University of CINCINNATI

Forecasting Google Stock Price

ARIMA, KNN & Neural Networks

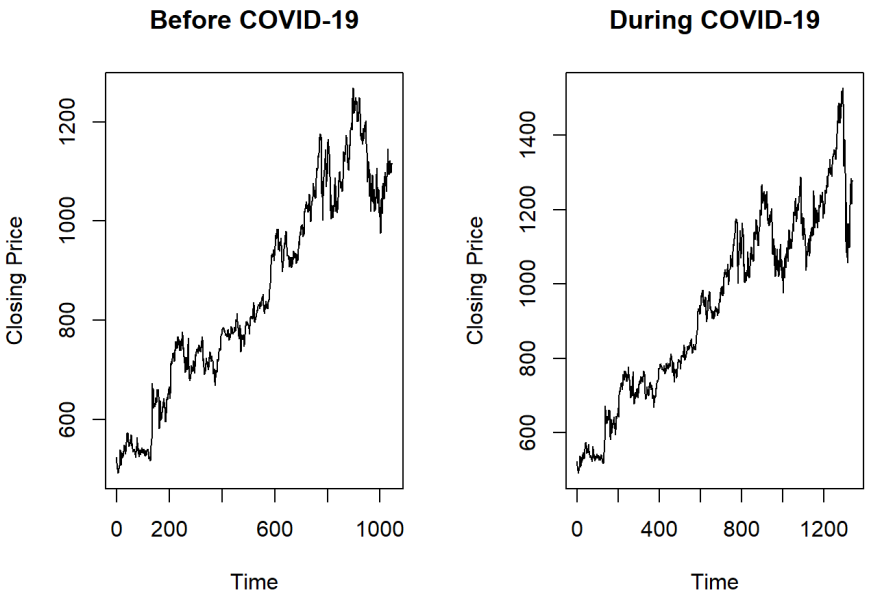
Synopsis

The goal of this project is to predict the future stock price of Google using various predictive forecasting models and then analysing the various models. The dataset for Google stocks is obtained from Yahoo Finance using Quantmod package in R. The timeline of the data is from 2015 till present day(4/26/2020).

Introduction

A forecasting algorithm is a process that seeks to predict future values based on the past and present data. These historical data points are extracted and prepared trying to predict future values for a selected variable of the dataset. During market history there have been a continuous interest trying to analyze its tendencies, behavior and random reactions. This continuous concern to understand what happens before it really happens motivates us to continue with this study. We shall also try and understand the impact of **COVID-19** disaster on the stock prices.

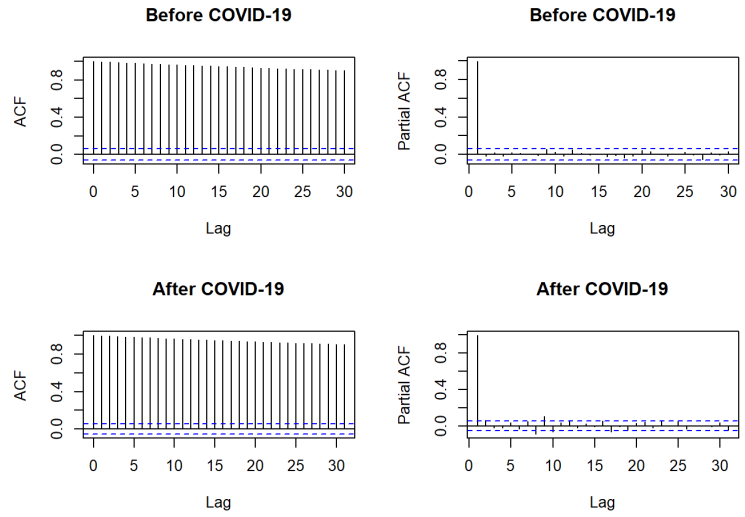
Graphical Representation of Data



Vipul Mayank | Varun Varma | Sayak Chakraborty | Priyanka Pavithran

ARIMA Model

Let us first analyze the ACF and PACF Graph of each of the two datasets.



We then conduct an **ADF (Augmented Dickey-Fuller)** test and **KPSS (Kwiatkowski-Phillips-Schmidt-Shin)** test to check for the stationarity of the time series data for both the datasets closing price.

```
##
##  Augmented Dickey-Fuller Test
##
## data:  tsData_before_covid
## Dickey-Fuller = -2.8718, Lag order = 10, p-value = 0.2093
## alternative hypothesis: stationary
```

Code

```
##
##  Augmented Dickey-Fuller Test
##
## data:  tsData_after_covid
## Dickey-Fuller = -3.7522, Lag order = 11, p-value = 0.02138
## alternative hypothesis: stationary
```

From the above ADF tests, we can conclude the following:

- For the dataset before COVID-19, the ADF tests gives a p-value of **0.2093 which is greater than 0.05**, thus implying that the time series data is **not stationary**.
- For the dataset after COVID-19, the ADF tests gives a p-value of **0.01974 which is lesser than 0.05**, thus implying that the time series data is **stationary**.

```
##
##  KPSS Test for Level Stationarity
##
## data:  tsData_before_covid
## KPSS Level = 12.468, Truncation lag parameter = 7, p-value = 0.01
```

Code

```
##
##  KPSS Test for Level Stationarity
##
## data:  tsData_after_covid
## KPSS Level = 15.692, Truncation lag parameter = 7, p-value = 0.01
```

From the above KPSS tests, we can conclude the following:

- For the dataset before COVID-19, the KPSS tests gives a p-value of **0.01 which is less than 0.05**, thus implying that the time series data is **not stationary**.
- For the dataset after COVID-19, the KPSS tests gives a p-value of **0.01 which is less than 0.05**, thus implying that the time series data is **not stationary**.

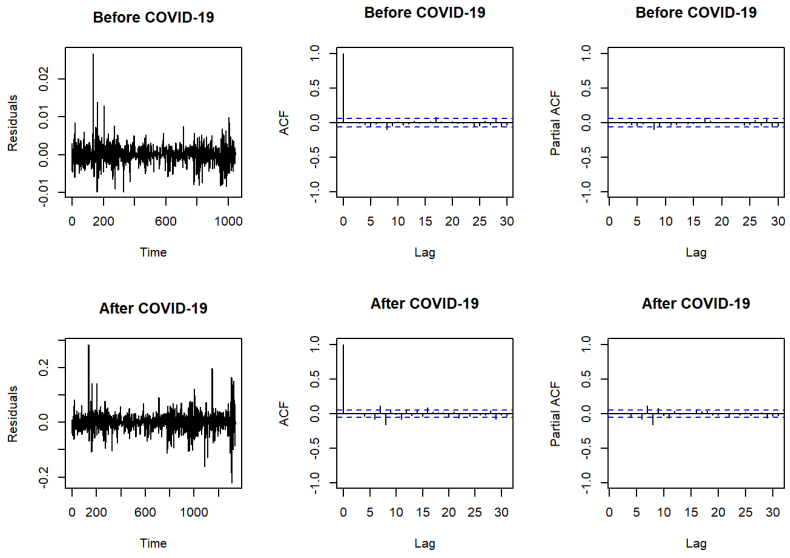
Thus, we can conclude from the above two tests that the time series data is **not stationary**.

We then use the **auto.arima** function to determine the time series model for each of the datasets.

From the auto.arima function, we conclude the following models for the two datasets:

- **Before COVID-19: ARIMA (2,1,0)**
- **After COVID-19: ARIMA (1,1,1)**

After obtaining the model, we then perform residual diagnostics for each of the fitted models.



From the residual plot, we can confirm that the residual has a mean of 0 and the variance is constant as well . The ACF is 0 for lag> 0, and the PACF is 0 as well.

So, we can say that the residual behaves like white noise and conclude that the models ARIMA(2,1,0) and ARIMA(1,1,1) fits the data well. Alternatively, we can also test at a significance level of 0.05 if residual follow white noise using the Box-Ljung Test.

Code

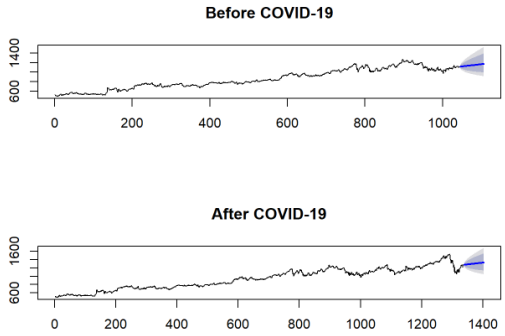
```
##
##  Box-Ljung test
##
## data:  modelfit_before_covid$residuals
## X-squared = 0.0052952, df = 1, p-value = 0.942
```

Code

```
##
##  Box-Ljung test
##
## data:  modelfit_after_covid$residuals
## X-squared = 8.6593e-06, df = 1, p-value = 0.9977
```

Here, the p value for both the models is greater than 0.05 . Hence, at a significance level of 0.05 we fail to reject the null hypothesis and conclude that the residual follows white noise. This means that the model fits the data well.

Once we have finalized the model for each of the datasets, we can then forecast the prices of the stock in the future days.

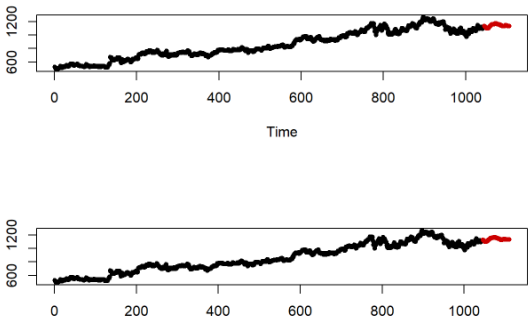


KNN Regression Time Series Forecasting Model

KNN model can be used for both classification and regression problems. The most popular application is to use it for classification problems. Now with the tsfknn package KNN can be implemented on any regression task. The idea of this study is illustrating the different forecasting tools, comparing them and analysing the behavior of predictions. Following our KNN study, we proposed it can be used for both classification and regression problems. For predicting values of new data points, the model uses 'feature similarity', assigning a new point to a values based on how close it resembles the points on the training set.

The first task is to determine the value of k in our KNN Model. The general rule of thumb for selecting the value of k is taking the square root of the number of data points in the sample. Hence, for the data set before COVID-19 we take k = 32 and for the dataset after COVID-19, we take k = 36.

Code



We then evaluate the KNN model for our forecasting time series.

```
##       RMSE       MAE      MAPE
## 44.046959 33.780280  3.170659
```

```
##       RMSE       MAE      MAPE
## 45.970317 35.782351  3.362729
```

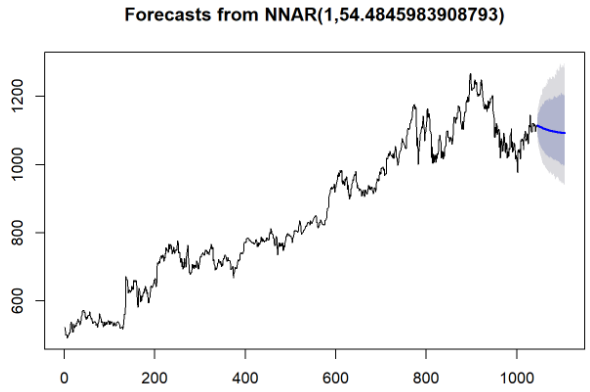
Feed Forward Neural Network Modelling

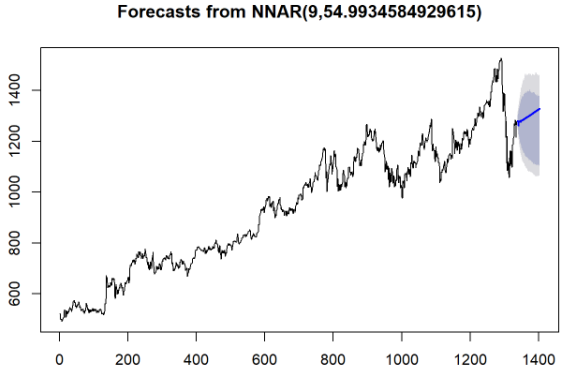
The next model which we would try, and implement is a forecasting model with neural networks. In this model, we are using single hidden layer form where there is only one layer of input nodes that send weighted inputs to a subsequent layer of receiving nodes. The nnetar function in the forecast package fits a single hidden layer neural network model to a timeseries. The function model approach is to use lagged values of the time series as input data, reaching to a non-linear autoregressive model.

The first step is to determine the number of hidden layers for our neural network. Although, there is no specific method for calculating the number of hidden layers, the most common approach followed for timeseries forecasting is by calculating is using the formula:

$$** N(hidden) = Ns / (a * (Ni + No)) **$$

where Ns: Number of train samples Ni: Number of input neurons No: Number of output neurons a : 1.5^-10





We then analyze the performance of the neural network model using the following parameters:

Code

```
##                     ME     RMSE      MAE          MPE     MAPE      MASE
## Training set 0.1218763 13.02254 8.772272 -0.007829055 1.023443 0.9957889
##                     ACF1
## Training set 0.02226194
```

Code

```
##                     ME     RMSE      MAE         MPE     MAPE      MASE
## Training set 0.085951 14.64451 9.797227 -0.01463471 1.030593 0.9358857
##                     ACF1
## Training set 0.01603359
```

Vipul Mayank | Varun Varma | Sayak Chakraborty | Priyanka Pavithran

Comparison of all models

We now analyse all the three models with parameters such as **RMSE (Root Mean Square Error), MAE (Mean Absolute Error) and MAPE (Mean Absolute Percentage Error)**.

Summary of Models for data before COVID-19

Model	RMSE	MAE	MAPE
ARIMA	13.08	8.81	1.02
KNN	44.04	33.78	3.17
Neural Network	13.01	8.77	1.02

Summary of Models for data after COVID-19

Model	RMSE	MAE	MAPE
ARIMA	16.64	10.44	1.09
KNN	45.97	35.78	3.36
Neural Network	14.71	9.82	1.03

Thus, from the above summary of model performance parameters, we can see that Neural Network Model performs better than the ARIMA and the KNN Model for both the datasets. Hence, we will use the Neural Network Model to forecast the stock prices for the next two months.

Final Model: Before COVID-19

We now forecast the values for March and April using the data till February and then compare the forecasted price with the actual price to check if there is any significant impact that can attributed because of COVID-19.

Code

Show 10 entries Search:

	Date	Actual.Values	Forecasted.Values
	All	All	All
1	2020-03-02	1389.109985	1114.41320638701
2	2020-03-03	1341.390015	1113.63010207459
3	2020-03-04	1386.52002	1112.86965352889
4	2020-03-05	1319.040039	1112.1312154885
5	2020-03-06	1298.410034	1111.41415951951
6	2020-03-09	1215.560059	1109.3852463583
7	2020-03-10	1280.390015	1108.7477611786
8	2020-03-11	1215.410034	1108.1287584478
9	2020-03-12	1114.910034	1107.52770463608
10	2020-03-13	1219.72998	1106.94408094182

Showing 1 to 10 of 40 entries Previous 1 2 3 4 Next

From the table we can see that the actual values of Google Stock in general are a bit higher than forecasted values during the month of March and April. Thus, we can say that Google has still performed considerably well inspite of this global pandemic.

Vipul Mayank | Varun Varma | Sayak Chakraborty | Priyanka Pavithran

Final Model : After COVID-19

We now forecast the values for May and June using the data till April to get an idea of future stock price of Google.

Code

Show 10 entries Search:

	Date	Price
	All	All
1	2020-04-27	1279.84354494492
2	2020-04-28	1276.78845538086
3	2020-04-29	1261.30579864621
4	2020-04-30	1274.80664409745
5	2020-05-01	1276.83305724482
6	2020-05-02	1277.96260267176
7	2020-05-03	1277.64865334086
8	2020-05-04	1280.003342512
9	2020-05-05	1280.29131939828
10	2020-05-06	1276.18314329299

Showing 1 to 10 of 65 entries Previous 1 2 3 4 5 6 7 Next

From the table, we can conclude that the prices of Google Stock will continue to rise and perform well in the coming months of May and June.