Statistical Analysis Plan
Stats Group 7 : Rankeen Rouf, Lisa Wang, Javier Cervantes, Kelly Tong

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):** Specific information related to each bond like ISIN, Duration, YTM, etc. 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, S&P, and Bloomberg's own 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.

Modeling

Utilizing this dataset, we aim to delve into two key research questions:
1. Which factors contribute to a bond's credit rating? (Outcome Variable: BB_COMPOSITE)
2. Can we predict a certain bond's credit spread based on a company's fundamentals and the market's sentiment related to that company? (Outcome Variable: credit_spread)

Research Question #1

Model Choice: Ordinal Regression

Inference or prediction: This is an inference problem. For inference problems, it is of utmost importance that the predictors (independent variables) selected make intuitive sense. The selection of predictors directly impacts the interpretability and generalizability of the model.

Variable Selection:

- Bond specific predictors: duration, credit spread
- Company specific predictors: market capitalization, ticker, industry sector
- Sentiment Score: this variable encompasses all the sentiment metrics we've obtained
- Fundamentals: a broad category that encompasses diverse predictors related to fundamental metrics for each company

Interaction term:

1. Sentiment Score and Sector: Through our EDA report, we have noticed that the marginal impact of sentiment on our outcome variable varies within distinct sectors

2. Debt to assets ratio and Sector: Given the nature of the different businesses we expect to observe that the marginal effect of leverage on our outcome variable will not remain constant across sectors.

Research Question #2
Model Choice: Linear Regression

Inference or prediction: The second research question is primarily geared toward prediction rather than inference. So the variable selection is more flexible, allowing for the consideration of a broader range of potential predictors. We prioritize predictive accuracy and may utilize feature engineering and machine learning techniques. Model assessment will involve metrics like MAE and RMSE to evaluate how well the model forecasts credit spreads.

Variable Selection:

- Bond specific predictors: duration, credit rating
- Company specific predictors: market capitalization, ticker, industry sector
- Sentiment Score: this variable encompasses all the sentiment metrics we've obtained
- Fundamentals: a broad category that encompasses diverse predictors related to fundamental metrics for each company

Interaction term:

1. Sentiment Score and Sector: Through our EDA report, we have noticed that the marginal impact of sentiment on our outcome variable varies within distinct sectors
2. Debt to assets ratio and Sector: Given the nature of the different businesses we expect to observe that the marginal effect of leverage on our outcome variable will not remain constant across sectors.

Addressing Potential Challenge

1. **Managing Influential Points**: Cook's distance will identify influential points for the dataset which could possibly lead to biased and inaccurate model fitting. We will first conduct a priori model without removing the identified influential points. Then, we will create another model which removes all the necessary influential points. Comparison on matrices such as model accuracy will be used to determine which model performs better.
2. **Resampling Imbalance Dataset:** There is an uneven distribution of unique company tickers in our raw data. Most tickers are associated with 1 to 10 bonds, but a few have between 60 and 70 bonds, creating an imbalance in the dataset. To address this issue, we can implement strategies for handling imbalanced datasets, such as weighted sampling, stratified sampling, or various resampling techniques.
3. **Filling Missing Values:** Several columns of the company fundamental data obtained from Yahoo Finance contained missing values. To tackle this challenge, we plan to conduct a thorough investigation to uncover the underlying reasons for these data gaps. Subsequently, we will undertake manual calculations and data imputation techniques to fill in the missing values, ensuring the integrity of the dataset.