Predicting Corporate Credit Ratings with Machine Learning Algorithms

Ola Berger Bungum (40730)

A thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science in Finance
Stockholm School of Economics

Abstract

This thesis investigates the performance of machine learning models in predicting long-term issuer credit ratings, relative to the that of traditional statistical modeling approaches. Our dataset consists of 3,992 ratings by S&P, Moody's and Fitch of American non-financial, non-governmental companies, in the period 1 January 2010 through 1 September 2016. 20% of the dataset is used strictly as an out-of-sample set, in order evaluate the models' performance. We find that our best-performing machine learning model, the ExtraTrees algorithm, achieves an accuracy of 37% when predicting over 16 classes – significantly better than our highest performing statistical method, multiple discriminant analysis, which had 27% accuracy. When predicting over 6 and 2 separate classes, the best-performing models achieved accuracies of 70% and 92%, respectively. These results are in line with previous research on the topic, but are the results of training on a significantly larger dataset. Whereas our results, and past studies show that a relatively high degree of accuracy is possible, the specific implications and possible applications are still unclear.

Acknowledgements

I want to thank Michael Halling, Associate Professor at SSE, for kindly supervising this thesis, providing his time and input.

Supervisor: Michael Halling, Associate Professor, Stockholm School of Economics
Keywords: Credit ratings, machine learning, rating agencies, credit risk
JEL Classification: G11, G24, C63

Contents

List of Tables

List of Figures

List of Abbreviations

- CRA - Credit Rating Agency
- NRSRO - Nationally Recognized Statistical Rating Organization
- SEC - U.S. Securities and Exchange Commision
- IOSCO - International Organization of Securities Commissions
- ANN - Artificial Neural Networks
- ANOVA - Aanalysis of Variacnce
- BPN - Backpropagating Neural Netork
- PCA - Principal Component Analysis
- PC - Princpal Component
- MSE - Mean Square Erorr
- MDA - Multiple Discriminant Analysis
- SEC - Securities and Exchange Commission
- DT - Decision Tree
- SVM - Support Vector Machines
- SGD - Stochastic Gradient Descent
- OLS - Ordinary Least Squares
- DT - Decision Trees
- LDA - Linear Discriminant Analysis
- MDA - Multiple Discriminant Analysis
- LR - Logistic Regression
- GP - Gaussian Processes

1 Introduction

This chapter introduces the research topic, credit ratings, machine learning, and the relevance of credit rating modeling. We explain the motivation for choosing the research topic, define the thesis' central research question and specify the objective of the study. Further we briefly describe the methodology that we will use to answer the research question. Finally, this chapter concludes with a high-level summary of structure of the thesis and its chapters.

1.1 Background and motivation

For any investment or financial transaction, risk is a major consideration and affects the decisions of investors, lenders, borrowers and other market participants (IOSCO, 2003). In corporate credit markets, credit ratings have become one of the primary references for financial institutions and other investors to assess credit risk (IOSCO, 2003). Credit ratings are the opinion of a rating agency about the credit quality of a bond issuer or a particular debt security, summarized as a grade according to a predefined scale (Bennell et al., 2006).

Credit rating agencies (CRAs), such as Standard & Poor's Financial Services (S&P), Moody's Investor Service (Moody's) and Fitch Ratings (Fitch) are some of the providers of credit ratings. CRAs are specialized companies, with the resources and knowledge to gather and analyse the data needed to assess credit quality (Bennell et al., 2006).

Typically, if the risk of lending to a borrower is high, investors will require a higher compensation. This compensation usually comes in the form of higher effective interest rates (Surkan and Singleton, 1990), but can even take other shapes, such as more lender-favourable terms or rights (Frank, 2009). Given the recognition and extensive use of credit ratings by the financial industry and regulators, CRAs are instrumental in determining the cost of borrowing for debt security issuers, giving them an influential position in financial markets (Maher and Sen, 1997).

The CRAs state that both quantitative and qualitative factors are considered in their assessment and use public. Both company-published and public domain information, as well as proprietary information, such as company-provided data and information from meetings with the issuer's management (Standard & Poor's Financial Services LLC, 2016).

There are many CRAs, but the largest ones are S&P, Moody's, and Fitch (Frank, 2009). However, obtaining a rating from any of these CRAs is both costly and time-consuming, as such an analysis is performed by experts (Hajek and Michalak, 2013). This expense translates into lower profit for the company and its investors. Furthermore, CRAs have been heavily criticised for issuing misleading and untimely ratings. The CRAs' role in exacerbating the 2008 financial crisis, and scandals such as Enron's bankruptcy in 2001, serves as stern reminders not to trust credit ratings blindly. In the specific case of Enron, it had an investment-grade credit rating up until five days before it filed for bankruptcy

(White, 2010). Lastly, CRAs have faced criticism for the opaqueness of their methodology and limited information about which variables are considered.

One way of lowering the costs, improve the timeliness of credit ratings, and create more transparency, could be to automate the rating procedure. Within academic literature, there has been extensive research on attempting to model credit ratings using publicly available information and a variety of statistical techniques. More recently, using machine learning techniques for this purpose have been studied in academic literature.

Machine learning has become increasingly pervasive in many disciplines and applications, and have gained substantial attention from academia and the industry. Whereas classical statistical methods are focused on theory-driven hypothesis testing, machine learning has a more data-driven approach to modelling. Furthermore, machine learning models often have relaxed assumptions on the structure of the data and can model complex non-linear relationships. However, machine learning models are often criticised for being 'black-boxes', from which it is hard to derive meaningful economic implications. Thus, as a theoretical tool, machine learning models are limited in what they can tell us about the underlying structure of the data.

Machine learning models are used for a multitude of tasks, but are frequently used for classification problems. Within finance, credit ratings is one of the most apparent classification problems. There has been numerous studies attempting to model credit ratings using machine learning models, with varying degrees of success. Modelling credit ratings is a relevant topic both for academic understanding and for the financial industry. Whereas machine learning models will likely not replace the role of CRAs, they do present an interesting topic. In particular, it allows us to examine how much information publicly available accounting data holds about credit ratings. In turn, this also allows us to assess the value that CRAs provide in the act of rating a company. Furthermore, it could serve as a lower-cost method of assessing credit risk, or even for predicting credit rating changes and corresponding bond yield changes.

The research topic could also be of great interest to the credit rating industry, as machine learning could help CRAs automate some of their processes, saving costs, or provide valuable information to their analysts, which could aid them in making more informed decisions, increasing the quality of the ratings.

1.2 Research problem and goal

Credit ratings have been modelled using various statistical and machine learning techniques. These attempts have done so with varying degrees of accuracy, but seem to exhibit the trend that machine learning models show greater promise in modelling ratings with high accuracy.

Generally speaking, two types of ratings exist; issuer ratings and issue ratings. Key to credit ratings is the default risk - in short, the capacity and willingness of the bond issuer to meet its financial commitments on a timely basis (Huang et al., 2004).

However, most of the literature on the subject has limited itself to investigating one or a few different modelling approaches at the time. Furthermore, in many studies, collecting a large sample has seemingly proved difficult for researchers. Moreover, many reports and articles have joined multiple of the ratings into groups (e.g. investment-grade, speculative), instead of considering the full range of possible ratings, which seems to devalue the relevance of their results. As such, it appears that there is room for improvement, both regarding methodology, and increasing the relevance of results.

In light of these observations, this thesis seeks to answer the research question of how well machine learning models can predict credit ratings, and if they can do so with higher accuracy compared to traditional statistical methods. Lastly, if they are better – we are interested in which model yields the highest accuracy.

As part of attempting to answer this research question, this thesis will model credit ratings from S&P, Moody's and Fitch based on public financial data, using different statistical and machine learning techniques. The goal is to have a comparable benchmark of the different approaches, on the prediction accuracy. This thesis will attempt to alleviate some of the issues from past studies, collect a large sample size and to use as many ratings classes as possible. Another goal of the thesis is also to implement an automated process for selecting modelling parameters. In the spirit of scientific rigour and academic integrity, another goal of the thesis will be to elaborate in-depth on the methodology used, such that others can reproduce or develop on the study.

Lastly, the final goal of the thesis is to provide the reader with an understanding of both credit ratings and the applied machine learning approaches.

The goal of thesis is not so much to determine which variables are most relevant in the credit rating process, nor is it necessarily to create more transparency about the ratings process. Machine learning would be inconvenient way to answer those questions, as many machine learning techniques, generally have little room for logic interpretation of its parameters.

On the contrary, this thesis seeks to investigate, the accuracy that machine learning models can achieve in modeling credit ratings, and which of the different machine learning algorithms in the scope of this thesis, have the highest accuracy. Whereas this is not distinctly unique, compared to previous studies, we do believe this thesis differentiates itself from other studies, by combining best practice from past studies, along with a significantly larger dataset compared to any previously published study within this field.

1.3 Overview and thesis outline

The thesis starts by, in Chapter 2, giving the reader an understanding of what credit ratings are, their role in financial markets, and how the rating process is structured.

Chapter 3, explores both traditional statistical and machine learning modelling approaches that have been used to model credit ratings. In the chapter, some basic knowledge of the models and their characteristics is described, along with some challenges of using these types of models for modelling credit ratings. The chapter seeks to give the reader some background on how the machine learning algorithms work and explain how these are different from statistical approaches.

Chapter 4 is dedicated to a review of previous literature on the topic. It will go over some past attempts to model credit ratings, both using statistical and machine learning techniques. We briefly explain and discuss the methods of the different authors and how these relate to the work presented in this thesis.

Chapter 5 describes the methodology used in this thesis. We describe how we apply the modelling techniques in practice, and how we find modelling parameter. Secondly, we describe how the modeling is done, and which techniques are employed and how. Lastly, we go through the performance evaluation methods used to evaluate the different models.

In chapter 6, we first detail the data collection process, how we process the data and our considerations regarding this. We then present and describe the data that we have collected and go through the data processed prior to any modeling, and present relevant data from this process.

In chapter 7, the main analysis of the thesis is presented. Here, we show the main findings of the different modelling approaches and their intermediate results. In the end, we present an overview of the performance of the different models.

Finally, in chapter 8, we conclude the thesis with the conclusions that we can draw from the analysis in chapter 7. We summarize the findings, discuss the results, and consider the implications, recommendations, and potential applications of the findings. Lastly, we discuss the validity of the results, methodological weaknesses and potential improvements, and suggestions for further research.

1.4 Summary

In this chapter, we have introduced the research topic. Risk is a major consideration, when investing in debt securities, and as many market participants use credit ratings as a proxy measure for risk, CRAs have a large influence in the market. Obtaining a rating is both time-consuming and expensive, and the CRAs have faced significant criticism for not providing timely and unbiased ratings. One way to alleviate some of these issues would be to automate the process, using machine learning algorithms, as these have proven to be suitable models for predicting credit ratings based on public information. This thesis seeks to compare machine learning models and statistical models, in order to create a benchmark, and hence identify which machine learning algorithm provides the highest accuracy. The structure of the thesis has been described, and in the next chapter, we will give background on the topic of modeling credit ratings, and how these can be modeled.

2 Credit Ratings and Credit Rating Agencies

2.1 Introduction

This chapter provides background information about credit ratings, the CRAs and their methodology. We start by defining credit ratings, describe their role in credit markets, who the different users of credit rating are and how they use them. We proceed by describing the rating scales of the different CRAs, the general credit rating process, and describe the variables that are considered in a credit rating. Lastly, we describe the criticism that CRAs have faced.

2.2 Credit Ratings

There is no single industry definition or standard defining what a credit rating is. The U.S. Securities and Exchange Commission (SEC) states that credit ratings reflect a CRA's opinion, of the creditworthiness of a particular company, security, or obligation (U.S. Securities and Exchange Commission, 2003). Moody's define creditworthiness as the ability and willingness of an obligor to make full and timely payment of amounts due on a security over its life (Moody's Investors Service, 2004).

More than 150 CRAs exist worldwide (Langohr and Langohr, 2012). Each has its focus, ratings scale and methodology. For more than a century, these agencies have been providing their opinions on bonds and the companies that issue them. Over time, credit ratings have become increasingly important for users and providers of debt financing (Langohr and Langohr, 2012). As we will see in this section, credit ratings affect markets in a multitude of ways, among others, issuers' access to capital, transaction structures, and the ability of certain institutional investors to make particular investments (U.S. Securities and Exchange Commission, 2003).

The primary users of credit ratings are bond investors who use the rating as a measure of the creditworthiness of issuers and hence the riskiness of securities they issue (IOSCO, 2003). This gives CRAs a vital role in credit markets and their opinions affect the marketability and yields of bonds (Kaplan and Urwitz, 1979). However, as we will see in this section, credit ratings have more wide-ranging implications for market participants, beyond just issuers and investors (IOSCO, 2003). For instance, ratings can also have implications for non-bond related contracts between issuers and private contractors, as well as for agreements between banks and issuers.

The three largest CRAs, S&P, Moody's and Fitch are all classified as nationally statistical rating organisations (NRSROs) by the SEC, giving their ratings certain entitlements. The SEC is the enforcing part of the financial regulatory body in the United States, whose responsibility it is to protect investors, market integrity and to facilitate capital formation (U.S. Securities and Exchange Commission, 2016). Both the SEC and its European equivalents, are frequent users of credit ratings. For instance, the SEC have more lax requirements for bond prospectuses, if the issuer is rated by an NRSRO (Frank, 2009). In Europe, although the uses of ratings in regulatory contexts historically have been less common (Langohr and Langohr, 2012), they now have significant regulatory implications. For instance, the Basel II framework specifies that ratings from approved agencies can be used by banks, when calculating capital reserve requirements (Bank for International Settlements, 2005). The regulatory uses of credit ratings go beyond these few examples, and we will look further into these in this chapter.

In determining a credit rating, S&P, Moody's and Fitch consider both quantitative factors, such as sales, earnings, and leverage, as well as qualitative factors such as market position and reputation. These are through a scoring system, converted into a single score on a scale of credit ratings. Whereas the process and specific factors considered differ between agencies, there are significant similarities in the method and process between the larger agencies (IOSCO, 2003).

The credit rating industry has been subject to controversy over the past few decades. Literature suggests that the large credit CRAs have played a significant role in exacerbating some high-profile bankruptcies, and in the 2008 financial crisis (White, 2010). As the firms being rated are also the ones paying the CRAs, their objectivity has been called into question, further supporting the need for automated, objective credit rating alternatives (Frank, 2009).

2.3 Rating agencies

CRAs are the companies that assess the creditworthiness of corporate and government entities, and the fixed-income securities that these issue. CRAs provide investors and lenders with an understanding of the risk faced when purchasing the bonds of a fixed income security issuer. As such, a credit rating is an evaluation of the ability and willingness of a borrower to pay their financial commitments to the lenders under the terms of the issue, with the purpose of decreasing the asymmetric information between the two parties (Langohr and Langohr, 2012).

CRAs, in their present form, did not emerge until about 100 years ago. In the 19th century, its place was filled by three separate types of institutions; credit *reporting* agencies, the specialised business press and investment banks (White, 2010). The first to bring these parts under one roof was John Moody in 1909, who established Moody's Investor Service. Even S&P was has deep roots in the history of CRAs – The Poor Company, a prominent specialised business press company, merged with Standard Statistics to become the S&P we know today. Investment banks, as underwriters, used their good reputation incite the confidence of investors (Langohr and Langohr, 2012), but this function was in part taken over by the CRAs, who offered an independent evaluation.

2.4 The Need for Credit Ratings

Credit ratings exist as a piece of information at the intersection where supply and demand for capital meets. Providers and users of ratings agree that they are an opinion of an entity's creditworthiness. In this section, we explore the economic function that credit ratings fill, and look at how different market participants use ratings.

2.4.1 Economic Function of Credit Ratings

Credit ratings reduces friction in matching users and providers of capital, as they satisfy certain needs of both parties, respectively. Capital providers need information about the quality and risks of their investments and users of capital need access to said capital. By satisfying these needs, credit ratings facilitate optimal decisions for investors and borrowers. Credit ratings can in a sense be seen as a way to reduce transactions costs, as they for investors reduce the cost of information and, for borrowers, reduce the cost of market access (IOSCO, 2003).

A prospective borrower will have more information about its creditworthiness than potential lenders, as they have access to information, which is not publicly available, as well as a more specialised understanding of the market and economic conditions they are operating within. This asymmetry of information between the two parties puts lenders in a position, where they gain from selectively disclosing information that would favourably bias the opinion of outsiders.

Such information asymmetries lead lenders to insist on being rewarded for taking upon the risk of such asymmetries impacting them adversely, which translates into higher capital costs for borrowers (Frank, 2009).

A primary rating function is to objectively measure the credit risk, about a certain issue or issuer and to resolve the information asymmetries that exists between lenders and borrowers. As such, the CRAs, by providing an unbiased opinion, with access to some non-public information, alleviates some of this asymmetry and minimise the higher capital costs of borrowers and higher investment evaluation costs on the part of lenders. This is the primary value of credit ratings - reducing the transaction costs and market friction, that can occur in such transactions, by reducing information asymmetry. This is also why regulators condone the existence of CRAs - in fact, they share some of the same goals in promoting a well-functioning capital market.

This works as the CRAs are held accountable by both sides involved in a debt transaction. For borrowers to be willing to pay for a credit rating, they must believe that it lowers the capital costs more than the costs of the rating itself. If they do not believe that it has a positive financial impact, they will simply choose not to be rated. For the rating to have any information for the lenders, they must believe that the rating has useful information about the credit quality of the potential investment. Evidently, the degree to which lenders feel that a particular CRA does have any useful information, depends on the reputation and performance of the CRAs in accurately assessing credit quality. If a CRA should consistently fail in correctly assessing the credit quality, they will no longer be of value to lenders, and thus lose its value to borrowers.

As a secondary function, credit ratings exist as a means of comparison between issuers and issues. Ratings give market participants a common standard, used to refer to credit risk by (Langohr and Langohr, 2012).

2.4.2 Users of Credit Ratings

As mentioned, credit ratings play a significant role in the decisions made by a multitude of market participants, even beyond lenders. Below, we look at how different market participants use credit ratings.

Issuers/Borrowers

Issuers use credit ratings for a number of reasons. These include improving the marketability and pricing of their securities (U.S. Securities and Exchange Commission, 2003), as a means to increasing their trustworthiness to investors and other counterparties. Furthermore, some investors have a preference for bonds with a rating, either due to the lower need for evaluation and monitoring or for regulatory reasons. As such, issuers use ratings simply to advance their access to capital (Becker and Milbourn, 2010).

Investors

Mutual funds, pension funds and insurance companies are among the largest owners of debt securities, and most retail participation in debt markets takes place through these fiduciaries (U.S. Securities and Exchange Commission, 2003). These entities are substantial users of credit ratings, as they as regulated entities under U.S. law, in many instances are prohibited from purchasing debt securities rated below a certain rating (Langohr and Langohr, 2012). Furthermore, in addition to helping investors understand the risks and uncertainties of investments, the independent opinion of creditworthiness that CRAs provide makes it easier for investors to compare different potential investments, while saving them the costs of doing their own analysis to evaluate risk prospects.

Brokers, Underwriters and Investment Banks

To a large extent, brokers and underwriters use credit ratings in a similar fashion as investors. Also, many underwriters have what they call 'rating advisory groups', who assist clients in selecting appropriate CRAs for their offerings and guide issuers through the rating process. Also, ratings have significant importance in over-the-counter (OTC) markets, where brokers and investment banks use credit ratings to determine appropriate counterparties and collateral levels (Frank, 2009). Furthermore, many of these firms are themselves issuers of bonds and debt instruments, such as the heavily discussed collateralized mortgage obligations (CMOs) (U.S. Securities and Exchange Commission, 2003).

Private Contractors

In financial and non-financial contracts, credit ratings are extensively used in so-called 'rating triggers'. These special clauses are triggered in the event of specified rating actions, such as an issuer's rating falling below a certain threshold. These can give counterparties and lenders the right to terminate the contract, accelerate credit obligations or have the rated entity post collateral (Langohr and Langohr, 2012). Such rating triggers can have severe implications, as they can exacerbate liquidity strains for issuers, who are already faced with a deteriorating credit quality (U.S. Securities and Exchange Commission, 2003).

Regulators

In the U.S., ratings have been a part of legislation since 1931, when it was ruled that banks could not hold bonds rated lower than BBB (Sinclair, 2008). During the past few decades, regulators, have increasingly used credit ratings to help monitor the risk of investments held by regulated entities such as banks and funds (U.S. Securities and Exchange Commission, 2003). Today, the use of ratings in regulation is widespread, in both federal and state laws in the U.S., as well as in EU law (Langohr and Langohr, 2012). In the U.S., the largest CRAs are recognised as NRSROs, giving their ratings a certain entitlements for regulatory purposes. In the European Union, the credit ratings of banks determine their capital reserve requirements, under the Basel II directive (Bank for International Settlements, 2005). Essentially, financial regulators use credit ratings and CRAs to outsource their judgements (White, 2010). The regulatory frameworks concerning credit ratings have contributed to the significance of CRAs in credit markets, in which they are now of central importance.

2.5 Credit Rating Methodology

2.5.1 Types of Credit Ratings

There are two types of credit ratings; *issue* credit ratings and *issuer* credit ratings. This thesis focuses on the latter; issuer ratings.

Issue and issuer credit ratings use identical symbols, but their definitions do not entirely correspond to each other. In essence, issuer ratings reflect only the risk of default, whereas issue ratings also incorporate views of the loss given default - or in other words, how much investors can recover given that an issuer default on an obligation (Standard and Poor's, 2008).

Issuer Ratings

'Issuer Ratings' rate the issuer/company as a whole, regardless of the particular debt instrument. It is not specific to a particular financial obligation, but rather provides an overall assessment of a company's creditworthiness (Standard and Poor's, 2008).

Issue Rating

'Issue credit' ratings, also called bond ratings, are the CRAs opinion about credit risk about a specific financial obligation or security.

2.5.2 Ratings Scales

CRAs summarise their opinions about the creditworthiness of obligors in ratings that are represented by a grade, from a set scale. The goal of these grades is to represent a group within which the credit quality and risk characteristics are roughly the same (Langohr and Langohr, 2012).

Even though the ratings scale and definitions vary between the agencies, the rating categories are in industry practice considered as being more or less comparable (Frank, 2009).

Table 2.1 describes the long-term issuer credit rating scale used by S&P. Their rating scale is divided into several categories ranging from the famed AAA rating, reflecting the strongest credit quality, to D, reflecting that the issuer is in currently in default or a state where payment default is imminent or unavoidable. Issuers who are rated in the top four categories AAA to BBB are said to be 'investment grade' while anything below it is said to be 'non-investment grade', and are even referred to as 'speculative', 'high-yield issues' or 'junk bonds'.

In addition to the a letter grade, the addition of a '+' or '-' as suffix gives an additional indication of the credit quality of the issuer in question. We also notice the two ratings 'R' and 'SD' which describe specific credit default or near-default situations. The rating 'R' is assigned to issuers who are under regulatory supervision, due to some aspect of its financial condition. The rating 'SD' stands for 'Selective Default', describing a situation where an issuer has failed to repay only a subset of their financial obligations but continues to pay the remainder. Short-term ratings are slightly different and for S&P these range from A-1 to D, and indicate the credit quality on a short-term time horizon. These will not be further elaborated upon, as this thesis is concerned with long-term issuer ratings.

As mentioned, the rating scales and definitions vary between agencies. In Table 2.3, below, the rating scales of the different agencies, and their inter-agency equivalents are presented, along with grade classifications.

2.5.3 Credit Rating Process

The credit rating process and methodology vary between CRAs. Some agencies use purely quantitative models and statistical analysis to form their rating. However, the larger CRAs all combine quantitative and qualitative factors (IOSCO, 2003).

Obtaining a rating is a laborious, time-consuming and intrusive process, where the issuer must be prepared to submit a vast quantity of data, and have its senior management attend several meetings with the CRA, share confidential information, conduct facility

Table 2.1: S&P Long-Term Issuer Ratings Definitions

Rating	**Definition**
AAA	Obligor has extremely strong capacity to meet its financial commitments. 'AAA' is the highest issuer credit rating assigned by S&P Global Ratings.
AA	Obligor has very strong capacity to meet its financial commitments. It differs from the highest-rated obligors only to a small degree.
A	Obligor 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	Obligor has adequate capacity to meet its financial commitments. However, adverse economic conditions or changing circumstances are more likely to lead to a weakened capacity of the obligor to meet its financial commitments.
BB	Obligor 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 which could lead to the obligor's inadequate capacity to meet its financial commitments.
B	Obligor 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	Obligor is currently vulnerable, and is dependent upon favorable business, financial, and economic conditions to meet its financial commitments. An
CC	Obligor 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	Obligor 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 / D	Obligor is in default on one or more of its financial obligations including rated and unrated financial obligations but excluding hybrid instruments classified as regulatory capital or in non-payment according to terms.
NR	An issuer designated 'NR' is not rated.
(+) / (-)	The ratings from 'AA' to 'CCC' may be modified by the addition of a plus (+) or minus (-) sign to show relative standing within the major rating categories.

Source: S&P Global Ratings Definitions, Standard & Poor's Financial Services LLC (2016)

Table 2.3: Long-Term Issuer Credit Ratings by Different Agencies

Group	**S&P**	**Fitch**	**Moody's**	**Description**
1	AAA	AAA	Aaa	Prime
2	AA+	AA+	Aa1	High grade
3	AA	AA	Aa2	
4	AA-	AA-	Aa3	
5	A+	A+	A1	Upper medium grade
6	A	A	A2	
7	A-	A-	A3	
8	BBB+	BBB+	Baa1	Lower medium grade
9	BBB	BBB	Baa2	
10	BBB-	BBB-	Baa3	
11	BB+	BB+	Ba1	Non-investment grade
12	BB	BB	Ba2	Speculative
13	BB-	BB-	Ba3	
14	B+	B+	B1	Highly speculative
15	B	B	B2	
16	B-	B-	B3	
17	CCC+	CCC+	Caa1	Substantial risks
18	CCC	CCC	Caa2	
19	CCC-	CCC-	Caa3	
20	CC	CC	Ca	Extremely speculative
		C		Default imminent
-	R	DDD	C	In default
	SD	DD	-	
	D	D	-	

tours, and have staff ready to respond to follow-up questions that the CRA might have (Langohr and Langohr, 2012).

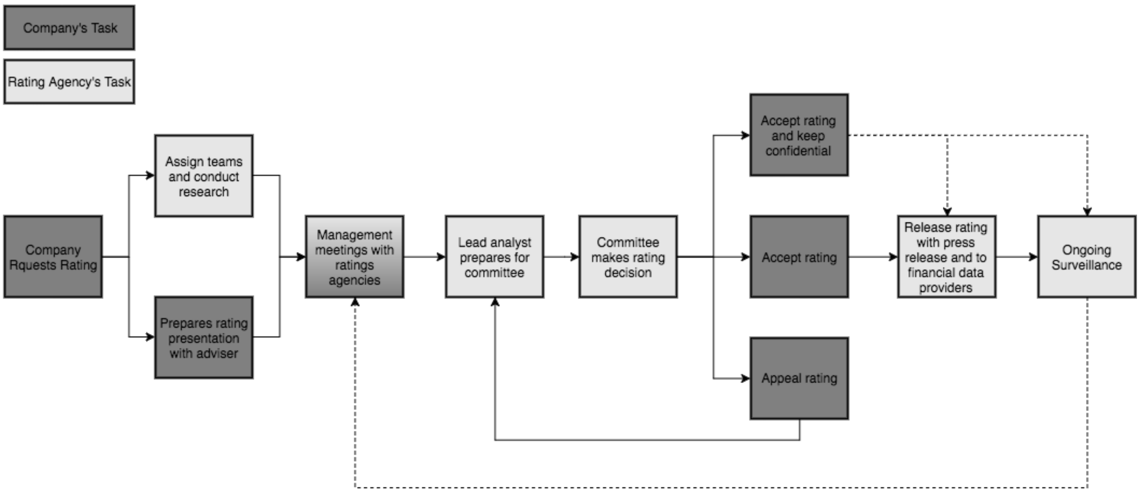
Before a debt issue, the issuer typically contacts a CRA and requests a phone call or meeting with a representative from the CRA, who will give information on the rating process and the costs.

From there on, the credit rating process is very much driven by the CRA. Submitting in documents to the CRA. These typically include:

- Relevant information on the company and its industry
- Descriptions of operations, products, and risk management
- Business plan
- Five years of audited annual financial statements
- Interim financial statements
- Draft registration statement or offering memorandum

This sets the CRA's process in motion, and a team of analysts are assigned to the case. Typically, an analyst will cover only one or two industries, in order to be sufficiently specialised. They do basic research and prepare the meetings with management. An overview of the credit rating process can be seen in figure 2.1.

Figure 2.1: Overview of a Credit Rating Process


Company's Task
Rating Agency's Task
Company Rquests Rating
Assign teams and conduct research
Prepares rating presentation with adviser
Management meetings with ratings agencies
Lead analyst prepares for committee
Committee makes rating decision
Accept rating and keep confidential
Accept rating
Appeal rating
Release rating with press release and to financial data providers
Ongoing Surveillance

Source: Langohr and Langohr (2012)

After this, interviews are held with the management of the company. Their purpose is to review the company's operational and financial plans, management policies, and other factors that could impact their credit quality.

After this, the research on the part of the CRA commences, and the analyst team will try to determine an appropriate credit rating, by bringing together understanding, data and methods.

The analyst then submits her recommendation to the rating committee - an internal committee composed of a lead analyst, managing directors and junior analytical staff. They examine the recommendation and its arguments and decide upon a final rating. The CRA communicates the rating decision to the issuer and underwriter. The issuer now has three options. They accept the rating, or they can appeal the rating, and give supporting arguments and documentations as to why it should be revised. Additionally, if the issuer strongly disagrees with the rating, but cannot present adequate arguments as to why it should be revised, they can decide to withdraw the rating request and refuse its publishing (Langohr and Langohr, 2012).

If the issuer accepts the rating, the CRA then publishes a press release, along with the rating report, while notifying financial information providers of the new rating. From there on, the rating will be monitored for one year. If the issuer has requested it, the CRA will continue to survey the issuer or issue, and revise the credit rating, in case financial conditions change (Standard and Poor's, 2008).

2.5.4 Credit Rating Methodology

CRAs look at some factors to assess the overall business and financial risk profile of an issuer and issue. While business risk factors involve fundamental analysis, dependent upon a large degree of subjective judgement, the financial risk factors involve looking at financial ratios over time (Standard and Poor's, 2008).

CRAs start by looking at the business risk of the company, where they evaluate the business of the company, the factors affecting the country, market and the issuer's competitive position within it. Finally, management is evaluated on their performance (Standard and Poor's, 2008).

Having evaluated the business risk, CRA proceed to analyse the financial risk of the issuer. As mentioned, this is a more quantitative and objective measure than assessing business risk, where financial ratios are used. To adjust for industry and company specific accounting policies and standards, several adjustments are made by analysts (Standard and Poor's, 2008).

Below are the factors used for assessing both business and financial risk are listed, and subsequently described briefly, based on the description by S&P.

Business Risk

- Country risk
- Industry Risk
- Competitive position
- Profitability and peer group comparisons
- Management review

Financial Risk

- Accounting characteristics and information risk

- Accounting characteristics and information risk
- Cash flow adequacy
- Capital structure and asset protection
- Liquidity and short-term factors
- Debt maturity schedules

Source: Standard and Poor's (2008)

Country risk

The operating environment in the particular country can have a large impact on the creditworthiness of issuers, both directly and indirectly. CRAs look beyond the credit rating of that country to evaluate country risk, including the impact of government policy on the issuer's business and financial environment (Standard and Poor's, 2008).

Industry risk

The degree of operating risk facing a company depends on the dynamics of the industry in which it participates. CRAs analyse the strength of industry prospects and the competitive factors affecting it, including growth prospects, cyclicality, technological change, labour unrest, regulatory interference, and changes in demand and supply (Standard and Poor's, 2008).

Competitive position

To assess the competitive position of an issuer, CRAs look at key factors, specific to the industry, in which it operates. Company size and diversification also play a role, but the CRAs stress that there is no fixed size criterion for certain ratings. However, size is often correlated to rating levels, as larger companies benefit from economies of scale, translating into a stronger competitive position (Standard and Poor's, 2008).

Profitability and peer group comparisons

Profitability is an important factor in credit quality assessment. A company generating higher operating margins has greater ability to meet financial obligations and withstand business adversity, while also attesting to asset values. CRAs also compare with peer companies on key profit metrics, to assess how the issuer is performing (Standard and Poor's, 2008).

Management review

Management is assessed for its role in determining operational success and its risk tolerance. CRAs use both track record of managers as well as the interviews conducted. This is a highly subjective process. The plans and policies of management are assessed, in their realism, their state of implementation, and how well they are executed or enforced.

Governance, risk tolerance and financial policies

The financial risk profile is, in part, determined by governance policies and procedures, the company's appetite for financial risk and its financial policies, concerning accounting practices, capital spending levels, debt tolerance, merger activity and asset sales (Standard and Poor's, 2008).

Accounting characteristics and information risk

Financial statements and disclosures serve as the CRAs' primary source of information about the financial condition and performance of companies. Accounting characteristics are reviewed, to determine whether ratios and statistics derived from the statements can be used to appropriately measure performance and position.

Analytical adjustments are often made to better portray reality and to make the ratios and statistics comparable to peer group companies (Standard and Poor's, 2008).

Cash flow adequacy

Although there is usually a strong relationship between cash flows and profitability, earnings is an accounting concept, and debt obligations must be serviced in cash. Thus, CRAs evaluate the debt-servicing capabilities by analysing cash flow patterns.

Cash flow analysis is according to Standard and Poor's (2008) the single most critical aspect of credit rating decisions, and is even more important when they are rating speculative-grade issuers, as they often have limited sources of alternative financing, which can be raised to service debt, and they are thus reliant on generating cash internally.

Capital structure and asset protection

CRAs will conduct a review of the issuers capital structure, which encompasses both the level and mix of debt types. Their analysis goes beyond reported debt and includes items such as leases, pension and medical liabilities, guarantees, and contingent liabilities.

Liquidity and short-term factors

Other factors, which are not in the other categories are examined as part of this category. The potential impact of adverse outcomes is considered, along with the management's contingency plans for these. Such outcomes include legal problems, lack of insurance coverage and covenants in loan agreements. In essence, this is an examination of how stress affects the company and its capability to sustain strain in the short run.

Debt maturity schedules

CRAs will look at the repayment scheduling of existing debt, to assess how reliant the issuer is on bank financing, as well as how the timing of servicing of existing debt, coincides with forecasts.

These factors help score the issuer on a 1-5 scale on both the business and financial risk profile, which are in turn used to formulate an anchor rating.

Table 2.4: Table of Typical Ratings Outcomes

	Financial Risk Profile				
Business Risk Profile	**Minimal**	**Modest**	**Intermediate**	**Agressive**	**Highly Leveraged**
Excellent	AAA	AA	A	BBB	BB
Strong	AAA	AA	A-	BBB-	BB-
Satisfactory	AAA	BBB+	BBB	BB+	B+
Weak	BBB	BBB-	BB+	BB-	B
Vulnerable	BBB	B+	B+	B	B-

Source: Standard and Poor's (2008)

However, as mentioned, there are some factors contributing to the outcome being different than the rating shown in Table 2.4, which is a function of the CRA's methodology, the analysts' evaluation and the rating committee's opinion.

2.6 Criticisms of Rating Agencies

Credit ratings have been widely accused of having severe involvement in the reasons behind and events leading to the 2007-2008 global financial crisis (White, 2010). In particular, the largest point of criticism was their involvement in giving investment-grade ratings to CDOs and MBSs, although these were backed by low-quality loans. While criticism of CRAs is not new, there is still substantial debate surrounding CRAs, due to their significant role in financial markets.

2.6.1 Conflicts of Interest

Issuers Paying for Ratings

It is often argued that CRAs have a clear conflict of interest, as they serve to masters; the issuers and the investors. While investors want the ratings to be objective and as accurate as possible, issuers want ratings to be better, to increase capital access and lower their cost of debt (Langohr and Langohr, 2012). As it is the issuers who pay the fees for ratings the worry is that this shifts the emphasis to commercial gains, rather than the provision of objective and unbiased measures of creditworthiness.

Ancillary Business

In addition to their core rating business, the large CRAs have developed ancillary businesses such as rating assessments services, risk management and consulting services (U.S. Securities and Exchange Commission, 2003).

These ancillary businesses create another potential conflict of interest for CRAs, and there are concerns that rating decisions are affected by whether the issuer purchases any of these additional services. In many ways, this concern is similar to that, which is prevalent in other professional services industries, such as auditing firms supplying consulting services or investment banks both conducting equity research, while also providing transaction services (U.S. Securities and Exchange Commission, 2003).

Familiar Relationships with Management of Issuer

In the rating process, CRAs will meet with management of the company they are rating. The worry is that CRAs could open themselves up to adverse influences and the vulnerability of being misled, having adverse effects on the accuracy and bias of their ratings (Frank, 2009).

2.6.2 Accuracy and Timeliness

The accuracy and timeliness of ratings have been under large scrutiny, especially in the wake of Enron and other high-profile bankruptcies. In the case of Enron, despite CRAs having been aware of the company's issues for months prior, Enron's rating had remained investment-grade until a few days before declaring bankruptcy (Frank, 2009).

2.6.3 Competition

The credit rating industry is dominated by S&P, Moody's, and Fitch. This has lead to concerns surrounding the state of competition in the industry. There has been much discussion about the three largest CRAs' statuses as NRSROs, and that this impedes competition, as it creates larger barriers to entry (U.S. Securities and Exchange Commission, 2003).

Furthermore, Fitch has in the past accused S&P and Moody's of using anti-competitive practices such as 'notching', where they will consistently give an issue a lower rating, unless that issuers other ratings are also rated by the same CRA (U.S. Securities and Exchange Commission, 2003).

Becker and Milbourn (2010), investigates the effect of increased competition on rating quality. Their results suggest that competition has lead to lower-quality ratings. As such, the question of whether the industry has too little or too much competition, appears to be an unresolved question in academic literature.

2.7 Summary

The general view in literature is that the role played by CRAs is conducive to the efficient operation of financial markets. As we have seen ratings are used for a variety of purposes, among a multitude of market participants. It is firmly established that CRAs have a substantial impact on the functioning of markets, and thus their opinions hold a lot of power.

While a large number of CRAs exist worldwide, the industry is dominated by the three largest firms, being S&P, Moody's and Fitch. They all use a combination of both subjective and objective, quantitative measures, in assessing the creditworthiness and credit quality of issuers and debt issues.

CRAs have faced substantial criticism both in the wake of high-profile bankruptcies, such as Enron, as well as for its role in exacerbating the 2007-2008 global financial crisis. This calls into question the objectivity, accuracy and timeliness of credit ratings.

With the concerns surrounding credit ratings, one can make an argument for a need to develop an alternative method for evaluating creditworthiness or some form of auditing of the quality of the CRAs' ratings. Such a method could involve better statistical modelling, which could contribute to more objective ratings and improved timeliness.

In the subsequent chapter, the methods used to model credit ratings are explored.

3 | Modelling Techniques

3.1 Introduction

In this chapter, we will look at some of the different types of modelling techniques that have been used in past studies to model credit ratings, and those that will be used in this thesis.

We start by considering the more classical statistical models, their characteristics, and their strengths and weaknesses in modelling credit ratings.

After considering some of the statistical techniques, we will review some selected machine learning approaches that will be employed in this thesis.

As we will discuss in this thesis, there is no clear line on what separates statistical techniques from machine learning techniques, but for the purposes of this thesis, the techniques considered machine learning are clearly specified.

Lastly, this chapter will explain some general techniques used in data mining. Data mining is somewhat different from classical statistical methods, as it concerns itself less with econometric theory of how models should be constructed, but takes its outset in data, and uses specific techniques as a means for creating models that are both effective, and generalizable.

The aim of this section is to give the reader an understanding of the different approaches that have been used to model credit ratings in past studies, and the ones that will be used in this thesis. This will be of importance in understanding the chapters containing the literature review, methodology, and analysis.

3.2 Statistical Techniques

In this section, I will go through a few of the models, which belong to the class of traditional statistical techniques, that have been used for modelling credit ratings. While this list is not exhaustive, it aims, to give an understanding of some of the ways to model credit ratings.

3.2.1 Logistic Regression

Logistic Regression (LR) has long been a common technique in statistics and econometrics. Logistic regression is an especially appropriate model when the response variable (i.e. the variable, which we are trying to predict) is categorical.

A linear regression model outputs a continuous response variable through the linear combinations of predictor variables (Kennedy, 2013). In distinguishing a dichotomous outcome, we want to reduce this output to 0 or 1. LR achieves this by applying a logistic transformation, which transforms the output from $[-\infty, +\infty]$ to a probability such that $g(x) \in \{0, 1\}$.

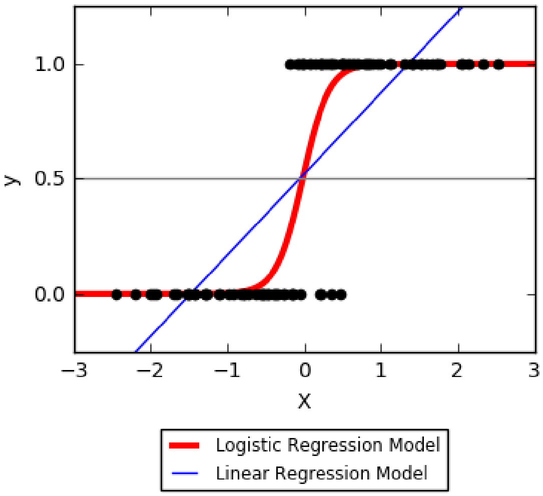
$$g(x) = \log \frac{\pi_g}{1-\pi_g} = \boldsymbol{x_i'\beta} \tag{3.1}$$

where π_g is the probability of belonging to a class, and the term $\frac{\pi_g}{1-\pi_g}$ is called the odds ratio. This is the logit transform link function, which is used to relate the probability of class membership to a linear function of the input variables. There are multiple other link functions such as the probit function. However the logistic function is the easiest to interpret and the differences in performance are small (Kennedy, 2013).

The vector of coefficients, $\boldsymbol{\beta}$, are estimated using maximum likelihood estimation (MLE), which is an iterative optimisation function that iteratively 'guesses' coefficient value to maximise the log likelihood (Kennedy, 2013).

A graphical comparison of linear and logistic regression is shown in figure 3.1

Figure 3.1: Logistic vs. Linear Regression



As an extension of the binomial logit, it can also be made into a multinomial logit model, that has the ability model multiclass problems, such as credit ratings. In the multinomial logit model, the log-odds of each response is assumed to follow a linear model.

$$g(x_{ij}) = \log \frac{\pi_{ij}}{\pi_{iJ}} = \alpha_j + \boldsymbol{x_i'\beta_j}, \tag{3.2}$$

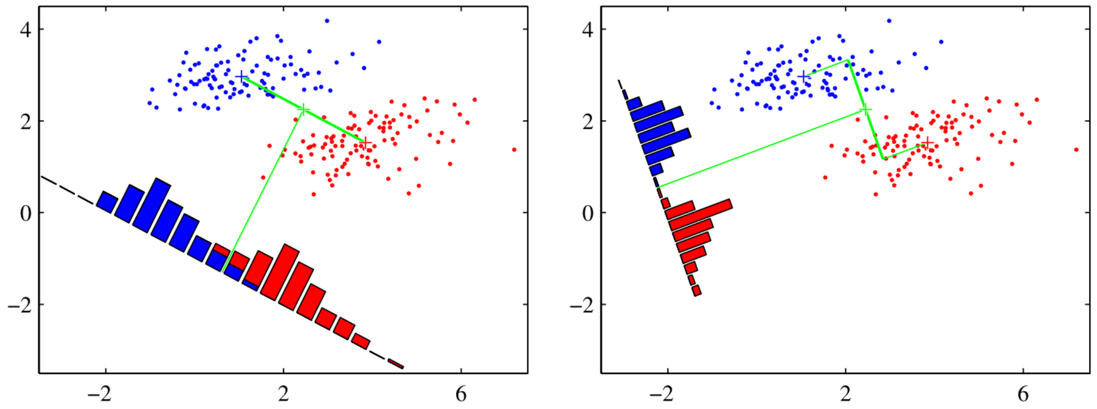
where α_j is a constant and and β_j is a vector of regression coefficients for $j = j \in [1, J-1]$

3.2.2 Linear Discriminant Analysis (LDA)

In short, LDA separates classes by finding the linear combinations of features which best separates them, using these linear combinations as a projection vector.

Figure 3.2 shows two plots, illustrating how LDA differs from normal comparisons of the means. The left plot shows the distributions of the two classes when projecting the features onto the line joining the class means. The difference is clear in this example; in the left plot, there is a significant overlap between the two distributions of classes, which

Figure 3.2: Illustrating LDA separation vs. Means Comparison



Source: Reproduction from Bishop (2006)

leads to poor separation. When using LDA, we see that the corresponding projection, shows virtually no overlap in the class distributions, leading to improved class separation.

Multiple Discriminant Analysis (MDA) is much like LDA, but with more than two classes, and a solution that is a projection space that simultaneously has the best joint separation of groups in the multivariate space.

3.3 Machine Learning Techniques

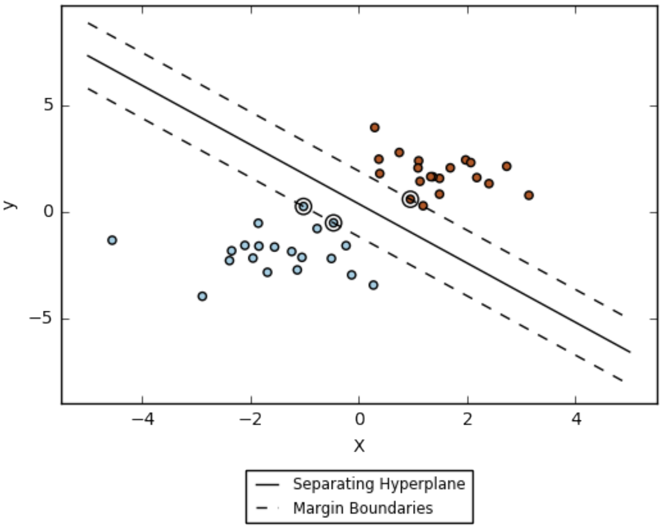
3.3.1 Support Vector Machines

Support Vector Machines (SVMs) use instances from the training data as support vectors, to outline a class-separating hyperplane in the feature space. Its optimisation sets the margin, which maximises the Euclidean distance from the separating hyperplane to the closest support vector. An important part of SVMs is using a so-called kernel to map the input variables to a higher dimensional feature space, such that it becomes linearly separable (Kennedy, 2013).

In Figure 3.3 we see how the SVM works, by creating a separating hyperplane, which maximises the margin between the input features as support vectors. It is important to note here, that this is a simple example, where the data is easily linearly separable. If it had not been linearly separable, the SVM would, using a kernel function, map the input vectors to a higher dimensional space (e.g. x^2, or any other higher-dimensional space), to make the classes linearly separable.

In general, SVMs perform well across many different domains. They have also been successfully applied to the problem of predicting credit ratings.

Figure 3.3: Illustration of SVM



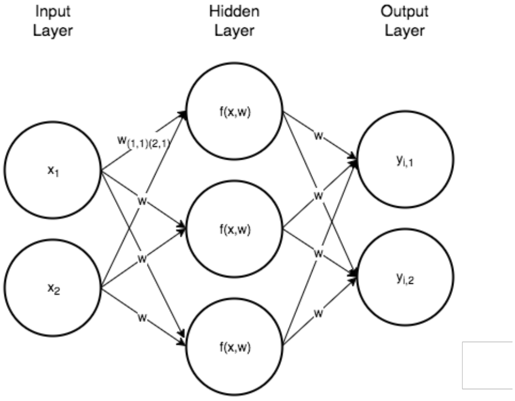
3.3.2 Neural Networks

Neural networks is a class of machine learning which learns relationships from data, and have been used extensively in a variety of applications. They consist of processing units, known as nodes (in some cases described as neurons or perceptrons), which are organised into layers. Each node between each layer is interconnected with every node in the next layer, with different weights. The weights that are the parameters, estimated in the model.

It starts with an input layer, where the input features are passed to it. The data then propagates throughout the network, layer by layer between the input and output layers (these are referred to as 'hidden' layers), until it arrives at the output layer.

Within each node is an activation function that, based on the values passed to it from the previous interconnected nodes and their associated weights, gives the value of the nodes, which is then passed on to each node in the subsequent layer.

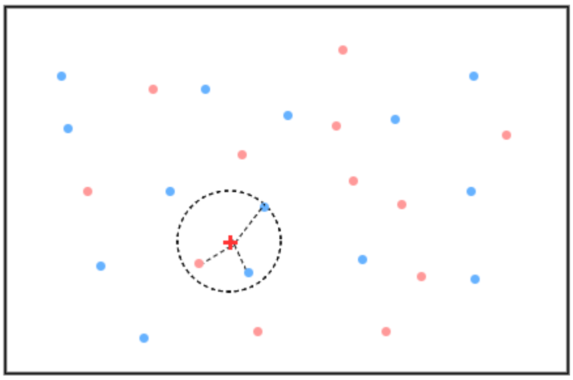
Figure 3.4: Illustration of Neural Network with two input variables, one hidden layer, and binary response variable



3.3.3 Nearest Neighbours

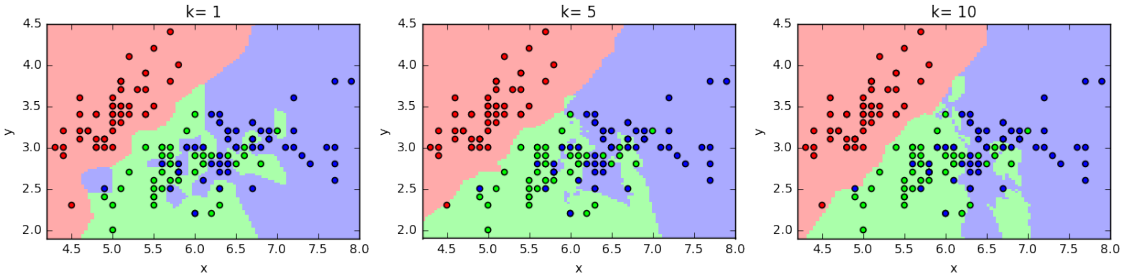
Nearest Neighbour type classifiers are memory-based and do not require a model to be fitted by estimating parameters. The k-Nearest Neighbour algorithm (KNN), given an input, x_0 in the feature space, finds the nearest, in terms of Euclidean distance, k training points for $x_{(r)}, r = 1, \dots, k$ (Hastie et al., 2009) and classifies the input instance as the class, which the majority of those neighbours belong to. This is illustrated in Figure 3.5, where we can see that when $k = 3$, the input feature set would classify the instance shown as belonging to the blue class.

Figure 3.5: Illustration of k-Nearest Neighbor Algorithm with $k = 3$



In practice, the KNN-algorithms, establishes 'decision boundaries' in the feature space. In Figure 3.6, we see how the decision boundaries are made up, based on the training data points shown, and as a function of how many of the nearest points it should consider. Given a set of input variables, the kNN-model will look at which decision boundary in the feature space, that the set lies within and classify it as such.

Figure 3.6: Illustration of k-Nearest Neighbor Decision Boundaries with different values for k

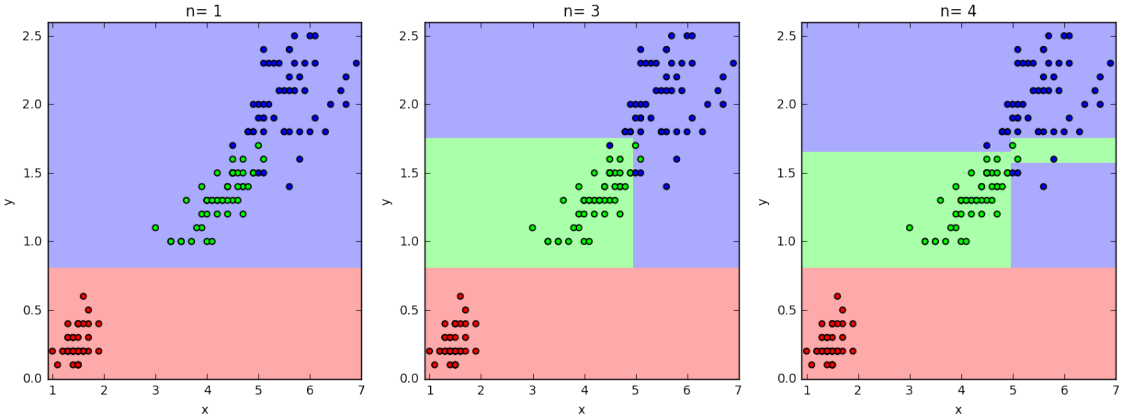


3.3.4 Decision Trees

Tree-based algorithms partition the feature space into a set of rectangles and then fit a simple model in each one. They are conceptually simple, yet powerful models (Hastie et al., 2009).

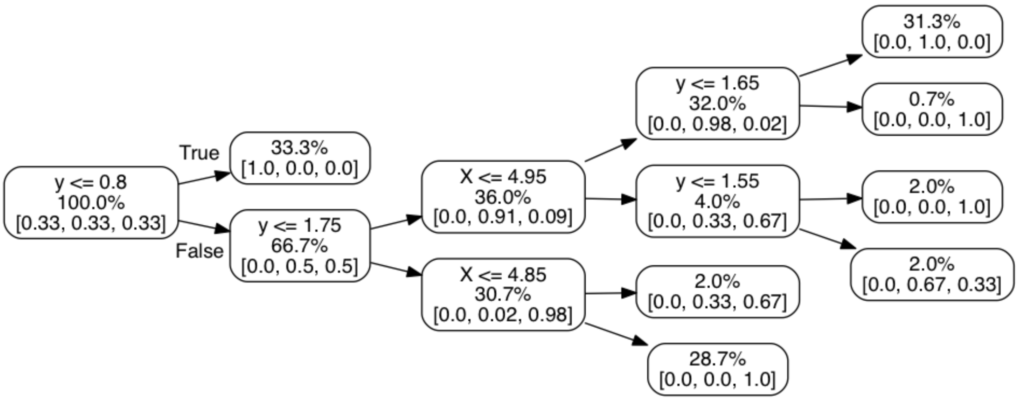
In Figure 3.7 we see how this works in practice. Here, a decision tree classifier has been trained using some the test data. Here we see exactly, that the algorithms draw rectangles

Figure 3.7: Illustration of Decision Tree Algorithm, with Different 'Depths'



to separate the data. The parameters, that the algorithms estimates are the points in the feature space, where the rectangles are bounded. In the leftmost graph, the tree depth has been set to 1, meaning there can be only one parameter separating the data. Here the algorithm chooses to draw the decision boundary line at $y = 0.8$, separating the data into the 'red' and 'blue' classes. Of course, this is less useful as we have three classes of data in this example. We improve the result further by allowing a depth of 3 levels in the tree, as shown in the middle graph. Lastly, this can be improved slightly by drawing another rectangle, by increasing the maximum depth to 4 levels.

Figure 3.8: Decision Tree Shown as Graph



In Figure 3.8, we see how the tree, which is also in the rightmost plot in Figure 3.7, is constructed and which parameters are calculated. It is much like a rule-based algorithm for creating decision boundaries.

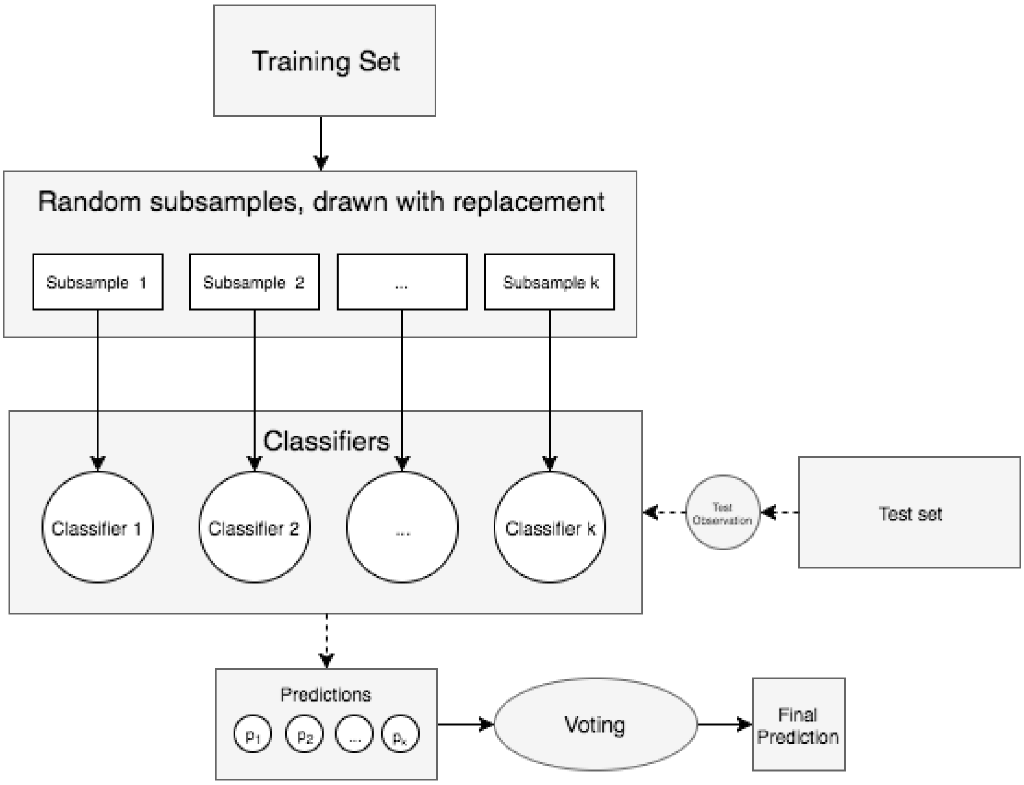
3.3.5 Ensemble Techniques

In machine learning, ensemble methods are not algorithms in themselves, but rather meta-methods, which combine multiple algorithms into one. The goal of doing this is to get models with higher accuracy or greater stability. In this subsection, we will discuss some of the ensemble methods used in machine learning.

Bootstrap Aggregation (Bagging)

Bootstrap aggregation, also known as *'bagging'*, is a meta-algorithm, made up of several different classifiers, who each 'vote', with equal weight, on which class the observation should be assigned to. In training each of the classifiers, a random subsample of the data is drawn, hence the name 'bootstrap'.

Figure 3.9: Bootstrap Aggregation Visualisation

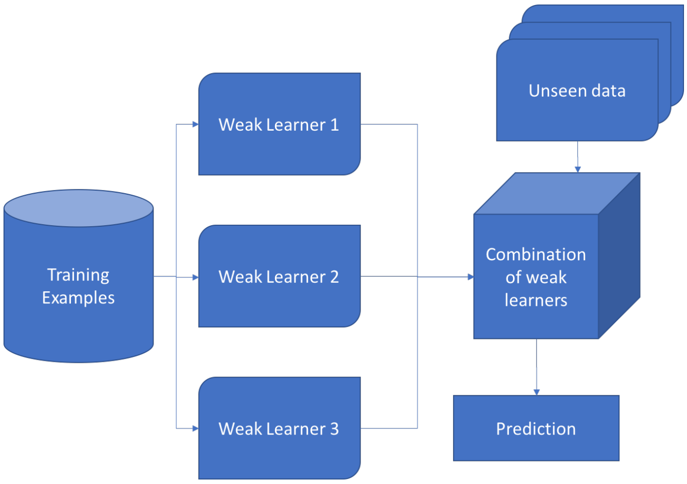


Source: Maheshwari (2016)

Boosting

Boosting is another type of ensemble technique, in which the meta-classifier is incrementally trained, where the result of each classifier re-weights the data to emphasise the data points, that the previous models has misclassified, and gives them more weight in the next classifier. Thereby, these points, are taken more into account, in the final classifier, which becomes increasingly better at classifying the 'corner cases'. It uses weak classifiers - if one used strong classifiers, the strongest one would take the vast majority of the weight, and thereby cancel out the weaker classifiers.

Figure 3.10: A boosting model made of multiple weak learners



3.3.6 Hyperparameters

In machine learning, most algorithms have some set of configuration parameters, that determines how the model will behave. In order to distinguish these from standard model parameters estimated when training the model, they are called hyperparameters. These express high-level properties of how the algorithm in the model should behave (Quora, 2017).

For instance, this could be the step size that should be used in a SGD-based optimization routine, or the maximum number of leaves should be in a decision-tree model.

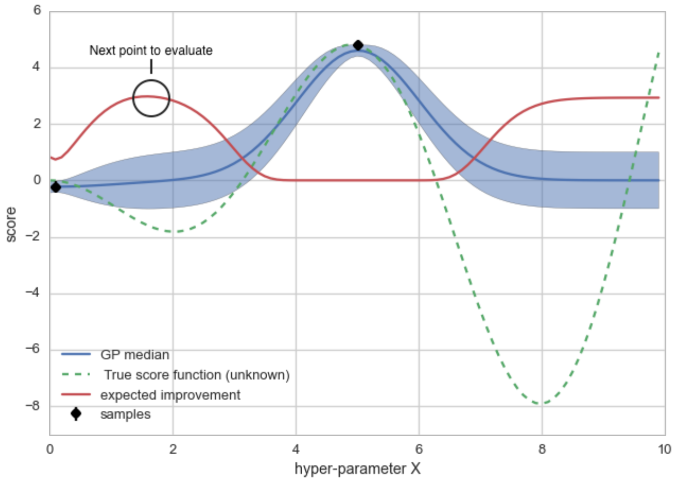
3.3.7 Hyperparameter optimization

Having established what hyperparameters are, we also need to know how to set them for each model. This is often a complex task, and can often take expertise rather than a structured approach. There are a few different strategies available; manual search, an exhaustive grid search, a random search or Bayesian optimization. The first can be time-consuming and require expert knowledge or insight. Grid and random search can be extremely computationally expensive, as they need to evaluate many different model configurations, with algorithms that are already computationally expensive. Lastly, using Bayesian optimization of hyperparameters you adjust the hyperparameters of a model, based on previously tested hyperparameter values.

Under Bayesian optimization, we develop a model for the metric that we want to optimize as a function of the hyperparameter(s). The model is then evaluated, at which point our model believes it will gain the highest metric score, and tests that hyperparameter. Based on the output, it then updates its expectations, develop a new model and then test it at the newly found point. This is also known as sequential model-based global optimization.

In our case, we use Gaussian processes as a means to model the objective function, and we optimize at each point for the expected improvement, and iteratively update our expectations with every observation. We illustrate this below in Figure 3.11, where we see the blue line as a fitted model to our two observations. Based on this, we calculate where the largest expected improvement is, and evaluates that point.

Figure 3.11: Bayesian Hyperparameter Optimization with Gaussian Processes



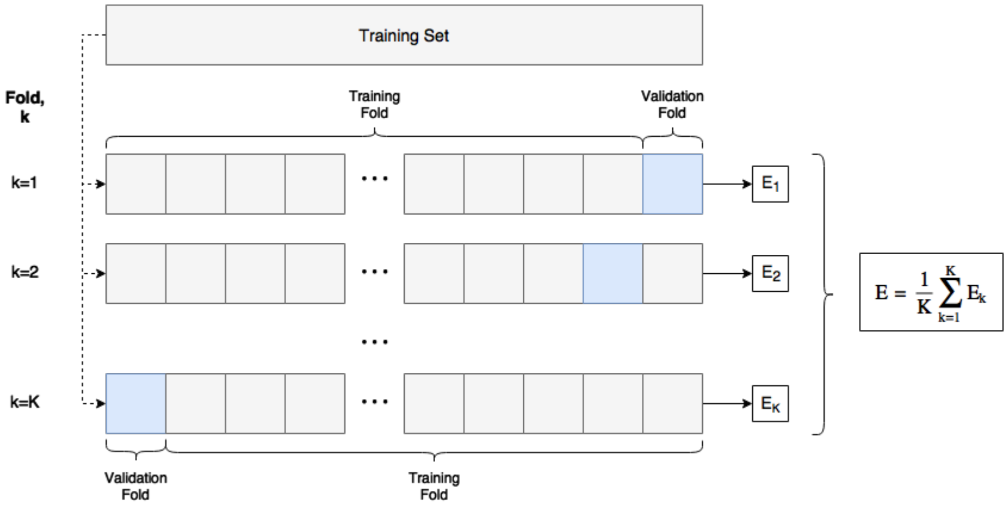
Source: Reproduction from Head (2015)

When this is done in practice, we optimize over a number of hyperparameters simultaneously, which is much like the simple illustration in this subsection, but in a higher-dimensional space.

3.3.8 k-Fold Cross Validation

In the k-Fold Cross Validation technique we split the training sample into k random parts. When a model is trained, we train and test over the data k times using the $k-1$ parts for training and the last part for evaluating the model. Effectively, this is a method for giving the models different data to be trained on each time, and leaving the remaining part out as a holdout sample. This is done in order for our training model to not be as overfitted, with respect to a specific dataset. Instead, we now have 10 different models which we can evaluate, and thus also find the statistical uncertainties of the model's metrics, as we now have a sample of the models.

Figure 3.12: k-Fold Cross Validation



3.3.9 Train-Test Splitting

The performance of the models is generally done by measuring the ability of the model to correctly determine the rating of observations for which the true rating is known. This generally is done by dividing the sample into two parts - a training and a test, or hold-out, sample, which is 'unseen' by the model. This is considered the appropriate way of validating the model, as models can be poorly generalisable due to either over- or under-fitting of the model.

3.3.10 Information Leakage

Information leakage in machine learning is when information from the test set has 'leaked' into the model. In essence, this implicates that the trained model, has already 'seen' the data it we are using to evaluate it with. This often results in some degree of overfitting and poor generalization of the model, meaning that the metrics obtained from evaluating the model on the test set is not valid, as it not truly out-of-sample (Brownlee, 2016). Information leakage can occur from a number of different reasons, and it can sometimes be hard to identify when it occurs. For instance, if one were to standardize the dataset, or do a cross-sectional operation on it, using information from the training set that information would be embedded into the entire dataset, and hence affect the entire dataset - even after subsequently splitting it into training and test. In order to avoid information leakage, the first step after obtaining the data, is to split it into training and testing datasets, and keep these apart.

3.4 Comparison Measures

3.4.1 Accuracy

Classification accuracy is the number of correct predictions made divided by the total number of predictions made (Brownlee, 2014),

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.3}$$

3.4.2 F-measure

The F-measure (or F1 score) is a measure of accuracy. In evaluating machine learning models, we often consider the measures precision and recall. Precision, is the rate of true positives to total positives, whereas recall, also known as specificity, measures the proportion of true positives to the sum of true positives and false negatives. However, an increase in one of these metrics often comes at a cost to the other. Hence, the F-measure is useful in re-conciliating these, and can be seen as the weighted average between these two.

$$Precision = p = \frac{TP}{TP + FP} \tag{3.4}$$

$$Recall = r = \frac{TP}{TP + FN} \tag{3.5}$$

$$F1 = 2\frac{pr}{p + r} \tag{3.6}$$

3.4.3 Cohen's Kappa

Cohen's kappa measures the agreement between two raters, in their ability to classify N items into C mutually exclusive classes. It is done to measure how well a trained classifier can predict a set of values, compared to a naive classifier, which simply assigns classes at chance, only knowing the class distribution (Viera and Garrett, 2005).

Kappa can take on values in the range of $k \in [-1, 1]$, where 1 is equal to perfect agreement, and any values less than one implies less than perfect matrix agreement.

Kappa can be negative, but this implies that the two classifiers agreed less than would should be expected by chance.

The equation for k is:

$$\kappa = \frac{p_0 - p_e}{1 - p_e} = 1 - \frac{1 - p_0}{1 - p_e}, \tag{3.7}$$

where p_0 is the relative observed agreement, and p_e is the hypothetical probability of chance agreement, using the observed data to calculate the random classification probabilities.

In Table 3.4.3, we see the interpretations of different Kappa values, as specified by Viera and Garrett (2005). According to the kappa values, for any of our models to be good predictive models, they should have a kappa value of 0.40 or higher.

Table 3.1: Interpretations of Kappa Values

Kappa Value	**Interpretation**
$[0,00,0.20[$	Poor agreement
$[0.20,0.40[$	Fair Agreement
$[0.40,0.60[$	Moderate Agreement
$[0.60,0.80[$	Good Agreement
$[0.80,1.00[$	Very Good Agreement

3.5 Summary

In this chapter, we have gone through some basics of using both statistical methods and machine learning techniques for classification tasks in general. We have also explained how we can measure the effectiveness of the models.

Whereas this chapter serves as an introduction to machine learning techniques, there exists a plethora of other techniques within this field, which are equally interesting, and relevant to this area of reasearch. Many of these other techniques, have been used for many interesting problems successfully, and could potentially contribute to solving the problem studied in this thesis. However, these are outside the scope of this thesis.

Having built a basic understanding of what machine learning is and how it is applied to the problem of predicting credit ratings, we will in the next chapter look at how researchers have done so in past studies.

4 | Literature Review

In this chapter, we consider some of the past literature on the topic of modeling credit ratings. We both look at some of the older literature, which primarily uses statistical methods, and more recent literature, which also uses machine learning algorithms. We attempt to provide the reader with a coherent overview of the previous research that have been made on the topic, and present the most relevant articles, summarize their method and key findings.

4.1 Introduction

Modeling credit ratings is an old topic within literature. Fisher (1959), used statistical techniques to analyse industrial bond ratings. Furthermore, Horrigan (1966) used linear regression and accounting data to model credit ratings. West (1970), expanded on this, also using statistical techniques. Ederington (1985) continued this path of research and benchmarked various regression techniques.

Most of the early research in modeling credit ratings was done using statistical techniques. In the 1980s Machine Learning appeared as a branch of computer science and it was soon thereafter applied to the topic of credit ratings. There does not exist, as such, a clear distinction between statistical and machine learning techniques, but it should rather be seen as a continuum of different techniques, ranging from statistical hypotheses to computational pattern recognition (Frank, 2009).

Statistical classification techniques classification have existed for far longer than machine learning methods. Whereas they take their point of departure from traditional probability statistics, machine learning draws on ideas from a diverse set of disciplines: artificial intelligence, probability, statistics, computational complexity, information theory, and philosophy (Gibert et al., 2008).

In short, the techniques share some similarities, but there are differences that separates the various methods into the two categories. Much like this thesis, past research has also compared statistical and machine learning techniques, for the purpose of modeling credit ratings. As such, this scope of this thesis is not novel in itself, yet we believe that certain elements of methodology and data characteristics of previous studies could be improved upon.

4.2 Statistical Methods

The first work in the area of bond rating modeling was by Fisher (1959), who used linear regression to explain the variance of risk premiums between bonds, where the risk premiums was defined as the difference between the bond's yield-to-maturity (YTM), and the risk-free interest rate. He hypothesized that a bond's YTM was a function of the issuer's risk of default and the bond's marketability, and that risk of default could be estimated by looking at historical financials, historical debt repayment performance and the debt-to-equity ratio.

Both Horrigan (1966) and West (1970) took this idea a step further and used linear regression to predict credit ratings. Horrigan (1966) considered the six highest rating classes by S&P and Moody's. He used accounting data in a two step approach – he first regressed bond ratings on 15 financial ratios, in order to select the most relevant variables. Subsequently, he regressed the ratings on these selected variables he achieved a 58% accuracy on Moody's ratings and 52% on S&P's ratings.

West (1970) critically commented on the work of Horrigan (1966), as he claimed the original model by Fisher (1959) was more theoretically sound, and replicated Fischer's model, using one additional variable, and achieved a 62% accuracy.

Pinches and Mingo (1973), instead of using linear regression, studied predicting credit ratings using multiple discriminant analysis (MDA), to increase the prediction accuracy. For their final model, they achieved a 60% accuracy.

Kaplan and Urwitz (1979) criticized all the work done up until now, as they all failed to take into account the ordinal nature of bond rating (i.e. that they were groups on an increasing scale), and instead implemented a multivariate probit regression model, and achieved 69% accuracy.

A number of studies has since then been conducted, using statistical models and most of these show prediction accuracy in the range 50-70%, most using 6 rating classes, ignoring the '+' and '-'suffixes of the ratings.

Researchers have tested a number of financial variables, however those proving to be most robust are measures of size, leverage, capital intensiveness, ROI, earnings stability and debt coverage (Sprengers et al., 2006).

Few have managed to surpass the 70% limit mark on prediction accuracy, and one can debate whether simply using 6 classes is sufficient for the prediction to be relevant for real-world applications. Recently, however, a number of researchers have attempted a new approach to predicting credit ratings, through machine learning models, with the aim of improving accuracy and increasing the granularity of the predictions.

4.3 Machine Learning Algorithms

While the earlier work was focused on using traditional statistical methods, in combination with economic theory to model ratings, Dutta and Shekhar (1988) were some of the first to use machine learning techniques for this purpose. Since then, multiple research papers have been published on the topic, each exploring different techniques and nuances.

The early work of Dutta and Shekhar (1988) used neural networks to predict credit ratings, and found that these outperformed their traditional statistical counterparts. They found that their neural network model was able to attain a 83% accuracy in identifying "AA" from "non-AA" ratings. They used linear regression as a benchmark case, which achieved less than 50% accuracy.

Surkan and Singleton (1990) also used neural networks, but on two groups of rating classes by Moody's, and achieved an 88% accuracy, which significantly outperformed benchmark tests using multiple discriminant analysis.

Kim et al. (1993) compared a backpropagating neural network model against benchmark models using linear regression, multiple discriminant analysis, rule-based systems and logistic regression, and found that the neural network far outperformed the other models when distinguishing between 6 ratings classes, scoring 55% accuracy, where the other models all scored around 40%. This was a very interesting study due to the number of different models investigated. However, the sample consisted of a mere 168 ratings.

Generally, neural networks have often been suggested models, when it comes to modeling credit ratings. Other noteworthy studies, such as Moody and Utans (1994), Kumar and Bhattacharya (2006), and Frank (2009), have all proven that machine learning, and in particular neural networks, perform well for this purpose, and are more effective than classical statistical approaches.

As mentioned, there exists a large number of machine learning models other than artificial neural networks, which are useful for predicting credit ratings. More recently, a number of other algorithms, have been investigated for the purpose of predicting credit ratings, one of them being support vector machines (SVM).

Huang et al. (2004) showed that the performance of neural network and SVM models are in fact comparable, and that SVM models in some cases outperform neural networks, when tested on the same datasets. One of their conclusions was that since SVM models are less computationally expensive to train than neural networks, while delivering similar performance, SVM's could be preferable in practice than ANNs.

Ye et al. (2008) investigates the accuracy of two different types of SVM algorithms to predict over 19 different classes of credit ratings, and achieves impressive results, with up to 64% accuracy – considerable improvements over their benchmark algorithms, which are bagged decision trees and a probit model. Lee (2007) also investigates SVM on a large Korean dataset, and achieves a 78% accuracy, predicting over 5 classes of ratings.

Wu et al. (2014) combine multiple algorithms in an ensemble classifier, where each model is trained separately, and is then combined into a hybrid, where the four models 'vote' on which class should be assigned to the observation. Predicting over 9 classes of ratings, in a Taiwanese dataset, they achieve an accuracy of 60%. Interestingly however, one of the single models, a bagging decision tree model, actually performs slightly better at 61%. Their finding shows that there is definitely some potential in ensemble classifiers, however that single-approach models with a boosting or bagging technique, can be more accurate.

From the literature review, it is clear that machine learning has been extensively applied to the credit rating problem, and that it often offer accuracy improvements over classical statistical techniques. However, comparing rating accuracies across studies can be hard, given that each use different datasets and different numbers of rating classes.

4.4 Conclusion

In Table 4.1, a selection of previous studies, their modeling approaches and the results achieved are summarised. We see that some of these have impressive prediction accuracies.

Here, it is important that we consider the sample sizes used, and the number of classes they predict. Generally, a lower number of classes is easier to predict and thus yields higher model accuracy. In this study, an attempt is made on having as many categories as possible, in order to maintain the relevance of the findings.

Table 4.1: Selected Past Studies

Author	Major method	Dataset	Samples	Classes	Prediction Accuracy (%)
Horrigan (1966)	OLS	US	352	6	59
West (1970)	OLS	US	313	6	62
Pinches and Mingo (1973)	MDA	US	180	5	60
Kaplan and Urwitz (1979)	Probit	US	327	6	69
Dutta and Shekhar (1988)	ANN	US	47	2	88
Surkan and Singleton (1990)	ANN	US	146	2	88
Kim et al. (1993)	ANN	US	168	6	55
Moody and Utans (1994)	ANN	US	797	16	30
Huang et al. (2004)	SVM	S. Korea	74	5	80
Kumar et Bhattacharya (2006)	ANN	US	129	6	79
Lee (2007)	SVM	S. Korea	3,017	5	78
Ye et al. 2008	SVM	US	1,570	19	64
Frank (2009)	ANN	US	153	4	75
Wu et al. (2014)	Ensemble	Taiwan	11,616	9	60

4.5 Summary

In this chapter, we have provided some background on the problem of predicting credit ratings, and reviewed the past literature written on the subject, both with respect to statistical techniques and machine learning algorithms. We have also provided some perspective on the weaknesses in some of these studies, and how this study attempts to differentiate itself from this already-conducted research.

5 Methodology

This chapter describe testing setup, and the methodology used for training and evaluating the different models. Additionally it describes which specific models, we evaluate. First, we go through the setup. We then describe how we find and set optimal hyperparameters for the models. We then describe the modeling process, and lastly we elaborate on the testing procedure for evaluating the individual models' performance.

5.1 Setup

After extracting the data with Bloomberg and Excel, data processing, cleaning and subsequent modeling was done using Python. The primary packages used were `pandas`[1], `scikit-learn`[2] and `scikit-optimize`[3].

`pandas` was used for data manipulation, cleaning, and for making tables and figures.

`scikit-learn` was used for model training and evaluation. It is an open-source package for Python, which includes a number of different machine learning algorithms and support functions.

For hyperparameter optimization the `scikit-optimize` package was used. This package has a number of different optimization functions, but for the purposes of this thesis, Gaussian Processes optimization was used.

5.2 Hyperparameter Optimization

When training machine learning models, the models themselves have input parameters, which affect their behaviour, which are referred to as hyperparameters. Tune a model's hyperparameters, for maximum model performance, can be very difficult and time-consuming.

Naturally, one could perform a complete search of the input space, to find the optimal parameters, however, for some models, there are sometimes more than 5 hyperparameters that can be adjusted, rendering the input space massive. This, in combination with machine learning algorithms generally being computationally expensive functions, makes this an impractical method of optimization. Instead, one can attempt to model the model evaluation measure as a function of the hyperparameter input space, and thereby do an approximated optimization of the hyperparameters.

This is the method we have chosen to employ in this thesis for selecting our hyperparameters. We use a technique called Gaussian Processes, which optimizes for the expected improvement in cross-entropy of the prediction and actual output vector. In practice this allows us to rather efficiently optimize the models, without spending multiple hours waiting for an exhaustive grid search to compute. If we had done a full grid search, we could

[1] http://pandas.pydata.org/

[2] http://scikit-learn.org/

[3] https://github.com/scikit-optimize

potentially have obtained better hyperparameters, but they might have been the same as the ones found by the process described.

Before training each model, we use each of the training sets, and find 3 optimal sets of hyperparameters for each of the different rating classes. We then store this, and use it as needed when training the models.

5.3 Model Training

Having identified the optimal hyperparameters for a given algorithm, we proceed to train the different statistical and machine learning models.

The statistical methods and machine learning method we will use in this thesis are:
Statistical Methods

- Logistic Regression
- Multiple Discriminant Analysis

Machine Learning Algorithms

- k-Nearest Neighbors
- Support Vector Machines
- Artificial Neural Networks
- ExtraTrees (Decision-tree algorithm with bagging)
- AdaBoost (Decision-tree algorithm with boosting)

On a practical level, we for each rating class group instantiate a classifier of the given type and pass them its unique configuration from the hyperparameters previously found. We then pass them the training datasets, consisting of both features and labels.

After the models have completed training we proceed to model evaluation.

5.4 Model Evaluation

When evaluating the data we evaluate them on four different measures, both in-sample (i.e. with the training data), and out-of-sample (i.e. test data). On a practical level, we pass them the features of the labels it should predict. This returns a vector of predictions, which we can then use to benchmark against the true labels.

The four measures we use to compare the models on are Accuracy, F1 Score, Cohen's Kappa and "1-off Accuracy". 1-off accuracy is simply the accuracy of where the prediction is at most one rating away from the true rating.

After evaluating the models, we present evaluation metrics for each model, as well as a confusion matrix for the "Class I" grouping, showing the predicted and true ratings in a matrix, as a percentage of the class total.

5.5 Summary

In this chapter we have described our setup for doing the testing, which we perform using the Python packages pandas and sklearn. We have elaborated on how we select and set hyperparameters for each of the models, which we do using scikit-optimize and Gaussian Processes optimization. We then described the modeling process and which models we will be attempting. Lastly we have gone through how we will evaluate the model performance using four different metrics.

6 Data

6.1 Introduction

In this chapter, we describe the data selection and collection methodology. We specify the data that we need to collect and then describe the steps that we go through to obtain our raw dataset. Next, we describe how we process the raw data, before we use it for modelling.

We have attempted to thoroughly document how data was collected in order to preserve transparency, and for helping future researchers validate or expand on this study. We feel it is important to stress that, to our knowledge, we managed to extract a rather unique dataset. Compared to similar studies within the same field, we have not found others who match our dataset with respect to the number of observations, for U.S. ratings.

Subsequently, we describe the data that we have collected, and present selected results from the pre-processing stages.

6.2 Data Selection and Collection

The data we need in order to answer the research question is the credit ratings over a given time period, and relevant accounting figures of the issuer at the time of rating. Having the issuer's accounting figures allows us to observe factors likely to affect credit ratings or signal creditworthiness. This raw data is then further processed, prior to being used for modeling. The exact data points we extract are detailed below, and have been selected based on information from the CRAs and past literature on the topic.

The data was collected from Bloomberg, in part through the terminal application and in part through the Excel plug-in. We used Bloomberg's RATC function to list all ratings published by S&P, Moody's, and Fitch in the period 1 January 2010 - 1 September 2016. Only U.S. companies were included in the search, and we deliberately excluded financial companies and government entities. The reason for doing so is that financial companies often have radically different accounting practices than companies from other industries, and that the rating of government entities would be significantly influenced by the possibility of the U.S. government covering some obligations in the event of bankruptcy.

Having collected these ratings, we needed to know which companies that were publicly traded at the time of rating, as detailed financial information would likely not be available for private companies, due to lower reporting requirements. To determine this, we extracted the last observed stock price at the date of the rating for the issuers in question. If the company hadn't been publicly traded within the preceding 30 days, the price would not return a value, and hence we could determine whether the company had been traded or not.

After excluding non-public firms from the list of ratings, we enriched it with financial data from the issuers' most recent quarterly financial statements, preceding the rating.

The specific data that the ratings were enriched with are presented in Table 6.1.

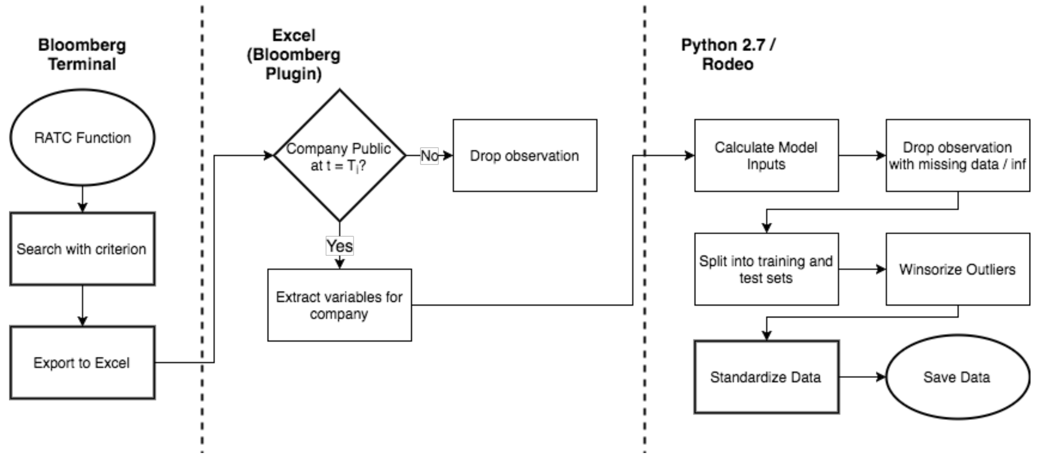
Table 6.1: Add caption

Cash Flow Statement	**Balance Sheet**	**Income Statement**
Trailing 12M Free Cash Flow	Total Common Equity T-1 year	Trailing 12M Earnings for Common Equity
Trailing 12M Cash From Operations	Short-term Debt	Trailing 12MNet Income
	Long-term Debt	Trailing 12M Net Sales
	Current Assets	Trailing 12M Operating Income
	Current Liabilities	Trailing 12M Pre-tax Income
	Average Total Assets	Trailing 12M Amortization and Depreciation
	Average Total Invested Capital	Trailing 12M Total Interest Expense
	Cash and Near-cash Items	Earnings Per Share
	Marketable Securities and Other Short-term Investments	
	Account Receivables	
	Short and Long Term Debt	
	Total Assets	
	Total Invested Capital	
	Total Common Equity	
	Total Equity	
	Total Liabilities	
	Net Fixed Assets	

Having isolated the final sample, containing both the ratings and financial information about the issuers, we can proceed with processing the data further through calculating the actual model inputs and cleaning the data.

The data collection procedure as described is illustrated in Figure 6.1.

Figure 6.1: Data Collection Process



This concludes our data collection. We now have the ratings for the study period, and the associated financial data for the same period as the rating.

6.3 Data Preparation

In the data preparation step, we prepare the raw data that we have collected to be used in the models. This consists of six parts. First, we calculate the financial ratios. Then, we clean the data. We then group the ratings into categories of different granularity. We then split the data into training and test sets. Subsequently we winsorize the data, in order to eliminate the effect of outliers on model performance. Lastly, we standardize all data, through demeaning and scaling.

6.3.1 Calculating financial ratios

In order to make the data comparable across companies of different scales, we calculate ratios, instead of using the financial data directly. We look at four different categories of financial ratios; liquidity, profitability, cash flow adequacy and capital structure. For inspiration on which input to include in the models, we considered information from both the CRAs, as well as from other researchers in the topic, such as Frank (2009), Lee (2007), Kumar and Bhattacharya (2006) and Kim (2005).

The ratios that we calculate using the raw data are presented in Table 6.2, and the formulae for the calculated inputs are shown in Appendix A.

Table 6.2: Calculated inputs

Liquidity	**Profitability**	**Cash Flow Adequacy**	**Capital Structure**	**Other**
Cash Ratio	Net Income	Solvency Ratio	Debt Ratio	log(Sales)
Quick Assets to Total Asets	Operating Income	Interest Expense to Sales	Common Equity to Total Invested Capital	
Current Ratio	Operating Margin	Times Interest Earned	Debt to Equity	
Current Assets to Total Assets	Pre-Tax Margin	Debt Coverage	Long-term Debt to Capital	
	After Tax Profit Margin	Operating Cash Flow to Sales	Total Equity to Total Assets	
	Asset Turnover	Cash Flow Return on Assets	Long-term Deb to Fixed Assets	
	Return on Invested Capital Before Tax			
	Return on Equity			
	Retun on Assets			
	Profitability Ratio			
	Sales to Net Worth			

6.3.2 Data cleaning

After calculating the ratios, we do a number of data cleaning operations, in order to have a dataset which is sufficiently clear of errors.

We start by cleaning up the text data that are the ratings themselves. Some of these have extra information, such as whether the rating was unsolicited or details about credit outlook, which is irrelevant for the purposes of this thesis. Thus, we remove this additional data. Additionally we remove observations of ratings classes that are outside the scope of this thesis, such as the ratings that indicate a withdrawn rating, no rating or some type of default (”WR”, ”NR”, ”D” and ”SD”). Lastly, we removed any observations, which had missing data, or any data, which had missing values or NaN as a value.[1]

[1] This occurred as a consequence of calculating the ratios, when e.g. a division-0 error occurred

6.3.3 Grouping the ratings

One worry with regards to modeling each class of credit ratings, is that there are some classes, for which there are few observations. For instance the rating 'AAA', there are just a handful of companies who are given that rating, such as Microsoft and Johnson & Johnson, as of December, 2016.

Having these underrepresented classes can be a problem for machine learning algorithms. There are some different strategies for overcoming them, but we chose to simply group the rating with low class frequency with its nearest class. Additionally, we want to be able to compare our results with the results of those papers with a lower number of classes, and therefore we group the ratings into three sets of groups; 'Class I', 'Class II' and 'Class III', going from most to least granularity. The specific class grouping can be see in Table 6.3.

Table 6.3: Ratings Class Groupings

Credit Rating	**Class-I**	**Class-II**	**Class-III**
AAA, AA+, Aaa, Aa1	1	1	1
AA, Aa2	2		
AA-,Aa3	3		
A+, A1	4	2	
A, A2	5		
A-, A3	6		
BBB+, Baa1	7	3	
BBB, Baa2	8		
BBB-, Baa3	9		
BB+, Ba1	10	4	2
BB, Ba2	11		
BB-, Ba3	12		
B+, B1	13	5	
B, B2	14		
B-, B3	15		
CCC[*], Caa[*], CC,C, Ca	16	6	

6.3.4 Split the data into training and test sets

After grouping the ratings, we split the data into training and test sets for each of the ratings classes. The training set will be used for training the models, and the test set will act as an out-of-sample trial to evaluate the predictive power of the models.

We use a 80/20 split ratio, and we draw the two subsample pseudo-randomly using a stratified split, meaning that the training and test sets have identical ratings distributions. We do this in order to ensure that all rating groups are populated. This is especially critical for the classes with a low number of observations.

It is crucial that this split is made before winsorization and standardization, as there could be risk of information leakage if it had been done after that. The concept of information leakage is described in further depth in chapter 3.

6.3.5 Winsorizing

In order to handle extreme observations, either as a result of outlier in the raw data, or very small denominators in the calculation of ratios, the data is winzorised. By winsorizing, we do not lose any observations from the dataset. As a result, values that fell outside the 5^{th} and 95^{th} percentile were corrected to the values at each percentile respectively.

We start by winsorizing the training sets. We store the parameters from this and apply them to the test sets.

6.3.6 Standardizing

Standardization is a common requirement for the input data of a number of machine learning models. Hence, we use a scaling function, which removes the mean and scales the data to unit variance. Again, we start by standardizing the training sets, and use these parameters to standardize the test sets.

6.4 Raw Data Overview

The list of credit ratings was downloaded, and we excluded the companies not listed at the date of rating.After this, our dataset consisted of 4,407 individual ratings. We then removed the non-applicable ratings, leaving us with 4,246 samples. Lastly, we removed samples with missing data or errors, which left us with a final dataset consisting of 3,992 ratings. This information is also presented in Table 6.4.

Table 6.4: Number of observations per step

Step	Description	Observations
1	Initial list	4,407
2	Removed non-applicable ratings	4,246
3	Removed observations with missing data	3,992

6.5 Rating Category Summary

Figure 6.2 shows the credit rating frequency distribution for the final dataset of the 3,992 ratings. We see that extremely few companies fall into the AAA and AA rating categories, and the bulk of the mass being in the middle categories.

We then group these ratings into the different classes, as described in subsection 6.3.3. In Table 6.5, 6.6, and 6.7, the distributions of each of these classes are displayed.

Figure 6.2: Distribution of Credit Ratings

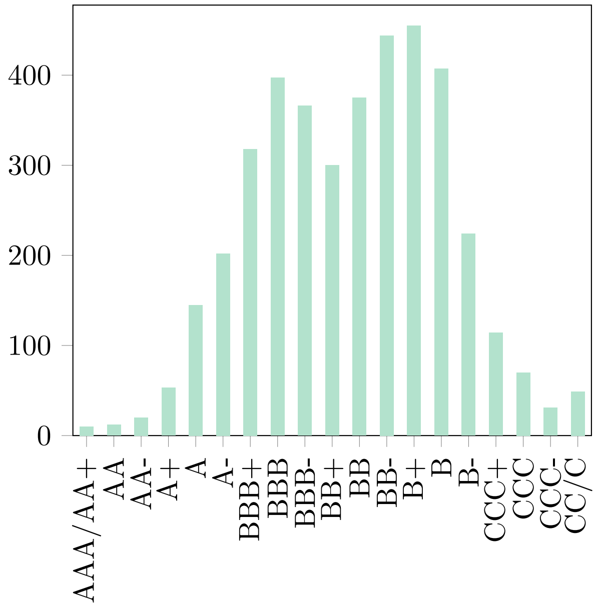


Table 6.5: Rating Class I Distribution

Class-I	Ratings	Rating Count	Percentage
1	AAA/AA+	10	0.3%
2	AA	12	0.3%
3	AA-	20	0.5%
4	A+	53	1.3%
5	A	145	3.6%
6	A-	202	5.1%
7	BBB+	318	8.0%
8	BBB	397	9.9%
9	BBB-	366	9.2%
10	BB+	300	7.5%
11	BB	375	9.4%
12	BB-	444	11.1%
13	B+	455	11.4%
14	B	407	10.2%
15	B-	224	5.6%
16	CCC+,CCC,CCC-,CC,C	264	6.6%

Table 6.6: Rating Class II Distribution

Class-II	Ratings	Rating Count	Percentage
1	AAA/AA+, AA, AA-	42	1.1%
2	A+, A, A-	400	10.0%
3	BBB+, BBB, BBB-	1081	27.1%
4	BB+, BB, BB-	1119	28.0%
5	B+, B, B-	1086	27.2%
6	CCC+,CCC,CCC-,CC,C	264	6.6%

Table 6.7: Rating Class III Distribution

Class-III	Ratings	Rating Count	Percentage
1	AAA/AA+, AA, AA-, A+, A, A-, BBB+, BBB, BBB-	1523	38.2%
2	BB+, BB, BB-, B+, B, B-, CCC+,CCC,CCC-,CC,C	2469	61.8%

6.6 Data Preparation

In order to be able to generalize the results obtained from the models, we preprocess the data by first calculating the model inputs from the raw inputs in our dataset. We then proceed by winsorizing and standardizing the model inputs.

In Table 6.8 we see the summary statistics of the model input data before it has been standardized, yet after it has been winsorized.

Table 6.8: Summary Statistics of Model Input Data

Input Variable	Mean	Std. Dev.	Min	Max	Median
ROE	0.070	0.388	-0.970	0.975	0.099
PROFIT_MARGIN	0.020	0.144	-0.409	0.227	0.044
LT_DEBT_TO_CAP	0.510	0.277	0.094	1.183	0.466
CURRENT_RATIO	1.754	0.942	0.582	4.091	1.517
ROA	0.021	0.081	-0.203	0.146	0.033
OPERATING_MARGIN	0.078	0.145	-0.335	0.305	0.091
PRE_TAX_MARGIN	0.037	0.167	-0.458	0.286	0.061
PRE_TAX_ROIC	0.051	0.139	-0.305	0.290	0.063
CURRENT_TO_TOT_ASSETS	0.324	0.193	0.063	0.705	0.303
QUICK_TO_TOT_ASSETS	0.187	0.123	0.029	0.455	0.165
DEBT_TOT_ASSETS	0.372	0.191	0.091	0.804	0.344
COMMON_EQY_TO_TOT_ASSETS	0.306	0.205	-0.145	0.631	0.323
TOT_EQY_TO_TOT_ASSETS	0.324	0.200	-0.130	0.644	0.343
ASSET_TURNOVER	0.853	0.570	0.203	2.257	0.693
SOLVENCY_RATIO	0.120	0.136	-0.184	0.407	0.111
SALES_TO_TOT_EQY	2.661	3.474	-3.757	12.220	1.718
CASH_RATIO	0.441	0.449	0.017	1.662	0.281
DEBT_TO_EQY	1.147	2.002	-3.215	6.873	0.799
TOT_EQY_AND_LT_DEBT_TO_FIXED	4.395	5.182	0.763	20.350	2.069
INT_EXPENSE_TO_SALES	0.040	0.048	-0.015	0.175	0.023
TIMES_INT_EARNED	4.483	10.458	-16.282	33.493	2.726
CF_TO_DEBT	0.364	0.373	-0.016	1.510	0.238
CF_TO_SALES	0.156	0.126	-0.011	0.468	0.124
CASH_RETURN_ON_ASSETS	0.091	0.055	-0.007	0.207	0.086
LOG_SALES	8.098	1.350	5.814	10.647	8.061

6.6.1 Winsorizing

When winsorizing, we store the winsorization parameters for the training set, and apply these to the test sets. In Table 6.9, we show the winsorization parameters.

Table 6.9: Winsorization parameters

	Lower Bound	Upper Bound
ROE	-0.970	0.975
PROFIT_MARGIN	-0.409	0.227
LT_DEBT_TO_CAP	0.094	1.183
CURRENT_RATIO	0.582	4.091
ROA	-0.203	0.146
OPERATING_MARGIN	-0.335	0.305
PRE_TAX_MARGIN	-0.458	0.286
PRE_TAX_ROIC	-0.305	0.290
CURRENT_TO_TOT_ASSETS	0.063	0.705
QUICK_TO_TOT_ASSETS	0.029	0.455
DEBT_TOT_ASSETS	0.091	0.804
COMMON_EQY_TO_TOT_ASSETS	-0.145	0.631
TOT_EQY_TO_TOT_ASSETS	-0.130	0.644
ASSET_TURNOVER	0.203	2.257
SOLVENCY_RATIO	-0.184	0.407
SALES_TO_TOT_EQY	-3.757	12.220
CASH_RATIO	0.017	1.662
DEBT_TO_EQY	-3.215	6.873
TOT_EQY_AND_LT_DEBT_TO_FIXED	0.763	20.350
INT_EXPENSE_TO_SALES	-0.015	0.175
TIMES_INT_EARNED	-16.282	33.493
CF_TO_DEBT	-0.016	1.510
CF_TO_SALES	-0.011	0.468
CASH_RETURN_ON_ASSETS	-0.007	0.207
LOG_SALES	5.814	10.647

6.7 Summary

In this chapter, we have seen how the initial dataset has been obtained, enriched with financial information, reduced from 4,407 ratings to 3,902 ratings through cleaning, and subsequently been categorized into the different class groups of ratings that will be used for modeling. The processes of obtaining a dataset of such size and quality, has been an iterative process and a large part of this thesis, so it is our hope that this documentation is useful.

In the next chapter, we will present the results from training and testing the models on the data obtained in this chapter.

7 | Analysis

7.1 Introduction

In this section, we present the results obtained through following the methodology described in chapter 5, and investigate the performance of the different models and techniques described in chapter 3. We consider each algorithm separately, and then contrast the findings to previous research, and the relative performance to the other algorithms applied in this thesis. Furthermore, we display the hyperparameters from our optimizations and discuss these briefly. Lastly, we summarize by presenting an overview of the results obtained and discuss high-level implications.

7.2 Data and Analysis of Results

7.2.1 Statistical Methods

Logistic Regression

For Logistic Regression there are no hyperparameters to tune, and as such we proceed directly to the model results. Looking at the results, we see that the Logistic Regression achieves an accuracy of 27.4% in the ”Class I” grouping, with a Kappa of .2, meaning a relatively low rate of agreement between the predicted and actual ratings. The most relevant study to compare our findings on Logistic Regression to would be Kaplan and Urwitz (1979), who achieved a 69% accuracy rating over 6 classes, using a probit regression. We achieved roughly 54% accuracy, also predicting the 6 classes (”Class II” grouping), so significantly lower. It is unclear what the exact cause of this is, yet can most probably be attributed to a difference in methodology.

In Table 7.1 we see a confusion matrix, showing the actual rating in the rows, and the predicted rating in the columns. Here, we see that the model make substantial classification errors. For instance, the observations with an actual rating in group 1, were predicted to be in group 5. Again, it is unclear what the exact cause of this is, but it could tell us that there are some structural aspects of the data, that makes it unsuited for logistic regression.

The Logistic Regression model that we made, have a sub-par performance, compared to both our other models, and past research. As such, it is a less useful model to base conclusions upon.

Table 7.1: Confusion Matrix for Logistic Regression

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.25	0.00	0.25	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.18	0.45	0.09	0.00	0.18	0.00	0.00	0.00	0.00	0.00	0.09	0.00	0.00
5	0.00	0.00	0.00	0.00	0.38	0.07	0.17	0.28	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.07	0.15	0.28	0.38	0.05	0.07	0.00	0.00	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.02	0.05	0.05	0.23	0.41	0.17	0.03	0.00	0.02	0.02	0.02	0.00	0.00
8	0.00	0.00	0.00	0.01	0.03	0.05	0.20	0.38	0.15	0.01	0.05	0.09	0.03	0.00	0.00	0.00
9	0.00	0.00	0.00	0.01	0.00	0.05	0.11	0.33	0.21	0.10	0.04	0.07	0.05	0.03	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.05	0.18	0.12	0.18	0.02	0.13	0.23	0.05	0.03	0.00	0.00
11	0.00	0.00	0.00	0.00	0.00	0.03	0.07	0.07	0.12	0.07	0.13	0.29	0.17	0.05	0.00	0.00
12	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.02	0.10	0.02	0.08	0.28	0.29	0.08	0.01	0.06
13	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.01	0.03	0.02	0.10	0.22	0.30	0.24	0.01	0.03
14	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.02	0.00	0.05	0.07	0.30	0.41	0.04	0.06
15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.04	0.24	0.33	0.13	0.22
16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.17	0.08	0.70

Table 7.2: Metrics for Logistic Regression

	Accuracy	Kappa	F1 Score	1-off Accuracy
Class I	0.274	0.201	0.261	0.636
Class II	0.539	0.386	0.534	0.966
Class III	0.845	0.670	0.844	1.000

Multiple Discriminant Analysis

For Multiple Discriminant Analysis there are no hyperparameters to tune, and as such we proceed directly to the model results. The results are presented in Table 7.3.

We achieve an accuracy of 26.9% in the "Class I" grouping, which is a reasonable accuracy. Comparing the "Class II" accuracy of 53% to that of Pinches and Mingo (1973), who obtain 60%, we again see that we have achieved a lower score for a comparable model. We suspect that the disparity is due to methodological differences.

Looking at the confusion matrix in Table 7.4, we also see that the model despite having a relatively high 1-off accuracy measure struggles with certain groups, and makes predictions multiple notches away from the actual rating. In this case, the model seems to be heavily influenced by the distribution of the training data, predicting many ratings to be in the most frequent classes.

Furthermore, considering the Kappa measures, the model does not perform well and actually falls below 0.2, indicating a low degree of agreement between our predicted and actual ratings.

To summarize, the MDA model does not perform well under our setup, as it has consistently low measures, and appears to have a low degree of precision, as it predicts ratings to be multiple notches away from the true rating. Furthermore, our performance metrics are lower than what should be expected, considering past research.

Table 7.3: Metrics for MDA

	Accuracy	Kappa	F1 Score	1-off Accuracy
Class I	0.269	0.197	0.264	0.618
Class II	0.528	0.374	0.525	0.962
Class III	0.839	0.656	0.838	1.000

Table 7.4: Confusion Matrix for MDA

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.50	0.00	0.25	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.09	0.00	0.09	0.36	0.09	0.00	0.27	0.00	0.00	0.00	0.00	0.09	0.00	0.00	0.00
5	0.00	0.03	0.03	0.03	0.28	0.07	0.21	0.28	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.10	0.10	0.17	0.25	0.28	0.05	0.05	0.00	0.00	0.00	0.00	0.00	0.00
7	0.00	0.02	0.00	0.02	0.05	0.03	0.28	0.36	0.14	0.03	0.05	0.03	0.00	0.00	0.00	0.00
8	0.00	0.01	0.00	0.01	0.05	0.04	0.25	0.34	0.13	0.04	0.03	0.08	0.03	0.00	0.00	0.00
9	0.00	0.00	0.00	0.03	0.00	0.08	0.08	0.29	0.16	0.11	0.07	0.11	0.04	0.00	0.03	0.00
10	0.00	0.00	0.00	0.00	0.00	0.03	0.18	0.15	0.17	0.05	0.13	0.23	0.03	0.02	0.00	0.00
11	0.00	0.00	0.00	0.03	0.00	0.03	0.05	0.08	0.11	0.04	0.19	0.31	0.12	0.04	0.01	0.00
12	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.03	0.07	0.02	0.12	0.33	0.21	0.10	0.06	0.01
13	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.02	0.03	0.02	0.12	0.22	0.25	0.23	0.02	0.04
14	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.02	0.00	0.06	0.09	0.28	0.39	0.06	0.05
15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.07	0.20	0.22	0.22	0.24
16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.21	0.13	0.57

7.2.2 Machine Learning Models

kNN

For the k-Nearest Neighbor algorithm, there is one hyperparameter to optimize for; n_neighbors. n_neighbors is how many of the nearest points the algorithm should choose for evaluating which class the test example should be predicted to be. Here it is in each case 1, meaning that it only looks at the single, closest example in the training set, in order to predict the class of the test case.

We see that kNN performs rather well, with an accuracy of 33%, and quite better than the statistical methods. There is no directly relevant comparison in the past studies we have considered, but a 65% accuracy in the "Class II" grouping, is comparable, albeit a little lower, to that achieved by Kaplan and Urwitz (1979). Looking at at 1-off acurracy for the "Class II" grouping, we see that we achieve a 96% accuracy, which can be considered impressive, when taking into account the simplistic and non-parametric nature of the kNN algorithm.

The kNN algorithm seems to offer a convenient and easy-to-understand method for predicting credit ratings, and achieves a descent accuracy. The results we obtain are not directly comparable with past studies we've considered but fall within the range of results achieved for machine learning algorithms.

Table 7.5: Optimal Hyperparameters for the kNN Algorithm

	n_neighbors
Class I	1
Class II	1
Class III	1

Table 7.6: Confusion Matrix for kNN Algorithm

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0.50	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.50	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.25	0.50	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.09	0.09	0.00	0.00	0.45	0.09	0.09	0.09	0.00	0.00	0.00	0.00	0.09	0.00	0.00	0.00
5	0.00	0.03	0.00	0.10	0.31	0.28	0.14	0.07	0.03	0.00	0.00	0.03	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.23	0.28	0.30	0.05	0.05	0.05	0.00	0.00	0.05	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.08	0.22	0.31	0.20	0.11	0.08	0.00	0.00	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.01	0.01	0.03	0.16	0.38	0.27	0.05	0.01	0.03	0.03	0.01	0.01	0.00
9	0.00	0.00	0.00	0.00	0.00	0.01	0.07	0.29	0.32	0.16	0.08	0.03	0.01	0.03	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.03	0.03	0.10	0.08	0.33	0.23	0.08	0.07	0.02	0.02	0.00
11	0.00	0.00	0.00	0.00	0.01	0.00	0.03	0.01	0.09	0.13	0.37	0.23	0.09	0.03	0.00	0.00
12	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.02	0.07	0.04	0.19	0.29	0.16	0.10	0.06	0.02
13	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.05	0.03	0.08	0.26	0.34	0.15	0.04	0.01
14	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.02	0.07	0.04	0.05	0.22	0.33	0.21	0.02
15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.02	0.02	0.11	0.13	0.36	0.07	0.24
16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.04	0.13	0.15	0.66

Table 7.7: Metrics for kNN Algorithm

	Accuracy	Kappa	F1 Score	1-off Accuracy
Class I	0.333	0.269	0.332	0.731
Class II	0.648	0.536	0.649	0.962
Class III	0.877	0.742	0.878	1.000

SVM

For the SVM algorithm there are only two parameters to optimize for; C and gamma. For SVMs there are a number of kernel functions available. We chose the RBF kernel, due to its speed and accuracy, compared to other functions. The RBF is a non-linear kernel, which seems like an appropriate choice considering the data.

Using SVM as an algorithm we obtain high metrics on all groupings, which can be seen in Table 7.10. For "Class I" we obtain an accuracy of 37% and a 1-off accuracy of 77%. Looking at the confusion matrix in Figure 7.9, we also see that the majority of prediction seems to be grouped relatively tighly around the diagonal, meaning that for the most part, when the model does make a wrong prediction, it is not many notches off. There are however some ratings, especially those in the ends of the scale which seem troublesome for the SVM algorithm. For instance, ratings belonging to Group 2 are predicted to be in group 4 and 6.

Whereas the accuracy of the SVM algorithm, being higher than the kNN algorithm in the "Class I" grouping, it is about the same for the "Class II" grouping. The most relevant study to compare this to would be that of Ye et al. (2008), who achieves a significantly higher accuracy of 64% across 19 categories, so in that regards, our results are disappointing. Ye et al. (2008) creates industry-specific models, include more variables and for multiple historical years. Implementing these into our methodology would likely have generated results comparable to Ye et al. (2008).

To summarize, whereas the SVM algorithm achieves a higher accuracy than the models tested before this, there appears to room for improvement, due to the higher accuracies found in comparable studies.

Table 7.8: Optimal Hyperparameters for the SVM Algorithm

	C	gamma
Class I	20.000	0.160
Class II	12.296	0.144
Class III	20.000	0.062

Table 7.9: Confusion Matrix for SVM Algorithm

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0.50	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.50	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.25	0.50	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.09	0.00	0.00	0.64	0.09	0.00	0.09	0.00	0.00	0.00	0.00	0.09	0.00	0.00	0.00
5	0.00	0.03	0.00	0.10	0.24	0.38	0.03	0.14	0.03	0.00	0.00	0.03	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.15	0.28	0.25	0.12	0.10	0.03	0.03	0.03	0.03	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.05	0.16	0.39	0.20	0.08	0.05	0.05	0.03	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.00	0.01	0.01	0.23	0.42	0.24	0.01	0.03	0.03	0.03	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.21	0.38	0.12	0.07	0.07	0.00	0.04	0.01	0.00
10	0.00	0.00	0.00	0.00	0.00	0.03	0.02	0.05	0.10	0.37	0.22	0.18	0.03	0.00	0.00	0.00
11	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.03	0.08	0.12	0.41	0.27	0.05	0.01	0.00	0.00
12	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.04	0.04	0.19	0.35	0.22	0.07	0.02	0.01
13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.02	0.10	0.31	0.37	0.14	0.02	0.00
14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.04	0.04	0.06	0.29	0.30	0.20	0.02
15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.09	0.20	0.36	0.09	0.22
16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.02	0.11	0.09	0.74

Table 7.10: Metrics for SVM Algorithm

	Accuracy	Kappa	F1 Score	1-off Accuracy
Class I	0.367	0.305	0.364	0.767
Class II	0.646	0.531	0.645	0.961
Class III	0.889	0.764	0.889	1.000

Artificial Neural Networks

Continuing to Artificial Neural Networks, we here do not optimize hyperparameters, as neural networks are incredibly computationally expensive functions, and it would simply be impractical for the level of hardware used for this thesis. This is of course disappointing, considering the potential that ANNs holds, when looking at the past literature. Instead, we attempted to manually tune the hyperparameters, using steps described by Frank (2009). As such, we cannot reasonably expect the ANN models we created to perform at their peaks, as our parameters are likely sub-optimal. The specific values we chose can be seen in Table 7.11.

We see that our network consists of one input layer of 500 nodes, and then another 5 layers with 500 nodes each, making it a so-called "deep" neural network. Through manual hyperparameter tuning, we found that the 'ReLU' activation function (Rectified Linear Unit), showed the highest in-sample performance, and thus this was chosen for the neural network model.

In Table 7.13, we see the performance metrics for ANN, showing an accuracy of 37% in "Class I". Performing at the same level as ExtraTrees, ANN is the most accurate algorithm, when considering "Class I"-groupings. However, when it comes to Group II and III, ExtraTrees, turns out to be more effective. Comparing it to past research, it would probably be the best comparison to compare it to the study by Kumar and Bhattacharya (2006), who achieved an accuracy if 79% when predicting over 6 classes – significantly higher than our "Class II" measure of 66%. Our hypothesis would be that this discrepancy is due to the lack of proper hyperparameter optimization.

In summary, the ANN model is the most accurate one in our sample, when looking at the performance metrics of the "Class I" groupings. However, it leaves a lot to be desired, as we do not achieve as strong results as past researchers, and as we are unable to properly optimize the model.

Table 7.11: Hyperparameters for the ANN Algorithm

	activation	beta_1	beta_2	epsilon	hidden_layer_sizes	max_iter
Class I	relu	0.9999	0.98	1e-08	(500, 500, 500, 500, 500, 500)	1000
Class II	relu	0.9999	0.98	1e-08	(500, 500, 500, 500, 500, 500)	1000
Class III	relu	0.9999	0.98	1e-08	(500, 500, 500, 500, 500, 500)	1000

Table 7.12: Confusion Matrix for ANN Algorithm

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.50	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.75	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.27	0.45	0.00	0.00	0.09	0.18	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.10	0.48	0.17	0.17	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.30	0.28	0.25	0.10	0.03	0.05	0.00	0.00	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.08	0.09	0.27	0.30	0.16	0.06	0.02	0.03	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.00	0.03	0.03	0.11	0.46	0.27	0.05	0.03	0.03	0.01	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.01	0.05	0.22	0.44	0.15	0.05	0.04	0.01	0.01	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.03	0.23	0.35	0.20	0.07	0.07	0.03	0.00	0.00
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.09	0.16	0.40	0.28	0.04	0.00	0.00	0.00
12	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.07	0.04	0.25	0.33	0.15	0.11	0.01	0.01
13	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.04	0.12	0.25	0.27	0.22	0.03	0.02
14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.01	0.04	0.07	0.29	0.34	0.16	0.04
15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.11	0.47	0.20	0.18
16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.06	0.08	0.17	0.68

Table 7.13: Metrics for ANN Algorithm

	Accuracy	Kappa	F1 Score	1-off Accuracy
Class I	0.368	0.307	0.365	0.780
Class II	0.660	0.552	0.655	0.985
Class III	0.890	0.767	0.890	1.000

ExtraTrees Algorithm

The ExtraTrees algorithm is, a decision-tree based ensemble algorithm which implements a meta estimator that fits a number of randomized decision trees on various sub-samples of the dataset and uses parameter averaging to improve the predictive accuracy (scikit-learn, 2017). In short, it is simply a bagged decision-tree model.

We obtain the optimized hyperparameters listed in Table 7.14 for the three class models.

Table 7.14: Optimal hyperparameters for ExtraTrees algortihm

	max_depth	max_features	min_impurity_split	min_samples_leaf	min_samples_split	n_estimators
Class I	57	20	1.000e-01	1	2	80
Class II	78	25	1.000e-01	1	2	100
Class III	100	25	1.000e-09	1	2	100

We find that the ExtraTrees algorithm performs quite well compared to the other models in our sample, and achieves an accuracy level of 37% for the "Class I" grouping, and 70% for the "Class II"- grouping. Whereas we don't have an exact comparison from past literature, we see that our results are close to those of Kumar and Bhattacharya (2006).

Table 7.15: Metrics for ExtraTrees Algorithm

	Accuracy	Kappa	F1 Score	1-off Accuracy
Class I	0.368	0.307	0.364	0.822
Class II	0.697	0.599	0.697	0.985
Class III	0.915	0.819	0.915	1.000

Interestingly, the ExtraTrees algorithm is a rather simplistic one, and it is interesting that it is possible to achieve such high performance figures with this model. Compared to more advanced and computationally expensive models, that we have tested, ExtraTrees is performing exceptionally well.

To summarize, ExtraTrees is one of the highest performing machine learning algorithms in our tests, and its results is comparable to that of past research. The fact that a relatively simple model can yield this performance, seems to indicate that there are some clear patterns in the data, which can be extracted and used for prediction, without extensive effort.

AdaBoost

The optimized hyperparameters for the AdaBoost algorithms are presented in 7.16, being the learning_rate for the optimization function and the number of weak learners (n_estimators).

In Table 7.18 we present results for the AdaBoost algorithm. Here we see that the model seems to perform worse than any of the other models, with only a 21% accuracy score. This can possibly be attributed to the complex relationships between the variables and also the high dimensionality of the input space. As AdaBoost is an ensemble method, which build a large number of "weak learners", and then combines these, it could be hard for it to correctly represent the complex relationships between the variables.

Table 7.16: Optimal Hyperparameters for the AdaBoost Algorithm

	learning_rate	n_estimators
Class I	0.648040	38.0
Class II	0.314516	57.0
Class III	1.125911	226.0

Table 7.17: Confusion Matrix for AdaBoost Algorithm

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00	0.25	0.25	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.27	0.00	0.00	0.55	0.00	0.00	0.00	0.09	0.09	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.31	0.03	0.03	0.52	0.00	0.07	0.03	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.12	0.00	0.03	0.70	0.00	0.03	0.00	0.10	0.00	0.00	0.03	0.00
7	0.00	0.00	0.00	0.00	0.03	0.02	0.09	0.67	0.02	0.05	0.02	0.09	0.02	0.00	0.00	0.00
8	0.00	0.00	0.00	0.00	0.04	0.00	0.05	0.70	0.01	0.05	0.10	0.03	0.01	0.00	0.01	0.00
9	0.00	0.00	0.00	0.00	0.01	0.01	0.04	0.53	0.04	0.08	0.14	0.04	0.04	0.01	0.03	0.01
10	0.00	0.02	0.00	0.00	0.02	0.02	0.05	0.40	0.07	0.07	0.17	0.08	0.07	0.00	0.05	0.00
11	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.37	0.04	0.04	0.15	0.11	0.13	0.03	0.07	0.04
12	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.16	0.04	0.04	0.13	0.20	0.12	0.03	0.17	0.08
13	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.09	0.00	0.04	0.18	0.22	0.15	0.05	0.19	0.07
14	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.00	0.02	0.07	0.16	0.17	0.07	0.32	0.15
15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.04	0.04	0.13	0.00	0.44	0.29
16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.19	0.02	0.32	0.45

Table 7.18: Metrics for AdaBoost Algorithm

	Accuracy	Kappa	F1 Score	1-off Accuracy
Class I	0.213	0.136	0.177	0.522
Class II	0.449	0.273	0.426	0.931
Class III	0.866	0.715	0.866	1.000

7.2.3 Comparison of Results

In Table 7.19, we present the accuracy measured across different models and class groupings. Measuring on accuracy, we see that the ExtraTrees algorithm is the highest scoring in all groupings, yielding accuracy metrics of 36.8%, 69.7% and 91.5%, respectively. It should be noted that the ANN model performs with an equal accuracy in the "Class I" grouping, yet cannot match the ExtraTrees algorithm on the subsequent groupings. Furthermore, ExtraTrees as an algorithm, is far less computationally expensive than ANN, and as such, one could consider it superior – at least for the methodology of this thesis.

Looking at the two statistical models tested, namely Logistic Regression and Multiple Discriminant Analysis, we found that these generally performed worse than the machine learning methods, with the exception of the AdaBoost algorithm. These two achieved sub-30% accuracies in "Class I", compared to the above 30% that all machine learning models, with the exception of AdaBoost, managed to achieve. Based on the past research, examined in the literature, this was the expected outcome, and intuitively it makes sense that the machine learning models are better able to reflect the non-linear relationships between the input variables. Even though the machine learning models displayed a higher accuracy than the statistical models, we in many cases failed to achieve as high accuracies as some of the articles examined in the literature review. Albeit disappointing, considering the large dataset in this study, it is perhaps not that surprising, considering the focus on a single algorithm of the other articles.

Table 7.19: Accuracy Metrics Across Models

	Class I	Class II	Class III
LR	0.274	0.539	0.845
MDA	0.269	0.528	0.839
KNN	0.333	0.648	0.877
SVM	0.367	0.646	0.889
ANN	0.368	0.660	0.890
ExtraTrees	0.368	0.697	0.915
AdaBoost	0.213	0.449	0.866

Considering the Kappa measures described in Table 7.20, the same hypothesis of machine learning models being better predictors, compared to statistical models, is also supported. The statistical models perform around the .20 mark, which indicates a low degree of agreement.

Table 7.20: Kappa Metrics Across Models

	Class I	Class II	Class III
LR	0.201	0.386	0.670
MDA	0.197	0.374	0.656
KNN	0.269	0.536	0.742
SVM	0.305	0.531	0.764
ANN	0.307	0.552	0.767
ExtraTrees	0.307	0.599	0.819
AdaBoost	0.136	0.273	0.715

7.2.4 Summary

In this section, we have seen and evaluated the performance of the different statistical and machine learning models that we have tested on our dataset. Our findings support the hypothesis that the machine learning models we have tested outperform statistical methods, in terms of the performance metrics we have chosen to consider.

Our results are for some cases on a comparable level in terms of accuracy, to past studies, yet for other we have achieved results with somewhat lower accuracies. We have discussed a number of reasons for why this could be the case, which is most probably mainly methodological differences. We do in fact, in our ANN model, achieve better results than those of Moody and Utans (1994), who, like us, predicted 16 classes using ANN. Moody and Utans (1994) achieved a 30% accuracy, whereas we managed to achieve 37% accuracy.

One of our models, specifically ANN, appears to show a greater promise, as it performs well, despite being properly optimized due to computational restrictions. This could suggest that there is some degree of improvement possible. For the other models, considering the findings of Ye et al. (2008), SVMs might also hold a lot of potential. Whereas ExtraTrees performed well in our tests, we do not see it having as great a potential for improvement as ANNs and SVMs.

We will discuss and comment further on our results and the thesis as a whole in the subsequent chapter.

8 Conclusion

8.1 Introduction

In this chapter, we briefly go over the results, and their implications for answering the research question. We will then attempt to answer our final conclusion with regards to the research problem, based on the data in our analysis. Furthermore, we will discuss our results, and make the recommendations that we see relevant.

Lastly, we discuss how our conclusion fits into the field of predicting credit ratings, and we will discuss any relevant issues or improvements to the methodology used in this thesis. Lastly, we discuss other research questions that this thesis has triggered, and why these could be interesting for further study.

8.2 Conclusion

Looking at our results, it seems that using machine learning for modeling credit ratings, is most certainly viable. We managed to collect an impressive dataset in terms of size, and although it failed to markedly improve upon the highest accuracies achieved in past literature, it has shown that predicting credit ratings, and doing so with higher accuracy than statistical models and little pre-existing knowledge of the field is possible. This does suggest that much of the information that exist in credit ratings, is already actually contained in historical and current accounting figures.

In our opinion, the results presented make a convincing case that machine learning models are better able to predict credit ratings, showing approximately a 10%-point increase in accuracy, when predicting over 16 classes. These findings are in line with the findings of previous papers.

8.3 Recommendation

As per our conclusion, it seems that credit ratings can be modeled somewhat accurately using machine learning algorithms, and that machine learning algorithms are generally better at doing so than statistical methods. As such, any pure prediction application of credit ratings should in our view implement a machine learning algorithm over a statistical approach. Furthermore, actually implementing a machine learning model over a statistical model is not harder today, due to the many software tools available, and the large body of literature on machine learning.

However, due to the nature of machine learning algorithms as effectively "black box"-models, where one cannot easily deduce how the different variables contribute to the final prediction, they might be less useful for application in *understanding* why certain credit ratings are given. For those purposes, statistical models will remain easier to interpret.

Nonetheless, considering the still relatively low accuracy rates on the full spectrum of credit ratings (in this thesis referred to as the "Class I" grouping), of somewhere between 30% and 40%, it is clear that the models from this thesis should most probably not be implemented directly to any kind of business-critical application, for which accuracy is paramount. Nonetheless, a reasonably good credit rating prediction model certainly could have some degree of value, in an industrial application. Furthermore, it seems there are still leeway to improve upon the models, which would also increase their usefulness. We have already discussed potential applications, but it is worth noting that for some of these, 100% accuracy may not be a strict requirement.

Addressing the big question – *can rating agencies be replace by models* – the answer is a resounding "maybe", in our view. In the current state, most probably not, but it is not unthinkable that in the future one could prove a mathematical model to exhibit higher reliability and timeliness than that of the CRAs, while also holding the trust in the market from the market participants. With more transparency and a generally higher availability of data, it could be conceivable that machine learning models and AI could be better at seeing through complicated accounting structures, spotting accounting fraud, and in general making sure that bond issuers are thoroughly investigated and evaluated, on an objective level, such that market participants can remain confident in the functioning of financial markets. For the time being, it seems more likely that machines increasingly will, as we are seeing in any other industries, assist humans in making qualified decisions. As such, the rating agencies will not be replaced overnight, but it will be a slow transition toward more automation, and better decisions – a win for society and the financial system.

As of applications for this technology right now, one could imagine these machine learning models providing better insights for companies about their creditor and contractors, and with an automated system would be able to plan and act accordingly, should any of their creditors' creditworthiness change. One could also imagine a bond trading strategy, where one would trade an issuer's bonds or stocks based on predictions of up- or donwgrades.

To summarize, our recommendation based on the findings of this thesis is not to replace the rating agencies with machines, but rather look at how machine learning can be more accurately applied in this field, and how machine learning can assist those evaluating credit ratings make more informed decisions.

8.4 Discussion

This thesis has explored a well-studied, yet small, some would even say "fringe" area of finance. As such, there is little "best practice" to refer to in this field. Nonetheless, we feel that we have made our case for why it is a topic relevant studying, and have attempted to give our contribution to this field. We see that one of our largest contributions, has been testing some of the techniques described in other research papers, on a large dataset, in a structured manner, and have at the same time detailed how our dataset was obtained, such that this study can either be replicated or improved upon by others interested in the field.

Nonetheless, there are some methodological improvements that could be fairly easily implemented by researchers who would wish to improve upon this study. For starters, this study has had an aim of benchmarking different algorithms, and as such been less focused on optimizing the performance of each models, prepocessing the data thoroughly, and tailor the input to the specific needs of the different algorithms. As such, we find it highly likely that it would be possible to obtain significantly superior results through going in-depth with a single algorithm, and spend effort on optimizing the results, through a combination of input processing, variable selection, hyperparameter optimization and using ensemble techniques.

Whereas the dataset in itself is impressive, due to its size, it is also likely that it could use more thorough cleaning, and some logic for handling ”corner cases”, in which the calculations will yield numbers that are not informational to the credit rating classification. For instance, if a company has negative equity, and at the same time has negative net income, it's return on equity would figure as positive - which a model should link to a good credit rating, which should most probably not be the case in this specific situation.

In our results, we believe that Artificial Neural Networks, shows great promise for optimizing results on. It would probably have been possible to obtain higher accuracy, had we been able to employ sufficent computational power to properly optimize the hyperparameters of our ANN model. Furthermore, different types of neural networks, such as convolutional neural networks, have proven themselves useful in a number of applications. These, more sophisticated types of neural networks could most probably be used successfully for the purposes of predicting credit ratings.

To finish our discussion, we want to state that it seems clear from our findings, as well as the findings of the studies preceding ours, that there exists a lot of information in accounting figures about the creditworthiness about issuers. If we can already now, like Ye et al. (2008), with 80% accuracy predict credit ratings within 1 rating, the value created by the CRAs does not seem to be in having the knowledge and techniques to make credit ratings. Instead, their value seems more to be in inciting confidence in the market, their special regulatory status, and of course, the human factor, which cannot be ignored when it comes to finance.

8.5 Future research

This thesis has opened a number of new and interesting questions. Based on both the actual availability of data, and the findings of previous research it seems that there are many possibilities for researchers to investigate the topic of machine learning and credit ratings further.

For starters, there exists a plethora of other machine learning models, which are constantly being used for new and sensational applications. These could could be interesting to apply to the problem of predicting credit ratings.

Secondly, many improvements in machine learning models is done by preprocessing the inputs and so-called feature extraction. What's more, is that there is more and more data becoming available. Future research is needed in how to better process inputs and extract features from data, as well as research into how other, non-accounting data could help improve the rating accuracy.

Lastly, it seems that some earlier studies have already obtained impressive results. For instance the findings of Ye et al. (2008) are incredibly interesting from this point of view, and it would be useful for the field to investigate if and how some of their approaches can be encoded into a ”best practice” or if their approach could be improved upon, in order to gain even better results.

Appendices

A Equations

Table A.1: Equations of Variables Used

Variable	Description
$Profit\ Margin = \frac{Net\ Income}{Net\ Sales}$	The ratio is an indicator of total margin to cover expenses and potentially yield a profit for the shareholders
$Operating\ Margin = \frac{Operating\ Income}{Revenue}$	This ratio is an indicator of the firms operating efficiency and measure which part of a company's revenue that is available after paying for variable costs of production e.g. wages and raw materials
$Return\ on\ Equity = \frac{Net\ Income}{Shareholder's\ Equity}$	The ratio is a important profitability measure that display the productivity of equity as reported in the balance sheet. The measurement provides indication of a firms ability to raise equity-capital that serve as a cushion for the debt-holders (financial statement analysis a practioners).
$LT\ Debt\ to\ Capital = \frac{LT\ Debt}{LT\ Debt+Total\ Equity}$	The ratio yields the firms relative balance between debt and equity of the firms long term financial obligations.
$Current\ Ratio = \frac{Current\ Assets}{Current\ Liabilities}$	The current ratio is an indicator to which extent liabilities that will be due soon are covered by assets that are expected to be converted to cash within the approximate same period of time.
$Return\ on\ Assets = \frac{NetIncome}{AverageTotalAssets}$	The ratio measure the profitability of relative to its asset-base and thus can be interpreted on how well the firm manages it's assets.
$\log\left(Net\ Sales\right)$	log-transformed net-sales that provides a levelized measurement of the firms ability to generate revenue.
$Pretax\ profit\ margin = \frac{Pre-tax\ icome}{Net\ sales}$	A company's earnings before tax as a percentage of total sales or revenues. The higher the pretax profit margin, the more profitable the company.
$Return\ on\ invested\ capital = \frac{Pre-tax\ income}{Avg.\ invested\ capital}$	Indicator that measure to which extent the company efficiently allocates funds to profitable investments regardless of source the of financing.

Table A.1: Equations of Variables Used

Variable	Description
$Cashflow\ to\ sales = \frac{Cash\ from\ operations}{Net\ sales}$	The ratio explains the firms ability to convert sales to cash
$Cash\ ratio = \frac{Cash\ and\ equivalents}{Current\ Liabilities}$	The cash ratio is the ratio between cash and equivalents to current liabilities and measure the firms ability to meet its short term financial liabilities
$Debt\ to\ equity = \frac{ST\ Debt+LT\ Debt}{Total\ Equity}$	A measure of a firms financial leverage
$Current\ to\ Total\ Assets = \frac{Current\ Assets}{Total\ Assets}$	Indicates the extent of total funds invested for the purpose of working capital and throws light on the importance of current assets of a firm
$Quick\ to\ Total\ Assets = \frac{Acct.\ Rec.+Cash\ and\ Equiv.+Mkt.\ Sec.\ and\ ST\ Inv.}{Total\ Assets}$	An indicator of a company's short-term liquidity and financial strength or weakness.
$Debt\ to\ Total\ Assets =$	Provides a measure of the proportion of debt a company has relative to its assets. It gives an indication of the amount of leverage being used by a company.
$Common\ Equity\ to\ Total\ Assets = \frac{Common\ Equity}{Total\ Assets}$	Expresses the share of the total assets that common stockholders are entitled to.
$Total\ Equity\ to\ Total\ Assets = \frac{Total\ Equity}{Total\ Assets}$	This ratio expresses the proportion of total assets financed by the owner's equity capital. It provides an indication of a company's leverage
$Asset\ Turover = \frac{Net\ Sales}{Avg.\ Total\ Assets}$	This ratio measures the amount of revenue generated for every dollar's worth of assets. It is useful for determining how efficiently and effectively management uses its assets to generate revenues
$Solvency\ Ratio = \frac{Net\ Income+Amort.\ and\ Depr.}{Total\ Liabilities}$	This ratio measures a company's ability to meet its long-term obligations. In general, the lower a company's solvency ratio, the more likely it is to default on its debt obligations.
$Sales\ to\ Total\ Equity = \frac{Net\ Sales}{Total\ Equity}$	This ratio provides a sense of a company's creditworthiness as it measures the number of sales dollars generated with each dollar of investment
$Total\ Equity\ and\ LT\ Debt\ to\ Fixed\ Assets = \frac{Tot.\ Eqt.+LT\ Debt}{Fixed\ Assets}$	Measures the extent to which the firms fixed assets are finance through equity and debt.

Table A.1: Equations of Variables Used

Variable	**Description**
$Interest\ Expense\ To\ Sales = \frac{Int.\ Exp.}{Net\ Sales}$	A useful metric for comparing the efficiency of a company's interest expenditure between companies in the same industry
$Times\ Interest\ Earned = \frac{Operating\ Income}{Interest\ Expense}$	Measures a company's ability to pay its debt obligations and an indication of the number of times a company can cover its interest charges with its pre-tax earnings.
$Cash - flow\ to\ Debt = \frac{Cash\ from\ Operations}{LT\ Debt+ST\ Debt}$	This ratio compares a company's cash flow to its revenues which provides a measure of the company's ability to generate cash from its current operations.
$Cash - flow\ to\ Sales = \frac{Cash\ from\ Operations}{Net\ Sales}$	An efficiency ratio that rates actual cash flows to company assets without being affected by income recognition or income measurements.
$Cash\ Return\ On\ Assets = \frac{Cash\ from\ Operations}{Avg.\ Net\ Sales}$	This ratio measures the cash a company can generate in relation to its asset size

References

Bank for International Settlements, 2005, Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework, Technical report, Bank for International Settlements.

Becker, Bo and Milbourn, Todd, 2010, How did increased competition affect credit ratings?

Bennell, Julia A., Crabbe, David, Thomas, Stephen, and ap Gwilym, Owain, 2006, Modelling sovereign credit ratings: Neural networks versus ordered probit.

Bishop, Christopher, 2006, *Pattern Recognition and Machine Learning*, Springer Science+Business Media, LLC, 1 edition.

Brownlee, Jason, 2014, Classification Accuracy is Not Enough: More Performance Measures You Can Use.

Brownlee, Jason, 2016, Data Leakage in Machine Learning.

Dutta and Shekhar, 1988, Bond rating: a nonconservative application of neural networks, In *IEEE International Conference on Neural Networks*, pages 443–450. IEEE.

Ederington, Louis H., 1985, Classification Models and Bond Ratings, *The Financial Review*, 20(4):237–262.

Fisher, Lawrence, 1959, Determinants of Risk Premiums on Corporate Bonds, *Journal of Political Economy*, 67(3):217–237.

Frank, Simon J., 2009, *Predicting Corporate Credit Ratings using Neural Network Models*, PhD thesis, University of Stellenbosch.

Gibert, Karina, Rodas, Jorge, and Gramajo, Javier, 2008, AI versus Statistics: Some common topics.

Hajek, Petr and Michalak, Krzysztof, 2013, Feature selection in corporate credit rating prediction, *Knowledge-Based Systems*, 51:72–84.

Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome, 2009, *The Elements of Statistical Learning*, volume 1 of *Springer Series in Statistics*, Springer New York, New York, NY.

Head, Tim, 2015, Bayesian optimisation for smart hyperparameter search.

Horrigan, James, 1966, The Determination of Long-Term Credit Standing with Financial Ratios, *Source Journal of Accounting Research*, 4:44–62.

Huang, Zan, Chen, Hsinchun, Hsu, Chia-Jung, Chen, Wun-Hwa, and Wu, Soushan, 2004, Credit rating analysis with support vector machines and neural networks: a market comparative study, *Decision Support Systems*, 37(4):543–558.

IOSCO, 2003, Report on the Activities of Credit Rating Agencies, Technical report, Internatioanl Organization of Security Commisions.

Kaplan, Robert S and Urwitz, Gabriel, 1979, Statistical Models of Bond Ratings: A Methodological Inquiry, *The Journal of Business*, 52(2):231.

Kennedy, K, 2013, *Credit Scoring Using Machine Learning*, PhD thesis, Dublin Institute of Technology.

Kim, Jun Woo, Weistroffer, H. Roland, and Redmond, Richard T., 1993, Expert systems for bond rating: a comparative analysis of statistical, rule-based and neural network systems, *Expert Systems*, 10(3):167–172.

Kim, Kee S., 2005, Predicting bond ratings using publicly available information, *Expert Systems with Applications*, 29(1):75–81.

Kumar, Kuldeep and Bhattacharya, Sukanto, 2006, Artificial neural network vs linear discriminant analysis in credit ratings forecast, *Review of Accounting and Finance*, 5(3):216–227.

Langohr, Herwig and Langohr, Patricia, 2012, *The Rating Agencies and their Credit Ratings*, John Wiley & Sons, Inc., Hoboken, NJ, USA.

Lee, Young-Chan, 2007, Application of support vector machines to corporate credit rating prediction, *Expert Systems with Applications*, 33(1):67–74.

Maher, John J. and Sen, Tarun K., 1997, Predicting Bond Ratings Using Neural Networks: A Comparison with Logistic Regression, *International Journal of Intelligent Systems in Accounting, Finance & Management*, 6(1):59–72.

Maheshwari, Manish, 2016, Ensemble of Weak Learners — Manish Maheshwari.

Moody, J and Utans, J, 1994, Principled Architecture Selection for Neural Networks: Application to Corporate Bond Rating Prediction, *Neural Networks in the Capital Markets*.

Moody's Investors Service, 2004, Guide to Moody's ratings, rating process, and rating practices, Technical report, Moody's Investors Service.

Pinches, George E. and Mingo, Kent A., 1973, A MULTIVARIATE ANALYSIS OF INDUSTRIAL BOND RATINGS, *The Journal of Finance*, 28(1):1–18.

Quora, 2017, What are hyperparameters in machine learning ?

scikit-learn, 2017, scikit-learn Documentation.

Sinclair, Timothy J., 2008, *The New Masters of Capital: American Bond Rating Agencies and the Politics of Creditworthiness*, Cornell University Press, Ithaca, United States, 1 edition.

Sprengers, Mark-Alexander, Van Den Berg, Jan, and Waltman, Ludo, 2006, BOND RATING CLASSIFICATION A Probabilistic Fuzzy Approach.

Standard & Poor's Financial Services LLC, 2016, S&P Global Ratings Definitions, Technical report, Standard & Poors.

Standard and Poor's, 2008, Corporate Ratings Criteria 2008, Technical report, Standard and Poor's, New York, NY.

Surkan, A.J. and Singleton, J.C., 1990, Neural networks for bond rating improved by multiple hidden layers, In *1990 IJCNN International Joint Conference on Neural Networks*, pages 157–162. IEEE.

U.S. Securities and Exchange Commission, 2003, Report on the Role and Function of Credit Rating Agencies in the Operation of the Securities Markets, Technical report, U.S. Securities and Exchange Commission.

U.S. Securities and Exchange Commission, 2016, What We Do.

Viera, Anthony J and Garrett, Joanne M, 2005, Understanding interobserver agreement: the kappa statistic., *Family medicine*, 37(5):360–3.

West, Richard R, 1970, An Alternative Approach to Predicting Corporate Bond Ratings, *Journal of Accounting Research*, 8(1):118.

White, Lawrence J., 2010, Markets: The credit rating agencies, *The Journal of Economic Perspectives*, 24(2):211–226.

Wu, Hsu-Che, Hu, Ya-Han, and Huang, Yen-Hao, 2014, Two-stage credit rating prediction using machine learning techniques, *Kybernetes*, 43(7):1098–1113.

Ye, Yun, Liu, Shufen, and Li, Jinyu, 2008, A Multiclass Machine Learning Approach to Credit Rating Prediction, In *2008 International Symposiums on Information Processing*, pages 57–61. IEEE.