Big Messy Data Project Proposal

David Fitzpatrick (df347), Jane Xiong (wx77), Jiaming Liu (jl4286)

October 3, 2021

1 Motivation

The operating performance of a publicly listed company has a significant effect on its stock price. However, it has been a challenge to achieve accuracy on identifying the key factors of operating variables that sway stock price. Structuring a model and performing an analysis of informative data, we will be able to find potential methods to improve the practice of interpreting fundamentals. Therefore, we believe it is valuable to explore how the operation of a public company affect its stock returns and utilize the exploration to gain a fresh insight into the impact operating factors have on financial market.

2 What we are doing differently

Many groups have access to the data we are using on kaggle and try to use it to predict stock returns. We have found a flaw in their notebooks due to confusion around the date variable in the dataset. the date variable is the calendar year which all of the revenue etc. were earned. We have found other notebooks using this data from year t to calculate returns for year t+1. The issue with this method is different companies filing their information at different times so for many stocks there is information bleed. Information bleed is where these notebooks are trading on information they would not actually know until months later. We plan to correct this by only selecting companies who report around the same time period, finance companies, and only predicting returns for the months after we know the information would be public. We plan to investigate the correlations within the dataset for any particularly interesting connections between economic variables and finance companies performance. We then plan to test a variety of linear, Support Vector Machine, and Decision tree based models for the ability to predict stock performance between information releases.

3 Dataset

The dataset we are going to use is the NYSE SP500 company fundamentals Kaggle database that contains the stock prices, the fundamental financial metrics, and the descriptions of each security from 2010 to 2016. But due to the issue of financial report filing time mismatch mentioned above, we reduce the scope of the stocks to finance companies only. Since only finance companies report around the same time period. In addition, we believe macroeconomic condition might also have impact on the company performance. To test this, we will use Federal Reserve data, for example, yield curve, to test this correlation.