Predicting Corporate Credit Ratings: A Comparative Study Between Ordered Probit, Neural Network and Random Forest

Daniel Larsson 23363 & Filip Wikander 23364

May 14, 2017

STOCKHOLM SCHOOL OF ECONOMICS
B.Sc Thesis in Finance

Abstract

This thesis compares the prediction accuracy for corporate credit ratings between three different models. The two first models, a traditional statistical model called ordered probit and a machine learning model called artificial neural network has been used with success before. The third model, a machine learning model called random forest is implemented and compared to the previous models. The random forest model accurately predicts 66% of all credit ratings in a holdout samples outperforming ordinal probit (58%) and artificial neural network (63%). McNemar's test validates that the accuracy of the random forest is significantly different from the ordinal probit at a 0.1% significance level and artificial neural network at a 5% significance level. The random forest also provides evidence that market value and equity volatility are important when predicting S&P credit ratings.

Keywords: Credit ratings; Machine learning; Ordered probit; Random Forest; Neural network
Tutor: Christian Huse

Contents

1 Introduction

In the financial market there are a myriad of entities providing analysis of companies. Among these, a few are recognized by the government as statistical rating organizations (NRSRO) giving them the right to set credit ratings on corporations, bonds and countries. Three companies dominate the business of credit ratings and their ratings have been shown to provide more information than what is publicly available, thus affecting both debt and equity prices (Kliger & Sarig 2000). These three companies are Standard & Poor's (S&P), Moody's Investor Services and Fitch Ratings and they are credit rating institutions. What they do is a fundamental and qualitative analysis of a company and in the end the company is given a credit rating, which is a measurement of the company's to repay its debts. While the three credit rating institutions have different labels for their measurement of default risk, they have the same implications. S&P which our thesis will focus on, gives out ratings in the form of AAA down to C where AAA is the best possible rating and means the company has a very low risk to default regardless of the financial cycle and for each step down in the ratings, the risk for default increases. The credit ratings are divided into investment- and speculative grade where everything rated below BBB is considered speculative. While looking at historical default rates might not predicative of the future, the difference in defaults between investment grade bonds and speculative grade bonds is staggering. In the annual global corporate default study by Vazza & Kraemer (2015) that is based on average cumulative corporate default rates from 1981-2015, an investment grade rated company leaps a 0.98% default risk during a five-year horizon and 3.14% risk during a fifteen-year horizon while a speculative grade rated company has a 15.24% default risk under a five-year horizon and 24.75% risk under a fifteen-year horizon. The credit ratings therefore have implications on how much entities are willing to lend, what interest rate the issuer of bonds must pay to creditors and investment in speculative grade bonds is also generally not allowed by depository institutions in the US (Office of the Comptroller of the Currency 1990) which makes credit ratings an important tool of information for investors.

The ratings are not given out for free and companies must pay the credit raters to get their ratings. This means that important information to investors are not available for companies either not being able to afford or not wanting to pay for a rating. Therefore, being able to accurately predict what rating a company has would be of great value to investors and could potentially be used to find underpriced and overpriced bonds. Research during the last years have also focused on the different biases involved in the rating process and how the competition between the agencies result in inflated ratings for certain firms. Bolton et al. (2012) showed that ratings agencies can increase their profits by deviating from their standards and provide inflated ratings during periods of time, and Fracassi et al. (2016) provided evidence of systematic optimism and pessimism among individual analyst which in turn affects the final ratings. A good model of the rating process could consequently be of help in finding the cases in which credit rating agencies deviates from their standard because of competitive reason or as a result of individual analyst sentiment. The subject of predicting credit ratings has been attempted under several decades, with early attempts using statistical models such as OLS and ordered probit. As computational power has grown, the traditional statistical models have been exchanged for more powerful machine learning algorithms. These models do not rely on any underlying assumptions of the data and are therefore strong when the distribution of data is unknown, as is the case of credit ratings and their underlying risk of default. For our thesis, we have decided to compare the statistical method ordered probit with two machine learning algorithms to predict credit ratings.

The first machine learning algorithm is called an artificial neural network and has been used to predict credit ratings and in previous studies it has been shown to significantly outperform traditional statistical models. The artificial neural network attempts to mimic the brains ability to quickly adapt to new information by utilizing previous knowledge and the idea is to learn from fundamental data what categorizes a company in a certain credit rating category. The problem with the neural network is that while it may be an accurate classifier, it works like a black box. The user is thus only presented the output and no information about how it made its decision. The inability to understand the neural network is a common criticism of the model and for predicting credit ratings it is important to understand what variables a credit rater looks at, which leaves much to be desired from this model (Huang et al. 2004). Since the neural network is hard to understand, we propose a second algorithm called a random forest. The random forest is a rule-based algorithm with clear visibility of how it reaches a decision and provides the user better insight into what variables are of importance in the model. It has previously been successful in predicting consumer credit risk (Khandani et al. 2010), which bear resemblance to how predicting credit ratings work. Our goal with this thesis is to investigate if the random forest will outperform traditional statistical models and provide at least equal accuracy as the neural network, while at the same time providing valuable information into what data S&P looks at when giving out corporate credit ratings.

1.1 Previous Literature

The history of predicting bond ratings includes a big variation of statistical techniques and variables. One of the first examples was a study conducted by Horrigan in 1966 in which he used an ordinary least squares (OLS) model to predict the credit ratings. Horrigan used the data from companies who did not get their credit rating adjusted during the period of 1959-64 to predict the ratings of companies which had their rating adjusted or was not previously rated. Both ratings from S&P and Moody´s were used and two models were subsequently estimated. The purpose of the study was to test as to whether financial ratios could be used to predict bond ratings, and only two non-accounting variables were used in the model (the total assets and a dummy for the subordination of the bond). The model included ratios for long-term solvency, short term capital-turnover, long-term capital-turnover as well as a profit margin ratio. The models managed to correctly predict 52% of the holdout sample of the S&P rated bonds and 58% of the holdout sample of the bonds rated by Moody's, which led Horrigan to conclude that financial ratios can be used to determine bond ratings (Horrigan 1966).

Horrigan's study was later criticized by West (1970) for lacking theoretical reasoning behind the choice of variables. The conclusion was also called into question as most of the predictive power of the model was due the non-accounting variables included in the model. West instead decided to base his model upon the work of Fisher in his paper *Determinants of Risk Premium on Corporate Bonds*. According to Fisher the average risk premium of a bond is dependent upon two main factors, the risk that the firm will default on its bond and the marketability of the bond (Fisher 1959). The higher the risk premium and the lower the marketability of the bond the higher the risk premium. Fisher uses the value of bonds outstanding as a proxy for the marketability of the bond. He then uses earning variability, period of solvency (number of years for which the firm has not defaulted on a bond) and the equity to debt ratio to determine the risk that the firm will default on its bond. The model used by West used the same variables as Fishers study and was slightly better at predicting credit ratings than Horrigan's model but due to the small number of observations no model could be proven to outperform the other. West also concluded that Horrigan's model is easier to use and therefore the more reasonable option in choosing between the two even though the model based on Fishers reasoning was more theoretically satisfying.

One important factor to note regarding both previously mentioned models is that they use ordinary least squares to predict bond ratings. This classification method assumes that bond ratings are on an interval scale. In other words, the difference between for example an AAA and an AA rating would be identical to the difference between a BB and a B rating. It is however not clear that this assumption is valid and neither of the two studies provide any evidence as to the validity of the assumption. Pogue and Soldofsky sidesteps the assumption of an interval scale by estimating the probability of one rating over the other rating. The study concluded that leverage and profitability were the most useful variables when predicting credit ratings (Pogue & Soldofsky 1969). This classification method does not assume an interval scale but does on the other hand only predict which of two ratings a bond is given. This approach does therefore provide less information than the previous models and the usefulness of a model which can only predict two ratings can be questioned.

1.1.1 Multiple Discriminant Analysis

A more effective way to deal with the assumption of interval scale was implemented by Pinches & Mingo (1973) as they used multivariate discriminant analysis to predict bond ratings. Their sample consisted of 180 bonds rated by Moody's during the period of 1967-1968. A holdout sample consisting of 48 firms were randomly generated and the model was developed using the existing 132 firms. Out of the five possible ratings the model correctly predicted approximately 60% of the ratings in the holdout sample. The model had significant predictive ability but the performance varied across the different ratings as for example not a single Baa rated bond was correctly classified by the model. The study concluded that the legal status (subordinate or not) of the bond is the single most significant variable. Subordination alone would correctly have classified 83.3% of the bonds in the holdout sample if one were to classify them into investment grade and non-investment grade (BBB or above). Years of consecutive dividend and issue size were according to the model variables of great importance while net income plus interest divided by interest and net income divided total assets were of lesser importance. The model used did consequently suggest that credit ratings at this point in time were more concerned about the stability of the firm rather than profitability.

The multiple discriminant model was also used by Belkaoui (1980) as he further developed the model used by Pinches and Mingo. Belkaoui criticize the previous research for lacking an economic rationale in the selection of variables. The study hypothesizes that the investment quality is determined by the interaction of three general variables, which are firm-, market- and indenture-related variables. The firm related variables consists of size and coverage factors. Size provides protection to bondholders through the size of underlying assets in the firm while coverage factors is an indication of the firm's ability to service its debt in the future. The variables total assets, total debt, long-term debt divided by total invested capital and short term-term debt divided by total invested capital were included to cover the size factors. Current assets divided by current liabilities and a coverage ratio were included to capture the coverage factors. Market related variables were thought of as being captured well by investor expectations and stock price to common equity was included to cover the market related variables. Lastly, the indenture-related variables depict the legal status of the bondholder and was included in the model through a dummy variable indicating the subordination of the bond. The sample consisted of 275 bonds rated by S&P during 1978. Out of the six available ratings the model correctly predicted the rating of 65.9% of the bonds in the holdout sample.

1.1.2 Ordered Probit

The multiple discriminant analysis handles categorical outcomes well but it does not consider the ordinal nature of the ratings. An AAA rating is for example indicative of lower credit risk than an

AA rating and so on. The models used by Pinches and Mingo, and then Belkaoui does not take this into account, instead each category is treated as independent from the other. Kaplan & Urwitz (1979) uses a modified maximum likelihood estimation to predict bond ratings. The model is called ordered probit and assumes that the classification is the result of a latent variable (McKelvey & Zavoina 1975). In the case of credit ratings, the assumption is that credit ratings are a way to convey ordered information (default risk) which is a much more plausible assumption than with the previously used models (Kaplan & Urwitz 1979). Kaplan and Urwitz believe the structure to be the following. A bond rater attempts to set bond ratings by measuring the risk for default of said bonds, but due to possessing inadequate information, said rater can only set the ratings on an ordered scale. That is, AAA is less risky that AA and so on. In the end, the rater would hope that AAA was the least defaulted bonds. The ordered probit model does not make any underlying assumption about the distance between the different ratings, only that AAA is to be less risky than AA and so on. The variables included in the model can be divided into five categories interest coverage ratios, capitalization (leverage) ratios, profitability ratios, size variables and stability variables. Most of the variables used has been used in previous studies except for the stability variables which included the variation coefficient of total assets, the variation coefficient of net income, an accounting beta and an unsystematic accounting risk measure. The authors reasoned that for firms which earnings are highly correlated with the market, then earnings fluctuations could be understood as a market variation which should not have a big of an impact on the long-term riskiness of the firm's bond. The beta and the unsystematic risk measure are likely not variables used by the rating agencies but they are likely an aggregation of the firms operational and financial risk which are something highly relevant to the rating agencies. All financial ratios used five year average the ratios for interest coverage, capitalization and profitability were adjusted for industry.

The model was estimated using 140 newly issued bonds during 1970-1971 rated by Moody and was then used to predict a holdout sample consisting of 64 newly issued bonds from the same time period. The model correctly predicted 69% of bonds in the holdout sample. One important factor which is being overlooked is the skewed distribution of given ratings among the holdout sample. The rating A is given many of the issues and a model which only predicts the rating A would in this case correctly predict 56% of the ratings, and 75% of all correct predictions by the Kaplan and Urwitz model are due it predicting rating A. Cross validation was made by using the model to predict the rating of seasoned bonds, during which it only correctly predicted 43% of the ratings. The authors conclude this is due to the rating agencies not updating their rating but looking at the data it is clear that Kaplan and Urwitz model is overestimating the number of bonds rated A, and is consequently performing worse. The difference in prediction accuracy between the two samples were consequently not only due to lag in agency ratings but also due to differences in rating distributions between the two samples. It is not clear why the two samples have different rating distributions but could likely be due to the sample of newly issued bonds during 1970-1971 not being representative of rating distributions in general, which would result in a biased estimator.

The ordered probit model was also utilized by Blume et al. (1998) in their paper *The Declining Quality of U.S. Corporate Debt: Myth or Reality*. The ratings given by rating agencies had been declining and this study examines whether this was due to a decline in quality or due to more stringent standards in the rating agencies. The variables included in the model were based upon previous research and ten "key" financial ratios published by S&P. The authors also include a variable for firm value and the two risk measures from the market model, the market beta and the standard error. The logic behind these variables is that firms with higher market value is thought to be older and more diversified which will result in lower credit risk. The risk measures have been proven to be good indicators of credit rating in previous studies, and even though the market model captures equity risk this can be thought to be correlated to the credit worthiness of the company. The sample consisted only of investment grade bonds which means that there were four possible ratings for each firm. The study did not use a holdout sample and the correctly predicted ratings in the sample were around 59 %. The study concludes that the declining ratings given to U.S. corporations are due to more stringent standards by the rating agencies rather than a decrease in the quality of the debt.

1.1.3 Artificial Intelligence and Neural Networks

During the latest years, a lot of research has been devoted to the use of artificial intelligence to predict credit ratings as compared to the statistical models previously used. One of the first studies to use machine learning to predict corporate credit ratings was made by Kim et al. (1993) when they compared an artificial neural network to the use of more traditional statistical models. The sample consisted of 110 firms rated by S&P during 1988 and the holdout sample consisted of 60 firms rated by S&P during 1989. Firms rated below B was not included in the sample which meant six possible ratings for each firm. The variables used were the same as the ones used by Belkaoui (1980). The predictive ability of the artificial neural network with 55.17% correctly predicted ratings was clearly higher than the regression analysis, multiple discriminant analysis and logistic regression with 36.21%, 36.20% and 43.10% correctly predicted respectively. The authors conclude that artificial neural perform better at predicting corporate credit ratings than the other traditional statistical model. One reason for this could according to the authors be that artificial neural networks does not require any a priori assumptions regarding the distribution or the functional form of the data which the other models do. Later research has supported the conclusion with Kumar & Bhattacharya (2006) showing that artificial neural networks outperform linear discriminant analysis and Bennell et al. (2006) providing evidence that artificial neural networks also outperforms ordered probit models.

2 Methodology

2.1 Ordered Probit

The ordered probit is a statistical model used for ordered data which does not assume a linear relationship with equal distance between ordered categories. Other models such as OLS assumes an equal distance between ordered categories, which is not useful when the outcome has a natural order but no quantitative interpretation between the outcomes. An example of a natural ordering without quantitative interpretation would be asking how a person feels about their own health where the possible outcomes are: excellent, good, neutral, poor. It is known that excellent is better than good and good is better than neutral, but it is hard to quantify the linear relationship between them. This is important when predicting credit ratings since they are categories of an ordinal nature representing default risk. While the true default risk between the categories is unknown it is unlikely that the default risk increases in equal increments between them. What is more likely is that default risk between rating AAA and AA may not be the same as the difference in default risk between category AA and A. By capturing the unequal distances between ratings, the ordered probit model better replicates the nature of the bond rating process compared to OLS which just assumes an equal distance between the categories. It was first used by Kaplan & Urwitz (1979) on bond ratings and has since then been used in most other studies (Blume et al. 1998, Dimitrov et al. 2015) regarding prediction of credit and bond ratings. This is therefore the model we will benchmark our machine learning models against and the computations for this model will be done using the oprobit function in Stata.

The mathematical representation of the ordered probit model is the following. If it is assumed that the unobservable variable y^* follow the relationship of equation 1 where $x^{(j)}$ is the vector of inputs for firm j and β is a vector of regression coefficients.

$$y^* = x^{(j)}\beta + \varepsilon \tag{1}$$

Since it is not possible to observe y^* we can only observe its categorical value. In our model the latent variable y^*will be representing the credit risk and the categorical values are the corresponding credit ratings. To estimate the credit ratings, the ordered probit will estimate y^* and then use certain cutoff points to assign different ratings (equation 2) where μ_i represents the cut off point for category i. In estimating the variable $y^*$the maximum likelihood estimation is used as the ordinary least squares would result in a biased estimator (McKelvey & Zavoina 1975).

$$y = \begin{cases} 0 & if \quad y \leq \mu_1 \\ 1 & if \quad \mu_1 < y^* \leq \mu_2 \\ 2 & if \quad \mu_2 < y^* \leq \mu_3 \\ \vdots & \quad\quad \vdots \\ k & if \quad \mu_k < y^* \end{cases} \tag{2}$$

2.2 Artificial Neural Networks

An artificial neural network, commonly known as a neural network is an attempt to mimic how the brain works in mammals or more specifically the human brain. The idea behind it stems from the fact that the brain is a highly complex, nonlinear and parallel computer capable of processing information in a different and much more effective way than conventional digital computers. For example, the brain routinely identifies and recognizes objects in its surroundings in approximately 100-200 milliseconds, whereas much simpler tasks can take several hours to days for a traditional digital computer (Haykin 2004). It is a fact that the brain is made up of cells called neurons and from this point the underlying theory on simulating the brain was first proposed by McCulloch & Pitts (1943) which bound together the field of neurophysiology with mathematical logic. A neuron is a cell receiving, processing and transmitting signals from our senses and a neural network tries to replicate their function in a digital environment. Instead of getting information from the senses the artificial neurons get signals from a data set, for example financial data and then process it with the goal of answering a predefined question such as what credit rating should a company with this financial data get assigned to.

While the computational abilities are yet nowhere near human capacity in terms of generalizing, specific tasks such as pattern recognition can be achieved with great results by putting together a network of artificial neurons, such as identifying written numbers with 0.21% error rate (Wan et al. 2013). Just like a human, the network also learns from experience and gets better from each sample. By storing knowledge about previous samples the network has a good ability to generalize which means it is good at predicting unknown samples if it is within the same task. Haykin (2004) who has written an influential book on neural networks describes it as an adaptive machine and it resembles the brain due to:

1. Acquiring knowledge from its surroundings by a learning process
2. Connection between the neurons, known as synaptic weights, allow for storage of knowledge.

The ability to learn from samples and not depending on underlying assumptions about distributions makes artificial neural networks a powerful model and it has already been successfully used within finance for predicting credit ratings as mentioned earlier (Kim et al. 1993, Bennell et al. 2006).

For our neural network, we are using a multi-layer perceptron classifier, in short MLP classifier, which has been applied successfully to a diverse amount of complex problems by training them on samples including both input and outputs. More specifically the MLP classifier from the SciKit Learn library will be used as it is widely used both within the academic setting and within industries (Pedregosa et al. 2011). The MLP classifier consists of two passes through the network, forward propagation and backward propagation. In the forward propagation, the input variables are applied on the network and propagates through the different layers while keeping weights fixed until finally resulting in an output in the last layer. The backward propagation then goes back through the layers changing the weights as to minimize the difference between the actual and desired output. This is then iterated over and over until a minimum difference has been reached.

2.2.1 Artificial Neuron

A neuron is the basis for what builds a neural network and figure 1 shows the model of a single neuron, which consists of an input layer with variables from the data set, one neuron for calculations and an output layer. There are three basic elements of importance in this model.

1. There is a set of weights, called synaptic weights, connecting the input signal to the neuron. Thus, at input signal x_j a synapse j is connected to neuron k by the synaptic weight Θ_{kj}.
2. An adder sums the input signals weighted by their respective synaptic weight
3. An activation function, also called a squashing function, limits the output to usually between 0 and 1.

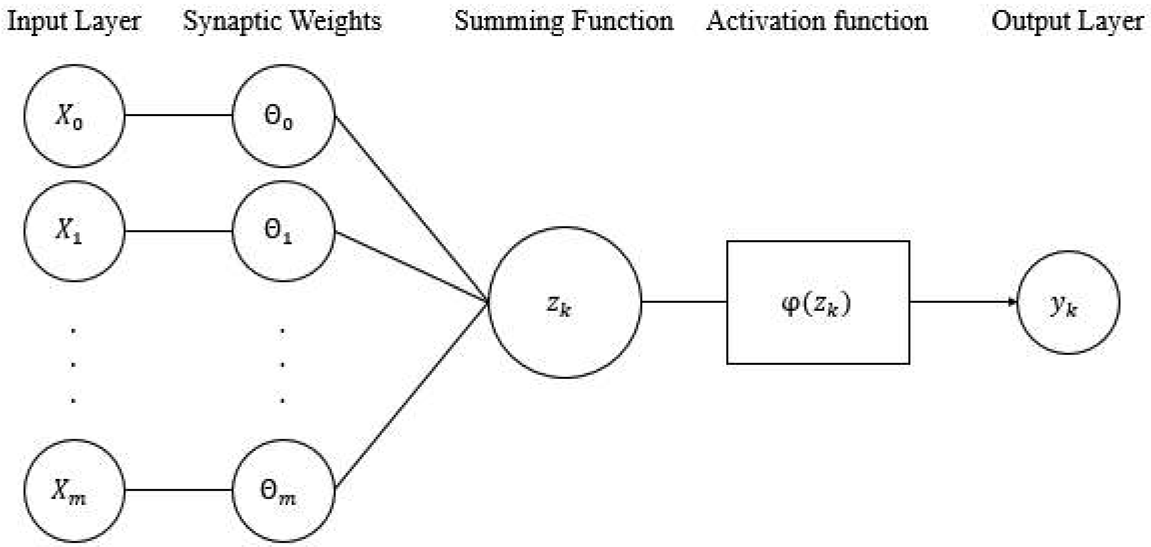
There is also a bias unit $x_0 = 1$ with the weight Θ_{k0} which helps shifting the activation function by a negative or positive amount to better fit the data, like b in $y = ax + b$ which shifts the intercept of the line to better fit data. A mathematical representation of the Neuron can be given by equation 3 and 4.

$$z_k = \sum_{j=0}^{m} \Theta_{kj} x_j \tag{3}$$

$$y_k = \varphi(z_k) \tag{4}$$

where $x_0, x_1 ..., x_m$ are the input signals and $\Theta_{k0}, \Theta_{k1}, ..., \Theta_{kj}$ are the synaptic weights relating to neuron k. z_k is the sum of the synaptic weights and input signals from 0 to m thus including the bias unit. $\varphi(z_k)$ is the activation function which limits the output between 0 and 1 and usually consists of

Figure 1: Artificial Neuron



a sigmoid function (equation 5) And the activation function then has a threshold function (equation 6).

$$p(z_k) = \frac{1}{1+e^{-z_k}} \tag{5}$$

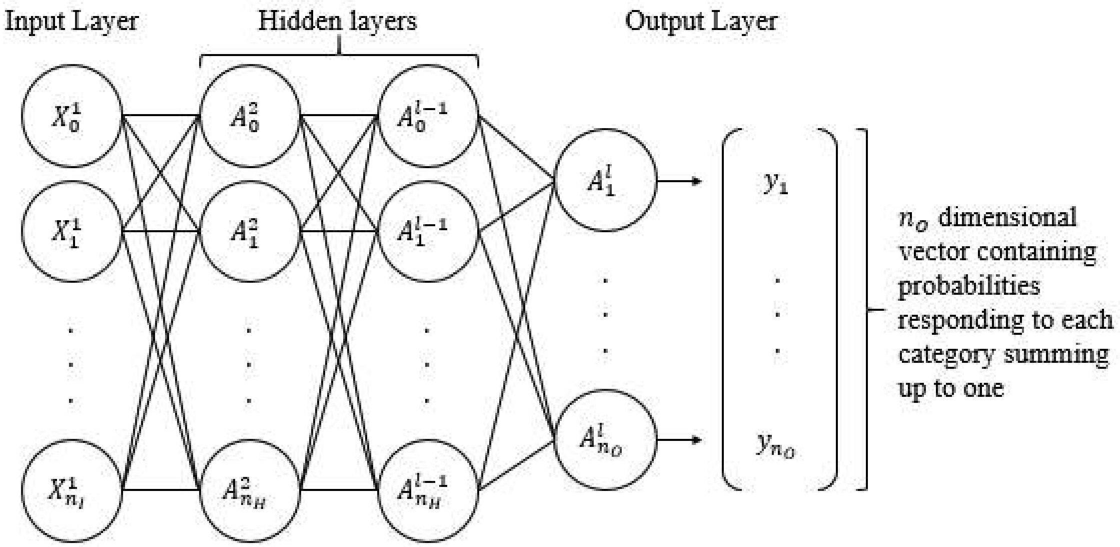
$$y_k = \begin{cases} 1 & if \quad z_k \geq 0 \\ 0 & if \quad z_k < 0 \end{cases} \tag{6}$$

An example of how the activation function works would be if the model was designed to predict if a stock would go up or down the next day. If the input were financial variables, such as price to earnings and historic volatility of the stock and the computation in the neuron led to $y_k = 1$ it would mean the model believes the stock is going up, while a 0 means it will go down. A visual representation of an artificial neuron is shown in figure 1.

2.2.2 Neural Network

A neural network consists of several neurons put together**.** While there are always one input and one output layer, there can be *n* hidden layers with *k* neurons in them. The hidden layer's act as a feature detector, meaning as the training of the network progresses they will discover previously unknown features in the data. This layer is called hidden because the user will not see what computations happen in this stage and it is therefore referred to as a black box. This is one negative side of the model, since unlike traditional statistical methods it is very hard to know why the model chooses

Figure 2: Artificial Neural Network



the features it does. The optimal structure of these layer should ideally be $n_I > n_H > n_O$ where n_I is the number of input signals, n_H is the number of neurons in the hidden layers and n_O is the number of outputs and is equal to the number of categories (Braspenning et al. 1995). For our model, we went with a 13-8-8-6 structure for our network. The size and number of hidden layers was derived from trial and error since there is no method for getting the optimal number of neurons and layers in a hidden network. For data preparation, we also used a function in Python called robust scaler which normalizes the data and takes care of outliers since this makes the backpropagation reach minimum faster. The true strength from this network is the ability to learn from experience using forward and backward propagation, which will be explained below.

What is different in the neural network model compared to a single neuron is $a_k^{(l)}$where a is the activation of neuron k in layer l. The mathematical representation is shown in equation 7 where equation 8 is the first layer after the input layer, and for the following layers up until the output layer in equation 9.

$$a_k^{(l)} = \varphi(z_k^{(l)}) \tag{7}$$

$$z_k^{(l)} = \sum_{j=0}^{n_I} \Theta_{ki}^{(l)} x_i \tag{8}$$

$$z_k^{(l)} = \sum_{i=0}^{n_H} \Theta_{ki}^{(l)} a_i^{(l-1)} \tag{9}$$

Where $x_0, x_1 ..., x_m$ are the input signals, $\Theta_{k0}^{(l)}, \Theta_{k1}^{(l)}, ..., \Theta_{ki}^{(l)}$ are the synaptic weights relating to neuron k from layer l to $l+1$ and $a_1^1, a_1^2, ..., a_k^l$ being the activation of neuron k in layer l. The number of neurons in the hidden layers are n_H and the number of neurons in the output layer is n_O.

For our network, we are using the Softmax function built into Python for computation of the activation's in the output layer. The strength of this activator is that it forces our outputs to sum to one, meaning the output will be a n_O dimensional vector containing the probabilities for a company belonging in each of the credit rating categories. A visual representation of the mathematics is shown in figure 2.

2.2.3 Forward Propagation

Forward propagation means that the network is moving forward through each layer and doing calculations in each neuron. It is done by first initializing the model with randomized weights. This helps breaking the symmetry of the model and prevents the hidden neurons from generating the exact same values, which would be the case if the weights were initialized at for example 1. The forward propagation method then moves forward through the layers activating the functions in each layer while keeping the weights fixed, finally resulting in output probabilities summing up to one for our six classes of credit ratings. As told before, the benefit of using a neural network is that the model can learn from experience by changing the synaptic weights between the neurons to better fit the data and this is done through the method backpropagation.

2.2.4 Backpropagation

To understand this method, first we have a set of learning samples $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), ..., (x^{(j)}, y^{(j)})$ where $x^{(j)}$ is a vector of all inputs for sample j and $y^{(j)}$ is the corresponding output for sample j. For example, if Microsoft, one of two AAA rated companies where in the learning sample, $x^{(j)}$ would be the input variables for Microsoft and $y^{(j)}$ would be the label AAA.

A cost function measures the difference between the actual output and desired output and the bigger the difference the bigger the cost is. The goal of the backpropagation algorithm is to minimize the cost by minimizing the difference between actual and desired output. This is done by changing the weights on each neuron and minimization of the function is done through minimizing the derivative of the cost function. In short, what backpropagation does is to calculate backwards through the layers to find the difference in each neuron and then changing their weights to better fit the data. This is done until a local minimum is reached and the derivative of the cost function is close or equal to zero. The minimum is not global since the cost function is not convex, but rather represented as a three-dimensional space that can be resembled to a landscape with hills and valleys. Since there are several valleys, there can therefore by several minima's. To avoid being stuck at a bad local minimum, the model initializes several times until the best possible convergence is reached, that is it finds the deepest valley. For neural networks, there are several cost functions to choose from depending on what the goal is with the model and for ours the cost function is a part of the Softmax function built into Python which outputs the categories as probabilities.

2.2.5 Generalization

Approximately 35 % of our data set is the holdout sample and has been withheld during the entire learning session. This part of the data set is therefore equal to meeting new unknown samples and test the models ability to generalize. A problem with neural network is that it can easily overfit the training data meaning it will be able to explain 100 % of the samples in the training set by creating a nonlinear function which covers every input and output, but then be bad at generalizing. To prevent this a technique called regularization that is built into the cost function of Softmax is used and it penalizes large weights leading to less overfitting and better generalization. The model is then tested on the holdout sample and the models effectiveness is measured in how many correct samples it can predict.

2.3 Random Forest

2.3.1 Decision Tree

Decision Trees are rule-based systems for classifying data and works by splitting data into a series of questions to narrow down the possible choices. It can be resembled to a tree with branches continuously splitting until forming a tree crown.

The concept of using decision trees to get a better understanding of decisions and their impact has been around for a long time but in modern times referring to algorithms the credit goes to Morgan & Sonquist (1963) for the first publication of a regression tree where people were divided into annual earnings categories based a set of variables, for example if they were a college graduate or not. Since then several publications have been done on the subject but it was the publication of Classification and Regression Trees (CART) by Breiman et al. (1984) which ignited the interest into the subject of decision tress. This book provides a thorough explanation of decision trees by developing on both the practical and theoretical sides of the subject and the ideas published in this book are still used today. The strength of using a decision tree model is that requires little preparation of data, is computationally cheap and is a white box model. A white box model means that it is easy to observe and understand what happens in each step, which is the opposite of a black box model such as a neural network. The disadvantages are that the model easily overfits the data and is sensitive to noise in the data, meaning small variations can lead to entirely different trees. A single tree is thus referred to as a weak learner, meaning that in most cases its predictions are just barely above chance. These problems are mitigated using bagging and random forests, which will be explained later on, and in turn random forest becomes one of better machine learning algorithms to date. For our study, we will be using a model based on the methodology of classification trees since we are trying to get our model to predict a set of predetermined classes. The CART methodology for classification trees works does just like the MLP classifier in neural network by requiring learning samples to create a model.

The difference from a neural network is that a decision tree splits the data set into a series of binary questions based on the input variables with the goal of asking a question where the majority of the observations in one of the answers belong to the same class. That means:

1. If all observations in the set D is of the same class or if the number of observations in D is very small the tree will reach an endpoint.

2. If D is very large and contains several classes, the decision tree algorithm will split the set of observations based on the best fitting input variable and repeat step one until an endpoint is reached.

An example of a classification tree would be determining if a person is eligible for a loan and examples of questions would then be: Does the person have a job, more than X years in present job, are there any previous remarks of payments due, is current debt more than X percent of the person's assets. These questions would be generated by finding the best fitting input variables and the question created would be represented as a splitting threshold in the tree. From each question two new nodes would be created until finally ending in a binary decision of eligible or not eligible for a loan. Even though the classification is divided up into binary answers, it can classify more than two classes. For example, if one were to identify fruits, a certain branch on the tree could go toward round fruits such as apples and oranges while another branch could go toward bananas.

2.3.2 Gini Impurity

In order to choose the optimal parameters to split at, the Gini impurity rule is used. What the algorithm does is to maximize homogeneity in each split, meaning it want as many classes of the same type as possible in each node, this in turn minimizes the probability of misclassification. A simple example would be if previous remarks of payments due was the only deciding factor for getting a loan or not. If previous remarks were noted as a 1 in the data set and no remarks were noted as 0, the gini rule would choose the split threshold to be input $x_{previous\,remarks} \leq 0$, thus maximizing homogeneity in each node.

If the learning sample $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), ..., (x^{(j)}, y^{(j)})$ where $x^{(j)}$ is a vector of all inputs for sample j and $y^{(j)}$is the corresponding output for sample j to train. Let t_{parent} be the parent node to the two child nodes t_{left} and t_{right} with the goal of finding the best splitting threshold $x_i^{(j)R}$ for the input variable x_i where $i = 1,...,M$ variables in vector j, resulting in $x_i^{(j)} \leq x_i^{(j)R}$. The best splitting value is the one that maximizes homogeneity at nodes t_{left} and t_{right} and is measured by the gini impurity function $i(t)$ where t is the node. The impurity function is constant for any t_{parent} and to maximize for homogeneity would therefore be to maximize $\Delta i(t)$, making the impurity as low as possible in the child nodes Timofeev (2004).

$$\Delta i(t) = i(t_{parent}) - Pr_{left} i(t_{left}) - Pr_{right}(t_{right}) \tag{10}$$

where $\Delta i(t)$ is the difference between the impurity measure for the parent node t_{parent} and the weighted sum of the impurity measures for the children nodes t_{left} and t_{right}. Pr_{left} are the proportion of samples that goes into node t_{left} from t_{parent} and Pr_{right} is the proportion of samples that goes into node t_{right}. The gini function $i(t)$ is defined in equation 11.

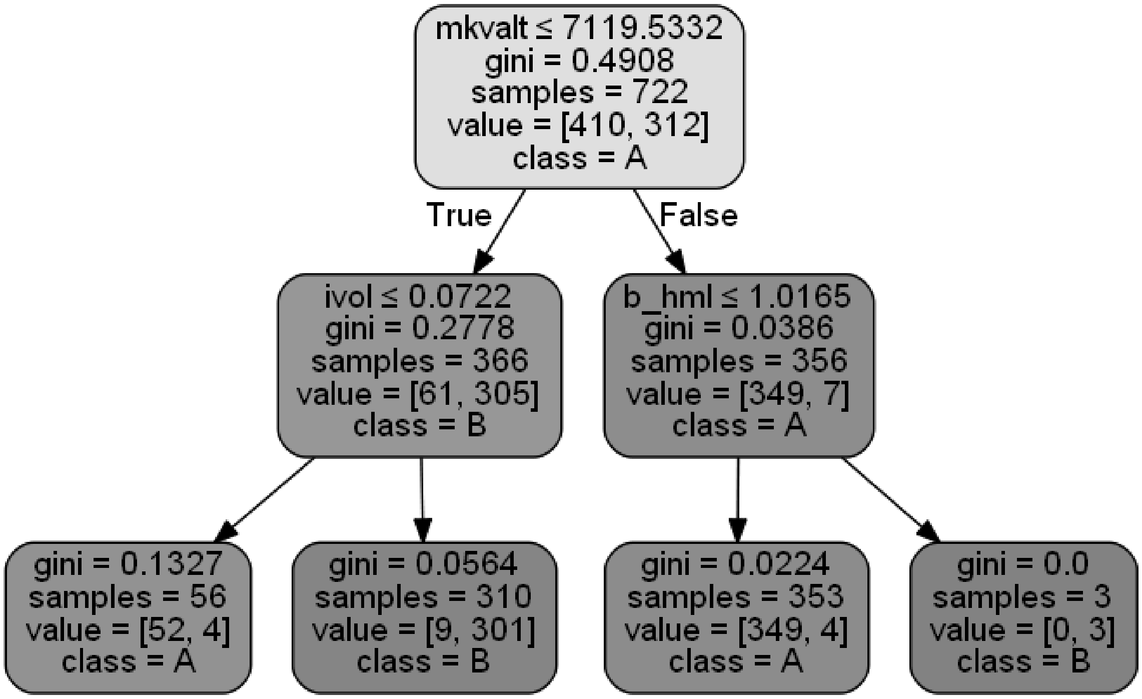
$$i(t) = 1 - \sum_{k=1}^{K} (p(k|t))^2 \tag{11}$$

where $p(k|t)$ is the percentage of observations of class k in node t and k is $1,...,K$ classes.

2.3.3 A Simple Decision Tree

The structure and inner workings of a decision tree can be viewed in figure 3. The decision tree is trained on 722 firms which are rated only A or B and it has been limited to only three levels. The actual decision trees used will have more possible ratings and a lot more levels, in order to make better decisions. The first decision rule of the tree is whether the firm has a market value of above 7119.5 million USD and this value has been chosen for the split since it maximizes homogeneity in both children nodes using the gini impurity algorithm. When the variable is a continuous one such as market value then an iterative process is used until the value which minimizes the gini impurity has been found. If the market value is less than the threshold value in the sample, it gets placed in the left node, while if false it gets placed in the right node. In the left node, there are 366 samples with 61 having received the rating A and 305 having received the rating B. The homogeneity of these firms are consequently higher than with the firms we started with. The increase in homogeneity can also be seen by the decrease of gini impurity which went from 0.49 in the first node to 0.28 in the second one, thus the sample is purer than before. The decision in this node regards the idiosyncratic volatility of the firms equity and as with the first decision threshold, firms for which the statement is true will be assigned to the left child node and if false it will be assigned to the right child node. As the tree has been limited to only three levels the most probable rating in this node will be the rating chosen, while a larger decision tree with more levels would

Figure 3: A Simple Decision Tree



continue until all samples in a node belong to the same class. Following the decision from the top we can see that the model will classify small firms with high idiosyncratic volatility as rated B and small firms with a low volatility as A. Going down the other branch of the tree, big firms with a high minus low beta will be predicted to receive a B and big firms with a low high minus low beta will be predicted to receive an A.

2.3.4 Bagging

Bagging is short for bootstrap aggregating and was first proposed by Breiman (1996). It is a method for reducing the variance of a models estimation prediction function. What bagging does is to bootstrap multiple versions a predictor, for example a classification tree. The bootstrap method draws random samples with replacement from the original data set creating new data sets of the same size. The sampling with replacement means that there will be duplicates within these new data sets and it is therefore not equal to the original one. The new data sets are then used to create new decision trees, thus creating an ensemble of trees. The trees are given a vote with equal weight for all trees, with the vote being their result of the classification and the majority vote becomes the output of the model. By letting an ensemble of trees vote instead of a singular tree, variance can be reduced, which improves the classification accuracy of the model. The ensemble of trees can be thought of as the board of a firm consisting of several different individuals. Each of the members have different experiences (samples) and based on these experiences each person will make a decision. The votes of all the board members are counted and the majority wins. The advantage of having several board members is that if one of them happens to make an absurd decision it is likely that the others will not and the poor decision of the one member will not win, making the decision more stable and less affected by unusual experiences (outliers) by one or a few board members. When Breiman used this method on classification trees for several data sets, he saw a reduction in misclassification rates ranging from 6% to 77%.

Given a set of learning samples $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), ..., (x^{(j)}, y^{(j)})$ where $x^{(j)}$ is a vector of all inputs for sample j and $y^{(j)}$ is the corresponding output for sample j. For $b = 1, ..., B$ where B is the number of bootstraps, sample with replacement B data sets.

Then train B decisions trees using the learning samples drawn in random order with replacement $(x_1^{(1)}, y_1^{(1)}), (x_1^{(2)}, y_1^{(2)}), ..., (x_B^{(j)}, y_B^{(j)})$ and take the majority vote as the prediction.

So, to take the previous example of determining if a person is eligible for a loan. With bagging, many decision trees would be generated from the original data set, thus creating an ensemble of voters. The input data would then go through all trees and let each one vote whether a person is eligible or not eligible for a loan. The category with over half the votes would be the majority vote which the model outputs.

2.3.5 Creating a Random Forest

A random forest is a refinement of bagging with the difference being that a random forest builds a large set of de-correlated trees and then averages them. The point of doing this is to further reduce variance. The idea of random forests was first published by Breiman (2001) and is a continuation of his work on bagging and further increases the accuracy for predictions.

Random forest draws bootstrap samples from the training data just like in bagging, but at each node in the tree it selects a random set of variables that is less than the total number of input variables, thus if M is total amount of input variables $m < M$ variables are chosen at random. The variable that provides the best split using gini impurity gets used to do a binary split on that node. This process is then repeated until reaching an endpoint. This prevents the bootstrapped decision trees from following the same splitting rules and makes for a stronger model when generalizing, by introducing randomness into the model and could be compared to a group of board members all having different experiences.

This is the algorithm we will be using for our data set to predict credit ratings and the computation is done using RandomForestClassifier in Python. The model requires little tweaking and is one of the leading algorithms in classification today. Compared to neural networks which is a black box model, where it is hard to know what is going on between the input and output layers, the random forest provides tools for understanding what input variables are important.

2.3.6 Variable Importance

The effects of variables are of course more complicated to distinguish in a random forest then in a single decision tree but there are still methods for understanding which variables are the important ones in the classification process. The most common one is referred to as gini importance or mean decrease gini. The importance of a variable is the sum of the gini decrease in all nodes in which the variable appears weighted by the probability of reaching that specific node. In a random forest the variable importance is averaged over all trees to get the importance of each individual feature in the forest (Louppe et al. 2013).

2.4 Ordinal Classification

The prediction of credit ratings is an ordinal classification problem as credit ratings are not on an interval scale but are still containing ordinal information. The artificial neural network and the random forest classifiers previously outlined does however not take this ordinal information into account. To classify ratings without considering that an AA rating is higher than an A rating means a lot of valuable information will be lost which would consequently affect the predictive abilities of the models negatively. To solve this, we will use the approach outlined by Frank & Hall (2001). Their strategy is to convert the multi-class classification to a series of binary classifications. As explained before the output of our machine learning models are a vector containing the probabilities for each rating. To use ordinal information of the rating each probability is converted to the probability of the rating being higher than a certain rating. The target variables of our model will therefore be vector of length $k-1$ where k is the number of possible ratings which contains dummies whether the rating is above a certain threshold. Assuming there are only three possible ratings, the target value of a firm given the lowest rating would thus be a vector of all zeroes $\begin{pmatrix}0, & 0\end{pmatrix}$ and a firm given the highest would result in a vector of all ones $\begin{pmatrix}1, & 1\end{pmatrix}$. A firm given the middle rating would result in a vector of a one and a zero $\begin{pmatrix}1, & 0\end{pmatrix}$ as the rating is above the lowest rating but not above the second lowest. The output of our models will thus be the probability of each dummy being a one. An output of $\begin{pmatrix}0.89, & 0.32\end{pmatrix}$ should therefore be interpreted as an 89% probability of being above the lowest rating and a 32% probability of having a rating above the second lowest rating. The probability of this fictional firm to be given the lowest rating is consequently 11% (1-0.89) a 57% (0.89-0.32) probability of having the middle rating and a 32% probability of having the highest rating. In more general terms the probability for firm i the receive a rating is given the equations 12, 13 and 14 where $x^{(j)}$ is the vector of inputs for firm j, $y^{(j)}$ is the actual rating for firm j and R_m is rating m with the ratings ranging from 1 to k.

$$P(R_1|x^{(j)}) = 1 - P(y^{(j)} > R_{1|}|x^{(j)}) \tag{12}$$

$$P(R_m|x^{(j)}) = P(y^{(j)} > R_{m-1}|x^{(j)}) - Pr(y^{(j)} > R_m|x^{(j)}) \quad 1 < m < k \tag{13}$$

$$P(R_k|x^{(j)}) = P(y^{(j)} > R_{k-1}|x^{(j)}) \tag{14}$$

The firm is then given the rating with the highest probability. This classification procedure is used to transform both the artificial neural network and the random forest to become ordinal classifiers. This means that the ordinal nature of the ratings is utilized in a more efficient way and will consequently be an improvement to our models.

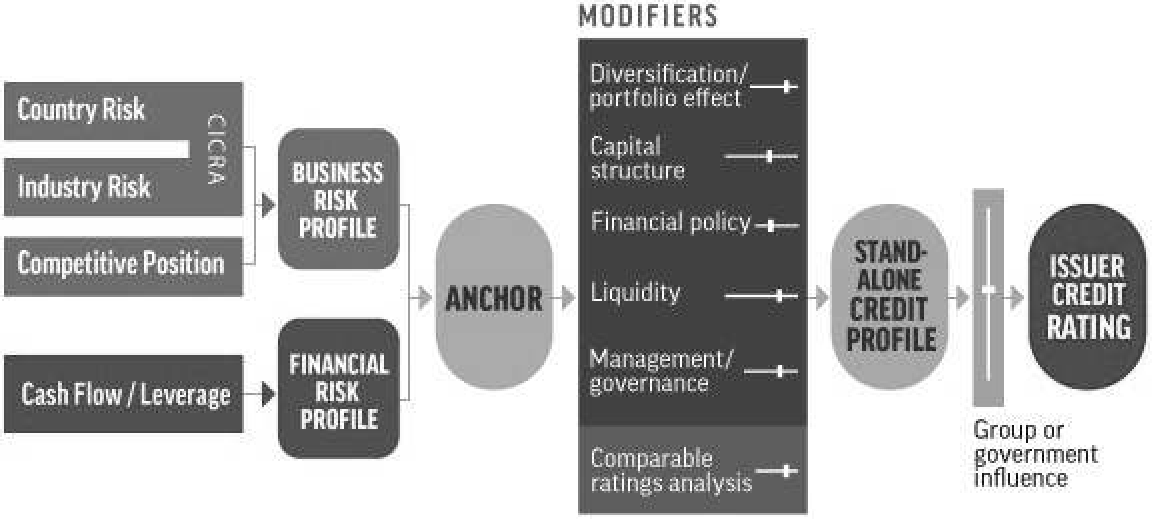
2.5 Variable Selection

2.5.1 S&P Corporate Criteria Framework

The variable selection in previous research has varied greatly and there is no general conclusion regarding which variables to include. One important thing to remember is what it is in fact that the models are modeling. The aim of this paper is not to find the perfect model for classifying different firms based on their credit risk or bond spreads, but rather to model the classifying process of S&P. The variable selection must therefore be based upon their credit rating process. The exact procedures of the ratings agencies are of course not made available to the public but some information about the methodology is available. The rating process of S&P when rating a non-financial corporate issuer follows according to themselves the *Standard & Poor's Corporate Criteria Framework* (figure 4). The analysis begins with assessing the firm's business risk profile which consists of country risk, industry risk and competitive position. The financial risk profile is created thereafter and these two-risk profile forms the anchor of the rating. Based on the anchor the modifiers can then notch the rating up or down depending on the outcome of the analysis. A firm with an anchor of BBB would for example be given two positive notches if its capital structure were to be considered a one on a five-grade scale with one being the most positive. After the modifiers are considered and the anchor has been notched, the stand-alone credit profile can be made. This is an assessment of the credit risk of the firm without considering any external support which the company might receive from parent companies or other owners. Next the firms group is identified and a group credit profile is established and the firm's strategic importance to the group is established (Standard Poor's 2016*c*). Lastly, the stand-alone credit profile and the group influence is combined to form an issuer credit rating.

The country risk aims to capture the risk that arises for a non-governmental entity to do business within a specific country (Standard Poor's 2016*b*). The risk of doing business within a country is analyzed by looking the economic risk, the institutional risk, the financial risk and the payment culture/rule-of-law risk. A firm which has a high percentage of its revenues from risky countries will be considered more likely to default in its bonds. The industry risk is determined by looking

Figure 4: S&P Corporate Criteria Framework (Source: (Standard Poor's 2017))



at two different factors, the cyclicality of the industry and the competitive risk and growth (Standard Poor's 2016*d*). The cyclicality is important as it highlights the firm's ability to service its debts during a cyclical downturn. The competitive risk and growth looks at the forces operating within the industry. In other words, what the outlook is for making profits and if there are any treat to the current or future profitability of the industry. The competitive position of a company relates to company specific factors which increases or offsets the country risk and the industry risk. This considers the firms competitive advantage, scale, scope and diversity, operating efficiency and profitability. The profitability is measured both as the level of profitability and as the variation of the profitability. The numbers for the profitability are five year averages which includes forecasted numbers for the two coming years, and the variation of profitability consists of the data from the last seven years as this is usually an adequate amount of time to capture a business cycle. The profitability is evaluated in the context of the industry within which the firm operates.

The financial risk profile measures the firm's ability to service its debts. According to Standard & Poor (2017) the pattern of cash flow generation in relation to its cash is the best indicator of financial risk. The analysis uses a wide range of ratios but it relies heavily on cash flow ratios. The relative importance of certain numbers and ratios depend on the values of other ratios and adjustments can also be made depending which industry the firm operates within. The diversification/portfolio modifier becomes relevant for companies with multiple product lines. This effect can improve a company's rating if it is believed that the multiple product lines within the company increases its credit strength. Whether the diversification leads to increased credit strength is largely dependent on the correlation between the different product lines. For a form to be notched up it must be believed that the diversification will lower earnings volatility throughout the business cycles. The capital structure modifier considers risk factors which are not related to cash flow or leverage. This can for example be the maturity of the loans or currency risks relating to bonds in foreign currencies. The cash flow consideration takes into account estimates for two years into the future but do not take into account risks further away in the future. To asses longer term risks S&P looks at the financial policy of the firm. The assessment looks at the financial discipline, the likelihood of managements leverage tolerance and likelihood of event risk, and the financial policy framework, firms with more comprehensive and transparent financial policies are more likely build sustainable credit quality. A firm will default on its debt no matter how profitable it is if the liquidity is inadequate which makes liquidity an important modifier. The liquidity assesses the firm ability to service its debts and its working capital needs for the next twelve months (Standard Poor's 2016*a*). The liquidity modifier effectively puts a cap on the ratings which can be given to certain firms as all investment-grade companies must have liquidity which is considered at least adequate by S&P. The liquidity ratios are absolute and not relative to industry peers as inability to service its debt will lead to default no matter the industry. The management/governance modifier is a broad analysis which encompasses strategic positioning, risk management/financial management, organizational effectiveness and governance. The last modifier compares the rating given to a firm to similar firms. This last step takes a holistic view and looks at all the other risk profiles and modifiers to make some minor final adjustments before the stand-alone credit profile is created.

2.5.2 Variables

The variables included in the models considers variables used in previous research and tries to capture the factors considered during S&Ps credit rating process. The variables are listed as they appear in process with variables covering the business risk profile first followed by the financial risk profile and lastly the modifiers.

Business Risk Profile

Market Value (Million USD): A high market value of a company is often related to older more established companies with more varied sources of revenue and can therefore be thought of as capturing some of the business risk and diversification (Blume et al. 1998). Market value has been found to have a positive relation with credit ratings in previous studies (Kaplan & Urwitz 1979, Horrigan 1966).

Net Operating Income to Sales: This profitability measure is included to capture the competitive position of the firm as firms with a good competitive position will be more profitable.

Variation of EBITDA: The coefficient of variation of EBITDA using the previous seven years of data. Previous research have used variation of sales (West 1970, Kaplan & Urwitz 1979) but according to Standard & Poor's (2017) the variation of profitability is of greater importance than variation of sales and is therefore included when assessing the competitive position of a firm.

Market Beta: One important factor in determining industry risk is the cyclicality of the industry Standard & Poor's (2017) and the market beta from the Fama-French three factor model (Fama & French 1997) is therefore included.

High Minus Low Beta: The market beta has been included in previous research (Kaplan & Urwitz 1979, Blume et al. 1998) because variations due to market variations and cyclicality should be interpreted differently than variation due to firm specific factors. The same logic can be applied the other risk factors included in the Fama-French three factor model and the high minus low beta is therefore included in the model.

Small Minus Big Beta: By the same logic as with the previous variable the small minus big beta from the Fama-French three factor model is included.

Idiosyncratic Volatility: The standard errors of the Fama-French model is included as a measure of volatility which is related to firm specific factors. While the risk factor betas are related to the industry risk, the idiosyncratic volatility is more correlated to the firm specific risk in the business risk profile.

Financial Risk Profile

Debt to EBITDA: Debt to EBITDA is capturing the companies leverage and ability to service its debts. The ratio is also considered as one of the core ratios used by Standard & Poor's (2017) when assessing the financial risk profile.

Interest Coverage Ratio: The interest coverage ratio is defined as EBITDA divided by the total interest and related expenses. The ratio is indicative of a firm's ability to service its debts and have been found to be correlated with credit ratings in previous research (Blume et al. 1998).

Debt to Assets: The ratio of debt to assets is included as it captures the leverage of the firm and has been found to have predictive abilities in previous research (West 1970, Blume et al. 1998).

Cash Flow from Operation to Debt: The ratio of cash flow to total debt is one of the supplemental ratios used by Standard & Poor's (2017) when assessing the financial risk profile.

Modifiers

Dividends: The dividends to assets captures some of the firm's financial policy as well as the competitive risk and growth since firms in mature industries are more likely to have higher dividend payouts.

Liquidity: The liquidity modifier considers the firm's ability to service its short-term debt commitments and the ratio of current assets to current liabilities will be used to capture this. This ratio has been found in previous research to be correlated to a firm's credit rating and liquidity (Belkaoui 1980).

We do not in any way believe the above-mentioned variables to be capturing all considerations made by the rating agencies nor that these are the exact variables used by the agencies. The beta variables are for example not likely used by agencies themselves but it does on the other hand capture some of the information considered in the process. The rating agencies do not only look at the data for the current year which is why all financial ratios are three-year rolling averages. S&P uses five year rolling averages including the two previous years, the current year and predictions of the two coming years (Standard Poor's 2017). As predicted values for future years are not available to people outside of S&P the closest approximation is the three year moving average. An important consideration is the industry in which the firm operates and all financial ratios (except for the liquidity ratio) will be adjusted for industry. The adjustment will be done in the same way as previous research Horrigan (1966), Kaplan & Urwitz (1979), Blume et al. (1998) where each industry is defined by its two number sic code. An average weighted by market value is calculated for each ratio and the adjusted was made by taking the ratio minus the weighted average of the industry divided by the weighted average of the industry.

2.6 Sample Selection

The sample consists of all firms with a S&P long-term issuer credit rating and annual financial data in the Compustat IQ database during the period 2010-2015. The credit ratings are by month and was matched to the month after annual financial data was released. For example, if the financial data was made publicly available in February the long-term issuer credit rating of March would be used. This was done to allow S&P to incorporate the information of the annual data into their ratings. The risk factors from the Fama-French model was taken from the Beta Suite of the Wharton Research Data Service. This generated a data set of 4560 observation with 910 unique firms. The rating process of financial and insurance companies differ significantly from other corporations (Standard Poor's 2017) and these companies were thus removed from the data set leaving 3785 observations from 755 firms. After removing observations with missing data, sample was reduced to 3487 observations from 697 firms. From this sample a holdout sample consisting of 1200 observations was randomly generated (interested readers may request a list of firms in the holdout sample by the authors). The holdout sample was created so that even though each firm may have observations from different years in the data set, all observations from a company is in either in the holdout sample or the learning sample to avoid biased results. The learning sample consists of

2287 observations from 451 firms and the holdout sample consists of 1200 observations from 246 firms.

Previous research have mainly used the rating of specific bond issues (Horrigan 1966, West 1970, Kaplan & Urwitz 1979, Belkaoui 1980) but this study is instead using the long-term issuer credit rating of the issuer. The rationale behind choosing the rating of the issuer instead of the issue is based on both availability of data and the rating process of issues. The process of rating single issued starts with the rating of the issuer and the issue can thereafter be notched up or down depending on the legal status of the specific issue (Standard Poor's 2008). The credit rating of a single issue is therefore anchored to the rating of the issuer. To correctly be able to predict the rating of the issuer is consequently a vital step in predicting the credit rating of an issue. To use the issuer rating instead of the issue rating also avoids the problems of different maturities of bonds, hybrid bonds, subordination of bonds and secured debt.

S&P use nine different rating categories for their long-term issuer credit rating (AAA, AA, A, BBB, BB, B, CCC, CC and C) and the rating D for companies which have not met their obligations (Standard Poor's 2016*e*). The ratings can also include a plus or a minus sign to show relative standing within the category. Previous research have not included the relative standing within the categories (Horrigan 1966, West 1970, Kaplan & Urwitz 1979, Kim et al. 1993) and for example regulatory agencies do not recognize the plus and minus signs in their considerations which means that the discrete benefits are the greatest within the broader rating categories (Kisgen 2006). Due to a low number of observations in the highest rating category (AAA) this category was merged with the second highest category (AA). The three lowest categories (CCC, CC and C) did also contain a low number of ratings and these were thus combined with the firms who have not met their obligations (D) and form the lowest category of our model. For simplicity, the lowest rating category will from here on be referred to as CCC even though it also contains the ratings CC, C and D.

3 Results

3.1 Descriptive Statistics

There are six possible categories for each firm in our sample and the distributions of the two samples are shown in table 1. The distribution between categories are clearly not identical as a lot more firms are rated BBB as compared to for example AAA/AA. The distributions of ratings are however similar in the learning and the holdout sample. Even though some categories are clearly more common in our sample a classifier which only predicted the most common rating would correctly predict only 36.67% of the firms in the holdout sample, which can be compared to the study made

Table 1: Distributions of Ratings

	Learning Sample		**Holdout Sample**	
Rating	Frequency	%	Frequency	%
AAA/AA	79	3.45%	28	2.33%
A	410	17.93%	209	17.42%
BBB	853	37.30%	440	36.67%
BB	608	26.59%	287	23.92%
B	312	13.64%	226	18.83%
CCC	25	1.09%	10	0.83%
Total	**2287**	**100%**	**1200**	**100%**

by Kaplan & Urwitz (1979) in which 56.25% of the holdout sample would have been correctly predicted using this simple classifier. An overview of the variables and the complete distribution of the ratings is available in table 9and 10 in the appendix.

3.2 Ordered Probit

The predictions of the ordered probit model is shown in table 2 and the estimated coefficients are available in table 12 in the appendix. The model correctly classifies approximately 58% of the holdout sample and approximately 98% of all firms in the holdout sample was classified within one category of their actual rating. One rating that stands out in terms of type I and II errors is the BBB rating. The rating is the most common one with 37% of the firm being given this rating in the holdout sample (table 1). The low type I error indicates that firms which have received the BBB rating will probably be correctly predicted by the model, while the type II error is the percentage of firms who have incorrectly received a BBB rating. The big discrepancy between the two numbers means even that though firms which have been given BBB ratings are often correctly predicted, a firm predicted as BBB by the model is not that often correct. The reason is that the model overestimates the number of BBB in the holdout sample. The model predicts the BBB rating for 54% of all firms while only 37% of the firms in holdout sample was given this rating. The low type I error of the BBB rating are therefore not a result of the model being good at recognizing this rating but rather a consequence of the models overestimation of the BBB rating. The tendency of the ordered probit model to overestimate the occurrence of the most common rating was also present in the model used by Kaplan & Urwitz (1979).

The opposite effect is true for the A rating as the type I error is high while type II error is low (table 2). The reason for the low type II error is consequently the fact that few firms are given the A rating by the model. Only 9.5% of firms in the holdout sample are predicted to be given the rating A while the actual number 18% (table 1) which is almost the double. Seeing as the ratings BBB and

Table 2: Predictions of the Ordered Probit Model

	Predicted Rating						
Actual Rating	AAA/AA	A	BBB	BB	B	CCC	**Type I Error**
AAA/AA	**11**	16	1	0	0	0	**61%**
A	11	**70**	128	0	0	0	**67%**
BBB	5	22	**357**	51	5	0	**19%**
BB	2	5	102	**132**	45	1	**54%**
B	0	1	6	83	**125**	11	**45%**
CCC	0	0	0	0	7	**3**	**70%**
Type II Error	**62%**	**39%**	**40%**	**50%**	**31%**	**80%**	**42%**

Note: Type I error is defined as the total number of firms given a specific rating divided by the number of these firms which were incorrectly predicted to get another rating. Type II error is defined as the number of firms predicted to be given a certain rating divided by the number of firms which were not given that rating.

A are next to each other in the scale it is not unreasonable to assume that the model is incorrectly predicting the BBB rating when it should in fact be the A rating. The highest and the lowest rating are the most unusual in the holdout sample (table 1). The type I and type II errors of the highest and the lowest category are high compared to the other ratings (table 2) but the two types of errors are however not different from each other. The model is thus not over or underestimating the number of firms in these ratings, but is on the other hand not doing a good job of distinguishing firms with AAA/AA or CCC rating from the rest of the firms.

3.3 Artificial Neural Network

The predictions of the artificial neural network are shown in table 3. The model correctly classifies approximately 63% of the ratings in the holdout sample and the prediction was within one category of the correct rating for approximately 99% of the firms in the holdout sample. The CCC rating stands out with a high type I and type II error and the classifier is thus doing a poor job differentiating between these firms and the other ones. One possible explanation is the low number of firms rated CCC in both the learning and the holdout sample (table 1) and the lack of observations results in poor predictive ability. On the other side of the scale is the AAA/AA rating has a high type I error but a low type II error. The reason for this is that the model is underestimating the frequency of the AAA/AA ratings. In general, the prediction accuracy of firms rated with one of the two most unusual ratings (AAA/AA and CCC) were lower than for firm given any other rating. The classifier is also overestimating the occurrence of the BBB rating in the holdout sample a little bit, but does apart from this not show any clear bias in the predictions.

Table 3: Predictions of the Artificial Neural Network

	Predicted Ratings						
Actual Ratings	AAA/AA	A	BBB	BB	B	CCC	**Type I Error**
AAA/AA	**13**	15	0	0	0	0	**54%**
A	5	**107**	97	0	0	0	**49%**
BBB	0	47	**326**	65	2	0	**26%**
BB	0	4	77	**156**	50	0	**46%**
B	0	1	3	71	**146**	5	**35%**
CCC	0	0	0	0	8	**2**	**80%**
Type II Error	**28%**	**39%**	**35%**	**47%**	**29%**	**71%**	**37%**

Note: Type I error is defined as the total number of firms given a specific rating divided by the number of these firms which were incorrectly predicted to get another rating. Type II error is defined as the number of firms predicted to be given a certain rating divided by the number of firms which were not given that rating.

3.4 Random Forest

The random forest classifier is the one which achieves the highest classifying accuracy with 66% of the firms correctly classified in the holdout sample and 99% of the firms within one rating from the actual rating (table 4). As with previous classifiers the random forest struggles with the highest and lowest rating. Both the AAA/AA rating and the CCC rating have high type I errors but comparatively small type II errors, which indicates that the model is underestimating the occurrence of these two ratings in the holdout sample. The most likely reason for this is due to the fact the AAA/AA and CCC ratings only makes up a few percent of the learning and holdout data sets. While training the model, a low amount of learning samples in these categories will mean the model rather favor adapting towards the larger classes. As explained earlier, the random forest model relies on optimal splits of data to minimize impurity and thus maximize homogeneity in the samples. This means that a class like AAA/AA which only makes out 2.33 % of the learning data set will have a very low probability of being split since it unlikely to be the split that achieve maximum homogeneity on both children nodes and thus minimizing impurity. This in turn affects the generalization of the model for samples which belong in these categories, since the splitting rules will be adapted for categories with more samples in them. This can be seen in the table 4 where our model predict 20 out of 28 of the AAA/AA samples as A, which can be due to our model understanding that the data should with high probability be at least A but due to the low amount of learning samples in AAA/AA being unsure about predicting them in that category and thus settling for A. Unsurprisingly BBB is once again the category with the lowest type I errors being only 23% being the category with most predictions in. What is surprising is the overall low type I and II errors across categories A down to B, which shows that this model is quite good at

Table 4: Predictions of the Random Forest

	Predicted Ratings						
Actual Ratings	AAA/AA	A	BBB	BB	B	CCC	**Type I Error**
AAA/AA	**6**	20	2	0	0	0	**79%**
A	7	**99**	103	0	0	0	**53%**
BBB	0	26	**341**	71	2	0	**23%**
BB	0	2	72	**189**	23	1	**34%**
B	0	1	3	66	**156**	0	**31%**
CCC	0	0	0	1	8	**1**	**90%**
Type II Error	**54%**	**33%**	**35%**	**42%**	**17%**	**50%**	**34%**

Note: Type I error is defined as the total number of firms given a specific rating divided by the number of these firms which were incorrectly predicted to get another rating. Type II error is defined as the number of firms predicted to be given a certain rating divided by the number of firms which were not given that rating.

correctly predicting credit ratings.

3.5 Comparison of models

All three models have shown ability in predicting credit ratings by all being better than simply choosing the most common rating, which would have resulted in 36.67% correctly predicted. But between the models, they have also showed significant difference in their predicting abilities with ordered probit correctly classifying 58% of all ratings, neural network 63% and random forest 66%. The models did all show strengths in different categories of the prediction with random forest being the best overall. The ordered probit model performs worse than both the neural network and random forest in all categories except for BBB, the largest of them, also resulting in it being the worst performing model.

3.5.1 Model Biases

The ordered probit model did show evidence of being biased towards the most common rating (BBB) to a much greater extent than the other two. The reason for this is that the ordered probit model appears to have problems distinguishing between A, BBB and BB rated firms and therefor predicts the one with most observations, while the other two classifiers appear to be better at distinguishing between these three categories. One probable reason is that there exist some non-linear relationships between the A, BBB and BB ratings which cannot be detected by the ordered probit model but is detected by the machine learning models.

Among the machine learning models, the neural network is significantly better at predicting the categories AAA/AA and CCC that has fewer samples in them. In these categories, the neural network manages to get 100% of its predictions within one category and correctly classify 46% of AAA/AA and 20% of CCC. This is much better than random forest which only managed to correctly classify 21% of AAA/AA and 10% of CCC and only 92% of the predictions where within one category. It is important to note here that the small sample size of these categories means that there will be large differences in error even if the difference in actual predictions are not that many. Also, the small sample means that it is hard to know how big the difference truly is since the performance could average out over a larger sample. What is most likely causing the difference between the models are the way they are designed. As mentioned before random forests rely on optimal splits to maximize homogeneity in the children nodes, and thus the model is biased towards the categories with larger samples since its more likely to maximize homogeneity by dividing the categories with large sets of samples into different nodes. The neural network on the other hand rely on interconnections between its neurons to find hidden features in the data and while it is hard to know what is going on inside the hidden layers, the model is less likely to as biased as random forests against ratings with few observations. Overall the best performer was the random forest with strong consistency in the categories A down to B, where the samples sizes were larger. In these categories, it outperforms the neural network in all categories except for Type I errors in predicting A. This shows that a rule-based model might be the most optimal way for predicting credit ratings and it is also one the most intuitive ways as well. The intuition is that these models can be seen as a group of credit raters looking over financial factors for companies and deciding on what factors best differentiates companies in the different categories.

3.5.2 Estimated Probabilities

Another important factor when looking at our three models is the probabilities which the models estimates. For a good model these probabilities would be close to the percentage correctly predicted in the sample. This has important implications about the models and their real-world use, since no one would trust a model which only gets for example 20 % correct predictions when it says its 90 % sure while a model which was 90 % correct when it was 90 % certain would be a very strong model which could be used in real financial applications. A comparison of the three models (table 5) show that the ordered probit model is in general outputting lower probabilities than the two other models, which is consistent with the higher rate of misclassifications. The probabilities of the ordered probit model are fairly consistent with the accuracy when the probabilities are between 40%-70%. For probabilities above 70% the accuracy starts decreasing which is likely due to the small number of observations. The neural network was the model with the highest probabilities but the outputted probabilities are in almost all cases higher than the actual accuracy. The probabilities of the neural network are consistent in that higher probabilities result in higher accuracy but the model is clearly too confident in the classifications it is making. The random forest was not as

Table 5: Distribution of Correctly Classified by Estimated Probabilities

	Ordered probit		**Neural Network**		**Random Forest**	
Probability	Observations	Accuracy	Observations	Correctly	Observations	Accuracy
0.3-0.4	0	-	2	0%	17	53%
0.4-0.5	172	49%	12	42%	91	51%
0.5-0.6	602	57%	149	45%	320	58%
0.6-0.7	368	65%	219	62%	317	68%
0.7-0.8	37	59%	286	60%	240	75%
0.8-0.9	9	44%	284	68%	158	73%
0.9-0.99	7	43%	221	74%	56	66%
1	5	40%	27	52%	1	0%
Total	**1200**	**58%**	**1200**	**63%**	**1200**	**66%**

confident in the predictions and the probabilities were consequently more in line with the actual accuracy of the model. In table 5 it is shown that the random forest classifier appears to be correct about the probabilities except for the highest and lowest probabilities, which are also the categories with the fewest number of observations.

The models have probabilities higher than the actual accuracy for predicted ratings with a high estimated probability. There are few ratings which have such a high probability and this could be one reason for the discrepancy between the predicted probability and the actual accuracy. Another reason could be that the very high probabilities are a result of outliers in the holdout sample for which the models incorrectly believe they are certain about the rating. Outliers in the holdout sample are similar to outliers in the learning sample, as the model has only seen these kinds of firms being given a certain rating it incorrectly predicts the rating as being much more certain than it is. Disregarding the higher probabilities, the correlation between higher probability and higher accuracy exists in all three models.

3.5.3 McNemar's Test

To test the robustness of the performance of the models it would be useful to see if the difference in performance is statistically significant. Due to the discrete nature of ratings an assumption of normally distributed variables cannot be made which invalidates some of the most common statistical tests. A non-parametric test is instead more appropriate as it does not make any underlying assumptions with regards to the distributions of the variables. Most of the previous research within credit rating predictions that has been referenced has only tested one model and has therefore not included any statistical test to compare the performance of models. The only exception is Kim et al. (1993) who used a 2x2 contingency table to test for significance. Later research within similar areas have however chosen the slightly different McNemar's test (Bensic et al. 2005) and this is the

Table 6: McNemar's Test on the Ordered Probit Model and the Artificial Neural Network

$$H_0 : P(\hat{y}_{ANN}^{(j)} = y^{(j)} \cap \hat{y}_{OP}^{(j)} \neq y^{(j)}) = P(\hat{y}_{ANN}^{(j)} \neq y^{(j)} \cap \hat{y}_{OP}^{(j)} = y^{(j)})$$

$$H_1 : P(\hat{y}_{ANN}^{(j)} = y^{(j)} \cap \hat{y}_{OP}^{(j)} \neq y^{(j)}) \neq P(\hat{y}_{ANN}^{(j)} \neq y^{(j)} \cap \hat{y}_{OP}^{(j)} = y^{(j)})$$

$$\chi^2 = 12.28^{***}$$

		Ordered Probit	
		Correct	Incorrect
Artificial Neural Network	Correct	590	166
	Incorrect	108	336

$^{***}p<0.001, ^{**}p<0.01, ^{*}p<0.05$

Note: $y^{(j)}$ is the actual rating for firm j and $\hat{y}_M^{(j)}$ is the predicted rating for firm j from model M

Table 7: McNemar's Test on the Ordered Probit Model and the Random Forest

$$H_0 : P(\hat{y}_{RF}^{(j)} = y^{(j)} \cap \hat{y}_{OP}^{(j)} \neq y^{(j)}) = P(\hat{y}_{RF}^{(j)} \neq y^{(j)} \cap \hat{y}_{OP}^{(j)} = y^{(j)})$$

$$H_1 : P(\hat{y}_{RF}^{(j)} = y^{(j)} \cap \hat{y}_{OP}^{(j)} \neq y^{(j)}) \neq P(\hat{y}_{RF}^{(j)} \neq y^{(j)} \cap \hat{y}_{OP}^{(j)} = y^{(j)})$$

$$\chi^2 = 27.33^{***}$$

		Ordered Probit	
		Correct	Incorrect
Random Forest	Correct	592	197
	Incorrect	106	305

$^{***}p<0.001, ^{**}p<0.01, ^{*}p<0.05$

Note: $y^{(j)}$ is the actual rating for firm j and $\hat{y}_M^{(j)}$ is the predicted rating for firm j from model M

test which will be used in this section.

The McNemar's test is used on paired dichotomous data in a 2x2 contingency table and tests whether there is marginal homogeneity between the two pairs of data. In other words, if there is any statistical significance between the accuracy of two models tested. The null hypothesis of the test is that the probability of model one being correct and model two being incorrect is the same as the probability of model two being correct and model one being incorrect, and this is tested against the alternative hypothesis that the probabilities are different. The test statistic of the McNemar's test is shown in equation 15 where A is equal to the number of times model one is correct and model two is incorrect and B is equal to the number of times model two is correct and model one is incorrect. Under the null hypothesis the test statistic follows a chi-squared distribution with one degree of freedom.

$$\chi^2 = \frac{(A-B)^2}{A+B} \tag{15}$$

Table 8: McNemar's Test on the Random Forest and the Artificial Neural Network

$H_0 : P(\hat{y}_{RF}^{(j)} = y^{(j)} \cap \hat{y}_{ANN}^{(j)} \neq y^{(j)}) = P(\hat{y}_{RF}^{(j)} \neq y^{(j)} \cap \hat{y}_{ANN}^{(j)} = y^{(j)})$

$H_1 : P(\hat{y}_{RF}^{(j)} = y^{(j)} \cap \hat{y}_{ANN}^{(j)} \neq y^{(j)}) \neq P(\hat{y}_{RF}^{(j)} \neq y^{(j)} \cap \hat{y}_{ANN}^{(j)} = y^{(j)})$

$\chi^2 = 4.88^*$

		Artificial Neural Network	
		Correct	Incorrect
Random Forest	Correct	661	128
	Incorrect	95	316

$^{***}p<0.001, ^{**}p<0.01, ^{*}p<0.05$

Note: $y^{(j)}$ is the actual rating for firm j and $\hat{y}_M^{(j)}$ is the predicted rating for firm j from model M

The result of McNemar's test can be viewed in tables 6, 7 and 8. The tests show that the accuracy of both the machine learning models are different from the accuracy of ordered probit model at the 0.1% significance level (table 6 and table 7). The test also showed a significant difference between the accuracy of the random forest and the artificial neural network but only at the 5% significance level (table 8).

3.5.4 Robustness of Results

The results presented in this thesis is dependent on the variables used and the design of the models and alterations to any of the two could consequently lead to different results. Firstly, the design of the models has been done in a way which follows recommendations in previous research but nothing states that these are the optimal way of designing the models. The ordered probit model has looked the same for a long time and it does not include any design decisions which means it is unlikely to be improved by any changes in design. The performance of the machine learning models does however vary with the design of the models which means that changing the model design can improve performance. It is mainly the performance of the artificial neural network which varies greatly with the architecture of the layers and activation functions, and it is thus likely that the result of the artificial neural network can be improved by better model architecture. The fact that the artificial neural network can be improved should not be seen as a strength of the model, but rather as a weakness. The results are highly dependent on the architecture and there is no consensus regarding the optimal design which makes the model difficult to use in a satisfactory way. The random forest is in comparison much easier to implement and use, requiring little optimization of model parameters.

The variables included in our models were chosen as they were believed to capture and the rating process of S&P and also following the conclusions of previous researchers. It is however likely that these are not the optimal choice of variables and that there exist some other set of variables which would capture the rating process even better and thus achieve a higher accuracy. Even though some other set of variables could perform better this would not necessarily change the main conclusion of this thesis as it could affect all the tested models in a similar way and consequently not change their relative performance significantly.

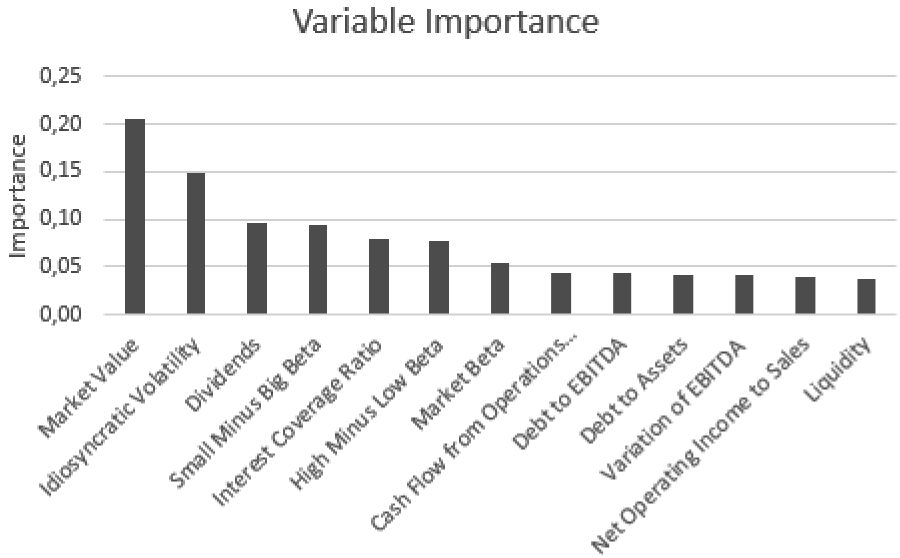
3.6 Importance of Variables

With the ordered probit model it is easy to get significance of all the included variables. The significance does however only provide information about if the variable has explanatory power but not how much explanatory power. This makes it difficult to understand the economic importance of the variables (Blume et al. 1998). The neural network is a black box model, meaning it is hard to know what is going on between the input and output layers. This makes it difficult to know whether a variable helps the model by uncovering some hidden feature in the data set or not. This is a great weakness of the model when trying to understand what variables are of importance when predicting credit ratings. The random forest does on the other hand provides the tools to see what input variables are of importance for the model.

The importance of each of the variables included were calculated using the mean decrease gini method outlined previously and the result is displayed in figure 5. From the graph, it is obvious that market value is the most important variable by a wide margin. The intuition behind this is that larger companies tend to be more stable and diversified with stable ratings, while smaller ones tend to be more prone to defaulting due to being concentrated within one area of business. The fact that market value is the most important variable is consistent with the findings of Pinches & Mingo (1975). The dividend variable is also one of the most important once in the model, and this variable does in some way capture some of the same considerations as the market value. Firms with high dividends are likely older more established firms who are operating within industries who are not expanding rapidly. A high dividend is also a signal from the management of the firm that they are confident in the firm's ability to cover its financial obligations.

One interesting observation from the variable importance is that the variables relating to equity volatility is more important than variables concerning leverage, profitability, liquidity or earnings volatility. The only financial ratios which appear to be as important as the equity variables are related to dividends and coverage. The equity volatility gives an indication of investors future expectations regarding the performance of the firm, and the result does consequently provide an indication of that S&P are more concerned with the volatility of future performance rather than current financial ratios or historic volatility. In some way, the market value and dividends are also incorporating the expectations of the future of investors and management respectively. The result

Figure 5: Variable Importance in the Random Forest



does consequently indicate that the issued credit ratings of S&P should be viewed more as an opinion regarding the future business risk of the firms rather than its current financial standing or even current business risk as that would have been reflected in some of the financial ratios.

In regards to the financial ratios the interest coverage ratio is the only who is as important as the other variables. Especially the liquidity ratios as well as the leverage ratios does not appear to be important factors when assigning credit ratings. One reason could be that liquidity is mainly a concern for firms which are close to default and there are few of these companies in the sample. Liquidity can consequently be an important consideration but only for a small number of firms which leads to the variable being given a low importance value.

4 Conclusion

The predictions of credit ratings have a long history with both methodology and variables improving along the way. Statistical models have long been the best performing alternatives with the ordered probit spearheading the way, but more recently as computational power has increased artificial neural networks have been shown to outperform the statistical models. Artificial neural networks are however only one of many machine learning models and one of the most opaque. We instead proposed the use of the random forest model which is more transparent and have shown good accuracy in similar tasks. In our study the random forest did indeed perform better than the two benchmark models with the random forest achieving a prediction accuracy of 66%, outperforming the artificial neural network and ordered probit with 62% and 58% accuracy respectively. The performance of the random forest was also statistically significantly different than the artificial neural network at the 5% level and the ordered probit at 0.1% level and showed better consistency than the other models in the difference between the estimated probabilities and the actual accuracy.

As for robustness, the performance of the models is dependent on the variables used and the design of the models, especially the neural network is highly dependent on the structure of the hidden layers. Other variables and a different design could likely improve the accuracy of the models somewhat but this would not necessarily alter the conclusion that the random forest is the preferred model.

Finally, the variable importance of the random forest model shed new light into what variables are of importance in credit ratings. While market value has been established as an important variable in many previous studies, our study is to our knowledge the first to provide evidence that equity volatility is a better indicator of credit ratings than financial ratios. Thus, the importance of market value and equity volatility seem be an important consideration for S&P when they are giving out their ratings. These exact variables may not be used by S&P but they nevertheless appear to incorporate the same information which S&P use when doing their analysis.

Our conclusion is that the random forest proved to be the best model for predicting credit ratings of the ones tested. The reasons being the better prediction accuracy, transparency and the fact that the model is easy to use. It has also helped shed new light into what credit raters look at when giving out corporate credit ratings.

4.1 Discussion

Machine learning models are alive and well. The move from statistical models to machine learning models have shown that it is possible to achieve higher accuracy when predicting credit ratings. As the machine learning models are continually improving, future studies could implement even more advanced models than the artificial neural network and random forest we implemented, which might improve accuracy even further. Another future study could also try to include variables which can capture the qualitative more subjective analysis the credit raters do when assessing companies. For example, variables relating to the management structure, the CEO and overall sentiment of a company could be included which might further increase accuracy and shed more light into what credit raters look at while at the same time providing a tool to increase the information investors have.

As mentioned in the introduction, some research has suggested that rating agencies sometimes diverge from their principles and consequently inflates or set ratings too low (Bolton et al. 2012, Fracassi et al. 2016). If this is true, then some of the ratings in our sample will have been misclassified by S&P even according to their own standards. An interesting topic would therefore be to follow incorrectly classified ratings from our model over time to see if there is a continued divergence or if the predictions of the random forest and the ratings given by S&P will converge.

References

Belkaoui, A. (1980), 'Industrial bond ratings: A new look', *Financial Management* **9**(3), 44–51.
URL: *http://www.jstor.org/stable/3664892*

Bennell, J. A., Crabbe, D., Thomas, S. & Ap Gwilym, O. (2006), 'Modelling sovereign credit ratings: Neural networks versus ordered probit', *Expert Systems with Applications* **30**(3), 415–425.

Bensic, M., Sarlija, N. & Zekic-Susac, M. (2005), 'Modelling small-business credit scoring by using logistic regression, neural networks and decision trees', *Intelligent Systems in Accounting, Finance and Management* **13**(3), 133–150.

Blume, M. E., Lim, F. & MacKinlay, A. C. (1998), 'The declining credit quality of us corporate debt: Myth or reality?', *The journal of finance* **53**(4), 1389–1413.

Bolton, P., Freixas, X. & Shapiro, J. (2012), 'The credit ratings game', *The Journal of Finance* **67**(1), 85–111.
URL: *http://www.jstor.org/stable/41419672*

Braspenning, P. J., Thuijsman, F. & Weijters, A. J. M. M. (1995), *Artificial neural networks: an introduction to ANN theory and practice*, Vol. 931, Springer Science & Business Media.

Breiman, L. (1996), 'Bagging predictors', *Machine learning* **24**(2), 123–140.

Breiman, L. (2001), 'Random forests', *Machine learning* **45**(1), 5–32.

Breiman, L., Friedman, J., Stone, C. J. & Olshen, R. A. (1984), *Classification and regression trees*, CRC press.

Dimitrov, V., Palia, D. & Tang, L. (2015), 'Impact of the dodd-frank act on credit ratings', *Journal of Financial Economics* **115**(3), 505–520.

Fama, E. F. & French, K. R. (1997), 'Industry costs of equity', *Journal of financial economics* **43**(2), 153–193.

Fisher, L. (1959), 'Determinants of risk premiums on corporate bonds', *Journal of Political Economy* **67**(3), 217–237.
URL: *http://www.jstor.org/stable/1827443*

Fracassi, C., Petry, S. & Tate, G. (2016), 'Does rating analyst subjectivity affect corporate debt pricing?', *Journal of Financial Economics* **120**(3), 514–538.

Frank, E. & Hall, M. (2001), A simple approach to ordinal classification, *in* 'European Conference on Machine Learning', Springer, pp. 145–156.

Haykin, S. (2004), *A comprehensive foundation*, Vol. 2.

Horrigan, J. O. (1966), 'The determination of long-term credit standing with financial ratios', *Journal of Accounting Research* **4**, 44–62.
URL: *http://www.jstor.org/stable/2490168*

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

Kaplan, R. S. & Urwitz, G. (1979), ‘Statistical models of bond ratings: A methodological inquiry’, *The Journal of Business* **52**(2), 231–261.
URL: *http://www.jstor.org/stable/2352195*

Khandani, A. E., Kim, A. J. & Lo, A. W. (2010), ‘Consumer credit-risk models via machine-learning algorithms’, *Journal of Banking & Finance* **34**(11), 2767–2787.

Kim, J. W., Weistroffer, H. R. & Redmond, R. T. (1993), ‘Expert systems for bond rating: a comparative analysis of statistical, rule-based and neural network systems’, *Expert systems* **10**(3), 167–172.

Kisgen, D. J. (2006), ‘Credit ratings and capital structure’, *The Journal of Finance* **61**(3), 1035–1072.
URL: *http://www.jstor.org/stable/3699317*

Kliger, D. & Sarig, O. (2000), ‘The information value of bond ratings’, *The Journal of Finance* **55**(6), 2879–2902.
URL: *http://www.jstor.org/stable/222405*

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

Louppe, G., Wehenkel, L., Sutera, A. & Geurts, P. (2013), Understanding variable importances in forests of randomized trees, *in* ‘Advances in neural information processing systems’, pp. 431–439.

McCulloch, W. S. & Pitts, W. (1943), ‘A logical calculus of the ideas immanent in nervous activity’, *The bulletin of mathematical biophysics* **5**(4), 115–133.

McKelvey, R. D. & Zavoina, W. (1975), ‘A statistical model for the analysis of ordinal level dependent variables’, *Journal of mathematical sociology* **4**(1), 103–120.

Morgan, J. N. & Sonquist, J. A. (1963), ‘Problems in the analysis of survey data, and a proposal’, *Journal of the American statistical association* **58**(302), 415–434.

Office of the Comptroller of the Currency, O. (1990), ‘Investment securities comptroller’s handbook (section 203)’.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011), ‘Scikit-learn: Machine learning in python’, *Journal of Machine Learning Research* **12**(Oct), 2825–2830.

Pinches, G. E. & Mingo, K. A. (1973), ‘A multivariate analysis of industrial bond ratings’, *The Journal of Finance* **28**(1), 1–18.
URL: *http://www.jstor.org/stable/2978164*

Pinches, G. E. & Mingo, K. A. (1975), 'The role of subordination and industrial bond ratings', *The Journal of Finance* **30**(1), 201–206.
URL: *http://www.jstor.org/stable/2978442*

Pogue, T. F. & Soldofsky, R. M. (1969), 'What's in a bond rating', *The Journal of Financial and Quantitative Analysis* **4**(2), 201–228.
URL: *http://www.jstor.org/stable/2329840*

Standard & Poor (2008), 'Criteria | corporates | general: 2008 corporate criteria: Rating each issue'.
URL: *http://www.standardandpoors.com/en*$_{U}S/web/guest/article/-$
$/view/type/HTML/id/1799871$

Standard & Poor (2016*a*), 'Criteria | corporates | general: Methodology and assumptions: Liquidity descriptors for global corporate issuers'.
URL:
http://www.standardandpoors.com/en$_{U}S/web/guest/article/-/view/sourceId/8956570$

Standard & Poor (2016*b*), 'General criteria: Country risk assessment methodology and assumptions'.
URL:
http://www.standardandpoors.com/en$_{U}S/web/guest/article/-/view/sourceId/8313032$

Standard & Poor (2016*c*), 'General criteria: Group rating methodology'.
URL:
http://www.standardandpoors.com/en$_{U}S/web/guest/article/-/view/sourceId/8336067$

Standard & Poor (2016*d*), 'General criteria: Methodology: Industry risk'.
URL:
http://www.standardandpoors.com/en$_{U}S/web/guest/article/-/view/sourceId/8304862$

Standard & Poor (2016*e*), 'S&p global ratings definitions'.
URL:
https://www.standardandpoors.com/en$_{U}S/web/guest/article/-/view/sourceId/504352$

Standard & Poor (2017), 'Criteria | corporates | general: Corporate methodology'.
URL: *http://www.standardandpoors.com/en*$_{U}S/web/guest/article/-$
$/view/type/HTML/id/1796819$

Timofeev, R. (2004), Classification and regression trees (CART) theory and applications, PhD thesis, Humboldt University, Berlin.

Vazza, D. & Kraemer, N. (2015), '2015 annual global corporate default study and rating transitions'.

Wan, L., Zeiler, M., Zhang, S., Cun, Y. L. & Fergus, R. (2013), Regularization of neural networks using dropconnect, in *'Proceedings of the 30th International Conference on Machine Learning (ICML-13)', pp. 1058–1066.*

West, R. R. (1970), 'An alternative approach to predicting corporate bond ratings', Journal of Accounting Research **8***(1), 118–125.*
URL: *http://www.jstor.org/stable/2674717*

Appendix

Table 9: Complete Distribution of Ratings

Rating	**Frequency**	**Percent**
AAA	22	0.63%
AA+	8	0.23%
AA	30	0.86%
AA-	47	1.35%
A+	113	3.24%
A	256	7.34%
A-	250	7.17%
BBB+	377	10.81%
BBB	542	15.54%
BBB-	374	10.73%
BB+	273	7.83%
BB	311	8.92%
BB-	311	8.92%
B+	230	6.60%
B	224	6.42%
B-	84	2.41%
CCC+	25	0.72%
CCC	4	0.11%
CCC-	2	0.06%
CC	3	0.09%
D	1	0.03%
Total	**3487**	**100%**

Table 10: Summary of Variables by Rating

					Percentile	
	Mean	**Std. Dev**	**Min**	**Max**	**25%**	**75%**
Market Value (Million USD)						
AAA/AA	160791	121509	1926	615337	59474	210859
A	35842	38342	226	213886	10519	45937
BBB	12749	17391	397	211447	3752	14602

Note: The variables in this table have not been adjusted for industry and they are also not moving averages

					Percentile	
	Mean	**Std. Dev**	**Min**	**Max**	**25%**	**75%**
BB	4358	5460	131	47319	1454	4776
B	1598	3256	33	48948	335	1738
CCC	1148	1573	33	7254	81	2217
Net Op. Income to Sales						
AAA/AA	0.2514	0.1014	0.0415	0.4517	0.1663	0.3234
A	0.2396	0.1299	-0.3583	0.7402	0.1530	0.3084
BBB	0.2044	0.1824	-3.4330	0.7150	0.1279	0.2812
BB	0.1701	0.1359	-1.5215	0.6992	0.0947	0.2187
B	0.1387	0.2855	-2.6277	0.7336	0.0655	0.2190
CCC	-0.0824	0.4408	-1.8451	0.5758	-0.0556	0.0731
Variation of EBITDA						
AAA/AA	0.2066	0.1953	0.0115	1.5457	0.0902	0.2581
A	0.2855	0.6130	-3.8813	8.0994	0.1232	0.3001
BBB	0.2334	4.94281	-170.4764	31.7782	0.1386	0.3482
BB	0.4772	6.6598	-129.7658	76.1185	0.1969	0.5060
B	3.9857	41.1732	-28.1296	829.985	0.2702	0.8925
CCC	2.2791	14.7104	-46.2649	62.4058	-1.7855	3.7492
Market Beta						
AAA/AA	0.7785	0.3111	0.3062	1.9160	0.5907	0.8182
A	0.8643	0.4063	0.1035	2.2414	0.5488	1.1061
BBB	0.9202	0.4034	0.2172	2.8682	0.5976	1.1740
BB	1.2589	0.4841	0.0948	2.9575	0.9258	1.5599
B	1.3535	0.5105	-0.4191	2.9164	1.0218	1.6330
CCC	1.3004	0.4655	-0.18221	2.1414	0.9701	1.5901
High Minus Low Beta						
AAA/AA	-0.1490	0.4454	-1.6283	0.3802	-0.2003	0.1117
A	0.2386	0.4723	-1.8814	1.0325	0.1252	0.5352
BBB	0.3593	0.4923	-2.0269	1.4458	0.2184	0.6304
BB	0.5298	0.6975	-2.4730	2.1463	0.2146	0.9646
B	0.7812	0.7768	-1.9945	2.4144	0.4519	1.2826

Note: The variables in this table have not been adjusted for industry and they are also not moving averages

					Percentile	
	Mean	**Std. Dev**	**Min**	**Max**	**25%**	**75%**
CCC	0.8322	0.5868	-0.6005	1.9298	0.2897	1.2200
Small Minus Big Beta						
AAA/AA	-0.3736	0.2710	-0.9734	0.3871	-0.5609	-0.1234
A	-0.0122	0.3031	-0.7574	0.9528	-0.1966	0.1288
BBB	0.1808	0.3713	-0.6564	1.9639	-0.0675	0.3938
BB	0.6223	0.4314	-0.5886	2.3576	0.3377	0.8863
B	0.7635	0.4968	-0.6475	2.3404	0.4286	1.0825
CCC	0.6653	0.4652	-0.0930	1.6988	0.3254	0.9421
Idiosyncratic Volatility						
AAA/AA	0.0646	0.0250	0.0413	0.1751	0.0513	0.0665
A	0.0728	0.0234	0.0420	0.2043	0.0567	0.0845
BBB	0.0834	0.0279	0.0421	0.2345	0.0646	0.0936
BB	0.1232	0.0400	0.0518	0.3258	0.0937	0.1443
B	0.1555	0.0455	0.0630	0.3687	0.1281	0.1730
CCC	0.1683	0.0510	1.010	0.3176	0.1369	0.1937
Debt to EBITDA						
AAA/AA	3.8022	2.5076	1.2391	13.1728	2.3798	4.0401
A	4.5445	2.6934	-4.5761	17.5203	2.6645	5.8036
BBB	4.4671	24.1334	-855.333	759.3333	3.3121	6.6206
BB	4.5367	43.2518	-698.4163	759.3333	3.6380	6.7284
B	18.3167	193.8448	-252.2525	4350.778	5.0080	9.7000
CCC	3.5635	48.4406	-226.2745	77.7423	-11.4809	22.3333
Interest Coverage Ratio						
AAA/AA	67.0202	81.9676	7.9344	419.4918	21.9911	80.8263
A	24.9296	40.2591	-28.3742	789.0637	9.7186	24.9056
BBB	19.6292	118.3432	-44.0905	2938.7180	6.3695	14.8993
BB	15.3254	98.2645	-24.2157	2796	4.4601	9.9204
B	4.3483	9.3248	-37.9432	163.7249	1.9424	4.4528
CCC	-1.5447	5.6561	-23.7024	2.4163	-2.1563	1.1877
Debt to Assets						

Note: The variables in this table have not been adjusted for industry and they are also not moving averages

					Percentile	
	Mean	**Std. Dev**	**Min**	**Max**	**25%**	**75%**
AAA/AA	0.5715	0.1422	0.3683	1.0037	0.4674	0.6220
A	0.5859	0.1590	0.2078	1.3744	0.4591	0.6946
BBB	0.6175	0.1577	0.2000	1.6015	0.4591	0.6946
BB	0.6385	0.2189	0.1520	2.3667	0.5116	0.7049
B	0.7817	0.2637	0.1731	2.0678	0.6089	0.7331
CCC	1.1002	0.3100	0.6586	1.8483	0.8541	1.2893
CFO to Debt						
AAA/AA	0.2710	0.1219	0.0604	0.6431	0.1879	0.3388
A	0.2292	0.1426	-0.1002	1.2134	0.1226	0.2892
BBB	0.1840	0.1197	-0.3951	1.1257	0.1082	0.2276
BB	0.1553	0.1143	-0.1948	0.7570	0.0836	0.1987
B	0.1053	0.1043	-0.1517	0.6348	0.0427	0.1428
CCC	-0.0128	0.0617	-0.2082	0.0825	-0.0279	0.0230
Dividends						
AAA/AA	0.0298	0.0129	0	0.0725	0.0203	0.0373
A	0.0191	0.0127	0	0.0962	0.0114	0.0240
BBB	0.0136	0.0184	0	0.4059	0.0056	0.0162
BB	0.0064	0.0145	0	0.2027	0	0.0078
B	0.0053	0.0247	0	0.3278	0	0.0016
CCC	0.0017	0.0038	0	0.0153	0	0
Liquidity						
AAA/AA	1.6504	0.6808	0.7735	3.7374	1.0801	2.1293
A	1.8162	1.1150	0.4027	9.5921	1.1163	2.2826
BBB	1.6823	0.9305	0.2054	7.1484	1.0513	1.0128
BB	2.0086	1.0139	0.3108	8.1064	1.3380	2.4429
B	1.9549	1.1576	0.2930	9.7166	1.1749	2.4011
CCC	1.4741	0.6752	0.6000	4.2264	1.1116	1.6894

Note: The variables in this table have not been adjusted for industry and they are also not moving averages

Table 12: Coefficients of the Ordered Probit Model

The estimation of this model used clustered standard errors since the same firms is likely represented several times in the data, which invalidates the assumption of zero conditional mean. This estimation is consequently based on the assumption that the errors of the same firm tend to covary across years. Even if the standard errors are not completely correct the model is still unbiased and will consequently not affect the models prediction accuracy.

Variables	**Coefficient**	**Std. Error**	**z-value**
Market Value	0.0001	0.0000	15.92***
Net Op. Income to Sales	-0.0679	0.0535	-1.27
Variation of EBITDA	-0.0029	0.0017	–1.74*
Market Beta	0.0845	0.1139	0.74
High Minus Low Beta	-0.6346	0.1004	-6.32***
Small Minus Big Beta	-0.3341	.0.1411	-2.37**
Idiosyncratic Volatility	-17.8086	2.0223	-8.81***
Debt to EBITDA	-0.0031	0.0021	-1.49
Interest Coverage Ratio	0.4104	0.0987	4.16***
Debt to Assets	-0.2579	0.1697	-1.52
CFO to Debt	0.3226	0.1400	2.30**
Dividends	0.1133	0.0560	2.02**
Liquidity	-0.0067	0.0493	-0.14
Cutoff Points			
μ_1	-6.0331	0.3589	
μ_2	-4.0328	0.2518	
μ_3	-2.4932	0.2072	
μ_4	-0.7222	0.1893	
μ_5	1.4845	0.3254	

$^{***}p<0.01, ^{**}p<0.05, ^{*}p<0.1$