DSCI 470 Project 3 Proposal

Ryan Rosiak and Grant Dawson

December 2021

1 Introduction

Our project will be to implement and assess the accuracy of a machine learning model for forecasting using the S&P 500 Index. Additionally, we will use the python library "yfinance" to interface with Yahoo Finance and acquire the data for our model. The S&P 500 Index holds stock data for the top 500 companies in the United States. We will use the opening day price, closing day price, high and low price of the day, and other attributes to predict future price fluctuations.

2 Library Analysis

The "yfinance" library will be our main interface for acquiring S&P 500 Index data from Yahoo Finance. Below, is list of attributes that will be used that are provided by the library.

- Ticker - ex: KO
  - Unique key for each stock given by the stock market.
- Period - ex: 1yd
  - Time frame for when the data is presented from the past year.
    Or provide:
  - Start - Date
  - End - Date
- Interval - ex: 1hr
  - Time increments of how often the data is captured.
- Prepost - boolean
  - Whether or not pre and/or post market/after hours are included.

The library provides more attributes but they are not very relevant. All of the data provided to us allows us to make candlestick graphs for each stock interval. The collection methods provided streamline the process of getting the data and creating great visualizations for the data acquired with minimal steps.

3 Attribute Analysis

The most important thing to look at is what the S&P 500 Index dataset provides us. Below, is a list of those attributes.

- Open
  - The price the time interval started at.
- Close

  - The price the time interval ended at.

- High
  - The highest price reported in the time interval.
- Low
  - The lowest price reported in the time interval.
- Adj Close
  - The closing price reported in the time interval factoring in anything that might affect the stock price after the interval ends.
- Volume
  - Number of shares sold/bought in the time interval

After careful analysis of the attributes of the dataset, it is important to derive questions to further our research. Some questions driving further analysis of the dataset and modeling are provided below.

- Are certain stocks/tickers easier to predict than others?
- Does the window in which the data comes from matter? Does the past year or the past 5 years affect the current year more?
- Does the interval have any affect? Does more data mean better results? Does it help to have more time between each data entry?
- Does factoring in pre/post market/after hours help?

4 Conclusion

Overall, predicting future price fluctuations in the S&P 500 Index can prove to be very useful in future trading practices and stock management. The "yfinance" library and Yahoo Finance dataset are tools that we will use to find out if modeling such a problem using machine learning can actually work and be beneficial for the others.