Microsoft Technical Stock Analysis

by Amit Bharadwa

October 2020

Mentor: Wayne Ang

Contents

1 Problem Statement

Predicting future stock prices can be incredibly profitable and not to mention, stock prices reflects a companies value. These two motivations alone make the stock market an interesting topic of research. There are many factors to consider when predicting a stock price and due to economic and social factors, stock prices are incredibly volatile and difficult to predict.

The technology company, Microsoft, has become one the leading pioneers in software, and many companies rely on their products to perform at the highest level on a day to day basis. Even-though there are many factors that influence stock price, can other companies influence Microsoft's future stock price and is it possible to predict Microsoft's adjusted close price using a machine learning model?

2 Introduction

Stock market prediction is the act of trying to determine the future value of a companies stock price which is traded on the financial market. Building a model that is able to predict the future value of a share price can be very difficult as there are multiple variable to consider. These variables include news of new products or services the company are offering, securing a new contract, takeovers or mergers, inflation, economical and political shocks, investor sentiment, current performance of the industry and many more. The mentioned variables are referred to as fundamental factors that impact a companies share price and that factors will not be used in this analysis. This report will focus on the technical aspect. This includes researching historical trends, charts and identifying statistical indicators. The underlying principle is that the market price reflects all information that could impact the market. For this reason, there is no need to look at the economic, fundamental or new developments since they're already priced in security[1].

The daily data collected for any company traded on the market follows a time series format. This means the price at each timestep is dependent on the previous timestep. This is crucial for predicting future values and will be discussed further in section 4.2. This report will analyse Apple (AAPL), Sony (SNE), Google (GOOGL) and HP (HPQ) stock prices to see if there is a statistical influence with Microsoft's (MSFT) stock price. By training three machine learning models with a training data set,a prediction will be made on Microsoft's future stock price, and a justification on which model proves to make the best prediction based on suitable model metrics.

3 Stock Data

This section will go into detail about the features of the data set, as well as visualising the data to identify any correlation between different technology companies. Using statistical methods, a conclusion will be reached on which companies have an influence on Microsoft's stock price and whether they should be included in training the machine learning models.

3.1 Understanding the Data

The data used in this project was obtained from the Yahoo fiance API and ranges from 01/01/2010 to 23/03/2020. As expected the data does not contain any missing values. The features of the data set include Date, High, Low, Open, Close, Volume and Adjusted Close. The difference between close price and the adjusted close price is that the close price is the stock price at the end of a trading day where the adjusted close price takes into account factors including the dividends, stock splits and new stock offerings [2]. For the rest of this project the adjusted close price will used for analyses and prediction.

3.2 Data Wrangling

As previously mentioned, the data used is a time series data set. In order to perform exploratory data analysis and reach a prediction for the future adjusted close price, the data will have to be in a particular format. This will require the Date feature to be the index of the data set and as a Pandas date-time object. By converting the adjusted close price as a 'float64' data type, the data set is ready to be explored for correlations and to perform statistical tests.

3.3 Exploratory Data Analysis

Understanding the relationships between the different technology companies is crucial to determine if there is a influence between different companies. By confirming through statistical tests, additional features can be added to the data set and in turn a robust predictive model.

Firstly by understanding the data and visualising any initial trends between the technology companies, the data can be explored further. Figure 1 below shows the general trend for each company by taking the rolling mean every 50 days. Clearly, Google is dominating the current market, however all companies seem to be following a positive trend.

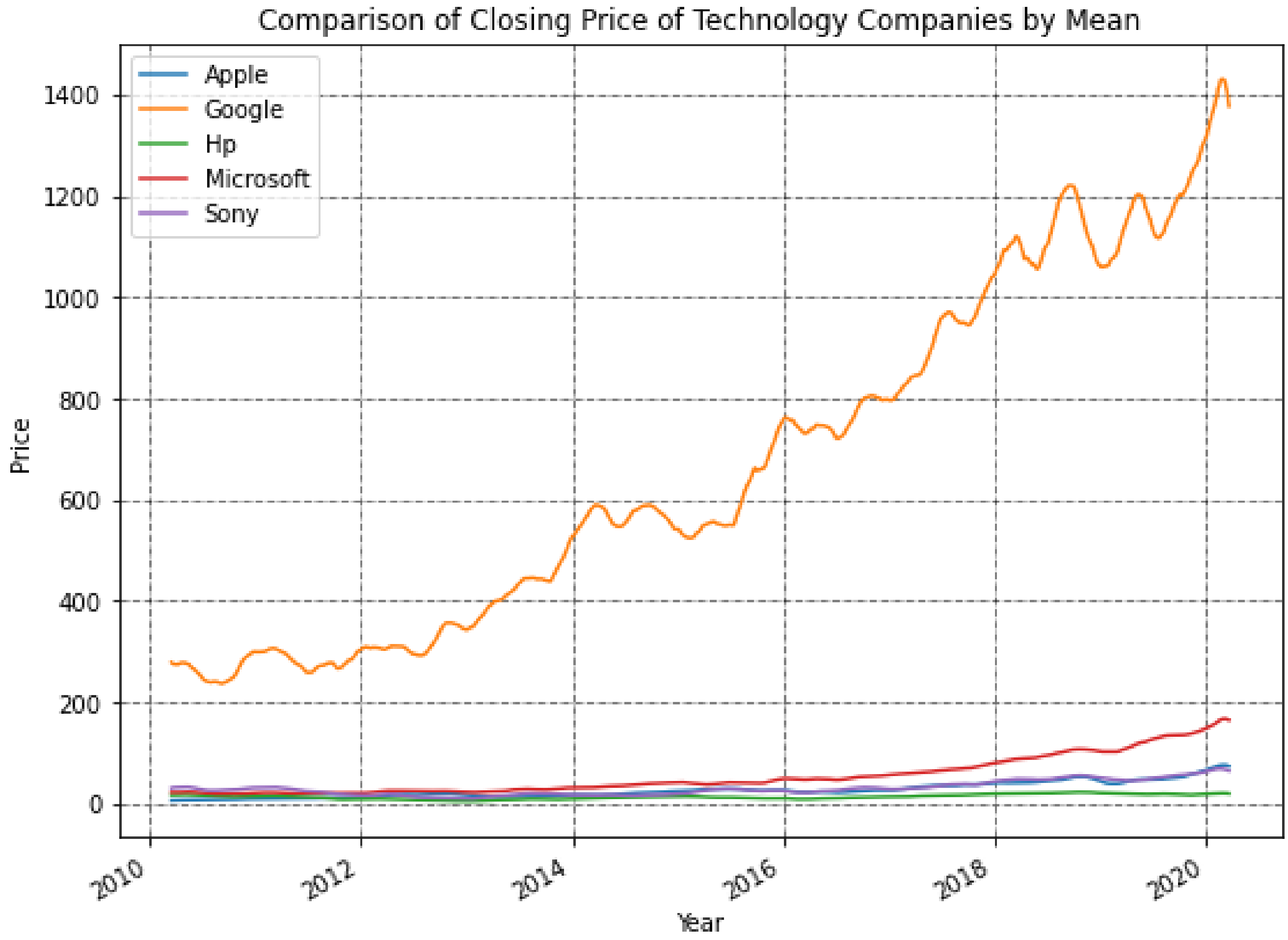

Figure 1: Comparison for all different companies based on the adjusted close price with a rolling window of 50 days.

With all companies showing a positive trend, we can identify their Pearson correlation coefficients with a heat map. The Pearson correlation coefficient is a numerical correlation between a dependent and independent variable. Figure 2 below shows the correlation coefficient between each company. As all trend lines are positive in figure 1, a positive correlation coefficient between each company is expected. The correlations between Microsoft and Apple,Google and Sony show a value greater than 0.8, where 1 would be a perfect fit. This however is not sufficient evidence that Microsoft's stock price is dependent on the respective companies.

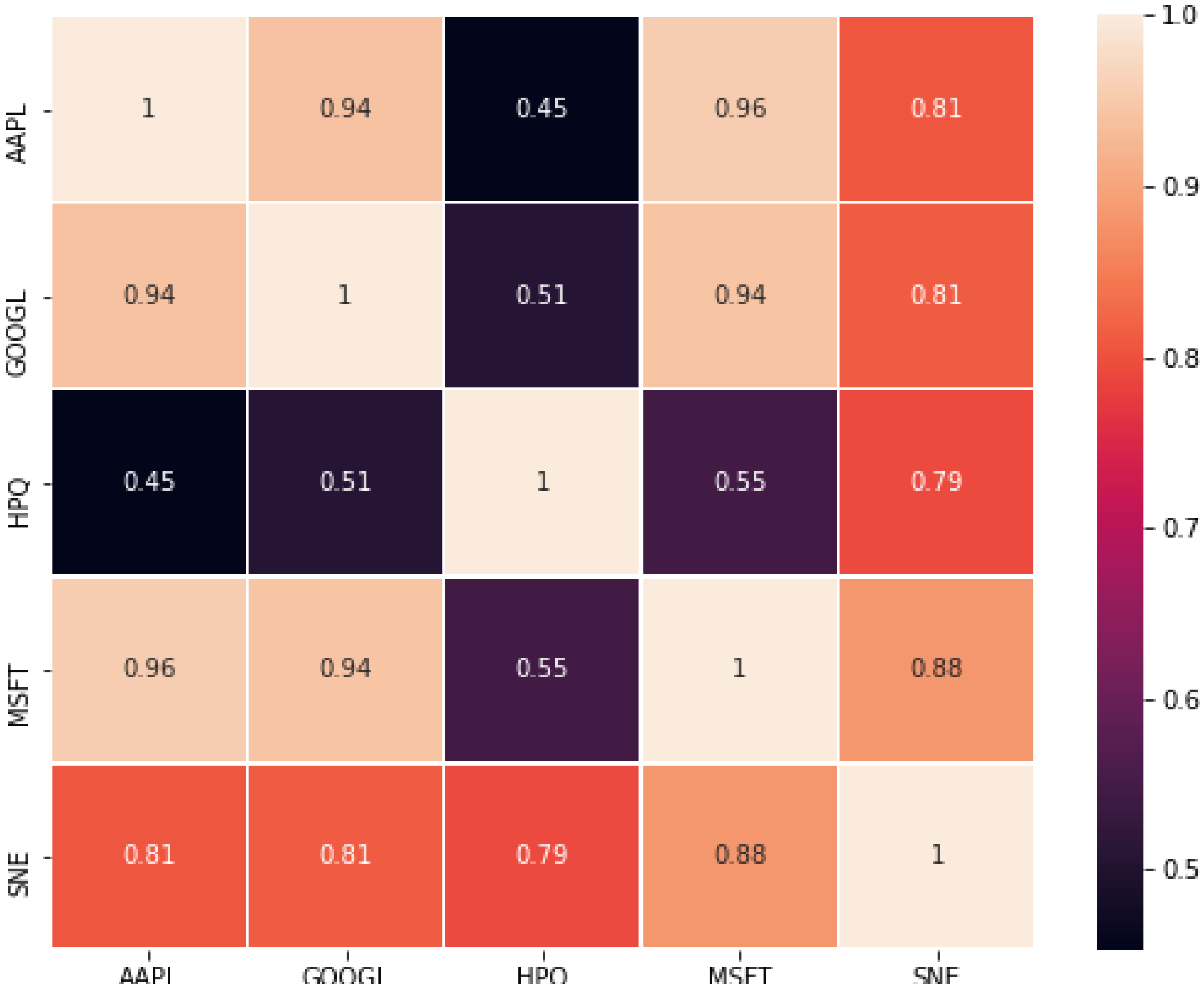

Figure 2: Correlation heat map for each company.

In order to identify if the companies are of significance to Microsoft, three steps need to be taken.

1. Identify if the adjusted close price follows a random walk for each company.
2. Perform Principle component analysis on all the companies to identify the explained variance.
3. Test the co-integration value between Microsoft and any company with a explained variance greater than 0.05.

For the first step we will use the Augmented Dickey Fuller test to identify if the companies adjusted close price follow a random walk with the following hypothesis test.

H_o : The companies adjusted close price follow a random walk.

H_1: The companies adjusted close price do not follow a random walk.

The significance value is calculated at a $p - value < 0.05$. Table 1 below shows the p-value for each of the companies and justifies each adjusted close price follows a random walk with noise.

	Apple	Google	HP	Microsoft	Sony
P-value	0.52	0.72	0.41	0.99	0.85

Table 1: P-value for adjusted close price of each company.

As all companies follow a random walk, we can continue to the second step. Using principle component analysis, we can calculate the explained variance and if the variance value is greater than 0.05, we can conclude the company's adjusted close price is statistically significant and can potentially influence Microsoft's stock price.

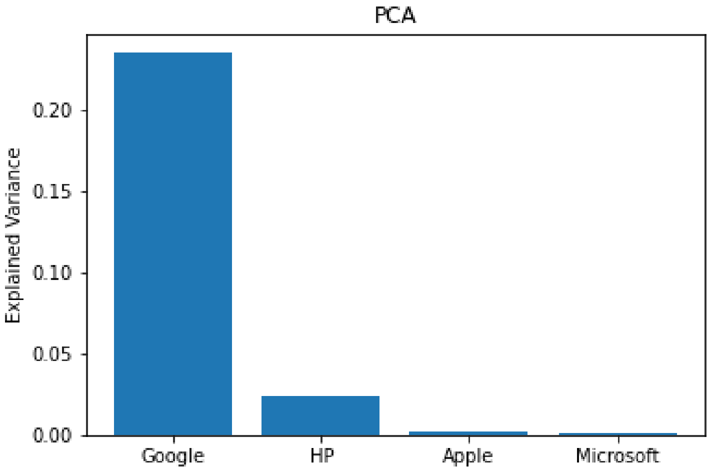

Figure 3: The explained variance value resulting from Principle Component Analysis.

From figure 3, the only company showing a statistically significant explained variance is Google. The other companies from this point will be withdrawn from the data set.

The final step is to inspect the co-integration value between Microsoft and Google. To do this, we need to add a constant to Microsoft adjusted close series and fit the series to an Ordinary Least-squares method. By computing the Augmented Dickey Fuller test on the difference between Google's adjusted close series and Microsoft's adjusted close series, we can come to a conclusion and justify if the two series are co-integrated. This is done with the following hypothesis test.

H_o : There is no evidence of co-integration between Microsoft and Google

H_1: There is a relation between Microsoft and Google's adjusted close price

The significance level for this test is a $p - value < 0.05$. The obtained value from the test was a **p-value = 0.76**.

This shows that Google did not have an influence on Microsoft's stock price and further analysis will continue with only Microsoft's adjusted close price.

4 Modelling

This section will go into detail about the steps required prior to modelling the data. This includes the preprocessing stage, where appropriate features will be created based on statistical methods for time series analysis. Further explanation will be presented on splitting the data so the model is not contaminated with the testing data set and the three machine learning models used to predict Microsoft's future stock price. Based on suitable model metrics, the best model will be chosen.

4.1 Prepossessing

In order to model time series data, the data needs to show stationary. Stationary means the statistical properties including the population mean (μ) and variance (σ^2) do not change over time. Stationary is important because many useful analytical tools and statistical models rely on it [3]. The model which requires stationary in this report is detailed in section 4.3.3.

To identify if Microsoft's adjusted stock price shows stationary, the return is computed. This is calculated by the percentage change between current and prior element. A visual representation of the returns can be seen in figure 4. As Microsoft's returns follow a normal distribution, this is sufficient evidence that the data is stationary.

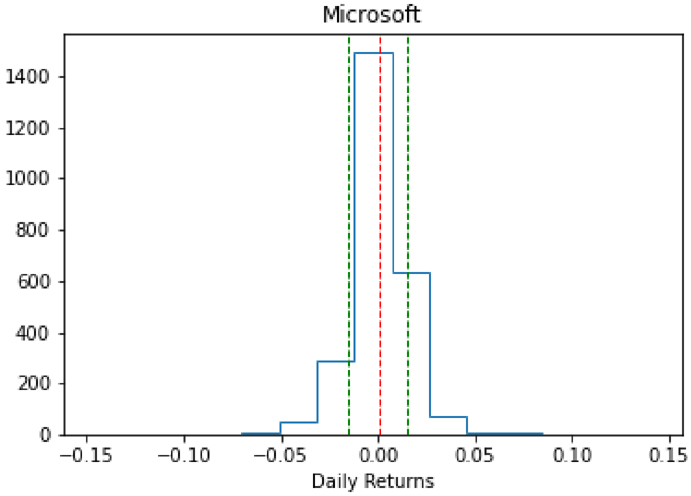

Figure 4: Daily returns for Microsoft's adjusted close price with $\mu = 0.000359$ and $\sigma^2 = 0.000203$.

In order to model the data, a training data set is required to train the model so future predictions can be made. This will require an additional feature. The auto correlation, represents the degree of similarity between a given time series and a lagged version of itself over successive time intervals [4]. Figure 5 shows the correlation for the lagged intervals up to 30 days. This clearly shows that the previous time step is dependent on the value of it's preceding time step.

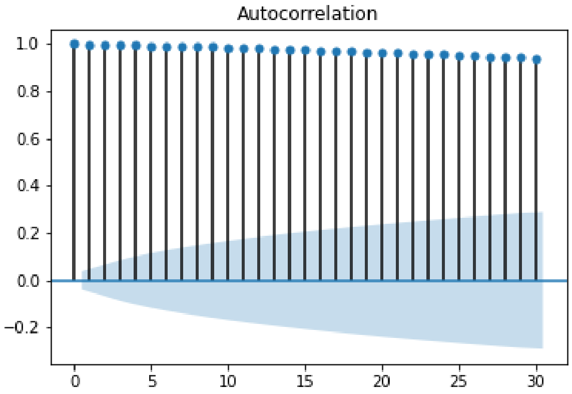

Figure 5: Autocorrelation for Microsoft's adjusted close price.

To identify a feature that will have the highest level of significance, the partial autocorrelation function will identify a relationship between an observation and with an observation at a prior time step. The relationships between intervening observations are removed. Figure 6 displays the result for the adjusted close price. A lag of 1 day shows the most significance and this feature will be included in the data set.

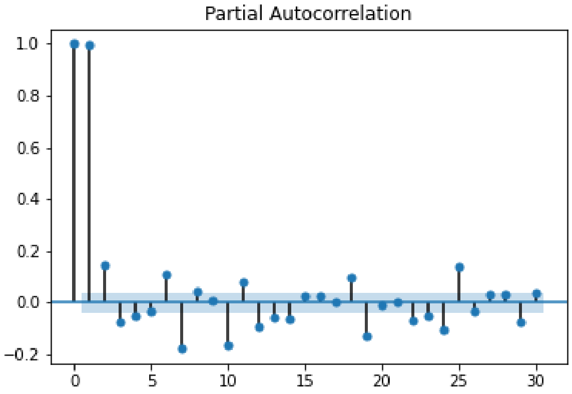

Figure 6: Partial autocorrelation for Microsoft's adjusted close price.

4.2 Training and Testing Data

To prevent the training data being contaminated with the test data, careful consideration needs to be taken to split the data. A time series splitter method is implemented to prevent contamination.The data from 01/01/2010 to 07/02/2020 is used for training each model and the data from 08/02/2020 to 22/03/2020 is used for testing.

4.3 Models

The next step to making predictions is training three machine learning models and identifying the best model. The three models used for testing and validation are Polynomial Linear Regression, XGBoost and ARIMA.

4.3.1 Polynomial Linear Regression

For the Linear regression model, the Scikit Learn module is used to predict Microsoft's stock price. In addition,a polynomial feature is added to the pipeline to improve accuracy. Polynomial features transform to create new versions of input variables for the predictive model [5].

A grid search cross validation method is used for this model. The hyperparameters used for modelling are the following:

$PolynomialDegree$: 1

The rest of the hyperparameters are set to their default value. The results for the Polynomial Linear Regression model can be seen in figure 7.

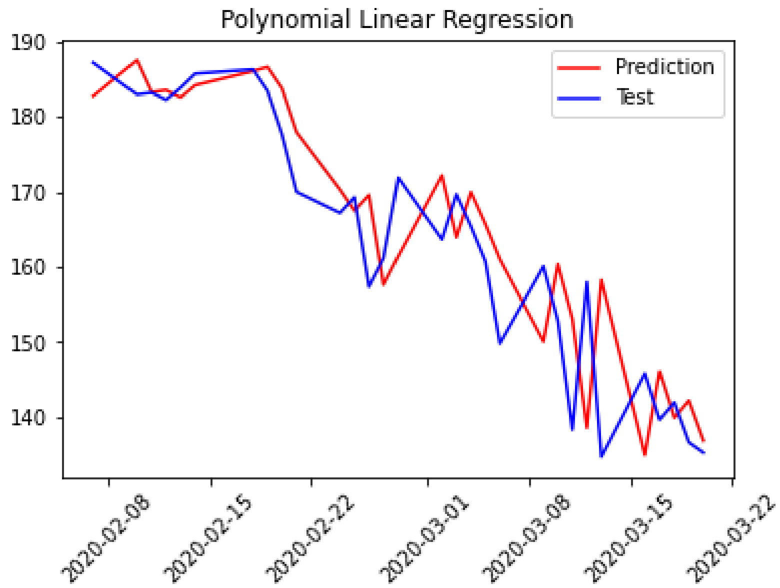

Figure 7: Forecast from Polynomial Linear Regression model.

4.3.2 XGBoost

XGBoost is an optimized distributed gradient boosting model. It provides a parallel to tree boosting. For modelling on the Microsoft data set, a randomized search cross validation is implemented to prevent the model being computationally expensive.

For this model a walk forward validation method is carried out. This method predicts a single value from the training data set, adds the new predicted value to the training data set and the process is iterated until a chosen value from the user. In this case it will be the number of days we want to predict.

The hyperparameters used for the final model are:

$subsample$: 0.03

$n\_estimators$: 500

max_depth: 16

$learning\_rate$: 0.5

$colsample\_bytree$: 0.37

The rest of the hyperparameters are set to their default value. The results from the XGBoost model can be seen in figure 8

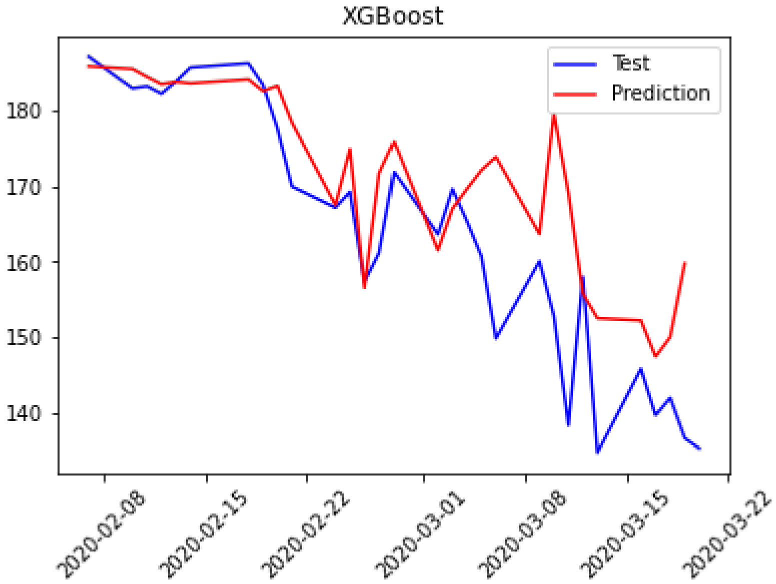

Figure 8: Forecast from XGBoost model.

4.3.3 AutoRegressive Integrated Moving Average (ARIMA)

The ARIMA model is a class of models that explains a given time series based on it's history. This refers to the the previous lags and lagged forecast errors. This enables the model to predict future values.

Similar to the XGBoost model, a walk forward validation method is used. Three hyperparameters are tuned for this model and are defined as:

p: Periods taken for the autoregressive model

d: Order of integration

q: The Periods in moving average

The hyperparameters choose for this model are: $p = 2$, $d = 2$, $q = 1$. The results from the ARIMA model can be seen in figure 9.

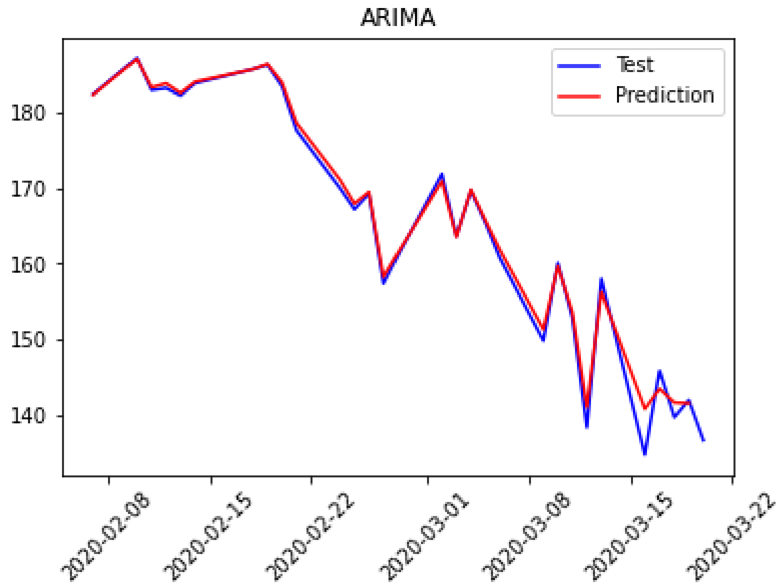

Figure 9: Forecast from ARIMA model.

4.4 Best Model

From figures 7,8 and 9 the ARIMA model shows the most accurate prediction. To determine the the best model for prediction, two model metrics are taken into consideration. The Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE). Table 2 showcases the values for each of the metrics.

	Polynomial Linear Regression	XGBoost	ARIMA
RMSE	8.61	11.34	8.10
MAE	6.62	7.52	6.24

Table 2: Model metrics for all three machine learning models.

By collective looking at figure 9 and table 2, the ARIMA model shows the most accurate prediction for forecasting Microsoft's adjusted close price.

5 Conclusion

This report identifies the necessary steps required to predict a stock price of any company on the market. The data involved 10 years worth of Microsoft's stock data from the Yahoo finance API and technology companies including Google, Sony, HP, and Apple. Exploratory data analysis was conducted on the data to determine if there was any relationships between the companies. By using statistical methods such as hypothesis testing ,Principle Component Analysis and the Augmented Dickey Fuller test, we were able to conclude that no other company had an influence on Microsoft's stock price. An additional feature with a lag of one day was added to the dataset which was derived from the partial autocorrelation function. Three machine learning models, which included Polynomial Linear Regression, XGBoost and ARIMA were trained and fitted to the training dataset. **The best model proved to be ARIMA model with a RMSE = 8.10 and MAE = 6.24**. Finally, there was not a company that showed a statistically significant co-integration value to conclude if another companies stock price had an influence on Microsoft's stock price, however the ARIMA model showed an accurate price prediction can be determined.

5.1 Assumptions

The following assumptions need to be taken into consideration for this report:

- The market has processed all available information and that it is reflected in the price chart.
- Price movement of stocks is dependent upon supply and demand.
- Limited to technical analysis only, no fundamental factors were taking account.

5.2 Future Work

For anyone would like to continue the work carried out in this report. The following statements identify the next steps:

- Compare other technology companies, outside this report, to identify co-integration.
- Attempt to forecast stock data using a deep neural network, specifically Long Short Term Memory (LSTM).

References

[1] Chen.J,17th May 2020,Guide to Technical Analysis, Investopedia, URL: https://www.investopedia.com/terms/t/technical-analysis-of-stocks-and-trends

[2] Bischoff B, 23rd May 2019, Adjusted Closing Price vs Closing Price, The Nest,URL : https://budgeting.thenest.com/adjusted-closing-price-vs-closing-price-32457.html

[3] Palachy S, 2019, Detecting stationarity in time series data, KD nuggets, URL: https://www.kdnuggets.com/2019/08/stationarity-time-series-data.html

[4] Smith T, 10th March 2020, Autocorrelation, Investopedia, URL : https://www.investopedia.com/terms/a/autocorrelation.asp

[5] Brownlee J, 28th August 2020, How to use polynomial features for machine learning, Machine Learning Mastery, URL : https://machinelearningmastery.com/polynomial-features-transforms-for-machine-learning/

[6] Prabhakaran S, ARIMA Model- Complete Guide to Time Series Forecasting in Python, machinelearningplus, URL: https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/