Exploratory Data Analysis Report - Group 7

Kelly Tong, Rakeen Rouf, Lisa Wang, Javier Cervantes

Data Overview

The research aims to estimate a company's credit risk using a dataset sourced on *September 22, 2023*. The dataset is a subset of the holdings within an ETF, exclusively representing companies listed in the S&P 500. It comprises **2,341** rows, with each row corresponding to a specific bond issued by an S&P 500 company. Across the dataset, there are **34** variables, which can be grouped into four distinct categories, each originating from different data sources:

1. **Bond information from iShare (13/34):** Information related to the bonds, including the issuer's name, industry sector, price, duration, yield to maturity, issuer's stock ticker, and market capitalization. Sourced from the USIG Ishares Credit Bond ETF.
2. **Company fundamentals from Yahoo Finance (11/34):** Company fundamentals, including various financial ratios (e.g., revenue, debt). Sourced from Yahoo Finance using the yfinance package.
3. **Credit ratings from Bloomberg (4/34):** Credit ratings from Fitch, Moody's, and S&P, and a composite credit rating. Sourced from the Bloomberg Terminal.
4. **Social sentiment indicators from Finhubb API (6/34):** Social sentiment indicators including the number of positive and negative mentions on Reddit last year. Sourced from the Finhubb API.

Utilizing this dataset, we aim to delve into two key research questions:

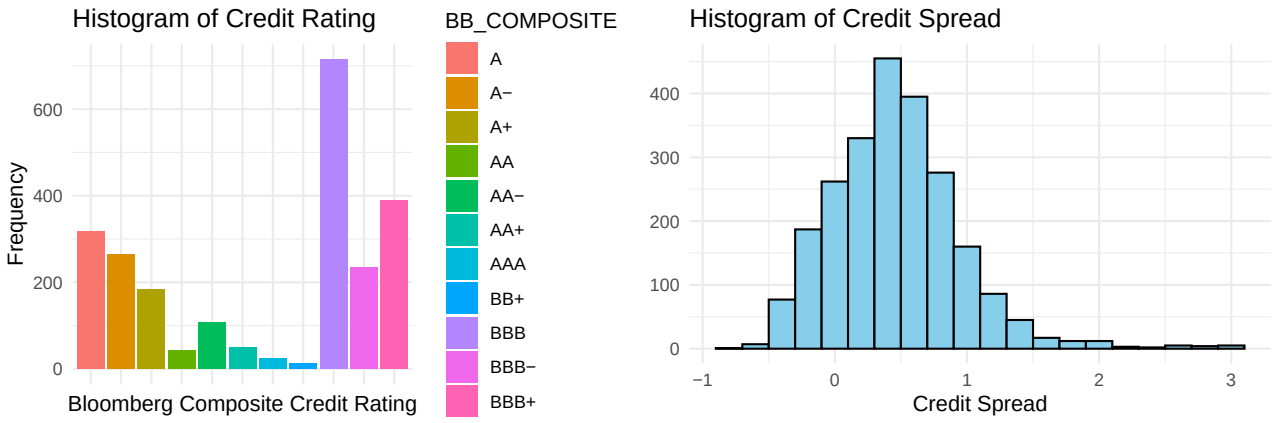
1. Which factors contribute to a bond's credit rating? (Outcome Variable: average_credit_rating)
2. Can we predict a certain bond's credit spread based on various metrics like the company's fundamentals and the market's sentiment related to that company? (Outcome Variable: credit_spread)

Outcome Variables

The `Credit Rating` is an ordinal categorical variable, and represents the composite credit rating issued to each company by Bloomberg. When we look at the distribution of Credit Ratings in Figure 1.1, we notice that it appears more scattered and less structured compared to the distribution of Credit Spreads in Figure 1.2. To make sense of this, we are considering combining certain categories of Credit Ratings. This could potentially help to reduce the variability and make the data clearer. We will discuss this further in the potential problem section. It's worth noting that the 'BBB' category has a much larger number of observations in the data set compared to other categories. This makes it a significant portion (30.5%) of the data.

The `Credit Spread` is a continuous outcome variable and represents the difference in yield or interest rate between a particular bond and a benchmark bond with similar characteristics but considered to be risk-free. For example, if a corporate bond yields 5% and a comparable risk-free government bond yields 3%, then the credit spread of the corporate bond is 2%. The distribution of credit risk in the dataset with respect to the Figure 1.2 below, appears to follow a normal distribution with a high positive skew. The mean Standard Deviation and Median of Credit Spread are as follows: 0.47, 0.49, 0.44.

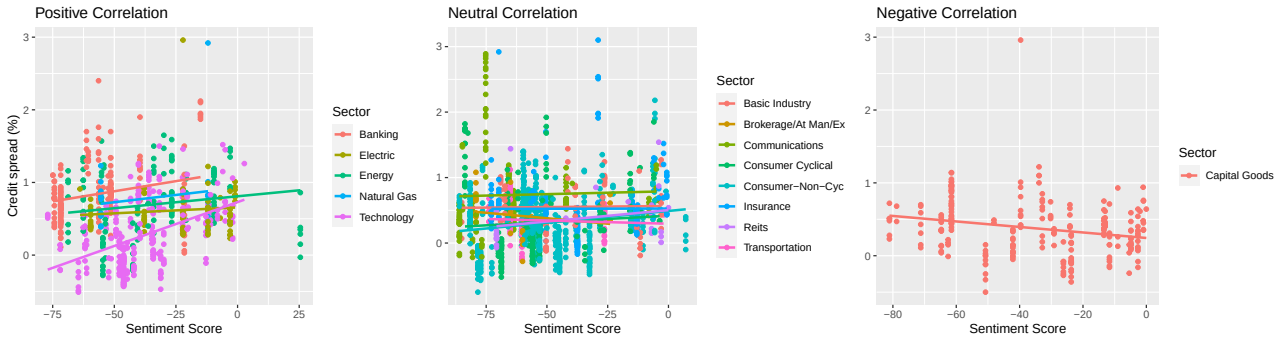
Figure 1.1, 1.2: Outcome Variables



Primary Relationships of Interest

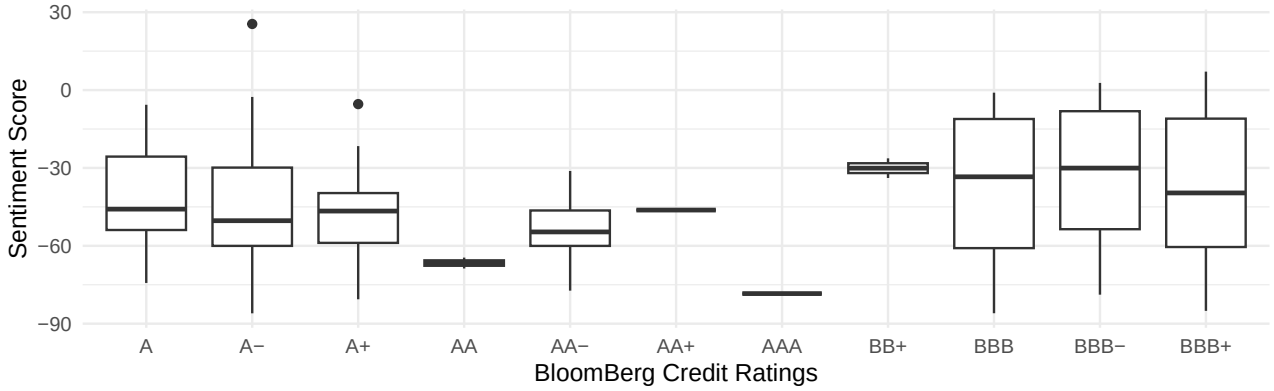
`Sentiment Score:` The sentiment score exhibits a notable influence on the behavior of credit spreads, varying significantly across different sectors. For instance, in Figure 2.1, 2.2, and 2.3, distinct sectors demonstrate varying correlations between Sentiment Score and Credit Spread. Some sectors display a positive correlation, while others show a negative or neutral association. Further examination revealed that sectors characterized by a more elastic demand, such as Technology and Banking, tend to exhibit a more pronounced positive correlation. Conversely, sectors with a more inelastic demand, like Basic Industry and Cyclical Consumer Industry, demonstrate a largely neutral response to the Sentiment Score.

Figure 2.1, 2.2 , 2.3: Sentiment Score Vs Credit Spread



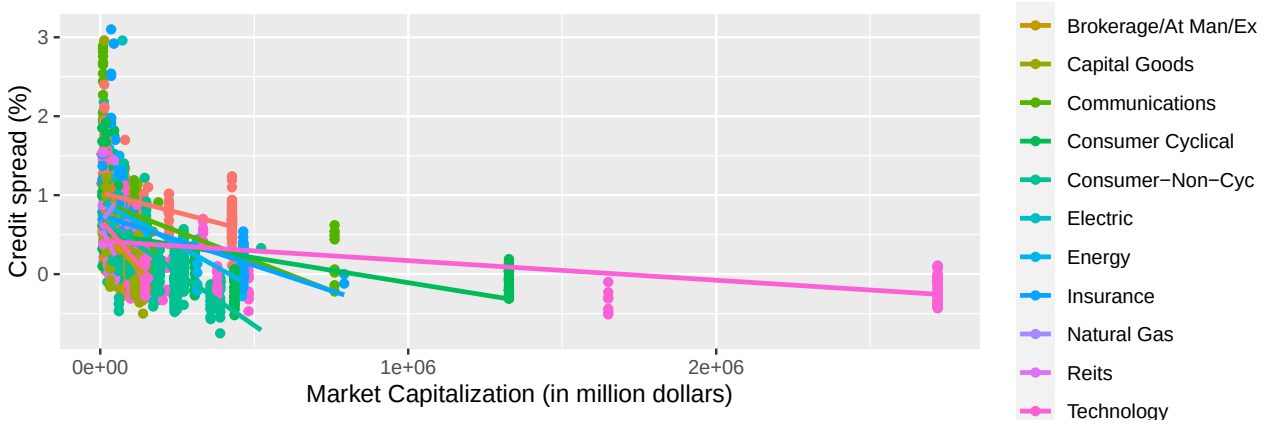
The data in Figure 2.4 suggests that companies with a credit rating of B (including BBB, BB+, etc.) exhibit a considerably broader interquartile range compared to those with an A rating. Additionally, all statistical measures in the box plot for A-rated companies are consistently lower when compared to their B-rated counterparts.

Figure 2.4: Sentiment Score Vs Credit Ratings



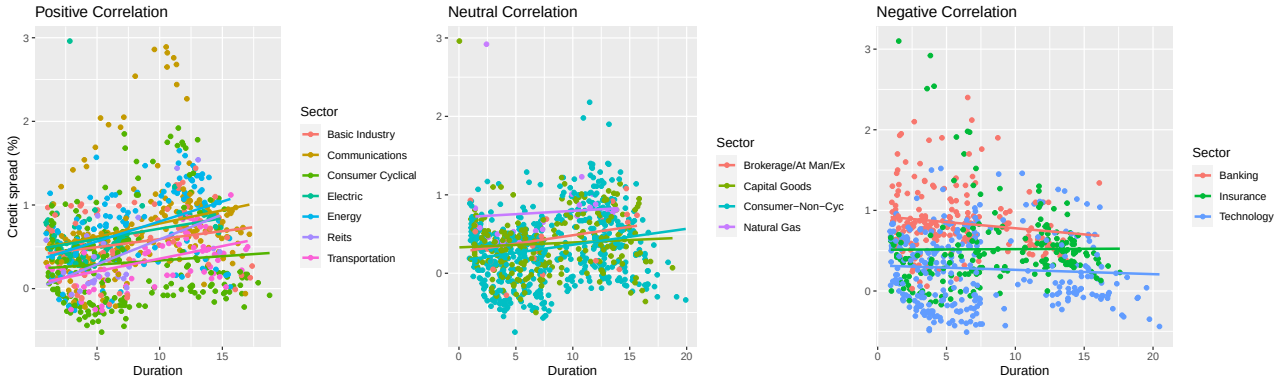
`Market Capitalization and Credit Spread:` Both visualizations seem to demonstrate an inverse relationship between Market Capitalization and Credit Spread. The significance needs to be investigated more. In almost all sectors, there is a relatively wide range of credit spread being represented in the lower rank Market Capitalization. This might be caused by the discrepancy between values of Market Capitalization, with some outliers that own very high Market Capitalization.

Figure 2.5: Correlation between Market Capitalization and Credit Spread



`Duration:` Duration in investment quantifies the degree to which a bond's or portfolio's price is affected by shifts in interest rates. The visualizations illustrate that the connection between duration and credit spread is not uniform and fluctuates across sectors. In sectors such as basic industry and energy, the relationship tends to be positive, given the stable cash flows and lower risk profile, which lead to the issuance of longer-duration bonds. Conversely, sectors like technology, characterized by substantial growth potential but less stability, and banking, sensitive to interest rate shifts, exhibit a negative correlation between duration and credit spread.

Figure 2.6, 2.7 , 2.8: Duration Vs Credit Spread



Other Characteristics

The dataset encompasses a range of additional fundamental technical indicators, including metrics like Rate of Amortization, EBITDA Margin, Operating Profit Margin, and others. A correlation matrix is created to demonstrate the relationship between all numeric fundamentals variables.

Figure 3.1: Correlation Matrix for All Numeric Variables

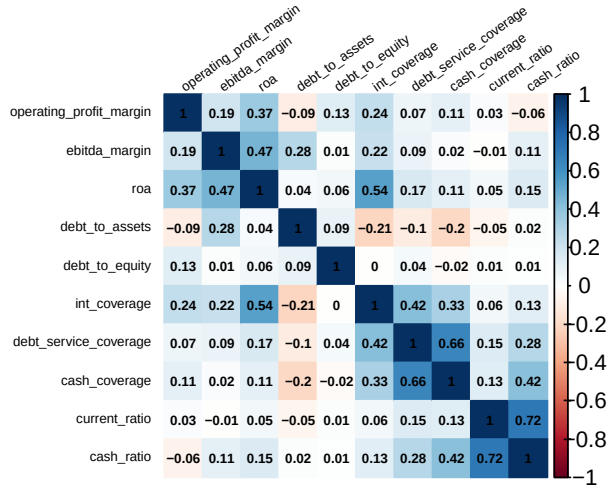


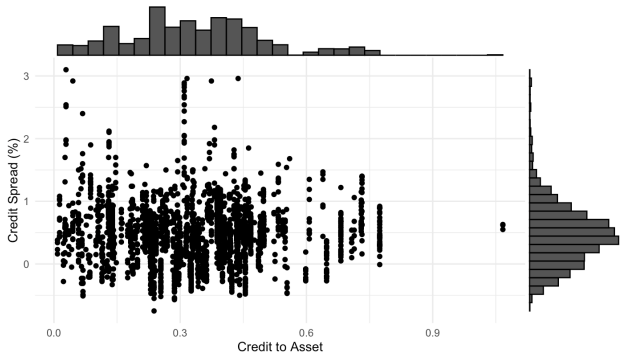
Table 1: Sample Predictor Variables Descriptive Statistics

Predictor Variables	**N = 2,341**
Interest Coverage (ratio)	Mean: 12.73, StDev: 35.19, Median: 8.04
Unknown	327
Debt Service Coverage (ratio)	Mean: 1.50, StDev: 3.05, Median: 0.60
Unknown	387
Cash Coverage (ratio)	Mean: 6.00, StDev: 11.10, Median: 1.98
Unknown	387
Current Ratio (ratio)	Mean: 1.36, StDev: 2.33, Median: 1.13
Unknown	350
Cash Ratio (ratio)	Mean: 0.37, StDev: 0.51, Median: 0.22
Unknown	350

Several columns in the company fundamental data gathered from Yahoo Finance contain missing values (Table 1). These missing values are typically derived from the company's balance sheet. We will conduct further investigation to understand the reasons for the data gaps and undertake efforts to manually calculate and fill in the missing values.

`Debt and Credit Spread:` While credit spread follows the bell-shape of a normal distribution, the density distribution of debt to assets presents multiple peaks and downturns. This suggests that the relationship between debt to assets and credit spread might depend on the industry sector. Upon further investigation, each sector was observed to have different relationships between the two variables, with some being positive and the other being negative. This could be a potential challenge if we want to incorporate debt to assets into our prediction model. It could impact accuracy and efficiency negatively, as well as creating potential over-fitting. It might be helpful to include an interaction term between debt to assets and sector.

Figure 3.2: Correlation between Credit Spread and Credit to Asset



From the density distribution of debt to equity (Figure 3.3), we can see that most data centered around "debt to equity equals 0." This suggests that there are a few outliers with debt smaller than -100 or larger than 150. Removing these outliers might present a better view of the relationship between the two variables. This again raises potential challenge on how to clean the variables and deciding what should be considered as an outlier.

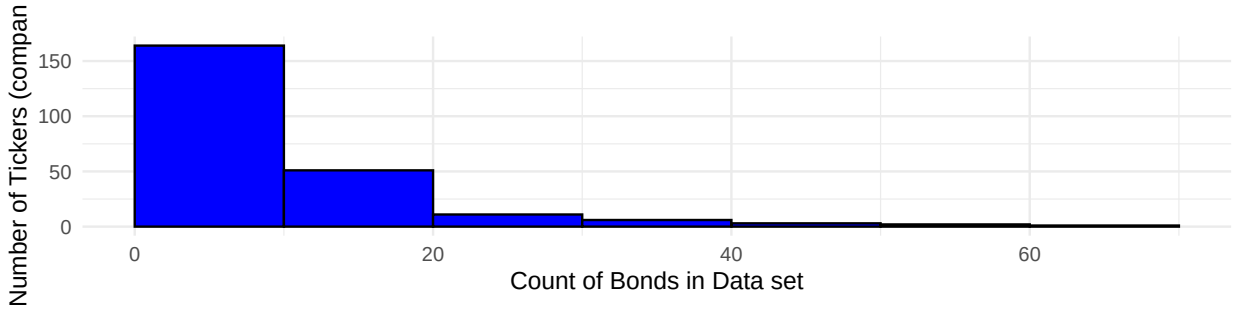
Figure 3.3: Correlation between Credit Spread and Credit to Equity

`Technical Ratios`: Interest Coverage, Debt Service Coverage, and Cash Coverage provide valuable insights into a company's creditworthiness and financial stability. A stronger financial position, as indicated by higher ratios, tends to correlate with a lower credit risk, potentially leading to narrower credit spreads and more favorable credit ratings. Cash ratio indicates a company's ability to pay off short-term liabilities with its cash and equivalents, while the current ratio measures its overall liquidity. Higher liquidity and solvency, reflected in these ratios, typically indicate lower credit risk, which may lead to narrower credit spreads and higher credit rating scores.

Several variables have been excluded from the statistical model. General bond information, such as the bond's name and unique identifier, will not be included. Additionally, certain variables that exhibit high correlations with each other, like the cash ratio and current ratio, or the count of positive/negative mentions and aggregate mention scores, will have only one representative variable retained in the model. Furthermore, while we gathered credit rating scores from three different institutions (Fitch, Moody's, and S&P), we will utilize the Bloomberg Composite Credit Rating, which is the equally weighted blend of all the scores, as the outcome variable for our analysis.

Potential Challenges

Figure 3.1: Histogram of Number of Bonds per Ticker



One challenge our team may encounter relates to the varying frequencies at which each unique company's ticker appears in our raw data. As depicted in the figure above, most tickers have between 1 and 10 bonds in our dataset. However, there are a few tickers at the higher end, with between 60 and 70 bonds. This imbalance leads to a data set where models trained on it might exhibit a bias towards companies with more bonds (observations). To address this concern, the team could implement strategies for handling the imbalanced dataset. Some potential approaches could include, Weighted Sampling, Stratified Sampling, Re sampling Techniques, Etc.

Another potential challenge arises from the low counts in the categorical variables, specifically in Bloomberg Composite Rating (the outcome variable) and Industry Sector. In both cases, there are classes with limited representation. For instance, Banking and Brokerage/Asset Managers/Exchanges collectively account for only ~10% of the total data. One argument can be made that these industries share enough similarities to be collapsed into a single category. By consolidating certain categories, the team aims to mitigate the issue of low counts.

In addition, extreme data points might influence the distribution of the values and model prediction. This could be a potential challenge when we are cleaning our variables since we need to make decisions on whether the data points are considered as outliers and whether they should be kept, excluded or normalized.

Once the adjustment are made, the team is confident that the resulting dataset will be of sufficient size to effectively address the research questions.