence machine learning, evolved and new methods were introduced such as neural network, different forms of support vector machines, and random forest. Some other studies took a different approach to predict bankruptcy. Shumway (2001) and Chava and Jarrow (2004) argue for a hazard model in contrast to a single-period model. Chava and Jarrow (2004) also investigated other aspects such as the importance of industry effects, predicting bankruptcy in financial firms, whether monthly observations outperforms yearly observations and the importance of accounting variables as predictive variables. Campbell, Hilscher, and Szilagyi (2006) followed by investigating how market ratios can improve the model as well as the use of a time lag in bankruptcy prediction. In addition to that, Campbell et al. (2006) also investigate the performance of financially distressed firms.

1.3.2. Mathematical Methods for Predicting Default

Default prediction started with single-period credit rating models proposed by Beaver (1966) and Altman (1968) focusing on accounting variables. Beaver's (1966) study, including 79 failed and 79 non-failed firms representing a balanced data set in 38 different industries, examined the predictive ability of ratios by univariate analysis. He suggested for future research that multiple ratio analysis possibly would predict even better than the single ratios. This leads to Altman (1968), who introduced the first multivariate study made of 66 publicly held manufacturing entities where half went bankrupt, and the other half did not. This represents a balanced data set. He merged a set of financial ratios into a five-factor model. This model is called the Z-score and presents multiple discrimination analyses (MDA) that predict bankruptcy if the firm's score falls within a specific range. So, having a Z-score of greater than 2.99 makes the firm falls into the non-distressed category where a Z-score below 1.81 results in the distressed category. The area between 1.81 and 2.99 is defined as the "zone of ignorance" or the "grey area" because of the weakness in classification for this range. The Z-score, with its 79% accuracy for the hold-out sample one year before failure, has become one of the most well-known bankruptcy prediction models, and it is still today taught at undergraduate as well as postgraduate levels all over the world. Since Altman (1968) introduced the "Z-score" model in 1968, many new and more complex bankruptcy prediction models have been developed.

Ohlson (1980) contributed to the field of default prediction as being one of the first to search for a probabilistic output from the bankruptcy prediction model. The model is called the O-score and is a logistic regression model with nine different factors where each has a related coefficient. Ohlson (1980) did also increase the number of firms included in the data set significantly compared to previous studies. The O-score model was built upon data from 2,058 non-bankruptcy and 105 bankruptcy firms in the period from 1970 to 1976, representing an unbalanced data set. The O-score uses a maximum likelihood function that seeks to give each observation a probability of default. The result of the O-score is not immediately interpretable, meaning there is a need for a transformation of the output. To convert the output into a probability, the following formula should be used, $p(failure) = \frac{e^{O-score}}{1+e^{O-score}}$. The O-score does not automatically split the observations into the binary classes, "default" or "non-default". Instead, there is a need for a cut-off determining in which probability interval the firm is being classified as default. Ohlson (1980) set this cut-off point at 0.038 as this point minimizes the sum of errors. This means that firms with a probability of default of 3,8% or higher are being classified as default. This leads to an accuracy of 85.1% one year prior to bankruptcy. However, it is possible to increase the overall accuracy by reducing the cut-off point. Though, this reduction will come with the cost of a higher proportion of type 1 errors.

During the 90s, the use of machine learning evolved, and new methods for classification problems were created. These methods could be used to default prediction, where the first method was neural network. One of the first to use the neural network method for default prediction was Wilson and Sharda (1994). They compared the predictive accuracy of neural network and MDA by using the sample from Moody's Industrial Manuals containing 65 bankruptcy firms and 64 non-bankruptcy firms matching on year and industry. This represents a balanced data set. Wilson and Sharda (1994) have five explanatory variables and decided to use similar ratios as Altman (1968). The results showed that neural network outperformed MDA in predicting accuracy. Neural network achieved an accuracy of 97.5%, while MDA only achieved an accuracy of 88.25%. The most accurate prediction result was found when the training and the testing sample was balanced. This means that if the sample composition had a higher proportion of non-bankruptcy firms, the two methods were worse to predict bankruptcy firms despite neural network still had a better result than MDA.

Following up on the evolution of technology, many new methods for classification problems developed. Baesens et al. (2003) investigated 17 different state-of-the-art classification methods. The data included eight different data sets containing information about consumer loans. Baesens et al. (2003) showed that neural network was the best classifier method in four out of the eight data sets while Radial Basis Function Least Squared-Support Vector Machine classifier was the best method in two out of the eight data sets. Furthermore, the article highlights the percentage correctly classified (accuracy) and the area

under the receiver operating characteristic curve (AUC) as different measures to evaluate accuracy. Lessmann, Baesens, Seow, and Thomas (2015) updated the article to include 41 different classification methods instead of the original 17 methods. Though, the data sets were very similar to the original ones from 2003. The study concluded that it was time to move away from logistic regression as the industry standard. In addition, it was shown that random forest, multilayer perceptron artificial neural network, and hill-climbing ensemble selection with bootstrap sampling were the best classification methods depending on different misclassification error cost.

There has also been a test of machine learning methods in comparison with more traditional models to predict corporate bankruptcy. Barboza, Kimura, and Altman (2017) investigate the methods, linear SVM, RBF SVM, boosting, bagging, random forest, neural network, logit and MDA, as well as their individual ability to predict corporate bankruptcy. The comparison was made on data from Compustat where the available data were split into two, in 1) a training sample in the period from 1985 to 2005, and in 2) a testing sample in the period from 2006 and until 2013. The training sample contained 449 firms that went bankrupt and the same number of healthy firms, indicating a balanced training sample. For the comparison, 11 explanatory variables were chosen where five of them originate from Altman (1968), and the remaining six variables originate from other studies. The variables were not normalized, so they had the calculated values of the ratios. The results showed that machine learning models outperform traditional models. Especially the machine learning methods; boosting, bagging, and random forest did well with all having accuracies over 85% for the testing sample. Furthermore, the study highlights that machine learning methods might be better to predict bankruptcy, but they do not necessarily explain why the company files for bankruptcy.

1.3.3. New Approaches for Predicting Default

The development of bankruptcy prediction has also led to aspects with different approaches than the traditional statistical method or the machine learning methods. At the beginning of the 00s, Shumway (2001) argued that single-period classification models, which he refers to as static models, are inappropriate for forecasting bankruptcies due to the nature of bankruptcy data. When applying a single-period model to predict bankruptcy, as Altman (1968) did, the analyst has to select when to observe each company's characteristics, because the model only considers one set of explanatory variables for each firm at a chosen time and as known most firms change from year to year. This may lead to unnecessary selection bias into the estimates. Bankruptcy data are multiple periods since bankruptcy arises occasionally, and analysts must use information from more than one financial year, for the given company, to estimate the models. Of these reasons, Shumway (2001) introduced a multi-period model which he refers to as a simple hazard model. The final sample contained 300 bankruptcies in the period from 1962 to 1992. The simple hazard model introduced can be thought of as a binary logit model where

the dependent variable is the time spent by a firm being in the healthy group. That way, the simple hazard model solves the complications of single-period models by explicitly account for time and exploit all available data for a given firm. So, the hazard models may produce more efficient out-of-sample forecasts than single-period models by utilizing much more data.

In the middle of the century, Chava and Jarrow (2004) confirmed the more accurate prediction of Shumway's (2001) hazard model in comparison to Altman's (1968) static model. This was done with the data of U.S. firms in the period from 1962 to 1999 with a total number of 1,461 bankruptcies whereas usually, it was no more than 300. This data set with yearly and monthly observation intervals were also used to other analyses. Among others, they found that the importance of industry effects appeared to be statistically significant in-sample, but it did not radically increase the out-of-sample accuracy. An extension of the hazard rate model was applied to financial firms and monthly observation intervals instead of the usual yearly observations. It was found that bankruptcy prediction for financial firms is more difficult to exercise than it is for non-financial firms, and that monthly forecasting increases the accuracy of all models in a statistically significant way. Finally, Chava and Jarrow (2004) demonstrated that accounting variables only add little predictive power when market variables are already included in the bankruptcy model.

Campbell et al. (2006) have the same starting point as Shumway (2001) and Chava and Jarrow (2004) in terms of a logit model with the same five variables. The data in the article is from Compustat with the use of monthly observations in the period from 1963 to 2003. It contains more than 10,000 U.S. firms. First, Campbell et al. (2006) seek to investigate how well market ratios can improve the model where he uses a time lag to examine bankruptcy prediction at long horizons. The result shows, as Chava and Jarrow (2004) also found, that market data was more important compared to accounting data. This applied particularly when the forecast horizon was increased. The second part of the article investigates the return of financially distressed firms. These firms deliver anomalously low average returns, according to Campbell et al. (2006), despite their high volatility and betas. In addition, these firms also tend to have small market capitalization and high book-to-market ratios which are factors that are included in the Fama and French three-factor model (1993). Fama and French (1993) argue that size and value stocks deliver abnormal high returns, but Campbell et al. (2006) show that it is not the case for financially distressed stocks.

1.3.4. Conclusion and Thesis Contribution

The literature review shows that there has been a significant development in the area for bankruptcy prediction. Altman (1968) set the standard back in the late 60s, and his method is still learned and used worldwide. Ohlson (1980) did also disrupt the field by making a logistic regression model giving each

firm a probability of default. The technological evolution and the spreading of personal computers made machine learning feasible for everyone. During the 90s and 00s, many new machine learning models arose which all tried to beat the previous ones. The accuracy did rise, but throughout the literature, there are different opinions on which models perform best. Our contribution to the field is to test five different machine learning models; logistic regression, neural network, linear SVM, RBF SVM and random forest on a large and realistic data set. The methodology of the thesis is very close to what Barboza et al. (2017) did in terms of selection of a balanced training set and testing the machine learning models on a more realistic imbalanced testing set. However, the thesis seeks to take the point of view of the credit lenders. Therefore, the analysis is also expanded to include a test for private firms where market variables are excluded. From our point of view, this would make the thesis more reliable and usable for credit lenders. Compared to Barboza et al. (2017), we believe that normalizing the ratios is closer to the state-of-the-art data science technique as well as having a more significant number of defaults like Chava and Jarrow states (2004).

Some studies had another approach than just being successful in terms of accuracy. Shumway (2001) concluded that a multi-period classification model, a simple hazard model, would be a better and more preferred method in predicting bankruptcies than a single-period model which was agreed upon by Chava and Jarrow (2004). Despite that, it is decided only to have data of one year of each observation in this thesis. The reason for this is that the focus of this study is more like a practical test of different machine learning models instead of the mathematical development of a hazard model. Chava and Jarrow (2004) also found that financial firms were more challenging to predict bankruptcy compared to non-financial firms. These findings will be used in this thesis to exclude financial firms in default prediction to make the prediction more valid for the rest of the companies. Further studies could do the test of machine learning models separately for financial firms. Finally, it was found by both Chava and Jarrow (2004) and Campbell et al. (2006) that monthly observations yielded better accuracy than yearly observations and that accounting variables did not add much predictive power when the model already included market variables.

1.4. Delimitation

The main delimitations made in the thesis will be described in the following, and minor delimitations will be done accordingly in the thesis in the part it fits. First and foremost, this thesis focuses on firms rather than individuals. Firms have, all else equal, more variables to analyse than individuals have. Furthermore, the firm aspect is more closely related to our academic area than the individual aspect concerning the courses we have taken, such as Financial Statement Analysis. Credit rating consists of different aspects, though this thesis focuses on a quantitative method to predict default rather than a qualitative method that considers the human perspective when determining whether the firm might

default or not. The way the quantitative method is carried out in this thesis is by predicting default with the use of five different classification machine learning methods; logistic regression, neural network, linear support vector machine, RBF support vector machine, and random forest. In machine learning, there are many different classifications methods, though the mentioned are chosen since the authors have received lessons in these methods and these methods cover most of the methods used by Barboza, Kimura & Altman (2017).

The data of the thesis is collected from listed firms in the USA. The reason for this is the possibility of obtaining a big data set for our analysis to get a more valid result. It is easier to collect data from listed firms due to the statutory requirements to publish financial statements. However, the thesis tends to analyse whether accounting data has predictive power to get the perspective from non-listed firms. The number of listed firms in Denmark is not that high, so it was decided to obtain data from the USA. The thesis collected data in the period from 1980 and up until 2015 to get a big data set. The reason for this period is first to obtain a lot of information but also to get it more general and robust. By general and robust it means that at least one business cycle is included in the data which will minimize the risk of just including data in either a recession or a boom.

Furthermore, the thesis is limited to analyse non-financial industries. The reason why financial sectors have been left out is because of the different regulations that are required of them and that they typically have very different accounting ratios compared to non-financial firms. This may lead to difficulties when predicting whether they default or not, which links to the findings from Chava and Jarrow (2004) who argued that financial firms are more challenging to predict default compared to non-financial firms, as mentioned in section 1.3.

Finally, the focus of the thesis will only be taken from the perspective of the credit lenders. This is to keep it simple concerning the analysis as well as to the interpretation.

2. Methodology

In this section, the methodology of the thesis will be described. The first part focuses on the applied theory of science. The second part describes the methodology of the thesis in relation to the data, hereunder the data collection and the quality assessment of the chosen methodology. This section ends in a third part that describes the structure of the thesis.

2.1. Applied Theory of Science Method

The scientific theoretical approach of this thesis depends on its chosen problem. The problem determines how the world are acknowledged. Therefore, it determines how the methodology of the thesis is arranged to solve the research question. The applied theory of science and its scientific theoretical frame is taken the positivistic paradigm as the starting point. The thesis aims at describing a concrete phenomenon concerning default prediction and based on the empiricism as well as the processing; the thesis aims to be able to place specific argumentation to causal connection.

The reason of the positivistic approach is the wish of describing the connections in the world. The positivism contains a realistic ontology. This means that the reality can be found “out there” in its pure form in the shape of legality independent of our acknowledge about it and where you should adjust yourself to fit. It is relevant for the thesis, due to the different conditions when a credit lender steps into a contract with a borrower. The borrower needs to follow the contractual obligations not to default and to determine whether this happens some clear frameworks state that. Furthermore, as mentioned in section 1.4, the data collected has been on listed firms, and these have several conditions and regulations to follow by law. The epistemology of the positivism is objective since the acknowledge of information happens without any consideration to who acknowledges it. This means that science is neutral to politics, religion, and ethics in the view of positivism which is also the case in this thesis. The described ontology and epistemology together result in a quantitative methodology. This can be seen in the thesis, where quantitative data collected and processed is obtained to create tests of different machine learning methods to predict default in firms. This is to plan causal connections that can be transmitted to decisions as choosing the most appropriate machine learning method when predicting one or five years prior to default (Holm, 2016, pp. 23-44).

The positivistic approach has made it possible to structure the problem in the thesis to the issues found in practice and how these problems can be solved. This is done by analysing and calculating with a focus on predicting default and more general use of empirical results. Furthermore, the theory of scientific method approach has made it possible to reflect critically in line with the process of the thesis. This reflection has made the relevance of the neo-positivistic approach clear. The neo-positivistic

approach is taking the basis from the positivistic approach. Though the paradigm of neo-positivism considers the aspect of humans as essential and should not be undervalued. On the other hand, for the positivism paradigm, the human aspect is not taken into account. This means that the ontology of the neo-positivism is limited realistic, and the epistemology is modified objective. The aspect of humans will be used in the different set of problems of the thesis because the quantitative results are not enough when determining the most appropriate machine learning approach. The quantitative testing results may conclude differently than with the view of humans since the measures, such as accuracy or ROC, may not give the full picture of the firm defaults or not. Concretely, the theory of science approach will be a combination of more than one paradigm. This is despite the opinion of some critics that argues about the realism of using more paradigms at the same time (Holm, 2016, pp. 64-74).

2.2. The Methodology of the Paper

The thesis makes use of mainly quantitative methods, as mentioned in section 2.1, where data has been collected to create machine learning methods to predict default within one or five years. The analysis results in quantitative findings that explain patterns by using the inductive procedure. This means using empiricism to produce theory such as using the data to determine which method should be applied to the prediction of default.

The thesis has used a desk study method since the study is done through research. The data collection is based on primary as well as secondary data. The primary data contains an interview with a senior analyst in a credit models department in a credit lender institution. This interview was done at the end of the process of the thesis. After the theoretical results were found, the interview was to create an insight into how to predict default in practice in Denmark. The interview followed a semi-structured approach with open questions. This was to not affect the answers of the interviewee and to let him talk freely about the credit models in their institution. The secondary data, on the other hand, is found from existing sources. It contains quantitative data obtained from Compustat and CRSP as well as qualitative data such as different peer reviews, articles, academic reports, and laws. This quantitative data is directly collected, cleaned, and prepared for this thesis, which is described further in section 2.2.1. To relate it to the theory of scientific method the quantitative part of the secondary data is connected to the positivistic paradigm, where the rest are in a more significant degree connected to the neo-positivistic paradigm where the human aspect is included.

2.2.1. Data Collection

The data used in this thesis is accounting data obtained from Compustat and market data collected from CRSP. As described in section 1.4, the period of interest is from 1980 to 2015. The raw data set includes the necessary accounting and market data, allowing to calculate the 20 selected variables. Fourteen of

the variables are accounting ratios, which only includes accounting data, and the remaining six variables include fully or to some extent market data. These variables are not randomly selected but cover the variables used in Altman (1968), Ohlson (1980), Shumway (2001), Chava and Jarrow (2004), Campbell (2006), and others. For the future prediction of default, these variables will be the explanatory variables. The full list of variables can be seen below and will be elaborated after.

Accounting variables	
Name	**Calculation**
WCTA	Working Capital / Total Assets
RETA	Retained Earnings / Total Assets
EBTA	Earnings Before Interest and Taxes / Total Assets
SLTA	Sales / Total Assets
CACL	Current Asset / Current Liabilities
NITA	Net Income / Total Assets
TLTA	Total Liabilities / Total Assets
FFOTL	Fund from Operation / Total Liabilities
X.NI	Relative change in Net Income $[(NI_t - NI_{t-1})/(\|NI_t\| + \|NI_{t-1}\|)]$
EBITDASL	Earnings Before Interest, Taxes, Depreciation, and Amortization / Sales
OCFTA	Operating Cash Flow / Total Assets
FESL	Financial Expenses / Sales
FDCF	Financial debt / Total Cash Flow
CLTA	Current Liabilities / Total Assets
Market variables	
Name	**Calculation**
METL	Market Capitalization / Total Liabilities
EXRET	Log (Firm return) – Log (value-weighted NYSE, AMEX & Nasdaq return)
RSIZ	Log (Firms Market Capitalization / Total NYSE, AMEX & NASDAQ Market Capitalization)
SIGMA	Monthly volatility over the last year (11-12 months)
NIMETL	Net Income / (Market Capitalization + Total Liabilities)
TLMETL	Total Liabilities / (Market Capitalization + Total Liabilities)

Table 2.1: An overview of the accounting and market variables and how they are calculated

WCTA is one of the variables used in Altman's (1968) Z-score, Ohlson's (1980) O-score, and in previous studies such as Chava and Jarrow (2004). It is a ratio measuring a firm's ability to cover its short-term financial liabilities by comparing the net liquid assets, also called the net working capital, of the firm to its total assets. A positive net working capital may indicate that the firm is able to pay its

short-term obligations and then has the potential to invest and grow. A negative net working capital may indicate that the firm has problems paying back creditors and in worst scenarios, the firm defaults.

RETA is another ratio used in Altman's (1968) Z-score and in previous studies such as Chava and Jarrow (2004). It is a measure of a firm's cumulative profitability over time against its assets. A high or increased RETA indicates the firm is able to continually retain more earnings increasingly. Since this ratio measures the cumulative profitability, the age of the firm is indirectly considered. The reason for this is that relatively young firms will probably get a low RETA because they do not have had time to build up their cumulative profits.

EBTA is another ratio used in Altman's (1968) Z-score as well as in studies such as Chava and Jarrow (2004). The use of EBIT is to focus attention on all income earned by the firm, operating as well as non-operating. This financial ratio compares the EBIT of the firm to its total assets invested in the company, which means it measures the true productivity of the firm's assets independent of any tax or leverage factors. The higher the EBTA ratio, the more profitable and effective is the firm to generate income from its assets.

SLTA, also called the asset turnover, is another measure used in Altman's (1968) Z-score and in previous studies such as Chava and Jarrow (2004). It measures the firm's ability to generate sales from its total assets. The higher the SLTA, the better the firm is to use its assets efficiently.

CACL, also called the current ratio, is one of the variables used in Ohlson's (1980) O-score and in previous studies such as Chava and Jarrow (2004). It measures the firm's ability to pay its short-term obligations. A low CACL indicates that the firm might not be able to pay its bills on time, while a high CACL indicates that the firm has enough cash and other current assets to meet its short term financial obligations.

NITA, also called return on assets (ROA), is used in Ohlson's (1980) O-score as well as in studies such as Chava and Jarrow (2004) and Cambell et al. (2006). The ratio measures how effectively assets are being used for generating profit. This ratio can be argued to give insight into how the earning of the firm is relative to its investments. A high NITA compared to similar firms or to the firm's required rate on return is to prefer since it indicates the firm is earning more on less investment.

TLTA, also called the debt ratio, is another variable in Ohlson's (1980) O-score and it has also been used in studies such as Chava and Jarrow (2004) and Cambell et al. (2006). It measures the financial risk of a firm by determining the proportion of a firm's assets that are financed with the debt of creditors

rather than equity. An increasing TLTA ratio indicates that a firm is either unwilling or unable to pay back its debt which in the end could lead to the firm defaults at some point in the future.

FFOTL is another measure in Ohlson's (1980) O-score. It measures the firm's ability to pay back its debt using only funds from operations. A low FFOTL indicates the FFO of the firm can cover a smaller percentage of the total liabilities which means the firm needs a longer horizon to cover all its total liabilities. A higher FFOTL indicates that the firm is in a stronger position regarding paying back its debt from its operating income, and hence the lower will the firm's credit risk be.

X.NI is a measure used in Ohlson's (1980) O-score. It measures the relative change in net income, indicating the development in net income from one year to another. When X.NI is positive, then the firm's net income is growing and vice versa. This measure gives the credit lender an indication of how the trend is in net income for the specific firm.

EBITDASL is a measure comparing earnings before interests, taxes, depreciation, and amortizations with its revenue to evaluate a firm's profitability. This ratio is most preferably when it is high, as it indicates the firm's ability to keep its earnings at a decent level by keeping certain expenses low.

OCFTA measures the firm's ability to generate operating cash flow from its total assets which means the amount of operating cash flow the firm generates for every dollar of assets invested in the company. The higher the OCFTA, the more efficiently the firm uses the assets.

FESL measures the proportion of the financial expenses constitute of the sales. The ratio is preferred when it is low as it indicates the sales of the firm is greater than the financial expenses.

FDCF measures the firm's ability to cover its financial debt by the cash flow of the firm. A ratio over 1 indicates that the total financial debt is higher than the cash flow of the firm from the given year. All things being equal, it is preferred to have a ratio as low as possible.

CLTA measures the firm's ability to cover its short-term financial obligations by comparing the firm's current liabilities with its total assets. Therefore, the higher the CLTA, the greater the risk of the firm defaults.

METL is a measure used in Altman's (1968) Z-score as well as in studies such as Chava and Jarrow (2004). It includes an accounting variable as well as a market variable. This ratio measures how many times the market value of the firm exceeds the total liabilities. A higher ratio is preferred, all things being equal.

EXRET is a measure of the excess return of the firm. This measure has been used in Chava and Jarrow (2004) and in Cambell et al. (2006). It measures the market return of the firm in comparison to the market. Therefore, this measure indicates how the firm is performing compared to the given index.

RSIZ is a measure of the relative size of the firm. This measure has been used in Chava and Jarrow (2004) and Cambell et al. (2006). It measures the size of the firm compared to the market. The higher the firm's market capitalization compared to the total market capitalization, the more likely it is for the firm to outstand problems that may arise in the future.

SIGMA is a measure of the volatility of the firm's stock. This was used in Chava and Jarrow (2004) and Cambell et al. (2006). It is measured by taking the monthly stock volatility over the last year (11-12 months), and it indicates how risky the firm is. The higher the volatility, the higher the risk of the firm defaults.

NIMETL is used in the paper of Campbell et al. (2006). It is a similar measure to NITA. The difference is in NITA net income is divided by the book value of total assets which is the sum of the book value of equity and the book value of liabilities, whereas in NIMETL net income is divided by the sum of the market value of equity and the book value of liabilities.

TLMETL is a measure used in the study of Campbell et al. (2006). It is a similar measure to TLTA. The difference is that in TLTA total liabilities are divided by the book value of total assets, whereas in TLMETL total liabilities are divided by the sum of the market value of equity and the book value of liabilities. TLMETL measures the per cent of a firm's valuation that is made of liabilities.

The next part describes how the process has been from having the raw data from the databases to the final data, which is used in the rest of the thesis. This process is typically illustrated by a continuous process which starts with the raw data and then moves forward to data cleaning, data preparation, and afterwards, some kind of modelling of the data. However, the process for this thesis was more like a cycle where the data was cleaned, prepared, and afterwards evaluated several times.

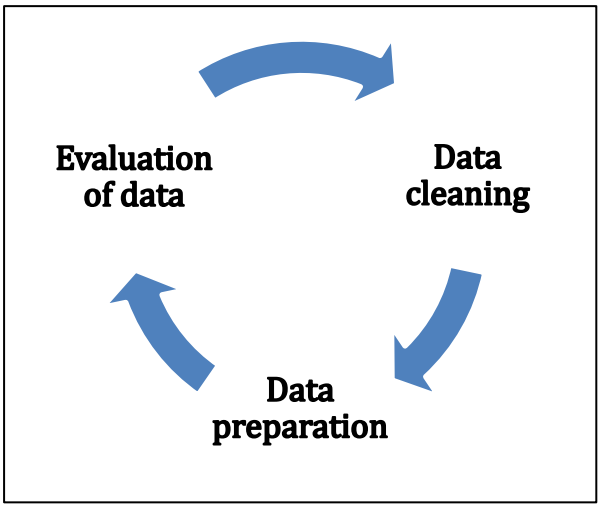


Figure 2.1: Illustration of the data processing

2.2.1.1. Data Cleaning

Before data cleaning, the total number of observations was over 193,000. The first thing to do was to remove all missing observations. It was found that the variable OCFTA had many empty fields because the number for operating cash flow was not reported before 1987. Therefore, it was decided to begin the data period in 1987.

The raw data included a code for the industry, which is called the standard industrial classification (SIC-code). The SIC-code can be divided into ten different industry classes (NAICS Association, n.d.). The split between the ten different industries can be seen in appendix A figure A1. As mentioned in section 1.4, it was decided to remove all observations within the category of financial firms. In addition, it was also decided to remove observations within the industry of "Agriculture, Forestry and Fishing", "Construction", and "Other". The argument here was that these three industries combined only account for around 1.7% of the total observations, and it would be simpler if the number of industries was reduced. Finally, the industry "Retail Trade" and "Wholesale Trade" were combined in one industry called "Retail & Wholesale Trade". Figure 2.2 shows the final split of the industries.

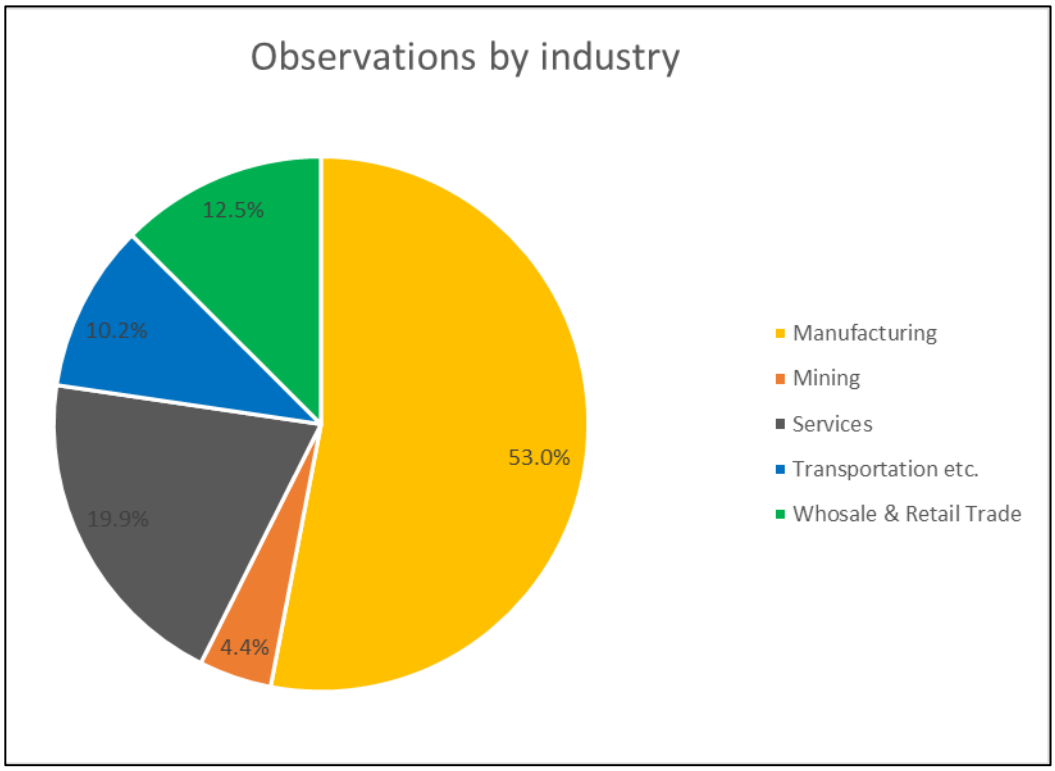


Figure 2.2: The distribution of the observations divided into industries after the cleaning process

The last process in the cleaning part was to remove all firms, only having one observation. Some of the variables, such as the change in NI, requires two observations from different periods from the same firm to be able to be calculated. If the variables, calculated on behalf of more observations, could not give a valid result, the observation was deleted.

The processing of data cleaning results in a data set with over 92,000 observations including more than 10,900 firms. Among those, 1,344 firms file for bankruptcy which equals 12.33% of the firms. However, the firms that are going to default also have several fiscal years in which they do not default. These fiscal years are also included in the data. Overall, it is only 1,46% of the observations that are going to default within the next year and 5,33% of the observations that are going to default within the next five years. These numbers show how unbalanced the data is regarding "default" or "non-default" before the following cleaning step.

It was chosen to get a balanced training set in terms of "default" and "non-default". This training set is used to build a model that should be tested on a realistic unbalanced testing set which is typically done in data science. The split of the data set belongs to the process of data preparation, why it will be described in section 2.2.1.2. The argument for the balanced training set is supported by the complications when running machine learning methods such as support vector machine on a training set containing all available data. To create a balanced testing set, all "default" observations were kept

and matched by industry and year with "non-default" observations. This procedure follows Barboza et al. (2017). It resulted in the same amount of observations in the category "default" as in "non-default".

2.2.1.2. Data Preparation

The first step in the preparation of the data was to calculate the chosen variables. Table 2.1 shows all the variables and how these are calculated. Some of the raw data did not have the exact information needed, e.g. the data did not contain information about FFO. A proxy was calculated for these variables to get a good estimation. A list of the variables where a proxy has been estimated can be seen below.

$$FFO = Net\ income + \ depreciation\ and\ amortization$$

$$Financial\ expenses = Interest\ and\ related\ expenses$$

$$Financial\ debt = debt\ in\ current\ liabilities, total + long\ term\ debt, total$$

$$CF = operating\ income\ after\ depreciation - interest\ and\ related\ expenses - income\ taxes, total - dividends\ common/ordinary$$

The data contained information about which firms default within one and five years as well as a bankruptcy date which defines the time the firm defaults. Many firms had a bankruptcy date there were more than one year after the last fiscal year of the firm which means that the referred firm did not appear with "default" in the field stating whether the firm defaults within the next year or not. Of this reason, the same field was changed to be calculated on behalf of the bankruptcy date and the last fiscal year of the firm. This methodology resulted in any firms with a bankruptcy date had a "default" in the field for the last fiscal period. The same method was applied to the field, stating whether the firm defaults within five years or not. This means a defaulted firm might have up to five fiscal periods with "default" in the field. However, some of the defaulted firms did not record for the whole time period of five years before they defaulted, which means that the average number of fiscal years was lower than five.

To be able to calculate the variables EXRET and RSIZ, additional market data was needed. This data was not initially available in the data set. However, it was possible to extract the market return and the market capitalization from CRSP for the period 1980-2015. After that, the procedure was to match the return and market capitalization to the correct fiscal year for each observation and then calculate EXRET and RSIZ.

When creating classification models, it is essential to make models that are robust so they can be used in the future. This study follows the normal procedure to split the data into a training and testing data set. The training set is from 1987, as the data from 1980 to 1986 was deleted in the data cleaning process, to 2005, while the testing set is from 2006 to 2015. This split makes the testing result more reliable compared to just testing on the same data, and it results in both the training and the testing set containing at least one business cycle.

Feature scaling is the method in the preparation process to scale the variables into the same range. The reason to do so is the risk of a very large difference in the values for the variables. For instance, if we compare the variables WCTA and METL in the data set. WCTA is measured on a scale from -11.6 to 1.0, while the values for METL are measured on a scale from 2.2 to 1,537,071. The scaling of all the explanatory variables can be seen in appendix A figure A2. This means that most machine learning methods will give the most focus to the variable with the widest scaling. In addition, some machine learning methods calculate a distance and this distance will be dominated by the variables with the widest scaling. When doing feature scaling, there are mainly two different methods which are normalization and standardization. Normalization has been chosen which uses the following formula:

$$x_{norm}^{(i)} = \frac{x^i - x_{min}}{x_{max} - x_{min}}$$

By applying this formula to all ratios for each observation, all explanatory variables get values between 0 and 1. This will cause different variables to be equally weighted in the models.

2.2.2. Quality Assessment

The quality assessment is evaluated on validity, reliability and sufficiency. Validity is an assessment of what the data and the result can be used to or what it covers (Olsen, 2003). The data has been collected from 1980 to 2015. This is a long period taking different decades, hereunder various fluctuations, and several business cycles into account. It is as up-to-date as possible from the platforms it is obtained from. Though later access to data would be more preferably. Despite this, the validity is high since the thesis success to measure what it wanted to.

As mentioned, the primary method is the quantitative approach which demands high reliability. Reliability is about the robustness of the data in relation to the way it is collected. It means that the test returns in the same outcome on repeated tests (Olsen, 2003; Carmines & Zeller, 1979/2011). This is the case because the relevant data is obtained, as mentioned, from Compustat and CRSP. These organisations have high credibility because they collect information listed firms have reported at a platform. This increases data reliability. Though, some information could not be found at the platforms as was elaborated in section 2.2.1. Overall, the data collected will be defined as highly reliable.

Finally, sufficiency is a question of whether the test with its sub-questions is suitable to answer the research question (Olsen, 2003). By answering the research question, machine learning methods have been trained to give the most precise models for each method. These models have been evaluated based on the two evaluation measures, the accuracy and the AUC. Then, some methods can use further variable selection to determine which variables are most important for the model. Furthermore, to evaluate default methods in terms of credit risk management, it is investigated how credit lender

calculate their credit risk, and why they prefer credit risk model which can fit into the regulations for the IRB approach.

2.3. The Structure of the Paper

The thesis has been divided into seven different sections. Section 1 is the introduction of the thesis, including the motivation of writing in this field, the research question, the literature review, and the delimitation. Section 2 determines the methodology of the paper, hereunder the applied theory of science, and the data collection. The latter has great importance in this thesis since the data is used to build machine learnings models to determine the most appropriate default prediction model for the different data sets. Section 3 describes the theory used in the thesis. It includes the risk of management, among these the regulation in credit risk management and a general description of the machine learning models for predicting default. Section 4 determines and analysis the empirical results which are done by building machine learning models based on a training data set and then test the model using a testing data set. This section is separated into four parts regarding the four different data sets. Every part is finished by an analysis of which model performs best on the given data set. Section 5 is a comparison of the models found in section 4. The first part of this section includes a general comparison of the models by using the evaluation measures. It is followed by a discussion of these measures for evaluating the models. The second part includes an analysis and discussion of variable selection, especially with the focus of logistic regression and random forest. The third part includes a small discussion of the predictive power of market variables in relation to predicting default for non-listed firms. Section 6 takes the knowledge from the previous sections and transfers it into the credit risk management of credit lenders on the Danish market. It includes a part with some of the regulations on the Danish market, as well the internal advantages of a better credit rating model. The second part discusses the result of the thesis in relation to the literature and whether it is possible to move away from logistic regression as the industry benchmark. Finally, section 7 concludes and put the thesis into perspective.

3. Risk Management

Risk management, concerning finance, is a process of identifying, analysing, accepting or modifying uncertain situations for the firm here with the focus on credit lenders offering loans to firms. The purpose of risk management is to quantify the potential risk for loss and then act in the best interest of the firm with that knowledge. The credit analysis is a central part of the risk management for credit lenders. In general, there are two approaches to analyse credit. The first one is to look at historical data from the borrower as well as the credit portfolio and revealing insight from it. The second one is to simulate the expected cash flow from the borrower and on behalf of that analyse the borrower's creditworthiness. There are estimations and judgment in both methods that can be inaccurate, but this thesis will focus only on the first approach.

The following consists of two parts. The first part contains a more general view of risk management, including some chosen part of the regulation as well as how to determine credit risk. The second part includes different methods in determining the probability of default as a measure which is used in risk management. This part is separated in two and includes 1) already existing default prediction methods such as univariate analysis, Z-score, O-score and credit ratings, and 2) classification methods in machine learning such as logistic regression, neural network, support vector machine, and random forest.

3.1. Regulation in Risk Management

Firms within the finance sector are governed by several different regulations. This prevails especially for credit lenders that are subject to requirements, restrictions, and guidelines to create transparency between the credit lender institution and the borrower. For most countries, there is a local regulation for the credit lenders to follow. However, these regulations are mainly built upon the Basel Accords, which sets the international recommendations for credit lender regulation.

The present international regulation is Basel II which is a three-pillar system. The first pillar is the requirement and regulation concerning capital, risk coverage, and leverage. The second pillar concerns the supervisory review process. The third pillar concerns market discipline. This thesis focuses on the first pillar and more specific the risk coverage among these credit risks and how to calculate it (Bank for International Settlements). In Basel II there are two main methods to calculate the credit risk, which are the standardized approach and the Internal Rating-Based (IRB) approach. The standardized approach assigns some risk weights to different credit lenders. These weights typically come from the rating assigned by external rating agencies such as S&P. As table 3.1 shows, the better rating the borrower has achieved from the rating agencies; the smaller weight is assigned to the credit risk for the

credit lender. If the firm does not have a rating, the standard weight will be 100% under the standardized approach. The second approach, the IRB, relies on the internal estimates concerning credit risk for the credit lenders. The component that should be calculated is the probability of default (PD), the loss given default (LGD), exposure at default (EAD), and the maturity of the loan. These four components, combined with the correlation between corporate exposure, can be calculated into the credit risk (Bank for International Settlements, 2019a).

The table provided below illustrates the risk weighting of rated corporate claims, including claims on insurance companies. The standard risk weight for unrated claims on corporates will be 100%. No claim on an unrated corporate may be given a risk weight preferential to that assigned to its sovereign of incorporation.

Credit assessment	AAA to AA-	A+ to A-	BBB+ to BB-	Below BB-	Unrated
Risk weight	20%	50%	100%	150%	100%

Table 3.1: The table from Basel accords for determining risk weights under the standard approach

Source: CRE 20.17

3.1.1. Credit Risk

According to Basel Committee on Banking Supervision (2000), the definition of credit risk is: "Credit risk is most simply defined as the potential that a bank borrower or counterparty will fail to meet its obligations in accordance with agreed terms." In other words, the credit risk is the risk that the loans are not being repaid to the full extent, which will imply a loss for the credit lender. Due to regulation and the risk of losing money to bad lenders, the credit lender has an obligation and incentive to analyse and measure the credit risk. For banks and other credit lenders, this includes estimating the PD, EAD, and LGD. These two estimates seek to evaluate the risk of the borrower not being able to meet its contractual obligation as well as to evaluate how much the credit lender risk of losing in case the borrower defaults.

The PD is very likely the most important question in the analysis of the credit risk. It is essential to have a valid estimate for all loans, but it might also be the most difficult estimate to make. Sometimes external conditions interrupt the models which previously gave a good estimate for the PD. However, models calculated on historical data are still the most frequent methods to calculate the PD. This will also be the method used in this thesis. Some of the models used in the past, and new machine learning methods to predict default will be described in section 3.2.

The second thing the credit lender needs to estimate in the analysis of credit risk is the potential losses in case a given firm file for bankruptcy. This is known as the LGD which is dependent on two factors namely the exposure at default (EAD) and the recovery rate.

First of all, the recovery rate is highly affected by what kind of security the credit lender has in the asset from the company. Bank loans are typically safer than other kinds of debt, like bonds, because the loans have collateral before the bonds in the liquidation order. Another implication for the recovery rate is what type of asset the credit lender has security in. A large proportion of financial assets like cash and stocks will imply that the recovery rate will be high. In case the assets are primarily material assets, then the recovery rate will drop to a medium level. However, if the defaulted firm has a lot of intangible assets, then these assets might be worthless or sold for a very small amount in proportion to what it is worth in ”the books”. All these things are worth considering when a credit lender should determine the recovery rate for a borrower. The EAD is calculated by the credit lender by taking the outstanding of the principal amount. This means, other things being equal, that a loan with a longer term to maturity will have a higher EAD compared to a loan with a shorter term to maturity. Finally, the LGD can be calculated as the EAD minus the recovery rate for default (Petersen & Plenborg, 2012, pp. 271-297).

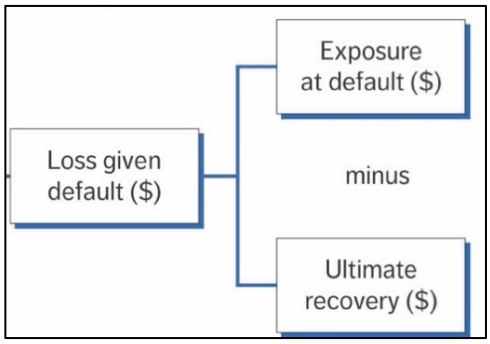


Figure 3.1: How to calculate loss given default

The thesis focuses on predicting whether a company defaults or not. Therefore, only one component from above will be elaborated, namely the PD. The next section includes two parts, a definition of already existing default prediction methods as well as different machine learning methods for predicting whether a company defaults or not.

3.2. Predicting Default

3.2.1. Default Prediction Methods

How to measure and analyse the riskiness of the loans is something that has been elaborated extensively over time. It has not only been directly on predicting the PD, but the essence has been the same, which is giving a valid estimation of how risky the loans of the credit lenders are. Several methods have been

proposed over time. The methods can be categorized as statistical methods that can be used for the prediction of corporate defaults by using financial ratios. Default prediction methods allow credit lenders to analyse a large number of firms fairly quickly and at a low cost. There are different types of default prediction methods used for this purpose that will be elaborated below among others univariate analyses, Altman's (1968) Z-score, Ohlson's (1980) O-score, and credit rating.

3.2.1.1. Univariate Analysis

A univariate analysis is probably the simplest way of analysing financial data. It analyses the predictive ability of ratios one at a time. Beaver (1966) is one of the first to study this area. He had 30 ratios which were divided into six "common element" groups where only one ratio from each group was selected as a focus for the analysis. The purpose was to see which ratios could predict failure and how many years in advance, the forecast could be made. The chosen ratios were cash flow to total debt, net income to total assets, total debt to total assets, working capital to total assets, current ratio, and the no-credit interval which is defined as defensive assets minus current liabilities to fund expenditures for operations.

A comparison of mean values of the ratios, called the profile analysis, was computed for the failed firms as well as for those of comparable firms that did not fail. This was computed for a period five years prior to default. Figure 3.2 shows the level and trend in the six ratios are poor for firm failing relative to nonfailing firms. The profile analysis is not a predictive test but rather a convenient way of outlining the general relationship between the failed and non-failed firms. Univariate predictions may end up giving different forecasts for the same firm depending on the chosen ratios. By using a multivariate approach, using several ratios, this may outcome the problem (Beaver, 1966).

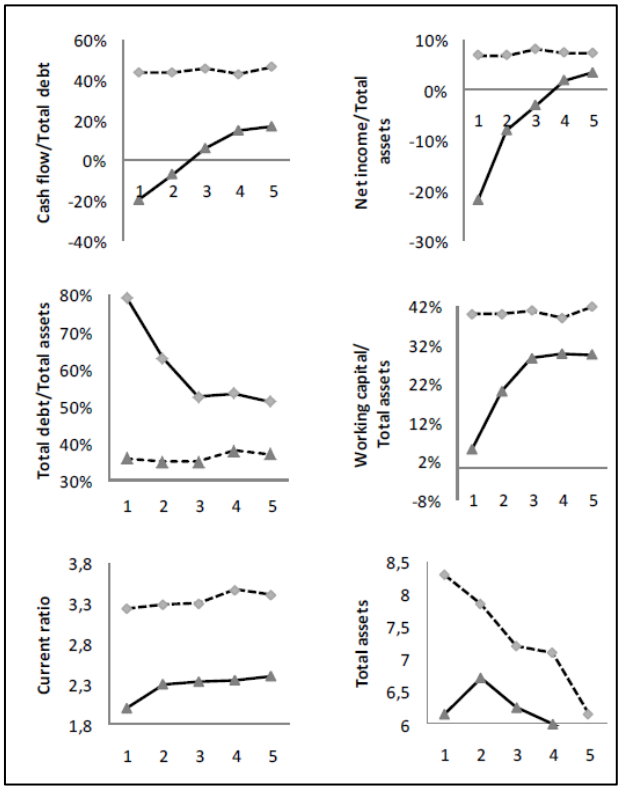


Figure 3.2: Comparison of mean values of six selected financial ratios for bankruptcy firms and non-bankruptcy firms five years prior to bankruptcy

Source: Beaver (1966)

3.2.1.2. Z-score – a Multi Discriminant Analysis

Multivariate prediction models add several ratios together to end up with a score. One of the first to do so in the field of default prediction was Altman (1968) with his Z-score, which is well-known all over the world and is still applied and taught today. The Z-score is a multiple discriminant analysis (MDA) used to classify public manufacturing firms into bankruptcy or non-bankruptcy groups. The method derives a linear combination of five financial ratios which in Altman's (1968) opinion best distinguish between "default" and "non-default" firms. It uses profitability, leverage, liquidity, solvency, and efficiency to predict whether a firm has a high risk of going default. The Z-score looks as follows:

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1.0X_5$$

where

- $Z = overall\ index$
- $X_1 = Working\ capital/Total\ assets$
- $X_2 = Retained\ earnings/Total\ assets$
- $X_3 = EBIT/Total\ assets$
- $X_4 = Market\ value\ of\ equity/Book\ value\ of\ total\ liabilities$
- $X_5 = Sales/Total\ assets$

Altman (1968) identified two cut-off points, 1.81 and 2.99. A Z-score higher than 2.99 indicates a low risk of default, whereas a Z-score below 1.81 indicates a high risk of default. A Z-score in the interval between 1.81 and 2.99 is defined as a grey area which leads to further analysis. Since its publishment in 1968 has the Z-score been revaluated by Altman several times by analysing other firms in other time periods.

3.2.1.3. O-score – a Logit Analysis

In MDA models, the standardized coefficients cannot be interpreted as the slopes of a regression, and therefore it does not indicate the relative importance of the different variables. With these difficulties of MDA models in mind, Ohlson (1980) applied, as one the first, the logistic regression model to the studies of default prediction. The logit regression model is used to estimate the probability of default based on several predictor variables. In practice, the benefit of the logit model is that it does not need the restrictive assumptions that are for the MDA method. Assumptions such as the independent variables must be normally distributed, and the variance-covariance matrices should be equal between the default and non-default groups. This is often violated when applying to default prediction problems. Furthermore, logit models allow working with disproportional samples (Sabato, 2008). The logit model fits with the problems of predicting default with a binary dependent variable. A more general description of logistic regression will be done in section 3.2.2.1. The O-score defines the probability of default as:

$$Probability\ of\ default = \frac{1}{1 + e^{-y}}$$

Where e is the base of the natural logarithms, and y is found by the following formula:

$$y = -1{,}32 - 0{,}407Size + 6{,}03\frac{TL_t}{TA_t} - 1{,}43\frac{CA_t - CL_t}{TA_t} + 0{,}076\frac{CL_t}{CA_t} - 2{,}37\frac{NI_t}{TA_t} - 1{,}33\frac{FFO_t}{TL_t}$$
$$+ 0{,}285Z - 1{,}72X - 0{,}521\frac{NI_t - NI_{t-1}}{|NI_t| + |NI_{t-1}|}$$

Size is equal to the log of total assets time t divided by Gross National Product price index level, TL is equal to total liabilities, and TA is equal to total assets. CA is equal to current assets, CL is equal to current liabilities, NI is equal to net income, and FFO is equal to funds from operations. Z and X are two dummy variables, where Z equals 1 if a net income is negative for the last two years and 0 otherwise, and X equals 1 if total liabilities exceed total assets and 0 otherwise. The score results in a number between 0 and 1 where any value bigger than 0.5 implies that the firm is going to default within one year (Ohlson, 1980, pp. 119-125).

3.2.1.4. Credit Rating

A credit rating system is a calculated evaluation of the creditworthiness of a borrower – the likelihood that the borrower will be able and willing to meet its financial commitments and to pay back the loan within the terms of the loan agreement. There exist several different credit rating systems all other the world but because this paper focuses only on corporations, the credit score systems for individuals and

others are not taken into account. The credit rating for determining the creditworthiness of a corporation is generally done by a credit rating agency such as Standard & Poor's (S&P), Moody's, or Fitch. The idea is the same for all the different credit rating systems – to rank the firm concerning its creditworthiness. Though, the different credit rating agencies do not necessarily have the same interpretation of this creditworthiness. S&P defines creditworthiness as only the probability of default where nothing else matters. This means that the recovery rate, the proportion the credit lenders end up with after the borrower has defaulted, has not been taken into account. Moody's, on the other hand, defines creditworthiness as the expected losses. Expected losses include the probability of default as one part, but it also includes EAD and recovery rate as the second part. This means that one firm may have a good credit rating according to S&P but a poor one according to Moody's (Salmon, 2011). These credit ratings affect the opportunity for the firm to be approved for a loan and might also affect the interest rate on loans. For the agencies which assign a rating to the firm, different aspects are looked at where each aspect is given a subjective weight to its importance. The agency, first of all, considers the firm's history regarding debt and paying off its debts. If the firm has a bad history of paying off debts, then the rating will be affected negatively. Second, the agency also considers the firm's future economic potential. In case it looks bright, the rating tends to be high, whereas if the future economic potential does not look too positive the rating will fall (Petersen & Plenborg, 2012, pp. 271-297).

S&P has its S&P Global Ratings which is a forward-looking opinion about the creditworthiness. It contains letters, numbers, words, or combinations of these in each rating scale to summarize its opinion. The general-purpose credit rating, referred to as the "traditional" credit rating, can be either short-term or long-term. Short-term issue credit ratings are usually appointed to those obligations considered as short-term in the relevant market commonly up to 1 year, and long-term issue credit ratings are assigned to the rest. The short-term issue credit ratings consist of six steps going from A-1 (excellent) to C before ending up at D, where the latter indicates the firm defaults on one or more of its financial obligations. See appendix B table B1 to see the table with definitions. The long-term issue credit ratings, on the other hand, consists of nine steps only containing letters starting at AAA (excellent) to C and then D, where the latter again indicates the firm defaults on one or more of its financial obligations. See appendix B table B2 for definitions (Standard & Poor's, 2019).

3.2.2. Machine Learning

Machine learning is an application of artificial intelligence where it is about obtaining knowledge from data by measuring patterns in the data of the same category and identifying features that separate the data into dissimilar groups. Systems that can learn from data in a manner of being trained are designed by computing. With both time and experience, the systems may learn and improve without being explicitly programmed. Machine learning methods have been used across a wide range of research fields

among others medicine, engineering, advertising, and predicting bankruptcy (Barboza, Kimura, & Altman, 2017; Bell, 2014) where the latter will be the focus of this paper.

The machine learning algorithm is either supervised or unsupervised learning, where the latter is not relevant for this thesis. Supervised learning is when working with a set of labelled data which means that each data point has a class. In this case, the classes are "default" or "non-default" which are used to classify the new data points to either one of them. So, for every observation in the data, you have an input as well as an output object. The data set is split into a training set and a testing set. A testing set should be used to test the algorithm developed from the training set (Bell, 2014). Below will the following supervised classification models; logistic regression, neural network, support vector machine and random forest be introduced. It is the same methods that afterwards will be tested to predict default.

3.2.2.1. Logistic Regression

Logistic regression is a predictive analysis method from the field of statistics that is borrowed by machine learning. It is used in the original form for binary classification problems – whether the firm defaults or not defaults given a set of explanatory variables. It models the probability of belonging in a given class, and therefore the final result of each observation in the logistic regression model should be between 0 and 1. So, it used a logistic sigmoid function which is characterized by an S-shaped curve. The sigmoid function transforms high negative numbers into numbers close to 0 and high positive numbers close to 1, which is illustrated in figure 3.3. In addition to that, the sigmoid function intercepts the y-axis at 0.5, meaning a 50% probability of default. The full logistic regression function inclusive the sigmoid transformation with k explanatory variables can be written as

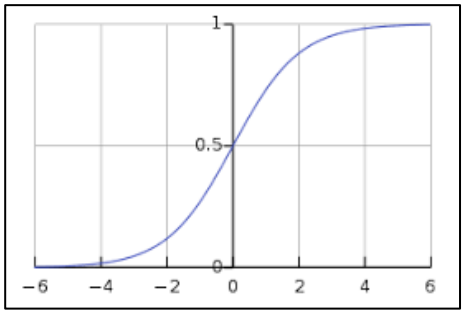


Figure 3.3: The sigmoid function

$$\hat{p} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_k x_k)}}$$

where $\hat{p}$ is the predicted output, e is the base of the natural logarithms, β_0 is the intercept term, and β_k is the coefficient for each input x_k. The coefficients β_0 and β_k must be estimated from the training data set by using the maximum likelihood estimation. The intuition of the maximum likelihood for logistic regression is to pursue values for the coefficients that minimize the error in the probabilities predicted by the model to optimize the best values from the training data set. It is done by the log-likelihood measure (Baesens B. , 2014, pp. 39-42). Another measure to determine the best model in logistic regression is the AIC. This is an approach used for model selection. It assumes no model is precise and therefore, the goal is to find the one closest to the true model. AIC is relative to other such measures which means it can only be used for model selection when the models are estimated on the same data set. It estimates the relative amount of information that is lost by a given model where the less

information lost by a model indicates a better model. The lower value of AIC the better model according to the measures (Sakamoto, Ishiguro , & Kitagawa, 1986). Ohlson (1980), with his O-score, is an example of the use of logistic regression when predicting default.

Logistic regression models are at the risk of being affected by multicollinearity if some of the explanatory variables are highly correlated. This can cause some of the variables to be insignificant and to have a wrong sign in the coefficient if the variable is strongly correlated with another variable in the model. One way to solve it can be by making a correlation matrix and then exclude the variables that are insignificant and correlated (Hastie, Tibshirani, & Friedman, 2009, pp. 122-124).

After the model has calculated all the weights for the variables, the logistic regression model should be tested. Here there is a need for a cut-off that separates the classes "default" or "non-default". Typically, it is decided to use 0.5 as the cut-off that separates the classes, but other alternatives can also be used (Swaminathan , 2018).

3.2.2.2. Neural Network

Neural network was created back in the 1940s, and it was the first method to classify on a larger scale. The theory has evolved multiple times, and today is artificial neural network powered by deep learning algorithms state of the art when it comes to image and speech recognition. Neural network has its inspiration from the human brain, which consists of approximately 100 billion neurons that are connected in a network (Freudenrich & Robynne, n.d.). Similarly, neural network has some input neurons that are connected in a network with several neurons in the hidden layers which decide the result of the output neurons. This output will be the result of the classification, e.g. result in either "default" or "non-default".

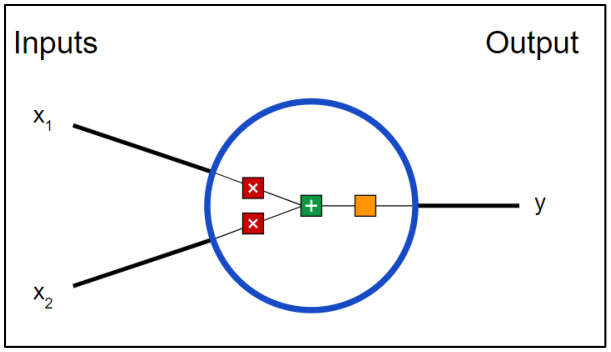


Figure 3.4: The process from the input neurons through a network with several neurons in the hidden layers, which decide the result for the output neurons

Neural network has a black box where several hidden layers occur each with several neurons within. It is through these hidden layers and neurons the decisions whether the classification results in "default"

or “non-default” are made. It will, in the following, be elaborated on what happens inside these hidden layers and neurons.

The input variables will be the 20 or 14 different variables described in section 2.2.1. In the neural network model, these input variables will be more or less activated dependent on what their values are. The range between 0 and 1 where 1 is fully activated and 0 is not activated at all. The activation is crucial in neural network because the output neuron should either activate the “default” or “non-default” neuron. The input variables are all being attached to a weight. The weight is a number that is being multiplied by the activation number of the input variable. This weighted sum gives a number of activation, which could be both positive and negative numbers. However, neural network wants to limit this activation number of the neuron to be between 0 and 1. Therefore a sigmoid function is added, as also done in logistic regression, to get the probabilistic output that maps the weighted sum into the range between 0 and 1. The last step to be able to calculate the next neuron is to apply a bias. The bias unit gives a threshold for when the neuron should be activated. The formula for calculating the next neuron will be:

$$\hat{y} = \sigma(w_1\alpha_1 + w_2\alpha_2 + w_3\alpha_3 \dots w_n\alpha_n + b)$$

where σ is the sigmoid function, w is the weight assigned to the input neuron, α is the activation number of the input neuron, and b is the bias. This function is calculated for all neurons in the hidden layer. Assuming there are two hidden layers, the result of the first hidden layer will impact the degree of activation for the second hidden layer. The last hidden layer will determine the activation of the output layer and decide whether the observation is being classified as a “default” or “non-default” (Amini, 2020). A visualisation of a network with two hidden layers can be seen in figure 3.5.

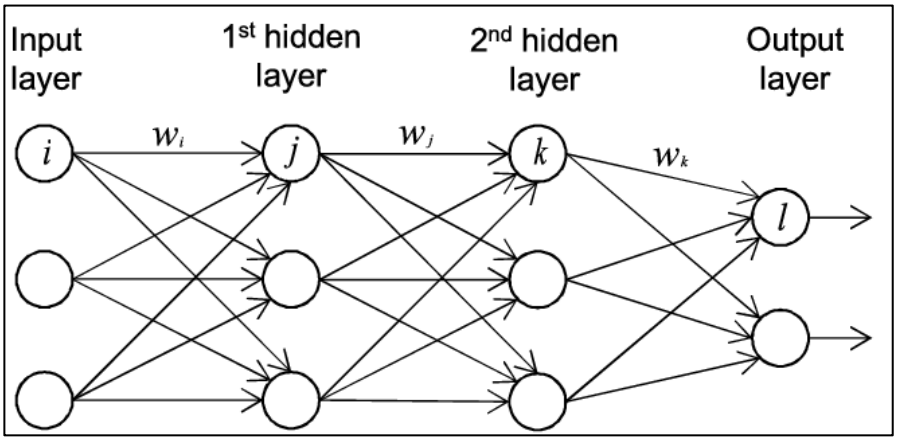


Figure 3.5: The process from the input layer to the output layer through a number of hidden layers

Training of Neural Network

When training a machine learning model, it is common to have a function to either minimize or maximize a factor. For neural network the cost function is the function that should be minimized. For every observation, the output layer gives an activation degree between 0 and 1 for both “default” and

"non-default". If the real class is "default" then it should have 1 in "default" and 0 in "non-default". The full cost function for the whole network can be written as:

$$Cost = \frac{1}{n}\sum_{i=1}^{n}\left(f\left(x^i; W\right), y^i\right)$$

The term for the predicted output is $f\left(x^i; W\right)$, while y^i is the actual output for the observation. The cost function then finds the average difference between what the neural network predicts and what the actual value is. This average should then be minimized by changing the attached weights and bias. Neural network seeks the minimum of the cost function with the use of a minimization function which can be gradient descent. However, the global minimum can be hard to find if the cost function is complicated and has more local minimums (Yiu, 2019).

To reduce the computational calculation in trying to find the minimum of the cost function, an algorithm called backpropagation is used in neural network. Simplified, this algorithm starts from the right side of the network, the output neurons, and move to the left until it hits the input neurons. Backpropagation looks at how the neurons of the last hidden layer should change to correctly classify the observation. This can be done by either changing the bias, changing the weight, or change the activation of the neuron in the hidden layer. This activation is a function of the previously hidden layer which means we then go back one layer and make this process again. This method is backpropagation, also called partial derivative, and will help the program to faster find the minimum of the cost function (Raschka & Mirjalili, 2017, pp. 412-417). In addition, a term called minibatch can be used to reduce computational calculation even further. The training data is divided into several parts, known as batches, where each batch has for instance 32 observations which is the most common amount of observations in neural network in R. Instead of running through the training set at once and changing the weights and bias on behalf of the whole training set the mini batches are used. One time a batch has passed through the network and changing the weights and biases, it is called an iteration. When the whole training data has passed through the network, it is called an epoch. The number of epochs is a parameter to tune in neural network because a higher number of epochs will fit the model closer to the training data (Sharma, 2017).

3.2.2.3. Support Vector Machine

The support vector machine has like the neural network a black box where the classification is made. The support vector machine aims at splitting classes with a hyperplane which in the primal version is written as

$$H0: w^T x + b = 0$$

Where w are the weights on features and x is the data points (support vectors). The objective of the hyperplane is to maximize the distance to the nearest training data point of any class to minimize the risk of misclassification. This is done by solving an optimization problem over w, known as the primal problem:

$$minimize_{w,b,\xi} \frac{1}{2} \sum_{j=1}^{N} w_j^2$$

where N is the size of the training data. At the nearest training data point of each class, another hyperplane occurs, which means there is a hyperplane placed on each side of the main hyperplane. The two hyperplanes are written as:

$$H1: w^T x + b = +1$$
$$H2: w^T x + b = -1$$

H1 indicates the hyperplane at the edge of class 2, where H2 indicates the hyperplane at the edge of class 1 as can be seen in figure 3.6. The distance from the first hyperplane, H1, to the origin equals $|b-1|/\|w\|$ where $\|w\|$ represents the Euclidean norm of w which is calculated as $\|w\| = \sqrt{w_1^2 + w_2^2}$. Similarly, the distance from the second hyperplane, H2, to the origin equals $|b+1|/\|w\|$. The goal is to get a function that returns +1 if the result of the function is positive which shows the data point is in one class and it returns -1 when the point is in the other class.

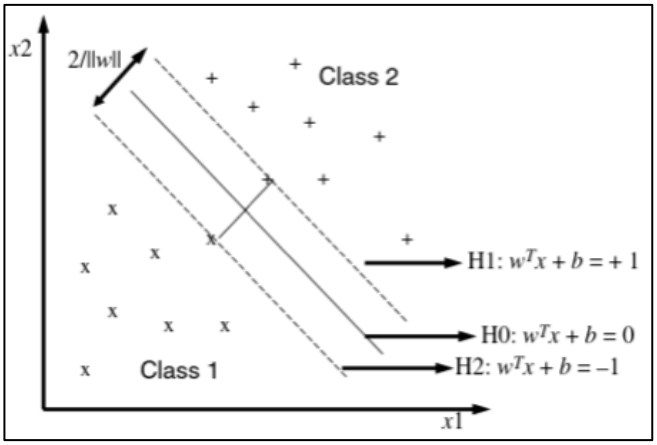


Figure 3.6: The three hyperplanes in SVM

The data points (vectors) that define the hyperplane are called the support vectors, and the distance between these and the hyperplane is called the margin. It is a technique that can be done with either a hard or soft margin. The difference between the two margins is the strictness of correct classifications. The soft margin allows not all individuals to be correctly classified, whereas this is not allowed for the hard margin. The strictness of the hard margin leads to the risk of overfitting the training data because of no flexibility to do misclassifications. It is known that economic variables are influenced by noise in empirical data and are often biased. This is the reason why the soft margin is regularly used. Therefore, when using the soft margin, the support vectors that define the hyperplane are those data points within the margin on the correct as well as on the wrong side of the hyperplane. The number of support vectors depends on how much misclassification is allowed. Allowing a large number of misclassifications will give a large number of support vectors and vice versa. Those data points on the correct side of the hyperplane but within the margin are support vectors that are correctly classified, but those data points on the wrong side of the hyperplane within the margin are support vectors that are misclassified.

Therefore, a large number of support vectors indicate a risk of a large number of misclassifications, and a low number of support vectors indicate a chance of a lower number of misclassifications. The soft margin will be used for this thesis to get a more robust model by adding an error term to the optimising problem over w:

$$minimize_{w,b,\xi} \frac{1}{2}\sum_{j=1}^{N} w_j^2 + C\sum_{i=1}^{n} e_i$$

C (cost) is the trade-off parameter between maximizing the margin and minimizing the error on the data. The larger the C, the more misclassification in the training set will be penalized. The last one is an error term, e_i, allowing misclassifications. If $e_i = 0$, then the individual i is correctly classified, and the second term disappears as if it was with the hard margin. If $0 < e_i \leq 1$, then i is inside the margin but at the correct side of the hyperplane therefore correctly classified. Finally if $e_i > 1$, then i is misclassified. The support vector machine can be either a linear classification, linear SVM, or a non-linear kernel classification such as RBF SVM (Baesens B. , 2014, pp. 58-61) (Bell, 2014, pp. 139-144).

The support vector machine classifiers can be written as either a primal or dual version where the latter is the most preferred when using kernels. Of this reason, the linear support vector machine classifier, as well as the non-linear support vector machine classifier, will be written in the dual version. The linear support vector machine classifier written as the dual version is as follows:

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i \left(x_i^T x\right) + b$$

by solving an optimization problem over α_i, known as the dual problem:

$$maximize_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \left(x_i^T x\right)$$

where y is the measured value, α_i is the Lagrangian multipliers stemming from the optimization (weight between 0 and cost), and x_i is the training data points, support vectors. Since support vectors are needed to construct the classification line, they will have a nonzero α_i but all other data points have a zero α_i. This is often referred to as the sparseness property of SVMs (Zisserman, 2015).

Non-linear support vector machine classification is characterized by a separation between classes that cannot directly be separated linearly due to the mix in the observation in each class. To be able to make a linear separation between the two classes with a hyperplane, the data should be transformed into a feature space by using the RBF kernel function. Thus, the feature space does not have to be explicitly specified. The non-linear dual version of a support vector machine can be formulated to learn a kernel classifier

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + b$$

by solving an optimization problem over α_i:

$$maximize_\alpha \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

where $K(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{\gamma^2}\right)$ when the kernel is an RBF SVM. Besides C, costs, RBF kernel includes an extra parameter to tune, which is γ (gamma) (Baesens B. , 2014, pp. 61-64; Zisserman, 2015).

Tuning the Model

As mentioned, linear SVM has C to tune, and RBF SVM has both C and gamma to tune. The tuning is done by the k-fold cross-validation, which is a statistical method used to estimate the skill of the machine learning model on new data. The parameter, k, is the number of groups a given data sample is to be split into. The most common one is k = 10, which means 10-fold cross-validation. The goal of cross-validation is to test the ability of the model to predict new data which was not used to estimate it. This is done to minimize the problems of overfitting or selection bias as well as to give an understanding of how the model will generalize to an independent testing data set. The cross-validation balances the importance of maximizing the margin against minimizing the error on the data to a number as close as possible to zero (Brownlee, 2018).

3.2.2.4. Random Forest

Random forest is created by a collection of classification trees to create a more robust model. First, a short introduction to classification trees will be done to get a basic understanding of how trees work. Second, random forest will be described, including an explanation of why the model becomes more robust when introducing random forest compared to classification trees.

Classification trees are often used when a data set is labelled, and the question is how new data points should be classified. A decision tree contains decision nodes where it starts with a root node, and then the data are split in two by using "if" statements. The goal is to pick nodes that give the best split possible. To determine the best split, the Gini Impurity can be used to maximize the gain in purity by minimizing the impurity. The latter appears when all observations are either one label or the other in the split. A higher Gini Gain indicates a better split. This can be done by the weighted decrease in the entropy measure:

$$max(gain) = 1 - \frac{m_1}{m} impurity_1 - \frac{m_2}{m} impurity_2$$

where m_k indicates the numbers of the label, "default" or "non-default", in the node, m indicates all observations and $impurity_k$ is defined by the entropy calculated for each node:

$$impurity_k = E(S) = -p_D \log_2(p_D) - p_{ND} \log_2(p_{ND})$$

where p_D and p_{ND} being the proportions of "default" and "non-default" respectively. The next nodes, children of the root note, only use data points that would take a particular direction which means to the left or the right of the root node. The way of building the next node is in the same way as for the root node whereby using the Gini Gain the best split is found. The next question that may occur is when to stop creating decision nodes, and the answer would be when all data points are equally good or if it has a Gini Gain of zero because then adding a decision node would not improve the decision tree. When this is the case, the node will be made as a leaf node which is the end node classifying any data point that reaches this node as in the same label. At the time when all possible branches in the classification tree end in a leaf node, the classification tree has been trained, and it can then be tested. So, the purpose in classification trees is to split observations into classes or labels of categorical dependent variables with a structure that looks like a tree and with as few as possible misclassifications. Classification trees suffer from instability because the classification trees may have high variability and the risk of overfitting. This instability of classification trees can be solved by introducing random forest (Baesens B. , 2014, pp. 42-48) (Zhou, 2019).

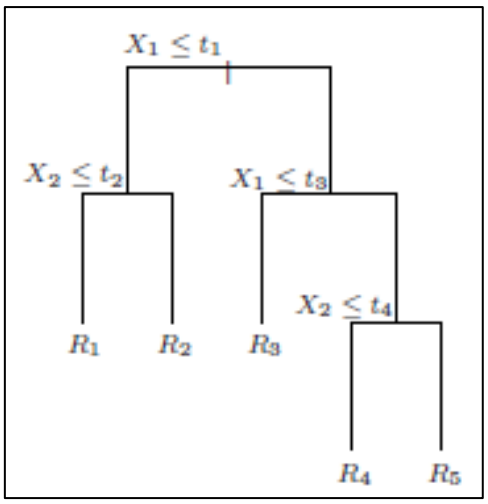


Figure 3.7: An example of a classification tree

The idea behind random forest is to average a collection of classification trees to build a more robust model with a better generalization performance and with less risk of overfitting. Random forest creates the collection of decision trees using each time a different training sample where each training sample is constructed by bootstrapping. Bootstrapping is a random sample with replacement which means that an element may appear multiple times in the one sample. This is repeated *k* times which is typically set to 500. Random forest uses the Strong Law of Large Numbers, which shows they always converge so that overfitting is not a problem and they produce a limiting value of the generalization error. Instead of evaluating all characters to determine the best split at each node, as done in classification trees, random forest only considers a random subset of those. In all, these factors may lead to better accuracy in random forest compared to a classification tree (Baesens B. , 2014, pp. 65-67) (Breiman, 2001) (Zhou, 2019).

3.2.2.5. General in Machine Learning

Type 1 and Type 2 Errors and Confusion Matrix

In machine learning the different methods are first trained on a data set called the training set, and afterwards, this model is tested on a data set called the testing set. After testing the model, different measures to evaluate the testing can be used. One of these measures is the accuracy as well as the table classifying how the prediction went compared to reality which is known as the confusion matrix. The accuracy tells the percentage of correctly predicted outcomes. Though, this may not give the complete picture of how good the model is. Therefore, the classification table is created to give an overview of where the model predicted correctly and most important, where the model predicted incorrectly. When the prediction is incorrect compared to the reality, it becomes either a type 1 or a type 2 error. Type 1 error, also known as false positive (FP), is where the predicted null hypothesis is rejected even though it was true. Type 2 error, which is also known as false negative (FN), is where the predicted alternative hypothesis is rejected even though it was true in reality. In this thesis, the null hypothesis is belonging to the "non-default" group where the alternative hypothesis is belonging to the "default" group. The critical error type is type 1 because the prediction tells the firm does not default, but it actually does. In this way, the credit lender may borrow money to a firm defaulting and ends up losing a proportion of the money.

Confusion Matrix		**Reality**	
		Non-default (positive) H_0 is true	**Default (negative) H_A is true**
Predicted	**Non-default (positive) H_0 is true**	Predicted correct (True positive)	**Type 1 error** (False positive)
	Non-default (negative) H_A is true	**Type 2 error** (False negative)	Predicted correct (True negative)

Table 3.2: Confusion matrix

Relative Operating Characteristic (ROC)

Another measure of evaluating the models is the relative operating characteristic (ROC) curve and the associated area under the curve (AUC). The ROC curve has the false-positive rate, also known as the rate of type 1 errors, on the x-axis and the true-positive rate on the y-axis. See figure 3.8 for illustration of the ROC curve. Therefore, the x-axis ranks the "default" on behalf of the probability of belonging in this class, and the y-axis shows the percentage of "non-default" excluded as a function of "default".

The goal is to have the ROC as far up to the top left corner as possible. This will increase the AUC, the area under the ROC curve, which is a measure to summarize the ROC. AUC will be a number between 0 and 1 where 0.5 indicates a random model and 1 indicates a perfect model that classifies all observation correct. If the AUC is 0.8, it tells that the model has an 80% chance of distinguishing correctly between the classes (Narkhede, 2018).

The difference between the accuracy and the ROC/AUC is the probability function which is used to a greater extent in ROC/AUC. The accuracy and the belonging type 1 and type 2 errors only show how the model classifies the observations for a given cut-off. Some of the models do not even need a cut-off but just classify the observations to either of the groups. ROC/AUC, on the other hand, plot the probabilities of belonging in a given class to visualize how well the model separates the two classes by probability and not only by classification.

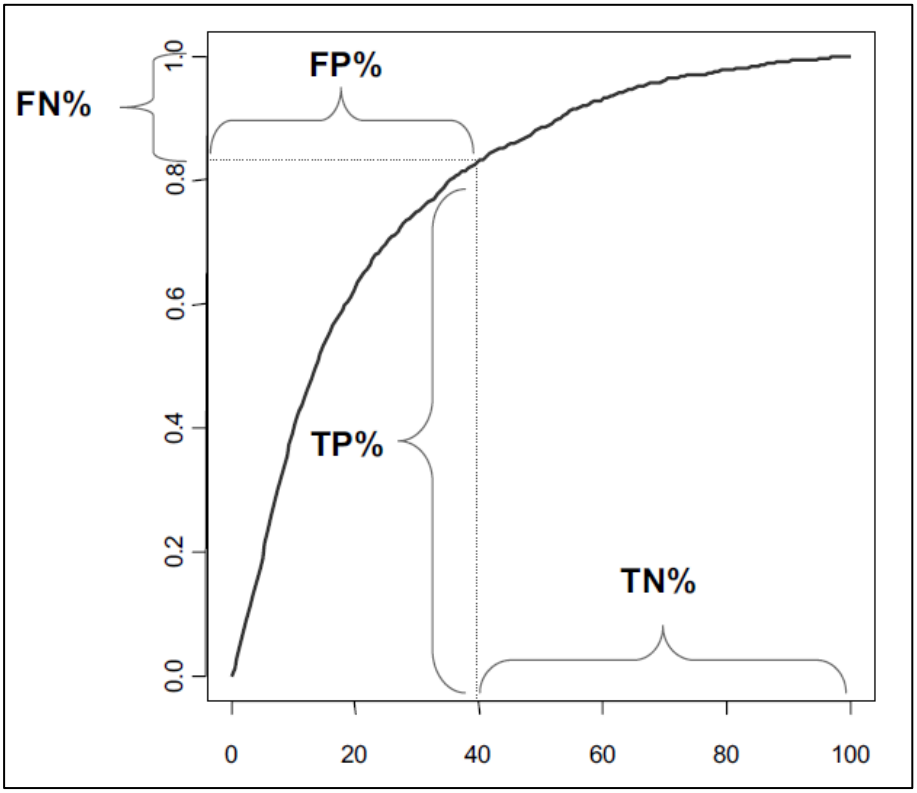


Figure 3.8: Schematic of a ROC showing how all four quantities of a confusion matrix can be identified on a ROC curve. Each region of the x- and y-axes have an interpretation with respect to error and success rates for non-defaulting (FP and TP) and defaulting (FN and TN) firms.

Source: Stein (2007)

Overfitting and Tuning

When training machine learning models, one of the most important aspects are to avoid overfitting. Overfitting occurs when a model fits the training data set too well, which means the model discovers

noise and try too hard to capture this in the training data set. It is illustrated in figure 3.9. This becomes a problem since the testing data set does not contain the same noise and then the ability of the model to generalize regarding a testing data set is negatively impacted. The more complex the model, the higher risk of overfitting and therefore a lower accuracy on the test will likely occur. There are different procedures to minimize the risk of overfitting and get a more robust model. Some of the procedures are to keep the model simple, to use regulation-algorithms, and/or to use cross-validation. By keeping the model simple, an approach could be to reduce the number of explanatory variables in the model using variable selection. Regulation-algorithms, on the other hand, is used to penalize complexity in the model by adding a parameter. Finally, cross-validation can be used when it is not possible to change model complexity or the size of the data set.

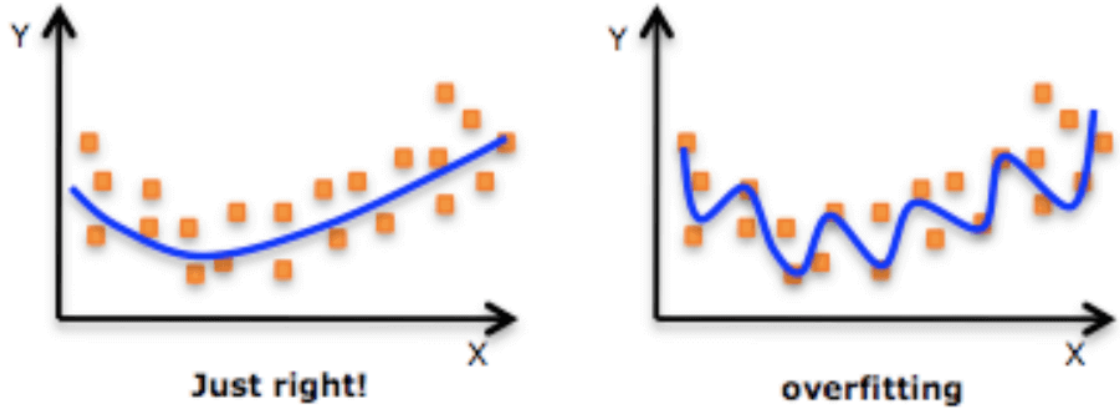


Figure 3.9: The difference between properly fitted and overfitting. The overfitted model is not going to be useful unless it is applied to the exact same data set because no other data will fall exactly along the overfitted line.

Source: Prakash (2018)

After training machine learning models, parameter tuning can be a way to build a model to optimally solve the machine learning problem. Parameter tuning is the choice of a set of optimal hyperparameters, such as the number of epochs in neural network, the cost and the gamma in SVM, and the number of trees in random forest.

4. Empirical Results

This section contains the empirical results of the different models followed by an analysis of these results. The first two parts include market as well as accounting variables, where the goal is to determine which model is the best when predicting default for listed firms. The two last parts only include accounting variables which are to investigate its predictive power and to analyse the impact on the market variables of the accuracy. This is to see if the empirical results could be transformed into non-listed firms. Both the investigation with market variables and the one without is separated in two time frames which are one year prior to default and five years prior to default. The time frame one year prior to default means that each firm-year is analysed, and the models decide whether the firm will be categorised as "default" or "non-default" within the next year. In the same way for the time frame five years prior to default it is analysed whether the firm defaults within a five-year period or not. Technically the models are fitted to the "non-default" category which means they predict "non-default" rather than "default". However, when dealing with binary classification, in the classes "default" or "non-default", the residual of "non-default" will always be "default". Likewise, when the models state a probability, this will be for the "non-default" group. To get the probability of the "default" group is simply to take one minus the probability of "non-default".

4.1. One Year Prior to Default Including Market and Accounting Variables

This part of the section contains the empirical results of the models made out of the training data one year prior to default, including both market and accounting variables. The training data set used to estimate these models includes 2,156 observations, 20 explanatory variables, and a binary dependent variable stating "default" one year prior to default and "non-default" otherwise. A correlation matrix can help to get an overview of how the variables in the training data are mutually correlated. The correlation matrix for this data is shown in table 4.1, and the correlation above $|0.5|$ is shown as the number for the given correlation. It can be a problem if too many variables are highly correlated as this indicate that the variables contain the same information. Most of the variables are correlated to a lesser extent, but there are some cases where the correlation is above $|0.5|$ which are the numbers shown in the table. RETA, EBTA, NITA, and OCFTA all have a correlation above $|0.5|$ with one another. These variables are mutually correlated, which makes sense since all of them are some kind of earnings over total asset ratios. This means that these four variables contain some kind of the same information. Especially the correlation at 0.89 between OCFTA and EBTA is very high meaning it will add very little extra information power to add both variables in the model.

	WCTA	RETA	EBTA	METL	SLTA	CACL	NITA	TLTA	EXRET	RSIZ	SIGMA	FFOTL	X.NI	NIMETL	TLMETL	EBITDASL	OCFTA	FESL	FDCF	CLTA
WCTA	1	-	-	-	-	0.51	-	-0.67	-	-	-	-	-	-	-	-	-	-	-	-0.74
RETA	-	1	0.6	-	-	-	0.56	-	-	-	-	-	-	-	-	-	0.67	-	-	-
EBTA	-	0.6	1	-	-	-	0.86	-	-	-	-	-	-	-	-	-	0.89	-	-	-
METL	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SLTA	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CACL	0.51	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
NITA	-	0.56	0.86	-	-	-	1	-	-	-	-	-	-	-	-	-	0.71	-	-	-
TLTA	-0.67	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	0.66
EXRET	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-
RSIZ	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-
SIGMA	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-
FFOTL	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-
X.NI	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-
NIMETL	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-0.7	-	-	-	-	-
TLMETL	-	-	-	-	-	-	-	-	-	-	-	-	-	-0.7	1	-	-	-	-	-
EBITDASL	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-0.78	-	-
OCFTA	-	0.67	0.89	-	-	-	0.71	-	-	-	-	-	-	-	-	-	1	-	-	-
FESL	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-0.78	-	1	-	-
FDCF	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-
CLTA	-0.74	-	-	-	-	-	-	0.66	-	-	-	-	-	-	-	-	-	-	-	1

Table 4.1: Correlation matrix for the variables in the data one year prior to default including market and accounting

The different machine learning models are then tested on the testing data set containing the same 20 explanatory variables but with 24,630 observations where 225 categorised as "default" and 24,405 as "non-default". The empirical results of each machine learning method will be elaborated and analysed in the next.

4.1.1. Logistic Regression

As known from section 3.2.2.1., logistic regression is a predictive analysis where the goal is to predict the probability of default for a firm, in this case, one year prior to default, by solving the binary classification problem.

The first model, LR1, is built with all 20 explanatory variables based on accounting variables as well as market variables. The model results in 12 statistically insignificant variables with a p-value higher than 0.05. The insignificant variables indicate that 12 of the explanatory variables have no significant effect on the dependent variable according to the results. A solution to the insignificant variables is to make a variable selection by taking out the least significant variable one by one until a model with no insignificant variables remaining is left. This may also result in a lower risk of overfitting the model. The results of this can be seen in table 4.2.

Model	Variable(s) included in the model	Log-likelihood	AIC	# of insignificant variables
LR1	WCTA, RETA, EBTA, METL, SLTA, CACL, NITA, TLTA, EXRET, RSIZ, SIGMA, FFOTL, X.NI, NIMETL, TLMETL, EBITDASL, OCFTA, FESL, FDCF, and CLTA	-998.711	2039.4	12
LR2	WCTA, EBTA, METL, SLTA, CACL, NITA, TLTA, EXRET, RSIZ, SIGMA, FFOTL, X.NI, NIMETL, TLMETL, EBITDASL, OCFTA, FESL, FDCF, and CLTA	-998.713	2037.4	11
LR3	WCTA, METL, SLTA, CACL, NITA, TLTA, EXRET, RSIZ, SIGMA, FFOTL, X.NI, NIMETL, TLMETL, EBITDASL, OCFTA, FESL, FDCF, and CLTA	-998.773	2035.5	10
LR4	WCTA, SLTA, CACL, NITA, TLTA, EXRET, RSIZ, SIGMA, FFOTL, X.NI, NIMETL, TLMETL, EBITDASL, OCFTA, FESL, FDCF, and CLTA	-998.831	2033.7	9
LR5	SLTA, CACL, NITA, TLTA, EXRET, RSIZ, SIGMA, FFOTL, X.NI, NIMETL, TLMETL, EBITDASL, OCFTA, FESL, FDCF, and CLTA	-999.028	2032.1	7
LR6	SLTA, CACL, NITA, TLTA, EXRET, RSIZ, SIGMA, FFOTL, X.NI, NIMETL, TLMETL, OCFTA, FESL, FDCF, and CLTA	-999.356	2030.7	6
LR7	SLTA, CACL, NITA, TLTA, EXRET, RSIZ, SIGMA, FFOTL, X.NI, NIMETL, TLMETL, OCFTA, FESL, and CLTA	-1000.071	2030.1	5
LR8	SLTA, CACL, NITA, TLTA, EXRET, RSIZ, SIGMA, FFOTL, X.NI, TLMETL, OCFTA, FESL, and CLTA	-1000.718	2029.4	4
LR9	SLTA, CACL, TLTA, EXRET, RSIZ, SIGMA, FFOTL, X.NI, TLMETL, OCFTA, FESL, and CLTA	-1001.029	2028.1	3
LR10	SLTA, CACL, TLTA, EXRET, RSIZ, SIGMA, FFOTL, X.NI, TLMETL, OCFTA, and CLTA	-1002.016	2028.0	3
LR11	SLTA, CACL, TLTA, EXRET, RSIZ, SIGMA, X.NI, TLMETL, OCFTA, and CLTA	-1004.254	2030.5	1
LR12	SLTA, TLTA, EXRET, RSIZ, SIGMA, X.NI, TLMETL, OCFTA, and CLTA	-1005.327	2030.7	0

Table 4.2: Table showing 12 different logistic regression models on the data one year prior to default including market variables

When only focusing on maximizing the log-likelihood measure the first model, LR1, including all 20 explanatory variables, are the preferred one with a log-likelihood of -998.711. However, this model includes 12 insignificant variables that may overfit the model based on the training data, which might

not give a good general result in terms of the testing data set. Furthermore, this model is also the most complex one due to many variables. Another parameter, AIC, which also compare the models, can then be taken into account. According to AIC, the best model is LR10 due to the lowest measure of AIC. The reason for this is it takes the number of variables into account and the fewer variables, the better the model will be. Nevertheless, despite the lower log-likelihood measure compared to LR1 and the higher AIC measure compared to LR10, it can be argued that the last model, LR12 with no insignificant variables, is the most preferably model estimated out of the training data set. Hence, it has fewer explanatory variables and therefore, less complexity. LR12, including nine explanatory variables, can be written as:

$$probability\ of\ default = 1 - \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_9 x_9)}}$$

Where

$$\begin{aligned}\beta_0 + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_9 x_9 \\ = -6.849 + 2.817\frac{Sales}{Total\ assets} - 25.761\frac{Total\ liabilities}{Total\ assets} \\ + 5.228\big(Log(Firm\ return) - Log(Market\ return)\big) \\ + 2.019\left(Log\left(\frac{Firm\ market\ cap}{Market\ cap}\right)\right) - 3.901SIGMA + 0.529\Delta\ in\ net\ income \\ - 32.277\frac{Total\ liabilities}{Market\ cap + Total\ liabilities} + 6.432\frac{Operating\ CF}{Total\ assets} \\ - 3.365\frac{Current\ liabilities}{Total\ assets}\end{aligned}$$

Where SIGMA is equal to the monthly volatility over the last year. The model holds five accounting variables and four market variables. This model is then tested on the testing data set. With a cut-off point of 0.5, the test classified 3,020 observations as "default" and 21,610 observations as "non-default" and obtained an accuracy of 88.31%. Though, this does not give a complete picture of how good the model is to predict default. A confusion matrix stating the proportion of correctly classified as well as those misclassified for each category gives a more fulfilling picture of the test result. Table 4.3 shows this confusion matrix. Misclassifications can be split into type 1 and type 2 errors. The model contains 42 (18.67%) type 1 errors which mean 42 are classified as "non-default" but they actually "default". This is a problem because then the credit lender will lose money in the case they borrow an amount to the firm. Type 2, on the other hand, contains 2,837 (11.62%) cases where the model classifies firms to default where in reality, the firms did not default. This will result in the credit lender, not borrowing the money to firms that actually would be able to fulfil its contractual obligations. The ROC curve of the model can be seen in appendix A figure A3, and the AUC equals 0.9092. To decide whether this is a good or a bad result, a translation table is made to explain the AUC value for the models. The table can be seen in appendix B table B3. Regarding this table, the AUC measure of this logistic regression model is a good result.

Reality / Predicted	**Non-default** H_0 **is true**	**Default** H_A **is true**
Non-default H_0 **is true**	21,568	42 (18.67%)
Default H_A **is true**	2,837 (11.62%)	183

Table 4.3: Confusion matric for logistic regression model one year prior to default including market variables

4.1.2. Neural Network

To test the data set with neural network the R package ANN2 is used. The description of this package is: "Training of neural networks for classification and regression tasks using mini-batch gradient descent" (Lammers, 2020). The package also allows for regularization, which is shortly described in section 3.2.2.5.

When the neural network was made, there was a focus on simplicity. This includes only one hidden layer with eight neurons within. The network has 20 input neurons in the first layer that indicate one for each explanatory variable. The next layer consists of eight neurons in the hidden layer. The last layer in the network is the output layer which only has two output neurons, "default" and "non-default".

One of the parameters to tune in a neural network is the number of epochs. Section 3.2.2.2 defined the number of epochs as the number of times the whole training data set has passed through the network. The higher number of epochs will result in the model being fitted better to the training data. Figure 4.1 shows the validation and training loss on the y-axis and the number of epochs on the x-axis. The number of epochs should be on the spot of the graph just before the validation loss hits a plateau. The graph shows that the validation loss (yellow line) is declining as the number of epochs increase, but after epoch number 50, the decline is reduced significantly. It is concluded that the number of epochs should not be more than 100 as this will overfit the model. However, it can be discussed whether the number of epochs should be 50, 100, or somewhere in between. The chosen number of epochs is 100, which is also the most used number in the model.

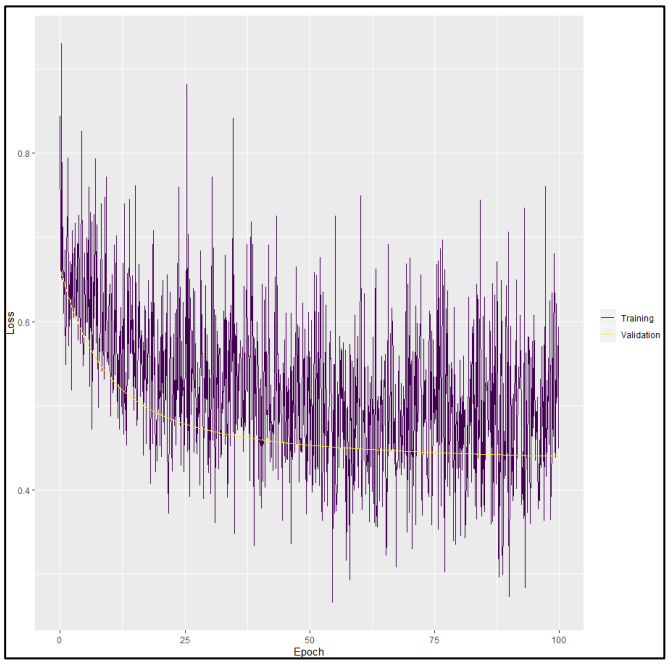


Figure 4.1: Graph showing the validation loss (the yellow line) on the data one year prior to default including market variables

The test sample was run through the network, and with a cut-off point of 0.5, the test result classified 2,793 observations as "default" and 21.837 observations as "non-default" with an accuracy of 89.26%. The confusion matrix showing the misclassifications is shown in table 4.4. The model has 40 (17.78%) type 1 error and 2,605 (10.67%) type 2 errors. The ROC curve of the model can be seen in appendix A figure A4 and the AUC equals 0.9047 which is a good result according to table B3 in appendix B.

Reality / Predicted	**Non-default** H_0 **is true**	**Default** H_A **is true**
Non-default H_0 **is true**	21,797	40 (17.78%)
Default H_A **is true**	2,605 (10.67%)	185

Table 4.4: Confusion matric for neural network model one year prior to default including market variables

4.1.3. Support Vector Machine

To refresh, the goal of SVM is to split the classes with a hyperplane that maximizes the distance to the nearest training data point to minimize the risk of misclassification. Two SVM models, the linear SVM and the RBF SVM will be estimated in R by installing and using the package e1071. This package is the first and most intuitive packages for estimating SVM in R.

4.1.3.1. Linear SVM

A linear SVM classification is characterized by a separation between classes that can be directly split linearly. As mentioned in section 3.2.2.3, linear SVM has one parameter to tune. This is C (cost) which is a trade-off parameter between margin and correct classification. While taking the error rate into consideration, it is most preferred to take a low C in a linear SVM model which indicates a lower risk of overfitting. A 10-fold cross-validation has been used to tune C. This 10-fold cross-validation balances the importance of maximizing the margin versus minimizing the error on the data to as close as possible to 0. The linear SVM model is tuned for C = $[10^{-4:4}]$. It might have been preferred if the linear SVM were trained in a wider range for its parameter, C, to return better results. Though this is time-consuming, since R on a personal computer might not be able to run this, and finally it looks like the error is ending up in a local and accepted minimum.

Table 4.5 shows the nine different models in the linear SVM. The best model to choose for linear SVM is the one with the lowest error of 0.2087 and a cost of 1. This model includes 1,101 support vectors, 551 "default" and 550 "non-default". Recall that support vectors are the number of data points within the soft margin of the hyperplane – the ones defining the hyperplane. The amount of support vectors depends on how much misclassification is allowed. With a large number of misclassifications, there will be a large number of support vectors and vice versa. Around 51% of the data points from the training data are support vectors which may indicate a risk of a large number of misclassifications. Though, the goal is not to find the lowest misclassification but rather the right number of misclassifications for the data being analysed.

Model	Cost	Error
LinearSVM1	0.0001	0.4044
LinearSVM2	0.001	0.2199
LinearSVM3	0.01	0.2097
LinearSVM4	0.1	0.2092
LinearSVM5	1	0.2087
LinearSVM6	10	0.2092
LinearSVM7	100	0.2097
LinearSVM8	1,000	0.2153
LinearSVM9	10,000	0.2421

Table 4.5: Table showing nine different linear SVM models one the data one year prior to default including market variables

The chosen model, linearSVM5, is then tested by using the testing data set. The test classified 2,865 observations as "default" and 21,765 observations as "non-default" and obtained an accuracy of 88.93%. The confusion matrix showing misclassifications can be seen in table 4.6. The model includes

43 (19.11%) type 1 errors and 2,683 (10.99%) type 2 errors. The ROC curve of this model can be seen in appendix A figure A5 and the AUC equals 0.9049, which is a good result according to table B3 in appendix B.

Reality / Predicted	**Non-default** H_0 **is true**	**Default** H_A **is true**
Non-default H_0 **is true**	21,722	43 (19.11%)
Default H_A **is true**	2,683 (10.99%)	182

Table 4.6: Confusion matric for linear SVM model one year prior to default including market variables

4.1.3.2. RBF SVM

RBF SVM is a non-linear SVM method characterized by a separation between classes that cannot be directly split linearly. The data in non-linear SVM methods should be transformed into a feature space by using a kernel function, in this case, the RBF kernel function. RBF SVM has two parameters to tune; C (cost) which is the same parameter as in linear SVM and gamma, which defines how far the influence of a single training example reaches. With a low value of gamma, every data point has a far reach which means that it takes the ones far away into consideration and vice versa. For this reason, a relatively low gamma would be preferred. The RBF SVM model is tuned for the same value of C as in linear SVM, C = $[10^{-4:4}]$, and value of gamma = $[2^{-2:2}]$. As mentioned above in linear SVM, it might have been more desirable to train the model in a wider range for each parameter, but due to the same mentioned reasons, this is not done. To tune these parameters, the 10-fold cross-validation is once again used.

This results in 45 different models where the best is RBFSVM21 with a gamma of 0.25, a cost of 1, and an error of 0.2139. The best model and the ones around can be seen in table 4.7. To see all 45 different models, see appendix B table B5. RBFSVM21 includes 1,381 support vectors, 734 “default” and 647 “non-default”. These support vectors still depend on how much misclassification is allowed. About 64% of the observations are support vectors which can be argued to be, to some extent relatively many support vectors. It may indicate a risk of a relatively large number of misclassifications compared to the linear SVM model.

Model	Gamma	Cost	Error
RBFSVM17	0.50	0.1	0.2932
RBFSVM18	1.00	0.1	0.3697
RBFSVM19	2.00	0.1	0.4559
RBFSVM20	4.00	0.1	0.5227
RBFSVM21	0.25	1	0.2139
RBFSVM22	0.50	1	0.2255
RBFSVM23	1.00	1	0.2658
RBFSVM24	2.00	1	0.3094
RBFSVM25	4.00	1	0.3716

Table 4.7: Table showing the best RBF SVM model and the models around this on the data one year prior to default including market variables

The chosen model, RBFSVM21, is tested by using the testing data set. The test classified 4,628 observations as "default" and 20,002 observations as "non-default". The estimated accuracy equals 81.85%, and the confusion matrix can be seen in table 4.8. The test results in 34 (14,22%) type 1 errors and 4,437 (18.18%) type 2 errors. The ROC curve of the model can be seen in appendix A figure A6, and the AUC equals 0.8987 which is a good result according to table B3 in appendix B, but it is the lowest among the models on the data set.

Reality / Predicted	**Non-default** H_0 **is true**	**Default** H_A **is true**
Non-default H_0 **is true**	20,147	34 (14.22%)
Default H_A **is true**	4,437 (18.18%)	191

Table 4.8: Confusion matric for RBF SVM model one year prior to default including market variables

4.1.4. Random Forest

To refresh, random forest makes a collection of trees which are built upon various if statements. The trees vote for which category the observation should belong to and the majority of the voters decide the classification. To test the data set with random forest, the R package "randomForest" is used. This package uses an algorithm that can be used for both regression and classification.

Like other machine learning methods, there are parameters to tune in random forest. Two parameters should be decided, namely the number of trees (ntree) and the number of variables sampled as

candidates at each split (mtry). To determine the number of trees, the training data is trained for 500 trees and plotted against the error rate on the model. Figure 4.2 shows the plot of the error rate (the black line), and it indicates that the last 200 trees are not needed for the model. The error rate is lower at 300 trees compared to 500 trees. To reduce computational calculation and to get a lower error rate, it is decided to use 300 trees in the model. Appendix C output C1 and C2 shows the summary for the model with 500 and 300 trees, and it demonstrates how the error rate declined from 19.81% to 19.53%. The second parameter to tune is mtry. It is common to set this parameter to the squared root of the number of variables. From that number, R can perform a test to see how the error rate will react if mtry is increased or decreased. It shows that the error rate will increase if the mtry is increased to 6 or reduced to 3. See appendix A figure A7 for the plot of the test. It is therefore decided to keep the mtry to 4 for the model which is also the closest number to the square root of the number of variables.

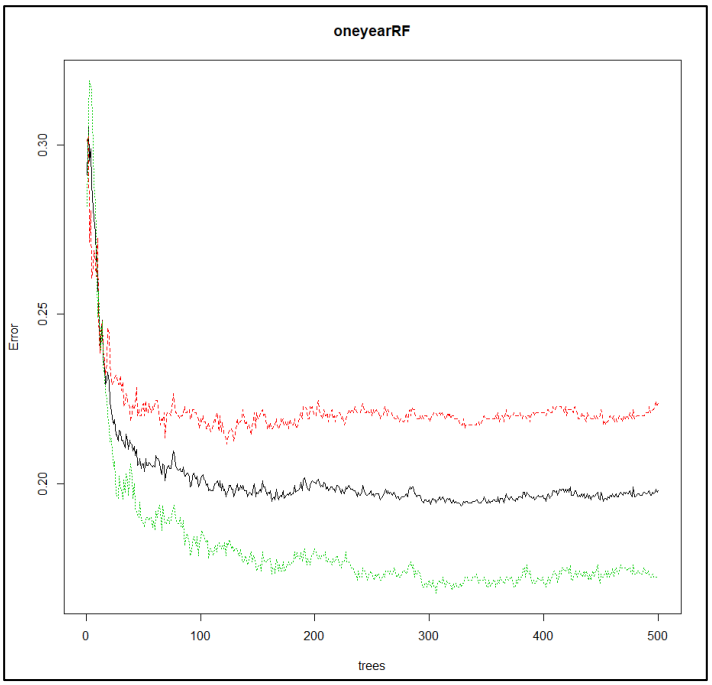


Figure 4.2: Figure showing the plot of the error rate (the black line) of random forest on data one year prior to default including market variables

Since the sample of variables available for each split is different, random forest can give an estimation of how important each variable is in the model. Figure 4.3 shows the relative importance of the variables in two measures. The first measure is the mean decrease in accuracy, which calculates how the accuracy will drop if one of the variables is omitted from the model. The second measure is the mean decrease in Gini which measures the average gain purity for the local split which means that it tells something about the composition of the trees and which variables have the highest gains in purity. For both measures, the more important variables are located to the right of the figure while the less important variables are found to the left of the figure. It is, to some extent, the same variables that are important independent of the measures. TLMETL and METL are in the top three for both measures and will have

the most impact if these variables where excluded. On the other hand, X.NI and SLTA will have the slightest impact if these variables were excluded. However, it is also worth mentioning that a high degree of correlation between some of the variables can have an impact on this importance plot. As the correlation matrix in section 4.1. shows there are variables with a correlation close to 0.9, which will have an impact on how they rank the mutual importance.

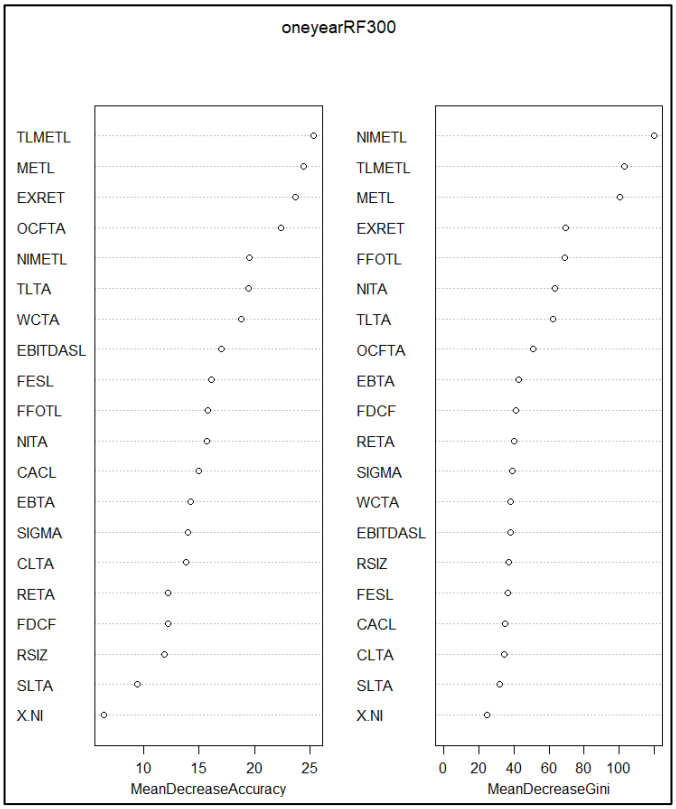


Figure 4.3: Figure showing the importance of each variable in random forest regarding the accuracy and the Gini index on the data one year prior to default including market variables

The model was run on the testing data set, and it classified 3,737 observations as "default" and 20,893 observations as "non-default". The empirical results obtained an accuracy of 85.52%. The confusion matrix can be as seen in table 4.9, where it is shown that the model has 32 (12,00%) type 1 errors and 3,539 (14.50%) type 2 errors. The ROC curve of the model can be seen in appendix A figure A8, and the AUC equals 0.9295 which is a very good result according to table B3 in appendix B, and it is the highest among the models on the data set.

Reality / Predicted	**Non-default H_0 is true**	**Default H_A is true**
Non-default H_0 is true	20,916	27 (12,00%)
Default H_A is true	3,539 (14.50%)	193

Table 4.9: Confusion matric for random forest model one year prior to default including market variables

4.1.5. The Machine Learning Methods Compared to One Another

This part gives a summary of the results between the five models as well as an analysis of which model performs best. The test was done on the testing set, which includes 24,630 firm-year observations with both accounting and market data as explanatory variables.

Model	Type 1	Type 2	Accuracy	AUC
LR	42	2,837	88.31%	0.9092
NN	40	2,605	89.26%	0.9047
LinearSVM	43	2,683	88.93%	0.9049
RBFSVM	34	4,437	81.85%	0.8987
RF	27	3,539	85.52%	0.9295

Table 4.10: Summary of the chosen models in each method and their most important results of type 1 errors, type 2 errors, accuracy and AUC on the data one year prior to default including market variables

Table 4.10 shows the most important results between the models. It can be concluded that neural network has the highest accuracy at 89.25% closely followed by linear SVM and logistic regression. Random forest and RBF SVM have both considerably lower accuracies at respectively 85.52% and 82.75%. From table 4.10, it is clear to say that type 2 error directs the accuracy percentage. However, as section 3.2.2.5 describes the critical ones are type 1 error as this is a situation where the credit lender gives credit to a company that defaults within the next year. Random forest is the model with the lowest number of type 1 errors at 27, but the range between the highest and lowest number for the models is not wide. Linear SVM, as the model with the highest number of type 1 errors, has 43 errors.

To evaluate which model is the best for this data, both type 1 and type 2 errors should be taken into account. However, it is fair to argue that the cost of a borrower who cannot pay back the loan is higher compared to the opportunity cost for a credit lender, that refuses to give credit to a borrower, who is not going to default within the next year. Another implication for this is, even though the borrower is not

going to default within the next year, there is still the risk of losing money if the borrower defaults in the following years. So, the question is how much more the model should penalize type 1 errors compared to type 2 errors. There is no simple answer to this question, and it will also depend on the risk appetite for the given credit lender. However, it is possible to analyse which model is the best given different weights multiplied type 1 errors which indicate the higher cost of type 1 errors relative to the cost of type 2 errors. Figure 4.4 gives an illustration of this.

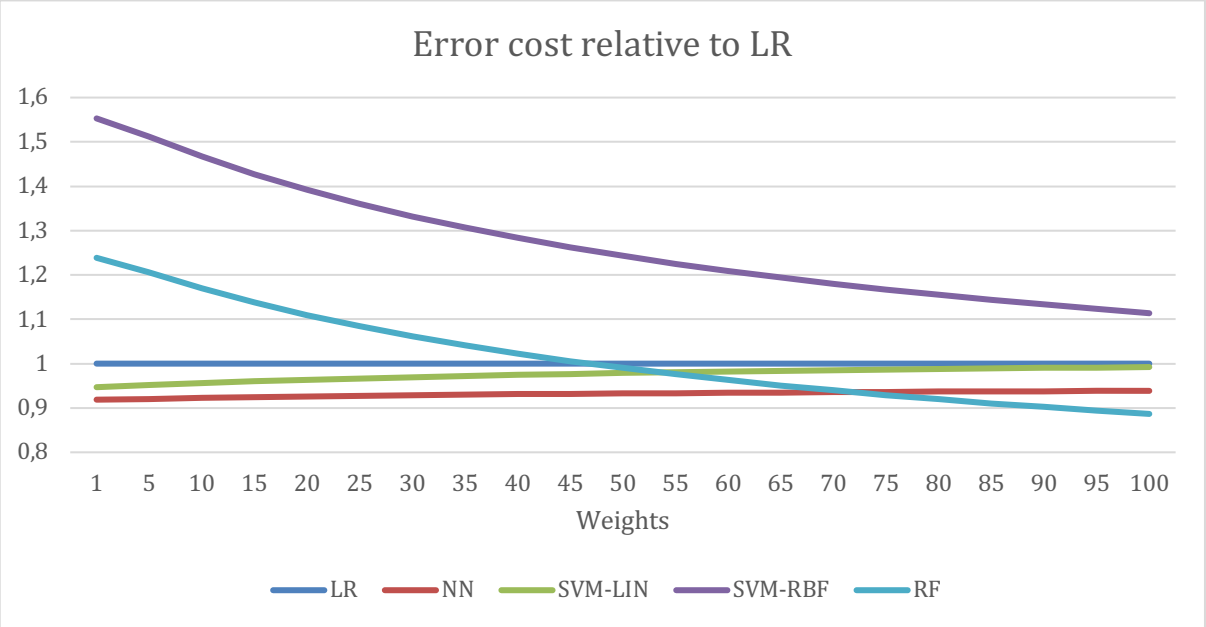


Figure 4.4: Figure showing the error cost relative to the logistic regression model on the data one year prior to default including market variables

The x-axis is the different weights, and if the weight is equal to ten, then type 1 error is ten times more costly compared to type 2 errors. The y-axis shows how each model is doing in terms of the cost of the errors compared to logistic regression. If the graph is lower than one, then the model is better than logistic regression at the given level of weights, while a graph higher than one means the model is worse compared to logistic regression. Logistic regression is chosen as the benchmark because it is the oldest and most simple model. The figure shows that neural network is the best model from the beginning where the costs of type 1 and type 2 errors are the same. This continues until the weight is equal to 75, where it is crossed by random forest. However, it can be argued that 75 times higher cost related to type 1 error compared to type 2 errors is very high. Therefore, it is decided that neural network is the best model to predict default within one year with accounting and market variables available in terms of the accuracy and the distribution of the error types.

Regarding the other measure to validate the models, AUC, the difference between the models is smaller compared to the accuracy. Random forest is the best model with an AUC of 0.9295, followed by logistic regression at 0.9092. RBF SVM is the one with the lowest value of AUC at 0.8987, and it was also the model with the lowest accuracy. All the ROC curves can be seen in figure 4.5, where random forest

(green line) is the line furthest to the top left corner, which indicates the best model. All the other lines follow each other very closely except for RBF SVM, which is below all the lines in the false positive rate interval from 0.1 to 0.4. Recall from section 3.2.2.5., that the false positive rate on the x-axis indicates the percentage of type 1 errors while the true positive rate on the y-axis indicates the percentage of "non-default" correctly classified.

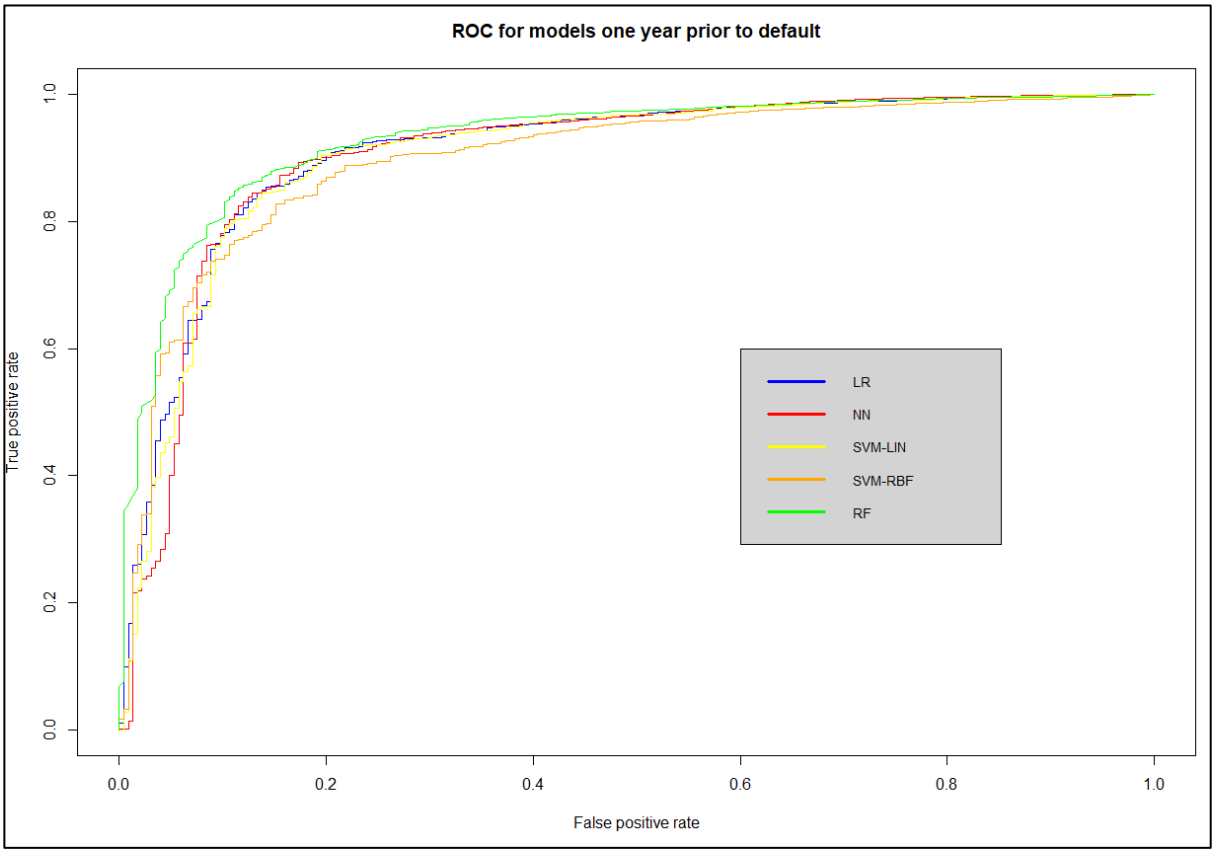


Figure 4.5: Figure showing the ROC curves of the chosen models in each method on the data one year prior to default including market variables

4.2. Five Years Prior to Default Including Market and Accounting Variables

The second part of this section contains the empirical results of the models created out of the training data five years prior to default, where market and accounting variables are included. The training data used for this part to estimate these models contains 8,286 observations, 20 explanatory variables, and a binary dependent variable stating "default" five years prior to default and "non-default" otherwise. The number of observations in the training set is almost four times as large compared to the training set for one year prior to default. The reason for this is, the number of observations which are categorized as "default" within the next five years must by nature be higher than within one year. In addition to that, the "default" observations were matched with "non-default" observations as described in section 2.2.1. which further increase the number of observations. The correlation matrix for the data can be seen in

table 4.11. The matrix follows the same method as described in section 4.1. with only showing correlation over |0.5| Even though the data is different from one year prior to default, it is almost the same pairs of variables that correlate over |0.5|. RETA, EBTA, NITA and OCFTA are still mutually correlated with some of the highest correlation among the numbers. The only new variable with a correlation above |0.5| is FFOTL, which is now correlated with EBTA and NITA.

	WCTA	RETA	EBTA	METL	SLTA	CACL	NITA	TLTA	EXRET	RSIZ	SIGMA	FFOTL	X.NI	NIMETL	TLMETL	EBITDASL	OCFTA	FESL	FDCF	CLTA
WCTA	1	-	-	-	-	-	-	-0.64	-	-	-	-	-	-	-	-	-	-	-	-0.61
RETA	-	1	0.62	-	-	-	0.59	-	-	-	-	-	-	-	-	-	0.63	-	-	-
EBTA	-	0.62	1	-	-	-	0.86	-	-	-	-	0.53	-	-	-	-	0.85	-	-	-
METL	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SLTA	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CACL	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
NITA	-	0.59	0.86	-	-	-	1	-	-	-	-	0.56	-	-	-	-	0.73	-	-	-
TLTA	-0.64	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	0.61
EXRET	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-
RSIZ	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-
SIGMA	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-
FFOTL	-	-	0.53	-	-	-	0.56	-	-	-	-	1	-	-	-	-	-	-	-	-
X.NI	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-
NIMETL	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-0.62	-	-	-	-	-
TLMETL	-	-	-	-	-	-	-	-	-	-	-	-	-	-0.62	1	-	-	-	-	-
EBITDASL	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-0.64	-	-
OCFTA	-	0.63	0.85	-	-	-	0.73	-	-	-	-	-	-	-	-	-	1	-	-	-
FESL	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-0.64	-	1	-	-
FDCF	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-
CLTA	-0.61	-	-	-	-	-	-	0.61	-	-	-	-	-	-	-	-	-	-	-	1

Table 4.11: Correlation matrix for the variables in the data five years prior to default including market and accounting

The testing data set contains 17,843 observations, where 559 of them are categorised as “default” and 17,284 of them are “non-default”. This testing set is also a bit different compared to one year prior to default, as the period has been limited until 2013. The reason for this is that it should be avoided to have observations where it was not certain whether the firm defaults or not within the next five years. The results for each machine learning method will be elaborated and analysed in the following.

4.2.1. Logistic Regression

LR1 is built with all 20 explanatory variables. The model results in nine statistically insignificant variables with a p-value higher than 0.05. Variable selection regarding insignificant variables will then be used to reduce the risk of overfitting. The results of this can be seen in table 4.12.

Model	Variable(s) included in the model	Log-likelihood	AIC	# of insignificant variables
LR1	WCTA, RETA, EBTA, METL, SLTA, CACL, NITA, TLTA, EXRET, RSIZ, SIGMA, FFOTL, X.NI, NIMETL, TLMETL, EBITDASL, OCFTA, FESL, FDCF, and CLTA	-5020.168	10082	9
LR2	RETA, EBTA, METL, SLTA, CACL, NITA, TLTA, EXRET, RSIZ, SIGMA, FFOTL, X.NI, NIMETL, TLMETL, EBITDASL, OCFTA, FESL, FDCF, and CLTA	-5020.179	10080	8
LR3	RETA, EBTA, METL, SLTA, CACL, NITA, TLTA, EXRET, RSIZ, SIGMA, FFOTL, X.NI, NIMETL, TLMETL, EBITDASL, OCFTA, FESL, and FDCF	-5020.224	10078	7
LR4	RETA, EBTA, METL, SLTA, CACL, TLTA, EXRET, RSIZ, SIGMA, FFOTL, X.NI, NIMETL, TLMETL, EBITDASL, OCFTA, FESL, and FDCF	-5020.274	10077	6
LR5	RETA, EBTA, METL, SLTA, CACL, TLTA, EXRET, RSIZ, SIGMA, FFOTL, NIMETL, TLMETL, EBITDASL, OCFTA, FESL, and FDCF	-5020.356	10075	5
LR6	RETA, METL, SLTA, CACL, TLTA, EXRET, RSIZ, SIGMA, FFOTL, NIMETL, TLMETL, EBITDASL, OCFTA, FESL, and FDCF	-5020.546	10073	4
LR7	RETA, METL, SLTA, CACL, TLTA, EXRET, RSIZ, SIGMA, FFOTL, NIMETL, TLMETL, OCFTA, FESL, and FDCF	-5021.326	10073	3
LR8	RETA, METL, SLTA, CACL, TLTA, EXRET, RSIZ, SIGMA, FFOTL, NIMETL, TLMETL, OCFTA, and FDCF	-5022.8	10074	2
LR9	RETA, METL, SLTA, CACL, TLTA, EXRET, RSIZ, SIGMA, FFOTL, NIMETL, TLMETL, and OCFTA	-5023.751	10074	1
LR10	RETA, SLTA, CACL, TLTA, EXRET, RSIZ, SIGMA, FFOTL, NIMETL, TLMETL, and OCFTA	-5025.692	10075	0

Table 4.12: Table showing ten different logistic regression models on the data five years prior to default excluding market variables

With the focus of the maximum log-likelihood measure, LR1, including all 20 explanatory variables where nine of them are insignificant, is the most prefered model with a log-likelihood of -5020.168. Though the model contains the most insignificant variables compared to the other models. The insignificant variables may lead to the risk of overfitting the model based on the training data, which might not give a good result in terms of the testing data set. Likewise, this model is the most complex one because of its many variables. Regarding the other measure, AIC, which compares models on the

same training data set, LR6 and LR7 would be the ones to prefer due to the lowest AIC. With this information in mind, despite the highest log-likelihood measure, it can be argued that LR10 is the most preferred one due to no insignificant variables and then fewer explanatory variables, and still a relatively low AIC compared to the other models. LR10 can be written as:

$$probability\ of\ default = 1 - \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_{11} x_{11})}}$$

Where

$$\begin{aligned}\beta_0 + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_{11} x_{11} &= -0.144 - 3.917\frac{Retained\ earnings}{Total\ assets} + 1.472\frac{Sales}{Total\ assets} + 4.486\frac{Current\ assets}{Current\ liabilities} \\ &- 14.932\frac{Total\ liabilites}{Total\ assets} + 1.182\big(Log(Firm\ return) - Log(Market\ return)\big) \\ &+ 1.903\left(Log\left(\frac{Firm\ market\ cap}{Market\ cap}\right)\right) - 3.273SIGMA + 7.940\frac{Funds\ from\ operations}{Total\ liabilities} \\ &- 14.838\frac{Net\ income}{Market\ cap + Total\ liabilities} - 38.052\frac{Total\ liabilities}{Market\ cap + Total\ liabilities} \\ &+ 8.462\frac{Operating\ CF}{Total\ assets}\end{aligned}$$

where SIGMA is equal to monthly volatility over the last year. This model includes 11 explanatory variables where six of them are accounting variables, and five of them are market variables. Seven of the 11 variables are similar to the ones chosen by the logistic regression model in one year prior to default. The four added variables are RETA, CACL, FFOTL and NIMETL, and the ones that have been dropped by this model is X.NI and CLTA. LR10 is then tested on the testing data set. The test classified 3,546 observations as “default” and 14,297 observations as “non-default” with a cut-off point of 0.5. The model achieved an accuracy of 81.09%. In table 4.13 is the confusion matrix shown. The model results in 194 (34.70%) type 1 errors and 3,181 (18.40%) type 2 errors. It is worth noticing how both types of errors had increased compared to one year prior to default. Especially type 1 error has a relatively higher percentages increase. The ROC curve of the model can be seen in appendix A figure A9, and the AUC equals 0.8185. To decide whether this is a good or a bad result, a new translation table is made to explain the AUC value for the models on the five years horizon. The table can be seen in appendix B table B4. Regarding this table, the AUC measure of this logistic regression model is a merely good result and considerably lower compared to the logistic regression model one year prior to default.

Reality / Predicted	**Non-default** H_0 **is true**	**Default** H_A **is true**
Non-default H_0 **is true**	14,103	194 (34.70%)
Default H_A **is true**	3,181 (18.40%)	365

Table 4.13: Confusion matric for logistic regression model five years prior to default including market variables

4.2.2. Neural Network

The design of the neural network at five years prior to default is the same as the neural network at one year prior to default with 20 input variables, one hidden layer with eight neurons within, and the two output neurons, "default" and "non-default".

When deciding the number of epochs in the model, figure 4.6 is used. It shows the number of epochs concerning the validation loos. It seems like the validation loss (yellow line) flatten out even earlier than the ones on the one-year horizon. Therefore, it is decided to use 50 epochs in the model to reduce the risk of overfitting as well as to reduce computational time.

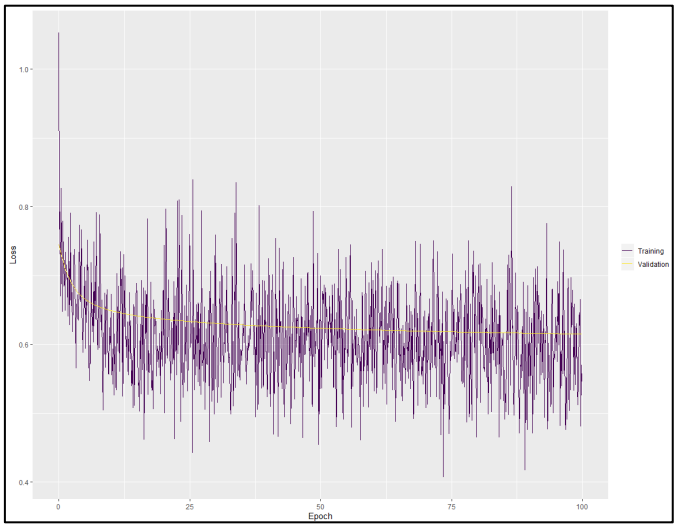


Figure 4.6: Graph showing the validation loss (the yellow line) on the data five years prior to default including market variables

The test sample was run through the network, and it classified 3,668 observations as "default" and 14,175 observations as "non-default". The empirical results obtained an accuracy of 80.57%, which is significantly lower than the test on a one-year horizon. However, this is not surprising as it is more difficult to predict correctly on a longer horizon. Table 4.14 shows the confusion matrix, where the model has 180 (32.20%) type 1 errors and 3,287 (19.02%) type 2 errors. It is worth to notice the high

percentage increase in type 1 error in the five-year horizon. The ROC curve of the model can be seen in appendix A figure A10, and the AUC equals 0.8075. This is a merely good result, according to table B4 in appendix B, and significantly lower compared to the neural network model one year prior to default and the worst on this data set.

Reality / Predicted	**Non-default H_0 is true**	**Default H_A is true**
Non-default H_0 is true	13,995	180 (32.20%)
Default H_A is true	3,287 (19.02%)	379

Table 4.14: Confusion matric for network model five years prior to default including market variables

4.2.3. Support Vector Machine

4.2.3.1. Linear SVM

This linear SVM model is, as for one year prior to default, tuned for C = $[10^{-4:4}]$ even though a wider range again could have been more preferred. The nine different models in the linear SVM can be seen in table 4.15. The best model to choose is the linearSVM8 with the lowest error of 0.3212 and a cost of 1000. This model contains 5,870 support vectors, 2,921 default and 2,949 non-default that are defining the hyperplane. Hence, 71% of the 8,296 data points in the training data set are support vectors which are considerably higher compared to linear SVM one year prior to default. This may indicate a risk of a relatively large number of misclassifications.

Model	**Cost**	**Error**
LinearSVM1	0.0001	0.3952
LinearSVM2	0.001	0.3343
LinearSVM3	0.01	0.3254
LinearSVM4	0.1	0.3245
LinearSVM5	1	0.3248
LinearSVM6	10	0.3245
LinearSVM7	100	0.3257
LinearSVM8	1,000	0.3212
LinearSVM9	10,000	0.4581

Table 4.15: Table showing nine different linear SVM models on the data five years prior to default including market variables

LinearSVM8 is then tested by using the testing data set. This test classified 3,058 observations as "default" and 14,785 observations as "non-default", and it obtained an accuracy of 83.57%. The confusion matrix, shown in table 4.16, indicates that the model includes 216 (38.64%) type 1 errors and 2,715 (15.71%) type 2 errors, which is a high increase in the type 1 error rate and a smaller increase in type 2 error rate. The ROC curve of the model can be seen in appendix A figure A11 and the AUC equals 0.8137 which is a merely good result, according to table B4 in appendix B, and still much lower than the linear SVM model one year prior to default.

Reality / Predicted	**Non-default** H_0 **is true**	**Default** H_A **is true**
Non-default H_0 **is true**	14,569	216 (38.64%)
Default H_A **is true**	2,715 (15.71%)	343

Table 4.16: Confusion matric for linear SVM model five years prior to default including market variables

4.2.3.2. RBF SVM

The RBF SVM model has two parameters to tune. It is, as for one year prior to default, tuned for C = $[10^{-4:4}]$ and value of gamma = $[2^{-2:2}]$. With this tuning 45 RBF SVM models are created where RBFSVM21 is the most preferred one due to the lowest error of 0.2902. Furthermore, it obtained a gamma of 0.25 and a cost of 1. The best model, as well as the ones around this model, can be seen in table 4.17. To see all 45 RBF SVM models, see appendix B table B6. RBFSVM21 includes 5,951 support vectors, where 3,111 of them are categorised as "default" and 2,840 are categorised as "non-default". Hence, 72% of the data points in the training data set are support vectors which are similar to the amount in linear SVM above and therefore, it also may indicate a relatively large number of misclassifications.

Model	Gamma	Cost	Error
RBFSVM17	0.50	0.1	0.3308
RBFSVM18	1.00	0.1	0.3802
RBFSVM19	2.00	0.1	0.4381
RBFSVM20	4.00	0.1	0.4937
RBFSVM21	0.25	1	0.2902
RBFSVM22	0.50	1	0.2965
RBFSVM23	1.00	1	0.3106
RBFSVM24	2.00	1	0.3310
RBFSVM25	4.00	1	0.3669

Table 4.17: Table showing the best RBF SVM model and the models around this on the data five years prior to default including market variables

The chosen model, RBFSVM21, is then tested against the testing data set. This test classifies 4,809 of the observations as "default" and 13,034 of the observations as "non-default". The test of the chosen model results in an accuracy of 74,65%. The confusion matrix seen in table 4.18 shows that the testing results in 137 (24.51%) type 1 errors and 4,387 (25.38%) type 2 errors. Both the percentage of type 1 and type 2 errors have increased relative to the RBF SVM one year prior to default. The ROC curve of the model can be seen in appendix A figure A12 and the AUC equals 0.8144 which is a merely good result, according to table B4 in appendix B, and close to the other AUC's five years prior to default.

Reality / Predicted	**Non-default** H_0 **is true**	**Default** H_A **is true**
Non-default H_0 **is true**	12,897	137 (24.51%)
Default H_A **is true**	4,387 (25.38%)	422

Table 4.18: Confusion matric for RBF SVM model five years prior to default including market variables

4.2.4. Random Forest

The random forest model is at first being trained with respect to the number of trees (ntree) and the number of variables sampled as candidates at each split (mtry).

Figure 4.7 shows the first random forest model being trained with 500 trees. At a close look, it is possible to see the error rate decreases just after 300 trees and increases before 500 trees. Therefore, it is decided to use 400 trees in the random forest model five years prior to default. Mtry is another parameter in the model, which is typically the squared root of the number of variables. Mtry is set to four as the starting point, and it is tested how the error rate will react if mtry is increased or decreased. Both ways of changing mtry resulted in higher error rates which indicate that four is a good number for the parameter. The figure of the tuning with respect to mtry can be seen in the appendix A figure A13.

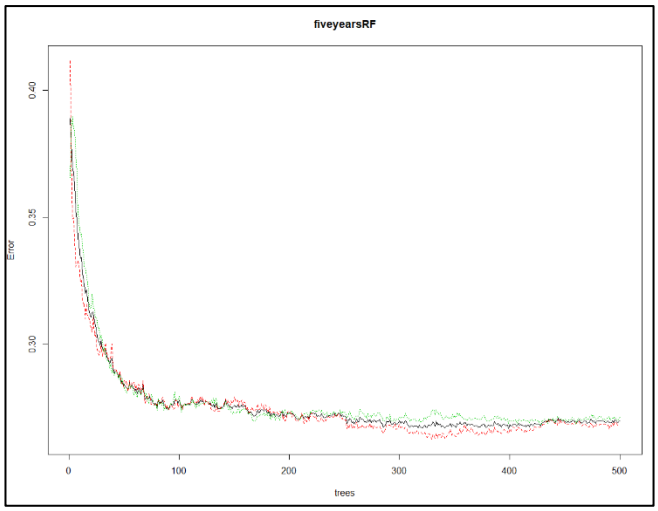


Figure 4.7: Figure showing the plot of the error rate (the black graph) of random forest on data five years prior to default including market variables

Figure 4.8 shows the relative importance among the variables included in the model for two measures which was described in section 4.1.4. It shows that OCFTA, TLMETL, and METL are all on top 4 for both measures of variables importance. On the other hand, X.NI and FDCF are both on the bottom for the two measures. One interesting point may also be that some variables are placed significantly differently on the importance of one year prior to default compared to the five years prior to default. One example is SLTA, which is the second-lowest importance one year prior to default, but the variable seems more valuable in this model.

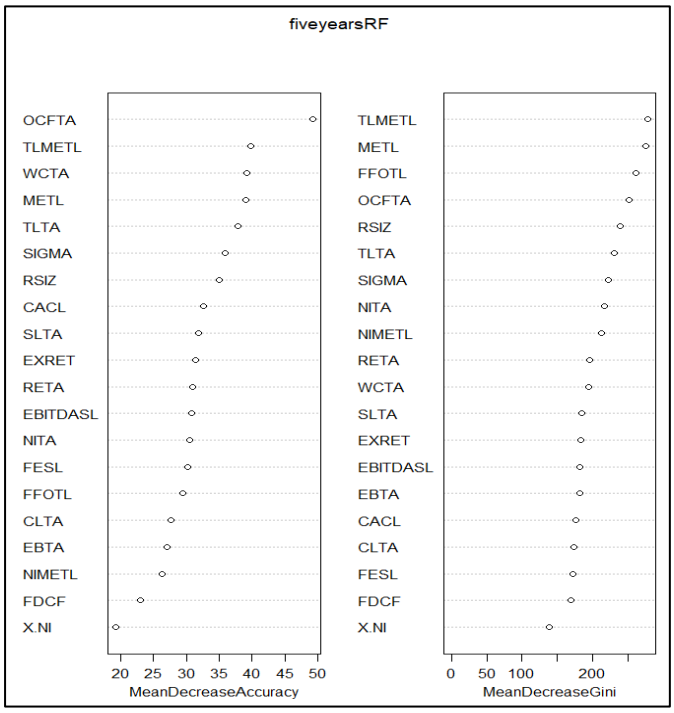


Figure 4.8: Figure showing the importance of each variable in random forest regarding the accuracy and the Gini index on the data five years prior to default including market variables

The random forest model was run at the testing sample and classified 4,565 as "default" and 13,278 as "non-default". This result gave an accuracy of 76,01%, which is nearly ten percentage-points lower compared to one year prior to default. The confusion matrix, which can be seen in table 4.19, demonstrated 137 (24,51%) type 1 errors and 3,289 (23,97%) type 2 errors. It is worth to notice the percentage of type 1 and type 2 errors are close to one another compared to other models on five years prior to default. The ROC curve of the model can be seen in appendix A figure A14 and the AUC equals 0.8367 which is a good result, according to table B4 in appendix B, and especially compared to the other AUCs, since it is the best one on this data set. However, it is still considerably lower compared to the random forest model one year prior to default.

Reality / Predicted	**Non-default** H_0 **is true**	**Default** H_A **is true**
Non-default H_0 **is true**	13,141	137 (24,51%)
Default H_A **is true**	4,143 (23,97%)	422

Table 4.19: Confusion matric for random forest model five years prior to default including market variables

4.2.5. The Machine Learning Methods Compared to One Another

This part gives a summary of the results for the five models built on the data from five years prior to default. The part finishes with an analysis of which model performs best on the testing sample.

Model	Type 1	Type 2	Accuracy	AUC
LR	194	3,181	81.09%	0.8185
NN	180	3,287	80.57%	0.8075
LinearSVM	216	2,715	83,57%	0.8137
RBFSVM	137	4,387	74.65%	0.8144
RF	137	4,143	76.01%	0.8367

Table 4.20: Summary of the chosen models in each method and their most important results of type 1 errors, type 2 errors, accuracy and AUC on the data five years prior to default including market variables

Table 4.20 shows the most important results between the models. It can be concluded that linear SVM has the highest accuracy at 83.57%, followed by logistic regression and neural network. All the models have accuracies that are considerably lower compared to one year prior to default, but this is to be expected. It is much harder to classify all five firms-year correct up to default compared to just classify the last firm-year correct. Therefore, it is a fairly decent accuracy of 83,57% for the linear SVM model. However, the table also shows that the model with the highest accuracy is also the one that is worst at predicting default. Linear SVM has 216 (38,64%) type 1 errors but has the highest overall accuracy. Random forest and RBF SVM have both lower accuracies but only respectively 137 and 140 type 1 errors.

When analysing which model performs best in terms of the distribution of type 1 and type 2 errors, it is worth mentioning the difference in the empirical results of one year prior to default compared to five years prior to default. First of all, there are more observations in the default category for five years prior to default. This is because a firm that is going to default will have up to five firm-year observations where it belongs in the category “default”. This implies that the number of firms with type 1 errors are considerably smaller than the number of type 1 errors itself. In addition, when analysing the error cost of the models, it is better to have four years where the borrower fulfils its payment and one year where the firms default, compared to the situation where there is just one year up to default. All this will impact the error cost of type 1 errors to be smaller five years prior to default compared to one year prior to default. This should be remembered when analysing the error cost of the models. Type 1 errors are still more expensive than type 2 errors but not as much as one year prior to default.

The error cost of the models is visualized in the same way as in section 4.1.5. The cost of the different models is compared to the cost of logistic regression. Different weights are multiplied with type 1 errors to adjust for the error type being more expensive compared to type 2 errors.

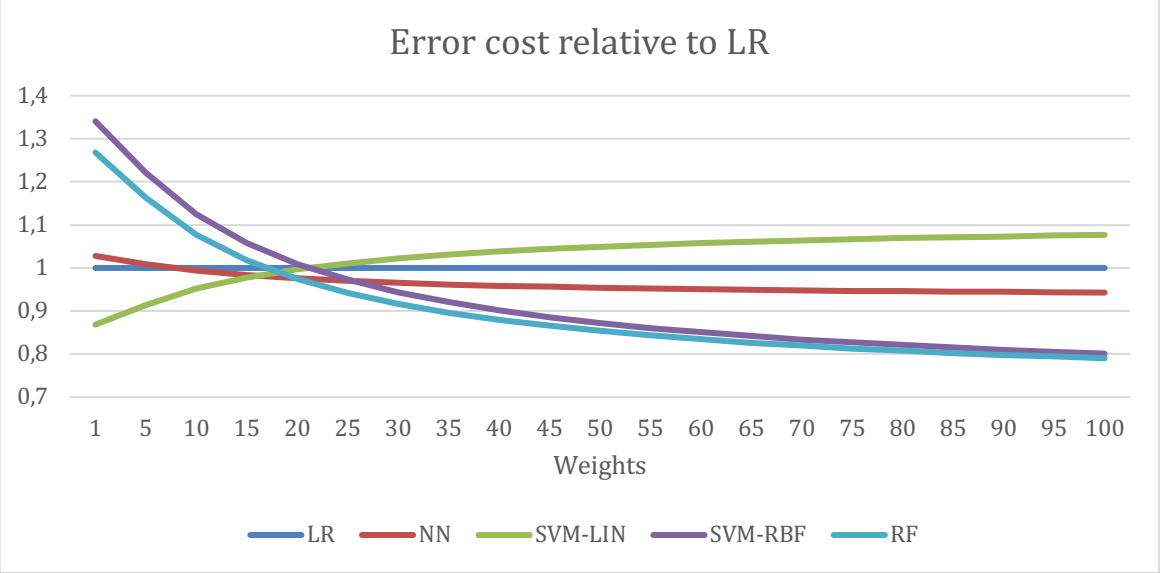


Figure 4.9: Figure showing the error cost relative to the logistic regression model on the data five years prior to default including market variables

Figure 4.9 shows the loss relative to logistic regression on the y-axis and the weights on the x-axis. When the weight is equal to 1, it means that type 1 and type 2 errors are equally costly, and thereby the accuracy determines which models are the best. From the beginning of the graphs, linear SVM is the best model until the weight equals to 15, where neural network overtakes the position. Neural network remains the best model until the weight equals 20, where random forest overtakes the position as the best model. It can be difficult to select one model as the best for this test sample, as it depends on how hard type 1 errors should be penalized. This comes down to the risk appetite for the credit lender and the historical cost of default compared to the opportunity cost of not giving credit to a non-default borrower. However, if one model should be selected as the best from this test sample, linear SVM is chosen since this model is by far the best model until the weight is equal to 15. Another argument for this is that the selected weight for the error cost must be lower at five years prior to default compared to one year prior to default, due to the relatively smaller error cost for type 1 errors on the five-year horizon. However, if the goal was just to predict "default" correctly random forest would be the most preferred model. It could also be argued that linear SVM could not be the best model since the error rate for type 1 errors is nearly 39% and another model with a lower error rate for type 1 errors should be used instead. Depending on the weights of the different error types, neural network or random forest would then be selected as the best model.

Regarding the other measure to validate the models, the AUC, the difference between the models in this measure is smaller compared to the accuracy. Random forest is the best model with an AUC of 0.8367,

where the rest of the models are within a very little AUC spread. However, all the AUCs have declined substantially compared to one year prior to default. All the ROC curves can be seen in figure 4.10. Again, the line for random forest (green) is furthest to the top left corner, which indicates the best model. However, for a false positive rate over 0.1, the models are very close, and there are some cases where other models than random forest are best. This also shows that the highest AUC does not necessarily mean that the model is best given all available threshold.

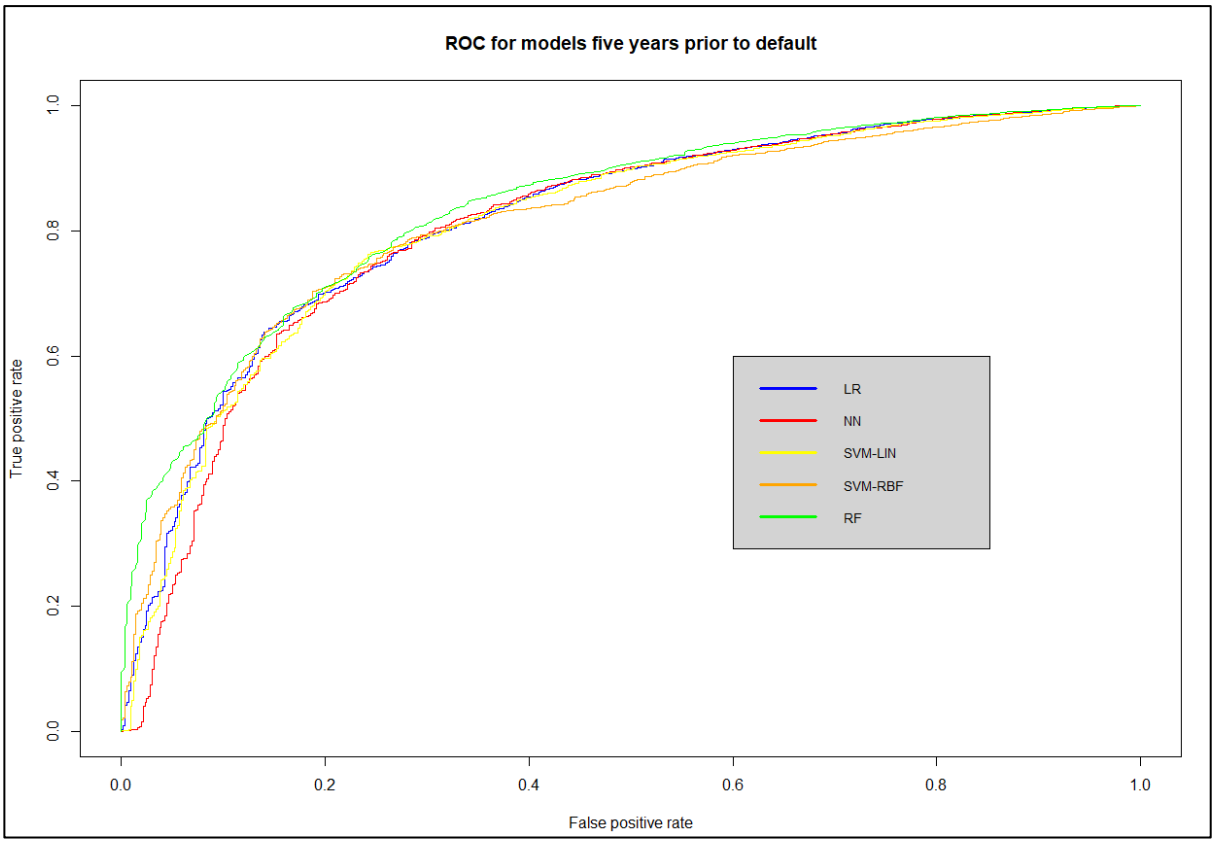


Figure 4.10: Figure showing the ROC curves of the chosen models in each method on the data five years prior to default including market variables

4.3. One Year Prior to Default Including ONLY Accounting Variables

The third part of this section contains the empirical results of the models created out of the training data set one year prior to default, where only accounting variables are included. The training data used for this part holds 2,156 observations similar to the training data in one year prior to default with both accounting and market variables. The difference is the fewer explanatory variables which are reduced from 20 to 14 since the market variables are excluded. Table 4.21 shows the correlation matrix of the remaining 14 variables. The correlation numbers are the same as in section 4.1, with the only difference of exclusion of the six market variables. The four earning to asset ratios RETA, EBTA, NITA, and OCFTA are still mutually correlated.

	WCTA	RETA	EBTA	SLTA	CACL	NITA	TLTA	FFOTL	X.NI	EBITDASL	OCFTA	FESL	FDCF	CLTA
WCTA	1	-	-	-	0.51	-	-0.67	-	-	-	-	-	-	-0.74
RETA	-	1	0.6	-	-	0.56	-	-	-	-	0.67	-	-	-
EBTA	-	0.6	1	-	-	0.86	-	-	-	-	0.89	-	-	-
SLTA	-	-	-	1	-	-	-	-	-	-	-	-	-	-
CACL	0.51	-	-	-	1	-	-	-	-	-	-	-	-	-
NITA	-	0.56	0.86	-	-	1	-	-	-	-	0.71	-	-	-
TLTA	-0.67	-	-	-	-	-	1	-	-	-	-	-	-	0.66
FFOTL	-	-	-	-	-	-	-	1	-	-	-	-	-	-
X.NI	-	-	-	-	-	-	-	-	1	-	-	-	-	-
EBITDASL	-	-	-	-	-	-	-	-	-	1	-	-0.78	-	-
OCFTA	-	0.67	0.89	-	-	0.71	-	-	-	-	1	-	-	-
FESL	-	-	-	-	-	-	-	-	-	-0.78	-	1	-	-
FDCF	-	-	-	-	-	-	-	-	-	-	-	-	1	-
CLTA	-0.74	-	-	-	-	-	0.66	-	-	-	-	-	-	1

Table 4.21: Correlation matrix for the variables in the data one year prior to default including only accounting

The binary dependent variable states “default” if the firm is going to default within the next year and “non-default” otherwise. The created machine learning models are then tested on the same testing data set, as described in section 4.1. The results for each machine learning method will be elaborated and analysed in the following.

4.3.1. Logistic Regression

LR1 is built with all 14 explanatory variables. This model ends up with nine statistically insignificant variables with a p-value higher than 0.05. A variable selection regarding insignificant variables is then used to minimize the risk of overfitting. The results of this can be seen in table 4.22.

Model	Variable(s) included in the model	Log-likelihood	AIC	# of insignificant variables
LR1	WCTA, RETA, EBTA, SLTA, CACL, NITA, TLTA, FFOTL, X.NI, EBITDASL, OCFTA, FESL, FDCF, and CLTA	-1133.822	2297,6	9
LR2	WCTA, RETA, EBTA, SLTA, CACL, TLTA, FFOTL, X.NI, EBITDASL, OCFTA, FESL, FDCF, and CLTA	-1133.828	2295.7	8
LR3	WCTA, RETA, EBTA, SLTA, TLTA, FFOTL, X.NI, EBITDASL, OCFTA, FESL, FDCF, and CLTA	-1133.941	2293.9	6
LR4	WCTA, EBTA, SLTA, TLTA, FFOTL, X.NI, EBITDASL, OCFTA, FESL, FDCF, and CLTA	-1134.312	2292.6	5
LR5	WCTA, EBTA, SLTA, TLTA, FFOTL, X.NI, OCFTA, FESL, FDCF, and CLTA	-1134.918	2291.8	4
LR6	WCTA, EBTA, SLTA, TLTA, FFOTL, X.NI, OCFTA, FDCF, and CLTA	-1135.656	2291.3	3
LR7	WCTA, EBTA, SLTA, TLTA, FFOTL, X.NI, OCFTA, and CLTA	-1136.472	2290.9	2
LR8	WCTA, SLTA, TLTA, FFOTL, X.NI, OCFTA, and CLTA	-1137.272	2290.5	0

Table 4.22: Table showing eight different logistic regression models on the data one year prior to default excluding market variables

When taking the maximum log-likelihood measure into account, the best model is LR1 with a log-likelihood of -1133.822, including all 14 explanatory variables. When taking the other measure, AIC, into account, the best model is LR8 which has the lowest AIC measure. LR8 is the model with the lowest log-likelihood measure but with the best AIC measure as well as being the only model without any insignificant variables left. Furthermore, LR8 is then the least complex model compared to the seven others. So, despite the worst log-likelihood measure, LR8 is chosen to be the best one also because the log-likelihood measures do not differ that much from each other. LR8 can be written as

$$probability\ of\ default = 1 - \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_7 x_7)}}$$

where

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_7 x_7 = -21.596 + 2.702\frac{Working\ capital}{Total\ assets} + 1.972\frac{Sales}{Total\ assets} - 35.228\frac{Total\ liabilities}{Total\ assets} + 15.246\frac{Funds\ from\ operations}{Total\ liabilities} + 1.191\Delta\ in\ Net\ income + 9.906\frac{Operating\ CF}{Total\ assets} - 4.423\frac{Current\ liabilities}{Total\ assets}$$

This model includes seven explanatory variables where five of them are the same as in the logistic regression model in section 4.1.1. The two added variables are WCTA and FFOTL. LR8 is then used on the testing data set. The test classifies 5,330 observations as "default" and 19,300 observations as "non-default". The model reached an accuracy of 78.92%. This accuracy is significantly lower compared to the one obtained on the similar test including market variables. The confusion matrix showing the misclassifications can be seen in table 4.23. The test results in 43 (19.11%) type 1 errors and 5,148 (21.09%) type 2 errors. It can be noticed that the number of type 1 errors are similar to those in section 4.1.1., but on the other hand, type 2 errors have had a high percentages increase. The ROC curve of the model can be seen in appendix A figure A15 and the AUC equals 0.8542 which is a merely good result according to table B3 in appendix B, but it is considerably lower compared to the model on the same time horizon but including market variables.

Reality / Predicted	**Non-default** H_0 **is true**	**Default** H_A **is true**
Non-default H_0 **is true**	19,257	43 (19.11%)
Default H_A **is true**	5,148 (21.09%)	182

Table 4.23: Confusion matric for logistic regression model one year prior to default excluding market variables

4.3.2. Neural Network

The design of this neural network is closely related to the previous ones, but this time with only 14 input variables since the market variables are excluded. It has one hidden layer with eight neurons within and the two output neurons, "default" and "non-default".

It is tested how the number of epochs will impact the model and whether the number of epochs should be reduced to prevent overfitting the model. Figure 4.11 shows the plot of the validation loss (yellow line) in relation to the number of epochs. The validation loss is still declining at 100 epochs, indicating that the number should not be less than 100. On the other hand, the figure also shows that the declining trend is diminishing which implies that 100 epochs are a good number to choose for the model.

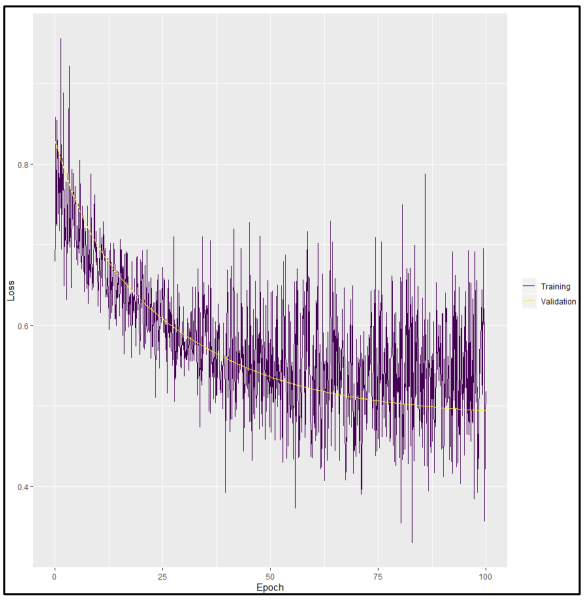


Figure 4.41: Graph showing the validation loss (the yellow line) on the data one year prior to default excluding market variables

The test sample was run through the network, and it classified 5,069 observations as "default" and 19,561 observations as "non-default". The empirical results obtained an accuracy of 79.91%, which is significantly lower than the test on the data, which included market variables. Table 4.24 shows the confusion matrix, and the model has 51 (22.67%) observations as type 1 errors and 4,898 (20.07%) observations as type 2 errors. It is worth noticing an increase in both type 1 and type 2 errors compared to the data including market variables. Especially in type 2 errors which increased by over nine percentage points. The ROC curve of the model can be seen in appendix A figure A16 and the AUC equals 0.8438 which is merely good according to table B3 in appendix B, but it is still significantly lower than the neural network model on the same time horizon including market variables and it is the lowest among the tested models.

Reality / Predicted	**Non-default** H_0 **is true**	**Default** H_A **is true**
Non-default H_0 **is true**	19,510	51 (22.67%)
Default H_A **is true**	4,898 (20.07%)	174

Table 4.24: Confusion matric for neural network model one year prior to default excluding market variables

4.3.3. Support Vector Machine

4.3.3.1. Linear SVM

Once again, the linear SVM model is tuned for C = $[10^{-4:4}]$, and nine models are created by the tuning. They can be seen in table 4.25. Out of these nine models, linearSVM7 is the best one with the lowest error of 0.2445 and a cost of 100. This model includes 1,305 support vectors where 654 are categorised as "default", and 651 are categorised as "non-default". Thus, 60.53% of the data points in the training data set are support vectors. This is somewhat higher compared to the same data including market variables why this can indicate a relatively large number of misclassifications.

Model	**Cost**	**Error**
LinearSVM1	0.0001	0.4406
LinearSVM2	0.001	0.2672
LinearSVM3	0.01	0.2496
LinearSVM4	0.1	0.2449
LinearSVM5	1	0.2459
LinearSVM6	10	0.2463
LinearSVM7	100	0.2445
LinearSVM8	1,000	0.2468
LinearSVM9	10,000	0.2579

Table 4.25: Table showing nine different linear SVM models on the data one year prior to default excluding market variables

LinearSVM7 is then tested using the testing data set. The test classified 4,882 observations as "default" and 19,748 observations as "non-default" with an accuracy of 80.75%. The accuracy has decreased compared to the same test including market variables. The confusion matrix in table 4.26 shows that the testing on the model results in 42 (18.67%) type 1 errors and 4,699 (19.25%) type 2 errors. Notice, that the type 1 errors are completely the same as for the test including market variables, but type 2 errors, on the other hand, have increased over eight percentages-point. The ROC curve of the model can be seen in appendix A figure A17, and the AUC equals 0.8642. This AUC is merely good according to the table B3 in appendix B, and it is still lower than the linear SVM on the same time horizon including market variables.

Reality / Predicted	**Non-default** H_0 **is true**	**Default** H_A **is true**
Non-default H_0 **is true**	19,706	42 (18.67%)
Default H_A **is true**	4,699 (19.25%)	183

Table 4.26: Confusion matric for linear SVM model one year prior to default excluding market variables

4.3.3.2. RBF SVM

The RBF SVM method is tuned for C = $[10^{-4:4}]$ and gamma = $[2^{-2:2}]$. This tuning turns out in 45 RBF SVM models where RBFSVM22 is the one to prefer due to the lowest error of 0.2385. This model obtained a gamma of 0.50 and a cost of 1. The model and the models around this are shown in table 4.27. To see all 45 RBF SVM models, see appendix B table B7. RBFSVM22 includes 1,407 support vectors where 745 of them lay in the classification "default", and the other 662 lay in the classification "non-default". Thus, 65.26% of the data points in this training data set are support vectors which is not much higher than from the same test including market variables. This can indicate the risk of a relatively large number of misclassifications.

Model	**Gamma**	**Cost**	**Error**
RBFSVM18	1.00	0.1	0.2937
RBFSVM19	2.00	0.1	0.3419
RBFSVM20	4.00	0.1	0.4235
RBFSVM21	0.25	1	0.2412
RBFSVM22	0.50	1	0.2385
RBFSVM23	1.00	1	0.2454
RBFSVM24	2.00	1	0.2616
RBFSVM25	4.00	1	0.2871
RBFSVM26	0.25	10	0.2440

Table 4.27: Table showing the best RBF SVM model and the models around this on the data one year prior to default excluding market variables

RBFSVM22 is then tested using the testing data set. This test classifies 5,206 observations as "default" and 19,424 observations as "non-default". It results in an accuracy of 79.45%, which is slightly lower than the one obtained in the similar test including market variables. The confusion matrix where misclassifications are highlighted can be seen in table 4.28. It shows that the testing results in 40 (17.78%) type 1 errors and 5,021 (20.57%) type 2 errors. Both types of errors have increased slightly in percentages compared to the similar test including market variables. The ROC curve of the model

can be seen in appendix A figure A18 and the AUC equals 0.8713 which is a merely good result according to table B3 in appendix B. This AUC is close to the AUC of the RBF SVM model on the same time horizon including market variables, and it is also among the highest AUC on the data set for the tested models.

Reality / Predicted	**Non-default H_0 is true**	**Default H_A is true**
Non-default H_0 is true	19,384	40 (17.78%)
Default H_A is true	5,021 (20.57%)	185

Table 4.28: Confusion matric for RBF SVM model one year prior to default excluding market variables

4.3.4. Random Forest

The random forest model is at first being trained with respect to the number of trees (ntree) and the number of variables sampled as candidates at each split (mtry).

Figure 4.12 shows a plot of the error rate in relation to the number of trees. It seems like, after 100 trees, the error rate does not fall significantly but only has some small movement up and down. Between 300 and 400 trees, it can be argued that there is some kind of local minimum. Therefore, it is decided to use 350 as the number of trees in the model. As mentioned in section 4.1.4, it is common to take the square root of the number of variables as mtry. With only 14 variables now, it is decided to keep the starting point of the test to four mtrys, and then it was increased and decreased to see how the error rate reacts. Appendix A figure A19 shows the result that both an increase and a decrease in mtry would increase the error rate.

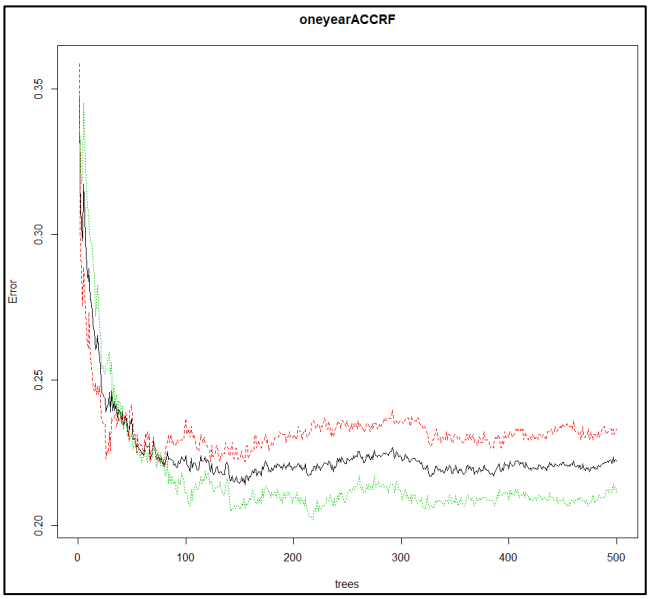


Figure 4.15: Figure showing the plot of the error rate (the black graph) of random forest on data one year prior to default excluding market variables

The random forest model can show the importance of variables that are included in the model. Figure 4.13 shows a plot of the 14 explanatory variables and their importance in the measures mean decrease in accuracy and mean decrease in Gini. The most important variable is for both measures TLTA, while the least important variables are X.NI, RETA, and SLTA. It is interesting to notice that TLTA has overtaken the leading role as the most important accounting variable compared to the data, which included market variables. Previously both OCFTA and WCTA were more important than TLTA in terms of accuracy, but it is no longer the case when the market variables are excluded. Even though the market variables are excluded there is still a risk of the importance of the variables does not give a true and fair ranking due to the correlation between the variables. As described in the correlation matrix in table 4.21, four of the variables are highly correlated, which will have an impact on their importance stated here.

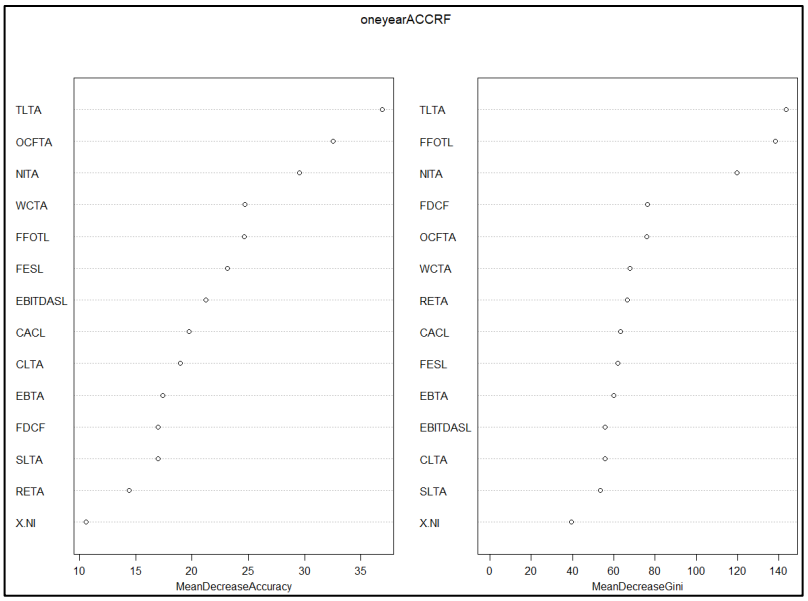


Figure 4.13: Figure showing the importance of each variable in random forest regarding the accuracy and the Gini index on the data one year prior to default excluding market variables

The random forest model was run at the testing sample, and it classified 4,155 as "default" and 20,475 as "non-default". This result gave an accuracy of 83.70%, which is not far from the previous result of 85.70% for market and accounting variables. The confusion matrix, which can be seen in table 4.29, demonstrates 42 (18.67%) observations as type 1 errors and 3,972 (16.28%) observations as type 2 errors. This is a high percentage increase in type 1 errors from 12% in the data set including market variables to 18.67% in this data set excluding market variables. However, type 1 errors are still among the lowest for the tested models. The ROC curve of the model can be seen in appendix A figure A20, and the AUC equals 0.8925 which is a good result according to table B3 in appendix B. It is the best among the models for this data set. However, it is still lower than the random forest model at the same time horizon including market variables.

Reality / Predicted	**Non-default** H_0 **is true**	**Default** H_A **is true**
Non-default H_0 **is true**	20,433	42 (18.67%)
Default H_A **is true**	3,972 (16.28%)	183

Table 4.29: Confusion matric for random forest model one year prior to default excluding market variables

4.3.5. The Machine Learning Methods Compared to One Another

This part gives a summary of the results for the five models built on data from one year prior to default only including accounting variables. The part finishes with an analysis of which model performs best on the testing sample.

Model	Type 1	Type 2	Accuracy	AUC
LR	43	5,148	78.92%	0.8542
NN	51	4,898	79.91%	0.8438
LinearSVM	42	4,699	80.75%	0.8642
RBFSVM	40	5,021	79.45%	0.8713
RF	42	3,972	83.70%	0.8925

Table 4.30: Summary of the chosen models in each method and their most important results of type 1 errors, type 2 errors, accuracy and AUC on the data one year prior to default excluding market variables

Table 4.30 shows the most relevant testing information for the five models. The result shows a small difference in the type 1 error where RBF SVM, as the best model, has 40 and neural network, as the worst model, has 51. On the other hand, the difference in type 2 errors are significantly larger. The best model is random forest which only has 3,972 type 2 errors corresponding to an error rate of 16.28%. The worst model between the five tested is logistic regression which has 5,148 type 2 errors corresponding to an error rate of 21.09%. It is clear to see that type 2 errors are those that lead the accuracy. This means that random forest has the highest accuracy at 83.70% with a sizeable lead over linear SVM and neural network at respectively 80.75% and 79.91%. It is worth noticing that especially logistic regression, neural network, and linear SVM have lost a significant proportion of their accuracy with only accounting variables available. Previously their accuracies were around 88-89% where it now is around 79-81%. Random forest, on the other hand, has only "lost" about two percentage-point when the market variables were excluded. It is clear to say that the exclusion of market variables affects some models more than others.

The same method, as described in section 4.1.5, is used to analyse the best model build on data one year prior to default including only accounting variables. Different weights are multiplied with the type 1 errors to compare how the models manage to classify correctly at different error costs. The result can be seen in figure 4.14.

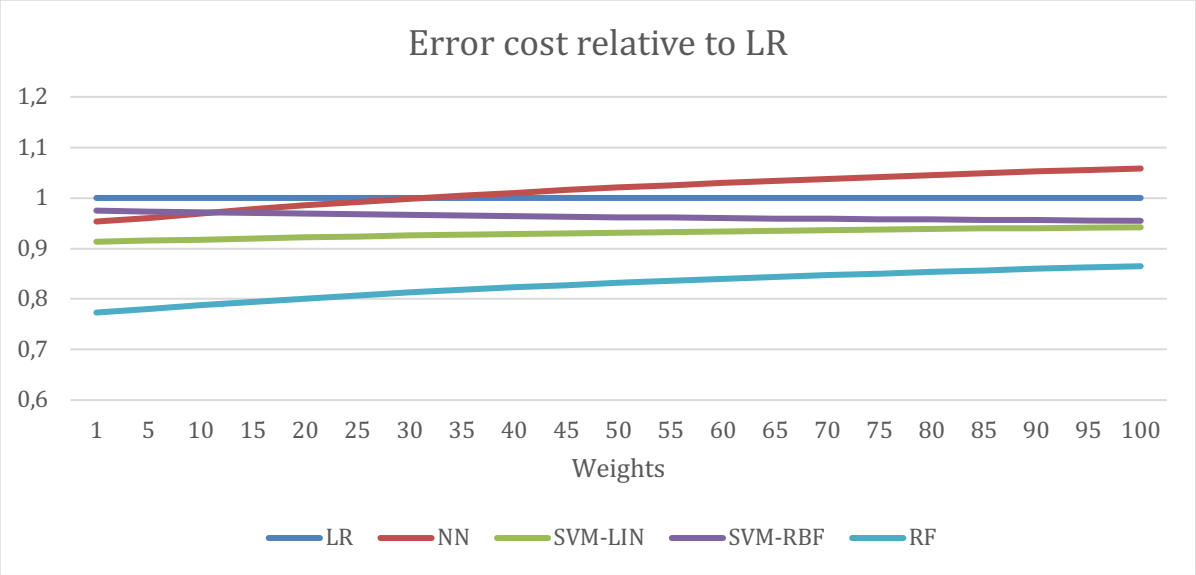


Figure 4.14: Figure showing the error cost relative to the logistic regression model on the data one year prior to default excluding market variables

The figure shows that for this data logistic regression model is the worst model when type 1 and type 2 errors are equally weighted, which was also demonstrated by the lowest accuracy. The logistic regression model is only crossed by neural network when the weight is equal to 35, which means type 1 errors are 35 times more costly than type 2 errors. On the other hand, random forest seems to be the best model through all weights up to 100. Only RBF SVM comes close when the weight increases as this model have the lowest number of type 1 errors and thereby the best model to classify the firms defaulting. However, when the model has more than 750 more type 2 errors and only two less type 1 errors compared to random forest it is fair to select random forest as the best model on the data including only accounting variables one year prior to default.

Regarding the other measure to validate the models, AUC, the best models are random forest and RBF SVM at respectively 0.8925 and 0.8713. neural network and logistic regression are the worst models according to AUC with only 0.8438 and 0.8542, respectively. The AUC also shows a larger spread among the models on this comparison relative to the comparison of the models including market variables. This also supports the argument that some models are more affected when the market variables are excluded. All the ROC curves can be seen in figure 4.15. It shows random forest as the best model for all false-positive rates as its green line is furthest to the top left corner at all times where RBF SVM is the only model that comes close at the beginning of the graph.

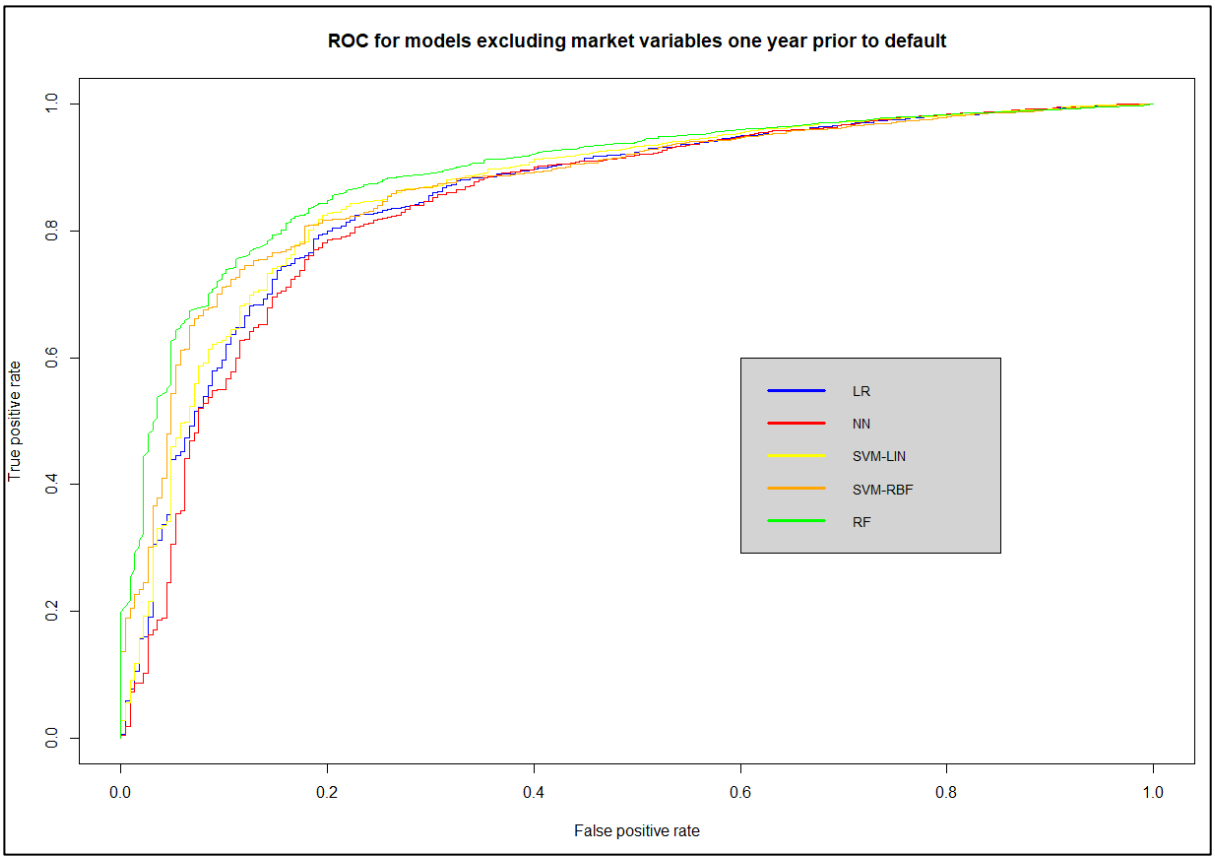


Figure 4.15: Figure showing the ROC curves of the chosen models in each method on the data one year prior to default excluding market variables

4.4. Five Years Prior to Default Including ONLY Accounting Variables

The fourth and last part of this section has obtained empirical results for the test of the models that are created out of the training data set five years prior to default, where only accounting variables are included. The training data set used for this part contains 8,286 observations, like in section 4.2. The data has 14 explanatory variables as well as a binary dependent variable that states "default" one year prior to default and "non-default" otherwise. As in section 4.2, the observations increased compared to one year prior to default because of the number of observations that are categorized as "default" within the next five years are by nature higher than within one year. The correlation matrix can be seen in table 4.31. It only shows the correlation pair with a value above |0.5|. The numbers are similar compared to the correlation matrix in section 4.2, though, with the exclusion of the six market variables. The four earning to asset ratios RETA, EBTA, NITA and OCFTA are still mutually correlated.

	WCTA	RETA	EBTA	SLTA	CACL	NITA	TLTA	FFOTL	X.NI	EBITDASL	OCFTA	FESL	FDCF	CLTA
WCTA	1	-	-	-	-	-	-0.64	-	-	-	-	-	-	-0.61
RETA	-	1	0.62	-	-	0.59	-	-	-	-	0.63	-	-	-
EBTA	-	0.62	1	-	-	0.86	-	0.53	-	-	0.85	-	-	-
SLTA	-	-	-	1	-	-	-	-	-	-	-	-	-	-
CACL	-	-	-	-	1	-	-	-	-	-	-	-	-	-
NITA	-	0.59	0.86	-	-	1	-	0.56	-	-	0.73	-	-	-
TLTA	-0.64	-	-	-	-	-	1	-	-	-	-	-	-	0.61
FFOTL	-	-	0.53	-	-	0.56	-	1	-	-	-	-	-	-
X.NI	-	-	-	-	-	-	-	-	1	-	-	-	-	-
EBITDASL	-	-	-	-	-	-	-	-	-	1	-	-0.64	-	-
OCFTA	-	0.63	0.85	-	-	0.73	-	-	-	-	1	-	-	-
FESL	-	-	-	-	-	-	-	-	-	-0.64	-	1	-	-
FDCF	-	-	-	-	-	-	-	-	-	-	-	-	1	-
CLTA	-0.61	-	-	-	-	-	0.61	-	-	-	-	-	-	1

Table 4.31: Correlation matrix for the variables in the data five years prior to default including only accounting

The test data is as well a bit different for five years compared to one year because of the period has been limited until 2013. The empirical results for each machine learning method will be elaborated and analysed in the following.

4.4.1. Logistic Regression

LR1 includes all 14 explanatory variables where four of them are statistically insignificant, with a p-value higher than 0.05. The empirical results of the variable selection regarding insignificant variables can be seen in table 4.32.

Model	**Variable(s) included in the model**	**Log-likelihood**	**AIC**	**# of insignificant variables**
LR1	WCTA, RETA, EBTA, SLTA, CACL, NITA, TLTA, FFOTL, X.NI, EBITDASL, OCFTA, FESL, FDCF, and CLTA	-5242.170	10514	4
LR2	WCTA, RETA, EBTA, SLTA, CACL, NITA, TLTA, FFOTL, X.NI, EBITDASL, OCFTA, FESL, and CLTA	-5242.172	10512	3
LR3	RETA, EBTA, SLTA, CACL, NITA, TLTA, FFOTL, X.NI, EBITDASL, OCFTA, FESL, and CLTA	-5242.180	10510	2
LR4	RETA, EBTA, SLTA, CACL, NITA, TLTA, FFOTL, X.NI, OCFTA, FESL, and CLTA	-5243.693	10511	1
LR5	RETA, EBTA, SLTA, CACL, NITA, TLTA, FFOTL, X.NI, OCFTA, and CLTA	-5244.914	10512	0

Table 4.32: Table showing five different logistic regression models on the data five years prior to default excluding market variables

With the maximum log-likelihood measure in focus, the best model is LR1 with a log-likelihood of -5242.170, which is a slightly better measure than LR2. LR1 is the model with the most statistically insignificant variables and hence the most risk of overfitting the model, which may result in poor results. When changing the focus to the AIC measure, LR3, on the other hand, are the model to prefer due to its lowest measure. Though the AIC measures do not differ much from each model, it can be argued to choose LR5 as the most preferred model because it does not have any statistically insignificant variables even though it has the lowest log-likelihood measure. LR5 can be written as

$$probability\ of\ default = 1 - \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_{10} x_{10})}}$$

Where

$$\begin{aligned}\beta_0 + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_{10} x_{10} \\ = -10.916 - 3.760 \frac{Retained\ earnings}{Total\ assets} + 3.827 \frac{Earnings\ before\ interest\ and\ taxes}{Total\ assets} \\ + 1.054 \frac{Sales}{Total\ assets} + 2.862 \frac{Current\ assets}{Current\ liabilities} - 3.783 \frac{Net\ income}{Total\ assets} \\ - 22.789 \frac{Total\ liabilities}{Total\ assets} + 9.097 \frac{Funds\ from\ operating}{Total\ liabilities} + 0.327 \Delta \text{ in Net income} \\ + 9.225 \frac{Operatting\ CF}{Total\ assets} - 1.876 \frac{Current\ liabilities}{Total\ assets}\end{aligned}$$

This model includes ten explanatory variables. Compared to section 4.2 with the same time frame but including market variables, LR5 contains six of the same accounting variables. LR5 adds four variables; EBTA, NITA, X.NI, and CLTA. When comparing to section 4.3 with different time frames but the same number of starting explanatory variables, LR5 includes six of the variables. The added variables are RETA, EBTA, CACL, and NITA, and the ones that have been dropped by this model is WCTA. LR5 is then tested on the testing data set. The test classifies 5,438 observations as "default" and 12,405 observations as "non-default". The model obtained an accuracy of 70.69%, which is significantly lower than the previously obtained accuracies regarding logistic regression. The test results in 175 (31.31%) type 1 errors and 5,054 (29.24%) type 2 errors. These misclassifications can be seen in the confusion matrix in table 4.33. The number of type 1 errors are decreased compared to section 4.2.1, while type 2 errors have increased by more than ten percentage-point. The ROC curve of the model can be seen in appendix A figure A21 and the AUC equals 0.7684 which is acceptable according to table B4 in appendix B. This AUC is significantly lower compared to the logistic regression model five years prior to default including market variables.

Reality / Predicted	**Non-default** H_0 **is true**	**Default** H_A **is true**
Non-default H_0 **is true**	12,230	175 (31.31%)
Default H_A **is true**	5,054 (29.24%)	384

Table 4.33: Confusion matric for logistic regression model five years prior to default excluding market variables

4.4.2. Neural Network

The design of this neural network is similar to the one in section 4.3.2, with only 14 input variables since the market variables are excluded. It has one hidden layer with eight neurons within and the two output neurons, “default” and “non-default”.

It is tested for how many epochs the network should run. Figure 4.16 shows a result of the validation loss (yellow line) on the y-axis and the number of epochs on the x-axis. It is clear that 100 epochs are too many as the validation loss has stagnated long before. Instead, it is decided to use 50 epochs in the model as this reduces the risk of overfitting; meanwhile, it is the same number used on five years prior to default which included market variables.

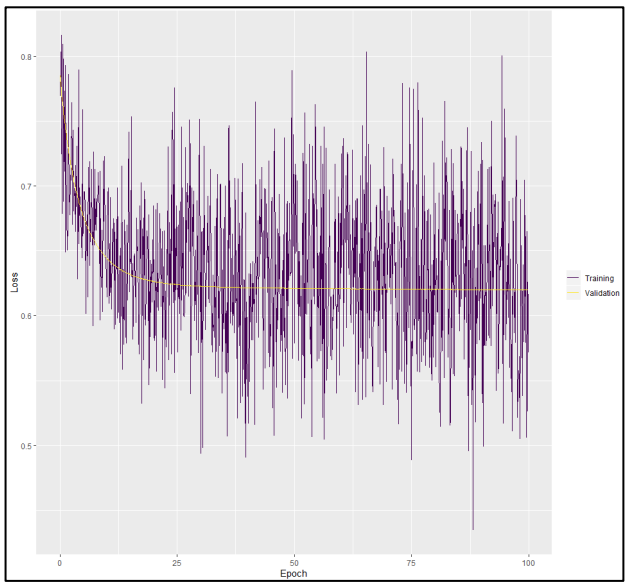


Figure 4.66: Graph showing the validation loss (the yellow line) on the data five years prior to default excluding market variables

The test sample was run through the network, and it classified 5,590 observations as "default" and 12,253 observations as "non-default". The empirical results obtained an accuracy of 69.80%, which is a significantly lower result compared to the other three networks in this thesis. Table 4.34 shows the confusion matrix, and the model has 178 (31.84%) type 1 error and 5,210 (30.14%) type 2 errors. The increase in type 2 errors compared to the neural network model five years prior to default, including market variables, is worth noticing. The error rate of type 2 errors increased with more than 11 percentages points, where the error rate for type 1 errors is nearly the same. The ROC curve of the model can be seen in appendix A figure A22 and the AUC equals 0.7583 which is acceptable according to table B4 in appendix B. The AUC is considerably lower compared to the similar model five years prior to default including market variables, and it is the worst among the tested models.

Reality / Predicted	**Non-default** H_0 **is true**	**Default** H_A **is true**
Non-default H_0 **is true**	12,075	178 (31.84%)
Default H_A **is true**	5,210 (30,14%)	381

Table 4.34: Confusion matric for neual network model five years prior to default excluding market variables

4.4.3. Support Vector Machine

4.4.3.1. Linear SVM

This linear SVM model is tuned for C = $[10^{-4:4}]$, and it results in nine models which can be seen in table 4.35. The most preferred model is linearSVM6 with the lowest error of 0.3455[1] and a cost of 10. This results in 6,572 support vectors where 3,287 of them are "default", and the rest 3,285 of them are "non-default". Thus, 79% of the data points in the training data set are support vectors which are the highest proportion compared to the previous SVM models why it might indicate a relatively large number of misclassifications.

[1] See output in appendix D code D4.3 for the precise decimals

Model	Cost	Error
LinearSVM1	0.0001	0.4357
LinearSVM2	0.001	0.3520
LinearSVM3	0.01	0.3484
LinearSVM4	0.1	0.3460
LinearSVM5	1	0.3455
LinearSVM6	10	0.3455
LinearSVM7	100	0.3471
LinearSVM8	1,000	0.4714
LinearSVM9	10,000	0.4605

Table 4.35: Table showing nine different linear SVM models on the data five years prior to default excluding market variables

LinearSVM6 is then tested using the testing data set. The test classifies 5,208 observations as "default" and 12,635 observations as "non-default" with an accuracy of 71.89% which compared to the previous linear SVM tests are significantly lower. The confusion matrix showing the proportion of misclassifications can be seen in table 4.36. The table shows the model results in 183 (32.74%) type 1 errors and 4,832 (27.96%) type 2 errors. Type 1 errors have increased slightly compared to section 4.2.3.1, while type 2 errors have increased by almost eight percentage-point. The ROC curve of the model can be seen in appendix A figure A23 and the AUC equals 0.7650 which is acceptable according to table B4 in appendix B. Once again, this AUC is lower compared to the similar model five years prior to default including market variables.

Reality / Predicted	**Non-default** H_0 **is true**	**Default** H_A **is true**
Non-default H_0 **is true**	12,452	183 (32.74%)
Default H_A **is true**	4,832 (27.96%)	376

Table 4.36: Confusion matric for linear SVM model five years prior to default excluding market variables

4.4.3.2. RBF SVM

The RBF SVM model is tuned for C = $[10^{-4:4}]$ and gamma = $[2^{-2:2}]$. This tuning resulted in 45 RBF SVM models where RBFSVM21 is the model to prefer due to the lowest error of 0.3044. It obtained a gamma of 0.25 and a cost of 1. This model and the ones around this model can be seen in table 4.37. To see all 45 RBF SVM models, see appendix B table B8. RBFSVM21 has 5,890 support vectors where 3,021 of them are defined as "default", and the other 2,869 are defined as "non-default". So, 71% of the

data points in the training data set are support vectors which are slightly higher compared to the chosen RBF SVM model in section 4.2.3.2. This indicates a risk of relatively large numbers of misclassifications.

Model	Gamma	Cost	Error
RBFSVM17	0.50	0.1	0.3201
RBFSVM18	1.00	0.1	0.3301
RBFSVM19	2.00	0.1	0.3611
RBFSVM20	4.00	0.1	0.4148
RBFSVM21	0.25	1	0.3044
RBFSVM22	0.50	1	0.3046
RBFSVM23	1.00	1	0.3110
RBFSVM24	2.00	1	0.3158
RBFSVM25	4.00	1	0.3279

Table 4.37: Table showing the best RBF SVM model and the models around this on the data five years prior to default excluding market variables

RBFSVM21 is then tested using the testing data set. This test classifies 4,992 observations as "default" and 12,851 observations and "non-default" with an accuracy of 73.41% which is lower than the precious results in RBF SVM. The confusion matrix shown in table 4.38 presents the testing results in 156 (27.91%) type 1 errors and 4,589 (26.55%) type 2 errors. Compared to section 4.2.3.2 type 1 errors increase almost by three percentage-point while type 2 errors only increase by 1.5 percentages-point. The ROC curve of the model can be seen in appendix A figure A24 and the AUC equals 0.7790 which is an acceptable result according to table B4 in appendix B, and it is among the best for this data set. However, it is still lower than the similar model five years prior to default including market variables.

Reality / Predicted	**Non-default H_0 is true**	**Default H_A is true**
Non-default H_0 is true	12,690	156 (27.91%)
Default H_A is true	4,594 (26.55%)	403

Table 4.38: Confusion matric for RBF SVM model five years prior to default excluding market variables

4.4.4. Random Forest

The random forest model is being trained with respect to the number of trees (ntree) and the number of variables sampled as candidates at each split (mtry).

The model is at first being trained with 500 trees. Figure 4.17 shows a plot of the error rate for the model on the y-axis and the number of trees on the x-axis. It can be seen that the error rate is declining when the number of trees is increased. Therefore, it is decided to keep the number of trees at 500 for this model. In relation to mtry it is tested how the error rate will react if mtry were increased or decreased from the starting point of four variables as candidates at each split. For both an increase or decrease in mtry, the error rate would increase, which means mtry should be four for this model.

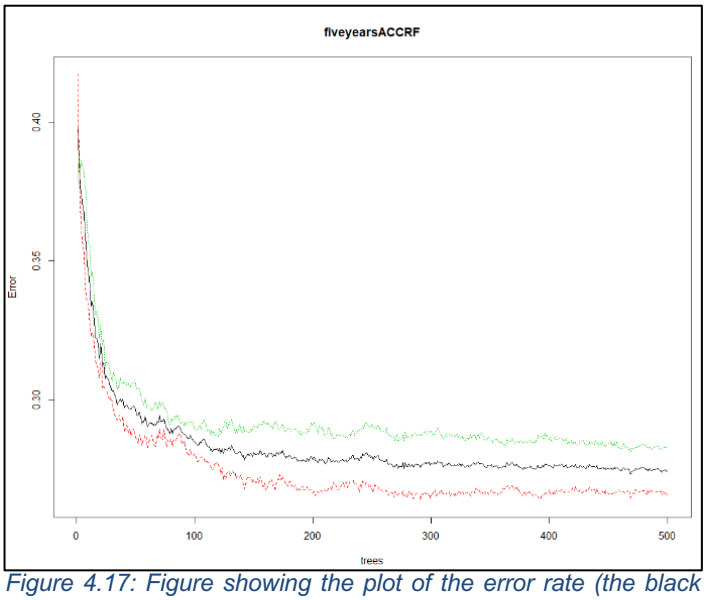


Figure 4.17: Figure showing the plot of the error rate (the black graph) of random forest on data five years prior to default excluding market variables

Random forest can show the mutual importance of the variables included in the model. This is visualized in figure 4.18 at the two measures, which are the mean decrease in accuracy and mean decrease in Gini. If this figure is compared to figure 4.13 based on the model for one year prior to default including only accounting variables there are similarities between the top and bottom of the score. TLTA is still the most important variable, while X.NI is the least important variable in both random forest models.

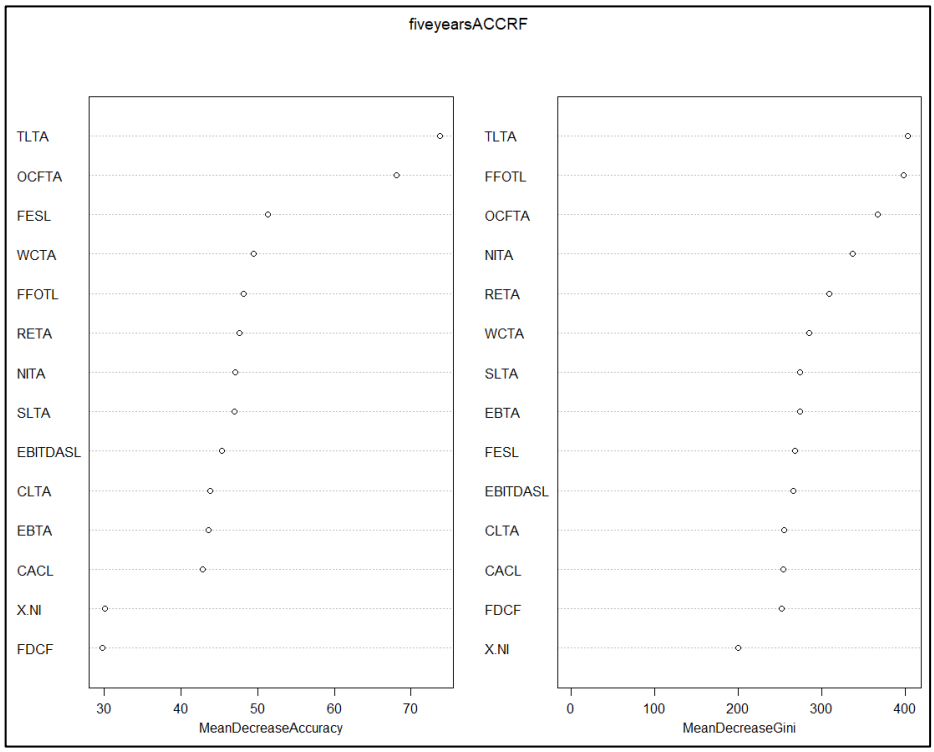


Figure 4.18: Figure showing the importance of each variable in random forest regarding the accuracy and the Gini index on the data five years prior to default including market variables

The random forest model was run at the testing sample for five years prior to default and classified 4,696 as "default" and 13,147 as "non-default". This result gives an accuracy of 75.05%, which is not far from the previous result of 76% for market and accounting data five years prior to default. The confusion matrix, which can be seen in table 4.39, demonstrates 157 (28.09%) type 1 errors and 4,294 (24.84%) type 2 errors. However, the confusion matrix shows a large percentage increase in type 1 errors compared to five years prior to default with market variables. This is important as type 1 error is the most costly error. The ROC curve of the model can be seen in appendix A figure A25 and the AUC equals 0.7988 which is merely good, according to table B4 in appendix B, but the best AUC for this data set. However, it is still lower than the similar model five years prior to default including market variables.

Reality / Predicted	**Non-default** H_0 **is true**	**Default** H_A **is true**
Non-default H_0 **is true**	12,990	157 (28.09%)
Default H_A **is true**	4,294 (24.84%)	402

Table 4.39: Confusion matric for random forest model five years prior to default excluding market variables

4.4.5. The Machine Learning Methods Compared to One Another

This part gives a summary of the most important results for the models built on data five years prior to default only including accounting variables. The part finishes with an analysis for which model is best on the data.

Model	Type 1	Type 2	Accuracy	AUC
LR	175	5,054	70.69%	0.7684
NN	178	5,210	69.80%	0.7583
LinearSVM	183	4,832	71.89%	0.7650
RBFSVM	156	4,589	73,41%	0.7790
RF	157	4,294	75.05%	0.7988

Table 4.40: Summary of the chosen models in each method and their most important results of type 1 errors, type 2 errors, accuracy and AUC on the data five years prior to default excluding market variables

Table 4.40 shows type 1 and type 2 errors, the accuracy, as well as the AUC for the models. Regarding type 1 errors, it is random forest and RBF SVM which performs best with respectively 157 and 156 type 1 errors. Regarding type 2 errors, it is also random forest and RBF SVM which performs best with respectively 4,294 and 4,594 type 2 errors. Therefore, the accuracy of these two models is the highest compared to all five models. It is worth to notice how the accuracy of logistic regression and neural network has changed significantly compared to section 4.2.5. For the data, including accounting variables, these two models had the highest accuracies among the five models. Now, when the data does not include market variables, these two models have the lowest accuracies. Both models had dropped around ten percentage-points in accuracy when the market variables were excluded. For random forest and RBF SVM, the drop in accuracy is only around two percentage points when the market variables were excluded.

The same method, as described in section 4.1.5, is used to analyse the best model on data five years prior to default only including accounting variables. The discussion in section 4.2.5 about the cost of type 1 errors compared to the cost of type 2 errors is also relevant in this part. So, even when type 1 errors are more costly than type 2 errors, the effect is greater one year prior to default compared to five years prior to default with or without market variables. Different weights are multiplied to type 1 error to compare how the models mutually manage to classify correctly. The result can be seen in figure 4.19.

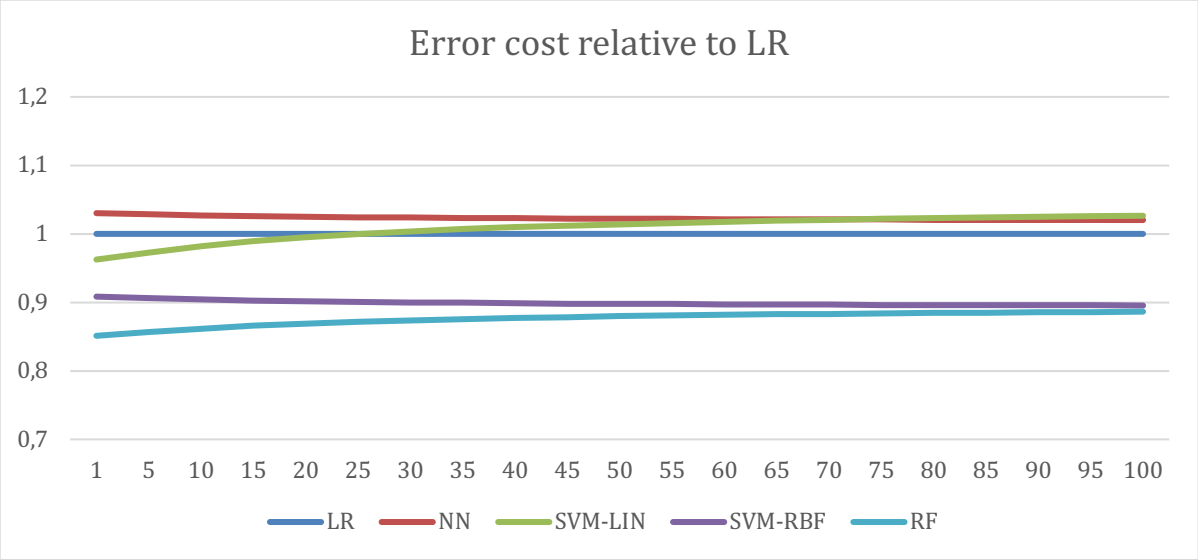


Figure 4.19: Figure showing the error cost relative to the logistic regression model on the data five years prior to default excluding market variables

The figure shows that random forest is the model with the lowest error cost among the five tested models up to weight equal to 100. The only one that gets close is RBF SVM when the weight is increased since the model has one less type 1 error. On the other end of the scale, logistic regression and neural network have the highest error cost when the weight is equal to one. When the weight is increased to over 30 linear SVM crosses logistic regression and becomes the model with the second-highest error cost. When the weight is higher than 70 linear SVM becomes the model with the highest error cost among the models. Therefore, it can be discussed which model is worst at classifying "default" and "non-default" for this data. However, it is not difficult to select the best model for five years prior to default only including accounting variables. Random forest has the highest accuracy, and it is the model with the lowest error cost up to type 1 errors are 100 times more costly than type 2 errors. This means random forest will be selected as the best model for this data in terms of the distribution of type 1 and type 2 errors.

Regarding the other measure to validate the models, AUC, the best models are random forest and RBF SVM at respectively 0.7988 and 0.7790. This means that random forest has a nearly 80% chance of separating correct between the classes. Neural network is the worst model, according to AUC, with only 0.7583. All the methods have their lowest AUC on this data set compared to the previous ones. This indicates the lack of market variables and the long horizon make it harder to separate between the classes. All the ROC curves for this data set can be seen in figure 4.20. Again, the line for random forest (green) is furthest to the top left corner, indicating the best model. RBF SVM is the only model that comes close at a false-positive rate around 0.2

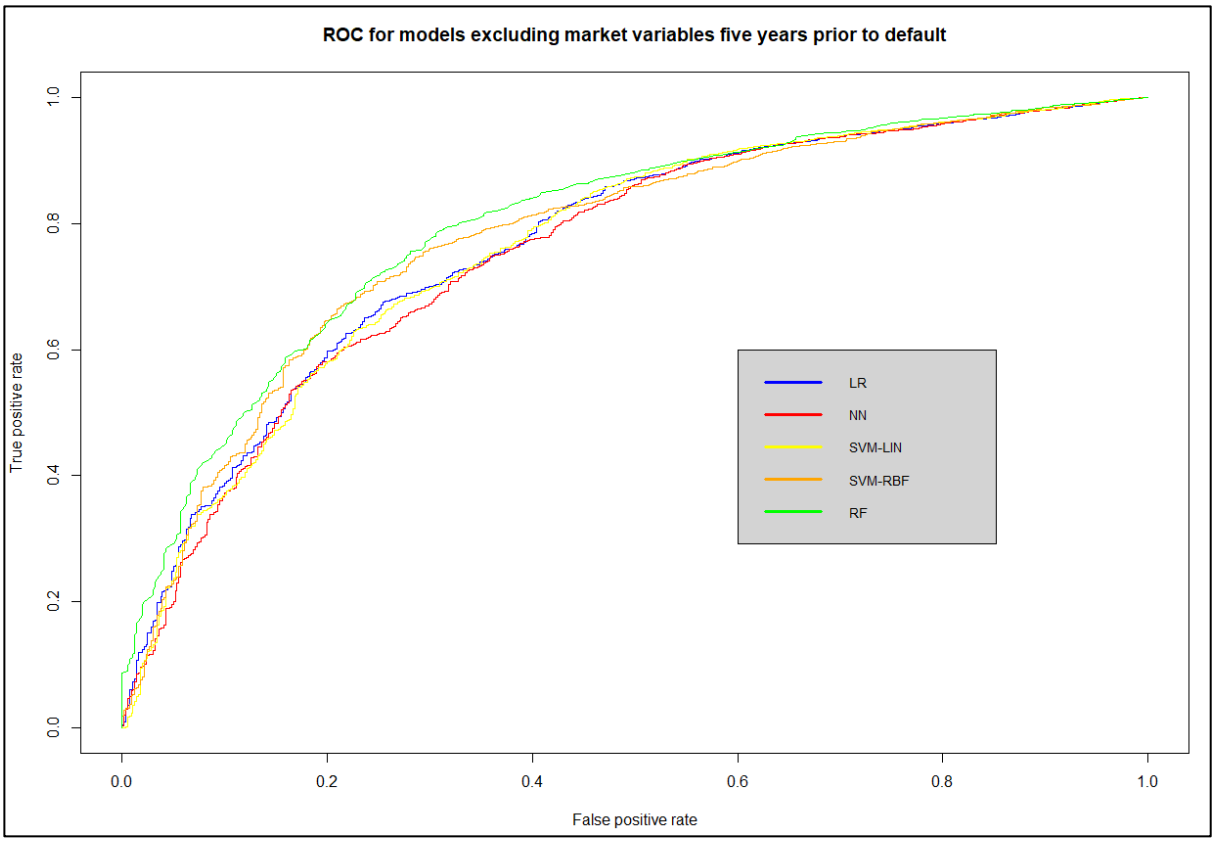


Figure 4.20: Figure showing the ROC curves of the chosen models in each method on the data five years prior to default excluding market variables

4.5. Sub Conclusion

This section analysed the empirical results of five different machine learning methods on four different data sets. First, it was found when predicting default one year prior to default based on data including accounting and market variables that two models were found to be the best depending on which measure to use. In terms of the distribution of type 1 and 2 errors, neural network is the best model until a weight of around 70, where random forest becomes the better one. When looking at the ROC curve and the AUC random forest is the model to prefer. Second, when predicting default five years prior to default based on data including accounting and market variables, the best methods changes. When looking at the distribution of type 1 and 2 errors, linear SVM, with the highest accuracy, is the best model until a weight of 15 where neural network becomes the best model until a weight of 20. With a weight higher than 20, random forest ends up as the best model. When looking at the ROC curve and the AUC, random forest is the best model. Third and fourth, when predicting default excluding market variables random forest is the best method in every aspect. It has the highest accuracy, the best distribution of type 1 and type 2 errors, and the best AUC for both the one year and fives years horizon.

This shows that the choice of measure to determine the best model has an impact on which model is chosen. The accuracy and distribution of type 1 and type 2 errors change the method of the best model when using different data sets whereas the ROC curve and the AUC indicate the whole time that random forest is the best model. Further elaboration and discussion of this can be found in the following.

5. Comparing the Machine Learning Models

This section makes an overall analysis and discussion of the models. The first part compares the models between data sets by looking at the accuracy and distribution of the error types and the ROC curves and the AUC. These evaluation measures, accuracy and AUC, are discussed regarding their advantages and disadvantages to see which one is the most preferred to use. The second part first separates the financial ratios into four categories, and the remaining market variables are in a fifth category called market information. This is followed by further analysis and discussion of the variable selection where the importance of the variables is discussed from the knowledge given in logistic regression and random forest. This part ends with discussing whether the industry levels should have been included in the thesis. The third and last part is a short discussion of whether market variables add predictive power.

5.1. Comparing the Models Between the Data sets

This part gives an overall comparison of how the five models perform on the four data sets. For every model and data set there is different measures for evaluating the performance of the models are created. The first part evaluates the models in terms of the accuracy and distribution of type 1 and type 2 errors. It is done by making a ranking of the models for each of the four data sets. The second part evaluates the models in terms of their AUC and ROC curves. Finally, the last part discusses which measure is the most important criteria when evaluating the models.

5.1.1. Accuracy and Distribution of the Error Types

The ranking of the models concerning the accuracy and distribution between type 1 and type 2 errors can be seen in table 5.1, where 1 indicates the best model and 5 indicates the worst model. The decision rule for determining which ranking each model achieves is based on the graph called “Error cost relative to LR” in the comparison part of each data set. The weight, which is used, should be lower five years prior to default compared to one year prior to default due to the argument that type 1 errors are relatively more costly one year prior to default as discussed in section 4.2.5. It is decided to use weights equal to 30 one year prior to default and weights equal to 10 five years prior to default. The way to get the ranking is to find the relevant weight, 10 or 30, and then go up on a vertical line until the first line is crossed. The first line and the belonging model will be ranked one, and the second line and the belonging model will be ranked two and so forth until it hits the last line.

Table 5.1 summarises this method for all the data sets and gives a total score for each of the five methods. For the first data set, neural network is the best method to classify default followed by linear SVM. For the second data set, linear SVM is the best model followed by neural network. For the last two data sets, it is random forest which is the best method to classify default followed by respectively

linear SVM and RBF SVM. However, neural network is the worst and second-worst classification method on the last two data sets while being among the best on the first two data sets. On the other hand, random forest is the second-worst method in the first two data sets while being the best model in the last two data sets. This also shows how some methods are more dependent on the market variables to make solid classifications. Neural network, logistic regression, and linear SVM all have significantly lower accuracies on the data sets excluding market variables compared to the same horizon including market variables. RBF SVM performs very badly on the first two data sets and the accuracy also falls on the last two data sets but not as much as the other three methods. Random forest, on the other hand, has the most impressive development from the first to the last data sets. The accuracy only fell around 1-2 percentage-points when the market variables were excluded. In addition, the method was also among the best for all data sets to classify defaulted firms correct and thereby reduce type 1 errors.

Model	One year incl.	Five years incl.	One year excl.	Five years excl.	Total
LR	3	3	5	4	15
NN	**1**	2	4	5	12
LinearSVM	2	**1**	2	3	8
RBFSVM	5	5	3	2	15
RF	4	4	**1**	**1**	10

Table 5.1: Table summarising the ranking of the models on each data set with the focus on the accuracy and the distribution of type 1 and 2 errors. This summarising ends in a total score for each of the five methods

Linear SVM is the method with the total lowest score indicating the best performing model, which is seen in table 5.1. Therefore, it could be argued that it is the best overall method in terms of accuracy and the distribution between type 1 and type 2 errors. One argument that speaks against this is the investigation on how close or far the models are from each other. This indicates how much the error cost will fall from changing from one model to a better one. One example is one year prior to default excluding market variables, illustrated in figure 4.14, where four of the models are relatively close while random forest is by far the best model. At the weight equal to 30, random forest is more than 11 percentage points better than the second-best model, which in this case is linear SVM. On the other hand, the situation is the opposite five years prior to default including market variables, where it is linear SVM which is the best model. This is illustrated in figure 4.9, and it shows that the model is only around four percentage points better than the second-best model. All this together shows that linear SVM might not be the best model even though the method has the lowest total score. The distances between the models are not shown in the simple ranking meaning some models can be "lucky" to get a better ranking but with an error cost very close to a worse ranked model. All this together shows that there is not one classification method that is superior on all data sets, and therefore, the best method changes between the data sets.

5.1.2. Best Model in terms of ROC and AUC

The other measure to validate the models is the AUC which is defined as the area under the ROC curve. All the AUCs for the five models on the four data sets can be seen in table 5.2. The result shows that random forest is the best method for all the data sets in terms of AUC. All the other methods seem to have AUCs close to each other with only neural network having a notable lower AUC on the two last data sets.

Model	One year incl.	Five years incl.	One year excl.	Five years excl.
LR	0.9092	0.8185	0.8542	0.7684
NN	0.9047	0.8075	0.8438	0.7583
LinearSVM	0.9049	0.8137	0.8642	0.7650
RBFSVM	0.8987	0.8144	0.8713	0.7790
RF	**0.9295**	**0.8367**	**0.8925**	**0.7988**
Average	0.9094	0.81816	0.8652	0.7739

Table 5.2: Table summarising the AUC of the models on each data set

However, recall figure 4.10 from section 4.2.5, which shows all the ROC curves in one diagram, the result might not be so unambiguous. From section 3.2.2.5 it was learned that the x-axis shows the false positive rate which is equal to the percentage of type 1 errors, and the y-axis shows the true positive rate which is equal to the percentage of "non-default" correctly classified. The graph shows that for some values of type 1 error rates the other models would have the same percentage of "non-default" classified correctly. From a type 1 error rate between 10% and 25%, all the models are fairly close to each other. Random forest seems to have the edge over the other models in the very small values for the type 1 error rate, which results in the highest AUC among the models.

In the analysis of the accuracy of the models, it was found that some methods were more affected when the market variables were excluded compared to others. This can also be shown by the ROC curves of the models. Figure 5.1 shows the ROC curves of random forest and neural network including and excluding market variables. The red and the green lines are the ROC curves including market variables for respectively neural network and random forest. Likewise, the pink and dark green lines are the ROC curves excluding market variables for neural network and random forest. If a type 1 error rate equals to 20% is used, the red and green lines are pretty close to one another. This means that the models are equally good at classifying "non-default" correct at roughly 90%. However, the result of the models is not the same on the data set excluding market variables. If the same type 1 error rate equals to 20% is used, the pink line is well below the dark green line. This indicates that random forest is significantly better than neural network at classifying "non-default" correct at this rate. It seems like the drop in type 1 error rate, from including to excluding, is twice as large for neural network compared to random

forest. This difference is larger for lower values of type 1 error rates but will diminish when the type 1 error rate is extremely high. In general, the distance from the pink line to the red line is larger than the distance from the dark green line to the green line. All this together shows that neural network as a method is affected significantly more when the market variables are excluded.

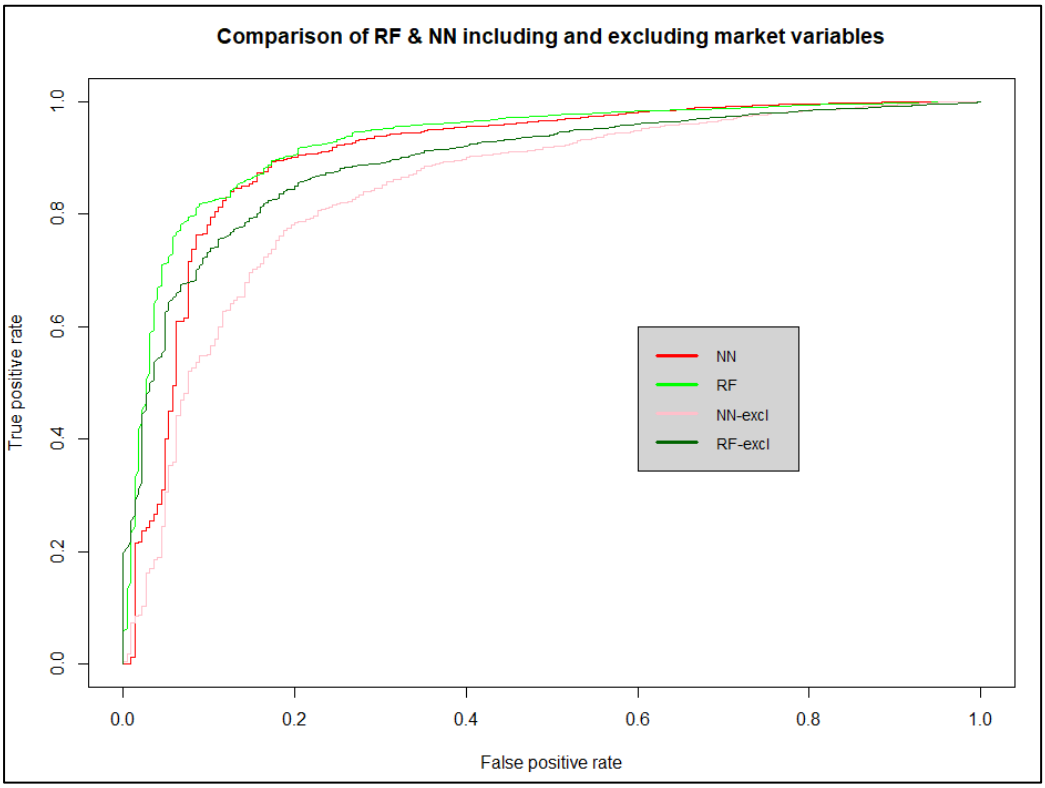


Figure 5.1: Figure showing the ROC curves of neural network and random forest, each method shows two ROC curves – one including market variables and another excluding market variables. NOTE: the ROC curves are on the same time horizon, one year prior to default.

5.1.3. Discussion of the Difference between Accuracy and AUC

The previous two parts showed that the result is different, whether the accuracy or the AUC is used to evaluate the models. In the ranking of the models in terms of accuracy and the distribution of type 1 and type 2 errors, linear SVM had the lowest total score indicating the best method on average. However, the model did also have some of the highest percentages of type 1 errors. On the other hand, random forest had the highest AUC for all the four data sets but had some of the lowest accuracies for the first two data sets. It can be difficult to select the best model as it might change from one data set to another, and it is not consistent between evaluation measures. However, there is more to say to this subject if the evaluation measures are investigated further.

There are some fundamental differences between accuracy and AUC, which lead to the different results of the best model. Basically, the accuracy focus on the total error rate, where the AUC focuses on the average error rates for type 1 and type 2. The accuracy is pretty simple as it gives the percentage correctly classified observations, but this accuracy percentage is only a snapshot of the truth at a given

cut-off. If the cut-off is changed, the accuracy would also change. In fact, it is only in logistic regression where the cut-off is explicitly set. For the other methods, the program chose the cut-off itself. However, it is possible to change the cost of the error types before the classification. This will change the result of the accuracy and thereby, the distribution of type 1 and type 2 errors. In this thesis, there has been a huge focus on the difference between type 1 and type 2 errors. The weights were added to give a more realistic total error cost of the model. In addition, the added weights also worked as penalisation to the models which tend to classify more observation as "non-default", when they actually "default", and thereby increase type 1 error. All this helps to give a more advanced view on which models were the best. However, it is worth noticing that all this effort to add weights to type 1 errors are made with the same cut-off. The accuracy does not tell anything about how the method performs if the cut-off is changed since it is a snapshot of one possible outcome.

The combination of the ROC curve and the associated AUC measure is very different compared to accuracy. Recall figure 4.5 from section 4.1.5, which illustrated all ROC curves one year prior to default including market variables. The graph illustrates the percentage of "non-default" correctly classified as a function of different type 1 error rates. An increase in the type 1 error rate will increase the percentage of "non-default" correctly classified. This also means that the "default" and "non-default" group are equally weighted in the ROC curve even though the data set is imbalanced. The test data for one year prior to default consist of 225 observations in the "default" group and 24,405 observations in the "non-default" group. This means that a one per cent change in the "non-default" group is equal to around 244 observations while one per cent change in the "default" default group is equal to around 2.3 observations. The ROC curve has rates in the axis, which means that the classes are equally weighted when the AUC is calculated. Said in another way, the "non-default" group on the y-axis in the ROC curve mainly determines the accuracy but the "default" group on the x-axis is equally important when it comes to the AUC measure. All this together explains the difference between accuracy and AUC and how the measures can result in different results.

The next question that arises is, what is the most important measure to evaluate the methods? It was elaborated on how the accuracy and the AUC resulted in different methods being the best among those tested. In addition, it was discussed why the measures had different results for the same models. Is the accuracy or the AUC the most important criterion when evaluating the models?

One of the advantages of ROC and AUC over the accuracy and type 1 and type 2 errors is the amount of information contained in the ROC curve. It shows all possible error rates split into the two classes. The accuracy only shows one possible outcome. Another advantage of ROC and AUC is the interpretation of a specific threshold. If the credit lender only allows having 10% of the "default" group being misclassified, the best model is the one which gives the highest true positive rate. In figure 5.2, the model will be random forest as this model has the line located above all the other at a type 1 error rate equal to 0.1.

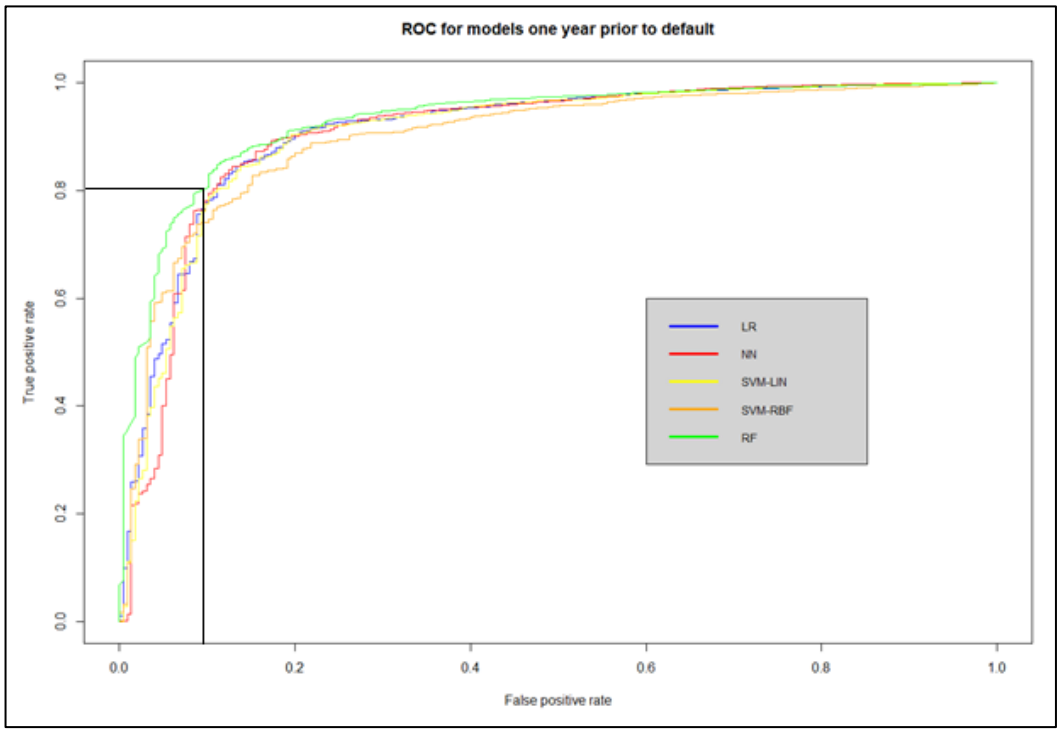


Figure 5.2: Figure 4.5 from section 4.1.5. with an illustration of a threshold of 10%

On the other hand, one may argue that AUC has some large disadvantages. First of all, the highest AUC might not be the most preferred model. For instance, when looking at the same graph, it seems like for type 1 error rates over 20%, all the models, except for RBF SVM, have roughly the same per cent of "non-default" correctly classified. The difference in the AUC comes from the very low values of type 1 error rates where random forest faster achieves a higher true positive rate. If the credit lender is not interested in having a model with such a low rate of type 1 error, the other models are just as good as random forest. In addition, the shaping of the ROC should be more decisive due to the fact there might be a situation where the highest AUC is not necessarily the best model for the credit lender. If ROC curves for two models cross each other, one model might be the best for one false positive rate, while the second model might be the best for a different false positive rate. The second disadvantage is that the AUC tells very little about the percentage of correctly classified observations when the data set is imbalanced. The last disadvantage of the AUC and also the ROC curve is the lack of easy interpretability compared to the accuracy. It takes some time to learn to interpret the ROC curve correctly and how to understand the AUC measure.

The accuracy and the distribution of type 1 and type 2 errors have some of the opposite advantages and disadvantages. The measure is easy to interpret since most people can understand the percentage of correctly classified. In addition, the distribution of type 1 and type 2 errors gives a good overview of how the errors are distributed. One of the disadvantages is the accuracy only shows one possible outcome of the model. Another disadvantage is the focus on one class if the data set is imbalanced. A high accuracy does not necessarily mean that the model performs well since it might just predict more observation in the large class. This is especially the case if the error cost for one type is higher than the other. The advantages and disadvantages of the two measures are summarized in table 5.3.

	Advantages	Disadvantages
ROC & AUC	It contains a lot of information in the ROC curve.	Hard to interpret
	Classes equally weighted	Highest AUC might not be the best model.
	ROC curve shows the relation between TP and FP.	Do not show the percentage correctly classified.
Accuracy	Easy to interpret	Lack of information – only one snapshot of the potential outcome.
	Overview of the absolute distribution of type 1 and type 2 errors	The highest accuracy might not be the best model.

Table 5.3: Summary of the advantages and disadvantages of the two evaluation measures, AUC and accuracy

There is no clear answer on which measure is the best to evaluate the models. The two measures complement each other and give different information about the result of the model. However, the accuracy might be the one that needs the most additional information to support the measure. Especially in the case, like this thesis, where type 1 errors are more costly compared to type 2 errors.

5.2. Variable Selection

The goal of this thesis is to make a multivariate rather than a univariate analysis regarding default prediction. As Beaver (1966) mentioned, the use of more ratios in one analysis may outcome the problem of univariate analysis where the prediction may end up giving different classifications for the same firm depending on the ratio chosen. Though, it is important to emphasize that the aim should not be to select too many variables why the variable selection is important. It is a way to keep the model simple, hence minimize the risk of overfitting when training the model. The following separates the financial ratios in four main categories and one category for the market information which are not categorised as financial ratios. Then a further variable selection is analysed and discussed with the main focus of the methods, logistic regression and random forest, that gives an indication of the importance

of the selected variables. Finally, the part ends with a discussion of the predictive power of market variables.

5.2.1. All Selected Variables

The starting point of the two data sets that includes market and accounting variables is 20 explanatory variables. These have been selected with the information from previous research studies regarding default prediction as described in section 2.2.1. The 20 variables are separated into 14 accounting variables and six market variables. Therefore, the two data sets, which only include accounting variables has 14 explanatory variables since the six market variables are excluded.

5.2.1.1. Financial Ratios

All 14 accounting variables, as well as three of the market variables, are calculated as financial ratios. A ratio by itself does not yield much information about the health of the firm. Though, when the ratios are compared to ratios from similar firms, to the firm's previous ratios, or to the required rate of the firm's return, they yield a wealth of information. The ratios of the thesis are split into four main categories: 1) profitability ratios; 2) liquidity ratios, 3) leverage rates, and 4) efficiency ratios.

Profitability Ratios

Profitability ratios are used to demonstrate a firm's ability to generate earnings relative to its revenue, operating costs, balance sheet assets, and shareholder's equity. In this thesis, the following six profitability ratios have been chosen:

- RETA (Retained earnings to total assets)
- EBTA (Earnings before interest and taxes to total assets)
- NITA (Net income to total assets)
- X.NI (Relative change in net income)
- EBITDASL (Earnings before interest, taxes, depreciation, and amortization to sales)
- NIMETL (Net income to the sum of market capitalization and total liabilities)

These ratios are put in this category since they are giving information about the future existence of the firm as well as the ability of the firm to ensure a satisfying return to shareholders.

Liquidity Ratio

Liquidity ratios are used to demonstrate a firm's ability to pay its short-term financial obligations, also known as total current liabilities, without raising external capital. In this thesis, three liquidity ratios have been chosen:

- WCTA (Working capital to total assets)
- CACL (Current asset to current liabilities)
- CLTA (Current liabilities to total assets)

These ratios are put in this category since they relate to the availability of cash and other current assets that can be converted into cash fast and cheap to cover current liabilities such as accounts payable, short-term debt, and other current liabilities.

Leverage Ratio

Leverage ratios are used to demonstrate a firm's ability to meet its financial obligations by looking at how much capital comes in the form of debt to finance its operations. In this thesis, the five following leverage ratios have been chosen:

- TLTA (Total liabilities to total assets)
- FFOTL (Funds from operations to total liabilities)
- FDCF (Financial debt to total cash flow)
- METL (Market capitalization to total liabilities)
- TLMETL (Total liabilities to the sum of market capitalization and total liabilities)

These ratios are put in this category since they evaluate the financial risk of the firm on a longer time horizon. Firms rely on a combination of equity and debt, and knowing the proportion of debt held by a firm is useful when evaluating whether it can pay back its debt as it comes due.

Efficiency Ratio

Efficiency ratios are used to demonstrate a firm's ability to use its assets and to manage its liabilities effectively in the existing period. In this thesis, the three following efficiency ratios have been chosen:

- SLTA (Sales to total assets)
- OCFTA (Operating cash flow to total assets)
- FESL (Financial expenses over sales)

These ratios are put in this category since they measure the time it takes to generate cash or income in relation to the total assets of the firm. This is not completely the case for FESL. Though FESL is put into this category since it did not match with any other categories, and the ratio shows how efficient the company is to generate revenue in relation to its financial expenses.

Common for all the financial ratios is that they are hard to use across industries since they have different conditions – they do not have the same asset base, same capital structure nor the same level of revenue in relation to its size.

5.2.1.2. Market Information

The data sets including market variables have six variables added to the originally 14 accounting variables. Three of them are mentioned above since they are categorized as financial ratios, whereas the other three market variables are categorized as market information:

- EXRET

- RSIZ
- SIGMA

These variables give each an indication on how the firm performs in terms of return, size and volatility in relation to the market.

All these variables should be seen as a whole when combining them rather than each separately since the use of multivariate analysis has been used in this thesis.

5.2.2. Further Variable Selection

To discuss the variable selection, the attention is turned towards logistic regression. All the models, except logistic regression, includes all the explanatory variables in the given data set. It is only the p-value of the coefficient in logistic regression that brings insight into whether a specific variable is significant. The chosen p-value of 0.05 has been used to determine whether the variable should be included or excluded from the model. The final logistic regression models for each data set can be seen in table 5.4.

Model	**# of variables**	**Variables included**
One year incl.	9	SLTA, TLTA, EXRET, RSIZ, SIGMA, X.NI, TLMETL, OCFTA, and CLTA
Five years incl.	11	RETA, SLTA, CACL, TLTA, EXRET, RSIZ, SIGMA, FFOTL, NIMETL, TLMETL, and OCFTA
One year excl.	7	WCTA, SLTA, TLTA, FFOTL, X.NI, OCFTA, and CLTA
Five years excl.	10	RETA, EBTA, SLTA, CACL, NITA, TLTA, FFOTL, X.NI, OCFTA, and CLTA

Table 5.4: Final logistic regression models for each data set showing number of variables and which variables are included in the final model

All the four models have three similar variables: two efficiency ratios, SLTA and OCFTA; and one leverage ratio, TLTA. Recall, that the prediction in R predicted "non-default" rather than "default" why the signs are opposite of the intuition in logistic regression. The variables are multiplied by a coefficient to show the weight of the variable in each equation and hence their predictive power in the model. Though, it should be remembered that the data is normalized which means that the coefficients cannot be interpreted directly regarding the real numbers. SLTA is multiplied with a coefficient between 1.054 and 2.817, TLTA is multiplied by a coefficient between -35.228 and -14.932, and OCFTA is multiplied by a coefficient between 6.432 and 9.906. It can then be argued that these variables have a general predictive power when predicting default, and TLTA is the one with the highest impact on the model, all things being equal since the weight is so high compared to the others. Though, among these three

variables, TLTA and OCFTA are highly correlated with respectively two and three other variables in all the data sets which can be seen in the introduction of each empirical result. TLTA is negatively correlated with WCTA and positively correlated with CLTA, whereas OCFTA is positively correlated with RETA, EBTA and NITA.

When comparing the two models based on the data including accounting and market variables, there are some similarities in the chosen variables aside from the three mentioned above. There are additionally four similar market variables with the same signs: three market information, excess return (EXRET), relative size (RSIZ), and volatility (SIGMA); and one leverage ratio, TLMETL. Excess return is multiplied by 5.228 and 1.182, relative size is multiplied by 2.019 and 1.903, volatility is multiplied by -3.901 and -3.273, and finally, TLMETL is multiplied by -32.277 and -38.052. These four variables can then be argued to have a general predictive power in predicting default when market variables are available. Recall that TLMETL is similar to TLTA with the difference in the use of the market value of equity instead of the book value of equity when calculating total assets. TLTA is still a part of both models even though market variables are included, and therefore TLMETL is available. TLMETL followed by TLTA have the highest impact in their models regarding the coefficient. This is an indication of the importance of the knowledge about the capital structure when predicting default on data where market variables are included. Furthermore, in the logistic regression model created on the data set one year prior to default including market variables, CLTA is included even though it is highly positively correlated with TLTA which is also included in the model. See the correlation table 4.1 in section 4.1. These are the only variables that are mutual highly correlated for that model. For the logistic regression model build on the data set five years prior to default including market variables, OCFTA is highly positively correlated with RETA, while TLMETL is highly negatively correlated with NIMETL. See the correlation table 4.11 in section 4.2.

When comparing the models excluding market variables with the models including market variables on the same time horizon, there are some similarities regarding the chosen variables aside from the previously mentioned ones. The two models, one year including market variables and one year excluding market variables, has two additional similar variables with the same sign: one profitability ratio, X.NI; and one liquidity ratio, CLTA. X.NI is multiplied by 0.529 and 1.191, and CLTA is multiplied by -3.365 and -4.423. In the model, one year prior to default excluding market variables, TLTA is highly correlated with WCTA as well as CLTA. See the correlation table 4.21 in section 4.3. The other two comparable models, five years including market variables and five years excluding market variables, have three similar additional variables: one profitability ratio, RETA; one liquidity ratio, CACL; and one leverage ratio, FFOTL. RETA is multiplied by -3.917 and -3.760, CACL is multiplied by 4.486 and 2.862, and FFOTL is multiplied by 7.940 and 9.097. In the model five years prior to default excluding market variables, TLTA is highly positively correlated with CLTA, and

OCFTA is highly correlated with RETA, EBTA, and NITA. Therefore, this model is in the risk of being affected by multicollinearity. See the correlation table 4.31 in section 4.4. Both models excluding market variables have three similar variables, FFOTL, X.NI, and CLTA. Therefore, it can be argued that these have a general predictive power when predicting default on data excluding market variables. Recall that TLTA is highly correlated with CLTA, and since TLTA is one of the variables that are included in all the logistic regression models, the inclusion of CLTA can be discussed. One of the assumptions in logistic regression is that the variables should not be highly correlated, which is not fully fulfilled in this paper. However, this is neither fully fulfilled in papers such as Barboza et al. (2017).

As mentioned, random forest includes all the explanatory variables that are in the given data set. Though, the importance of each variable can be analysed and discussed since it gives the mean decrease accuracy and the mean decrease Gini. Comparing all four models the profitability ratio, X.NI is the variable with the lowest importance of each model in relation to the accuracy as well as the Gini except the model five years prior to default excluding market variables which placed X.NI second-lowest according to the accuracy. Therefore, the importance of the variable can be discussed, whether it brings something to the model or not. However, this variable is included in the three out of four models in logistic regression, and it is not highly correlated with any of the other variables in the data sets. The correlation can have an impact on the importance of each variable. It can be argued not to make sense to exclude this from the data set since the models should on a be built on the same information.

With the above in mind, it can be discussed whether the methods that do not have the option to determine the importance of each variable or to define the variables which are insignificant, should have had an additional variable selection. Too many variables in the model may lead to overfitting and then a lower accuracy for the model. Though, the accuracies have been somewhat satisfying. When predicting default one year prior to default, the accuracies are between 78.92% and 88.97%, and when predicting default five years prior to default, the accuracies are lowered to be between 69.81% and 81.09% because of the higher uncertainty.

5.2.3. Industry Level

As it is shown until now, different methods are to prefer when different data sets are studied, which means no method is the best for every case it studies. This is also the case in different industries. The industries can be so different in the way the capital structure, asset base, etc. are. Therefore, it can be discussed whether each industry in the thesis mentioned in section 2.2.1.1 should have each its own model, or if they could have been divided further into different groups. The senior analyst (2020) in a credit lender institution described their credit models as many different models. The reason for this is that the lenders belong to different industries such as services, manufacturing, and retail. They cannot be compared in relation to financial ratios and the risk they are bearing since it is not the same.

Furthermore, historical information regarding paying back the loans as well as the manager's economic situation are taken into account. These different investigations result in the need for different models.

5.3. Including or Excluding Market Variables

This part sums up the comparison of those models including market variables and those models excluding market variables. The performance of all the models decreases when the market variables are excluded. This gives an indication of the relatively high predictive power of market variables. When looking at the accuracy and the distribution of the error types in section 5.1.1, it can be seen that methods that are best for predicting default when market variables are included are not necessarily the best ones for predicting default when market variables are excluded, and vice versa. Recall that the performance of neural network becomes a lot worse and random forest becomes relatively better when market variables were excluded. Overall regarding the data sets in this thesis, the power of market variables is concluded to be relatively high.

Though, market data is only available for listed firms, and since the biggest part of all firms in Denmark, and the world for that matter, are non-listed, models without market variables are needed for analysing the creditworthiness of the firm. These models only include accounting data, and the approach to collect this data is different from this thesis. They are in the need for the borrower to share given financial ratios which means there is a direct connection between the borrower and the credit lender institution regarding the needed information (Analyst, 2020). Therefore, data used to create the models are in practice internal information, the credit lender itself has collected through time rather than public statistical information.

5.4. Sub Conclusion

This section analysed and discussed the empirical results in section 4 regarding the measure to determine the best performing model, variable selection, and whether market variables have predictive power. First, it is found that linear SVM has the best overall ranking in terms of accuracy, but this result does not show every aspect since the method is only barely the most accurate model for one data set. In terms of the AUC and the ROC curve, random forest is the best model for all the data sets. It is also found that there are some fundamental differences between accuracy and AUC that leads to different results when deciding the best performing model. There is no clear answer on which measure to use when evaluating models. They complement one another by giving different information about the result of the model. Though, the accuracy might be the one that needs the most additional information to support the measure. Particularly, when the data set is imbalanced, and the type 1 errors are more costly compared to type 2 errors, as is the case for this thesis, or the other way around.

Second, it was found that the chosen ratios could be split into four different financial ratios containing the accounting variables and three of the market variables, and one group for the remaining market variables called market information. This is followed by a further variable selection. The different logistic regression models are compared in relation to the variables used. Here it is found that the accounting variables SLTA, OCFTA, and TLTA are selected for all the logistic regression models. It is also found that when market variables are included in the data set, the models include EXRET, RSIZ, and SIGMA. Though, when the market variables are excluded, the models add more accounting variables to compensate for the missing market variables. Second, the random forest models are compared in relation to the importance of the variables. Though, this importance can be discussed whether it brings some information about the models, since X.NI, the least important variable in three models in random forest, is included in three out of four models in logistic regression. Finally, the use of variables used on models for different industries is discussed since the ratios should be compared with firms within the same industry. The reason for this is that industries can be very different in the way the capital structure, asset base, etc. are.

Lastly, it is found that market variables have relatively high predictive power. The biggest part of all firms in the world are non-listed, and therefore they do not have market data available. Hence, models without market variables are needed for analysing the creditworthiness of the firm in practice. These models are created on the use of internal information the credit lender itself has collected.

6. Danish Market for Credit Lending

This section focuses on the Danish market for corporate credit lending. The empirical results of this thesis come on behalf of data from listed firms in the USA. Even though firms in the USA have different characteristics compared to Danish firms, the conclusions of the empirical results will still be transferred to the Danish market. It is assumed, when random forest was the best performing method on the data sets built upon USA firms, the result would be the same if the empirical result was built upon Danish firms. The focus will also be narrowed down to the data sets excluding market variables since most companies in Denmark do not have market variables available.

6.1. Power and Calibration for Credit Risk Models

This part takes the empirical results and reflect them to the risk management of the credit lender. The focus is on credit risk as this type is typically the most important risk measure for credit lenders. For the largest bank in Denmark, Danske Bank, the credit risk accounts for more than 70% of the total solvency requirement in 2018 (Danske Bank Group, 2018). As described in section 3.1.1, credit risk is the risk that the loans are not being repaid to the full extent, which will imply a loss for the credit lender. The banks in Denmark has an obligation to measure their credit risk since it determines the capital requirement, which should cover potential losses. The riskier the loans are in terms of credit risk, the more capital is needed for the bank.

In this thesis, the focus has been on testing the ability of different machine learnings methods to classify correct between “default” and “non-default”. However, the probabilistic output of the models has not been elaborated much. According to Stein (2007), there are two different main categories when it comes to model evaluation. These categories are power and calibration. The power is the ability of the model to separate between the classes which have been the primary focus of this thesis. The measures to evaluate the power of the model are the accuracy including the confusion matrix and the ROC curve including the AUC measure. The second category, the calibration of the model, is the matching and comparison of predicted probabilities of “default” with the observed indicators for “default” of the given classes (Nehrebecka, p. 4). The calibration is typically measured by the log-likelihood, which was seen in the result part of logistic regression. However, if the true calibration measure should be compared, the log-likelihood should be measured on the testing data instead of the training data. With only a minor change to SVM, all the methods result in a probabilistic output which can be used to test the calibration of the model. However, it would be very extensive work to include a test of the calibration of all the models. Therefore, it is assumed that the model with the better result in power also would have a better result in calibration. This is reasonable to assume since these models are already the best at separating among the classes. The assumption is supported by Stein (2007).

> ”This implies that a more powerful model will be able to generate probabilities that are more accurate than a weaker model, even if the two are calibrated perfectly on the same data set. This is because the more powerful model will generate higher probabilities for the defaulting firms and lower probabilities for non-defaulting firms due to its ability to discriminate better between the two groups and thus concentrate more of the defaulters (non-defaulters) in the bad (good) scores.”

The perfect calibration of the model can only be as good as the power allows it. Therefore, it is fair to assume that if the model has high power, it is also possible to calibrate the model well. With this in mind, the empirical results will be used to discuss the implication of risk management of the banks and more specific the credit risk. The result from the data set one year prior to default excluding market variables will be used since one year is the most common horizon to predict on, and the Danish market is dominated by non-listed firms.

A good model to predict “default” and “non-default” accurately has several benefits for the bank. Therefore, the credit risk models of banks are also a competition parameter. The next part focuses on two areas, namely capital requirements, including the use of IRB models as well as a more internal focus of the credit lender in terms of the advantages of a more accurate model.

6.1.1. Capital Requirement and the IRB Approach

As described in section 3.1, there are two different approaches for calculating the capital requirement concerning credit risk for the credit lender, which are the standardized approach or the IRB approach. For both approaches, the object is to calculate the risk-weighted asset (RWA). The RWA is used to calculate the capital requirement for both approaches. The current Basel accord states that total capital must be at least 8% of the RWA excluding a core equity conservativism buffer of 2,5% of RWA according to chapter 20.1 in BIS regulations (Bank for International Settlements, 2019c). The standardized approach calculates RWA on behalf of some predetermined weights multiplied the outstanding of the loans. The IRB approach calculates RWA on behalf of a specified formula and internally calculated values of PD, LGD, and EAD according to chapter 31.4 in BIS regulation (Bank for International Settlements, 2019b). Some would argue that credit lenders can use the standardized approach to calculate RWA and thereby, the capital requirement of the firm. If doing so, the credit lender should not bother finding a model that is both accurate and fulfilling the demands of an IRB model according to the Danish FSA. However, all five Systemically Important Financial Institutions (SIFI) use primarily the IRB approach (Erhvervs-, Vækst- og Eksportudvalget, 2018). When this is the case, there must be some kind of advantages using the IRB over the standardized approach.

In Denmark, most corporates have no external rating from any of the rating agencies. This means that the RWA will be 100% of the outstanding of the loan, under the standardised approach, since the standard weight for unrated corporates is 100% according to chapter 20.17 in BIS regulations (Bank for International Settlements, 2019a). According to a Danish expert group, the average weight under the IRB approach for larger risk exposures is around 40% (Erhvervs-, Vækst- og Eksportudvalget, 2018). This means that the capital requirement for corporates credit exposure under the standardized approach will be 2,5 times higher compared to the IRB approach. It might not be so drastic for other types of risk exposures. However, according to Sørensen (2013), the implementation of the IRB approach would make the RWA half the value after several years compared to continuing with the standardized approach. The decreased RWA from the IRB approach will lower the capital requirement for the credit lenders. This will allow the credit lenders to either have less capital reserved as a buffer or have a higher amount of lending, which will probably result in more earnings for the credit lender. However, there are several requirements needed to get permission from the Danish FSA to use the IRB approach.

One of the requirements is to have documentation of the rating system and the underlying model. This includes having data from a whole business cycle, documentation of validation, calibration, and validation methods for the model (The European Parliament and the Council of the European Union, 2013). In addition, other documentation like management reports and stress tests of the model is needed before the Danish FSA can allow the credit institution to use the IRB approach to calculate credit risk. Once the credit institution receives approval for the IRB approach, there is still demands to update and maintain the model (Analyst, 2020). However, it is nevertheless worth spending time on this since all SIFIs in Denmark use the IRB approach. In this thesis, it is shown how some models were more accurate compared to others. For the data set one year prior to default excluding market variables random forest was the best model while logistic regression was the worst model among those tested. From figure 4.14, it can be seen that random forest is significantly better than logistic regression for all given weights. If we assume the result can be transferred to the Danish credit institution market, the next question which may arise is whether this implies that the credit lenders are only allowed to use the model with the most accurate model validation. From the result of this thesis, it would be random forest since it reported a significantly better result. However, given the regulation from the Danish FSA, there are other conditions to take into account than just the model validation.

According to EU regulation for credit institutions, a statistical model must meet the following requirements. “The model shall have good predictive power and capital requirements shall not be distorted as a result of its use. The input variables shall form a reasonable and effective basis for the resulting predictions” (The European Parliament and the Council of the European Union, 2013, Article 174, a). The focus in this thesis is on the last part, which states that the input variables shall form a reasonable and effective basis for the result prediction. Recalling section 3.2.2 where the different

classification methods for predicting default were elaborated, it was found that several of the methods had a black box that determines the class of the observations. The black-box was investigated, which means that some kind of understanding is acquired of what happens inside the black-box. However, there is still no clear form of how the input variables produce a reasonable and effective bias for predicting correctly. The predictions are good and accurate, but it is very difficult to see how the different input variables affect the model as a whole. Only logistic regression shows a real measure for how the input variables affect the model as described in section 5.2.2. Neural network and SVM have a black box that makes it hard to find the true impact of each variable. For random forest the variables available for each split is different, which implies that there is no clear form for how the input variables produce the result of the predictions. Though, the output shows the importance of the variables according to the accuracy and the Gini. All this together shows, that even though random forest is significantly better as a model, it might not fulfil the technical requirements for statistical models to calculate the credit risk.

6.1.2. Internal Advantages of a Better Credit Risk Model

The last part discusses the result of the models in terms of the capital requirement with a focus on the difference between the IRB and the standardized approach. However, there are more areas where the credit risk model is used within the credit institution. This part discusses some of these areas, namely provision and the evaluation of a potential customer of the credit lender.

When lending corporates money there will be bad payers in all large portfolios. This means that the credit lender has some expected credit losses every year. It is the situation where the loan has been granted, but the borrower fails to fulfil its contractual obligation. The credit lender must be prepared for these losses, and it is exactly what the loan-loss provision does. It prepares the credit lender for borrower defaults on a proportion of the portfolio and set aside an amount for impairment losses. The credit risk model is also important when it comes to the size of the loan-loss provision for each year. The calculation for expected credit loss builds upon the PD and LGD for the portfolio. This means that a credit risk model that has a more accurate value for the PD gives the credit lender a more precise estimate for expected credit loss. Given the assumption that the best model in terms of power also will be the best model in terms of calibration, as discussed in section 6.1, random forest will generate more robust estimates for the probability of default. This will also imply that the expected credit loss for a better model will be more accurate compared to a worse model. Even though the accuracy is “only” about five percentage-point better for random forest compared to logistic regression, it would still make a great difference and give the credit lender a more accurate estimate for the loan-loss provision of the period.

Another area where the credit lender has an advantage with a better credit model is the evaluation of a potential customer before the loan will be granted. First of all, a more accurate model will imply the credit lender to give loans to the correct companies. It is essential for reducing the impairment losses to accept loan applications from good payers and reject loan applications from bad payers. This is very difficult to be certain about before the loan is paid off, but a more accurate model will help the procedure to evaluate potential customers and thereby reduce impairment losses. The goal for the credit lender is not to reduce the bad payers down to zero since the credit institution has its justification of existing to take risks. Though, the credit lender must have a transparent and accurate measure for its risk and here an accurate credit risk model can help to achieve it.

One last point where the credit lender can make use of a more accurate credit risk model is in the pricing of the loans. It is common to set the interest rate matching the risk of the firm borrowing the money. This means that a higher PD or LGD will result in higher interest for a potential customer. The estimates for PD and LGD are then crucial to be accurate to give a fair evaluation of the potential customer.

6.2. Results of the Thesis into the Perspective of the Litterature

This part puts the empirical results of the thesis into the perspective of the literature of similar studies. It will also give reasons why credit lenders might not be so ready to change their credit risk model to new and more accurate machine learning methods for predicting default.

In the literature, logistic regression is stated as the industry benchmark for credit scoring while new machine learning methods are being tested up against logistic regression (Lessmann, Baesens, Seow, & Thomas, 2015). Throughout this thesis, the classification result of logistic regression, neural network, linear SVM, RBF SVM, and random forest has been analysed on different data sets. For every data set, a comparison of the error cost is made with logistic regression as the benchmark of the analysis. Section 4.3.5 and 4.4.5 show how logistic regression has considerably higher error cost compared to primarily random forest. The error cost of random forest is nearly 20 percentage points lower than logistic regression in the data one year prior to default excluding market variables at the weight equal to 30. The AUC and the ROC curve show the same result and dominated logistic regression in all given thresholds for the same data set. These results are aligned with the ones in the literature where random forest is standing out as a very solid classification method (Lessmann, Baesens, Seow, & Thomas, 2015; Barboza, Kimura, & Altman, 2017). Lessmann et al. (2015) also argued that on behalf of their result, it is time to move away from logistic regression as an industry standard and instead towards new state-of-the-art classification methods. So, why do some Danish credit institutions not use random forest in its credit risk model? There might be two answers to this question – tradition and regulation. Credit institutions, and especially larger banks, are typically older organizations where experienced methods

are highly valued. This means that there might be a great aspect of tradition when determining the credit risk model. In addition, organisations like the Danish FSA, which must accept the credit risk model, are also influenced by tradition. However, in the past decade, several credit institutions have been through a digital transformation which means that the institutions are aware of the possibilities in machine learning. Nonetheless, the regulation is still a straitjacket when it comes to the flexibility of credit risk models. As discussed in section 6.1.1, there are requirements for the models of how the input variables should form reasonable bias for the result. In random forest the variables used at each split are different, and therefore it is questionable whether the method can be used as a credit risk model. On the other hand, logistic regression has some strong advantages when it comes to interpretability and how the input variables affect the result of the model as described in section 5.2.2. This implies that the credit institution might not change its credit risk method significantly before the regulation change.

6.3. Sub Conclusion

The previously parts show that larger credit institutions prefer the IRB approach over the standardized approach for measuring the credit risk on corporate exposure. In the IRB approach, the credit lender produces values for PD, LGD, and EAD from its credit risk model. The most important reason to use the IRB approach is the lower RWA which implies the credit lender with a lower capital requirement. There are several requirements in the implementation and use of the IRB approach. One of them is that the model should be valid and calibrated well.

It is shown that there more examples for the credit lender to benefit from a solid and accurate credit risk model. The credit institution needs to have a precise measure for the credit risk since it is the largest risk exposure for the firm. The credit risk model generates estimates which are used in several areas such as in the calculation of provision, evaluation a potential customer in terms of accepting the loan application and adjusting the interest rate, so it matches the risk.

The last part discussed how the literature plays a role in the findings of the thesis. It is shown that the literature with several articles also has random forest as one of the most solid classification methods. It is discussed why logistic regression cannot be rejected as the industry-standard even though other state-of-the-art classifiers outperform logistic regression numerous times. In this thesis, random forest also heavily outperformed logistic regression especially when it comes to the last two data sets which imply that the Danish credit institution can achieve a more accurate credit risk model by using this method. However, the traditions in the credit institutions and the regulation might prevent this from happening.

7. Conclusion

With the use of theory from finance and data science, the thesis performs a test on different machine learning methods to predict default. The test is built upon accounting and market data from firms in the USA from 1987 to 2015. The tested methods were logistic regression, neural network, linear SVM, RFB SVM, and random forest. The test was split into four different data sets regarding different time horizon and including or excluding market variables. The first data set tests the ability of the methods to predict default one year prior to default including market variables. The thesis finds that neural network is the best model in terms of accuracy and the distribution of type 1 and type 2 errors while random forest is the best model in terms of the ROC curve and AUC. For the second data set, linear SVM has the highest accuracy while random forest is still the best model in terms of the ROC curve and AUC. For the third and fourth data set excluding market variables, random forest is the best performing model in both the accuracy, the distribution of type 1 and type 2 errors, the ROC curve, and the AUC. The thesis thereby concludes that there is no method which is best in all data sets when it comes to the accuracy and the distribution of type 1 and type 2 errors. On the other hand, random forest is the best model in the ROC curve and AUC for all four data sets which means this model is the most preferred one if the goal is to separate correctly between the two classes. Overall the conclusion is that random forest, in general, is the most appropriate method when it comes to the empirical results on the data sets used in this thesis. It is also found that some methods are more affected when the market variables are excluded. This relates especially to logistic regression, neural network, and linear SVM. Random forest, on the other hand, does not lose much accuracy when the market variables are excluded. This indicates that this method might be better at predicting default for non-listed companies. The thesis also discussed the different measures to evaluate model performing, namely, accuracy and AUC. It is found that there is no clear answer on which measure to use since the two measures contain different information and complement one another. However, the accuracy is probably the measure which has the most problem standing alone, especially if the error cost of type 1 and type 2 errors are not equal.

It is also found that the variables in the model can be separated into five different categories, and some of the methods can measure the importance of the specific variable in the model. Especially in logistic regression, the model clearly explains how the individual variable contributes to the model and shows how some variables are consistently important among all four data sets. Furthermore, when market variables are excluded, the models seek to compensate for the missing variables by adding more accounting variables. Random forest shows the importance of the variables in the model. Though, this importance can be discussed, since it is found that X.NI in random forest has the least importance in three out of four models, but in logistic regression, it is included in three out of four models. Furthermore, it was found that the performance of all the models decreases when market variables are excluded, which gives an indication of the relatively high predictive power of market variables.

In the last part of the thesis, the focus shifted towards the Danish market for credit lending with the focus on non-listed firms. Of this reason, it is argued how the result from the models with data set excluding market variables can be transferred to the Danish market. It is found that due to the implementation of Basel II credit institutions prefer using the IRB approach over the standardized approach for calculating credit risk on its corporate portfolio. Both approaches are used to calculate the capital requirements for the credit lender, which is a function of the risk-weighted asset. The IRB approach uses its own credit risk model to calculate values for PD, LGD and EAD, while the standardised approach uses predetermined weights assigned to the outstanding of the loans. To do so, the credit lenders can reduce the total risk-weighted asset, which implies a lower capital requirement for the credit institution. However, it is shown that the implementation of the IRB approach has stringent regulatory demands. The section also discussed how the credit lender can benefit from a more precise credit risk model in other areas for instances when calculating the provision or evaluate a potential customer. A more precise credit risk model will imply that the calculation of provision would be more accurate, and the evaluation of a potential customer would help the credit lender to decide whether the loan application should be accepted and what the interest rate should be. Random forest had substantially lower error cost compared to logistic regression on the relevant data sets for the Danish market. This means that there might be great benefits to use this method to calculate its credit risk, and thereby determining the capital requirement, provision, and evaluation of the customer. Lastly is was discussed how the findings are put into perspective from the literature of default prediction. It is found that there are several examples in the literature of random forest being one of the best methods to predict default which also is the case in this thesis. This could support the statement as it is time to move away from logistic regression as the industry benchmark according to some papers in the literature. However, the regulation requires the input variables to form a reasonable and effective basis for the resulting predictions. This makes it difficult for credit institutions to use new state-of-the-art classifiers, like random forest, for predicting default, since it is not clear how the single input variables affect the result of the model.

The thesis has engaged in the field on how models can predict default for firms. As the results show, there is not one of the tested models which manage to classify all firms correct. An explanation of this could be that the accounting and market variables obtained for this thesis cannot predict all defaults. Sometimes there are market manipulation, accounting fraud, or external chocks, which make it very difficult to predict the future. During the first half of 2020, the COVID-19 crises have overtaken most parts of the world. This is a health crisis which is starting to be an economic crisis due to the lockdown of the world, resulting in an enormous chock to the global economy. The default rate has already increased, and it is expected to continuously grow (Fitch Ratings, 2020). Some would argue that the credit risk models have no possibilities to predict these firms going default under the COVID-19 crises

since the explanation of the default is primarily an external chock. Furthermore, there is no doubt that it would be harder to use the same historical models under the crises. However, the result of this thesis also accounts for a large crisis in the test result. The financial crisis in 2007-2009, which is considered as the most serious crisis since the great depression, is part of the testing sample in the thesis. This means that the result of the thesis is already affected by the most serious crisis since the great depression. Whether the present crisis is going to exceed the financial crisis is difficult to tell. However, it shows that the credit risk model should be able to predict well regardless of being in a boom or depression.

8. Bibliography

Altman, E. I. (1968, September). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance, 23*(4), pp. 589-609.

Amini, A. (2020, January 27). *Introduction to Deep Learning.* Retrieved Marts 2020, from Introduction to Deep Learning: http://introtodeeplearning.com/slides/6S191_MIT_DeepLearning_L1.pdf

Analyst, S. (2020, April 21). Interview with an Employee from a Danish Credit Institution. (A. Kjøller-Hanse, & S. S. Jensen, Interviewers)

Baesens, B. (2014). *Analytics in a Big Data World: The Essential Guide to Data Science and Its Applications* (Vol. 1). New Jersey: John Wiley & Sons, Incorporated.

Baesens, B., Gestel, T. V., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003, June). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society, 54*(6), pp. 627-635.

Bank for International Settlements. (2000, September). *Principles for the Management of Credit Risk.* Retrieved Marts 2020, from Bank for International Settlements: https://www.bis.org/publ/bcbs75.pdf

Bank for International Settlements. (2019a, December 15). *CRE20 - Standard approach: individual exposures.* Retrieved April 2020, from Bank for International Settlements: https://www.bis.org/basel_framework/chapter/CRE/20.htm?inforce=20191215

Bank for International Settlements. (2019b, December 15). *CRE31 - IRB approach: risk weight functions.* Retrieved April 2020, from Bank for International Settlements: https://www.bis.org/basel_framework/chapter/CRE/31.htm?inforce=20191215

Bank for International Settlements. (2019c, December 15). *RBC20: Calculation of minimum risk-based capital requirements.* Retrieved April 2020, from Bank for International Settlements: https://www.bis.org/basel_framework/chapter/RBC/20.htm?fbclid=IwAR0z-m7QmWnoft2WitJO-rmxY7zGkJe7utqqhOs99ROC6HwrZGGaFN_vsnQ

Bank for International Settlements. (n.d.). *Basel Committee on Banking Supervision reforms – Basel III.* Retrieved Marts 2020, from Bank for International Settlements: https://www.bis.org/bcbs/basel3/b3_bank_sup_reforms.pdf

Barboza, F., Kimura, H., & Altman, E. (2017, April 10). Machine learning models and bankruptcy prediction. *Expert Systems With Applications, 83*, pp. 405–417.

Beaver, W. H. (1966). Financial Ratios As Predictors of Failure. *Journal of Accounting Research, 4*, pp. 71-111.

Bell, J. (2014). *Machine Learning: Hands-On for Developers and Technical Professionals* (Vol. 1). Indiana: John Wiley & Sons, Incorporated.

Breiman, L. (2001). Random Forest. *Machine Learning, 45*, pp. 5-32.

Brownlee, J. (2018, May 23). *A Gentle Introduction to k-fold Cross-Validation*. Retrieved April 2020, from Machine Learning Mastery: Making Developers Awesome at Machine Learning: https://machinelearningmastery.com/k-fold-cross-validation/

Campbell, J. Y., Hilscher, J., & Szilagyi, J. (2006, July). In Search of Distress Risk. *National Bureau of Economic Research*.

Carmines, E. G., & Zeller, R. A. (1979/2011). Introduction. *Reliability and Validity Assessment*, 9-16.

Chava, S., & Jarrow, R. A. (2004). Bankruptcy Prediction with Industry Effects. *Review of Finance, 8*, pp. 537–569.

Compustat. (n.d.). Retrieved January 2020, from Warthon Research Data Services: https://wrds-web.wharton.upenn.edu/wrds/query_forms/navigation.cfm?navId=60&fbclid=IwAR2DZZ-gaMDdfB4tRX-zADhJxkux7kP_ZoV4MrZMyiwIEl7r1PFPKaC9TbQ

Danske Bank Group. (2018). *Internal Capital Adequacy Assessment 2018.* Retrieved April 2020, from Danske Bank: https://danskebank.com/-/media/danske-bank-com/file-cloud/2019/2/internal-capital-adequacy-assessment-2018.pdf

Erhvervs-, Vækst- og Eksportudvalget. (2018, Februar). *Effekter af Baselkomitéens Anbefalinger om Kapitalkrav til Kreditinstituttioner.* Retrieved April 2020, from Folketinget: https://www.ft.dk/samling/20171/almdel/ERU/bilag/108/1853762.pdf

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns. *Journal of Financial Economics*.

Fitch Ratings. (2020, Marts 30). *Europe's High-Yield Default Rates Rise as Credit Cycle Turns*. Retrieved Maj 2020, from Fitch Ratings: https://www.fitchratings.com/research/corporate-finance/europe-high-yield-default-rates-rise-as-credit-cycle-turns-30-03-2020

Freudenrich, C., & Robynne, B. (n.d.). *How Your Brain Works*. Retrieved Marts 2020, from HowStuffWorks: https://science.howstuffworks.com/life/inside-the-mind/human-brain/brain1.htm

Gissel, J. L., Giacomino, D., & Akers, M. D. (2007). A Review of Bankruptcy Prediction Studies: 1930-Present. *Journal of Financial Education, 33*, pp. 1-42.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Vol. II). Springer Series in Statistics.

Holm, A. B. (2016). *Videnskab i Virkeligheden - En Grundbog i Videnskabsteori* (Vol. 1). Samfunds litteratur.

Lammers, B. (2020, Marts 14). *Package 'ANN2'.* Retrieved Marts 2020, from The Comprehensive R Archive Network - CRAN: https://cran.r-project.org/web/packages/ANN2/ANN2.pdf?fbclid=IwAR3sBD9MABJ7GNBzh3QFwc_jPsHjOPn80G2h97gwFnWnoLCmhB8qYE98O-s

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015, May 14). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, pp. 124–136.

NAICS Association. (n.d.). *Search SIC Codes by Industry*. Retrieved Marts 2020, from NAICS Association: https://www.naics.com/sic-codes-industry-drilldown/

Narkhede, S. (2018, June 26). *Understanding AUC - ROC Curve*. Retrieved April 2020, from Towards Data Science: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Nehrebecka, N. (n.d.). *Probability-of-default curve calibration and the validation of internal rating systems.* Retrieved April 2020, from Bank for International Settlements: https://www.bis.org/ifc/events/ifc_8thconf/ifc_8thconf_4c4pap.pdf

Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research, 18*(1).

Olsen, P. B. (2003). Kapitel 10: Kvalitetsbeskrivelse. In P. B. Olsen, & K. Pedersen, *Problemorienteret projektarbejde* (pp. 189-205). Samfundslitteratur.

Petersen, C. V., & Plenborg, T. (2012). *Financial Statement Analysis.* Essex, England: Pearson Education Limited.

Prakash, O. (2018, July 5). *What are the key trade-offs between overfitting and underfitting?* Retrieved Marts 2020, from Quora: https://www.quora.com/What-are-the-key-trade-offs-between-overfitting-and-underfitting

Raschka, S., & Mirjalili, V. (2017). *Python Machine Learning.* Birmingham: Packt Publishing Ltd.

Sabato, G. (2008). Chapter 15: Managing Credit Risk for Retail Low-Default Portfolios. In N. Wagner, *Credit Risk : Models, Derivatives, and Management.* United States of America: CRC Press LLC.

Sakamoto, Y., Ishiguro , M., & Kitagawa, G. (1986). Reweived work: Akaike Information Criterion Statistics. *Journal of the Royal Statistical Society. Series D (The Statistician), 37*(4/5), pp. 477-478.

Salmon, F. (2011, August 9). *The difference between S&P and Moody's*. Retrieved Marts 2020, from Reuters: http://blogs.reuters.com/felix-salmon/2011/08/09/the-difference-between-sp-and-moodys/

Sharma, S. (2017, September 23). *Epoch vs Batch Size vs Iterations*. Retrieved Marts 2020, from Towards Data Science: https://towardsdatascience.com/epoch-vs-iterations-vs-batch-size-4dfb9c7ce9c9

Shumway, T. (2001, January). Forecasting Bankruptcy More Accurately: A Simple Hazard Model. *The Journal of Business, 74*(1), pp. 101-124.

Standard & Poor's. (2019, Sep). *S&P Global Ratings Definitions*. Retrieved Marts 2020, from S&P Global Ratings: https://www.standardandpoors.com/en_US/web/guest/article/-/view/sourceId/504352

Stein, R. M. (2007). Benchmarking default prediction models: pitfalls and remedies in model validation. *Journal of Risk Model Validation, 1*(1), pp. 77–113.

Swaminathan , S. (2018, Marts 15). *Logistic Regression — Detailed Overview*. Retrieved Marts 2020, from Towards Data Science: https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

Sørensen, R. L. (2013). *Større danske kreditinstitutters overgang til IRB-modeller: Effekt på solvens og risikovægtede aktiver.* KRAKAfinans - Finanskrisekommissionens sekretariat Teknisk arbejdspapir.

The Center for Research in Security Prices (CRSP). (n.d.). Retrieved January 2020, from Wharton Research Data Service: https://wrds-web.wharton.upenn.edu/wrds/query_forms/navigation.cfm?navId=118&_ga=2.139674867.1787231311.1589269159-47547912.1578994404&fbclid=IwAR0Hhcpif2hRTsXlsqWTTwg6jtndND0QLcYffFpx7F3NhvfCDJAFgchrI84

The European Parliament and the Council of the European Union. (2013, June 26). *Regulations.* Retrieved April 2020, from EUR-Lex: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32013R0575&from=DA

Wilson, R. L., & Sharda, R. (1994). Bankruptcy prediction using neural networks. *Decision Support Systems, 11*, pp. 545-557.

Yiu, T. (2019, June 2). *Understanding Neural Networks*. Retrieved Marts 2020, from Towards Data Science: https://towardsdatascience.com/understanding-neural-networks-19020b758230

Zhou, V. (2019, April 10). *Random Forests for Complete Beginners*. Retrieved 2020, from Victor Zhou blog: https://victorzhou.com/blog/intro-to-random-forests/

Zisserman, A. (2015). *Lecture 3: SVM dual, kernels and regression.* Retrieved Marts 2020, from C19 Machine Learning lectures Hilary 2015: http://www.robots.ox.ac.uk/~az/lectures/ml/lect3.pdf

9. Appendix

Survey of appendices

Appendix A: Additional figures in the thesis 1

Appendix B: Additional tables in the thesis 15

Appendix C: R output 22

Appendix D: R code screenshot of results 23