Copenhagen Business School

Master's Thesis

Advanced Economics and Finance

Neural Networks and Logistic Regression for Credit Rating Prediction: A Comparison of the Manufacturing and Retail Industry

Author
David Tobjörk

Supervisor
Michael Ahm

September 17, 2018

Number of standard pages: 44 ,Characters: 55 695, Student number: 040590DAT1

I started out writing a thesis together with Janid Abdellah, but no material has been shared between us and only materiel written by me was kept for this thesis. New thesis contracts were also submitted and the majority of the material was produced after terminating the partnership.

Abstract

Credit ratings are expensive to obtain and there is a risk of bias in the credit rating process. It is therefore interesting to analyze the credit rating process to see if it can be automated. In this thesis it has been investigated how the credit ratings of firms in the manufacturing industry as well as in the retail industry can be modelled using neural networks and logistic regression. Data sets of US credit ratings from Standard and Poor's for the period 1 January 2013 - 31 December 2017 were used together with publicly available accounting data. For both industries the neural network outperformed the logistic regression, in both terms of accuracy and Cohen's kappa. In most previous research using neural networks to predict credit ratings, focus have solely been on the predictive performance. In this thesis the neural networks was further analyzed using the connection weight approach. The analysis showed that total assets is a especially important explanatory variable for Standard & Poors's credit ratings on manufacturing firms. For the retail firms the cash flow from operations to current liabilities was found to be important. Furthermore, earnings per share was found to be among the most important variables for both industries.

Contents

1 Introduction **5**
1.1 Background . . . 5
1.2 Problem Statement . . . 6
1.3 Structure of the Thesis . . . 7

2 Literature Review **9**

3 Credit Ratings **12**
3.1 Introduction . . . 12
3.2 Credit Rating Methodology . . . 14

4 Modelling Techniques **17**
4.1 Logistic Regression . . . 17
4.1.1 Logistic Regression with Binary Response . . . 17
4.1.2 Logistic Regression with More than Two Response Classes . . . 19
4.2 Artificial Neural Networks . . . 21
4.2.1 Network Structure . . . 21
4.2.2 Fitting the Neural Network . . . 22
4.2.3 Starting Weights . . . 25
4.3 Overfitting and Underfitting . . . 25
4.3.1 Hyperparameters . . . 25
4.3.2 K-Fold Cross Validation . . . 26

5 Evaluation Metrics **27**
5.1 Accuracy . . . 27
5.2 Kappa . . . 27
5.3 The Connection Weight Approach . . . 28

6 Data **29**
6.1 Data Selection . . . 29
6.2 Data Preparation . . . 30
6.3 Grouping of Ratings . . . 32
6.4 Training and Test Set . . . 33

7 Methodology **34**
7.1 Software . . . 34
7.2 Hyperparameters and Model Training . . . 34
7.3 Model Evaluation and Analysis . . . 36

8 Modelling Results and Analysis **37**
8.1 Manufacturing . . . 37
8.2 Retail . . . 37
8.3 Comparison . . . 38
8.4 Variable Importance . . . 39

9 Conclusion and Further Work **44**

A Appendix **46**
A.1 Input Variables . . . 46
References . . . 48

Chapter 1

Introduction

This chapter introduces credit ratings and their purpose. Also, the topic of modelling credit ratings and the research question is presented. Finally, the structure and content of the remaining chapters are given.

1.1 Background

Credit ratings are opinions about credit risk given by credit rating agencies. These ratings are given to issuers such as corporations or governments as well as to individual issues such as corporate bonds. The scale used to express the agencies opinion on credit risk differ between the agencies, but often ratings are expressed on a scale of letter combinations (Standard & Poor's, 2015).

What is the purpose of credit rating agencies, that specialize in giving opinions about creditworthiness? Arguments related to information asymmetries in the credit market can be made. A borrower knows its creditworthiness better than the lender. The lender can get information by conducting its own research or use information from credit rating agencies. Ackerloff (1970) describes with an example of the used car market how market failure can occur when there is asymmetric information. If buyers can not distinguish between good cars and bad cars they are only willing to pay the average value. However, the sellers know the value and will only sell the bad ones with a higher price than the value. This in turn results in the buyers being willing to pay even less and the quality of cars decreasing further.

This can be translated to the credit markets, were lenders are unable to distinguish between high-risk and low-risk borrowers and therefore charge the same interest rate. The low-risk borrowers would suffer from having to pay the same interest rate as the high risk ones and lenders would allocate less than optimal resources to low-risk borrowers. Reality is somewhere between no information and perfect information. Here the credit ratings in addition to its own analysis allows the lender to distinguish between borrowers, but not perfectly (Adelson, 2012). Credit ratings are of value to the issuers of debt themselves by giving access to capital markets. Credit ratings help the issuer in communicating its credit risk and to get access to more investors. Furthermore, ratings help in anticipating interest rates on an issue (Standard & Poor's, 2015).

Another function of credit ratings is that they make it harder for one investor to make better judgment about creditworthiness than others. With credit ratings available for all market participants the market efficiency can be improved (Adelson, 2012).

However, the credit rating process tend to be costly as allot of human resources are spent on analyzing the issuer. The rating process is preformed by experts at the agencies and often involves quantitative as well as qualitative analysis. Moreover, it is a highly concentrated industry. There is three major players in the credit rating industry: Standard & Poor's, Moody's and Fitch. As of 31 of December 2016 they represented 48.9, 34.2 and 13.3 percent of the outstanding ratings respectively (U.S. Securities and Exchange Commission, 2017). The fact that the industry is highly concentrated and that the major rating agencies primarily generate revenue from the issuers have raised concerns and criticism of the industry. Furthermore, the financial crisis of 2007-2008 brought allot of attention to the credit rating agencies and their operations. Many suggest that too high credit ratings given to the new type of financial products was an important factor leading to the crisis (Hunt, 2009). Many pension funds and money market funds were limited to invest in products with the highest ratings (McLean & Nocera, 2011).

If the ratings process can be automated it would allow for cost saving and more transparency in the rating process. This has led to several studies on modelling of credit ratings using statistical and machine learning models. Previous studies have used models such as linear regression, logistic regression and neural networks. In the next chapter some of these previous studies are presented. The use of these type of modelling techniques should also be of interest to rating agencies. Even if they do not replace the judgment of the human rater they can offer a valuable additional tool in the rating process.

1.2 Problem Statement

This thesis further investigates neural networks and logistic regression for predictions of credit ratings. However, in this thesis the focus will be on differences when modelling credit ratings for different industries. There is good reason to further investigate this as credit rating agencies such as Standard & Poor's does in fact take into account industry specific characteristics (Standard & Poor's, 2018). Neural networks have shown great promise for modelling of credit ratings in previous research, but the models are often treated as "black boxes". In this study the neural networks will be analyzed further, which hopefully also leads to a better understanding of the credit rating process.

Two aspects are considered: 1) the preferred modelling technique and 2) the input variables. It is of interest to see if these aspects differ between industries and how. A deeper look at the neural networks will be taken, not only evaluating the predictive performance, but also analyzing the input variables. The thesis aims at answering the following questions:

- Is there a difference in importance of input variables between manufacturing and retail companies?
- Does the selected statistical technique differ between manufacturing and retail firms?

Here some limitations have been made. first, only two, but broad industries are considered. These are the manufacturing and retail industry. Second, the credit ratings modelled are those of Standard & Poor's. Other choices of credit ratings could have been made, but as noted above Standard & Poor's represents a large fraction of the outstanding credit ratings. Third, two modelling techniques are considered, neural network and logistic regression.

To determine the preferred statistical technique two evaluation metrics are used, the accuracy and Cohens kappa. The accuracy is commonly used in previous research on the topic. Cohens kappa is a measure that takes in to account some weakness of using the accuracy. To analyze the input variables to the neural networks, the connection weight approach is used. All these measure are further discussed in chapter 5.

1.3 Structure of the Thesis

The remainder of the thesis is structured in the following way. Chapter 2 discusses previous models used for credit rating predictions, variables used and other main findings. Chapter 3 gives some background information on credit ratings and rating methodology. Chapter 4 gives the theory for the two techniques considered in this study, logistic regression and neural networks. How to handle the problems overfitting and underfitting are also discussed. Thereafter chapter 5 presents the metrics used to evaluate and analyze the fitted models. Chapter 6 introduces the data used for the thesis. The data source as well as the choice of input variables and the data cleaning process are discussed. In chapter 7 the methodology is described. Chapter 8 presents and discusses the results for the fitted models and Chapter 9 concludes.

Chapter 2

Literature Review

In this chapter some of the previous research done on credit rating predictions are presented. In previous research many different types of modelling techniques and input variables have been considered. Early studies such as Horrigan (1966) and West (1970) used ordinary least squares (OLS) to predict corporate bond ratings. Instead of Linear regression Pinches and Mingo (1973) used multiple discriminant analysis and Ederington (1985) used logistic regression. Many of the more recent studies have focused on techniques that often are called machine learning techniques. For example Dutta and Shekhar (1988) used neural networks and Huang, Chen, Hsu, Chen, and Wu (2004) support vector machines.

Horrigan (1966) examined the value of using accounting data in the form of financial ratios for predictions of corporate bond ratings. Horrigan predicted Six rating classes from Standard & Poor's as well as Moody's. Since the ratings are given as letter combinations, they were translated into numeric form to allow for the use of OLS. The analysis was done by first selecting the variables highly correlated with the bond ratings and eliminating highly intercorrelated variables. Based on data of stable ratings between 1959-1964 multiple regressions were fitted. Finally the models were tested on two sets of data: 1) Firms that received ratings during 1961-1964 and 2) firms whose previous ratings were changed during 1961-1964. To put each observation in a discrete category the mean of the predicted values for each rating in the original data set was used. That is a prediction on a observation in the test was assigned the category with the closest mean. On the first data set Horrigan achieved an accuracy of 52% and 58% for Standard & Poor's and Moody's respectively. For the second data set the accuracy was 57% for Standard & Poor's and 54% for Moodys.

Fisher (1959) used linear regression to explain risk premiums of corporate bonds. Where the difference between the yield to maturity of the bond and the risk free rate was used as the risk premium. He showed that the risk premiums can be successfully determined by measures of marketability of the bonds, earnings variability, reliability of the firm in meeting its obligations and capital structure

Inspired by Fisher (1959), West (1970) argued that a better model than the one of Horrigan (1966) can be achieved by not only relying on one year of financial data in ratio form. Using a similar model and variables to Fisher West predicted Moody's bond ratings using six classes, and achived an accuracy of 62%.

instead of using linear regression as in previous studies, Pinches and Mingo (1973) tested the usefulness of multiple discriminant analysis for predicting corporate bond ratings. Pinches and Mingos data set consisted of Moody's bond ratings between 1967 and 1968, using only ratings of B or above and excluding the Aaa class due to few observations. That is in total they predicted five different classes of ratings. When testing the model on a holdout sample they achieved an accuracy of 65%.

Dutta and Shekhar (1988) began to examine if neural networks are suitable for predictions of corporate bond ratings. They used data of bonds and ten financial variables taken from the April 86 issues of the Valueline Index and the SS-P Bond Guide. Using two classes (AA or non-AA) they achieved and accuracy of 83% on their holdout sample.

Surkan and Singleton (1990) used neural networks and a data set of 18 telephone operating companies divested by AT and T in 1982. In their analysis they tried to predict whether a bond is rated Aaaa (Moody's highest rating) or one of A1, A2, A3 (the three lowest investment grade ratings of Moody's).They compared the accuracy of networks with one and two hidden layers. They concluded that a network with two hidden layers outperforms a single layer network, independent of the combination of the number of elements in the layers. furthermore, to evaluate the neural networks as a classification technique they used multiple discriminant analysis as a benchmark. Depending on number of of layers and elements neural networks achieved an accuracy between 65 and 88 percent on their holdout sample. While the accuracy for their multiple discriminant analysis was only 39%.

Huang et al. (2004) compared support vector machines with neural networks as well as logistic regression for bond rating predictions. Moreover, they analyzed the neural network and compared the Taiwan and US market using Garson's meassure of variable importance (Garson, 1991). They used two similar data sets, one data set of the Taiwan market and one for the US market. The Taiwan data set consisted of ratings from Taiwan Ratings Corporation (a partner with Standard & Poor's) and 16 financial variables. The data set for the US consisted of Standard & Poor's ratings and the same financial variables, except two that were not available for the

US data set. Both of these data sets consisted of a total of five rating categories, that were used as output variables. When testing the different techniques they considered both the full set of variables as well as a subset of only seven previously commonly used variables. Using 10-fold cross validation as well as leave one out cross validation they found that support vector machines and neural networks consistently achieved a higher accuracy than logistic regression. However, support vector machines and neural networks achieved similar results. They also found that a model with a small set of variables achieves comparable and sometimes better result than the larger model. For the US data set the best model used neural networks and achieved an accuracy of 80.75%. A model using support vector machines and with an accuracy of 79.73% was the best for the Taiwan data.

Kumar and Bhattacharya (2006) compared a three layer neural network with linear discriminant analysis for prediction of corporate bond credit ratings. As dependent variables they used Moody's ratings for 129 companies during the period January 2003 to June 2004. The ratings were divided into six categories and 25 variables were used as inputs. That is their network had 25 elements in the input layer and 6 elements in the output layer. When testing their neural network on a holdout sample it achieved and accuracy of 79% while the multiple discriminant analysis only achieved an accuracy of 33%.

Chapter 3

Credit Ratings

Before taking on the task of modelling credit ratings it is important to understand the agencies issuing them, the purpose of them and the framework used when giving a rating. Therefore this chapter will first give an introduction to the credit rating agencies and the types of ratings. Thereafter the purpose of credit ratings will be discussed and finally the framework of Standard & Poor's credit rating process will be presented.

3.1 Introduction

There are several definitions of credit ratings available, for example the US Securities and Exchange Commission use the following definition:

"A credit rating reflects a rating agency's opinion, as of a specific date, of the creditworthiness of a particular company, security, or obligation" (Langohr & Langohr, 2010).

The credit rating agencies also have their own definitions of credit ratings. Standard & Poor's (2015) define credit ratings as

"Credit ratings are opinions about credit risk. Our ratings express the agency's opinion about the ability and willingness of an issuer, such as a corporation or state or city government, to meet its financial obligations in full and on time."

A credit rating with respect to a specific financial obligation is called issue credit rating while a rating with respect to the obligor's overall is called issuer credit rating. Furthermore, ratings can be either short-term or long-term (Standard &

Poor's, 2014). In this study the long-term issuer credit ratings of Standard & Poor's are used. The opinion of the credit rating agency is often summarized by assigning a rating represented by symbols. The symbols and rating system differ between credit rating agencies. The symbols and definitions used by Standard & Poor's for long-term issuer ratings can be seen in Table 3.1 below:

Table 3.1: Standard & Poor's rating scale

Category	Definition
AAA	"An obligor rated 'AAA' has extremely strong capacity to meet its financial commitments. 'AAA' is the highest issuer credit rating assigned by S&P Global Ratings."
AA	"An obligor rated 'AA' has very strong capacity to meet its financial commitments. It differs from the highest-rated obligors only to a small degree."
A	"An obligor rated 'A' has strong capacity to meet its financial commitments but is somewhat more susceptible to the adverse effects of changes in circumstances and economic conditions than obligors in higher-rated categories."
BBB	"An obligor rated 'BBB' has adequate capacity to meet its financial commitments. However, adverse economic conditions or changing circumstances are more likely to weaken the obligor's capacity to meet its financial commitments. "
BB	"An obligor rated 'BB' is less vulnerable in the near term than other lower-rated obligors. However, it faces major ongoing uncertainties and exposure to adverse business, financial, or economic conditions that could lead to the obligor's inadequate capacity to meet its financial commitments."
B	"An obligor rated 'B' is more vulnerable than the obligors rated 'BB', but the obligor currently has the capacity to meet its financial commitments. Adverse business, financial, or economic conditions will likely impair the obligor's capacity or willingness to meet its financial commitments."
CCC	"An obligor rated 'CCC' is currently vulnerable and is dependent upon favorable business, financial, and economic conditions to meet its financial commitments."
CC	"An obligor rated 'CC' is currently highly vulnerable. The 'CC' rating is used when a default has not yet occurred but S&P Global Ratings expects default to be a virtual certainty, regardless of the anticipated time to default."
R	"An obligor rated 'R' is under regulatory supervision owing to its financial condition. During the pendency of the regulatory supervision, the regulators may have the power to favor one class of obligations over others or pay some obligations and not others."
SD and D	"An obligor rated 'SD' (selective default) or 'D' is in default on one or more of its financial obligations including rated and unrated obligations but excluding hybrid instruments classified as regulatory capital or in nonpayment according to terms. An obligor is considered in default unless S&P Global Ratings believes that such payments will be made within five business days of the due date in the absence of a stated grace period or within the earlier of the stated grace period or 30 calendar days. A 'D' rating is assigned when S&P Global Ratings believes that the default will be a general default and that the obligor will fail to pay all or substantially all of its obligations as they come due. An 'SD' rating is assigned when S&P Global Ratings believes that the obligor has selectively defaulted on a specific issue or class of obligations but it will continue to meet its payment obligations on other issues or classes of obligations in a timely manner. An obligor's rating is lowered to 'D' or 'SD' if it is conducting a distressed exchange offer."

Source: Standard & Poor's (2014).

3.2 Credit Rating Methodology

In this section focus is on the credit rating methodology of Standard & Poor's as these ratings are used in this study. The rating process involves a combination of qualitative and quantitative analysis. Information about the credit rating methodology is released in Standard & Poor's criteria documents. However, these documents only give a description of the framework used not the detailed information about the credit rating assessment. Therefore this thesis can also give more insight into the credit rating process.

When rating non-financial corporates Standard & Poor's credit rating process is divided into several different categories, this is described in Standard & Poor's (2018). First, The Business risk profile is assessed by looking at country risk, industry risk and competitive position. The Business risk profile is then combined with an assessment of the financial risk profile to form an anchor rating. This so called anchor can then be modified by considering several modifiers. Finally, group or government influence is accounted for in some cases to get the issuer credit rating. The credit rating process of Standard & Poor's is illustrated in Figure 3.1.

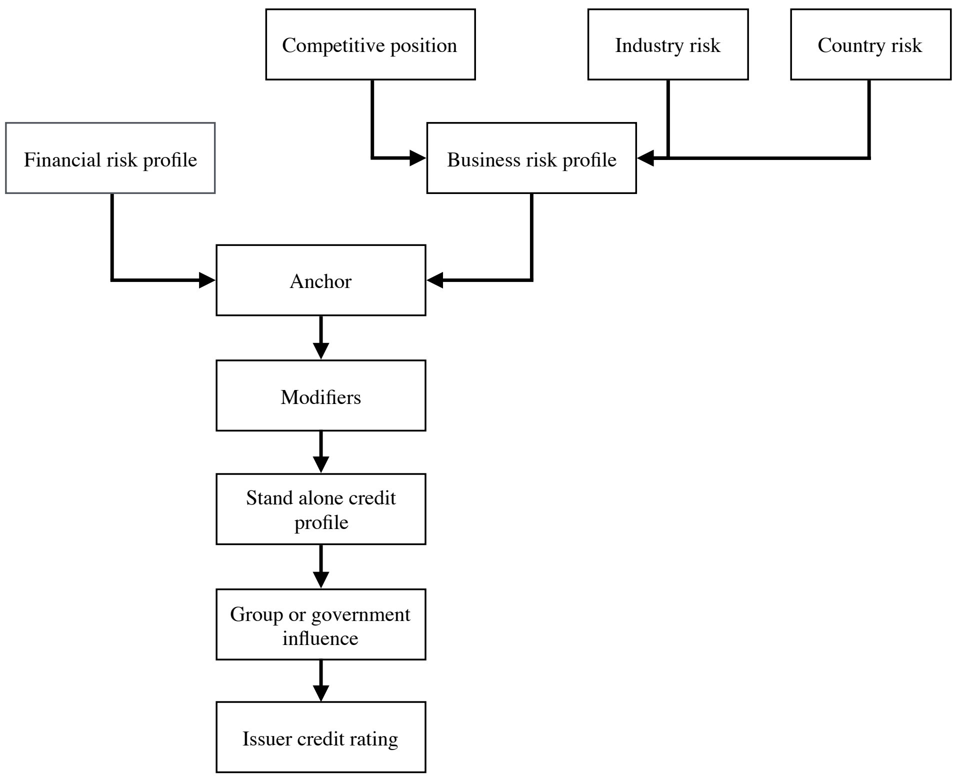

Figure 3.1: The credit rating process of Standard and Poor's.

Standard & Poor's assign a country risk that reflects factors such as economic risk, institutional and governance effectiveness risk, financial system risk, and rule of law/payment culture risk. If the company is exposed to more than one country the risk assessment will be a weighted average of the country risks. To asses the industry risk Standard & Poor's consider, cyclicality as well as competitive risk and growth. Together these factors give the risk of the industry. By combining the country risk and the industry risk Standard & Poor's assign the issuer's Corporate Industry and Country Risk Assessment.

The competitive position of the issuer is combined with the above mentioned corporate industry and country risk assessment to get the business risk profile. When determining the competitive position, four components are considered: competitive advantage, scale/scope/diversity, operating efficiency and profitability. The three first components determine the competitive position, that is then either confirmed or adjusted for profitability.

The business risk profile is combined with the financial risk profile to get the anchor rating. When assessing the financial risk profile it is taken into account how the company is funded and how the balance sheet is constructed. For investment-grade anchor ratings a higher weight is put on business risk profile and for speculative-grade a higher weight is put on financial risk profile.

After assigning an anchor, modifiers are taken into account to possibly modify the rating and get the so called stand-alone credit rating. These earlier not considered factors in the categories: diversification, capital structure, financial policy, liquidity, governance, and an overall assessment of these factors. Companies can get a upwards adjustment due to having diversified their business. In the capital structure category factors such as mismatch of its cash flows and sources of financing due to currency and maturity of debt is considered. These factors can result in an upwards as well as downwards adjustment. Upwards or downwards adjustments due to earlier not considered financial policy factors is the impact of the managements preferences for risk. The analysis of liquidity includes for example a quantitative analysis of the the sources and use of cash as well as a qualitative analysis of bank relationships and ability to handle rare events with high impact. Governance modification factors focus on qualitative assessment of the managements, board and owners competence with respect to aspects such as strategic competence and operational effectiveness. Finally, after considering the above mentioned modifiers an overall assessment of the modifications is done to possibly adjust the rating up or down.

A company can be part of a group, for example as a subsidiary, and therefore receive support from the group. For companies being part of a group Standard & Poor's evaluate the credit profile of the group and the likelihood of the company getting help from the group. The likelihood of getting help is evaluated by considering the importance of the company within the group. The stand-alone credit rating can be kept unadjusted or adjusted anywhere up to the group credit profile.

In addition to the above described framework, a stand alone credit profile below B is given if the issuers capital structure is sensitive and the issuer relies on beneficial business, economic, and financial conditions to fulfill its financial obligations.

Chapter 4

Modelling Techniques

In this chapter the two techniques considered for modelling the credit ratings, logistic regression and neural networks are presented. For these models the general setup, how the models are fitted to the data and how they can be used for prediction is described. The chapter also describes how to avoid problems with overfitting and underfitting .

4.1 Logistic Regression

One of the models considered in this thesis is logistic regression. Here the basic logistic model is introduced first, that is a model with a binomial response variable. However, in this study more than two different categories of credit ratings are predicted, therefore the chapter proceeds with explaining how this basic model can be modified to allow for more than two categories for the response. A similar presentation of logistic regression can be seen in Hosmer and Lemeshow (2000).

4.1.1 Logistic Regression with Binary Response

Logistic regression can be used when there is a data set with a binary response variable Y taking on values 0 or 1. Logistic regression models the probability that Y belongs to a given category. That is we model $P_r(Y = 1|X)$, were X is the explanatory variable. In short this is written as $p(X)$. To make sure that the probability falls between 0 and 1 we use the logistic function:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \tag{4.1}$$

In logistic regression the interpretation of the β_1 coefficient is not the same as with linear regression. With linear regression β_1 is the average change when X increases by one unit. For logistic regression a positive β_1 means that an increase of X increases the probability $p(X)$.However, as we do not have a linear relationship between $p(X)$ and X the size of the effect depends on the value of X. (4.1) can be rewritten as:

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X} \tag{4.2}$$

This is the odds, a high value meaning a high probability of $Y = 1$. As an example with odds of 1/4 we will have on average that 1 of 5 $Y = 1$. Increasing X by one unit will multiply the odds by e^{β_1}. By taking the logarithm of 4.2 we get the log odds or logit:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X \tag{4.3}$$

The logit is linear in X and a one unit increase of X does on average increase the logit by β_1.

The unknown parameters $\boldsymbol{\beta} = (\beta_0, \beta_1)$ are estimated using maximum likelihood. Using maximum likelihood the parameters are estimated such that the probability of obtaining the training data is maximized. This is done by writing the probability of obtaining the data as a function of the the parameters to be estimated and choose these such that the function is maximized. This function is called the likelihood function and can be found by using (4.1). As we have that $P_r(Y = 1|X) = p(X)$ and $P_r(Y = 0|X) = 1 - p(X)$, the contribution to the liklihood function of a pair (x_i, y_i) is $p(x_i)^{y_i}[1 - p(x_i)]^{1-y_i}$. As the observations are assumed to be independent the likelihood function can be written as:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} p(x_i)^{y_i}[1 - p(x_i)]^{1-y_i} \tag{4.4}$$

By taking the logarithm of (4.4) we get the log-likelihood function, which is easier to work with. This is possible as the logarithm is a strictly increasing function. Therefore the likelihood function and the log-likelihood function will have the same maximum. The log-likelihood is:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n}\{y_i ln[p(x_i)] + (1 - y_i) ln[1 - p(x_i)]\} \tag{4.5}$$

To find the parameters β_0 and β_1 (4.5) is maximized by differentiating it with respect to β_0 and β_1 and setting these equal to zero. This results in the following equations, called the likelihood equations:

$$\sum_{i=n}^{n}[y_i - p(xi)] = 0 \tag{4.6}$$

$$\sum_{i=1}^{n}[y_i - p(x_i)] = 0 \tag{4.7}$$

The estimated parameters $\hat{\boldsymbol{\beta}} = (\hat{\beta_0}, \hat{\beta_1})$ is the solution to these equations.

With the estimated parameters of the model the probability of an example belonging to the category where $Y = 1$ is predicted to be:

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 X}} \tag{4.8}$$

The logistic regression can be generalized to include a set of p predictors $X = (X_1, X_2, ..., X_p)$, generalizing (4.3) we have:

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p \tag{4.9}$$

This can be rewritten to:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2 ...+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_p X_p}} \tag{4.10}$$

4.1.2 Logistic Regression with More than Two Response Classes

Using multinomial logistic regression is possible when the response variable takes on more than two categories. For ease of exposition the case with three categories will be explained here, but the principle of the multinomial model is the same when there are more categories. In this case the response variable Y will be coded such that it takes on 0, 1 or 2. Similar to (4.3) we form two logits, with $Y = 0$ being the base line outcome. The p explanatory variables and the constant is represented by the column vector $\boldsymbol{x}$ of length 1+p. $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ are column vectors with elements being the parameters in the first and second logit respectively. Then the logits can be written as:

$$g_1(\boldsymbol{x}) = log\left[\frac{P(Y = 1|\boldsymbol{x}}{P(Y = 0|\boldsymbol{x}}\right] = \beta_{10} + \beta_{11} X_1 + \beta_{12} X_2 + ... + \beta_{1p} X_p = \boldsymbol{x'\beta_1} \tag{4.11}$$

$$g_2(\boldsymbol{x}) = log\left[\frac{P(Y=2|\boldsymbol{x}}{P(Y=0|\boldsymbol{x}}\right] = \beta_{20} + \beta_{21}X_1 + \beta_{22}X_2 + ... + \beta_{2p}X_p = \boldsymbol{x'\beta_2} \quad (4.12)$$

Using (4.11) and (4.12) it can be shown that:

$$P(Y=0|\boldsymbol{x}) = \frac{1}{1+e^{g_1(\boldsymbol{x})}+e^{g_2(\boldsymbol{x})}} \quad (4.13)$$

$$P(Y=1|\boldsymbol{x}) = \frac{e^{g_1(\boldsymbol{x})}}{1+e^{g_1(\boldsymbol{x})}+e^{g_2(\boldsymbol{x})}} \quad (4.14)$$

$$P(Y=2|\boldsymbol{x}) = \frac{e^{g_2(\boldsymbol{x})}}{1+e^{g_1(\boldsymbol{x})}+e^{g_2(\boldsymbol{x})}} \quad (4.15)$$

To form the likelihood function indicator variables Y_0, Y_1 and Y_2 are used. Were,

if $Y = 0$ then $Y_0 = 1$, $Y_1 = 0$ and $Y_2 = 0$

if $Y = 1$ then $Y_0 = 0$, $Y_1 = 1$ and $Y_2 = 0$

if $Y = 2$ then $Y_0 = 0$, $Y_1 = 0$ and $Y_2 = 1$

Now letting $p_j(\boldsymbol{x}) = P(Y = j|\boldsymbol{x})$, were $j = 0, 1, 2$ indicates the class and $\boldsymbol{\beta'} = (\boldsymbol{\beta_1'}, \boldsymbol{\beta_2'})$ is a vector holding the parameters of the model, the likelihood function can be written as:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n}[p_0(x_i)^{y_{0i}}p_1(x_i)^{y_{1i}}p_2(x_i)^{y_{2i}}] \quad (4.16)$$

Taking the logarithm of (3.16) gives the log-likelihood function which can be written as:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_{1i}g_1(\boldsymbol{x_i}) + y_{2i}g_2(\boldsymbol{x_i}) - ln(1 + e^{g_1(\boldsymbol{x_i})} + e^{g_2(\boldsymbol{x_i})}) \quad (4.17)$$

Taking the partial derivatives of the log-likelihood function with respect to each of the unknown parameters and setting these equal to zero we get the likelihood functions. letting $p_{ji} = p_j(x_i)$ these can be written as:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_{jk}} = \sum_{i=1}^{n} x_{ki}(y_{ji} - p_{ji}) = 0, \ j = 1, 2 \ and \ k = 0, 1, 2, ..., p \quad (4.18)$$

As in the case with a binary response variable the estimated parameters $\hat{\boldsymbol{\beta}}$ is the solution to these equations.

4.2 Artificial Neural Networks

The second technique considered is a backpropagation neural network with one hidden layer. In this section this technique is descibed in a similar way to Hastie, Tibshirani, and Friedman (2009).

4.2.1 Network Structure

A neural network can be represented by a network diagram as in Figure 4.1. The method got the name "neural network" as it first was developed as a model of the human brain. The circles in Figure 4.1 are called units or neurons and the connections can be thought of as synapses. The first layer is the input layer with p inputs X_ℓ. The middle layer is called the hidden layer and consists of M units Z_m. The final layer is the output layer with K units Y_k, where each unit represents a class to be predicted.

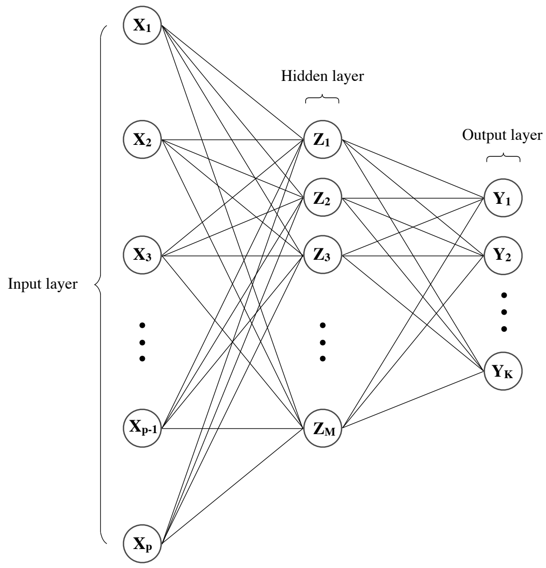

Figure 4.1: Neural network diagram.

The derived features in the hidden layer are modeled as a function of linear combinations of the input variables, that is:

$$Z_m = \sigma(\alpha_{0m} + \alpha_m^T X), m = 1, ..., M \tag{4.19}$$

Where, α_m is a vector of weights of length p and α_{0m} is a bias. The bias can be thought of as an additional input feeding into the units in the hidden layer. $\sigma(v)$ is called the activation function and gives the activation of the units in the hidden layer. The activation function is usually chosen to be the logistic function, that is $\sigma(\upsilon) = 1/(1 + e^{-\upsilon})$. The logistic function is illustrated in Figure 4.2.

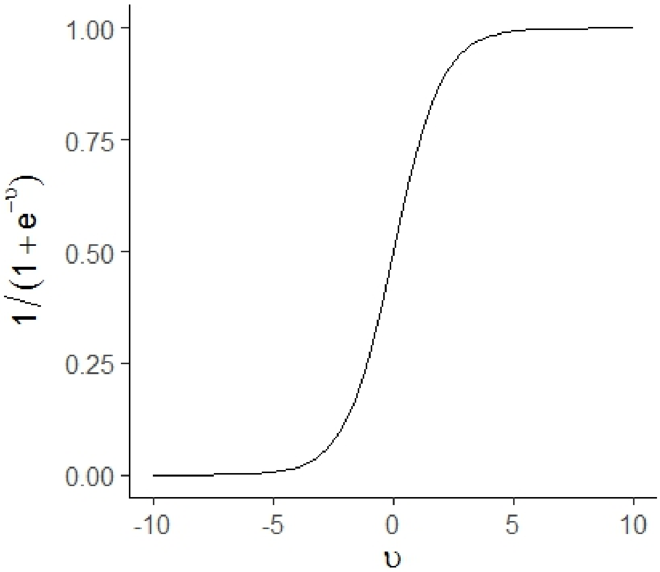

Figure 4.2: The logistic function.

The target Y_k is then modeled as a function of linear combinations of the units Z_m in the hidden layer:

$$T_k = \beta_{0k} + \beta_k^T Z,\ k = 1, ..., K \tag{4.20}$$

$$f_k(X) = g_k(T),\ k = 1, ..., K \tag{4.21}$$

where, β_k is a vector of weights and β_{0k} is a bias. $T = (T_1, T_2, ..., T_K)$, $Z = (Z_1, Z_2, ..., Z_M)$ and $g_k(T)$ is the output function that transforms the vector T to the final output. Here again the logistic function can be chosen as the output function. The unit in the output layer with the highest activation is then the classification of an observation x, that is:

$$G(x) = argmax_k f_k(x) \tag{4.22}$$

4.2.2 Fitting the Neural Network

The unknown parameters of the neural network are the weights and biases. The goal is to find these such that the model fits the training data well. To evaluate how well the model fits the data an error function is used. As error function often the mean squared error is used. With N observations $i = 1, ..., N$, the set of unknown parameters denoted by θ and the true activation denoted by y_{ik}, then the mean squared error function can be written as:

$$R(\theta) = \sum_{k=1}^{K} \sum_{i=1}^{N} (y_{ik} - f_k(x_i))^2 \tag{4.23}$$

The error function is minimized by gradient descent. This involves calculating the gradient of the error function, that is calculating the partial derivatives of the error function with respect to any weight and bias. By making small adjustments to the weights and biases in the opposite direction of the gradient a minimum of the error function can be found. An algorithm that can be used to do this is called backpropagation. To find the equations that are needed for the backpropagation algorithm the chain rule is used. In the case of the mean squared error function the error for an individual training example is:

$$R_i(\theta) = \sum_{k=1}^{K} (y_{ik} - f_k(x_i))^2 \tag{4.24}$$

With an observation x_i we have from (4.19) that $z_{mi} = \sigma(\alpha_{0m} + \alpha_m^T x_i)$ and we let $z_i = (z_{1i}, ..., z_{Mi})$. Using the chain rule the derivatives can be found:

$$\frac{\partial R_i}{\partial \beta_{km}} = -2(y_{ik} - f_k(x_i)) g_k'(\beta_k^T z_i) z_{mi} \tag{4.25}$$

$$\frac{\partial R_i}{\partial \alpha_{m\ell}} = -\sum_{k=1}^{K} 2(y_{ik} - f_k(x_i)) g_k'(\beta_k^T z_i) \beta_{km} \sigma'(\alpha_m^T x_i) x_{i\ell} \tag{4.26}$$

(4.25) can be written as

$$\frac{\partial R_i}{\partial \beta_{km}} = \delta_{ki} z_{mi} \tag{4.27}$$

where

$$\delta_{ki} = -2(y_{ik} - f_k(x_i)) g_k'(\beta_k^T z_i) \tag{4.28}$$

and (4.26) can be written as

$$\frac{\partial R_i}{\partial \alpha_{m\ell}} = s_{mi} x_{i\ell} \tag{4.29}$$

where

$$s_{mi} = -\sum_{k=1}^{K} 2(y_{ik} - f_k(x_i)) g_k'(\beta_k^T z_i) \beta_{km} \sigma'(\alpha_m^T x_i) \tag{4.30}$$

This can be written as

$$s_{mi} = \sigma'(\alpha_m^T x_i) \sum_{k=1}^{K} \beta_{km} \delta_{ki} \tag{4.31}$$

δ_{ki} and s_{mi} are called the errors of the output layer and hidden layer respectively. They measure how far of the units of the current model are.

The backpropagation algorithm can now be explicitly expressed.

1. Set the starting weights and biases θ.
2. Predicted values are computed using (4.19), (4.20) and (4.21). This is called the *Forward pass*.
3. The error in the out put layer δ_{ki} is calculated from (4.28) and then back propagated using (4.31) to get the error in the hidden layer. This is called the *Backward pass*.
4. The partial derivatives are then calculated using (4.27) and (4.29).
5. The weights and biases are updated according to (4.32) and (4.33) below.

The updates of the weights and biases for the r+1 iteration is done according to:

$$\beta_{km}^{r+1} = \beta_{km}^{r} - \gamma_r \frac{\partial R_i}{\partial \beta_{km}^{(r)}} \tag{4.32}$$

$$\alpha_{m\ell}^{r+1} = \alpha_{m\ell}^{r} - \gamma_r \frac{\partial R_i}{\partial \alpha_{m\ell}^{(r)}} \tag{4.33}$$

This process is then repeated, thereby getting closer to a minimum of the error function. As a stopping criteria for the learning process a maximum allowed value for the error function can be used as well as a predefined maximum number of iterations (Zell et al., 1998). γ_r is a small positive number called the learning rate, controlling how big the steps are in each update. Typically the learning rate is chosen to be a number between 0.1 and 1 (Zell et al., 1998). Later in this chapter it will be discussed how this parameter can be optimized. In (4.32) and (4.33) one training example i is processed before the weights and biases are updated. This is called online learning and it is the update function used in this study. Alternatively the updates can be done by batch learning, then all training examples are used before updating the weights and biases. For this the above equations are modified to (4.34) and (4.35).

$$\beta_{km}^{r+1} = \beta_{km}^{r} - \gamma_r \sum_{i=1}^{N} \frac{\partial R_i}{\partial \beta_{km}^{(r)}} \tag{4.34}$$

$$\alpha_{m\ell}^{r+1} = \alpha_{m\ell}^{r} - \gamma_r \sum_{i=1}^{N} \frac{\partial R_i}{\partial \alpha_{m\ell}^{(r)}} \tag{4.35}$$

4.2.3 Starting Weights

In the above describe backpropagation algorithm starting weights need to be set. The starting weights should be set to random values close to zero. With values close to zero the logistic function is close to linear (see 4.2) and a linear model. So the model is initially close to a linear model and increases the size of the weights as nonlinearities are needed. However, starting with exactly zero weights gives zero derivatives and the algorithm does not update the weights. Furthermore, large values tend to give poor solutions (Hastie et al., 2009).

4.3 Overfitting and Underfitting

The models are built with the purpose of making better predictions of the output variables based on the inputs, to know how well this is done the actual values can be compared to the predicted ones. However, when testing the model we want to see if it is *generalizing*, that is that the model performs well on inputs it has not seen before and not only on those used for training. For this reason we need to save data that is not used while training the model, a *test set*. But only using these two data sets can be a problem. Goodfellow, Bengio, and Courville (2016) explain why three data sets are needed. There is a risk of overfitting, that is that the model learns the training data well, but it is not generalizing. We want to control for overfitting as our model learns, this is done using a *validation set*. When there is overfitting the difference between the error of the model on the training set and testing set is large. In addition underfitting occurs when the training set is not learned well, the training error is large. By controlling the models ability to fit a large variety of functions (the models capacity) we can control if it is more likely to overfit or underfit.

4.3.1 Hyperparameters

This ability can be controlled by the settings of the models *hyperparameters*, settings that control the learning algorithm. Examples of hyperparamters that can be tuned in the neural network is the number of hidden units and learning rate. To choose hyper parameters and controlling for how well the model generalizes, an additional data set, the validation set is used. The validation set is not used during training process. The error measuring the genearlization plotted as a function of a hyperparameter generally has a U-shape, at one extreme we have low capacity and underfitting and at the other high capacity and overfitting. Somewhere between these extremes is the model with the optimal capacity, achieving the lowest generalization error. Note that whether a high or low value of the hyperparameter achieves a high or low capacity depends on the hyperparameter. So three sets of data are needed, the training set used for training the model, the validation set to check how well it generalizes while it learns and the test set that is used in the end to test the model when all choices of the model is done including hyperparameters (Goodfellow et al., 2016).

There are several methods that can be used to choose hyperparameters. For example, suitable hyperparameters can be searched for manually, using experience of the perticular application of the model. Another approach is to use an automatic procedure. Here an automtic precedure called *Grid search* will be used. Using Grid search a set of values for each of the hyperparameters to be tuned is selected. Thereafter, all combinations of the hyperparameters are trained on the training set and validated on the validation set. Finally, the set of hyperparameters that gives the lowest validation error is selected (Goodfellow et al., 2016).

4.3.2 K-Fold Cross Validation

when the amount of data is small the above mentioned procedure using a separate validation set can be replaced with what is called *k-fold cross validation*. The training data is split into K folds, where one of them is used for validation and the rest for training of the model. The training and validation is done K times each time using a different fold for validation. The average performance of the K folds is the overall measure of the performance (Zheng, 2015). Here the accuracy will be used as the measure of the performance during the cross validation. The K-fold cross validation can be used to generate the data while using one of the above mentioned methods for tuning the hyperparameters. In this study K-fold cross validation repeated three times will be used together with grid search.

Figure 4.3: K-fold cross validation.

Chapter 5

Evaluation Metrics

In this chapter the methods used to evaluate and compare how well the models are able to predict the credit ratings are described. Also, a measure used to analyze the contribution of each of the inputs in the neural networks is presented.

5.1 Accuracy

Accuracy is a simple measure of the performance of the model that is commonly used in previous research on credit rating predictions. The accuracy is the number of correct predictions divided by the total number of predictions (Zheng, 2015).

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \tag{5.1}$$

5.2 Kappa

Cohens Kappa (Cohen, 1960) was originally developed as a measure of agreement between two judges classifying cases into a set of categories. In the context of this study kappa measures the agreement between the model and the true credit ratings. Instead of only considering the accuracy Kappa takes into account the expected agreement by chance. The expected agreement by chance for the first class is simply the product of the judges proportions classified into the first class. Summing over all classes gives the overall expected agreement by chance. With p_o being the observed accuracy and p_e the expected agreement by chance, $k = 1, ...K$ indicating category, N the number of cases and n_{ik} the number of cases classified into category k by judge i. The kappa is:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{5.2}$$

where,

$$p_c = \frac{1}{N^2} \sum_{k=1}^{K} n_{1k} n_{2k} \tag{5.3}$$

The kappa is the proportion of agreement when chance agreement has been removed. When agreement by chance equals the accuracy $\kappa = 0$. Perfect agreement occurs when $\kappa = 1$. A negative kappa occurs when there is less agreement than by chance.

5.3 The Connection Weight Approach

In most of the previous research on the use of neural networks for credit rating predictions little attention is put on the difference in importance of input variables to the networks and neural networks have often been considered to be a "black box". However, Huang et al. (2004) did analyze the variable importance for neural networks predicting ratings in the US and Taiwan market using Garson's measure.

In this thesis the importance of inputs are analyzed using the connection weight approach suggested by Olden and Jackson (2002). Olden, Joy, and Death (2004) tested the connection weight approach as well as Garson's measure on simulated data sets were the true importance of the variables are known. In their study they showed that the connection weight approach consistently outperforms Garson's measure. The connection weight approach gives a measure of the relative importance/contribution of each input variable to each output category. It is calculated as the product of the input-hidden and hidden-output weights that is between each input and output neuron and summed across all hidden neurons. Having a neural network with J inputs, M hidden units and K outputs. Letting the weights between input j and and neuron m in the hidden layer be denoted w_{mj} and the weights between neuron m in the hidden layer and output k be denoted w_{mk}. Then the importance of input j on output k is:

$$Imp_{jk} = \sum_{m=1}^{M} w_{mj} w_{mk} \tag{5.4}$$

Chapter 6

Data

In this chapter, the data used in this study is described to be as transparent as possible and allow for the results to be validated. First it is described how the raw data was selected and thereafter how the raw data was prepared before it was used for modelling.

6.1 Data Selection

The data used are Standard & Poor's credit ratings and accounting data of companies in the United States. The accounting data was then used to calculate the ratios used as explanatory variables in the models. The financial ratios that were used are ratios that are commonly used in research on credit rating predictions, such as in Huang et al. (2004) and Kumar and Bhattacharya (2006). Two data sets of this type were prepared, one representing manufacturing companies and one representing retail companies.

The data was extracted from Bloomberg using the rating changes function (RATC) and the Excel plugin. As a first step one list of companies classified as manufacturing companies and one list of companies classified as retail companies were created in the Bloomberg terminal. To create these lists Standard Industrial Classification codes (SIC) was used. Companies with SIC codes 2000-3999 are classified as manufacturing companies. For retail companies, the retail trade category with SIC codes 5200-5999 was used.

Thereafter, a list of Standard & Poor's Long-term issuer credit ratings available was obtained for the period 1 January 2013 - 31 December 2017. Then using the Bloomberg Excel plugin the following accounting data from the companies annual reports was obtained:

Table 6.1: Accounting data downloaded from Bloomberg.

Cash Flow Statement	Balance Sheet	Income Statement
Cash flow from operating activities	*Revenue*	*Total assets*
	Total liabilities	*Operating income*
	Current assets	*EBIT*
	Current liabilities	*Interest expense*
	Inventories	*Costs of goods and services sold*
	Long-term debt	*Income tax expense*
	Short-term debt	*Net income*
	Total equity	*Earnings per Share*
	Common equity	
	Cash and equivalents	

6.2 Data Preparation

Some observations were missing accounting data and were therefore removed. Other observations had credit ratings recorded as NR, which stands for not rated, and were also removed. After removing these the manufacturing data set consisted of 604 observations and the retail data set of 143 observations. In addition some of the credit ratings contained additional information to the standard rating categories. This additional information was removed to only focus on the standard rating categories. For example an u suffix indicates an unsolicited rating, that is a credit rating initiated by some one else than the issuer. A + or - indicates the relative standing within a category.

Thereafter, the financial ratios were calculated using the financial variables presented in Table 6.1. By using financial ratios instead of accounting data directly the input variables are comparable across companies of different size. However, total assets and total liabilities were kept to measure the pure effect of size. The full set of 23 explanatory variables are (For definitions of these see Appendix A.1):

Total assets	Net profit margin	Asset turnover
Total liabilities	Pretax margin	Earnings per share
Debt ratio	Operating margin	Cash flow from operations to current liabilities
Current ratio	Gross profit margin so	Inventory to current assets
Quick ratio	Operating profitability	Total equity to total assets
Current assets to total assets	Return on assets before interest and tax	Sales to net worth
Solvency ratio	Return on assets after interest and tax	Cash to total assets
Return on equity	Debt to equity	

Summary statistics of the input variables in the manufacturing data set and retail data set are shown in Table 6.2 and Table 6.3 respectively.

Table 6.2: Summary statistic for the manufacturing data set.

Statistic	Mean	St. Dev.
TOTAL_ASSETS	15,271.830	44,315.870
TOTAL_LIABILITIES	9,678.013	30,590.890
EPS	1.679	3.947
DEBT_RATIO	0.339	0.192
OPERATING_MARGIN	0.081	0.148
RETURN_ON_ASSETS_BEFORE_INTEREST_AND_TAX	0.071	0.085
RETURN_ON_ASSETS_AFTER_INTEREST_AND_TAX	0.039	0.085
PRETAX_MARGIN	0.061	0.169
NET_PROFIT_MARGIN	0.046	0.157
CURRENT_RATIO	2.260	1.177
QUICK_RATIO	1.591	1.009
CURRENT_ASSETS_TO_TOTAL_ASSETS	0.399	0.159
INVENTORIES_TO_TOTAL_ASSETS	0.301	0.153
TOTAL_EQUITY_TO_TOTAL_ASSETS	0.358	0.239
OPERATING_PROFITABILITY	0.081	0.148
GROSS_PROFIT_MARGIN	0.334	0.201
SOLVENCY_RATIO	0.358	0.239
SALES_TO_NET_WORTH	3.196	15.970
CASH_TO_TOTAL_ASSETS	0.094	0.081
DEBT_TO_EQUITY	0.967	6.097
ASSET_TURNOVER	0.929	0.520
RETURN_ON_EQUITY	0.114	0.723
CASH_FLOW_OPERATIONS_TO_CURRENT_LIABILITIES	0.591	0.498

Table 6.3: Summary statistics for the retail data set.

Statistic	Mean	St. Dev.
TOTAL_ASSETS	8,853.900	14,864.030
TOTAL_LIABILITIES	6,256.585	10,253.900
EPS	0.370	7.109
DEBT_RATIO	0.405	0.272
OPERATING_MARGIN	0.052	0.103
RETURN_ON_ASSETS_BEFORE_INTEREST_AND_TAX	0.089	0.116
RETURN_ON_ASSETS_AFTER_INTEREST_AND_TAX	0.039	0.095
PRETAX_MARGIN	0.037	0.097
NET_PROFIT_MARGIN	0.019	0.085
CURRENT_RATIO	1.714	0.870
QUICK_RATIO	0.793	0.702
CURRENT_ASSETS_TO_TOTAL_ASSETS	0.402	0.179
INVENTORIES_TO_TOTAL_ASSETS	0.520	0.253
TOTAL_EQUITY_TO_TOTAL_ASSETS	0.247	0.276
OPERATING_PROFITABILITY	0.052	0.103
GROSS_PROFIT_MARGIN	0.336	0.135
SOLVENCY_RATIO	0.247	0.276
SALES_TO_NET_WORTH	17.125	110.763
CASH_TO_TOTAL_ASSETS	0.075	0.074
DEBT_TO_EQUITY	2.656	22.470
ASSET_TURNOVER	1.745	1.032
RETURN_ON_EQUITY	−0.258	6.942
CASH_FLOW_OPERATIONS_TO_CURRENT_LIABILITIES	0.528	0.447

Finally these inputs were scaled to have a mean of zero and a standard deviation of one. This was done to have a range of the starting weights in the neural networks that are meaningful (Hastie et al., 2009).

6.3 Grouping of Ratings

For the two data sets, manufacturing and retail, there is imbalance in the rating categories, see Figure 3.1. Due to his, first a broad grouping of the categories into investment grade and non-investment grade was made. Standard & Poor's ratings of BBB or above are investment grade and all below are non-investment grade. Classification into two broad categories have been done by several previous studies, making it interesting for comparison. However, it is also interesting to see how well the models classify rating categories that are close. As seen in Figure 3.1 the majority of the ratings belong to the categories: BBB, BB and B. Subsetted data sets only including these categories were formed for classification into these three categories. The data set for manufacturing then consisted of 500 observations and the data set for retail of 110 observations.

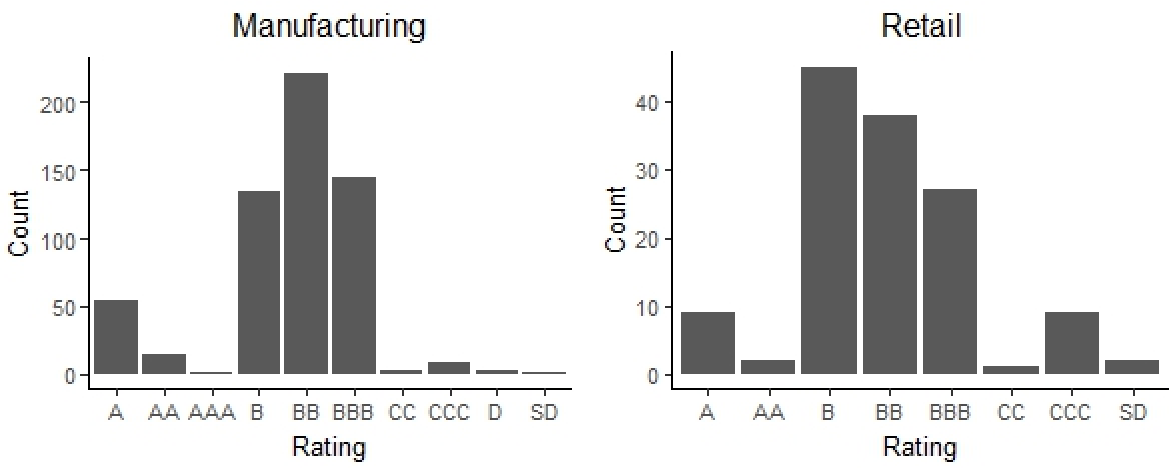

Figure 6.1: The distribution of ratings in the data sets

6.4 Training and Test Set

Finally, the data was split into a training set and a test set. The training set is used for training and cross validation while the test set is saved for testing of the final models. The data was split into 70% training and 30% testing. To preserve the overall rating distribution observations were randomly selected as trainig and testing within each rating category.

Chapter 7

Methodology

In this chapter, first the programming language and packages used are described. Then the steps taken when building and evaluating the models are presented.

7.1 Software

After obtaining the raw data set from Bloomberg using the Excel plugin, data preparation, modelling and evaluation was done using the statistical programming language R (R Core Team, 2017). The R packages *caret* (Kuhn, 2018), *nnet* (Venables & Ripley, 2002) and *RSNNS* (Bergmeir & Benítez, 2012) were used to build and evaluate the models. The caret package contain different functions that makes the modelling process efficient and allows for the use of a common syntax for different models. Moreover, the caret package has functions for cross validation and parameter tuning. For modelling of neural networks caret was used together with the RSNNS package. RSNNS provides an R interface to the Stuttgart Neural Network Simulator (SNNS) (Zell et al., 1998), that allows for building, training and testing of neural networks. The modelling with logistic regression was done using the caret package together with the function for logistic regression in the nnet package.

7.2 Hyperparameters and Model Training

The trained neural networks are single hidden layer networks. That is neural networks with three layers, a input layer, a hidden layer and a output layer. Two hyperparameters were tuned for the neural network, the nummber of hidden untis and the learninig rate. The number of hidden units in the neural network depends on things such as the number of inputs and output. One rule of thumb is that the number of hidden units should be somwhere between the size of the input layer and the output layer (Blum, 1992). The learning rate is choosen to be some small positive number, and as noted by Zell et al. (1998) typicall numbers are 0.1,...1.0. To find the hyperparameters, grid search was used together with 10-fold cross validation repeated three times, where the hyperparameters corresponding to the model with the highest accuracy was selected. Using grid search, a grid for the hyperparameters is defined. This grid is a set of hyperparmeters that are considered. The set of hidden units considerd were all integer values from number of outputs to the number of inputs. For the learning rate values considered are 0.1,...1.0.

After the hyperparameters being selected the neural network was trained using the whole training data set. When training the neural networks, backpropagation was used as learning algorithm and the logistic function as activation function. The starting weights for the backpropagation algorithm were randomized values close to zero. More precisely the default interval [-0.3, 0.3] of the Stutgart Neural Network Simulator was used. Furthermore, for the maximum number of allowed iterations while training and maximum training error the default options of 100 and 0 respectively were used.

For logistic regression no hyperparameters were tuned, the model was directly trained on the full training data.

With two different modelling techniques, two industries and two types of classification, there are eight models in total as described in the table below:

Table 7.1: Summary of trained models.

Model Name	Method	Industry	Classification
Model 1	Neural Network	Manufacturing	Investment grade / non-investment grade
Model 2	Neural Network	Manufacturing	BBB, BB and B
Model 3	Logistic Regression	Manufacturing	Investment grade / non-investment grade
Model 4	Logistic Regression	Manufacturing	BBB, BB and B
Model 5	Neural Network	Retail	Investment grade / non-investment grade
Model 6	Neural Network	Retail	BBB, BB and B
Model 7	Logistic Regression	Retail	Investment grade / non-investment grade
Model 8	Logistic Regression	Retail	BBB, BB and B

7.3 Model Evaluation and Analysis

After training both the logistic regression and neural network models they were evaluated on the test set. The models were evaluated using both the accuracy of the models as well kappa. Furthermore, the impact of different input variables for the neural network models were analyzed using the connection weight approach.

Chapter 8

Modelling Results and Analysis

In this chapter, the results from following the steps described in chapter 7 are presented and analyzed. First the predictive ability of the different models are compared using the accuracy and kappa. Finally, the neural networks fitted on the two industries are further analyzed using connection weights.

8.1 Manufacturing

For the model classifying the ratings into investment grade and non-investment grade, the logistic regression model achieved an accuracy of 83 % while the neural network achieved an accuracy of 86 %. Both models performed about equally well and the accuracy is similar to what Dutta and Shekhar (1988) presented for a neural network predicting two classes (83 %). Comparing the kappas of the models reveals a similar result, with the logistic regression having a value of 0.63 and the neural network a slightly higher value of 0.69.

When predicting three categories (BBB, BB and B) the neural network again performed slightly better than the logistic regression model. The logistic regression achieved an accuray of 62 % on the test set and the neural network 64 %. The kappa of the logistic regression is 0.41 and 0.4488 for the neural network.

8.2 Retail

Logistic regression achieved an accuracy of 92 % and the neural network 95 % for the investment grade\non-investment grade grouping. The kappas is 0.82 and 0.87 for the logistic regression and the neural network respectively. Again the neural network performed better than the logistic regression on the test set. However, the evaluation metrics were also in this case close.

For the grouping into three categories the results were similar. The neural network achieved a higher accuracy of 75 % while the accuracy for the logistic regression was 66 %. Also in terms of kappa the neural network with a value of 0.62 is slightly better than the logistic regression with a value of 0.49.

8.3 Comparison

For both industries the neural network has a higher accuracy and kappa. It seems that a neural network better predicts the credit ratings independent of industry. However, it should be noted that the simpler logistic regression is close to the neural networks in terms of accuracy and kappa.

It is interesting to note that the accuracy and kappa for both the logistic regression and the neural network is much higher for the retail firms than the manufacturing firms. One explanation for this can be that there is more quantifiable information in credit ratings for retail firms than for manufacturing firms. The credit rating agencies does not only use quantitative information, but also qualitative. The qualitative information is harder to capture in models such as neural networks and logistic regression. It is also possible that the firms in the retail data set are easier to differentiate, because of larger differences between firms. The results for the neural networks and logistic regressions are summarized in Table 8.1 and Table 8.2 respectively.

Table 8.1: Summary of neural network models

Industry	Classification	Accuracy	Kappa	Learning rate	Hidden units
Manufacturing	Investment grade / non-investment grade	0.8571	0.6931	0.1	10
Manufacturing	BBB, BB and B	0.6443	0.4488	0.2	19
Retail	Investment grade / non-investment grade	0.9487	0.8734	0.2	10
Retail	BBB, BB and B	0.7500	0.623	1.0	11

3

Table 8.2: Summary of logistic regression models

Industry	Classification	Accuracy	Kappa
Manufacturing	Investment grade / non-investment grade	0.8343	0.6318
Manufacturing	BBB, BB and B	0.6242	0.4060
Retail	Investment grade / non-investment grade	0.9231	0.8152
Retail	BBB, BB and B	0.6562	0.4899

8.4 Variable Importance

In this section the neural networks with three credit rating categories are analyzed. Using Olden's connection weight measure of variable importance. It is analyzed which input variables that are the most important in general for prediction. Also, the differences between the two industries are further analyzed to see if there are differences in the input variables that are the most and least important. The resulting connection weights for the neural network fitted on the manufacturing data are given in Figure 8.1 to 8.3 and for the retail data in Figure 8.4 to 8.6.

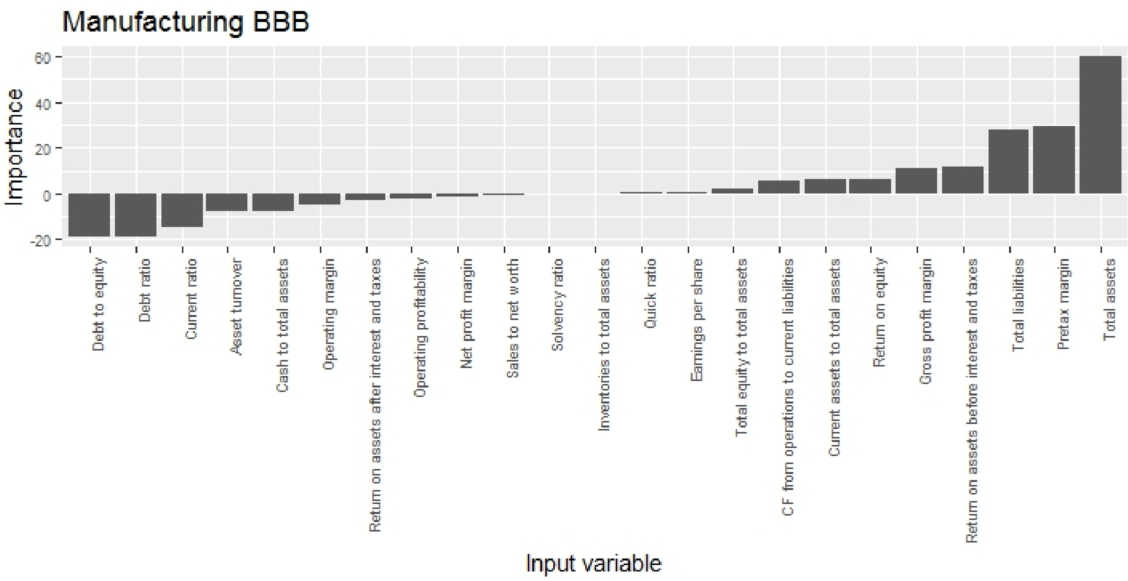

Figure 8.1: BBB class connection weights for three category neural network model on manufacturing data.

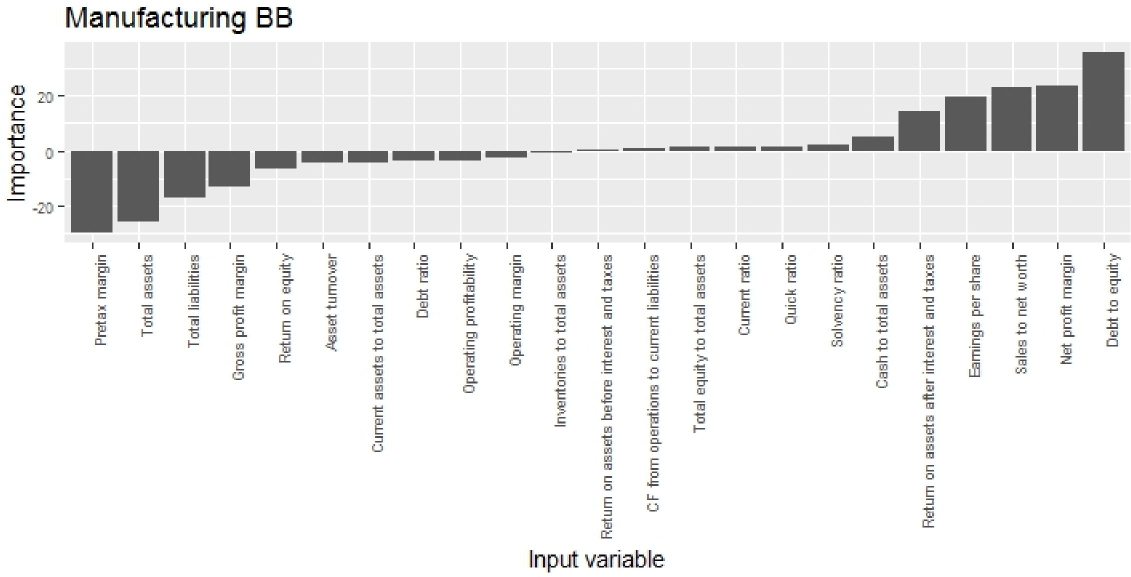

Figure 8.2: BB class connection weights for three category neural network model on manufacturing data.

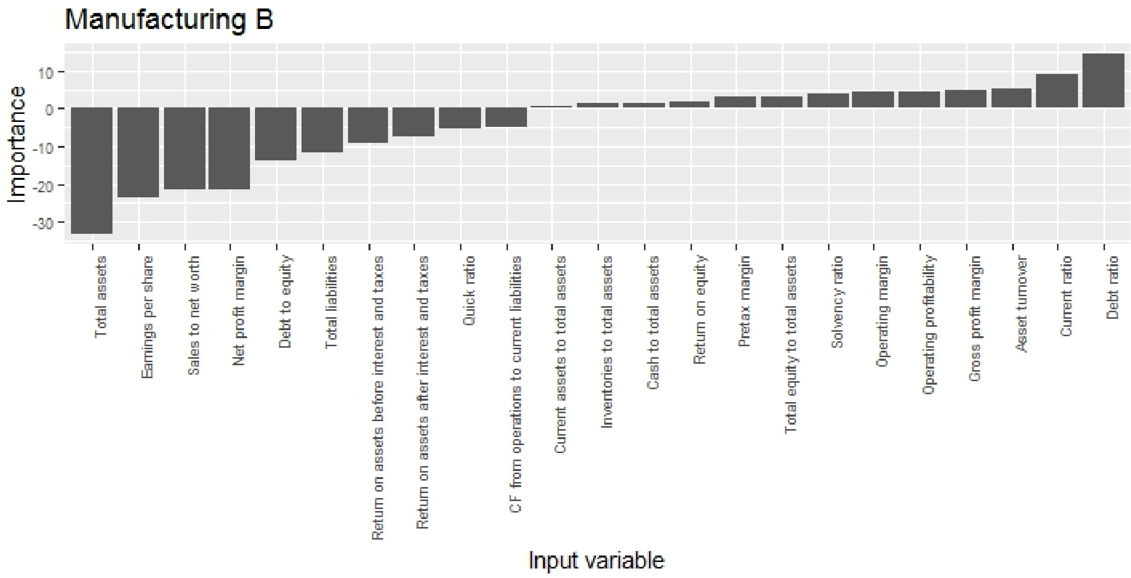

Figure 8.3: B class connection weights for three category neural network model on manufacturing data.

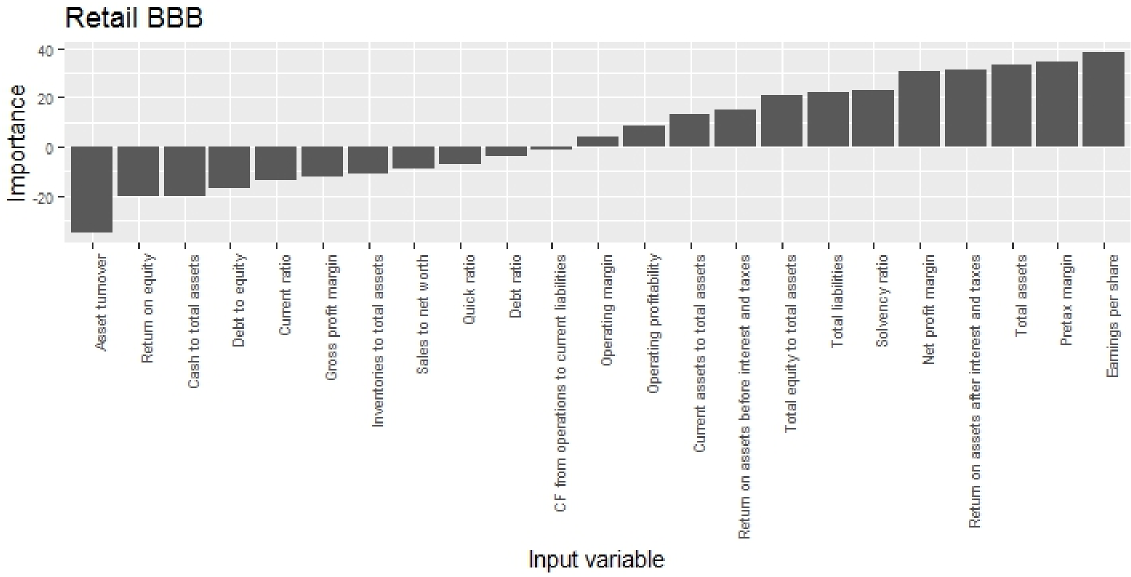

Figure 8.4: BBB class connection weights for three category neural network model on retail data.

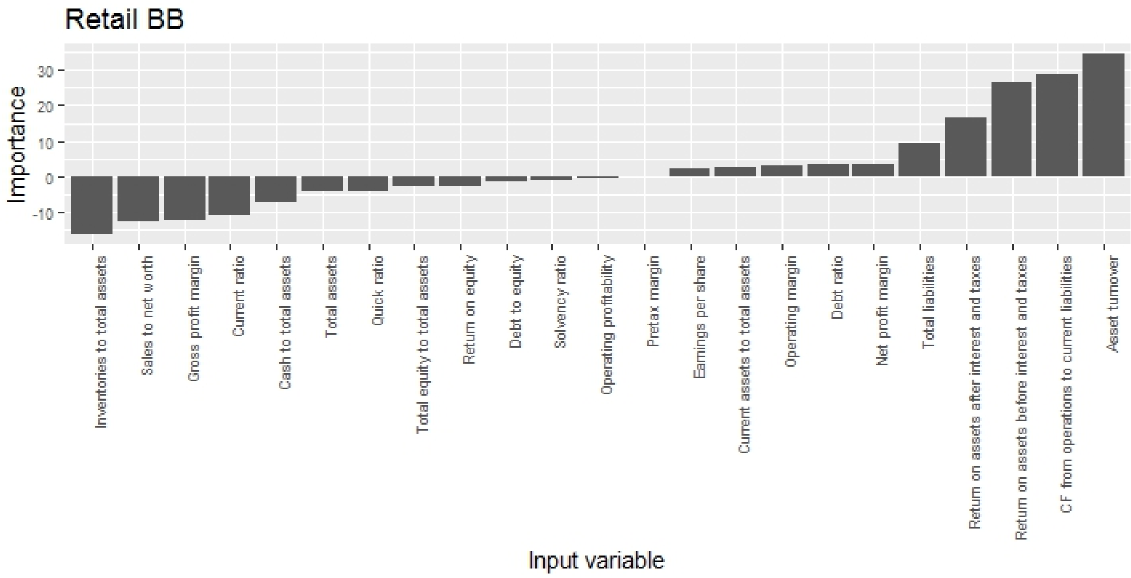

Figure 8.5: BB connection weights for three category neural network model on retail data.

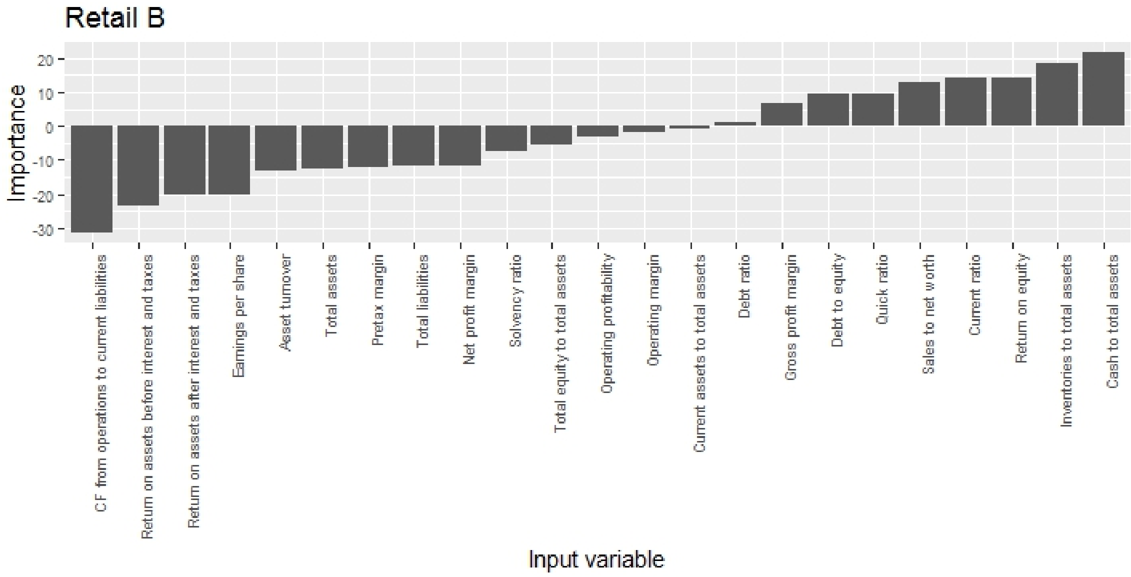

Figure 8.6: B class connection weights for three category neural network model on retail data.

Input variables that are contributing allot to one of the output classes also tend to do so for one or both of the other classes. For example, for the manufacturing firms total assets is important across all rating classes. Other important variables for the manufacturing firms are pretax margin, debt to equity and net profit margin.

For retail firms, asset turnover contributes the most for the BB class and second most for the BBB class among all variables. Another important input variable is cash flow from operations to current liabilities, it contributes most for the B class and second most in the BB class. Return on asset, both before and after tax, are also important variables.

Comparing the model for the manufacturing firms with the one of retail firms, shows some similarities and differences. The relative importance of the total assets is especially high in the manufacturing model. It is not among the least important in the retail model, but its relative importance in the retail model tends to be lower. This suggests that the size of the firm is important in the rating process, and especially for manufacturing firms.

As mentioned, cash flow from operations to current liabilities is an important variable for the retail model. While, it is not an important variable in the manufacturing model. The short term liquidity seems to be especially important in the rating process of retail firms. In general it is good for a firm to have a high cash flow from operations to current liabilities ratio. However, a low ratio is not necessarily a warning. A firm can have a lower ration due to investments that will generate more cash in the long run. Differences in the ratio for manufacturing firms might more often be due to things such as temporary investments and therefore not be as valuable information in the rating process.

Inventories to total assets is a fairly important variable for the retail firms. It contributes fifth most to the BB class and sixth most to the B class. But it is not important for the manufacturing firms, it is the 23 most important for BBB (least important), 22 for the B class and 21 for the BB class. The inventories to total assets is likely a more important factor for retail firms because they hold more inventories in general and there being bigger differences between firms. For a retail firm it is important to not hold too much inventory. The average inventories to total assets for the firms in the retail data set is 52 % and 32 % for the firms in the manufacturing data set.

There are also similarities between the two models. Earnings per share is a important variable for both. Earning per share contributes the most to the BBB class and fifth most to the B class for the retail model. Earnings per share also contributes the second most to the B class and sixth most to the BB class for the manufacturing model.

Chapter 9

Conclusion and Further Work

The main conclusions from this study will now be discussed and some guidance for further research will be given.

The use of neural networks and logistic regression for predictions of credit ratings of manufacturing firms and retail firms have been studied. The neural networks consistently outperforms the logistic regression in terms of the evaluation metrics considered. Independent of industry the neural network is better at predicting the credit ratings than the reference model logistic regression. This result is in line with previous studies investigating the use of machine learning for prediction of credit ratings. The models show that a large part of credit ratings can be explained by publicly available accounting data. Using two output classes the neural networks achieved an accuracy of 86 % and 95 % for the manufacturing and retail data set respectively. In the more demanding challenge of classifying into three neighboring classes the neural networks achieved an accuracy of 64 % and 75 % for the manufacturing and retail data set respectively. The usefulness of machine learning in explaining credit ratings is interesting to better understand the rating process, cutting costs and providing a way of giving an artificial rating for unrated firms. Furthermore, it is of interest to the credit rating agencies themselves. There is promise for the usefulness of machine learning techniques as a tool in the rating process.

Previous studies using neural networks for credit rating predictions tend to solely focus on the predictive performance of the networks. In this study the neural networks were further analyzed using Olden's connection weight approach. This allowed for a industry comparative study and a better understanding of the bond rating process. The total assets was found to be a especially important variable for predicting the credit ratings of the manufacturing firms. For the network fitted on the retail data set, cash flow from operations to current liabilities is an important variable, but not so for the network fitted on the manufacturing data set. Other variables such as earnings per share is among the most important variables for both the network fitted on the manufacturing data set and retail data set.

In future studies it would be interesting to further study the importance of the input variables to the neural network. In this study the ratings of Standard & Poor's were used, further research using ratings from other credit rating agencies is needed to see if the results hold for other ratings than Standard & Poor's. Moreover, a different set of input variables can be used. For example also including different types of non accounting data might give even better performance and new insight into the credit rating process. With data getting more accessible there is hope for finding models with better inputs and larger data sets can be used to get more reliable results.

Appendix A

Appendix

A.1 Input Variables

1. $Total\ Assets$
2. $Total\ Liabilities$
3. $Earnings\ Per\ Share$
4. $Debt\ Ratio = \frac{Long-Term\ Debt+Short-Term\ Debt}{Total\ Assets}$
5. $Operating\ Margin = \frac{Operating\ Income}{Revenue}$
6. $Return\ on\ Assets\ Before\ Interest\ and\ Taxes = \frac{Earnings\ Before\ Interest\ and\ Taxes}{Total\ Assets}$
7. $Return\ on\ Assets\ After\ Interest\ and\ Taxes = \frac{Net\ Income}{Total\ Assets}$
8. $Pretax\ Margin = \frac{Net\ Income+Tax\ Expense}{Revenue}$
9. $Net\ Profit\ Margin = \frac{Net\ Income}{Revenue}$
10. $Current\ Ratio = \frac{Current\ Assets}{Current\ Liabilities}$
11. $Quick\ Ratio = \frac{Current\ Assets-Inventories}{Current\ Liabilities}$
12. $Current\ Assets\ to\ Total\ Assets = \frac{Current\ Assets}{Total\ Assets}$
13. $Inventories\ to\ Total\ Assets = \frac{Inventories}{Total\ Assets}$
14. $Total\ Equity\ to\ Total\ Assets = \frac{Total\ Equity}{Total\ Assets}$
15. $Operating\ Profitability = \frac{Earnings\ Before\ Interest\ and\ Taxes}{Total\ Revenue}$
16. $Gross\ Profit\ Margin = \frac{Revenue-COGS}{Total\ Revenue}$

17. $Solvency\ Ratio = \frac{Net\ Worth}{Total\ Assets}$

18. $Sales\ to\ Net\ Worth = \frac{Revenue}{Net\ Worth}$

19. $Cash\ to\ Total\ Assets = \frac{Cash\ and\ Equivalents}{Total\ Assets}$

20. $Debt\ to\ Equity = \frac{Short-Term\ Debt+Long-Term\ Debt}{Total\ Equity}$

21. $Asset\ Turnover = \frac{Revenue}{Total\ Assets}$

22. $Return\ on\ Equity = \frac{Net\ Income}{Total\ Assets-Total\ Liabilities}$

23. $Cash\,Flow\,From\,Operations\,to\,Current\ Liabilities = \frac{Cash\ Flow\ From\ Operations}{Current\ Liabilities}$

References

Ackerloff, G. (1970). The market for lemons: Quality uncertainty and the market mechanism. *Quarterly journal of economics*, *84*(3), 488–500.

Adelson, M. H. (2012). The role of credit ratings in the financial system. *Standard & Poor's*.

Bergmeir, C., & Benítez, J. M. (2012). Neural networks in R using the stuttgart neural network simulator: RSNNS. *Journal of Statistical Software*, *46*(7), 1–26. Retrieved from `http://www.jstatsoft.org/v46/i07/`

Blum, A. (1992). *Neural networks in c++: An object-oriented framework for building connectionist systems*. New York, NY, USA: John Wiley & Sons, Inc.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37–46.

Dutta, & Shekhar. (1988). Bond rating: a nonconservative application of neural networks. In *Ieee 1988 international conference on neural networks* (Vol. 2, p. 443-450).

Ederington, L. H. (1985). Classification models and bond ratings. *Financial review*, *20*(4), 237–262.

Fisher, L. (1959). Determinants of risk premiums on corporate bonds. *Journal of Political Economy*, *67*(3), 217–237.

Garson, G. D. (1991). Interpreting neural-network connection weights. *AI Expert*, *6*(4), 46–51.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. (`http://www.deeplearningbook.org`)

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction, second edition*. Springer.

Horrigan, J. O. (1966). The determination of long-term credit standing with financial ratios. *Journal of Accounting Research*, *4*, 44–62.

Hosmer, D., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). Wiley-Interscience.

Huang, Z., Chen, H., Hsu, C.-J., Chen, W.-H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision support systems*, *37*(4), 543–558.

Hunt, J. P. (2009). Credit rating agencies and the worldwide credit crisis: the limits of reputation, the insufficiency of reform, and a proposal for improvement. *Columbia Business Law Review*, *2009*, 109.

Kuhn, M. (2018). caret: Classification and regression training [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=caret` (R package version 6.0-80)

Kumar, K., & Bhattacharya, S. (2006). Artificial neural network vs linear discriminant analysis in credit ratings forecast: A comparative study of prediction performances. *Review of Accounting and Finance*, *5*(3), 216–227.

Langohr, H., & Langohr, P. (2010). *The rating agencies and their credit ratings: what they are, how they work, and why they are relevant* (Vol. 510). John Wiley & Sons.

McLean, B., & Nocera, J. (2011). *All the devils are here: The hidden history of the financial crisis.* Penguin.

Olden, J. D., & Jackson, D. A. (2002). Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological modelling*, *154*(1-2), 135–150.

Olden, J. D., Joy, M. K., & Death, R. G. (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling*, *178*(3-4), 389–397.

Pinches, G. E., & Mingo, K. A. (1973). A multivariate analysis of industrial bond ratings. *The journal of Finance*, *28*(1), 1–18.

R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Standard & Poor's. (2014). *S&P global ratings definitions.*

Standard & Poor's. (2015). *Guide to credit rating essentials: What are credit ratings and how do they work?*

Standard & Poor's. (2018). *How standard & poor's rates nonfinancial corporate entities.*

Surkan, A. J., & Singleton, J. C. (1990). Neural networks for bond rating improved by multiple hidden layers. In *1990 ijcnn international joint conference on neural networks* (Vol. 2, p. 157-162).

U.S. Securities and Exchange Commission. (2017). *Annual report on nationally recognized statistical rating organizations.*

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth ed.). New York: Springer.

West, R. R. (1970). An alternative approach to predicting corporate bond ratings. *Journal of Accounting Research*, *8*(1), 118–125.

Zell, et al. (1998). Snns, stuttgart neural network simulator, user manual v. 4.2. *University of Stuttgart and University of Tübingen.* Retrieved from `http://www.ra.cs.uni-tuebingen.de/SNNS/`

Zheng, A. (2015). *Evaluating machine learning models.* O'Reilly Media, Inc.