Stock Market Analysis and Prediction

Ian MacDougall   
*Computer Science Undergraduate, Wentworth Institute of Technology*

***Abstract***

This report explores stock market data [1] using a data science oriented approach. It will address 5 questions that will be answered using the stock market data. It asks questions like what features of a stock correlate to a high price or how close stock prices of the same sector are related to each other. This report uses various python libraries to produce our response like pandas, numpy, scikit-learn, & lightgbm. With our data, it was clear that there are serval instances where there were clear patterns for an increase in prices. However, with our data, it seems more difficult to find instances where the pattern clearly reflected a decrease in price. This could be due to outside factors that are not directly related to the stock or even the company itself. Regardless it was interesting to work on and view this data, in the future I hope to work with more complex data like this and make better predictions of what make important features that lead to increased or decreased price.

***Keywords—Python, Clustering, Pandas***

**I.Introduction**

The Stock Market is a complex, wild system that many wish they could tame and understand. It is influenced by a wide range of factors and outside behaviors that all play into each stock’s price. I want to make analyzations of both predictions about how the market has & will act, and how the stock market has been affected regarding important events & purchases from important parties. A reason one might want to learn to predict the stock market based on its features & current price is so that they can take advantage of these stocks. Because if you had the ability to predict the stock market, you would have the ability to turn your money into a grand investment only through stock market purchasing and selling. Some research by Deepak Kuma, Pradeepta Kumar Sarangi, and Rajit Verma [3] suggest that with correct feature engineering and an integrated process stock market predictions could be considered more accurate.

**II.Datasets**

**A. Source of dataset**

This report uses one main data set which consists of the stock market data for stocks in the Standard & Poor's 500[1]. This source provides the stock price of stocks at SP’s 500 along with information about the companies. This information was found on Kaggle and was produced by a trusted & highly rated user, currently rated 19 out of 15,272 users who provide data to consistently to Kaggle.

The other data set used in this report consists of financial purchase made by United States Senators that are disclosed to the public [2]. This data set is gathered in a uniform manner on the site Senate Stock Watcher. The user gathers this data from the United State Senate Financial Disclosures (eFD), which is directly distributed by the United States Senate.

**B.Character of the datasets**

The dataset on the ‘sp500\_stocks’ is a 200-megabyte CSV file with around 1.88 million lines of data. It consists of the Date, Stock Symbol, Adjusted Close, Close, High, Low, Open, and volume. When using the data, the stock would be broken apart by the desired stock symbol. Due to the quality of this dataset, it did not have to be cleaned and consist of small amounts of null data. This data was combined with the ‘sp500\_companies’ dataset to create new features that may affect the price of the stock.

The dataset ‘sp500\_companies’ consists of items like Sector, Market cap, Revenue Growth, and location. This data did not need to be cleaned and often only needed to be adjusted for the needed stock.

The ’transactions’ dataset consists of the Senator’s name, party, transaction type, and asset type. This data has to be cleaned to only show strictly stock asset type transactions. It will then be further changed to represent the desired stock to analyze.

**III.Methodology**

Through the report various data science methodologies were used. We will go over them in the following subsections.

**A. Method A: graphing**

This was the simplest method used while working on the project. This was used when reviewing important dates that correlated to stock price data or when looking into Senator’s stock sales & purchases. I chose it as I needed to assess what happened around these important dates and how the price could have changed after the specified date. Thus, it was easiest to visual display what dates were important along with the stock price data.

**B.Method B: Light Gradient Boosting**

Light Gradient Boosting was used in this report to help discover what the most important factors were in determining a high stock close price. It uses decision trees to discover what is most important in determining a stock price. It slowly re-iterates over itself so that the best solutions are slowly approached.

**C.Method C : Clustering**

Clustering was used to evaluate patterns in the stock market and find similar stocks in the same sector. This is used to evaluate how closely related two stocks are by their calculated distance. Once visualized it is very easy to see how these clusters are separated

**IV.Results**

Since we asked questions about stocks multiple times, we will have a result for each question

**A.Result A**

**Q1)** What are the most relevant factors in producing a high return?

We discovered that the most relevant features in producing a high return are Volume, City, Current Price, Industry, and Market Cap.

**B.Results B**

**Q2)** What common patterns do we see in the market resulting in an increase of decrease in price?

With the clustered stock pattern data, we are able to see that the cluster is able to predict clusters that lead to an increase in price. However, there is no clear indication of what types of clusters predict a decrease in price.

**C.Results C**

**Q3)** How close is a stock’s price related to the other stocks in its subclass / sector?

The results from our clusters result in a clear visualization of how the stocks are separated. It is clear that there are various stocks that have a correlated stock price, in fact a large number of stocks have stock prices that correspond to each other.

**D.Results D**

**Q4)** How have tech giants been affected by the introduction of AI?

While there was not a significant amount of important AI events, when comparing the important events to the stock price of Microsoft it was clear to see that Microsoft had several shape increases after a few of these events.

**E.Results E**

**Q5)** Do people who affect our laws have stock options that correlate with price changes?

When looking at the Senators’ purchase history for Microsoft after 2019, most

**V.Discussion**

The report would have benefited from access to more feature data related to each stock on a specific date. The complexity of the stock market is difficult to capture only using a stocks price & corresponding factors, especially when creating accurate predictions. I would like to research more in-depth methods of working with my data and hopefully finding more feature rich data to use.

**VI.Conclusion**

This report was to analyze stock market data and answer several questions using that data. This report also gave me the opportunity and knowledge to begin working on analyzing stock market data and other data pieces like it. It also made it clear that the data is available that will allow us to analyze and possible predict patterns in data and analyze how events and purchase lead to affect the market. Thus, this report encourages those interested in the stock market to work with the data. As it is complex but can still yield very interesting results.

**Acknowledgment**

We acknowledge that there is not a variety of stock market data choices. However, the stock market data is the same for everyone on each day. Because of this we believe that it was appropriate that only one clean, usable file was needed.

There is more access to high level data, however this data is not free and requires a subscription to access. While it could lead to better results, for this report the free, usable data was more than enough to answer the questions we needed.

The number of significant events should have been larger. However, without clear guidance of what may be considered a date of interest or other recognized dates of events it is

difficult to choose what date can and should be considered important.

**References**

1. A. Maranhão, “S&P 500 Stocks (daily updated)” . Kaggle.com. https://www.kaggle.com/datasets/andrewmvd/sp-500-stocks. (accessed Nov. 8, 2024)
2. “Senate Stock Watcher”, senatestockwatcher.com. https://senatestockwatcher.com/api. (accessed Nov. 26, 2024)
3. D. Kuma, P. Kumar Sarangi, and R. Verma, “A systematic review of stock market prediction using machine learning and statistical techniques”, *Science Direct*, vol. 49, no. 8, pp. 3187-3191. Jan. 2022. Accessed: Nov. 29, 2024.