Norwegian School of Economics
Bergen, Fall 2019

Predicting Financial Distress in Norway

Using Logistic Regression and Random Forest Models

Guang Na Zhang and Fan Ye

Supervisor: Håkon Otneim

Master thesis, MSc in Economics and Business Administration

Major: Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

Abstract

Financial distress can be a highly costly and disruptive event, both on the level of the firm as well as for the society. Models to predict financial distress for this reason have been beneficial. In this thesis, we aim to develop a similar model which is applicable to Norwegian companies. Rather than solely focusing on bankruptcy predictions as previous research has done, we use financial ratios and other related company information, to predict whether firms are likely to enter financial distress within the next two years. Furthermore, we seek to identify early warning signs of financial distress in order for the management to start financial reconstruction in time.

A traditional and a more recent algorithm – logistic regression and random forest – were utilized in our analysis for their complementary properties. The models were created based on data provided by the Norwegian School of Economics where we selected a sample of 30 000 companies in the period from 2013 - 2016 after thorough cleaning of data.

We find very similar performance for both models where random forest shows slight superiority to logistic regression. Both models yield an AUC of ~ 0.65, and from the results obtained, it indicates that they are able to correctly predict ~ 60% of both healthy and financially distressed companies ahead of time.

Moreover, the results indicate that our models assign high importance to some commonly used ratios in the past, such as Size (Log of total assets), ROA, Retained earnings/Total assets, Total debt/Total assets and Debt/Equity. We also find Cash ratio and Net profit margin as important variables, which have been neglected previously. All these variables may contribute as warnings signs of financial distress when making predictions.

Acknowledgments

This thesis was written during the fall of 2019 at the Norwegian School of Economics (NHH), as a part of our MSc in Economics and Business Administration, with specialization in Business Analytics.

Although this work has been presented with its challenges, we are thankful to have had the opportunity to pursue this project, which has been a rewarding one. To our knowledge, previous studies addressing financial distress predictions in the Norwegian market have not been conducted. Thus, we hope that with our findings, improved models for the purpose of predicting financial distress in Norway can be developed.

We would like to thank the Centre for Applied Research at NHH (SNF) for providing data used in this research. We would also like to thank Dun & Bradstreet for additional data on bankruptcy and company credit rating.

Most of all, we would like to express our greatest gratitude to our supervisor, Assistant Professor Håkon Otneim, for his valuable guidance and support during this project. His advice has led to great improvement in the quality of our research.

Finally, we want to give our special thanks to our family and friends for their help, support and encouragement along the way.

Contents

List of Figures

List of Tables

1 Introduction

Recent bankruptcy statistics reveals that 6 311 Norwegian companies filed bankruptcy in 2019 – only 138 companies less than the number of bankruptcy filings during the financial crisis in 2008-2009 (Nervik, 2019). This increasing number of bankruptcies is of great concern and is detrimental to the Norwegian market if this trend continues. It reflects the vast number of companies who have experienced financial distress in the past couple years.

Broadly speaking, financial distress is characterized as a condition where the company is unable to meet their current financial obligations (Wruck, 1990). The cost of financial distress is high, not only for the company itself, but also for external stakeholders. This includes shareholders, investors, creditors, managers and employees (Chen & Merville, 1999). It results in forced asset selling, lost opportunity costs of projects, losses on outstanding debt, less productive employees, increased unemployment, resulting in volatility of the market. Thus, bankruptcy costs may not only adversely affect the company but also evolve into a social problem. Therefore, it is imperative to establish prediction systems which can monitor company efficiency and lower system risks to prevent a firm from entering financial distress.

Research on building financial prediction models of companies have been apparent from the 1930s (Bellovary, Giacomino & Akers, 2007). From the Z-score and O-score model to numerous machine learning methods, multiple approaches have been employed to construct models to predict financial distress. Most of the research that have been previously conducted, focuses on bankruptcy as the sole factor in financial distress predictions, as they attempt to evaluate creditworthiness of businesses (FitzPatrick, 1932; Altman, 1968; Ohlson, 1980). This creates a limited evaluation as it favors the perspective of creditors. Limiting financial distress to merely the final stage restricts the scope of application of prediction models. This gives insufficient time for a company to identify potential financial problems and adjust financial decisions accordingly. An analysis of UK firms showed that companies take up to three years to enter a state of bankruptcy (Tinoco & Wilson, 2013). Similar observations are made in the US market (Theodossiou, 1993), thus may be applicable to the Norwegian market. As such, if prediction models are only based on bankruptcy, the prediction only provides late stage analysis. However, we believe that in order to prevent financial distress, predictions must take into account companies in other stages of financial distress and not only bankrupted companies. These more inclusive models can provide early enough signs for companies to start financial reconstruction in time. From the perspective of a manager, it is particularly relevant as they can take precautionary action to steer the company away from financial distress.

Research on financial distress can be found on the international level, especially in the US. As for Norway, there are several bankruptcy prediction studies which have been conducted based on the Norwegian market (Bernhardsen, 2001; Aae & Hansen, 2017; Meese & Viken, 2019). However, to our knowledge, a model which predicts financial distress of companies in the Norwegian market has not been conducted. Moreover, given the Norwegian financial situation with an increasing number of bankruptcies in the country, it may be relevant that such prediction models are established. To this end, we aim to build a model that can effectively predict financially distressed companies in Norway. As such, there are two research questions to be addressed. Firstly, does the model have good prediction power which can be applied to real world situations? Secondly, which factors play a significant role to be used as warning signs of financial distress?

Based on two approaches, we develop two prediction models using financial statements from Norwegian companies. A more conventional algorithm, logistic regression, and a more recent machine learning algorithm, random forest, were utilized. The results show that our models hold predictive power, which may indicate financial distress and prevent companies from becoming bankrupt. Furthermore, we identify two novel warning signs which should be included in evaluating companies in financial distress.

1.1 Overview of Sections

This paper consists of seven sections. The next section reviews the literature within financial distress predictions, including the definition, previous models and commonly used ratios in the past. Section three describes the algorithms used and the methods of evaluation and validation. In section four, we create an indicator for financial distress, select variables and present our data treatment and modelling processes. The results of the models are presented in section five. Section six is dedicated to a discussion of the findings, limitations as well as suggestions for further research. In the final section, we conclude on our findings of this study.

2 Literature Review

In this section of our literature review, we will first introduce the definition of financial distress. We will also present the standard models of financial distress, in addition to more recent advances within the field. Moreover, we will include a brief overview of popular financial ratios used in the past. The insights above will form a theoretical foundation for our following empirical work.

2.1 Definition of Financial Distress

Financial distress is a broad term that can be interpreted in different ways. Generally speaking, financial distress refers to a financially dangerous stage which has the possibility to cease the activity of a business. Many researchers have done research on corporate financial distress, yet there is no unified definition or standard of it. Due to the vagueness of the term, financial distress has commonly been used interchangeably with *bankruptcy* in the literature (Dichev, 1998). They indicate both occurrence of defaulting events. However, there is a substantial difference between the two situations in terms of severity as well as in the sequence of events. Wruck (1990) suggested that financial distress refers to a situation where the firm has insufficient cash flows to cover the current debt obligations. Platt & Platt (2002) further emphasized that financial distress can be defined as a late stage of corporate decline, where bankruptcy usually is the last step. In the next paragraph we will see how bankruptcy and financial distress have been used differently in the past as criterion for financial failure.

FitzPatrick (1932) was one of the first to do research on financial status of companies. He considered bankruptcy as the criteria of a company's financial failure. Later, this criteria was used by many other researchers, such as Altman (1968) and Ohlson (1981). They used "file for bankruptcy" as the signal for failure. However, some researchers think it is too narrow to describe financial status of companies merely as bankrupt and non-bankrupt. Financial status is a continuous dynamic process, hence bankruptcy would not be an event that just shows up suddenly. Beaver (1966) was the first to bring financial distress to the stage. He broadened the concept of company failure from merely bankruptcy to "bankruptcy, bond default, an overdrawn bank account or non-payment of a preferred stock dividend". If any of the previous events occurred, the company has failed, or in other words, is in a financial distress stage.

Based on these studies, a company can be defined as financially distressed when it has serious complications in their operational processes, has insufficient cash flows to pay its debt on time and/or has declared bankruptcy.

2.2 Models

2.2.1 Early Studies

Financial statement analysis has been used to predict financial distress for a long time, and was primarily used by creditors to evaluate creditworthiness of its borrowers (Beaver, Correia & McNichols, 2011). Initial studies used financial ratios, which is the relative relationship between two values derived from financial statements of a company. These ratios were used as predictors due to their availability in the financial statements of the firms, which are commonly available to the public. The earliest evidence of using financial ratios in separating distressed and healthy firms is from the 1930s with the work of FitzPatrick (1932), where he compared 13 ratios of failed and successful firms.

The study of Beaver (1966) is considered the pioneering work on failure prediction models. He applied a univariate model with 30 financial ratios in which a classification model was carried out separately for each ratio, in order to find significant ratios to discriminate firms into failed and viable. The univariate analysis was conducted on a sample of 79 failed firms and 79 non-failing firms through a period of five years, where he investigated the predictive power of each ratio when applied 1-5 years prior to failure. Of the ratios examined, he found that Cash flow/Total debt and Net income/Total assets were good predictors for firm failure, with a respectively accuracy[1] of 78% and 72% for five years preceding failure. For the best predictor, Cash flow/Total debt, the probabilities of Type I and Type II errors in the first year prior to distress were equal to 5% and 22% respectively, while two years prior, these probabilities were 8% and 34% respectively. When predicting financial distress, Type I error refers to predicting a firm in distress while it is actually healthy, whereas Type II error refers to predicting a healthy firm, but it is actually in distress. Based on this model, Beaver further created four propositions that are still highly relevant today. He stated that the larger (1) the

[1] Calculated as overall accuracy: Correctly classified observations/All observations

reservoir and (2) the net liquid-asset flow from operations, the smaller the probability of failure. On the other hand, the larger (3) the amount of debt held and (4) the fund expenditures, the greater the possibility of failure.

Overall, Beaver's model seems to have a reasonable performance ability. Nevertheless, a univariate statistical analysis might suffer from a number of limitations (Cybinski 2003). To mention one, there is a high doubt associated to whether models with one variable are able to fully capture all relevant dimensions of a firm, as financial failures often are very complex. At the end of his paper, Beaver suggested that using a multivariate approach might provide a better model.

2.2.2 Altman Z-Scores

As an extension to Beaver's univariate model, Altman (1968) conducted the first Multivariate Discriminant Analysis (MDA) on prediction of financial distress. The model is called Z-score and is the most well-known model in the literature and is until now still widely used. In his research, Altman used 66 manufacturing firms in a period of 20 years (1946-1965) as samples. Based on the bankruptcy filings by the National Bankruptcy Act Chapter X, he divided the samples into two groups in which 33 firms were considered in a financial distress stage (bankrupt) and the other 33 were considered healthy (non-bankrupt). He further matched the samples of distressed and healthy firms, using the firm size and industry as criterion. Altman evaluated variables in a list consisting of 22 potential financial ratios, from which he ended up with five ratios that were the best predictors in terms of overall performance. The five ratios are presented together with the final model in Equation 2.1.

$$Z = 0.012X_1 + 0.014X_2 + 0.033X_3 + 0.006X_4 + 0.999X_5, \qquad (2.1)$$

where

X_1 = Working capital/Total assets

X_2 = Retained earnings/Total assets

X_3 = Earnings before interest and taxes/Total assets

X_4 = Market value of equity/Book value of total debt

X_5 = Sales/Total assets.

Altman's function gives a value of so-called Z-score where high values indicate healthiness of a firm and low values suggest a higher probability of financial distress. Firms with a Z-score above 2.99 would be deemed relatively safe, whereas firms with Z-score below 1.81 are considered to have a high possibility of failure. Scores between 1.81 and 2.99 are interpreted as the grey area, in which the model is not able to distinguish between healthy and bankrupt firms. Being in the grey area, there still exists a great possibility of company failure, hence one should exercise caution.

The results showed high predictive power one year before failure with an overall accuracy of 95%. The model's performance however dropped off considerably and down to 72%, 48%, 29% and 36% accuracy two, three, four and five years before failure, respectively. This indicates that when predicting more than two years prior to failure, guessing will yield better accuracy than Altman's model. Moreover, a criticism toward the accuracy of the model is that predictions are made in-sample, meaning that Altman predicted observations that was part of the data sample used to fit the model. Hence, high accuracy is expected and one could argue that 95% accuracy one year prior is not very impressive in this case. When the model was tested on a hold-out sample, the accuracy for t-1 was only 79%. Furthermore, the in-sample probabilities of Type I and Type II errors one year prior to distress were equal to 3% and 6% respectively, and 6% and 28% respectively for two years prior. Compared to Beaver's univariate model, the probabilities of the errors were significantly lower. Altman's Z-score model using MDA seems to be an improvement over the univariate model.

2.2.3 Ohlson O-Scores

Another famous financial distress model was created by Ohlson (1980) using logit analysis. In his paper, Ohlson highlighted some shortcomings regarding MDA. Among other things, he argued that the model's assumptions concerning distributional properties of the predictors in MDA may not be realistic as it requires normal distribution of all variables. Accordingly, the assumed equality of variance-covariance matrices between distressed and healthy firms becomes questionable. He further argued that matching distressed and healthy firms based on size and industry is somewhat arbitrary. Firm size should rather be used as a variable when predicting, he argued.

Ohlson (1980) stated that the use of logit analysis avoids the issues of restrictive assumptions. He collected data from the original bankruptcy filings and annual reports (10K) spanning from 1970 to 1976. The data included financial information of 105 bankrupt firms and 2 058 non-bankrupt firms. Just as prior studies, Ohlson utilized financial ratios for predicting financial distress. The final O-score model is as follows:

$$\begin{aligned} O - Score = {} & -1.32 - 0.407(SIZE) + 6.03(TLTA) - 1.43(WCTA) \\ & + 0.076(CLCA) - 1.72(OENEG) - 2.37(NITA) - 1.83(FUTL) \\ & + 0.285(INTWO) - 0.521(CHIN), \end{aligned} \tag{2.2}$$

where

SIZE = Log (Total assets/GNP price-level index)

TLTA = Total liabilities/Total assets

WCTA = Working capital/Total assets

CLCA = Current liabilities/Current assets

OENEG = 1 if total liabilities exceed total assets, 0 otherwise

NITA = Net income/Total assets

FUTL = Funds provided by operations/Total liabilities

INTWO = 1 if net income was negative for the last two years, 0 otherwise.

The interpretation of O-score is that the higher it is, the higher the probability of default. An O-score above 0.5 indicates potential failure within one year, while a O-score below 0.5 is considered as safe (Ohlson, 1980).

Ohlson computed three sets of estimates using his logit model. Model one predicted financial distress one year prior. Model two predicted two years prior, and model three three years prior. These models yielded an accuracy of 96.12%, 95.55% and 92.84% respectively. Same as for Altman, these accuracies were based on in-sample predictions. The overall accuracy of the model when tested on a hold-out sample was 85% one year prior.

Ohlson optimized the trade-off between type I and type II errors by assessing different cut-off points. He showed that for model one, 0.038 is the cut-off point that minimizes the sum of Type I errors and Type II errors. The model misclassified 12.4% of the healthy firms (Type I error) and 17.4% of the distressed firms (Type II error). For model two, the optimal cut-off point was 0.8, yielding an error rate of 8.6% for type I and 20.2% for type II. Comparing predictions one year prior, the error rates were considerably higher than Altman's (1968) study. One of the explanations Ohlson offered is the differences in sample sizes which makes it difficult to compare the two studies. Moreover, he argued that the "lead times" in reporting financial results affect the classifications and that errors would be reduced once subsequent annual reports are used to compute the relevant ratios. For two years prior, Ohlson's type II error was actually lower, while type I error was only slightly inferior compared to Altman's.

2.2.4 New Methods

Odom and Sharda (1990) were the first to utilize neural network technique for bankruptcy prediction. They built a neural network model and a MDA model as a standard for comparison based on Altman's study, using the same five financial ratios as Altman. The sample included 129 companies between 1975 and 1982, of which 65 went bankrupt and 64 were non-bankrupt, matched on industry and year. The result showed that, on the same hold-out sample, the accuracy of the neural network model was 81.81%, which was significantly higher than the result of Altman's MDA model of 74.28%.

Another popular new method used in financial prediction is the support vector machine (SVM). Min and Lee (2005) used SVM to predict company bankruptcy, where the optimal parameters were determined using grid search technique and 5-fold cross-validation. They found out that SVM model outperforms MDA, logic and back propagation neural network models.

2.3 Ratio Analysis

As we have seen from the studies by Beaver (1966), Altman (1968) and Ohlson (1980), financial ratios are commonly used as variables in prediction of financial distress. In 2007, Bellovary et al. conducted a thorough analysis of 165 financial distress studies published from 1965 to 2007, in which one of the analysis compared different variables used in past studies. From the 165 studies, they observed that a total of 752 different variables were used and 674 of the variables were utilized in only one or two studies, indicating a plentiful number of unique variables. The most common variable was found in 54 studies, and is the ratio of Net income/Total assets. The second most utilized variable is the ratio of Current assets/Current liabilities (Current Ratio), included in 51 studies. The ten most used ratios based on the analysis by Bellovary et al. (2007) are shown in Figure 2.1.

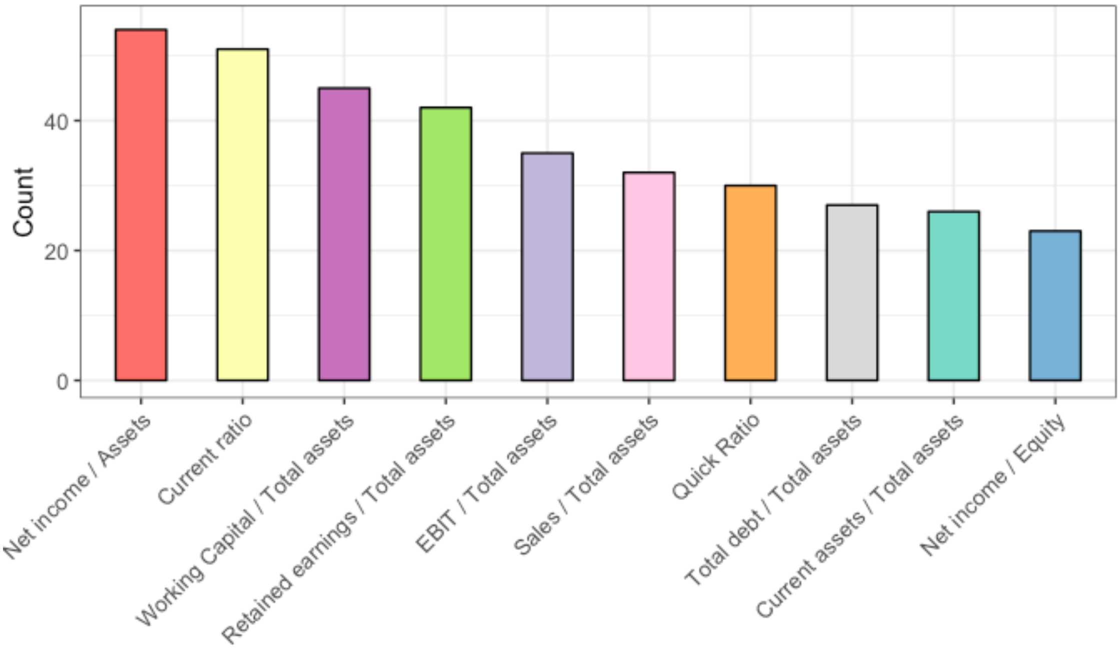


Figure 2.1: Distribution of the ten most used ratios previously

According to Petersen and Plenborg (2012), financial ratios can be categorized into four main categories: (1) Liquidity, (2) Profitability, (3) Leverage and (4) Activity/Efficiency. To explain each of them briefly, *liquidity* refers to the availability of liquid assets, i.e. cash, to pay its liabilities when they are due. *Profitability* is a company's ability to generate profits from its operations. *Leverage* (or gearing) is related to a company's capital structure and refers to the amount of debt in relation to equity. *Efficiency* is about how a company uses minimum of input to produce the highest amount of output. In general, we associate great profitability or efficiency with robustness in a company, while high gearing or poor liquidity often are connected with financial risk.

Looking at one ratio is insufficient when assessing a company's overall financial stability, hence multiple ratios should be considered (Petersen & Plenborg, 2012). One example is that liquidity ratios may be ideal, but if profitability or activity ratios are bad, then it creates a totally different situation. The ratios will be briefly explained and categorized in descending order, starting from the most popular ratio Net income/Total assets.

Net income/Total assets is also called Return on Assets (ROA). ROA measures the profitability of a company relative to its total assets. Companies with a high ROA are efficient in using assets to generate earnings, and are hence more likely to attract new investors, which in turn leads to growth and higher possibility of increased revenues.

$$ROA = \frac{Net\ income}{Total\ assets}$$

The second most common variable, Current ratio, is a liquidity ratio that measures a company's ability to cover its current liabilities (i.e. short-term obligations) with its current assets. Current ratio is also called the working capital ratio since the formula for working capital is current assets minus current liabilities. A ratio greater than 1 is usually a minimum because when above 1 it means that current assets exceed current liabilities. The ratio has a great importance to lenders.

$$Current\ ratio = \frac{Current\ assets}{Current\ liabilities}$$

Another liquidity ratio is Working capital/Total assets and it is the third most commonly used ratio in past studies. The ratio gives an idea of the amount of assets that is available for a company to run its day-to-day operations.

$$WC\ to\ Total\ assets = \frac{Working\ capital}{Total\ assets}$$

Next is Retained earnings/Total assets, which calculates the percentage of total assets funded by cumulative earnings. The ratio was first proposed by Altman in 1968 where he used it as a measure of leverage in his model. A high score indicates that assets are financed by retention of profits rather than from new capital or debt.

$$RE\ to\ Total\ assets = \frac{Retained\ earnings}{Total\ assets}$$

Earnings before interest and taxes (EBIT) to Total assets is another profitability ratio. It simply shows the profits that the company has generated from its total assets. The ratio is similar to ROA, but instead of using net income, EBIT is used.

$$EBIT\ to\ Total\ assets = \frac{EBIT}{Total\ assets}$$

The Sales/Total assets ratio is also known as the Asset Turnover ratio. It gives an indication of a company's ability to generate sales or revenue using its assets. A higher ratio is generally favored because it implies that the company is efficient in using its investments (i.e. assets).

$$Asset\ Turnover = \frac{Sales}{Total\ assets}$$

Next on the list is Quick ratio. It is very similar to the Current ratio which is described earlier in the sense that they both use current assets and current liabilities to measure liquidity. However, they differ because Quick ratio focuses on liquid assets, i.e. assets that can be quickly converted to cash, rather than all current assets. This is done by excluding inventory from current assets since inventory is generally more difficult to turn into cash.

$$Quick\ Ratio = \frac{Current\ assets - Inventory}{Current\ liabilities}$$

Total debt/Total assets is a leverage ratio that shows the proportion of total assets that is financed with debt. A high number of ratio implies high degree of leverage and increases the financial risk.

$$Total\ debt\ to\ Total\ assets = \frac{Total\ debt}{Total\ assets}$$

Current assets to Total assets defines the portion of total assets that is occupied by current assets. As current assets are essential for forming working capital, this is just yet another ratio for liquidity.

$$CA\ to\ Total\ assets = \frac{Current\ assets}{Total\ assets}$$

Net income/Equity is a profitability ratio for Return on Equity (ROE). This is an interesting ratio for shareholders as it reveals the profit a company generates with shareholders' invested money, thus it shows their percentage return accordingly.

$$ROE = \frac{Net\ income}{Shareholders\ equity}$$

Using ratios as variables has been a standard procedure for all financial distress prediction models to our knowledge as it has shown to yield good predictive results. Developing a prediction model, it was assumed that only testing the ten most popular ratios is somewhat inadequate because of the small number of variables this will represent. For this reason, we found it necessary to include more ratios. We selected six additional ratios that we suspected will contribute to enhancement of prediction performance once included. The selected ratios and their categorization are presented in Table 2.1. These ratios will not be further explained in detail, nevertheless, the categorization gives a good indication of which type of ratios they represent.

Liquidity	**Profitability**
Current ratio Working capital / Total assets Quick ratio Current assets / Total assets Cash / Current liabilities	Net income / Total assets (ROA) EBIT / Total assets Net income / Equity (ROE) Net income / Sales
Leverage	**Efficiency**
Retained earnings / Total assets Total debt / Total assets Debt / Equity Long-term debt / Total assets	Sales / Total assets Sales / Current assets Sales / Working capital

Table 2.1: Categorization of financial ratios

3 Methodology

The methodology section is divided into two main parts. In the first part, we present the theoretical framework for the algorithms that our models were built upon. The second part considers the methods of evaluation and validation. Here we also discuss the characteristics of our data and other factors that need to be taken into consideration.

3.1 Machine Learning Algorithms

In the following section, we explain the two classification algorithms – the logistic regression and random forest. Logistic regression is a traditional and simple model based on a generalized linear model. On the contrary, random forest is a complex non-linear model combined of two recent machine learning methods: the classification trees and the bagging algorithm. Their different strengths make them compelling to analyse in our research and to find out which performs better.

3.1.1 Logistic Regression

Our response variable *financial distress* is a qualitative variable that falls into one of the two categories, *distressed* or *healthy*. In logistic regression, these two categories are represented by 1 (distressed) and 0 (healthy).

For any X, logistic regression models the *probability* that Y belongs to a particular category, that is $p(X) = \Pr(Y = 1|X)$ written mathematically (James, Witten, Hastie & Tibshirani, 2017). The predictions of $p(X)$ must fall between 0 and 1 due to the nature of how probabilities work. It makes no sense if $p(X) > 1$ or $p(X) < 0$, which is why some methods such as linear regression are not appropriate for classification problems. For this reason, we seek to model $p(X)$ using a function that fulfils the criteria for probabilities, i.e. returns outputs between 0 and 1 for all values of X. James et at. (2017) mentions that this description is met by many functions, and one of them is the *logistic function*,

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}. \tag{3.1}$$

It is easy to see that no matter what values *X* or the $\beta_i's$ take, *p*(*X*) will return values between 0 and 1. Moreover, the logistic function will always produce an *S-shaped* curve within the range [0,1]. Figure 3.1 illustrates an example of a logistic curve.

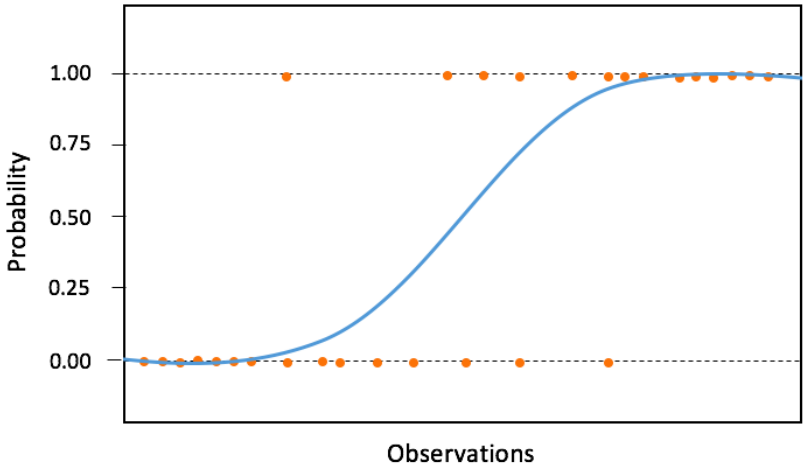


Figure 3.1: Illustration of a probability distribution using logistic regression

In our research, the assigned *p*(*X*) represents the probability that a company is financially distressed, which creates a useful indicator for whether to predict an individual as healthy or distressed. One might for example predict *distressed* for any company for whom *p*(*X*) > 0.5. Alternatively, if we want to be more sensitive in predicting companies who are at risk of default, we may lower the threshold, such as *p*(*X*) > 0.1. In other words, a company is predicted as distressed when probability for it is 10% or higher (James et al., 2017). After a bit of rearrangement of Equation 3.1, it becomes

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p. \tag{3.2}$$

The transformation is referred to as the logistic transformation of *p*(*X*), where the left-hand side is called the *logit* (James et al., 2017). We see that after performing the logit transformation, we obtain the standard linear regression model (the right-hand side). It shows that the underlying technique of logistic regression is quite similar to linear regression, hence the name *logistic regression*.

In the logistic function (Equation 3.1), the coefficients $\beta_i's$ are unknown. To fit the model and estimate the coefficients, we use the general method of *maximum likelihood*. We try to estimate $\beta_i's$ that when utilized in Equation 3.1, yield a predicted probability $\hat{p}(x_i)$ that corresponds as closely as possible to the actual response of the observation, that is close to 1 for all companies that are distressed, and close to 0 for all healthy companies. James et al. (2017) formalized this intuition using a mathematical equation called the *likelihood function*:

$$\ell(\beta_0, \beta_1) = \prod_{i:\mathcal{Y}_i=1} p(x_i) \prod_{i':\mathcal{Y}_{i'}=1} \left(1 - p(x_{i'})\right), \tag{3.3}$$

where the estimates of $\beta_0$and β_1 are chosen to *maximize* this function.

3.1.2 Random Forest

Decision trees, or more specifically *classification trees* (in the case of qualitative response), are the building blocks of the *random forest* model. Hence, we will go over classification trees before describing the method of random forest.

In general, decision trees go from observations about a predictor (branches) to conclusions about the target variable (leaves), where the *leaves* or *terminal nodes* are determined by predicting that "each observation belongs to the most commonly occurring class of training observations" (James et al., 2017). The task of growing a classification tree is done by a top-down approach known as *recursive binary splitting*. It begins at the top of the tree and then at each step of the tree-building process makes the best split at that particular step that maximizes class separation across the prediction space (James et al., 2017). Each split produces two new nodes further down on the tree.

Tree-based methods have the advantage of being simple and useful for interpretation because they mimic the human decision-making process (James et al., 2017). However, trees typically suffer from high variance, causing lower prediction accuracy. A common solution is to aggregate multiple trees which are then combined to yield a single averaged prediction (James et al., 2017). Some methods to grow multiples trees are bagging, boosting and random forest. We only consider random forest due to its technique of correlation treatment. An improvement with random forest over bagged trees is that only a random selection of m predictors is considered for every split in a tree. In this way, random forest decorrelates the trees, leading to a variance reduction that outperforms bagging models, which do not address the problem of predictor correlation (James et al., 2017). Typically, we select m equals to the square root of the total number of predictors (p).

There are several measures of separation ability in a node (node purity) that can be used to determine which variable and at what threshold to make the split, namely classification error rate, Gini index and cross-entropy (James et al., 2017). Throughout the paper we focus on the Gini index which is defined by

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}), \tag{3.4}$$

where $\hat{p}_{mk}$ is the proportion of observations in class k in node m. From the formula it is not hard to see that if all of the $\hat{p}_{mk}{}'s$ are close to zero or one, the Gini index goes towards zero. Small values of Gini indicate that observations in a node predominantly falls into a single category and we have a highly pure node.

To measure variable importance for estimating target variable, we use mean decrease in Gini index. It calculates a variable's total decrease in Gini when making splits, averaged over all trees. Since lower Gini index indicates higher node purity, higher mean decrease in Gini means higher variable importance.

3.2 Evaluation and Validation

In this section, we will first present the confusion matrix and then discuss the performance measures that can be derived from it that best suit our problem. Secondly, we consider our class imbalanced data and introduce a baseline for the classification threshold that we used under modelling. Lastly, we explain the method of cross-validation and how we performed final tests of our models.

3.2.1 Performance Measures

Confusion Matrix

A confusion matrix is often used to assess the performance of a classification model. The matrix in itself is not a performance measure as such, but almost all of the performance metrics can be calculated based on the numbers inside it. Table 3.1 shows an example of a confusion matrix.

		Actual	
		Distressed	Healthy
Predicted	Distressed	TP	FP
	Healthy	FN	TN

Table 3.1: Confusion matrix

For a two-class classification problem (e.g. distressed and healthy firms), the confusion matrix produces four different combinations of predicted and actual values, where two are referring to true predictions and two are false predictions. When working with financial distress, true predictions might be cases where distressed firms are predicted as distressed, known as *true positives* (TP), or cases where healthy firms are predicted as healthy, which is called *true negatives* (TN). When predictions are incorrect, we obtain *false positive* (FP) if a healthy firm is predicted as distressed, and *false negative* (FN) if a distressed firm is predicted as healthy. TP, TN, FP and FN that are just presented, are whole numbers, from which can be used to derive some useful rates.

True Rates

Some common rates to calculate when we have the true and false numbers inside the confusion matrix are the corresponding true and false rates. *True positive rate* (TPR) shows how often the model predicts financial distress out of all actual distressed observations, while *true negative rate* (TNR) is the number of predicted healthy firms in relation to all actual healthy firms. TPR and TNR are defined as follows:

$$TPR = \frac{TP}{TP + FN} \quad , \quad TNR = \frac{TN}{TN + FP} \tag{3.5}$$

The false rates consist of *false positive rate* (FPR) and *false negative rate* (FNR). FPR is the type I error that we briefly explained in section 2.2.1. It shows how often the model predicts distress out of all actual healthy observations. FNR on the other hand, is the type II error, and indicates the prediction of healthiness out of all companies that are actually distressed.

For any classifier, there is always a trade-off between TPR and TNR (James et al., 2017). If we for instance want more of the actual distressed firms to be classified as distressed (i.e. higher TPR), a natural consequence is that less of the actual healthy firms will be classified as healthy (i.e. lower TNR), since we have favored TPR and distressed firms to begin with. The trade-off is important in classification and needs to be considered carefully depending on the implications associated with the different error types. A consequence of type I error (FPR) is that the management spends too much time on avoiding financial distress in a healthy firm and thereby loses out on other relevant business activities, or a bank does not grant loan to a healthy firm and loses the opportunity of receiving potential interest. However, the cost related to type II error (FNR) is considered as much more severe. If a firm ignores the warning signs of financial distress, they might end up having huge financial problems which in turn destroys the trust of the company or even worse, results in an ending of the business. For a creditor, high FNR implies that they lose all or parts of the capital lent to the company. Due to the severity of the implications associated with FNR, it is important to lower the classification threshold in order to obtain a higher TPR, which consequently lowers the FNR.

ROC curve and AUC

Another method for evaluating performance is directly connected to the rates we just explained. By using the *ROC curve* (*Receiver Operating Characteristic curve*), we display the true positive rates versus the false positive rates for all possible thresholds (Figure 3.2). The associated AUC which is the *area under the (ROC) curve*, is an important metric for measuring the overall performance of a classifier, summarized over all possible thresholds (James et al., 2017). AUC represents the degree to which a model is capable of distinguishing between classes. The higher the AUC, the better the classifier is at predicting distressed firms as distressed and healthy firms at healthy.

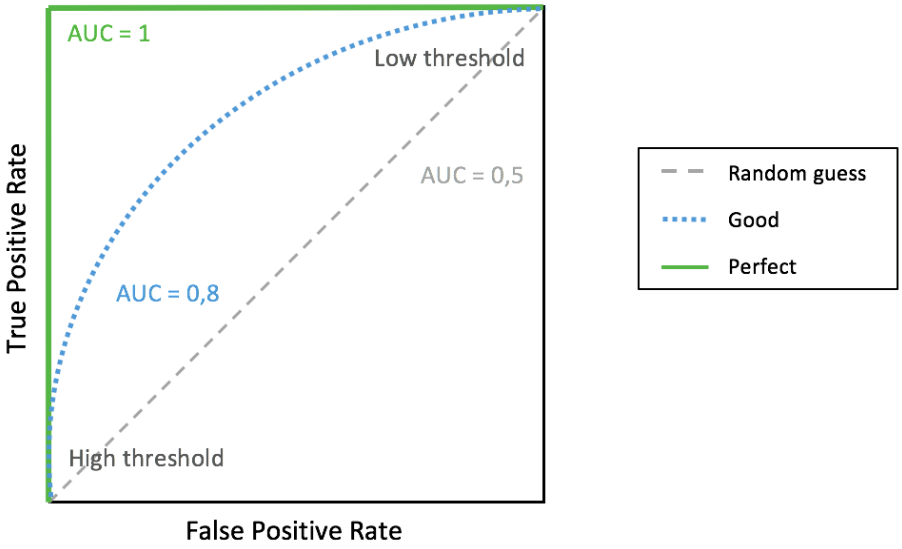


Figure 3.2: ROC curves

Figure 3.2 displays different ROC curves with their associated AUC values. As mentioned, the higher the area under the curve the better, an ideal ROC curve will therefore hug the top left corner. The green ROC curve follows closely to the corner, which means the model separates the classes perfectly and will yield an AUC of 1. In the optimal point to the top left corner, TPR is 100% and FPR is 0% (i.e. TNR = 100% because TNR = 100% - FPR). The blue dotted curve depicts a situation where the model has some ability to distinguish between classes, although not perfect. With an AUC of 0.5 as we see for the grey dashed line in the middle, the model has no discrimination capacity, i.e. it has same the probability as guessing randomly.

As ROC curves take into account all possible thresholds, it shows how varying the classification threshold impacts the TPR and FPR. At the bottom left of the curve, the threshold is high, indicating low TPR and low FPT (i.e. high TNR), whereas at top right, the threshold is low, indicating high TPR and high FPR (i.e. low TNR). This is one advantage of the ROC curve, because based on the purpose of the model, we can choose the TPR and TNR trade-off that we want (Moro, Cortez & Rita, 2014).

Overall Accuracy

Another popular performance measure which is used more frequently in the literature is the overall accuracy. It shows the proportion of correctly classified values out of all values, and is calculated by taking the sum of the true positives and true negatives divided by the total number of individuals tested, as shown in the formula below:

$$Overrall\ accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.6}$$

Overall accuracy is a good measure when the proportion of each class in the data set is somewhat balanced (Dong & Bailey, 2016). However, the measure becomes poor when the distribution of the response variables is skewed, i.e. when we have a majority of the response variable belonging to one class, because overall accuracy will not be able to reflect a model's ability to discriminate between classes, and will only be misleading. In the next section, we will shed light on the class imbalance problem and explain how we chose to deal with it.

3.2.2 Class Imbalance

A classifier with a threshold of 50% is known to yield the highest overall accuracy (James et al., 2017). This means that an observation is predicted as distressed if the probability for it is larger or equal to 50%, otherwise it is predicted as healthy. However, it requires the response variable to be somewhat balanced if we want the model to have the ability to discriminate between the two classes.

A problem that we faced was the fact that financial distress is a rare event, as it refers to the late stage of corporate decline (Platt & Platt, 2002). Consequently, the number of distressed firms is far less than healthy firms in our data set, causing a class distribution which is skewed by its nature. Problems arise when data are highly imbalanced. A traditional classification model of which the performance is determined by overall accuracy will tend to predict all firms as the majority class, so it makes less overall mistakes. If we assume that all new cases are assigned to the majority class, we will obtain a null error rate of 9.46% (actual distressed firms), and consequently an accuracy of 90.54%. Despite the high overall accuracy, the model is unreliable and useless if it is incapable of detecting the individuals in the minority class and discriminating between the two classes. As such, using a classification threshold of 50% may not be the best solution when data are imbalanced.

One approach to deal with the problem is to regulate the frequency of distressed and healthy firms, so that the data become balanced. This can be done by either undersampling the majority class or oversampling the minority class, which in practice requires that we remove an excessive amount of healthy firms or adding more copies of the distressed firms (He & Ma, 2013). However, Berg (2007) argued that the manipulated data are not representative of the real population and thus limits the accuracy and applicability in the real world. Instead, he changed the classifying threshold. Berg (2007) lowered the threshold to 10%, meaning that any firms with a higher or equal probability, would be classified as distressed. By lowering the threshold, it would produce a lower overall accuracy because more of the healthy firms would be predicted as distressed. At the same time, the model's ability to discriminate improves, which is far more important than the overall accuracy alone. Acknowledging this, we will not put too much weight on overall accuracy and primarily focus on AUC to evaluate the performance of our models. Moreover, we chose to do as Berg (2007) by lowering the threshold to 10% and using this as our baseline threshold.

3.2.3 Cross-Validation

It is important to validate the models that are created. We are interested in how well the fitted model works in predicting some previously unseen data (James et al., 2017). It should be stressed that training and validating a model on the same data is not optimal, as it often leads to overfitting. A consequence of overfitting is that the model performs poorly when predicting new data, even though the model has learned the data set very well. A good approach is therefore to hold out some data when training the model. For example, one can split the data set into a *training* set and a *validation* set, then fit a model using the training data and apply it on the validation data set to estimate how it performs when predicting the unseen data.

The method of using one training set and one validation set is called the *validation set approach*. An advantage of the method is its simplicity. Nonetheless, the drawback is that the validation error can be highly variable due to the randomness of the training/validation split. A method that deals with this problem is the *k-fold cross-validation* (k-fold CV) where validations are performed *k* times, each time using a different one of the *k* folds as the validation set. This results in *k* validation error estimates, which are then averaged to obtain the final validation error. Since the validation process is repeated several times, k-fold CV is more stable than the validation set approach method, which is why we used it to validate our models.

When choosing *k*, the bias-variance trade-off needs to be considered. Too much bias leads to underfitting, while too much variance leads to an overfitted model. In order to build a good model, finding a balance between bias and variance that minimizes the total error is therefore essential. With respect to k-fold CV, lower value of *k* results in lower variance, but higher bias, while higher *k* leads to lower bias, but higher variance. Given the consideration of the bias-variance trade-off, one should choose $k = 5$ or $k = 10$, as these have been shown empirically to result in validation error estimates that suffer from neither very high bias or variance (James et al., 2017). We chose to pick the lower *k* and thus performed 5-fold CV on our models, which is less computationally expensive, seeing that we were using a large data set.

The purpose of the validation phase under cross-validation is to select the best performing approach and estimate how well the model has been trained. The last step is the application phase where we apply our model to real-world data that were held out in the beginning and completely unseen under the whole training and cross-validation process. The idea is not to make any further changes or improvements to the model, but to see how it performs when applied to the real world.

4 Modelling

Before we could start building the models, we had to go through the processes of data treatment. In this section, we present our data set, explain how we selected variables and created an indicator for financial distress. Furthermore, we provide a walkthrough of the six steps of our data preprocessing.

4.1 Introduction to the Data

Our data was obtained from a database provided by the Centre for Applied Research (SNF) at the Norwegian School of Economics. The complete data set consists of accounting and company information for all Norwegian firms for the years 1992 to 2016, with some minor exceptions of companies that have been left out due to missing data. The database is a result of a collaboration between different organizations (Berner, Mjøs & Olving, 2016). While the majority of the accounting data was provided by Dun & Bradstreet (D&B), other company information was delivered by governmental institutions in Norway, such as Statistics Norway, Norges Bank and Brønnøysund Register Centre. For more details about the data, we refer to the working paper by Berner et al. (2016).

It should however be noted that we did not use the whole data set back from 1992, but instead limited the data from 2013 to 2016. Some of the reasons for not using data from all years are that new accounting standards have been implemented and there is a certain lack of observations during the initial years. More importantly, market conditions have changed. Consequently, we believe that using data from recent years will more effectively reflect the present economic environment for business, and hence should be used to build predictive models for financial distress. We have used companies in 2013 to train our models. While for the final testing, the trained models have been applied on companies in 2014.

4.1.1 Variable Selection

Previous studies found ratios relevant for predicting financial distress, thus we primarily used ratios as variables in our models. The ratios that we incorporated are discussed in section 2.3 about ratio analysis, and were calculated accordingly based on their formula and by employing the data from the income statement and the balance sheet.

Aside from the accounting ratios, we found some other company information interesting and that we wanted to test. We recall from the literature review part section 2.2.3 Ohlson's (1980) criticism on how size was used to match distressed and healthy firms, and thereby he concluded that size rather should be used as a variable. Another researcher, Shumway (2001), argued that market variables contain important information which is helpful for predicting financial distress. In his research, he found that the model consisting of both the financial ratios and market variables has improved performance compared to the model that consist of only ratios. Size is one of the market-driven variables he points out. Based on this, we chose to include firm size as a predictive variable in our models.

Firm size can be measured in several ways depending on how one defines it and considers as "big". We created two different variables as measures for firm size. The first one is log of total assets. Total assets are all the resources with economic value owned by a business. It is therefore natural to assume that total assets reflect company size, so the greater the total assets, the greater the size of the company. The second measure is related to number of employees. Companies with large number of employees are normally considered as big. However, it is not uncommon for big companies (with large total assets), such as big real estate companies to have a small number of employees.

Moreover, we included the variables of sector, number of shareholders, number of board members, and lastly, the number of female board members. By doing this, we wanted to test if there are any particular sectors that are more prone to financial failure, or whether a high number of shareholders or board members leads to diverse opinions within the company, which consequently impacts the corporate governance. It was also interesting to see whether having female board members has any impact on a company's performance.

In summary, this sums up to 22 independent variables. A complete list of the variables is presented in Table 4.1 below:

Variable name	Definition	Abbreviation	Category
x1	Net income / Total assets	ROA	Profitability
x2	Current assets / Current liabilities	Current ratio	Liquidity
x3	Working capital / Total assets		Liquidity
x4	Retained earnings / Total assets		Leverage
x5	Earnings before interest and taxes / Total assets		Profitability
x6	Sales / Total assets	Asset turnover	Efficiency
x7	Current assets - Inventory / Current liabilities	Quick ratio	Liquidity
x8	Total debt / Total assets		Leverage
x9	Current assets / Total assets		Liquidity
x10	Net income / Equity	ROE	Profitability
x11	Cash / Current liabilities	Cash ratio	Liquidity
x12	Net income / Sales	Net profit margin	Profitability
x13	Debt / Equity		Leverage
x14	Long-term debt / Total assets		Leverage
x15	Sales / Current assets		Efficiency
x16	Sales / Working capital		Efficiency
x17	Log of total assets		Other
x18	Number of employees		Other
x19	Number of shareholders		Other
x20	Number of board members		Other
x21	Number of female board members		Other
x22	Sector		Other

Table 4.1: Selected variables

4.1.2 Response Variable

Financial distress is not easily definable as we mentioned earlier in section 2.1, and it is a matter of definition when a firm is financially distressed. Based on previous research, we describe a firm as distressed if it has major struggles with operational activities, unable to pay its obligations when due and/or has filed for bankruptcy. It is an insolvent stage with rare chance of reversing the condition if no actions are taken.

However, unlike bankruptcy which the O-score and Z-score models are based on, having debt default or problems with daily activities is not so straightforward and easy to be observed. Insolvency does not necessarily mean that companies have no cash in their account. In fact, even bankrupted companies rarely have their cash balances literally falling to zero (Beaver et al., 2011). Furthermore, it is also inaccurate to focus on the proportion of liabilities to total assets. Liabilities is not in-and-of-itself a financial indicator of poor economic quality of a company. For some types of businesses, it is totally normal to have a large proportion of liabilities despite healthiness. Airlines are an example of such business. As one of the biggest airline companies in the world, United Airlines has a debt to asset ratio as high as 77.68% at the end of 2018 (United Airlines, Inc., 2018). However, great percentage of liabilities can be a sign of default risk, which is why liability occurs frequently in the financial ratios. Given that there is not anything specific to look for in accounting information that applies for all firms, it is problematic and risky to create our own measure for financial distress only based on information from financial statements. We instead utilized Dun & Bradstreet's credit ratings of companies, which are included in our data set.

D&B is a corporation that offers information on companies' creditworthiness. With more than 170 years' experience into what makes businesses fail, D&B has developed advanced scoring algorithms that combines a company's size and its balance sheet information to give an overall rating for companies around the globe (Dun & Bradstreet, 2013). Ratings are given based on a thorough estimation of default risk, and are hence practical as an indicator of financial distress that also has the meaning of defaulting events.

In the Nordic countries, D&B uses a Triple A rating system. An explanation of the rating codes is shown in Table 4.2. The estimated default rates were obtained from a study by Moody's (Hamilton, Ou, Kim & Cantor, 2007), as Dun & Bradstreet has not published the associated default rates for their rating system. Moody's uses a similar triple A system, thus we considered the default rates as comparable for the two credit rating agencies.

Triple A Rating	Description	App. Default rate
AAA	Strong creditworthiness	0.1 %
AA	Good creditworthiness	0.2 %
A	Creditworthy	0.25 %
B	Credit against security	7.5 %
C	Credit not advisable	30 %
Bankrupt/dissolved/liquidated		

Table 4.2: Rating system of Dun & Bradstreet. Approximate default rates obtained from Moody's.

With an estimated default rate of 30%, C-rated companies demonstrate very low credit quality and substantial risk of default. 30% default rate implies that the C-rated company has 30% of their debt being unpaid for an excessive amount of time. For this reason, we believe that having a C-rating is reasonably correspondent with the definition of a firm being financially distressed which is “unable to pay its debt when due”. However, recognizing that financial distress refers to a severe economic condition with rare chance of getting better, and that a C-rating could be a temporary one-year case, we only consider a C-rated company as financially distressed if it has been rated C for at least two consecutive years. Moreover, we identify bankrupt firms as financially distressed, as suggested by Beaver (1966).

To further justify our choice of using C-rating for two consecutive years as an indicator of financial distress, we analyzed the ratings of companies in the third year. Figure 4.1 displays companies' ratings in year 2015 after scoring C in both 2013 and 2014. We can immediately see the overwhelming results of NAs (missing values) and C, followed by some less frequent ratings, B and bankruptcy. If a company persists with a C-rating in the third year, it indicates that the financial situation did not improve and the company still has an estimated debt default of approximately 30%, which is then safe to infer that the company is financially distressed. A more interesting question is what the vast amount of NAs entails. After analyzing the data set more closely, we observed a pattern of companies with NA-rating being followed by NAs for consecutive years, majority of the cases, which implies that the company disappeared from the rating system and most likely went bankrupt. It is therefore reasonable to assume that the occurrence of NAs has close connection to firms filing for bankruptcy. Seeing that a large number of companies rated C for two years persists with C-score or go bankrupt (NA) and not turning better, it is fair to use this as an indicator of financial distress. Similar results were obtained when the same analysis was performed for other years. It is nevertheless worth noticing that not all companies with NA are going to be bankrupt, but is a pattern that we have observed, and we need to be careful before making any conclusions about NAs.

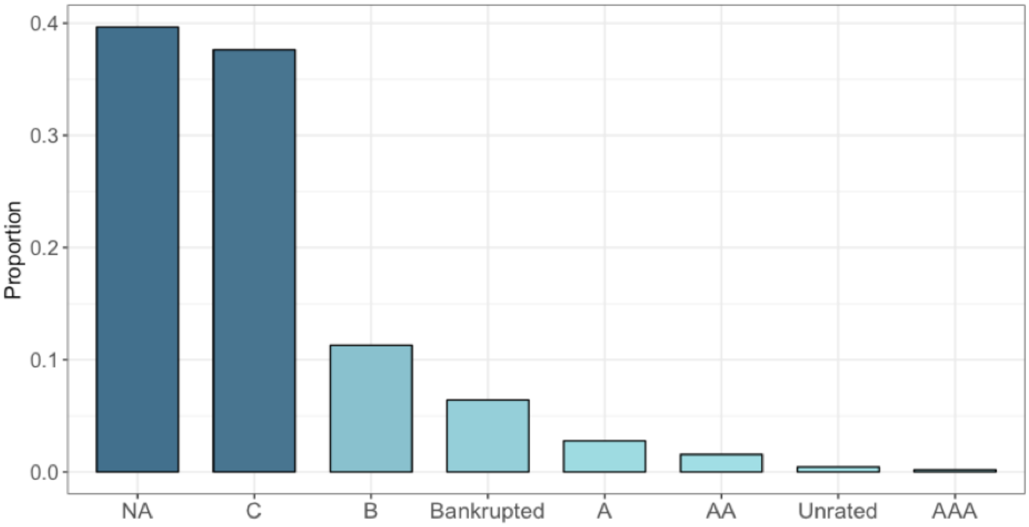


Figure 4.1: Distribution of credit ratings in year 2015 for companies that scored C in both 2013 and 2014

Predictions are made for whether or not a company will enter financial distress within the next two years. We believe this time frame is optimal because companies tend to stop submitting financial statements before they go into liquidation (Theodossiou, 1993), making it difficult to obtain enough data to predict right before distress. Another argument is the trade-off between early and late financial distress predictions, where predicting too much in prior weakens the reliability of the predictions, whereas predicting too late restrains stakeholders from acting on information on time. We consider that a two-year period gives business managers some time to react, and at the same time it is not too early that the model becomes unreliable.

Looking at the distribution of the response variable for companies in 2013, we observed that distressed companies constitute only a small fraction of all the observations. Distressed companies amount to approximately 10.33% of the data, thus the other 89.67% are all healthy companies. The skewed distribution between distressed and healthy firms is in line with the class imbalance problem we have described in section 3.2.2.

4.1.3 Data Preprocessing

Our data preprocessing consised of six steps: 1) data cleaning, 2) calculate variables, 3) data labelling, 4) selection of samples, 5) dealing with missing and extreme values and 6) removal of strongly correlated variables.

1) In the data cleaning part, companies that lack critical information for modelling or are outside the scope of the thesis were removed. For each year, there are two files for a company. One contains companies' accounting data, e.g. sales revenues and current assets. The file contains about 162 accounting variables. The other file includes company information, e.g. credit rating and number of board members. We only considered the firms in 2013, since they were used to build our models, and removed the firms that did not exist in both the accounting system and the company information system. Having information from both was essential for the modelling. The number of observations removed in this step was less than 0.1% of the total sample. Bankrupt companies in 2013 were also removed, which amounted to about 5% of the total sample. As they were already bankrupt, there was no point for further prediction.

The original data contain corporations with 42 different company forms including municipalities, church council and associations etc. These companies might be very different in terms of funding source, company structure or financial behavior from companies whose goal is to maximize profits. As such, we chose to limit our research object to companies that belong to one of the five most common company forms in Norway: "AS", "ASA", "ANS", "DA", "ENK"[2] (Altinn, 2019), which cover about 97% of the total data set.

[2] Description: limited share company, public limited company, unlimited company, shared liability and sole proprietorship, respectively.

2) The next step was to calculate the variables since ratios are not directly available from the accounting data. Companies in the accounting data set have vastly different sizes of asset, debt, income etc. Calculating ratios brings them on one scale and avoids the need for further normalization and standardization. Size of the companies was also included as a variable by utilizing companies' total assets. However, considering the large values of total assets compared to our ratios, we scaled them down using the logarithmic transformation (log10). Furthermore, we extracted some additional variables, for example, size of board, number of shareholders and sector category. We ensured that they were of the same scale as previous variables. The sector variable was further transformed into a categorical variable.

3) Creating labels for companies was the third step. As mentioned before, our objective is to predict financial distress within the next two years. Hence, the model uses data from t-2 (corresponds to year 2013 in our model building process) for the prediction and classifies companies that score C in both t-1 and t as distress, in addition to companies that become banktupt in either t-1 or t. The following table demonstrates the process of labelling (Table 4.3). After this step, 10.33% of the companies in 2013 were labelled as distressed, and 89.67% were labelled as healthy.

	t - 2	**t - 1**	**t**	**Response**
Company A	B	C	Bankrupt	Distressed
Company B	A	Bankrupt	NA	Distressed
Company C	B	C	C	Distressed
Company D	AA	A	B	Healthy

Table 4.3: Labelling of companies

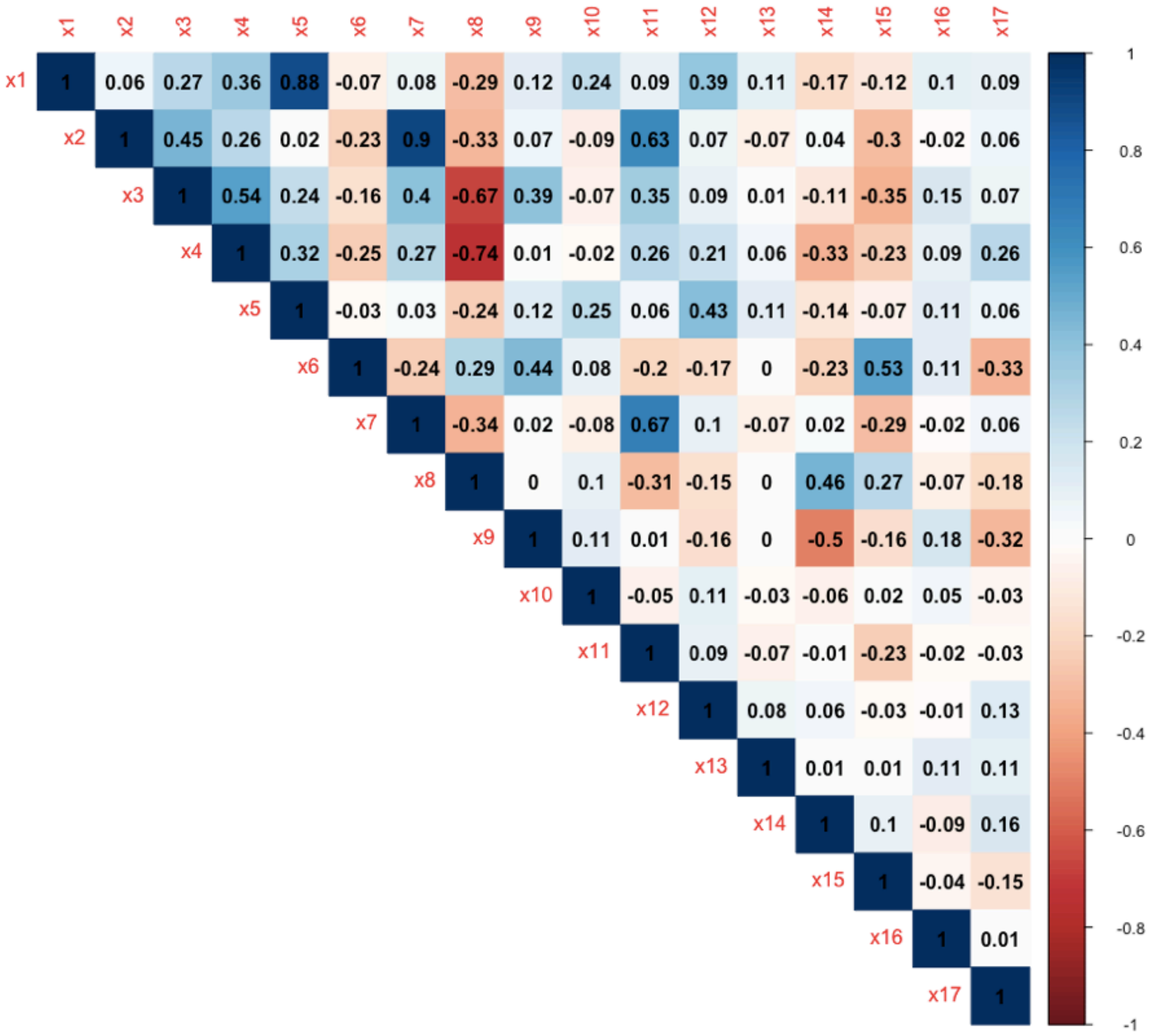
4) Taking into consideration the limited computing power of our computers, we decided to downsize our sample from more than 200 000 companies to 30 000 companies. We randomly selected 30 000 companies from our data set and did not change the distribution of healthy and distressed companies. Keeping the classes imbalanced ensures a representable sample of the population and that our model is applicable in the real world.

5) Despite that data from SNF have been quality checked, we noticed some variables with missing values. As a consequence, many companies have infinite or extreme values in their ratios, which can impact the following modelling process. One way of defining outliers is by using the interquartile range (IQR). Usually, the lower limit and upper limit of an non-outlier are 1.5 IQR from the first and third quartile (Deep, 2006). However, the influence of outliers is limited when we are working with a large data set. Moreover, recognizing that the real world is filled with outliers and that we seek to build models that are applicable to real populations, we wanted to ensure that we did not delete valuable observations. As such, we relaxed the standard to define a value as an outlier only when it is more than 10 IQR away from the first or third quartile, and we replaced the value with "NA".

Little and Rubin (2002) suggested that observations should not be deleted if a variable has more than 5% missing data. Five percent is a considerable amount, thus deleting this amount of observations omits the valuable information they might contain. Following the rule of Little and Robin, we checked all the variables. Even for the variable with most outliers and missing values, the number is less than 2% of the total observations. In applying this principle more carefully, we decided to only remove observations that had three or more missing values (NA). The rest of NAs were filled with the mean of the same variable. The total number of observations removed during this step was 4.3%. In this process, the proportion of distressed companies dropped from 10.33% to 9.46%. This shows that distressed companies have more missing values and outliers than healthy companies, which indicates that outliers contain some information. However, since the percentage change was small, it did not affect the distribution in any significant way.

6) Statistical inferences made about the data may not be reliable with the presence of high correlation between features. Hence, we identified the highly correlated variables in our data by visualizing a correlation matrix heat map (Figure 4.2). The darker the color, the stronger the correlation is. Dark blue and dark red represent strong positive and negative correlation respectively. Note that Figure 4.2: Heat map of the correlation matrix only includes the variables up to x17, because x18 - x21 have shown to yield small overall correlation and x22 (sector) is a categorical variable (see Figure A1.1 in appendix for the whole correlation matrix). We removed highly correlated variables by setting the threshold to ± 0.8, and we found a strong correlation between x1 and x5, and also between x2 and x7. For definition of the variables, we refer to Table 4.1.

The results are not so surprising. It is not hard to see that the correlated variables actually are very similar. The only difference between x1 (Net income/Total assets) and x5 (EBIT/Total assets) is the replacement of net income with EBIT, and for x2 (Current assets/Current liabilities) and x7 (Current assets - inventory /Current liabilities), the difference is that the latter excludes inventory from the current assets. Some companies simply do not have or have a small amount of inventory. For both of the cases where one variable is computed from the other variable, the extra variable does not convey extra information and rather represents a repetition and noise in the data. The solution was to remove one variable from each of the correlation-pairs, and we decided to remove the less popular ones, x5 and x7.



	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	x17
x1	1	0.06	0.27	0.36	0.88	-0.07	0.08	-0.29	0.12	0.24	0.09	0.39	0.11	-0.17	-0.12	0.1	0.09
x2		1	0.45	0.26	0.02	-0.23	0.9	-0.33	0.07	-0.09	0.63	0.07	-0.07	0.04	-0.3	-0.02	0.06
x3			1	0.54	0.24	-0.16	0.4	-0.67	0.39	-0.07	0.35	0.09	0.01	-0.11	-0.35	0.15	0.07
x4				1	0.32	-0.25	0.27	-0.74	0.01	-0.02	0.26	0.21	0.06	-0.33	-0.23	0.09	0.26
x5					1	-0.03	0.03	-0.24	0.12	0.25	0.06	0.43	0.11	-0.14	-0.07	0.11	0.06
x6						1	-0.24	0.29	0.44	0.08	-0.2	-0.17	0	-0.23	0.53	0.11	-0.33
x7							1	-0.34	0.02	-0.08	0.67	0.1	-0.07	0.02	-0.29	-0.02	0.06
x8								1	0	0.1	-0.31	-0.15	0	0.46	0.27	-0.07	-0.18
x9									1	0.11	0.01	-0.16	0	-0.5	-0.16	0.18	-0.32
x10										1	-0.05	0.11	-0.03	-0.06	0.02	0.05	-0.03
x11											1	0.09	-0.07	-0.01	-0.23	-0.02	-0.03
x12												1	0.08	0.06	-0.03	-0.01	0.13
x13													1	0.01	0.01	0.11	0.11
x14														1	0.1	-0.09	0.16
x15															1	-0.04	-0.15
x16																1	0.01
x17																	1

Figure 4.2: Heat map of the correlation matrix

4.2 Model Building

The following section describes how we selected the best performing approach of the models with respect to variables, hyperparameters and classification threshold by using cross-validation. The purpose of doing this is to use the best approach to refit the model on the whole training set in the end, resulting in the final model that will be used in our tests.

4.2.1 Choosing Variables

After fitting a logistic model with all the variables, we performed a chi-squared test on the model to assess whether the variables are able to explain the variation in the response variable. Chi-squared test considers each variable individually against the response variable assuming that the null hypothesis is true, i.e. there is no relationship between the response variable and the independent variable (Freitas & Freitas, 2013). The purpose is to find statistical relationships between two variables that are significant enough to reject the null hypothesis. We looked at the p-values from the results and noticed that the variables x10, x11, x12, x14, x20, x21 and x22 were of no significance (Table A2.1 in Appendix). It implies that there is not much value added when incorporating these variables to the model, hence we decided to remove them and only considered the remaining variables as candidates for inclusion in the model. Next, we looked at the variables' performance when seen in group (the multiple logistic regression) (Table A2.2 in Appendix) and fitted a new final model consisting of variables that show significance at the 0.1 level, i.e. variable x1, x4, x6, x8, x9, x13, x15, x16, x17 and x19. As a result, the model's mean AUC of the 5-fold cross-validation increased from 0.6593 (all variables) to 0.6613 (selected variables).

4.2.2 Hyperparameter Tuning

For the random forest model, we had to select the number of splitting variables *m* and the number of trees to grow in the "forest". James et al. (2017) mentioned that *m* typically is selected to be equal to the square root of *p* (total number of predictors). In our case, where we have 20 predictors, the square root equals to 4.47. We tested the model with *m* = 3, *m* = 4 and *m* = 5, in order to see if any performs better. Using cross-validation to test the different values of *m*, we decided to proceed with *m* = 4. Similarly, for the number of trees, we applied cross-validation to the random forest using different number of trees each time, and noticed no significant improvements of AUC after growing around 300 trees. We argue therefore that using 300 trees is sufficient to give good performance. Finally, we fitted the random forest model using all variables. Note that we have tested different combinations of variables to include in the model according to each variable's mean decrease in Gini (Figure 5.3 under Results), but noticed that using all variables was superior in terms of AUC.

4.2.3 Optimization of Threshold

Logistic Regression

We performed a 5-fold cross-validation on the logistic model first by using the baseline threshold of 10% and obtained an averaged (over the five folds) confusion matrix as shown in Table 4.4. The associated overall accuracy is 74.63%, which at first sounds not bad. The table shows that the model predicts healthy firms quite well with a TNR of 77.27%. For the distressed firms however, the model is only able to classify 48.38% correctly, implying that the model has little discriminatory power, as it tends to predict the over-represented category (healthiness). With a TPR of 48.38%, even guessing yields better results.

		Actual	
		Distressed	Healthy
Predicted	Distressed	254	1189
	Healthy	271	4042

Table 4.4: Confusion matrix with baseline threshold – Logistic regression

In order to say that the model has the ability to distinguish between the two classes, both TNR and TPR have to be higher than 50% as a minimum. As an attempt to improve the model, we tried to optimize the classification threshold while carefully consider the trade-off between TPR and TNR. We obviously want a higher TPR because, as outlined in section 3.2.1, it is costly to mistakenly predict an unhealthy firm as healthy (type II error). However, while making type II error is undesirable, it will be ignorant to force the TNR down to 50%, seeing that we have a large number of healthy firms (~90%) in our data set which as well reflects the real-world distribution. To put it another way, a small adjustment of TNR has large influences on financial distress predictions of the healthy firms. Taking into account that type II error is costly, but at the same time, the large number of healthy firms that can not be disregarded, we decided to optimize the threshold such that TPR = TNR. Figure 4.3 displays TPR and TNR over different thresholds. As the figure shows, around 8.6% is the optimal threshold where TPR and TNR are equal. Hence we changed the classification threshold to 8.6% and tested the model again.

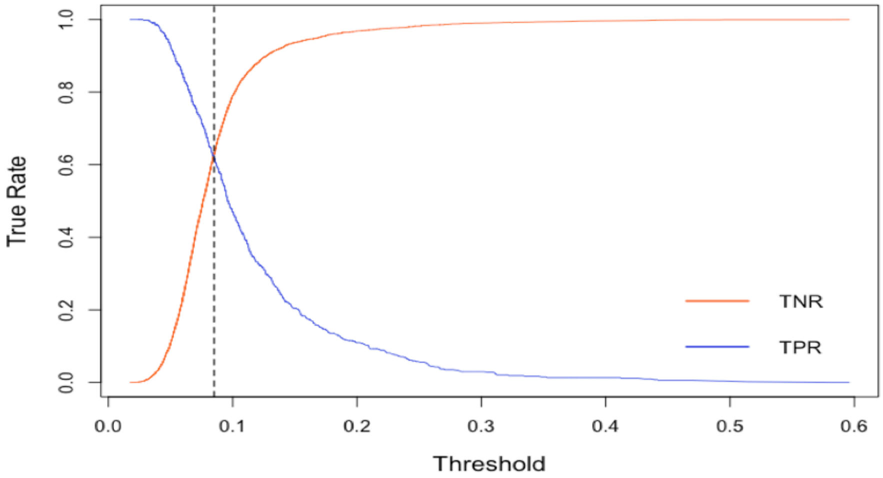


Figure 4.3: Optimization of threshold such that TPR equals TNR – Logistic regression

A new mean confusion matrix was obtained by optimizing the threshold (Table 4.5). We learn that TPR increases from 48.38% to 61.71%, while TNR decreases from 77.27% to 62.31%. This results in a drop of the overall accuracy that now is 62.25%, compared to the accuracy before that was 74.63%. It has nonetheless been argued before that overall accuracy is not a suitable indicator for our data, thus one should not put too much focus on it. What is more important is that both TPR and TNR are above 50%, indicating that the discriminatory power of the model has enhanced. Overall accuracy and AUC for all five folds are summarized in Table 4.6. The mean AUC is 0.6613 for the logistic model.

		Actual	
		Distressed	Healthy
Predicted	Distressed	324	1972
	Healthy	201	3260

Table 4.5: Confusion matrix with optimized threshold – Logistic regression

	Overrall accuracy		*AUC*
Fold	Basecase (10%)	Optimized (8.6%)	
Fold 1	0.7384	0.6163	0.6584
Fold 2	0.7488	0.6364	0.6509
Fold 3	0.7559	0.6196	0.6689
Fold 4	0.7437	0.6117	0.6681
Fold 5	0.7447	0.6285	0.6599
Average	**0.7463**	**0.6225**	**0.6613**

Table 4.6: Cross-validated results – Logistic regression

Random Forest

Similarly to the logistic model, the random forest model was cross-validated with five folds and by using the baseline threshold of 10%. This resulted in a confusion matrix as shown in Table 4.7. As we can see from the table, TPR is 60.08%, while TNR amounts to 65.31%. With both TPR and TNR being above 50%, this means that even with the baseline threshold, the model has some ability to discriminate between distressed and healthy companies. This result was further improved by optimizing the threshold, as we want a model that is relatively as likely to predict distressed firms correctly as to predict healthy firms correctly (TPR = TNR). Figure 4.4 illustrates the true rates and indicates that the optimal threshold lays around 9.5%.

		Actual	
		Distressed	Healthy
Predicted	Distressed	316	1815
	Healthy	210	3417

Table 4.7: Confusion matrix with baseline threshold – Random forest

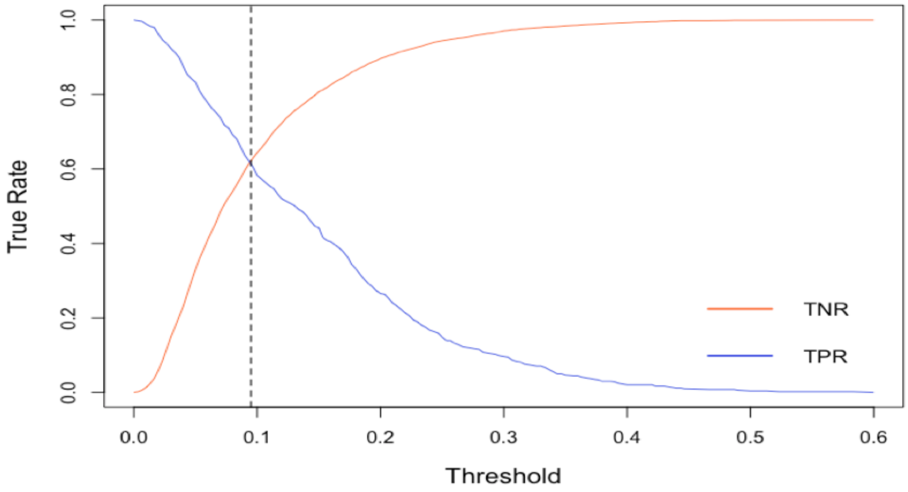


Figure 4.4: Optimization of threshold such that TPR equals TNR – Random forest

We fitted the model again using the optimized threshold and obtained a new confusion matrix as displayed in Table 4.8. The new model is able to predict 62.85% of the distressed firms correctly, while for the healthy firms 62.6% are predicted correctly. The results after optimization are clearly more balanced than with 10%-threshold that yielded a TPR and TNR of respectively 60.08% and 65.31%. Table 4.9 presents the results of the five folds. The mean AUC archived for the random forest is 0.6691.

		Actual	
		Distressed	Healthy
Predicted	Distressed	330	1962
	Healthy	195	3284

Table 4.8: Confusion matrix with optimized threshold – Random forest

	Overrall accuracy		*AUC*
Fold	Basecase (10%)	Optimized (9.5%)	
Fold 1	0.6503	0.6246	0.6634
Fold 2	0.6441	0.6224	0.6598
Fold 3	0.6620	0.6327	0.6982
Fold 4	0.6393	0.6170	0.6543
Fold 5	0.6460	0.6297	0.6696
Average	**0.6483**	**0.6253**	**0.6691**

Table 4.9: Cross-validated results – Random forest

5 Results

After selecting the best performing approach by choosing variables and parameters such as number of trees and classification threshold, we refitted both models on the whole training set and tested them on an unseen test set with 9 576 companies from 2014. We will first look at the test results compared to the cross-validation results, and then look into the models' ROC curves and important variables, before presenting a comparison of the models.

5.1 Logistic Regression

Logistic regression produces the confusion matrix shown in Table 5.1. We compare the results using the test set and cross-validation in Table 5.2. Not surprisingly, we observe that the performance of the model has dropped a little when using the test set. AUC has decreased from 0.6613 to 0.6407, whereas overall accuracy decreased from 62.25% to 59.83%. Consequently, both TPR and TNR dropped slightly. Other than that, the model's test results are quite comparable to the results from cross-validation.

		Actual	
		Distressed	Healthy
Predicted	Distressed	564	3479
	Healthy	368	5165

Table 5.1: Final confusion matrix – Logistic regression

	Accuracy	*TPR*	*TNR*	*AUC*
Cross-Validation	62.25%	61.71%	62.31%	0.6613
Test Set	59.83%	60.52%	59.75%	0.6407

Table 5.2: Comparison of CV results and test results – Logistic regression

Figure 5.1 illustrates the model's associated ROC curve. As the curve is on the left side of the diagonal line, it verifies that the logistic model has some ability to distinguish between distressed and healthy firms. The shape of the curve indicates that the model performs fairly well for the TNR (summarized over all thresholds). When TPR reaches 50%, TNR remains over 70%. However, as TPR rises (above 50%), larger cost is introduced for the TNR, because it is difficult to achieve high TPR without sacrificing large amount of TNR, which is true from what is characteristic of our data.

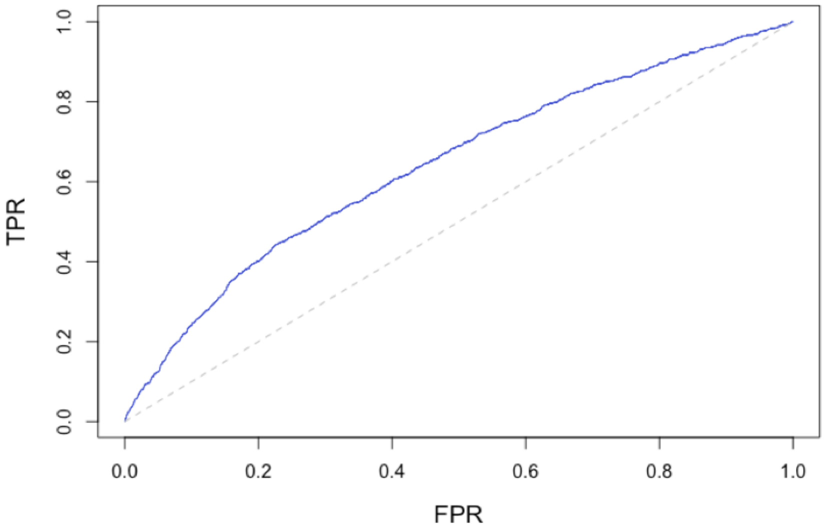


Figure 5.1: ROC curve – Logistic regression

Looking at the summary statistics of the final model (Table A2.3 in Appendix), logistic regression suggests that the variables Net income/Total assets, Retained earnings/Total assets, Total debt/Total assets, Current assets/Total assets, Debt/Equity, Log of total assets and Number of shareholders have the strongest significance, indicating that these variables might be important predictors for explaining financial distress.

Paying attention to the coefficients of the variables in the summary statistics, we observe that having higher Net income/Total assets, Retained earnings/Total assets, Log of total assets or Number of shareholders reduces the probability of financial distress. On the other hand, with increasing Total debt/Total assets, the probability of financial distress increases. The abovementioned connections are quite intuitive. However, we have also noticed two not so intuitive connections from the results: 1) Higher Current assets/total assets increases the probability of financial distress and 2) Higher Debt/Equity reduces the probability of financial distress.

The definition of current assets is cash and other assets that are easily converted to cash, such as accounts receivable and inventory. It is contradictory to conclude that a company with large amounts of cash is not able to pay back its debt. However, the observed relationship can be related to the cost of holding excessive cash (opportunity cost). Moreover, if a company has huge inventory over a long period of time, it can indicate that the company has problems with turning over the inventory and make sales, which might consequently force the company into financial failure.

When it comes to debt to equity, the result is conflicting from what we get from the other variable Total debt/Total assets which says that higher debt increases the chance of financial distress. It is generally agreed upon that companies with high debt are associated with high risk and are therefore more likely to be unhealthy. The coefficient for debt to equity is suggesting the opposite with its slightly negative (-0.016) coefficient. Our assumption to these contradictory results is that finding a balance between debt and equity is essential. Having a high D/E ratio could be dangerous from a creditor's perspective. For a company however, it is important to maintain a reasonable D/E ratio. Companies that are able to utilize external sources of finance are more likely to outgrow companies that only have access to their own financial sources. The optimal debt to equity ratio depends on the financial status of the company, industry as well as other factors. A ratio of 1 to 1.5 is generally considered as good (Mandan, 1978). From our data set, we observe that the median of debt to equity ratio for Norwegian firms is 1.48, which is a healthy number.

Note however that these are just potential explanations of the coefficients. Logistic regression is known to be an unstable model that produces varying coefficients each time. Hence the interpretation of the coefficients needs to be exercised with caution and it would be necessary to consider them together with the other mentioned variables indeed.

5.2 Random Forest

The confusion matrix that we obtained using random forest is displayed in Table 5.3. A comparison of the CV results and test results is presented in Table 5.4. Similarly to logistic regression, the performance dropped slightly when the model was applied to the test set. We observe that AUC has decreased from 0.6691 to 0.6558, while accuracy decreased from 62.53% to 60.05%. Moreover, it is noticeable that the true rates are not as balanced as the CV results, which can be argued is caused by a different distribution of the response variable in the test data. TPR is quite similar to before, TNR on the other hand has dropped with almost three percentage points.

		Actual	
		Distressed	Healthy
Predicted	Distressed	583	3477
	Healthy	349	5167

Table 5.3: Final confusion matrix – Random forest

	Accuracy	*TPR*	*TNR*	*AUC*
Cross-Validation	62.53%	62.85%	62.60%	0.6691
Test Set	60.05%	62.55%	59.77%	0.6558

Table 5.4: Comparison of CV results and test results – Random forest

The model's associated ROC curve is displayed in Figure 5.2. It looks very similar to the curve produced by logistic regression. Hence, the same logic applies for this ROC curve by random forest, that is the model performs well for TNR, but sacrifices larger amount of it when trying to achieve TPR higher than 50%.

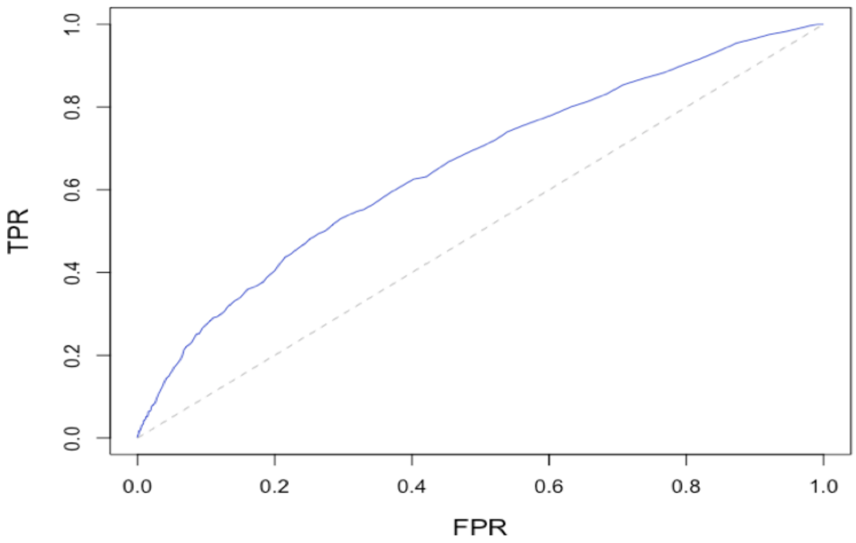


Figure 5.2: ROC curve – Random forest

Random forest returns the mean decrease in Gini according to each variable's importance in separating observations while making splits in the trees (Figure 5.3). The results are largely aligned with those we obtained using the linear regression. It suggests that Log of total assets, Retained earnings/Total assets, Net income/Total assets and Total debt/Total assets are important variables that contributes the most when reducing the Gini, i.e. same variables as logistic regression. However, in contrast to summary statistics produced by logistic regression, mean decrease in Gini does not indicate a relationship between the explanatory variable and the response variable as the model is non-linear. It only indicates the degree of importance. It is nonetheless reasonable to assume that higher Log of total assets, Retained earnings/Total assets and Net income/Total assets reduce the chance of financial distress when making splits in the trees. Total debt/Total assets on the other hand should be considered with care, as we already argued for that maintaining a reasonable debt ratio is positive.

The next important variables that follow are Cash/Current liabilities, Net income/Sales and Debt/Equity. It would be natural to assume that managing to have sufficient cash to cover its liabilities or high net income per kroner of sale is positive. Debt to equity is similarly as before considered as an important variable as well.

Current assets/Total assets and Number of shareholders that were significant in logistic regression are now ranked relatively low by random forest, which reduces their chance of being as important as before. As opposed to logistic regression, Number of shareholders is ranked as the second worst variable. The most likely reason why the number of shareholders was important in the previous model is the correlation of it with Log of total assets which is considered as highly important in both models. The correlation matrix shows that the correlation between these two variables is 0.4 (Figure A1.1 in Appendix).

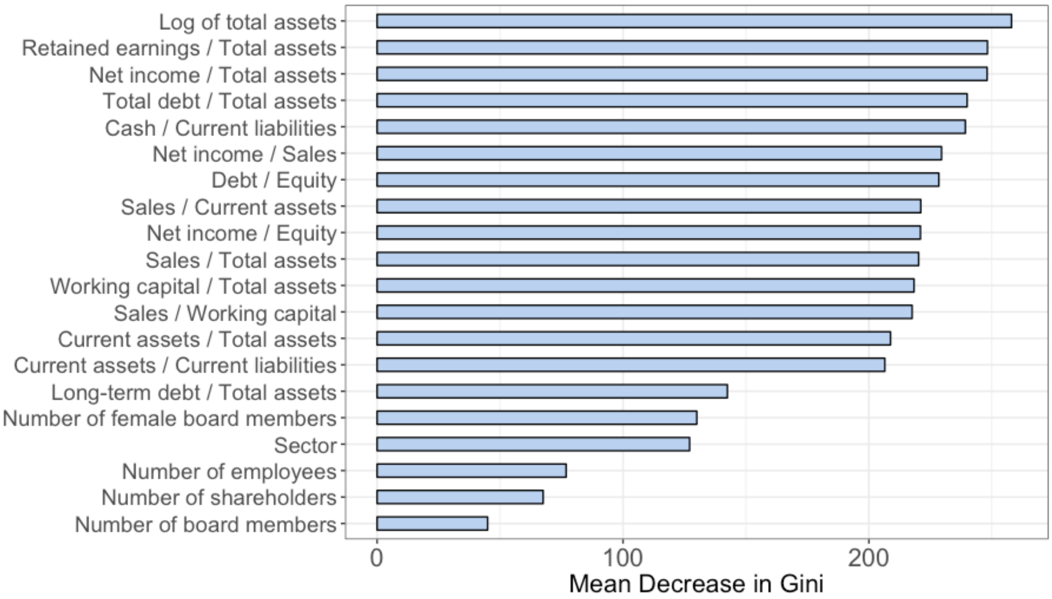


Figure 5.3: Variable importance in random forest

5.3 Comparison

The results from logistic regression and random forest are presented in Table 5.5. Both models were able to correctly classify approximately 59.8% of the healthy firms. For the distressed firms, logistic regression correctly classified 60.52%, compared to random forest that was able to predict 62.55% correctly. The results show that both models have some classifying power, but random forest is slightly superior because of the higher TPR, overall accuracy and AUC. Although random forest represents a slight improvement over logistic regression, we notice that overall error is quite high. In fact, 39.95% of the observations were classified incorrectly. The results are hence not very impressive.

	Accuracy	*TPR*	*TNR*	*AUC*
Logistic Regression	59.83%	60.52%	59.75%	0.6407
Random Forest	60.05%	62.55%	59.77%	0.6558

Table 5.5: Comparison of the results from logistic regression and random forest

The ROC curves from both models are displayed together in Figure 5.4. As we can observe, the curves are very close to each other, indicating similar performance of both models. For the overall TNR, both models perform fairly well. The performances for TPR are however not so well unless trading away larger amounts of TNR, causing poor overall accuracy if we want a reasonable TPR. The slightly higher curve of random forest indicates that this model is superior.

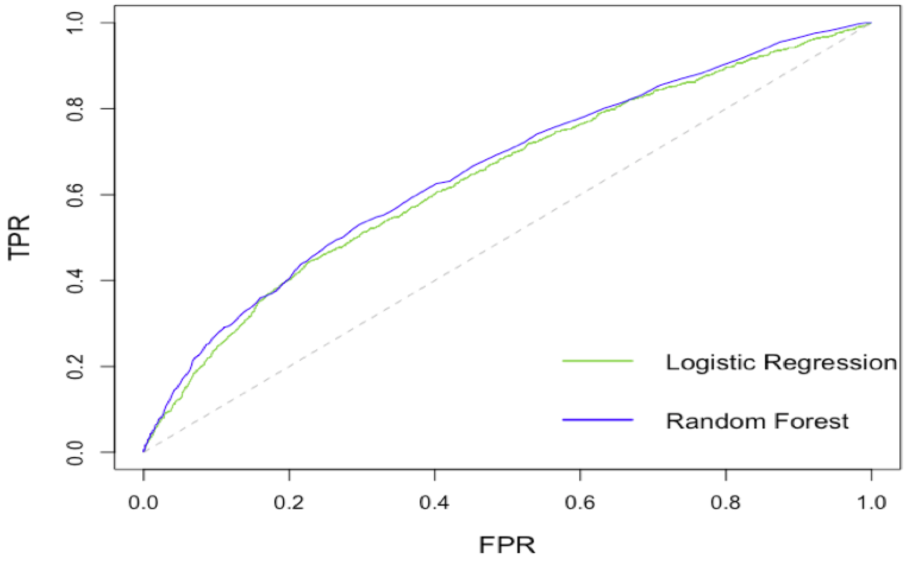


Figure 5.4: ROC curves of logistic regression (green) and random forest (blue)

6 Discussion

Developing models to predict financial distress aids in preventing companies from going bankrupt and contributes to maintaining a healthy market. In this thesis we have developed two models to predict financial distress using accounting data from Norwegian companies. The objective of this section is to provide an answer to our research questions and discuss our results in comparison to previous literature.

6.1 Models of Financial Distress

In our first research question, we aimed to address whether our model had a good prediction power which could be applied to real world situations. We obtained similar results for logistic regression and random forest, with AUC of respectively 0.6407 and 0.6558, and overall accuracy of 59.83% and 60.05%. Regarding the TPR and TNR, logistic regression was able to predict 60.52% correctly for the distressed firms, while 59.75% for the healthy firms. For random forest, these numbers were 62.55% and 59.77% respectively. Overall, random forest showed a slightly superior performance compared to the logistic model. The results indicate that our models have certain predictive power, but there is room for improvement.

The superior performance of random forest might indicate non-linear relationships in the data, which we believe is true after what we have argued about the debt ratios. Logistic regression is generally better suited when data is linearly separable. An advantage of random forest is the decorrelation procedure that the model imposes on each subtree, making it suited to process features that are correlated. Moreover, it seemed like random forest were superior in learning the complex patterns in the class-imbalanced data as less adjustment of the threshold (9.5%) was needed in order to maximize TPR and TNR compared to logistic regression that had to decrease the threshold to 8.6%. While random forest seemed to perform slightly better in our data, the complex model of 300 trees took considerably longer time to compute. Logistic regression has the advantage of being easier to interpret and has a faster computational time.

Comparing the performance of our models to previously created models at two years prior, we noticed that all of the three models: Beaver's univariate model, Altman's Z-score and Ohlson's O-score outperformed even our best random forest model in terms of overall accuracy, TPR and TNR. The different results may nevertheless be explained by a numerous of reasons, like the different definitions of financial distress and approaches of modelling and testing.

We have mentioned in section 2.1 how different financial distress has been used in the past due to the vagueness of the term. Altman and Ohlson solely looked at bankruptcy as the criteria of distress, whereas Beaver looked at it in a broader spectrum including companies at the stages before bankruptcy. Our choice of a broader term of financial distress is aligned with Beaver's definition and is motivated by the fact that we want an early prediction of financial distress. We believe that this has had an impact on the performance, as bankruptcy is an extreme situation which might be easier for the model to learn, as opposed to our extended definition.

Moreover, two of the aforementioned models, Beaver's model and Altman's Z-score, were created based on a balanced response variable. Modified distributions such that both responses are equally likely to be detected by the model might be argued are not representable of real-world distributions. This makes the models less applicable in real use, which is why we decided to keep the imbalanced distribution. Besides keeping the real distribution when sampling, we obtained a considerably larger sample size of data, in order to create a more representative sample of the population and to limit the influence of extreme observations. Another difference is our use of recent data. Previous models were built based on company data from their period. We assume that the performance of the previous models will no longer be as impressive when applied to companies today.

Lastly, but perhaps the most important reason for different results is the various approaches in testing the models. We have tested our models using an unseen test set which was held out from the beginning. The models of Alman and Ohlson however, were tested using in-sample data, meaning that the same data used to train the model was used to test the model. We learnt that when the models that predict one year prior to failure were tested out of sample, the performances were significantly worse. We do not have the exact numbers for two years prior however, but we can still argue that Alman's and Ohlson's two years prior *in-sample* accuracy of respectively 72% and 95.55% are not very impressive, compared to our best model's 60.05% *out-of-sample* accuracy.

6.2 Warning Signs of Financial Distress

One part of a successful model is to have strong predictive power, but another aspect is that it is imperative for companies to receive warning signs early enough to take action in time. Hence, in our second research question we aimed to identify warning signs based on the developed models. Here we have shown in both models that Net income/Total assets (ROA), Retained earnings/Total assets, Total debt/Total assets, Debt/Equity and Log of total assets are important variables when trying to predict financial distress, correlating with previous studies. ROA was included in both Beaver's and Ohlson's models. Altman used EBIT/Total assets which is strongly correlated with ROA. Retained earnings/Total assets was included by Altman, whereas Ohlson found size (Log of Total assets/GNP price-level index) and debt to equity as important variables. Interestingly, our models revealed two variables as novel warning signs of financial distress. The mentioned variables should be paid extra attention to while assessing the financial situation of a company.

As frequently suggested by other models (Beaver, 1966; Altman, 1968; Ohlson, 1980), profitability is crucial for long-term survival. Poor profits (ROA) is usually the first sign that a business is not doing well as it means that it struggles to sustain itself from internal funds. If a business is forced to raise money externally, it will raise its business risk. Additionally, if the accumulated profits are low (Retained earnings/Total assets), as a result of poor profits or because the profits rather are distributed as dividends, the company is also more likely to experience financial distress. A company should have a clear path to profitability and reinvest some of the profits in order to sustain itself and fund additional growth of the business.

Furthermore, we suggest companies to maintain a reasonable level of leverage (D/E). As mentioned before, a D/E ratio between 1 and 1.5 is generally considered as good, but the optimal ratio is dependent on several factors. Size of the company is one of the factors that should be taken into account. Companies with higher leverage ratio normally have higher risk of running into financial distress. However, companies that have access to large debts are usually large, hence large companies often have higher level of leverage, but large companies are also less likely to suffer from financial distress as we will see in the next paragraph.

Size (Log of assets) of the company matters because large companies usually are well-established, have more capital and can easily obtain financial support compared to small companies, and are therefore unsurprisingly more likely to survive. As such, when assessing the financial situation of a company, size should be taken into consideration.

In addition the the warning signs above, our models identify two novel warning signs of financial distress – warning signs which may have been overlooked in the past. These include Cash/Current liabilities (Cash ratio) and Net income/Sales (Net profit margin), which were found in random forest. The first novel warning sign of financial distress as highlighted by our model is cash ratio. An early sign of things that are going wrong is related to a company's liquidity and constant lack of cash. The cash ratio measures a company's ability to cover its short-term obligations using only cash, therefore, is an indicator of its ability to maintain a healthy cash flow. The significance of cash ratio once again proves that for a business *Cash is king*. Hence, we suggest companies to maintain a healthy amount of cash in the account in order to cover its current liabilities and avoid heading towards the state of financial distress.

Secondly, our model further highlights how the profits of a company can reflect their situation. In addition to ROA and Retained earnings/Total assets, net profit margin should be taken into account. Net profit margin reveals how effective a company is at converting revenue into profits. Companies with high profit margin are indicative of good cost control and correct pricing of their products. A low margin suggests the opposite, where prices are too low and/or expenses too high, indicating poor operational efficiency. The ratio is particularly useful when comparing companies within the same industry, as they have approximately the same cost structures (Bragg, 2019). In order to stay competitive in the market, we suggest companies to aim for a net profit margin not less than the industry average.

Lastly, it is worth mentioning that the presented signs of financial distress are symptoms of financial failure and must not be confused with the causes of it. The key is to catch the signs early, so the management can begin to identify the causes and take actions accordingly before it is too late.

6.3 Limitations

Although our models have relatively good performance in both cross-validation and test set, there are some limitations which need to be addressed. One of the main concerns is how and which variables were chosen. Firstly, the most popular financial ratios were chosen as variables for our models. However, the status of companies is a dynamic process and to fully describe the situation a company is facing, financial ratios may not be sufficient to accurately reflect the whole situation. For this reason, we also included some company information as variables. Although we have attempted to take the dynamic nature of a company into account, the data which we have access to is limited. Other factors such as internal processes, competition in the market and macroeconomic conditions can all affect the financial state of a company.

Another limitation is how we have defined financial distress. In this thesis, we created our own criteria for financial distress after carefully reviewing previous research. More specifically, the criteria we proposed was based on company ratings provided by Dun & Bradstreet. However, using criteria of financial distress based on the rating system of one company combined with our own subjective opinions, may be a naive solution. As such, this definition of financial distress may contribute to an uncertain prediction result.

In addition, although our models yield relatively good AUC, in terms of prediction accuracy it comes out weaker. The fact that our model is universal instead of only focusing on a certain industry or size of companies, may result in an unsatisfactory performance of the models. Due to this and the limitations described above, we have not yet found a suitable solution for having high TPR, TNR and overall accuracy. Despite this, we believe that the current attempt in developing a prediction model for financial distress provides some ideas for researchers in the future.

6.4 Future Directions

Along with the development of machine learning and deep learning fields, increasingly advanced modelling methods can be used in the financial prediction area. We are looking forward to the rise of more non-traditional models with better performance. Furthermore, data mining and data crawling techniques may help to gather additional information for modelling, including shareholding structures or even market signals. Conducting text mining and sentiment analysis of annual reports or business news could also be an exciting direction to take. Moreover, building models focusing on a specific size of companies or further investigating sector-specific features would also be interesting, giving greater application value.

7 Conclusion

The main objective of the thesis was to create a two-year model for financial distress prediction and provide early signs for companies heading towards such state, instead of focusing on bankruptcy predictions that we have seen apparent in the past. The created models utilize financial ratios and other features based on company information, and are applicable to all sectors and sizes of companies in the Norwegian market.

An exact definition of financial distress is non-existent. As such, the criteria for financial distress was based on a subjective opinion combining bankruptcy data and careful incorporation of company ratings provided by Dun & Bradstreet. In this paper, we have defined companies as distressed if they went bankrupt or were rated C for two consecutive years.

The logistic regression and the random forest method were chosen for their complementary properties. Logistic regression is a traditional and simple model, whereas random forest is a complex recent method. Both models were optimized with respect to AUC. Moreover, we maximized TPR and TNR by adjusting the classification threshold in order to equally balance them.

Our research indicates that random forest is slightly superior to the logistic regression model. Random forest and logistic regression obtained an AUC of 0.6558 and 0.6407 respectively. The two models that yielded similar performance were able to correctly predict ~ 60% of both healthy and financially distressed companies, indicating that the models have some class separation ability. Additionally, important variables that were derived from the models may represent warning signs of financial distress and should be considered when assessing the financial state of a company. The models assigned high importance to some previously used ratios, such as Size (Log of total assets), ROA, Retained earnings/Total assets, Total debt/Total assets and Debt/Equity. However, the models also indicated that Cash ratio and Net profit margin are important variables that have been neglected in the past. Taken together, we believe that our models provide some evidence for financial distress predictability and possible warning signs. With this knowledge and future investigations, we will better be able to develop models which predict financial distress in Norway.

References

Aae , E. L., & Hansen, M. A. (2017). *Bankruptcy prediction : the credit relevance of reclassfied financial statement ratios.* Master thesis, Norwegian School of Economics. Retrieved from http://hdl.handle.net/11250/2487912

Altinn. (2019, October 1). *Valg av organisasjonsform*. Retrieved December 1, 2019, from https://www.altinn.no/starte-og-drive/starte/valg-av-organisasjonsform/

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance, 23*(4), 589-609. Retrieved from https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

Altman, E. I., & Haldeman, R. G. (1977). ZETATM analysis A new model to identify bankruptcy risk of corporations. *Journal of Banking & Finance, 1*(1), 29-54.

Beaver, W. H. (1966). Financial Ratios As Predictors of Failure. *Journal of Accounting Research, 4*, 71-111. Retrieved from www.jstor.org/stable/2490171

Beaver, W. H., Correia, M., & McNichols, M. F. (2011). Financial Statement Analysis and the Prediction of Financial Distress. *Foundations and Trends® in Accounting, 5*(2), 99-173. Retrieved from http://dx.doi.org/10.1561/1400000018

Bellovary, J. L., Giacomino, D. E., & Akers, M. D. (2007). A Review of Bankruptcy Prediction Studies: 1930 to Present. *Journal of Financial Education, 33*, 1-42. Retrieved from www.jstor.org/stable/41948574

Berg, D. (2007). Bankruptcy prediction by generalized additive models. *Applied Stochastic Models in Business and Industry*, 129-143. Retrieved from https://doi.org/10.1002/asmb.658

Berner, E., Mjøs, A., & Olving, M. (2016). *Norwegian corporate accounts : documentation and quality assurance of SNF's and NHH's database of accounting and company information for Norwegian companies.* Working paper, SNF. Retrieved from https://openaccess.nhh.no/nhh-xmlui/handle/11250/2431354

Bernhardsen, E. (2001). *A Model of Bankruptcy Prediction.* Working paper, Norges Bank. Retrieved from Norges Bank: https://www.norges-bank.no/en/news-events/news-publications/Papers/Working-Papers/2001/200110/

Bragg, S. (2019, January 8). *Net profit margin*. Retrieved 12 December, 2019, from AccountingTools: https://www.accountingtools.com/articles/what-is-net-profit-margin.html

Chen, G. M., & Merville, L. J. (1999). An Analysis of the Underreported Magnitude of the Total Indirect Costs of Financial Distress. *Review of Quantitative Finance and Accounting, 13*(3), 277–293. Retrieved from https://link.springer.com/article/10.1023%2FA%3A1008370531669

Cybinski, P. (2003). *Doomed Firms: An Econometric Analysis of the Path to Failure.* London: Routledge. Retrieved from https://doi.org/10.4324/9781315199351

Deep, R. (2006). *Probability and Statistics: With Integrated Software Routines.* Elsevier Academic Press.

Dichev, I. D. (1998). Is the Risk of Bankruptcy a Systematic Risk? *The Journal of Finance, 53*(3), 1131-1147. Retrieved from https://doi.org/10.1111/0022-1082.00046

Dun & Bradstreet, Inc. (2013). *D&B Cross Border Insight Capabilities.* Retrieved from http://www.dnb-nederland.nl/data/sitemanagement/media/Brochures/D&B_International_Capabilities_22012014.pdf

FitzPatrick, P. J. (1932, October). A comparison of the ratios of successful industrial enterprises with those of failed companies. *Certified public accountant*, 598-605.

Freitas, L., & Freitas, A. P. (2013). *Multivariate Analysis in Management, Engineering and the Sciences.* Rijeka: InTech.

Hamilton, D. T., Ou, S., Kim, F., & Cantor, R. (2007). *Moody's Special Comment.* Moody's investor Service. Retrieved from https://www.moodys.com/sites/products/DefaultResearch/2006400000429618.pdf

He, H., & Ma, Y. (2013). *Imbalanced Learning: Foundations, Algorithms, and Applications.* Wiley-IEEE Press.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning* (8th Edition ed.). Springer. Retrieved from http://faculty.marshall.usc.edu/gareth-james/ISL/

Little, R. J., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data, Second Edition.* John Wiley & Sons, Inc. Retrieved from http://dx.doi.org/10.1002/9781119013563

Madan, B. K. (1978). *Report on a Study of the Debt-equity Ratio Norms.* University of Michigan.

Meese, E. N., & Viken, T. (2019). *Machine learning in bankruptcy prediction : utilizing machine learning for improved bankruptcy predictions in the Norwegian market with an emphasis on financial, management and sector statements.* Master thesis, Norwegian School of Economics. Retrieved from http://hdl.handle.net/11250/2611655

Min, J. H., & Lee, Y. C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications, 28*(4), 603-614. Retrieved from https://doi.org/10.1016/j.eswa.2004.12.008

Moro, S., Cortez, P., & Rita, P. (2014, June). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems, 62*, 22-31. Retrieved from https://doi.org/10.1016/j.dss.2014.03.001

Nervik, S. (2019, November 20). *Antallet konkurser i Norge øker - nærmer seg nivået under finanskrisen*. Retrieved December 15, 2019, from Nettavisen: https://www.nettavisen.no/okonomi/antallet-konkurser-i-norge-oker---naermer-seg-nivaet-under-finanskrisen/3423880728.html?fbclid=IwAR1D1pmA2q5qCMd67SRunm2CQfQHi-QAq_42kcDyK15zSkoRosRWZPPxbUE

Odom, M., & Sharda, R. (1990). A neural network model for bankruptcy prediction. *1990 IJCNN International Joint Conference on Neural Networks, 2*, 163 - 168. Retrieved from https://ieeexplore.ieee.org/document/5726669

Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research, 18*(1), 109-131. Retrieved from https://www.jstor.org/stable/2490395

Petersen, C. V., & Plenborg, T. (2012). *Financial Statement Analysis: Valuation, Credit analysis and Exective compensation.* Essex, England: Pearson Education Limited.

Platt, H. D., & Platt, M. B. (2002). Predicting corporate financial distress: Reflections on choice-based sample bias. *Journal of Economics and Finance, 26*(2), 184–199. Retrieved from https://doi.org/10.1007/BF02755985

Shumway, T. (2001, January). Forecasting Bankruptcy More Accurately: A Simple Hazard Model. *The Journal of Business, 74*(1), 101-124. Retrieved from https://www.jstor.org/stable/10.1086/209665

Theodossiou, P. T. (1993). Predicting Shifts in the Mean of a Multivariate Time Series Process: An Application in Predicting Business Failures. *Journal of the American Statistical Association, 88*(422), 441-449. Retrieved from https://doi.org/10.1080/01621459.1993.10476294

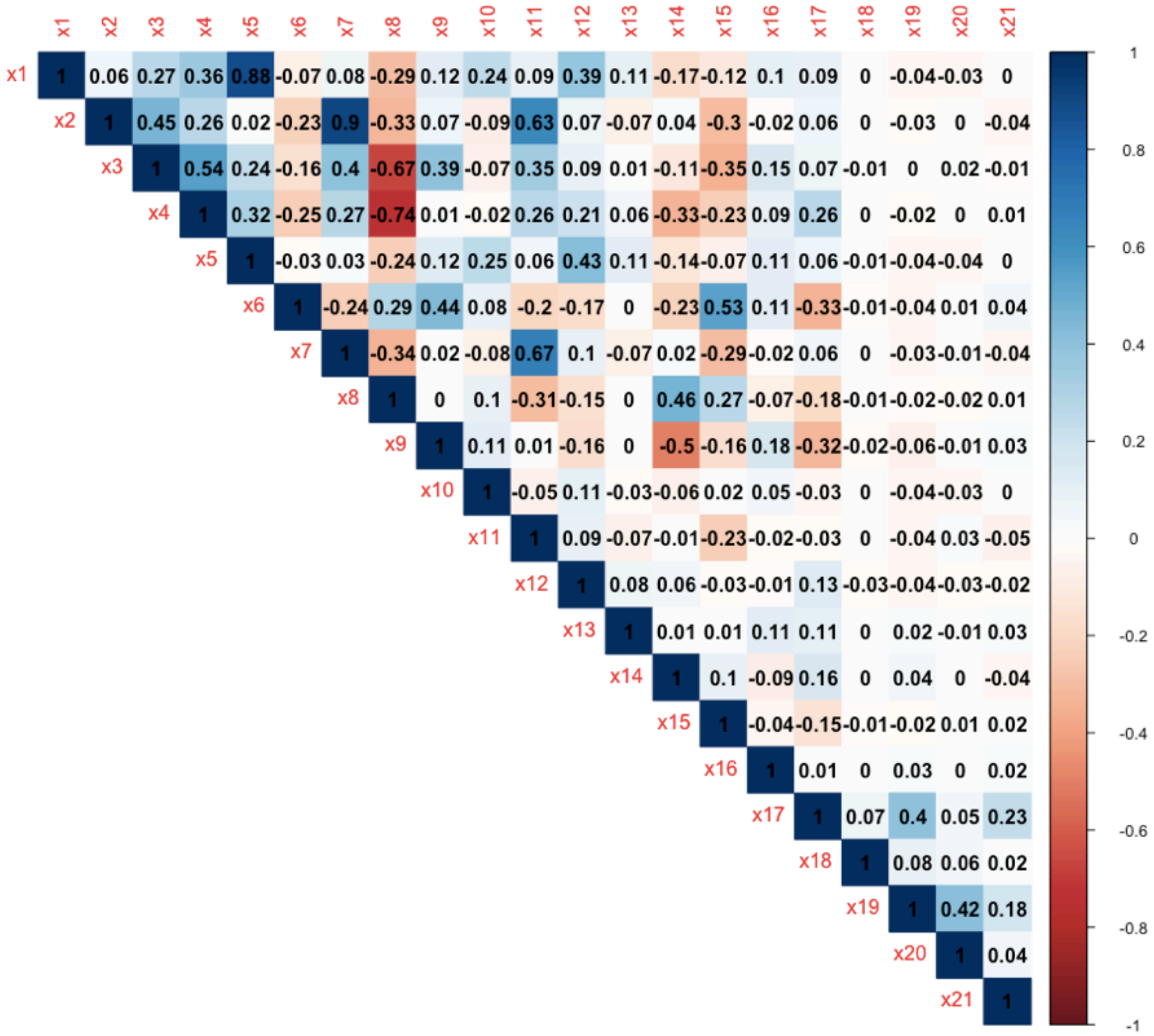
Tinoco, M. H., & Wilson, N. (2013). Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables. *International Review of Financial Analysis, 30*, 394-419. Retrieved from https://doi.org/10.1016/j.irfa.2013.02.013

United Airlines, Inc. (2018). *Annual Report on Form 10-K.* Washington, DC: United States Securities and Exchange Commission. Retrieved from http://ir.united.com/static-files/7764f912-7135-486b-aa5c-f729fa8b7cb3

Wruck, K. H. (1990). Financial distress, reorganization, and organizational efficiency. *Journal of Financial Economics, 27*(2), 419-444. Retrieved from https://doi.org/10.1016/0304-405X(90)90063-6

Appendix

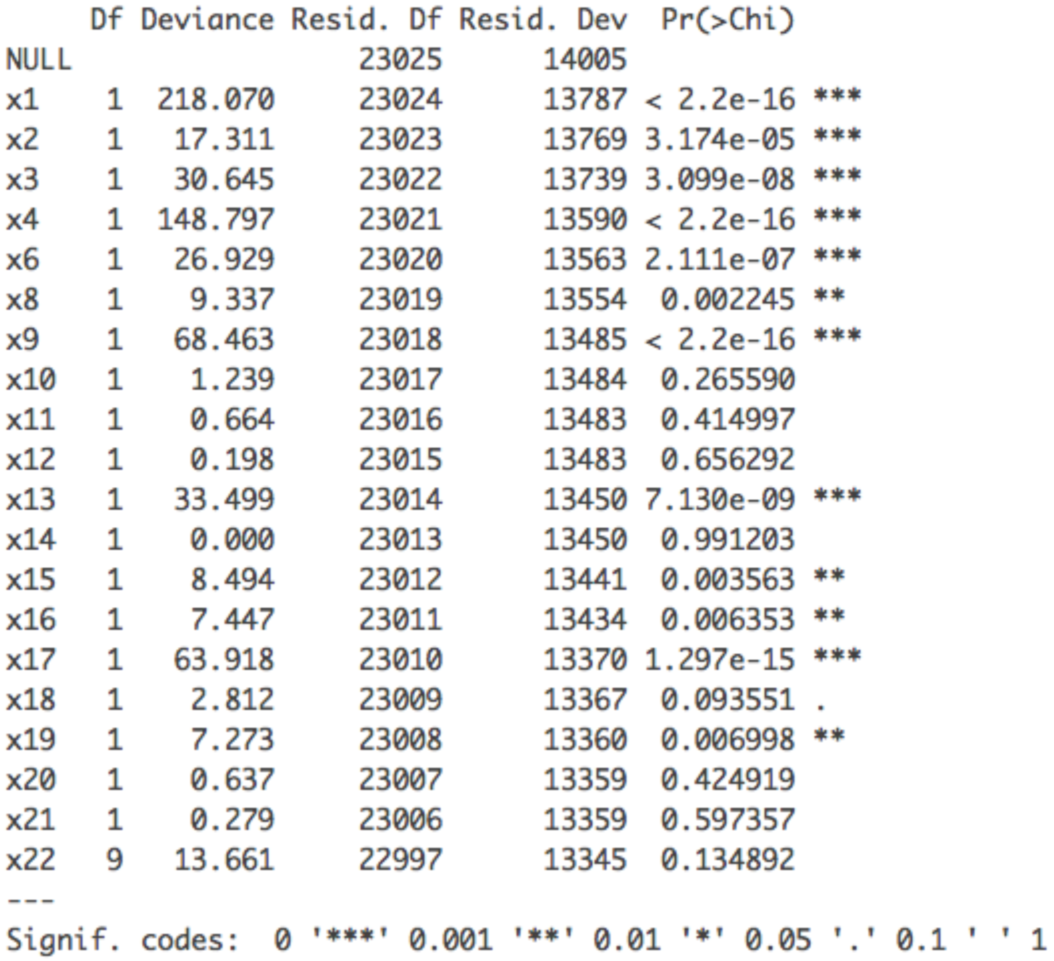
A1 Correlation Matrix



	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	x17	x18	x19	x20	x21
x1	1	0.06	0.27	0.36	0.88	-0.07	0.08	-0.29	0.12	0.24	0.09	0.39	0.11	-0.17	-0.12	0.1	0.09	0	-0.04	-0.03	0
x2		1	0.45	0.26	0.02	-0.23	0.9	-0.33	0.07	-0.09	0.63	0.07	-0.07	0.04	-0.3	-0.02	0.06	0	-0.03	0	-0.04
x3			1	0.54	0.24	-0.16	0.4	-0.67	0.39	-0.07	0.35	0.09	0.01	-0.11	-0.35	0.15	0.07	-0.01	0	0.02	-0.01
x4				1	0.32	-0.25	0.27	-0.74	0.01	-0.02	0.26	0.21	0.06	-0.33	-0.23	0.09	0.26	0	-0.02	0	0.01
x5					1	-0.03	0.03	-0.24	0.12	0.25	0.06	0.43	0.11	-0.14	-0.07	0.11	0.06	-0.01	-0.04	-0.04	0
x6						1	-0.24	0.29	0.44	0.08	-0.2	-0.17	0	-0.23	0.53	0.11	-0.33	-0.01	-0.04	0.01	0.04
x7							1	-0.34	0.02	-0.08	0.67	0.1	-0.07	0.02	-0.29	-0.02	0.06	0	-0.03	-0.01	-0.04
x8								1	0	0.1	-0.31	-0.15	0	0.46	0.27	-0.07	-0.18	-0.01	-0.02	-0.02	0.01
x9									1	0.11	0.01	-0.16	0	-0.5	-0.16	0.18	-0.32	-0.02	-0.06	-0.01	0.03
x10										1	-0.05	0.11	-0.03	-0.06	0.02	0.05	-0.03	0	-0.04	-0.03	0
x11											1	0.09	-0.07	-0.01	-0.23	-0.02	-0.03	0	-0.04	0.03	-0.05
x12												1	0.08	0.06	-0.03	-0.01	0.13	-0.03	-0.04	-0.03	-0.02
x13													1	0.01	0.01	0.11	0.11	0	0.02	-0.01	0.03
x14														1	0.1	-0.09	0.16	0	0.04	0	-0.04
x15															1	-0.04	-0.15	-0.01	-0.02	0.01	0.02
x16																1	0.01	0	0.03	0	0.02
x17																	1	0.07	0.4	0.05	0.23
x18																		1	0.08	0.06	0.02
x19																			1	0.42	0.18
x20																				1	0.04
x21																					1

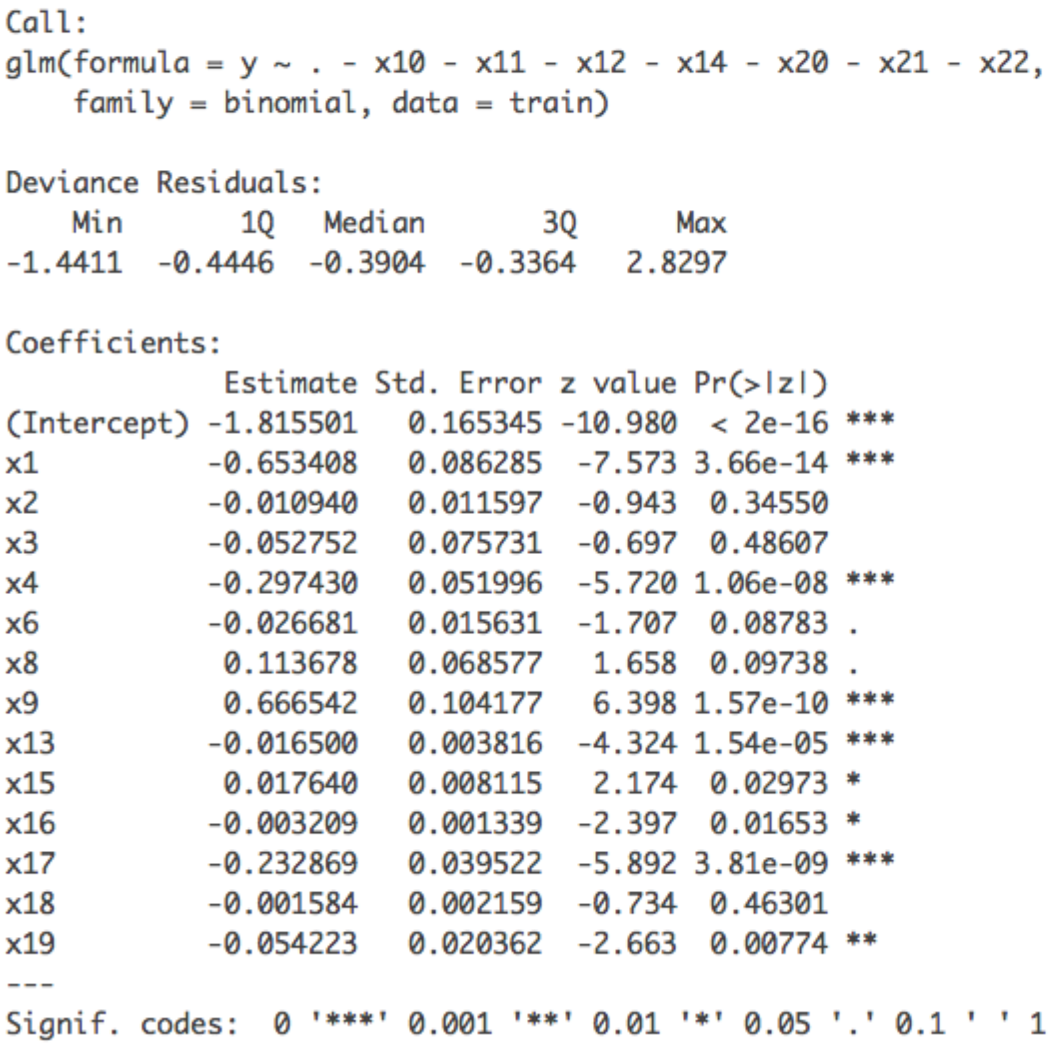
Figure A1.1: Correlation Matrix of variables from x1 to x21 (excluding the categorical variable x22)

A2 Logisitc Regression Results



```
     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                 23025      14005              
x1    1  218.070     23024      13787 < 2.2e-16 ***
x2    1   17.311     23023      13769 3.174e-05 ***
x3    1   30.645     23022      13739 3.099e-08 ***
x4    1  148.797     23021      13590 < 2.2e-16 ***
x6    1   26.929     23020      13563 2.111e-07 ***
x8    1    9.337     23019      13554  0.002245 ** 
x9    1   68.463     23018      13485 < 2.2e-16 ***
x10   1    1.239     23017      13484  0.265590    
x11   1    0.664     23016      13483  0.414997    
x12   1    0.198     23015      13483  0.656292    
x13   1   33.499     23014      13450 7.130e-09 ***
x14   1    0.000     23013      13450  0.991203    
x15   1    8.494     23012      13441  0.003563 ** 
x16   1    7.447     23011      13434  0.006353 ** 
x17   1   63.918     23010      13370 1.297e-15 ***
x18   1    2.812     23009      13367  0.093551 .  
x19   1    7.273     23008      13360  0.006998 ** 
x20   1    0.637     23007      13359  0.424919    
x21   1    0.279     23006      13359  0.597357    
x22   9   13.661     22997      13345  0.134892    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table A2.1 : Results from chi-squared test in logistic regression

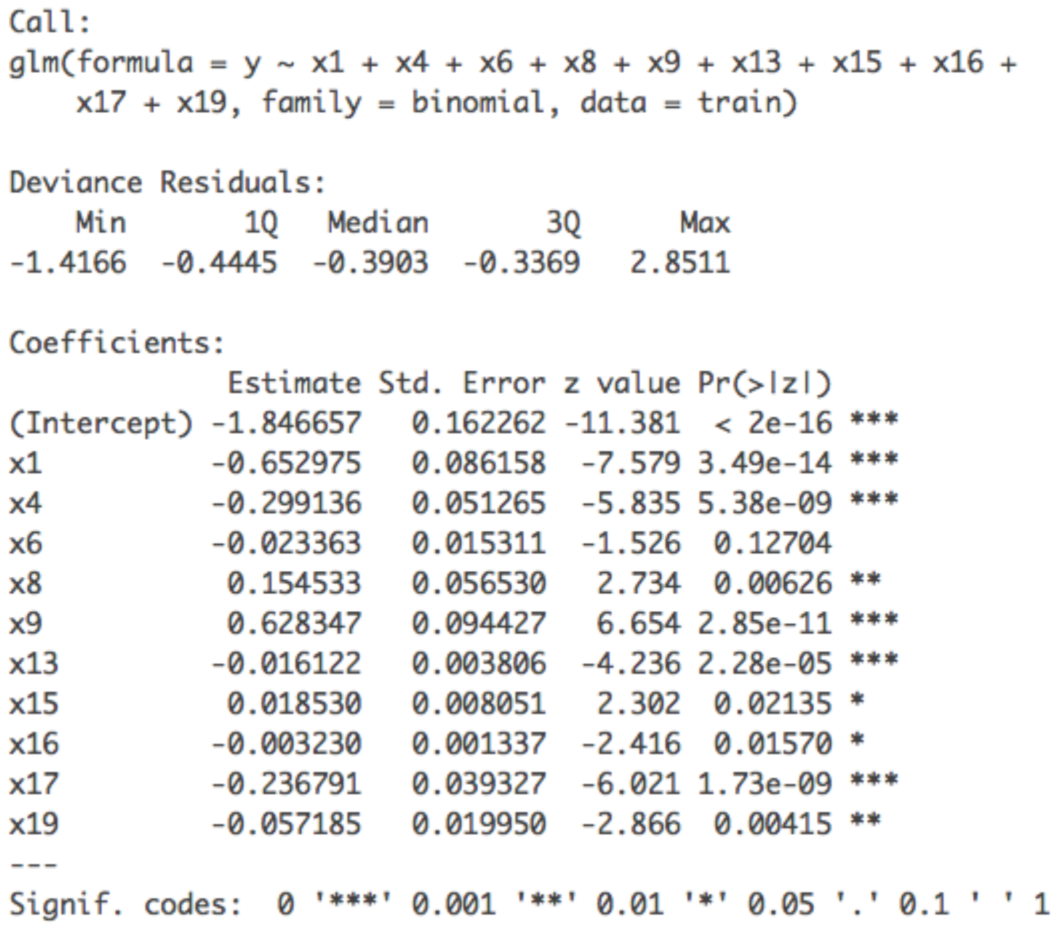


```
Call:
glm(formula = y ~ . - x10 - x11 - x12 - x14 - x20 - x21 - x22,
    family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4411  -0.4446  -0.3904  -0.3364   2.8297

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.815501   0.165345 -10.980  < 2e-16 ***
x1          -0.653408   0.086285  -7.573 3.66e-14 ***
x2          -0.010940   0.011597  -0.943  0.34550
x3          -0.052752   0.075731  -0.697  0.48607
x4          -0.297430   0.051996  -5.720 1.06e-08 ***
x6          -0.026681   0.015631  -1.707  0.08783 .
x8           0.113678   0.068577   1.658  0.09738 .
x9           0.666542   0.104177   6.398 1.57e-10 ***
x13         -0.016500   0.003816  -4.324 1.54e-05 ***
x15          0.017640   0.008115   2.174  0.02973 *
x16         -0.003209   0.001339  -2.397  0.01653 *
x17         -0.232869   0.039522  -5.892 3.81e-09 ***
x18         -0.001584   0.002159  -0.734  0.46301
x19         -0.054223   0.020362  -2.663  0.00774 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table A2.2: Results of logistic regression after removing insignificant variables according to chi-squared test



```
Call:
glm(formula = y ~ x1 + x4 + x6 + x8 + x9 + x13 + x15 + x16 +
    x17 + x19, family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4166  -0.4445  -0.3903  -0.3369   2.8511

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.846657   0.162262 -11.381  < 2e-16 ***
x1          -0.652975   0.086158  -7.579 3.49e-14 ***
x4          -0.299136   0.051265  -5.835 5.38e-09 ***
x6          -0.023363   0.015311  -1.526  0.12704
x8           0.154533   0.056530   2.734  0.00626 **
x9           0.628347   0.094427   6.654 2.85e-11 ***
x13         -0.016122   0.003806  -4.236 2.28e-05 ***
x15          0.018530   0.008051   2.302  0.02135 *
x16         -0.003230   0.001337  -2.416  0.01570 *
x17         -0.236791   0.039327  -6.021 1.73e-09 ***
x19         -0.057185   0.019950  -2.866  0.00415 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table A2.3: Results from final logistic model