D214 Data Analytics Graduate Capstone

Task 3: Executive Summary

Jonathon “Jon” Fryman

Data Analytics 01/01/2022

Student ID 00974544

Program Mentor: Lea Yoakem

C: 419-206-6989

Table of Contents

[Part I: Overview 2](#_Toc130220891)

[Part II: Problem Statement and Hypothesis 2](#_Toc130220892)

[Part III: Data Analysis Process 3](#_Toc130220893)

[Part IV: Findings 4](#_Toc130220894)

[Part V: Limitations 4](#_Toc130220895)

[Part VI: Proposed Actions 4](#_Toc130220896)

[Part VII: Expected Benefits 5](#_Toc130220897)

[Part VIII: Sources 5](#_Toc130220898)

# Part I: Overview

According to the New York Stock Exchanges available data, an average of $18.9 billion is traded on the stock market each day. With such a tremendous amount of money being invested, perspective investors stand to lose a substantial amount of funding if poor investments are made due to lack of research into the available data for a given stock.

# Part II: Problem Statement and Hypothesis

The stock market can often be riddled with uncertainty and filled with great fluctuation. To facilitate educated investing, it is essential to ascertain trends and seasonality that may exist within a given stocks performance prior to making a large investment.

Utilizing available historical stock information for the companies being analyzed, the objective is to develop a model capable of forecasting future stock performance with an accuracy that can provide a greater level of confidence for investors prior to committing large amounts of funding.

**Research Question**: To what extent can a company’s future daily per-share closing stock dollar value be accurately predicted?

**Hypothesis**: The hypothesis of the current data analysis is that the Mean Absolute Percentage Error(MAPE) score associated with each predicted stock closing price will be below 20%. A MAPE score below 20% indicates the difference between the dollar value for the predicted and actual stock closing price is smaller than 20% which is a generally considered good MAPE score.

**Null hypothesis**: The null hypothesis is that the MAPE score is 20% or greater. This would indicate less than 80% similarity on the predicted dollar value when compared to the actual stock closing price.  
  
**Alternate Hypothesis**: The alternate hypothesis is the predictions associated MAPE scores are less than 20%. A MAPE score in this range would indicate a greater than 80% similarity on the predicted dollar value in comparison to the actual stock closing price.

# Part III: Data Analysis Process

**Data Collection**: The data utilized was collected from <https://www.kaggle.com/datasets/evangower/big-tech-stock-prices>. This dataset contains historical market performance data for 14 American tech companies dating from January 2010 through December 2022. The data contained information related to date, opening price, closing price, daily low price, daily high price, adjusted closing price, and volume. The data set was downloaded as 14 unique .csv files with each file pertaining to a specific company’s data. A dataset consisting of the full range of dates beginning in January 2010 contained 3,271 historical trading days’ worth of data.

**Data Preparation**: Jupyter Notebooks was the primary tool used for this project. Prior to the .csv files being imported, importing the required Python libraries was required. For this analysis, the pandas, matplotlib, seaborn, sklearn, statsmodel, and pmdarima packages were required.

The 14 .csv files were then imported using pandas .read\_csv() function. A review was performed to ensure there were no missing or null values contained within the original 14 datasets. It was determined the files associated with Meta and Tesla did not contain a consistent number of historical datapoints in comparison to the remaining 12 files, and as a result these two companies had their data excluded from the analysis.

The remaining 12 files were then used to populate a new empty DataFrame that used the trading date as an index formatted into datetime format. The new DataFrame contained a column for each of the 12 companies. The columns were labeled using the company’s stock trading abbreviation and populated with the Closing price associated with the date found in the specific row’s index.

Additional Exploratory Data Analysis was performed in the form of generating histograms for each company’s historical stock prices. A line chart for each company was generated depicting their stock price over the available historical period. An Augmented Dickey-Fuller test was performed for each company’s data, which reflected no seasonality was found for any of the datasets. Additionally, the seasonal\_decompose() function was also used to review for seasonality.

A train/test split was performed to facilitate training the prediction model and having separate data available to perform validation of its performance.

Finally, the auto\_arima function was used to determine the appropriate p,q, and d values to be set for each company prior to fitting the training data to the respective model and generating predictions.

# Part IV: Findings

**Time Series Analysis**: It was determined that the closing price for all companies included for analysis could be forecast with an accuracy greater than 80% when compared to the actual price contained within the test validation data set. For many of the companies, the prediction accuracy was well above 90% when compared to the validation test data.

When the predictions were validated against the actual historical closing price associated with the predicted days:

* 99.3% of the 432 predictions had an accuracy >= 80% of the actual historical close price.
* 77% had >= 90% accuracy of the actual historical close price.
* 47% had >= 95% accuracy of the actual historical close price.
* The three predictions with < 80% accuracy had 79.2-79.9% accuracy in comparison to the actual historical close price.
* All three with accuracy below 80% were associated with Nvidia within a four-day span.

Additionally, the MAPE score associated with each predicted day for all 12 companies was below the 20% threshold required to accept the null hypothesis.

# Part V: Limitations

## The limitations of the performed analysis were that the predicted estimates were consistently lower than the stocks actual market price for a given day. In this case, it is beneficial to error on the side of caution and have the prediction be below an actual price as opposed to above. However, it would warrant additional fine tuning of the model to determine if performance can be improved to better align with the actual price more closely on a given day.

# Part VI: Proposed Actions

**Recommendation**: The recommended course of action based upon the completed analysis is that utilizing the developed model can assist with investment decisions if a company is seeking insight into an expected company's future market performance.

However, as stated in the previous section, it would be prudent to attempt fine-tuning of the model to improve predictions consistently being below the actual price. If a company’s market value significantly increases for a given period, the model may lag in reflecting accurate predictions at the new market value. This could cause an investor to underestimate any gains from their investment.

**Approach for future research**: Regarding future data study, there are two key points that would be essential to maintaining the accuracy of the current prediction model and improving it for additional use.

It would be prudent to facilitate the model being able to regularly update the collected historical data. This would allow predictions to continue being made for trading days in the future based upon historical data that is not yet currently available.

It would create additional opportunity to implement Improvement to the current prediction model that allows the collection of data for additional companies. This would facilitate predictions being generated for any company trading on the stock market that has historical data available for collection.

# Part VII: Expected Benefits

Utilizing the developed prediction models, an investor can confidently review the expected Closing price for a given trading day and have a general idea how well they can expect that stock to perform. This would facilitate an investor having the ability to roughly calculate an expected loss/gain for a given trading day and have an expectation how the funding they’ve invested will fluctuate. This would provide the ability to make prudent investment decisions in terms of the best time to buy or sell a specific stock to minimize loss or maximize gains depending on the situation.

# Part VIII: Sources

1. Gower, E. (2023, January 30). Big Tech Stock prices. Kaggle. Retrieved February 23, 2023, fromhttps://www.kaggle.com/datasets/evangower/big-tech-stock-prices
2. Trading & Data. NYSE. (n.d.). Retrieved March 20, 2023, from https://www.nyse.com/trading-data
3. Fryman, J. (2023, March 19). D214 Data Analytics Graduate Capstone Task 2: Data Analytics Report and Executive Summary