Estimating Credit Spreads with Gradient Boosted Regression Trees and Neural Networks

René Alby

MSc Computer Science

The University of Bath

April 2023

**Estimating credit risk premiums with neural networks**

Table of Contents

[1.1 Introduction 3](#_Toc131857691)

[1.2 Structure: 4](#_Toc131857692)

[2.1 Literature Review 4](#_Toc131857693)

[2.2 Credit analysis and predicting bankruptcy 5](#_Toc131857694)

[2.3 Machine learning applications to financial analysis 6](#_Toc131857695)

[Model Selection 8](#_Toc131857696)

[Artificial Neural Network Model 9](#_Toc131857697)

[Gradient Boosted Regression Tree model 10](#_Toc131857698)

[Requirements 11](#_Toc131857699)

[Data Sources 11](#_Toc131857700)

[Credit Risk Premiums as measured by Z-Spreads 11](#_Toc131857701)

[Feature Variables used to estimate the Z-Spreads 12](#_Toc131857702)

[Summary of Feature Variables 12](#_Toc131857703)

[Detail information regarding Feature Variables 13](#_Toc131857704)

[Methodology 17](#_Toc131857705)

[Data collection and compilation 18](#_Toc131857706)

[Calculating Z-Spreads 19](#_Toc131857707)

[Compiling the Research Data 21](#_Toc131857708)

[Training the ML Algorithms 22](#_Toc131857709)

[Training the Neural Network 22](#_Toc131857710)

[Training the Gradient Boosted Regression Tree model 23](#_Toc131857711)

[Results 23](#_Toc131857712)

[Neural Network Results 23](#_Toc131857713)

[Gradient Boosted Regression Trees Results 24](#_Toc131857714)

[Interpretation of Results 24](#_Toc131857715)

[APPENDIX I – SOFTWARE DEVELOPED TO SUPPORT THIS RESEARCH 25](#_Toc131857716)

[Graphical User Interface (GUI) 26](#_Toc131857717)

[Building and Training a ML Model 27](#_Toc131857718)

[Making Predictions and Exploring Results of trained models 27](#_Toc131857719)

[Exporting / Saving Results and Models 29](#_Toc131857720)

[Inspecting model weights 29](#_Toc131857721)

[APPENDIX II – Artificial Neural Network Model 30](#_Toc131857722)

[APPENDIX III – Gradient Boosted Regression Trees 32](#_Toc131857723)

[GLOSSARY OF FINACIAL TERMS 33](#_Toc131857724)

[REFERENCES 35](#_Toc131857725)

# 1.1 Introduction

A fundamental concept in financial economics is the risk/return trade off. Investors expect to earn a higher return on a risky investment as compared to an investment with a lower level of perceived risk. While some exceptions to this rule have been observed, generally this is a widely accepted phenomenon in financial markets (Ghysels, Santa-Clara and Valkanov, 2005). For example, the interest earned on a bank account is extremely low because the depositor has access to the funds immediately and the deposits are guaranteed by government backed insurance providers. The interest earned on a corporate bond, however, would generally be significantly higher than interest on a bank account because there is a risk the corporation may find itself unable to pay back the bond, and because unlike bank deposits, bonds cannot easily be converted into cash.

The difference in return between an investment that is risk-free versus an investment that is risky is referred to as the risk premium. For corporate bonds, there are various sources of risk that contribute to the risk premium. Macroeconomic factors such as inflation and central bank interest rates, as well as microeconomic factors such as the risk of corporate default. In his seminal 1959 study in risk premiums, Lawrence Fisher identified the following factors that contribute to microeconomic risk premiums: Earnings Variability, Period of solvency, Equity/Debt Ratio, Bonds Outstanding, Volume of Trading and Ratio of standard deviation in earnings to equity (Fisher, 1959). These variables capture two sources of risk: The risk that a corporate borrow will not be able to repay their loans, known as Credit Risk, and the risk that an investor will not be able to sell or redeem a bond for cash easily if needed, known as Liquidity Risk. Fisher found that approximately 75% of the risk premium on corporate bonds could be estimated using a simple linear regression model on the variables mentioned above.

Risk premiums are fundamental to pricing or valuing investments. The role of an investment analyst is to estimate the risk premium for any given investment. And yet, there is no model that can perfectly estimate the appropriate risk premium for a particular investment. The unpredictability of economic outcomes ensures estimating risk premiums remains as much an art as a it is a science. Due to the empirical nature of investment valuation, machine learning may lend itself nicely to problem of estimating risk premiums, as there are vast amounts of financial data readily available, and numerous predictors that can be fed into a machine learning model.

Much of the literature investigating machine learning for investment valuation has focused on estimating stock returns or picking stock outperformers. As far back as 1993, Kryzanowsi et al showed how a simple neural network trained on company fundamental data and macroeconomic data can perform well when selecting stocks that will outperform (Kryzanowski, Galler and Wright, 1993). More recently, machine learning has been applied to the problem of valuation and determining equity risk premiums (i.e. the risk premiums on stocks/shares). Gu et al. found that machine learning models significantly outperformed traditional ordinary least squares regression models in estimating equity risk premiums using a large number of fundamental valuation metrics (Gu, Kelly and Xiu, 2018). They attributed this outperformance to the machine learning models ability to handle nonlinearities and interactions between the predictive variables. Comparatively little research has been conducted in the application of machine learning corporate bond realm.

This study seeks to expand upon the work of Gu, Kelly and Xiu in empirical asset valuation, except with a focus on corporate bond risk premiums rather than equity risk premiums. A deep learning model and a Gradient Boosted Regression Tree model will be trained using fundamental valuation metrics such as corporate leverage and solvency ratios, individual bond terms and conditions, and observable market data including yields on short-term and long-term treasury bonds. The goal of this study is to assess how well machine learning models can predict risk premiums on corporate bonds as measured using Z-Spreads for vanilla corporate bonds. In financial industry jargon this is referred to as Credit Spreads.

Z-Spreads on vanilla corporate bonds were selected for the purpose of this study because they are widely observable measure of the credit, liquidity, and payment timing risk of a bond. Vanilla corporate bonds refers to bonds without any embedded options. Embedded options in bonds may have potentially distorting effects on the risk profile of bonds and would require significant amount of technical adjustment to remove their effects. By sticking to vanilla bonds we avoid these pitfalls.

The application of a machine learning model to estimate credit spreads on corporate bonds could be useful for asset managers for two reasons: First, if the model is found to be accurate in estimating appropriate risk premiums, it could potentially be applied to assess relative value, and to identify overpriced or under-priced bonds in the market. Second, it can serve as a tool to assess new bond issues that do not yet have observable market prices.

Much of the research in estimating credit risk employs market focused unsystematic risk variables such as the volatility of a bond issuer’s stock price, with many derivatives of this methodology referred eponymously referred to as Merton Models (Merton, 1973). However, many bond issuers are private corporations that do not have publicly listed stocks and therefore no stock price volatility to measure. This means these models cannot be applied to a large amount of the outstanding universe of corporate bonds. This research attempts to create a model that is applicable to the entire universe of outstanding bonds, focusing solely on systematic market risk measures such as general market volatility or interest rates.

## 1.2 Structure:

The structure of this dissertation is as follows…

# 2.1 Literature Review

The topic of this research intersects the fields of financial economics and computer science, specifically machine learning. This review will investigate the literature around credit risk premiums, how they are estimated and how they are measured. Then it will delve into the topic of machine learning in financial applications and the state of the art around machine learning in credit analysis.

## 2.2 Credit analysis and predicting bankruptcy

Arguably the first published investigation into the factors that contribute to bond risk premiums began with the seminal work of Lawrence Fisher (Fisher, 1959), who postulated that bond risk premiums, as measured by the yield on a bond less the yield on the pure rate of interest, depends on whether the firm will default and the bond marketability, and that this risk premium can be estimated using a multiple linear regression (ordinary least squares) of the logarithm of four variables: 1. The variance of the firm’s earnings over 9 years. 2. the length of time the firm has been operating without any default. 3. the market value of the firm’s equity relative to the size of its debt. 4. The market value of all publicly traded bonds the firm has outstanding, which fisher uses as a proxy for marketability. Fisher found that this model accounted for 81% of the variance in the log of risk premiums.

William Beaver posited that the use of accounting ratios could be useful as predictors of firm failure (Beaver, 1966). Accounting ratios standardise the financial results of companies, allowing for performance comparisons to be made. To test this hypothesis, Beaver took a sample of failed and non-failed firms from the Moody’s Industrial Manual, matching the, and split them by their industry type using Standard Industry Classification (SIC) codes and asset size as measured form their most recent financial statement before failure. The data was aligned to ensure failed and non-failed firms were in each industry and were similar in asset size, in a pairwise fashion. Beaver assessed 30 financial ratios related to firm cashflow, net income, debt levels, liquidity and inventory turnover, and selected one from each: Cash Flow to Total Debt, Net Income to Total Assets, Total Debt to Total Assets, Working Capital to Total Assets, Current Ratio and No Credit Interval. Through comparisons of these ratios between failed and non-failed firms, Beaver identified clear patterns which could be used as predictors of firm failure.

Edward Altman formalised the use of financial ratios for credit risk assessment using Multiple Discriminant Analysis (MDA) in what is now eponymously known as the Altman Z-Score (Altman, 1968). Similar to Beaver’s research design, Altman took a split of failed and non-failed firms. Altman assessed 22 financial ratios using correlation analysis to weed out ratios that did not add additional information to the model. Altman landed on a selection of 5 ratios: Working Capital / Total Assets, Retained Earnings / Total Assets, EBIT / Total Assets, Market Value of Equity / Book Value of Total Debt, Sales / Total Assets. Altman’s MDA model correctly classified 95 percent of the total initial sample of 33 firms, however model’s predictive capability decrease significantly as the time to bankruptcy increases. Edward Deakin combined Altman and Beaver’s models using an MDA model applied to Beaver’s set of financial ratios and achieved high predictive capability over longer time to bankruptcy measures by calculating the discriminant analysis in each of the five years to bankruptcy (Deakin, 1972).

James Ohlson extended the research, both by using much larger sample sizes than his predecessors, 2163 firms, and by employing logistic regression which gives a probability of firm failure as opposed to the binary outcome of MDA. It’s worth noting here that the logistic regression function is mathematically equivalent to the sigmoid activation function in a feedforward neural network machine-learning model. Ohlson finds that the statistically significant factors that can be used to determine probability of failure are : Firm size, measured as total assets / GNP price-level index (i.e. adjusted for inflation), Total Liabilities / Total Assets, OENEG which is a dummy variable that =1 if total liabilities exceeds total assets and zero otherwise, Net income / Total Assets, Funds from operations / total liabilities, Change in net income measures as NIt0 – NIt-1 / ABS( NIt0 ) + ABS( NIt-1 )

Zmijewski added to the accountancy based probabilistic models of bankruptcy by addressing some of the potential biases in bankruptcy data sets and by introducing a model based on the probit function. Using all listed companies on the NYSE from 1972-78, totalling in the thousands. The model results in a score that classifies firms above 0.5 as bankrupt and vice versa. The accuracy of his model was measured at above 95%.

Wu, Gaunt, Grey (2010) compare various bankruptcy probability models: Altman’s MDA method, Ohlson’s Logit method, Zmijewski’s Probit model, as well as the Shumway Hazard model and A model based on Black-Scholes option pricing. The research indicates that the Altman MDA model did not perform as well as the Probit and Logit models which in turn worked well on older data sets but not as well on more modern data sets. The Hazard model had a higher predictive capability than the others and this was attributed to its use of non-systematic market data (i.e. the company share price). As noted earlier, this type of model is only useful for companies with shares listed on the stock market, and therefor is out of scope for the research conducted in this dissertation. The same is true for Black-Scholes models. Wu et al. also construct a new model consisting of key elements from the other 5 models assessed and find it has a higher predictive capability.

While it has generally been considered axiomatic that investors will demand a premium on bonds that increases with risk of bankruptcy, a model for determining whether the bond risk premium actually exists or not was only proposed in 1999 (Elton, Gruber, Agarwal, Mann, 1999). They hypothesized that the risk premium (i.e. the ‘credit spread’) could be related to different tax treatment or other non-risk related variables. The model worked by calculating spot rates across corporate bond spot curve, effectively the yields at various cashflow payments on a bond. The model then priced the bond on a risk-neutral model. By additively factoring tax and liquidity differentials, any remaining differential can only be explained by a risk premium. The result is that they found a substantial risk premium exists in corporate bonds.

Much of the literature investigating credit spreads looks broadly at economic factors that impact general (i.e. systematic) levels of spreads for specific credit ratings categories. Andrew Davies found that credit spreads on AAA and BAA Moody’s rated bonds were positively correlated to the S&P 500 index. (Davies, 2008). Davies also found credit spreads are inversely related to the the risk-free rate, however this result is considered more controversial (See Li, X.-L., Li, X., and Si, D.-K., 2020 for a summary of contradicting evidence toward this). Another finding from Davie’s econometric model was that differing inflation regimes impacted high grade and low-grade credit in differing ways, with the latter more affected by high inflation regimes. Similarly, spreads are found to be inversely related to periods on economic expansion and periods of economic recession (Kao, 2000). Employing the multifactor vasicek model, Bhar (2010) finds that the systematic variability in US credit spreads can be mostly explained by 3 factors: 1. the stock market volatility as measured by the volatility index (VIX), 2. the stock market return as measured by the S&P 500 index, and 3. the so called ‘long bond rate’ i.e. the yield to maturity on the 10 year benchmark government bond.

In summary, there are numerous firm-level and macroeconomic variables that are relevant to bond risk premiums and corporate bond valuation. However, to the best of my knowledge, there is not a single valuation model that combines all these variables together to give a single estimate for a corporate bond. This dissertation will investigate the use machine learning against a broad set of firm-level and macroeconomic variables to gauge the capability of these models in credit valuation.

## 2.3 Machine learning applications to financial analysis

A great deal of research has been conducted in the field of machine learning applications to financial problems. This is unsurprising given financial analysis and machine learning are both data-driven processes. The majority of the research focuses on stock market prediction and trading. For a review of the literature in this field see Henrique, Sobreiro, Kimura (2019). To a lesser extent, machine learning applications to financial analysis have investigated foreign exchange rate prediction, commodity price prediction, macroeconomic prediction, portfolio management and credit risk. For a summary of the literature in these fields of research see Huang, Chai, Cho (2020).

Investigations into the use of artificial neural networks for stock selection can be traced back to the early nineties when Wong et. al. used a set of 32 company financial variables whilst applying a layer of rules based ‘expert knowledge’ (i.e. an ‘expert system’) to the data (Wong, Wang, Goh, 1992). After the expert rules were applied, the data was into an artificial neural network for predictions about stock prices for 24 stocks. Their model was highly accurate in predicting the direction of stock price movement, and to a lesser extent, the actual percent performance. The name they gave to their model was FuzzNet for fuzzy network.

Swales and Yoon (1992) compared the performance of multiple discriminant analysis (MDA) to that of an artificial neural network (ANN) to two samples of 58 companies and 40 companies from Fortune 500 and BusinessWeek respectively. For each group, the study analysed the content of the president’s letter to shareholders identifying themes of economic importance. The researchers used the frequency of theme and the percentage of the letter that was dedicated to the theme to feed both their MDA and ANN to produce a binary output of ‘good’ vs ‘bad’ as relates to the company share price performance. The ANN model without any hidden layers underperformed the MDA model at about a 52% success rate, but with additional hidden layers it outperformed the MDA model, with a 77% success rate compared to 65% success rate. They concluded that non-linear relationships between predictors were contributing valuable information to the model.

Kryanowski et. al. (1993) applied an ANN to predict stock prices returns on companies on 120 companies over a five year period from 1984-89. They trained the model on the stock return 1 year in the future, using their most recent 4 years trailing financial data, calculating ratios similar to Altman and Deakin, and trained on a set of seven macroeconomic indicators which include : Industrial production, gross domestic product, the McLeod Young Weir corporate long bond index, the 90 day treasury bill rate, the government of Canada long bond index, the consumer price index and the Montreal Exchange 25 index. Due to the small sample size, they directly encoded feature data into the model, rather than attempting to get the model to generalize the features. To do this, they encoded 3-bit combinations which translated to trends in the data, up-trend, down-trend or stable. Their tests achieved a 66.4% overall accuracy in predicting whether a stock’s return would be positive or negative in the following year.

More recently, Gu et. al. (2018) compared the success of various machine learning algorithms against a large universe of stocks and predictors to measure equity risk premiums. Their sample included nearly 30,000 individual stocks over a 60 year period, employing over 900 variables based on 94 predictors for each stock, 8 predictors calculated as interactions with the 94 independent variables, as well as 74 dummy variables reflecting industry sector. Employing such a large set of independent variables is novel to the economic literature which largely focuses on linear models which cannot consume so many inputs. The research is also novel in that they investigate and compare which predictors are most influential amongst the different models employed. They find the models are mostly in agreement about the set of predictors found to be the strongest. This adds to the robustness of the findings. Finally, they demonstrate significant economic gain in the use of machine learning algorithm as compared to liner models in stock selection.

While all of the above focus on investment valuation in equity markets, relatively less research has investigated machine learning applications in the bond market.

Bianchi , Buchner, Temoni (2019) investigated boosted regression trees, random forests, extremely random randomized regression trees and shallow and deep neural networks to predict bond risk premiums focusing narrowly on the treasury market, first by using yield curves as predictors, and then by adding several time series of macroeconomic variables to the model. They found neural networks performed well in the yield-only case, and even better when employing macroeconomic variables. The following macroeconomic variables were studied: real output and income, employment and hours, real retail, manufacturing and sales data, international trade, consumer spending, housing starts, inventories and inventory sales ratios, orders and unfilled orders, compensation and labour costs, capacity utilisation, price indices, interest rates and rates spreads, stock market indicators and foreign exchange rates.

Credit ratings are an area where machine learning has enjoyed a significant amount of focus and success. This not surprising given the categorical nature of credit rating, it’s a problem well suited to machine algorithms. For a review of the literature in this field, see Golbayani, Florescu and Chatterjee (2020). Golbayani et. al. also contributed to the literature with comparative study, employing bagged decision trees, support vector machines and multilayer perceptrons, and finding the decision trees had the best performance.

Credit spreads (i.e. risk premiums on corporate bonds) are an area that has received little attention in the field of machine learning. Kim, Kim and Jung (2021) compared various machine learning methods against linear regression models to predict the 1 year and 2 year corporate bond yield spreads. Yields were calculated by taking the volume-weighted average price of trades from the American TRACE reporting system. The independent variables they chose were : credit ratings, interest rate volatility, equity index volatility, bond maturity and coupon rate, and liquidity (measured as number of days bond traded in past 12 months). The models they employed were Ridge, multivariate adaptive regression splines, neural networks, random forest and support vector machines. They found neural networks to outperform all other methods.

7Liu, Zhou and Xiong (2021) proposed a novel depth-gated recurrent neural network with a self-attention mechanism for the prediction of credit spreads. Their research investigated the Long-Short-Term meory (LSTM) neural network architecture, the Gated Recurrent Unit (GRU) Neural Network architecture, and the Just Another NETwork (JANET) neural network architecture compared against linear models. They also implemented a Self-Attention mechanism that dynamically generates weights for different connections within neural network layers. Adaptive Momentum Estimation (Adam) optimizer and Extreme Gradient boosting (XGBoost) were employed for training, and the mean squared errors was selected as a loss function. Predictors included the Risk Free interest rate, credit spread term structure, stock index return, stock index volatility, exchange rates, Oil Prices, TED Spread, swap spread and the commodity price index. Credit Spreads in this strudy were taken on aggregate by using ICE BofAML credit indices. Their findinds findings suggest the self-attention mechanism can improve results in neural network architectures.

# Model Selection

The scope of this research is to see whether we can arrive at a reasonable valuation for a given corporate bond as measured by credit spreads, by using machine learning algorithms. This research is not meant to predict future spreads based on a time series of current spreads and other data. Therefore, recurrent neural network models were deemed out scope. Instead, based on the review of the literature, two models were selected as being highly performant for financial applications.

1. Artificial Neural Network model (aka multilayer perceptron)
2. Gradient Boosted Regression Tree model

## Artificial Neural Network Model

An Artificial Neural Network model is composed of layers of if interconnected Perceptrons. Perceptrons are a form of binary classifier composed of an input variable (X), a weight (W) and a bias / intercept (t).

W1X1 + W2X2 + b = 0

In an artificial neural network, Perceptrons are layered and transformed in such a way that the output of one perceptron passes through a non-linear ‘Activation function’ and is feed to all the Perceptrons in the next layer. This is why the model is also referred to as a Multilayer Perceptron model. The unit whereby the output of a perceptron is fed through an activation function is referred to as a neuron or a node in the artificial neural network model.

Diagram

Description automatically generated

Where each node is composed of a Weight (W) multiplied by the input data (X) plus a bias value (B) and then the result of this is passed through the activation function (Af)

**Diagram, venn diagram

Description automatically generated**

At the final layer, a single perceptron outputs a value that could be a classifier of some sort, a probability, or, for a regression problem, a single estimated value.

All artificial neural networks are composed of the following:

* **Input layer**: A Layer that receives the set of estimators or X-Variables for the model (also referred to as Features).
* **Hidden layers**: Stacked layers of Nodes where the output of each node is fed to all nodes of the next layer
* **Output Layer**: A single node that outputs the result of the calculation.
* **Activation Function**: Usually either a rectified linear unit (ReLU) function, a logistic function, or a linear function. The output of each node is fed through this function which serves to transform a linear model into a non-linear model.
* **Weights and Biases**: Each note has a weight between 0-1 which allows the model to be adjusted so that a single node feeds more or less data to the following nodes. The bias is added to the weighted value prior to the activation function being applied.
* **Loss function**: A Function that measures the output or prediction of the neural network against the actual expected value. The purpose of the model is to minimize the loss for any single estimate produced by the neural network. Therefore the artificial neural network is ultimately a complex optimisation problem.
* **Backpropogation**: refers to the process of feeding the calculated loss backwards through the neural network, at which point the weights of each node are adjusted using differential equations, so that the model will produce a closer estimate to the actual value given the same inputs. This is ultimately the model’s learning process.
* **Optimization algorithms**: As previously described, the backpropagation algorithm relies on differential equations (or gradient descent) to optimize the weights of the model. There are various algorithms that can improve this learning process. One is ADAM (Adaptive Momentum Estimation) which smooths the data by applying moving averages ultimately facilitating the process of gradient descent.

The implementation of the neural network algorithm for this research was performed using TensorFlow 2.10.0.

For more detail on this model refer to Appendix II

## Gradient Boosted Regression Tree model

A Gradient Boosted Regression Tree model combines multiple decision trees to arrive at a predicted value in a regression or classification problem. Decision trees work by taking data and repeatedly splitting it into subsets based on the values of each different feature (X-Variable) until it arrives at a ‘leaf’ node, referring to a final value of data or the final prediction. Each split attempts to minimize an impurity function, usually an entropy function for classification problems or the variance for a regression problem. Once several trees are created, predictions can be made my taking the average value produced by many trees, i.e. a random forest.

The Gradient Boosted Regression Tree model enhances on the random forest model by applying a gradient boosting algorithm. The model creates new trees by fitting the trees to the negative gradient of the loss function (or objective function), therefore optimizing each tree to arrive at a closer estimate of the final prediction. The process continues until a maximum level of trees has been applied, referred to as the number of estimators. The process of creating new optimized trees this way generally outperforms other decision tree models.

This research uses the XGBoost python library version 1.7.3 to implement the gradient boosted regression tree model

More detail on this model is contained within Appendix III

# Requirements

The below section will outline the requirements to achieve the objectives of this research.

## Data Sources

All data used in this research was sourced from one of the following sources:

1. Refinitiv: A London stock exchange owned data platform through which we have access through the University of Bath library.
2. The Bank of England official data published on their website.
3. The Office for National Statistics (ONS) official website
4. Yahoo Finance

## Credit Risk Premiums as measured by Z-Spreads

The Z-Spread (zero volatility spread) is a fixed amount that when added to the to the benchmark government spot rate curve will make a bond’s discounted future cashflows equal to the observed market price of the bond. It is a single numeric value that captures the excess risk and return investors currently demand to hold the bond as compared to ‘risk-free’ government bonds of the same currency. The formula for calculating the z-spread is as follows:

0 = MP -

Where :

* MP refers to the market price of the bond on the date you are calculating the z-spread
* CFx refers to the future bond cash flow at time ‘x’
* Y refers to the corresponding treasury spot yield at time ‘x’
* T refers to the T time to X expressed as an annual fraction
* Z refers to the Z-Spread we are solving for

Therefore, Z is the number that makes the market price of the bond, minus the sum of the series of future discounted cash flows, equal to zero.

Therefore, to calculate the bond Z-spreads we require :

1. Market Prices of Bonds
2. An interpolated Treasury Spot Curve

Throughout this research we will refer to Z-Spreads by the industry standard measure: Basis Points (bps). A basis-point is one one-hundredth of a percent expressed as an integer. So, 2.2% is 220 (bps)

## Feature Variables used to estimate Z-Spreads

A selection of company financial ratios, bond terms and conditions information and general market data were selected as estimators or features (to use machine learning jargon) for credit spread prediction. The list of items is presented here with detail below. Note that the order of these items reflects the order (or column number) that they were fed into the machine learning algorithms and therefore the order required to make predictions from the machine learning algorithms:

### Summary of Feature Variables

|  |  |
| --- | --- |
| **Data Item** | **Data Source** |
| 1.      Bond Time to Maturity | Refinitiv |
| 2.      1 Year Benchmark Government Bond Yield | Bank of England |
| 3.      10 Year Government Yield | Bank of England |
| 4.      5 Year Breakeven Inflation | Bank of England |
| 5.      ONS GDP Growth Estimate | ONS |
| 6.      FTSE 100 Standard Deviation (22 day rolling) | Refinitiv |
| 7.      FTSE 100 Return (22 day rolling) | Refinitiv |
| 8.      VIX Daily Close Price | Yahoo Finance |
| 9.      Senior Subordination (Y/N) | Refinitiv |
| 10.   Unsecured (Y/N) | Refinitiv |
| 11.   Secured (Y/N) | Refinitiv |
| 12.   Quick Ratio | Refinitiv |
| 13.   Current Ratio | Refinitiv |
| 14.   Interest Coverage Ratio | Refinitiv |
| 15.   Working Capital to Total Assets | Refinitiv |
| 16.   Debt to Asset Ratio | Refinitiv |
| 17.   Debt to Capital Ratio | Refinitiv |
| 18.   Debt to Equity Ratio | Refinitiv |
| 19.   Net Profit Margin | Refinitiv |
| 20.   Operating Cashflow to Debt Ratio | Refinitiv |
| 21.   Return on Equity using EBIT | Refinitiv |
| 22.   Return on Assets using EBIT | Refinitiv |

### Detail information regarding Feature Variables [[1]](#footnote-1)

1. **Bond Time to Maturity (years)**: The time to maturity is the remaining time until a bond reaches its maturity date. The maturity date is when the loan must be repaid to investors in full. The time to maturity is an important factor in determining the risk of a bond because longer maturity periods expose investors to more risks, such as inflation risk, credit risk, and interest rate risk. Longer time to maturity is associate with higher credit risk and vice versa. So longer times to maturity are associated with higher credit risk and therefore higher credit spreads.
2. **One Year Government Benchmark Yield**: The one year government yield is the interest rate that the government currently pays to borrow money for a period of one year. It is used as a benchmark for the pricing of short-term bonds and as an indicator of the general direction of interest rates. This is because government rates are viewed as ‘risk-free’ and therefore serve as the theoretical minimum rate of return that an investor should demand from an investment. They are risk free because governments can theoretically print money to pay off debts, so their bonds should theoretically have no default risk.
3. **Ten-Year Government Benchmark Yield**: Like the one-year benchmark, the ten-year government benchmark yield is the interest rate that the government pays to borrow money for a period of 10 years. As discussed above, the longer the time to maturity of a bond, the higher perceived risk the bond poses to an investor, so the yield on the government 10-year bond is typically higher than the yield 1-year government yield. The difference between short term interest rates and longer term interest rates is referred to as the ‘Yield Curve’. The shape of the yield curve can also be important in credit analysis. A normal yield curve, where long-term yields are higher than short-term yields, can indicate a healthy economic environment and therefore stable credit risk for corporate bonds with longer maturities. However, an inverted yield curve, where short-term yields are higher than long-term yields, can indicate a potential economic recession and increased credit risk for corporate bonds with longer maturities. By having both 1 year and 10-year government benchmark yields, the hope is to capture the ML algorithm can discover the predictive feature of the yield curve in estimating credit spreads.
4. **Five Year Breakeven Inflation:** The five-year breakeven inflation is the difference between the yield on a five-year inflation-linked government bond and the standard five-year government benchmark. It reflects the market's expectations for inflation over the next five years and is used to gauge the inflation risk associated with a corporate bond. Inflation is an important factor when considering a bond investment because inflation will erode the returns on a bond. Because bonds pay a fixed rate of interest over a period of time, an investor will want that amount to be above the inflation level over that time frame. This is referred to as the ‘real’ rate of return i.e., the return achieved after subtracting inflation. Therefore, as inflation expectations rise, so should yields and therefore potentially also spreads. An interesting element here is whether the inflation expectations are fully captured within the yield of government benchmark yields. Recall the credit spread reflects only the amount above the government yield. If inflation expectations are fully reflected in government yields, then the breakeven inflation should have no impact in predicting credit spreads, because inflation should be fully factored into the model.
5. **The Office for National Statistics GDP Growth Estimate**: A measure of the expected growth rate of the United Kingdom's economy as estimated by the economists at the Office for National Statistics. A company's creditworthiness is largely determined by its ability to generate revenue and cash to service its debts. These in turn are influenced by broader economic conditions. If the economy is growing, companies may have more opportunities to increase sales and profits, which can improve their creditworthiness and reduce the credit risk of their bonds. Conversely, if the economy is contracting or experiencing slow growth, companies may struggle to generate revenue and profits, which can increase their credit risk and make it more difficult to repay debt. So the GDP growth estimate provides insight into broader economic conditions in the UK which in turn may be associated with higher or lower credit spreads.
6. **FTSE 100 22-Day Rolling Standard Deviation of Daily Returns:** The Standard deviation is a measure of the volatility of the FTSE 100 index. It is used to assess the overall level of market risk associated with a corporate bond. Volatility may be an important predictor of credit spreads because market volatility is associated with moments of financial distress for companies. Therefore one could expect a positive correlation between stock market volatility and credit spreads. Note that if the bond issuer had a publicly listed stock price, this would almost certainly be a much better predictor of credit spreads than the broad FTSE 100 stock market. However, since not all bond issuers have publicly listed stocks, this estimator would not be available in all situations. In order to make the results of this research as general purpose as possible, no stock specific data was used as a predictor. This was calculated by taking the Refinitiv daily FTSE 100 values and calculating the daily return in excel. Then the rolling 22-day standard deviation was calculated off the daily returns.
7. **FTSE 100 22-Day Rolling Return**: A measure of stock market performance. This may be correlated to credit risk since the performance of the stock market reflects the financial performance of companies, and the general economic climate. Both of these factors should affect the ability of a company to generate revenue and repay debt. Therefore, the assumption here is that positive stock market performance should be associated with lower credit spreads and vice versa.
8. **VIX Daily Close Price**: The Chicago Board Options Exchange (CBOE) Volatility Index, commonly known as the VIX, measures the expected volatility of the S&P 500 index, which is a benchmark index of the US stock market. The VIX is unique in that it extrapolates the forward-looking volatility that is priced into stock options on the underlying index. This estimator is similar to item 7, except it is more forward-looking and it is calculated on US listed companies rather than UK companies. The VIX should theoretically serve as a barometer for global economic conditions.
9. **Senior** **Subordination**: Senior debt refers to the level of subordination for a particular debt instrument. A bond with senior subordination means it has priority over other forms of debt in the event of default or bankruptcy. It is considered less risky than other forms of debt and is therefore associated with a lower credit risk. Here a flag was used with 1 indicating a bond has senior subordination, whereas a 0 was used to reflect a bond that did not.
10. **Unsecured**: Unsecured debt is a type of bond that is not backed by collateral, i.e., assets pledged to secure the bond. This means that in the event of default, the bondholders do not have a specific claim on any particular assets of the company to be repaid. Instead these bonds rely solely on the company's ability to generate sufficient cash flows to repay the bond. It is considered more risky than secured debt and should therefore be associated with a higher credit risk and higher credit spreads
11. **Secured**: Secured debt is backed by collateral. It is therefore considered less risky than unsecured debt and should be associated with a lower credit risk and lower credit spreads.
12. **Quick Ratio**: The quick ratio, also known as the acid-test ratio, is a financial ratio that measures a company's ability to meet its short-term obligations using its most liquid assets, such as cash, marketable securities, and accounts receivable. The quick ratio is calculated by dividing the company's current assets minus its inventory by its current liabilities. This ratio is used to assess a company's ability to pay off its short-term debts without having to sell its inventory or other long-term assets. The ratio is a potential estimator to consider when assessing a company's ability to meet its debt obligations. A company with a high Quick Ratio should generally be considered to have a stronger ability to meet its short-term obligations, which can reduce the credit risk of its bonds. or the purpose of this research the Quick Ratio was taken directly as calculated by Refinitiv.
13. **Current Ratio**: The current ratio is also measure of a company's ability to meet its short-term obligations. It is calculated by dividing a company's current assets by its current liabilities. A company with a high current ratio is generally considered to have a stronger ability to pay off its short-term debts, which may reduce the credit risk of its bonds and therefore credit spreads. For the purpose of this research the Current Ratio was taken directly as calculated by Refinitiv.
14. **Interest Coverage Ratio**: The interest coverage ratio is a financial ratio that measures a company's ability to pay off its debt interest expenses using its earnings before interest and taxes (EBIT). It is calculated by dividing a company's EBIT by its interest expenses. A company with a high interest coverage ratio is generally considered to have a stronger ability to pay off its interest expenses, which can reduce the credit risk of its bonds and should therefore be associated with lower credit spreads. For the purpose of this research the Interest Coverage Ratio was taken directly as calculated by Refinitiv.
15. **Working Capital to Total Assets**: The working capital to total assets ratio is a measure of a company's ability to meet its short-term obligations. It is calculated by dividing a company's working capital (current assets minus current liabilities) by its total assets. A company with a high working capital to total assets ratio is generally considered to have a stronger ability to meet its short-term debt obligations, which can reduce the credit risk of its bonds. This is because a higher working capital to total assets ratio suggests that the company has more liquid assets available to meet its short-term obligations. For the purpose of this research, the ratio was calculated manually from Refinitiv Data as:

1. **Debt to Asset Ratio**: The debt to asset ratio is a financial ratio that measures the proportion of a company's total assets that are financed by its debt. Conversely, the Debt to Asset ratio can be seen as a measure of the burden of a company’s debt relative to its asset base. It is a measure of financial leverage, which is financial industry jargon meaning the debt load a company bears or the amount by which a company is financed by debt. In general, a lower debt to asset ratio may suggest that a company is less risky because it has a lower level of debt relative to its assets, which means it may have a greater ability to repay its debts in case of financial distress. However this may be complicated by the fact that a company that uses debt to generate new business may be more profitable, that a similar company that does not. So, measures of financial leverage should theoretically have non-linear relationships to credit risk and credit spreads. For the purpose of this research, the ratio was calculated manually from Refinitiv data as:

1. **Debt to Capital Ratio**: The debt to capital ratio is also a measure of a company's leverage. It is calculated by dividing a company's total debt by its total debt and equity (assets – liabilities). A higher ratio indicates a higher degree of financial risk and is therefore associated with a higher credit risk. However, like the Debt-to-Asset ratio, it could also suggest that the company has been able to leverage its capital to generate higher returns, so may not entirely be credit negative. For the purpose of this research, the ratio was calculated manually from Refinitiv data as:

1. **Debt to Equity Ratio**: The debt to equity ratio is also a measure of a company's leverage. It is calculated by dividing a company's total debt by its total equity similar to the above two ratios. For the purpose of this research, the ratio was calculated manually from Refinitiv data as:

1. **Profit Margin**: The profit margin is a measure of a company's profitability. It is calculated by dividing a company's income by its revenue. A higher net profit margin indicates a greater ability to generate profits and is therefore associated with a lower credit risk. A company with a high net profit margin is generally considered to be more profitable and financially healthy, which can reduce the credit risk of its bonds. Therefore a higher margin should theoretically be associated with lower spreads. Because taxes are idiosyncratic and depend on factors external to the company, the profit margin here was calculated on earnings before tax. For the purpose of this research, the ratio was calculated manually from Refinitiv data as:

1. **Operating Cashflow to Debt Ratio**: A financial ratio that measures a company's ability to generate enough cash flow to cover its debt obligations. It is calculated by dividing a company's operating cash flow by its total debt. Cash flow is more important than profitability to a credit investor because creditors are only concerned with the company’s ability to pay debt. A firm that generates lots of operating cash flow but is not profitable may still be viewed as low risk to a creditor. A high operating cashflow to debt ratio suggests that a company has sufficient cash flow to cover its debt obligations, which can reduce the credit risk of its bonds and therefore should be associated with lower credit spreads. For the purpose of this research, the ratio was calculated manually from Refinitiv data as:

1. **Return on Equity using EBT**: Return on Equity (ROE) is a financial ratio that measures a company's profitability by expressing its income as a percentage of shareholders' equity. ROE using EBT (Earnings Before Taxes) is potentially more relevant to a bond investor than net income. As previously mentioned, taxes are idiosyncratic and not easily comparable between companies in an international context, so should be removed for the purpose of this analysis. Also, credit investors are concerned with evaluating a company's ability to generate earnings to service debt, and EBT indicates how much income remains available after having paid interest but prior to paying taxes. Unlike dividends to shareholders, interest payments are made from pre-tax income, so EBT is a logical measure for a company’s earnings from the perspective of a bond investor. For the purpose of this research, the ratio was calculated manually from Refinitiv data as:

1. **Return on Assets using EBIT**: Similar to the above except using the company’s assets as a base, calculated manually from Refinitiv data as:

# Methodology

This section will outline what methodology was employed to accomplish the goals of this research.

## Data collection and compilation

The data requirements for this research posed a complex problem. Bond prices and therefore Credit Spreads change daily depending on economic conditions and supply and demand. The underlying company financial data used as predictors in this research is generally reported on a quarterly or semiannual basis. A method to collect bond prices, calculate z-spreads and associate the data with the underlying bond issues financial data at a correct point-in-time posed a considerable challenge.

First, the universe of corporate bonds is massive and therefore would need to be shrunk to an appropriate scale for this research. Therefore the universe of bonds was limited to only those that are GBP Sterling denominated. These bonds however could be issued by any corporation worldwide.

As previously mentioned, only vanilla corporate bonds were in scope for this research to avoid issues around bonds with embedded options or exotic features such as being inflation-linked. Therefore, using the Refinitiv platform, a screen for bonds was created with these conditions:

1. Universe = Corporate Bonds
2. Currency = GBP
3. Excluded Coupon Types: Variable, floating, zero/coupon, resettable, range, to-be priced
4. Exclude Convertibles bonds
5. Exclude Green Bonds (
6. Exclude Putable bonds
7. Exclude Callable bonds
8. Exclude bonds with Sinking Fund
9. Exclude inflation-linked bonds
10. Include inactive / non trading bonds

The result was a list of 596 bonds that was downloaded with corresponding terms and conditions information and is referred to as the ‘Master’ list.

Each bond was (painstakingly) loaded in the Refinitiv price viewer and the entire available monthly price series for the bond was downloaded in excel format. The prices downloaded corresponded to the last day of each month. The date range for prices was between Jan 31, 2003 and Jan 31, 2023, so the research spans the economic period from 2003 to 2023 or 20 years. In total 30,036 monthly bond prices were collected.

For each bond price series downloaded the corresponding parent company’s financial statements were also manually downloaded from Refinitiv. These were downloaded for as long a time-frame as available and with a financial reporting frequency that provided the most information, so first Quarterly if available, then Semi Annually if available, then Annually if available.

Once the data was downloaded, two complex tasks remained:

1. Calculate the Z-Spreads from all the bond prices
2. Create a table data that associates the bond z-spread to the corresponding company financial data and the required bond terms and conditions from the ‘Master’ list so it can be fed into the machine learning algorithms.

To perform these tasks, I developed a set of software tools for data aggregation, manipulation and calculation, and a GUI to navigate it all. This is available [here](https://github.com/Reneige/EstimatingCreditSpreadsViaMachineLearning). The details of this tool are available in Appendix I.

## Calculating Z-Spreads

The main input required in calculating the Z-Spread is the Government Spot Curve that corresponds to the date for which you are calculating the Z-Spread, i.e., the date that corresponds to the market price of the bond for this particular data point. The monthly spot curve data was sourced from the Bank of England website with data going as far back as 1970. However, to calculate Z-Spreads you require a spot curve with points (referred to as ‘tenors’ in industry jargon) that correspond to each future cash-flow date of the bond. The bank of England data provided monthly data points on the curve for the first few years, and then semi-annual data points up to 25 years. Beginning 2016, the curves were available with semi-annual data up to the 40-year point. Bonds may pay cash flows for periods much longer than this, so this posed a potential issue. Furthermore, the semi-annual data points available do not perfectly correspond to the dates of the cash flows of the bonds. So, a method of interpolation and extrapolation was required. The chosen method was a Cubic Spline which has been found to be a reasonable method of yield curve interpolation (McCulloch, 1975). Cubic splines are piecewise polynomial function used for interpolating smooth surfaces. Here the spline was implemented using an in-built method from the Python Pandas library..

No bonds had cash flows extending beyond 40 years, however, a handful had cash flows extending beyond 25 years for which there was no yield curve point prior to 2016 and extrapolation was required. To do this, the average difference between the 25 year curve tenor and the 40 year curve tenor was taking for all the data available, i.e. all the months from 2016 onward. This average was found to be -21 basis points.

Chart, scatter chart

Description automatically generated

To extrapolate the curve up to the 40 year point, an amount equal to the 25 year tenor minus 21 basis points was inserted into the data at the 40-year tenor point, and the cubic spline was performed up to that point. The python code written to perform curve interpolation code is available [here](https://github.com/Reneige/EstimatingCreditSpreadsViaMachineLearning/blob/master/valuation_curve_builder.py) and forms part of the wider application developed for this research. The resulting curves are hereby referred to as the ‘Valuation Curves’ as they are used to value the bonds to calculate Z-Spreads.

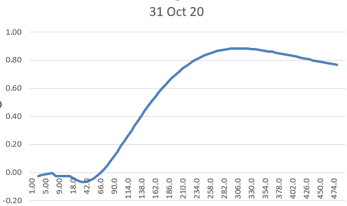
The below figures present the spot curve data interpolated but not extrapolated, interpolated and extrapolated, and an original non-interpolated, non-extrapolated curve covering the entire 40 year period. ADD FIGURE NUMBERS

A picture containing venn diagram

Description automatically generated

A picture containing venn diagram

Description automatically generated



Once the valuation curve is constructed we can continue to calculate the Z-Spread. To do this, for each bond and for each day that the bond has a price available, we project all the bonds interest payments (referred to as ‘coupons’ in industry jargon) and the final repayment of the loan. This is performed by taking the maturity date, the first coupon date, the issue date and the coupon frequency information from the Master list of bond data. The Valuation Curve that corresponds ot the price date of the bond is merged into the data, giving us the corresponding government benchmark yields on the cash flow dates. Then an iterative algorithm is applied to progressively find the Z-Spread that equates the discounted cash flow to the current market price of the bond. Once this value is found, the process continues for each bond and for each date. The resulting data is stored within the ‘Prices’ table of the database in the Python tool. The specific module that calculates this is available [here](https://github.com/Reneige/EstimatingCreditSpreadsViaMachineLearning/blob/master/ZSpreadCalc.py)

## Compiling the Research Data

With Z-Spreads now calculated, the table of research data to feed into the machine learning algorithms must be compiled. This is complex because it requires combining data from various sources and linking them by date. The methodology employed to do this was as follows. Code was written to aggregate all the bond price data, all company financial data, and all the economic data, and the master bond list which includes the bond static data, using Python Pandas (available [here](https://github.com/Reneige/EstimatingCreditSpreadsViaMachineLearning/blob/master/data_aggregator.py)). This data was injected into an SQLite database with a separate table for Prices, Financials, Master Bond List, and for each of the Economic Series : VIX, FTSE 100, ONS GDP Estimates, Yield Curves, Inflation Curves. Bond Prices and Z-Spreads were linked to the corresponding company financial data by using the Master List as a mapping table (Bond -> ISIN -> Company Name -> Company Financial Data). All of the required data was then queried from the database by joining the data sets on date. Crucially however, because financial data is reported with a lag, the financial data was first transformed so that the reporting dates pushed forward by 3-months. This means that the Price Data / Z-Spread that was ‘Joined’ between the prices table and the financial data table was actually for prices 3-months after the reporting date. 3-months seemed like a reasonable amount to lag the reporting data by given many company’s report quarterly.

The resultant data set was then adjusted as follows:

1. The data was trimmed to remove each line where there was missing feature.
2. Lines of data where the calculated z-spread was less than one year from the bond maturity date were removed. This is due to annualization of yields and spreads that below one year result in distorted values which would negatively impact the training process.
3. Some extreme values of financial ratios were removed because they could negatively impact the training of the neural network. Specifically, the rows of data where the Operating Cashflow to Debt ratio was above 50 were removed.
4. Rows with extreme and suspicious Z-Spread values were removed. Specifically, All Z-spreads above 1000 or below -10 were removed.

The final data set had 22,167 rows of training data.

## Relationship between predictors and Z-Spreads

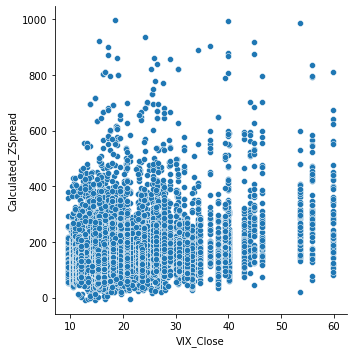
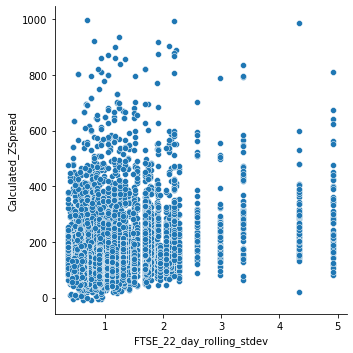
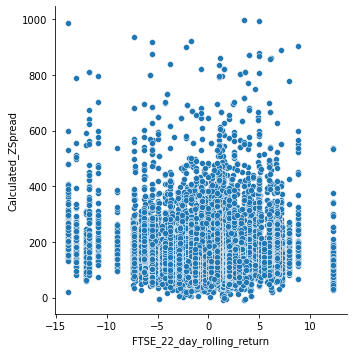
It can be useful to visualize estimators (X Variables) with respect to Labels (Y variables) to see if there are any clear relationships. Note some of the outliers have been removed from the below charts to allow the images to render properly.

### Market data variables

There does appear to be a slight negative relationship between Z-spreads and nominal yields and inflation. Low values of the VIX and FTSE Volatility are associated with lower credit spreads which seems logical.

Chart, scatter chart

Description automatically generatedChart, scatter chart

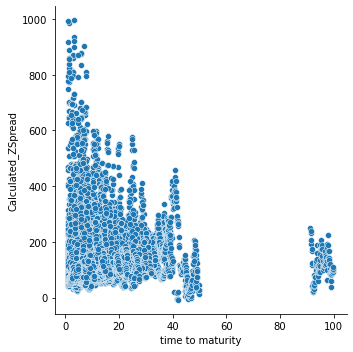
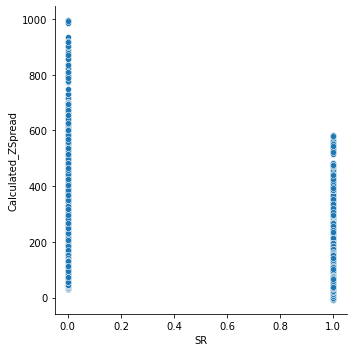
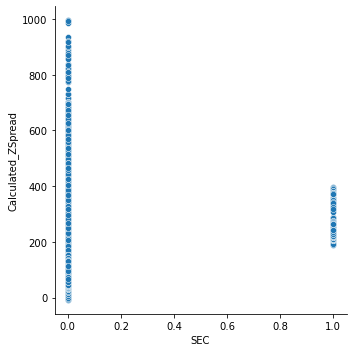
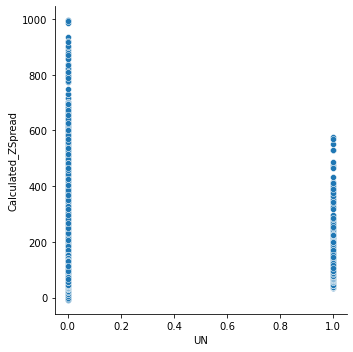
Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

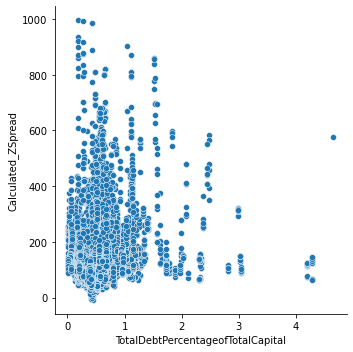
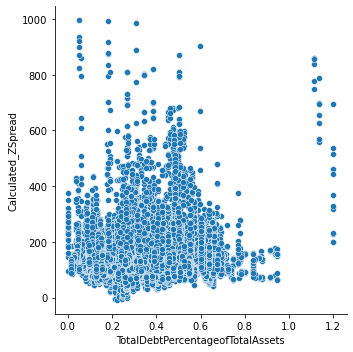
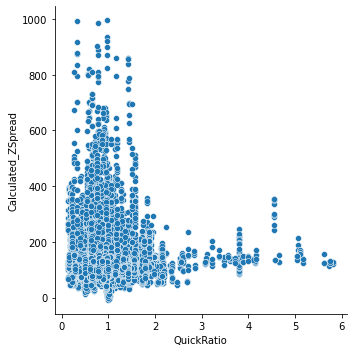
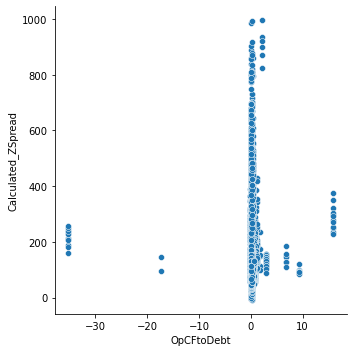
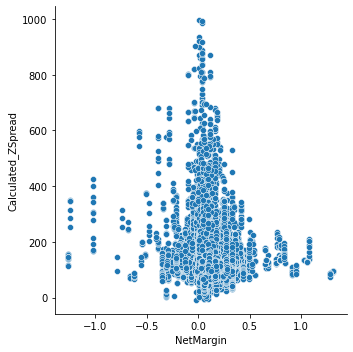
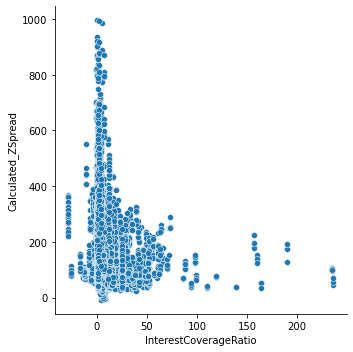
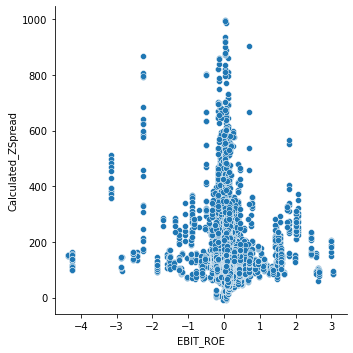
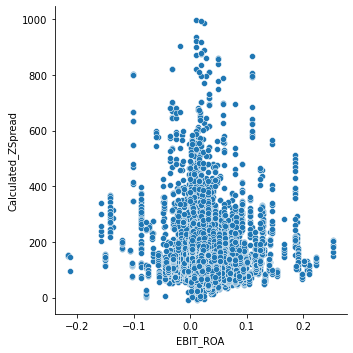
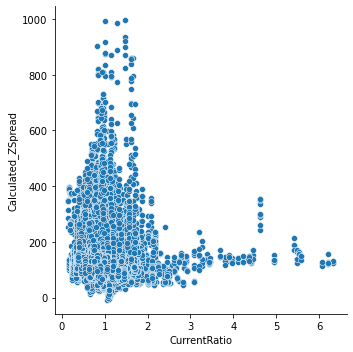
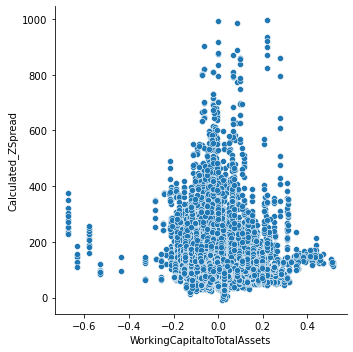
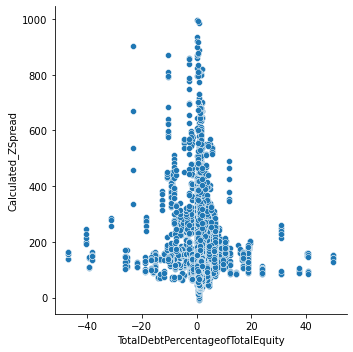
### Bond Static Data

Senior and Secured debt exhibit much lower Z-Spreads. This is logical is expected given the lower risk profile of those debt types. The unsecured category is less intuitive, exhibiting lower spreads. This category should theoretically be mutually exclusive from the secured category. This suggests the data collected from Refinitiv may not be completely reliable. The time to maturity exhibits very strong negative relationship with Z-Spreads which is very strange. However, this could be due to long term bonds being restricted to issuers deemed to be the most creditworthy.



### Financial Ratios

Many of the financial ratios exhibit clear negative relationships to credit spreads, as would be expected, given most of the ratios indicate financial distress at low levels.



## Training the ML Algorithms

For both models the data was split into a 80% train and 20% test set. For the Neural Network model, the 80% training data was further split into 25% validation data.

### Training the Neural Network

The neural network was implemented with TensorFlow V2.10.0

There is no rule book for training Neural Networks. The process is empirical and is best suited to trial and error, so it is therefore as much an art as it is a science. Given the 22 features/Estimators that were being used, I began with a network with 3 hidden layers that shrunk progressively from 22 neurons down to one. For all the hidden layers a Relu activation function was used. The output layer used a linear activation function given this is a regression problem to predict the single value of the bond’s Z-Spread. The loss function was set as the mean squared error for computational efficiency. However, the results were assessed with the mean absolute error for readability. I also employed 20% dropout nodes at each hidden layer to prevent overtraining and employed batch normalisation. I began training with around 100 Epochs. While it is standard practice to normalize the training data, I did not perform this step for the following reasons:

1. I wanted to preserve the values of the test data for model interpretation and exploration.
2. Most of the input data was in % or Ratio form which is already standardized.
3. The Boolean data for Senior/Secured/Unsecured was not suitable for standardization.
4. Standardization is meant to help prevent training sessions that initialize badly but it is clear from the slow learning rate when this occurs. So as a workaround, when the learning rate is visible too slow, the process can just be restarted until a better learning rate is observed.

Through trail and error, I quickly realised that dropout and batch normalization was not necessary. The noisiness of the training data ensured no overtraining was possible. In fact, the data could be trained for very long periods of time and the validation data would eventually flatline but never increase. I also discovered that a much deeper and wider network structure produced better results. The final Neural Network Model contained 8 hidden layers shrinking progressive with the following number of Neurons : 576, 288, 144, 72, 36, 18, 9, 3, 1.

Chart, histogram

Description automatically generated

### Training the Gradient Boosted Regression Tree model

The gradient boosted regression tree model was implemented using XGBoost v1.7.3. The exact same data was used except without a separate validation set, so just using the 80/20 train/test split. The metric used was the Mean Absolute Error. The number of estimators / trees was set to 8000 given the data does not appear to be prone to over training. However, the above 1000 the training displayed extremely diminished returns, as presented in the learning curve chart below. The learning rate was set to 0.05.

Chart

Description automatically generated

# Results

1. The testing data contained 4,435 rows.
2. The Mean value of the Z-Scores in the test set is 149bps
3. The standard deviation of the Z-Scores in the test set is 76bps.

## Neural Network Results

1. The mean absolute error on the test data was 16.97 bps
2. The standard deviation of the mean absolute error was 24.42 bps

This is a reasonable result which shows the neural network clearly has predictive capability in estimating the credit spreads of bonds based on input data.

## Gradient Boosted Regression Trees Results

1. The Mean Absolute Error on the test data was 8.38 Bps
2. The Standard Deviation of the Mean Absolute Error was 14.01 bps

This is an impressive result which shows the Gradient Boosted Regression Trees clearly has strong predictive capability in estimating the credit spreads of bonds based on input data.

## Feature importance within the Gradient Boosted Regression Tree model

XGBoost comes with inbuilt capability to enspect the feature importance. This represents the average number of splits in the trees attributed to a particular feature. However, the splits in themselves do not necessarily mean a feature is a strong predictor. The next section shows the model weight which is potentially a better tool for analysing the model.

Table

Description automatically generated

## Weights of GBRT model

Using the Eli5 Library, we can inspect model weights of the Gradient Boosted Regression Tree model. These weights here depict the average gain of a particular feature in the model. The list is presented from highest to lowest. Interestingly, the weights are quite difference from the importance presented above.

A screenshot of a computer

Description automatically generated with low confidence

## Inspecting individual predictions

Using the LIME Library for the Neural Network Model and using the Eli5 Library for the Gradient Boosted Regression Tree Model, we can analyse individual predictions of each model on the same set of data. This gives us an idea of how the algorithms are arriving at their predictions and, and which estimators are the most important for their predictions. Note that this functionality is built into the Research Tool that was developed for this software. To learn more about this tool including how to test it out on your own data, refer to Appendix I.

I will inspect 1 item where the models both predict accurately, and one item of each scenario where one model predicts accurately and the other does not.

### Both predicting accurately:

Graphical user interface, text, application

Description automatically generated

Chart, funnel chart

Description automatically generated

Table

Description automatically generated

### Gradient Boosted Tree is accurate but Neural Network is not:

Graphical user interface, text, application

Description automatically generated

Chart

Description automatically generated

Table

Description automatically generated

### Neural Network is accurate but Gradient Boosted Tree model is not

Graphical user interface, text, application

Description automatically generated

Chart, funnel chart

Description automatically generated

Table

Description automatically generated

# APPENDIX I – SOFTWARE DEVELOPED TO SUPPORT THIS RESEARCH

To facilitate this research, a software tool was developed in python. This tool is [available here on GitHub](https://github.com/Reneige/EstimatingCreditSpreadsViaMachineLearning/blob/master/data_aggregator.py). The tool provides the following functionality:

* Aggregate data in bulk and insert into database for the following data sets:
  + Bond prices
  + Company financial data
  + Bond Master List with bond static data
  + Yield curve data
  + Economic data
* Data exploration tools to ensure data was being captured and associated correctly.
* Generate a ML training data set on the fly using the data above.
* Import Training data via the clipboard.
* Train A Neural Network or Gradient Boosted Regression Tree model as specified in this dissertation.
* Browse the results of Testing data for either model.
* Save both ML models for later analysis.
* Analyse Regression Tree predictions with Eli5 library
* Analyse Neural Network predictions with LIME library
* Analyse Regression Tree model F-Scores and Weights

## Graphical User Interface (GUI)

The interface of the tool presents the list of bonds aggregated from the raw excel data that was downloaded during the data collection process. The data can that be browsed by right-clicking on a bond and selecting data items. This was created to test whether the SQL Queries and associated Python code were correctly linking bond price data / Z-Spreads to the correct underlying financial data on the correct date.

Graphical user interface, text, application

Description automatically generated

New data can quickly be added to the database by downloading files into the /data/ folder and selecting Database > Rebuild Database

Graphical user interface, text, application

Description automatically generated

When first running the tool, the valuation curve must be interpolated using the cubic spline algorithm as described in the Methodology > Calculating Z-Spreads Section of this paper. This is performed by selecting *Database > Build Valuation Curve DB* menu item as depictured above. Afterwards the bond Z-Spreads can be calculated using the *Database > Calculate Z-Spread Analytics* menu item.

## Building and Training a ML Model

To generate a full research set of data based on the features described in in the Requirements section of this paper, click on *Build ML Model > Build ML Training Data Set.* This will take several minutes to run and will perform all the SQL queries required to link the bond prices and static data to the corresponding time-shifted company financial data, and economic time series. Once complete, this data will appear in a new excel document. The data can then be cleaned and processed however the analyst would like. The final data set must be in this format : All the features should be on the left (X-Variables), and the labels (Y-Variables) should be the rightmost column of data. In this format, the user must capture the data from excel in the clipboard by pressing Ctrl-C. And then select *Build ML Model > Grab Training Data from Clipboard.* The tool will automatically adjust the input layers of the Neural Network and the Gradient Boosted Tree model to the number of X-Variables captured when doing this. You can also confirm the X-Variables are correct by selecting *Build ML Model > Display Training Data.* This will not display Labels/Y-Variables however, as they have already been sliced out of the data set at this point.

Graphical user interface, text, application

Description automatically generated

You can then select Number of Epochs or Number of Estimators depending on whether you are training a Neural Network or a Boosted Regression Tree model respectively.

Graphical user interface, application

Description automatically generated

Once set, select *Build ML Model > Train …* to train your respective model. The training status should display in the console or shell window.

## Making Predictions and Exploring Results of trained models

Once the training is complete, the tool will automatically pop up both a chart with the model learning curve, as well as a data viewer showing the results of the test set of data. The final two columns of the data viewer show the actual verses predicted Y-Variables (Labels) whereas all the other columns are the Estimators / X-Variables.

Here the user can now right-click and explore individual predictions using LIME for Neural Networks or Eli5 for Gradient Boosted Trees:

Chart

Description automatically generated with medium confidence

Here, for example, we have selected the explanation for the 120 bps Z-Spread predicted (against 127 bps Actual) with the trained Neural Network. The results window show the major elements contribution negatively to the Z-Spread (decreasing credit risk/return) is that it is a senior unsecured bond with interest coverage ratio above 7.35. The major items contributing positively to Z-Spread (increasing credit risk/return) is the low current ratio and the low operating cash flow to debt ratio.

Chart

Description automatically generated

## Exporting / Saving Results and Models

The results can also be saved to the database for future exploration, or exported to Excel for further analysis

Graphical user interface, text, application, table

Description automatically generated

If you must also save the trained ML models if you would like to use them in the future to make predictions or to explore predictions.

Graphical user interface, text, application

Description automatically generated

Note, Neural Networks are saved in a folder and contain various files and subfolders. When saving a neural network, you must create a folder first to save all this data. Gradient Boosted Trees, on the other hand, are saved as a single JSON file.

## Inspecting model weights

Once you have a trained model or have loaded a saved model, you can inspect the Weights and F-Scores of to get an idea as to what features the model is finding most important when making predictions:

Text

Description automatically generated

# APPENDIX II – Artificial Neural Network Model [[2]](#footnote-2)

Let X be the input data, with shape (n, d), where n is the number of samples and d is the number of input features.

Let Y be the output data, with shape (n, k), where k is the number of output classes or continuous values.

Let W be a set of weights, with shape (m, d), where m is the number of nodes in the current layer and d is the number of nodes in the previous layer.

Let b be a set of biases, with shape (m,), where m is the number of nodes in the current layer.

Let f be the activation function applied to the weighted sum of the inputs to a node.

Let L be the loss function used to measure the difference between the predicted output and the true output.

Let η be the learning rate, which determines the step size of the weight and bias updates during training.

**Step 1 : Forward propagation:**

The output of the first hidden layer can be calculated as follows:

*Z*1​=*XW*1​+*b*1​

where Z1 has shape (n, m1), where m1 is the number of nodes in the first hidden layer. We then apply the activation function f to Z1 to get the output A1:

*A*1​*=f(Z*1​*)*

We repeat this process for each subsequent hidden layer and the output layer. The output of the final layer is the predicted output Y\_hat:

*Z L*​=*AL*−1​*WL*​+*bL*

*Yhat​=f(ZL​)*

**Step 2 : Calculate Loss:**

We use the loss function L to calculate the difference between the predicted output Yhat and the true output Y:

*L(Y,Yhat)*

A typical loss function is the Mean Squared Error:

*MSE = (1/N) \* sum[(yi - fi)^2]*

**Step 3: Backpropagation:**

We now use backpropagation to update the weights and biases in the MLP. We start by calculating the gradient of the loss function with respect to the weights and biases in the output layer:

We repeat this process for each subsequent hidden layer and the output layer. The output of the final layer is the predicted output Yhat

Text, letter

Description automatically generated

Where  Is the ith row of 

We can then use these gradients to update the weights and biases in the output layer:

Text

Description automatically generated

Apply the chain rule calculating the gradient of the loss function with respect to the weights and biases for each previous layer:

Text, letter

Description automatically generated

# APPENDIX III – Gradient Boosted Regression Trees [[3]](#footnote-3)

The process of calculating gradient boosted regression trees is as follows.

Begin by defining an objective function (loss function). The objective function measures the accuracy of the model in fitting the training data. For regression problems, the most common objective function is the MSE (mean squared error) defined as:

MSE = (1/N) \* sum[(yi - fi)^2]

N = the number of training instances

yi = the i-th instance of the actual value (i.e. label)

fi = the i-th instance of predicted value (i.e. label)

Steps the model:

1. Start by Initializing the model with a constant value. Usually the mean of the target variable.
2. Fit a regression tree to the residuals of the current model. The residuals are the differences between the actual labels and the predictions of the current model. The goal of the tree is to learn the pattern in the residuals that the current model has not captured.
3. Add the new tree to the current model by updating the predicted values. Each new tree provides a correction to the previous predictions, which improves the overall performance of the model.
4. Repeat steps 2 and 3 until the objective function reaches a minimum or a maximum number of trees have been added.

Residual are defined as follows:

r = y - f

y = the vector of actual labels and

f = the vector of predicted labels from the current XBGR model.

We want to fit a regression tree to the residuals, which means finding the optimal splits that minimize the MSE. Define the MSE of a leaf node as:

MSE\_node = (1/N\_node) \* sum[(yi - fi)^2]

where N\_node is the number of instances in the leaf node

We can then define the total MSE of the tree as the sum of the MSEs of all the leaf nodes:

MSE\_tree = sum[MSE\_node]

The goal of fitting the tree is to find the optimal splits that minimize the MSE\_tree. The searches for the optimal splits by trying all possible splits for each feature and selecting the one that results in the greatest reduction in MSE\_tree. The reduction in MSE\_tree from a split can be computed as:

MSE\_reduction = MSE\_parent - (MSE\_left + MSE\_right)

where MSE\_parent is the MSE of the parent node before the split, MSE\_left is the MSE of the left child node after the split, and MSE\_right is the MSE of the right child node after the split.

Once we have found the optimal splits for all the features, we can build the tree by recursively partitioning the data based on the splits until we reach the maximum depth or the minimum number of instances per leaf node.

Finally, to add the new tree to the current model, we multiply its predictions by a learning rate (gamma) and add it to the previous predictions:

f\_new = f + gamma \* f\_tree

f\_new = the new vector of predicted labels

f\_tree = the vector of predicted labels from the new tree

gamma = the learning rate.

The learning rate controls the contribution of each new tree to the final prediction and can be tuned to balance the bias-variance trade-off. A smaller learning rate will result in a more conservative model with lower variance but higher bias, while a larger learning rate will result in a more aggressive model with higher variance but lower bias.

# GLOSSARY OF FINACIAL TERMS

**Asset**: Here used as a synonym for an investment

**Asset Valuation**: Specifying the percent return an investment should provide an investor in order to compensate for the perceived risks of holding that investment.

**Credit**: Here used as a synonym for corporate bond

**Credit Spread**: The difference in yield between a risky bond and a risk free government bond of equal maturity and currency

**EBIT**: Earnings before interest and taxes. A measure of business profitability before costs of servicing debt and taxes.

**EBIDTA**: Earnings before interest, depreciation, taxes, and amortization. A measure of business profitability before costs of servicing debt and taxes, and prior to some of the potential distortions of modern accounting methods.

**Enterprise Value**: The sum of the market value of a company’s debt and equity. Viewed as the total value of a business, or the cost of purchasing an entire business and therefore taking ownership of the company’s debt obligations.

**FFO**: Funds from operations. This represents the cash a business generates from conducting its primary business operations

**Fundamental Valuation**: Analysing financial health of a company to arrive at a valuation for the investments they issue.

**Liquidity**: The ease with which an investor can sell (liquidate) an investment.

**Leverage**: The amount of debt a company holds

**Maturity**: The expected date a bond will be repaid to investors

**Options**: A financial instrument or a feature of a bond that allows the holder to buy or sell an investment at a given price in the future.

**Option Adjusted Spread (OAS)**: A single discount rate added to each point of the risk-free term structure, which when used to discount a risky bond’s cash flows, makes the discounted cash flows equal to the observed market price of the bond after adjusting the price for any embedded options. The OAS can be viewed as a measure of a bond’s credit spread, however it also includes other risks such as liquidity risks. In the absence of embedded options, the OAS is equivalent to the Z-Spread.

**Risk Premium**: The percent amount an investment will compensate an investor as compared to a risk free investment.

**Solvency**: The risk a company will not be able to repay its debts

**Systematic Risk**: Risk variables that that impact the entire universe of an investment and therefore are inherent to the entire market

**Unsystematic Risk**: Risk that is specific to an individual investment or firm.

**Vanilla Bonds:** Bonds that contain no special price features suck as sinking funds or embedded options

**Yield**: Short for Yield-to-Maturity which is the expected return a bond should provide an investor if held to the maturity date of the bond.

# REFERENCES

Babbel, D. and Zenios, S., 1992. Pitfalls in the Analysis of Option-Adjusted Spreads. Financial Analysts Journal, 48(4), pp.65-69.

Bhar, R., and Handzic, N., 2010. *A multifactor model of credit spreads*. Asia-Pacific Financial Markets, 18(1), pp.105–127.

Bianchi, D., Büchner, M. and Tamoni, A., 2020. Bond Risk Premiums with Machine Learning. *The Review of Financial Studies*, 34(2), pp.1046-1089.

Beaver, W.H., 1966. Financial ratios as predictors of failure*. Journal of Accounting Research*, 4, p.71.

Buczynski, W., Cuzzolin, F. and Sahakian, B., 2021. A review of machine learning experiments in equity investment decision-making: why most published research findings do not live up to their promise in real life. *International Journal of Data Science and Analytics*, 11(3), pp.221-242.

Davies, A., 2008. Credit spread determinants: An 85 year perspective*. Journal of Financial Markets*, 11(2), pp.180–197.

Deakin, E.B., 1972. A discriminant analysis of predictors of business failure. *Journal of Accounting Research*, 10(1), p.167.

Elton, E.J., Gruber, M.J., Agrawal, D., and Mann, C., 2001. Explaining the rate spread on corporate bonds. *The Journal of Finance*, 56(1), pp.247–277.

Fama, E., 1970. Efficient Capital Markets: A Review of Theory and Empirical Work. The Journal of Finance, 25(2), p.383.

Fisher, L., 1959. Determinants of Risk Premiums on Corporate Bonds. Journal of Political Economy, vol. 67, no. 3, University of Chicago Press, 1959, pp. 217–237.

Ghysels, E., Santa-Clara, P. and Valkanov, R., 2005. There is a risk-return trade-off after all. *Journal of Financial Economics*, 76(3), pp.509-548.

Golbayani, P., Florescu, I., and Chatterjee, R., 2020. A comparative study of forecasting corporate credit ratings using neural networks, support vector machines, and decision trees. *The North American Journal of Economics and Finance*, 54, p.101251.

Gu, S., Kelly, B. and Xiu, D., 2018. Empirical Asset Pricing Via Machine Learning.  Chicago Booth Research Paper No. 18-04, 31st Australasian Finance and Banking Conference 2018, Yale ICF Working Paper No. 2018-09, Available at SSRN: https://ssrn.com/abstract=3159577

Harvey, C., 2021. Be Skeptical of Asset Management Research. *SSRN Electronic Journal*,. Available at Available at SSRN: <https://ssrn.com/abstract=3906277>

Henrique, B.M., Sobreiro, V.A., and Kimura, H., 2019. Literature review: Machine learning techniques applied to financial market prediction. Expert Systems with Applications, 124, pp.226–251.

Huang, J., Chai, J., and Cho, S., 2020. Deep Learning in Finance and banking: A literature review and classification. Frontiers of Business Research in China, 14(1).

Kryzanowski, L., Galler, M. and Wright, D., 1993. Using Artificial Neural Networks to Pick Stocks. *Financial Analysts Journal*, 49(4), pp.21-27.

Kao, D.-L., 2000. Estimating and pricing credit risk: An overview. *Financial Analysts Journal*, 56(4), pp.50–66.

Li, X.-L., Li, X., and Si, D.-K., 2020. Asymmetric determinants of corporate bond credit spreads in China: Evidence from a nonlinear ARDL model. *The North American Journal of Economics and Finance*, 52, p.101109.

Liu, X., Zhou, R., Qi, D., and Xiong, Y., 2022. A novel methodology for credit spread prediction: Depth-gated recurrent neural network with self-attention mechanism. *Mathematical Problems in Engineering*, 2022, pp.1–12.

McCulloch, J.H., 1975. The tax-adjusted yield curve. The Journal of Finance, 30(3), pp.811–830.

Merton, R.C, 1973. An Intertemporal Capital Asset Pricing Model. *Econometrica: Journal of the Econometric Society*, 41, 867-887

OpenAI, ChatGPT, 24 Jan. 2023, https://chat.openai.com/

Swales, G.S., and Yoon, Y., 1992. Applying artificial neural networks to investment analysis. *Financial Analysts Journal*, 48(5), pp.78–80.

Wong, F.S., Wang, P.Z., Goh, T.H., and Quek, B.K., 1992. Fuzzy Neural Systems for stock selection. *Financial Analysts Journal*, 48(1), pp.47–52.

Zmijewski, M.E., 1984. Methodological issues related to the estimation of financial distress prediction models. Journal of Accounting Research, 22, p.59.

1. OpenAI ChatGPT was used to assist writing with this section. [↑](#footnote-ref-1)
2. OpenAI ChatGPT was used to assist writing with this section. [↑](#footnote-ref-2)
3. OpenAI ChatGPT was used to assist writing with this section. [↑](#footnote-ref-3)