VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering

J Component report

Programme	**: B.Tech**
Course Title	**: Foundations of Data Analytics**
Course Code	**: CSE3505**
Slot	**: F1**

STOCK MARKET PREDICTION ANALYSIS

Team Members: Anchana. V | 20BCE1788

Aryaan Chauhan | 20BCE1214

Faculty: DR.A. SHEIK ABDULLAH **Sign:**

Date:

Index Page:

List of Abbreviations:

EWMA	Exponential Weighted Moving Average
ACF	Autocorrelation plot
PACF	Partial autocorrelation plot
ADF	Augmented Dicky Fuller Test
XGBoost	Extreme Gradient Boosting
LSTM	Long short-term memory
SES	Simple Exponential Smoothing
HES	Holt’s Exponential Smoothing
HWES	Holt Winter’s Exponential Smoothing
RMSE	Root means square error
MAPE	Mean absolute percentage error
MAE	Mean absolute error
VWAP	Volume-weighted average price
EMA	Exponential moving average
ARIMA	AutoRegressive Integrated Moving Average
AR	Auto regressive
MA	Moving average
AIC	Akaike Information criteria
MLE	Maximum likelihood estimation

List of Tables:

For data set we used data set from,

https://www.kaggle.com/code/kp4920/s-p-500-stock-data-time-series-analysis
in this data set there are total 126217 real time data, and from here we used data set for five companies Google, Amazon, IBM, Microsoft and Nike.

List of Figures:

ABSTRACT:

This Document is meant to delineate the features of a Stock Market Prediction using RStudio. Regardless of your investment strategy, fluctuations are expected in the financial market. Despite this variance, professional investors try to estimate their overall returns. Risks and returns differ based on investment types and other factors, which impact stability and volatility. To attempt to predict returns, there are many computer-based algorithms and models for financial market trading. Yet, with new techniques and approaches, data science could improve quantitative researchers' ability to forecast an investment's return.

There are many technological innovations that is being incorporated in top companies that are using quantitative trading with big data processing and machine learning technologies to provide real-time market analytics.

The main aim of this project is that it focuses on using different models that forecasts an investment's return rate and predict the trend of the market for upcoming years, months or days. The objectives would be to train and test an algorithm on historical prices, with as much accuracy as possible.

INTRODUCTION:

The project is developed on an idea, that if any investor wants to invest in some new company, then this project would be able to give an analysis of the stock price of a reputable company which has been in the business for a long time and has always faced changes. We also plan to predict the future prices of its stock which in turn will reflect on the working the company.

The entire idea of a stock market predictor, with the help of Machine Learning, is to gain significant number of profits by understanding the investments that would provide efficient and profitable returns. There are many factors involved in such market investments like physical, psychological, rational and irrational behaviour, and so on. All these factors combine to make the share prices Dynamic and Volatile. These predictions, basically help one to discover the future value of the company stocks and other financial assets traded on exchange.

The main objective of this project is to provide descriptive and predictive analysis of the stock price of a company by the analysis of the data provided of the company's stock price data over a period of time.

The methods and the model which we are going to use int this project is that for the analysis and prediction of stock prices, we will be using ARIMA (Autoregressive Integrated Moving Average) Modelling in R and Exponential Weighted Moving Average.

Stock Market Prediction is an extremely difficult task, and there are various methods to achieve such a task. Due to the fluctuating nature of the stock, the stock market is uncertain in various ways, making it a complex model.

The various methods that are discussed in this research paper, along with their advantages and Disadvantages, to achieve this feat are:

- Exponential Weighted Moving Average.
- ARIMA Model.

Exponential Weighted Moving Average:

Working model:

It is moving weighted average is a stock indicator with greater weight and significance on the most recent data points. It is used in technical analysis which reacts more significant to recent price changes.

Advantages:

Since it is based on the calculation with greater weight and significance on more recent trends, it gives a more realistic stock price cost compared to other models.

Disadvantages:

As this model depends on the data of the past and if there is any natural calamity, or abrupt changes that are going to affect the prices of the stocks then in this scenario this model wouldn't be able to give the accurate pricing and there will large deviation between the predicted stock price and the actual stock price.

ARIMA Model:

Working model:

Identified, estimated, and diagnosed with time-series data. Generate short-term forecasts. The future Value of a variable is a linear combination of past values and past errors.

Advantages:

- Robust and Efficient forecasting of financial time series. Small standard of error of the regression.
- Better understanding of the time series patterns

Disadvantages:

- Suitable for short-term predictions only.
- Captures only linear relationships.

LITERATURE SURVEY:

“Review of Data Pre-processing Techniques in Data Mining”

By Bhaya, Wesam..

Introduction

The given Research Paper talks about the possible Data Pre-processing techniques that can be used in order to obtain better and more efficient results. Even though this Research Paper talk about Pre-processing in the field of Data Mining, similar techniques can be applied in the field of Machine Learning for optimized results. Raw Data usually susceptible to Missing Values, noisy data, incomplete data, inconsistent data and outlier data. So, it is important for these data to be processed before being worked upon. Data Pre-processing deals with data preparation and transformation of the dataset and seeks at the same time to make knowledge discover more efficient.

Several techniques that can be used for Data Pre-processing are

- Cleaning
- Integration
- Transformation
- Reduction

The Research Paper elaborates on these techniques and when can they be utilized.

Proposed Techniques

1. **Data Cleaning**
   Row Records may have incomplete records, noise values, outliers and inconsistent data. Data Cleaning is used to find the missing values, smooth noise data, recognize outliers and correct inconsistent data.
   Missing values can be filled in the following ways
   - Ignore the Tuple
   - Fill the Values Manually
   - Use a Global Constant to fill the Missing Values
   - Use the Attribute Mean to Fill the Missing Values
   - Use the Attribute mean for all samples belonging to the same class as the given tuple
   - Use the most probable value to find the missing value

2. **Noise Data**
   Noise Data is a random error or variance in a measured variable. Noise Error means that there is an error data or outliers which deviates from the normal.

Possible Techniques include

- Binning – Smoothing stored data based on its “neighbourhood” which is the values around it. Sorted values are divided into a number of buckets or bins. They perform local smoothing
- Regression – Fitting the data into a function. Uses the line of best fit for two variables, so that each attribute can be used to predict the other
- Clustering – Grouping set of points into clusters according to a distance measure. Technique is used to detect outliers, since it is grouping similar points into a cluster.

3. Data Integration

This technique works by combining data from multiple and various resources into one consistent data and store, like in data warehouse. These resources can have multi database, files, or data cubes. In data integration, there are a number issues like - schema integration, redundancy and object matching which are important aspects.

4. Data Transformation

Includes transforming the data to suitable forms, it includes the following

- Smoothing
- Aggregation
- Generalization
- Normalization
- Data reduction
- Data Cube aggregation
- Attribute Subset selection
- Dimensionality reduction
- Numerosity reduction

Conclusion

Real world data tend to be incomplete, inconsistent, noisy and missing. Data pre-processing, which includes Data integration, Data transformation, Data Cleaning, Data transformation and Data Reduction, is one of the important matters for both data warehousing, mining, and Machine Learning. The study explains an overview of those techniques.

Deep Learning for Stock Market Prediction

By M. Nabipour , P. Nayyeri , H. Jabani and E. Salwana

ABSTRACT

Due to its intrinsic dynamism, non-linearity, and complexity, the prediction of stock group prices has long been appealing to and difficult for investors. This essay focuses on stock market group future predictions. From the Tehran Stock Exchange, four groups named diversified financials, petroleum, non-metallic minerals, and basic metals were selected for experimental evaluations. Based on ten years' worth of historical documents, information for the groupings was gathered. The value forecasts are made for the next 1, 2, 5, 10, 15, 20, and 30 days. Different machine learning techniques were used to forecast future stock market group values. We used artificial neural networks (ANN), recurrent neural network (RNN), long short-term memory, bagging, random forest, adaptive boosting (Adaboost), gradient boosting, and eXtreme gradient boosting (XGBoost) (LSTM). Each of the prediction models' inputs was given ten technical indicators. Finally, based on four indicators, the forecasts' findings for each technique were provided. The algorithm used in this paper with the highest model fitting ability, LSTM, exhibits more accurate findings. Additionally, Adaboost, Gradient Boosting, and XGBoost frequently compete fiercely for tree-based models.

SUMMARY

The prediction of stock groups values has always been attractive and challenging for shareholders due to its inherent dynamics, non-linearity, and complex nature. This paper concentrates on the future prediction of stock market groups.We employed decision tree, bagging, random forest, adaptive boosting (Adaboost), gradient boosting, and eXtreme gradient boosting (XGBoost), and artificial neural networks (ANN), recurrent neural network (RNN) and long short-term memory (LSTM). Finally, the results of the predictions were presented for each technique based on four metrics.

Stock Markets: An Overview
By Rjumohan Asalatha

Abstract

Stock markets are without any doubt, an integral and indispensable part of a country's economy. But the impact of stock markets on the country's economy can be different from how the other countries' stock markets affect their economies. This is because the impact of stock markets on the economy depends on various factors like the organization of stock exchanges, its relationship with other components of the financial system, the system of governance in the country etc. All of these factors are distinct for each country; therefore, the impact of stock markets on a country's economy is also distinct. Over the years, the Indian capital market system has undergone major fundamental institutional changes which resulted in reduction in transaction costs, significant improvements in efficiency, transparency and safety. All these changes have brought about the economic development of the economy through stock markets. In the same way, economic expansion fuelled by technological changes, products and services innovation is expected
to create a high demand for stock market development. The present paper is divided into two parts: in the first section, the evolution of international stock markets and the developments in Indian stock markets are briefly reviewed to help us understand how stock markets have emerged as the driving economic forces that they are today; and the second part presents a number of studies that review the impact of financial development, stock market development and its functions and its possible impact on economic growth.

Summary

The NSE review (2012) records that the Indian equity market has a nationwide trading network with over 4827 corporate brokers and about 10,165 traders registered with the SEBI. The stock markets have become an integral component of our economy. These stock markets aren't going to be replaced anytime soon and they are meant to stay for years to come. They will continue to remain the major driving force of the economy in many countries. But there are a few things the analysts anticipate. NYSE still remains the largest and inarguably the most powerful stock exchange in the world and its market capitalization is larger than that of Tokyo, London and NASDAQ. Also in all probability, stock markets may merge with each other in the years to come. Some analysts have even forecasted a single global stock market but this is highly unlikely in the recent future. Irrespective of the changes in stock markets, stock markets will continue to run the global economies for the long foreseeable future.

What is EMA? How to Use Exponential Moving Average With Formula

By
JAMES CHEN

Exponential Moving Weighted Average

Exponential moving weighted average is a stock indicator with greater weight and significance on the most recent data points. It is used in technical analysis which reacts more significant to recent price changes. Traders often use Exponential Moving Weighted Average algorithm to find more accurate stock price for their buying and selling of stocks. Based on the timeline of checking the stock price can use several different EMA lengths such as 10-day, 50-day and 200-day moving averages.

Summary:

Using the research paper we were able to under the use of EMWA and how is it used to calculate and predict next future prices of the stock market. It also shows the benefit of EMWA over other moving average algorithms.

"Exponential Smoothing Methods for Detection of the Movement of Stock Prices"

By Shaik Shahid, SK. Althaf Rahaman,

Introduction

Exponential smoothing is a technique for smoothing time data in order to predict the near future. Exponential smoothing's fundamental principle is to forecast future values by taking a weighted average of all previous values in our time series data. Exponential smoothing is a technique for forecasting time series data based on three factors: level, trend, and seasonal component.

Based on these methods, we have three types of smoothing techniques:

- **Simple Exponential Smoothing (SES)**
  The SES method can be used to anticipate future values of time series data that does not contain a trend or seasonality.
  The forecast equation for SES is:

  Forecast = Estimate level at most recent time point

$$F_{t+k} = L_t$$

the kth step ahead of SES forecast is simply the most recent Estimate of level (L), at time (t) for which we need to estimate the level using the level updating equation:

$$L_t = \alpha Y_t + 1 - \alpha\ L_{t-1}$$

The above equation states that the algorithm is learning the new level from the newest data it is seeing.

- **Holt’s Exponential Smoothing (HES)**

Holt’s Exponential Smoothing is also called Double Exponential Smoothing. The major goal of this method is to extend the SES setup to include a trend component. When there is a trend in the time series data but no seasonality, Halt's Exponential Smoothing might be used.

The forecast equation for Holt’s Exponential Smoothing is:

Forecast = Estimated level + Trend at most recent time point

$$F_{t+k} = L_t + KT_t$$

Here, we have two update equations. One for level and one for trend. Level updating equation,

$$L_t = \alpha Y_t + (1 - \alpha)\ Lt_{-1} + T_{t-1}$$

The above equation states that we are adjusting previous level by adding trend to it.

Trend updating equation,

$$T_t = \beta\ (L_t - L_t{-}1) + (1 - \beta)\ T_{t-1}$$

The above equation states that we are updating previous trend by using the difference between the most recent level values.

- **Holt Winter’s Exponential Smoothing (HWES)**

Any Winter’s Exponential Smoothing also called Holt-Winter’s Exponential Smoothing and even sometimes called Triple Exponential Smoothing. This method takes the idea of Holt’s method and adds a seasonal component to create the even more complex system. We assume here, that the series has a level, trend, seasonality with M seasons and noise.

The forecast equation for Winter’s Exponential Smoothing is:

Forecast = Estimated level + Trend + Seasonality at most recent time point

$$F_{t+k} = L_t + KT_t + S_{t+k-M}$$

In Winter’s Exponential Smoothing method, we have three smoothing constants and there are three updating equations.

Level updating equation,

$$L_t = \alpha \frac{\mathrm{Yt}}{\mathrm{St-M}} + (1 - \alpha)(L_{t-1} + T_{t-1})$$

Here, when we divide Y by S it means, we are de-seasonalizing the value of Y. In this level equation we are therefore updating the previous level, else of t-1 by adding the previous trend estimate, $T_{t-1}$and then combining with the de-seasonalized value of Y_t.

Trend updating equation,

$$T_t = \beta\ (L_t - L_{t-1}) + (1 - \beta)\ T_{t-1}$$

The above equation is similar to the one in Holt’s Exponential Smoothing method.

Seasonality updating equation,

$$S_t = \gamma \frac{\mathrm{Yt}}{\mathrm{Lt}} + (1 - \gamma)\ S_{t-M}$$

Here, we can see the Y_t is divided by the level component L_t. This gives the de-trended value of Y. So, the seasonal component St-M with the de-trended value of Y_t.

Smoothing constant (γ) is controlling the speed of adjusting the seasonality

Proposed Model

Stock price data-sets containing historical stock prices of two Indian IT businesses, Tata Consultancy Services Limited (TCS) and Hindustan Computers Limited (HCL), were downloaded from Kaggle.com for this time-series data analysis task.

Date, Symbol, Series, Prev Close, Open, High, Low, Last, Close VWAP, Volume, Turnover, Trades, Deliverable Volume, and percent Deliverable are among the attributes in the data collection.

The data set is split into two sections: training data and test data.

Data cleaning, data transformation, and data splitting are all steps in preparing the data collection for use in the model (train dataset, test dataset). The data had previously been cleansed and converted. The data from the previous three years is divided into two training sets and one testing set.

Results and Conclusion

The root means square error (RMSE), mean absolute percentage error (MAPE) and mean absolute error (MAE) metrics are used to evaluate the effectiveness of the methods.

According to the results, Holt's approach of exponential smoothing performed the worst on the TCS data-set, while the simple exponential smoothing method performed the worst on the HCL data-set. On the other hand, Holt-exponential Winter's smoothing method performed the best on both firms' data sets, outperforming the other two methods.

PROPOSED METHODOLOGY:

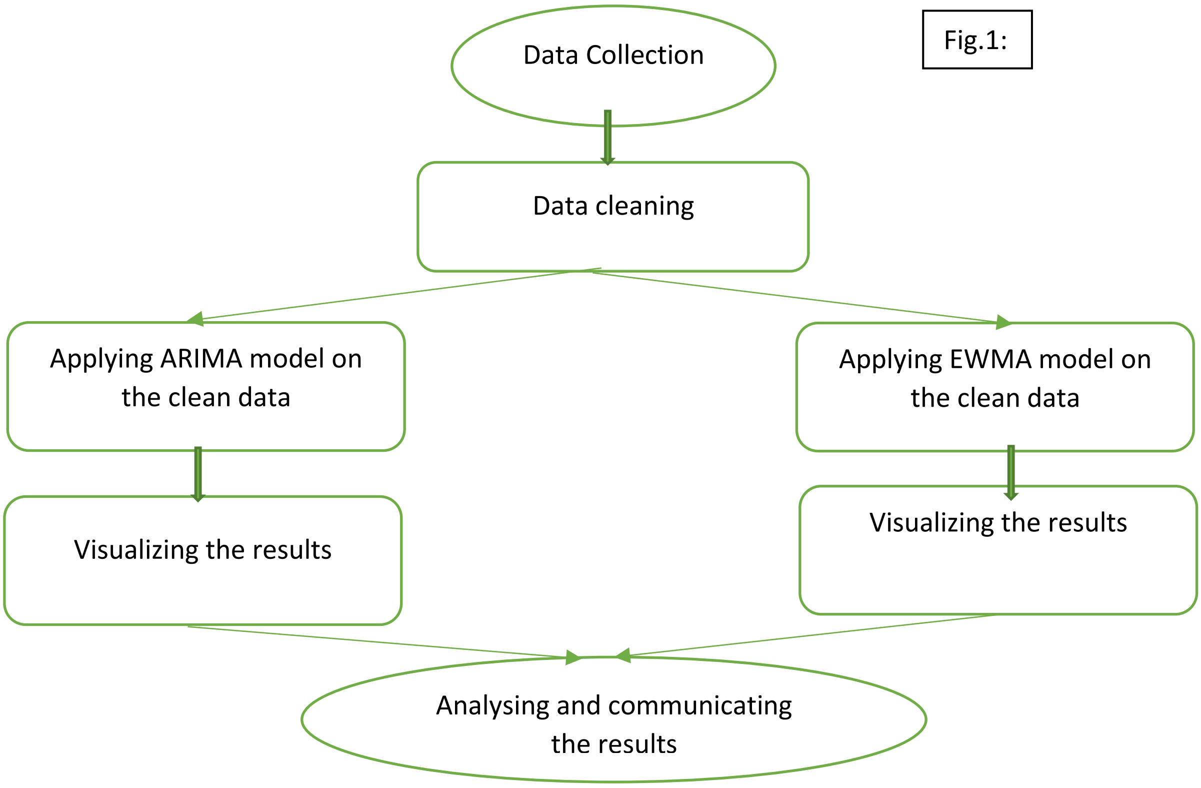

Data Collection
Data cleaning
Applying ARIMA model on the clean data
Applying EWMA model on the clean data
Visualizing the results
Visualizing the results
Analysing and communicating the results

Fig.1:

Exponential Weighted Moving Average:

The Exponentially Weighted Moving Average (EWMA) is a quantitative or statistical measure used to model or describe a time series. The EWMA is widely used in finance, the main applications being technical analysis and volatility modeling.

The moving average is designed as such that older observations are given lower weights. The weights fall exponentially as the data point gets older – hence the name exponentially weighted.

Formula:

$$EMA(today) = \left(Value(today) * \left(\frac{Smoothing}{1 + Days}\right)\right) + EMA(yesterday) * \left(1 - \left(\frac{Smoothing}{1 + Days}\right)\right)$$

Volatility can be estimated using the EWMA by following the process:

1. **Step 1**: Sort the closing process in descending order of dates, i.e., from the current to the oldest price.
2. **Step 2**: If today is t, then the return on the day t-1 is calculated as (S_t / S_{t-1}) where S_t is the price of day t.
3. **Step 3**: Calculate squared returns by squaring the returns computed in the previous step.

4. **Step 4**: Select the EWMA parameter alpha. For volatility modeling, the value of alpha is 0.8 or greater. The weights are given by a simple procedure. The first weight (1 – a); is the weights that follow are given by a * Previous Weight.
5. **Step 5**: Multiply the squared returns in step 3 to the corresponding weights computed in step 4. Sum the above product to get the EWMA variance.

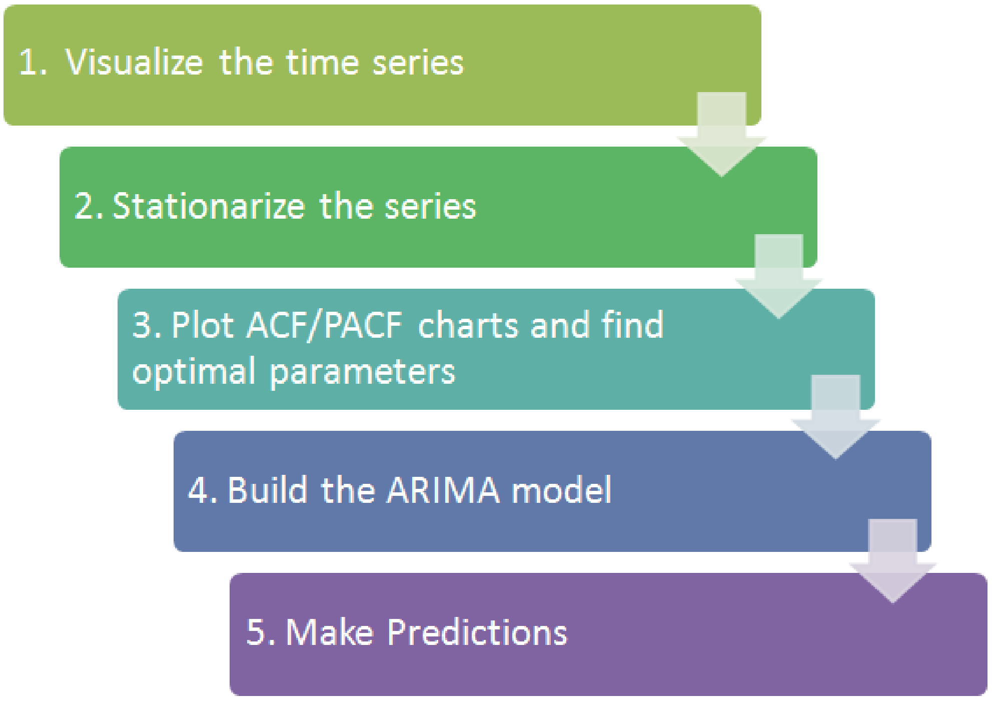
ARIMA (Autoregressive Integrated Moving Average)

An autoregressive integrated moving average, or ARIMA, is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends.

A statistical model is autoregressive if it predicts future values based on past values. For example, an ARIMA model might seek to predict a stock's future prices based on its past performance or forecast a company's earnings based on past periods.

KEY TAKEAWAYS:

- Autoregressive integrated moving average (ARIMA) models predict future values based on past values.
- ARIMA makes use of lagged moving averages to smooth time series data.
- They are widely used in technical analysis to forecast future security prices.
- Autoregressive models implicitly assume that the future will resemble the past.
- Therefore, they can prove inaccurate under certain market conditions, such as financial crises or periods of rapid technological change.



1. Visualize the time series

2. Stationarize the series

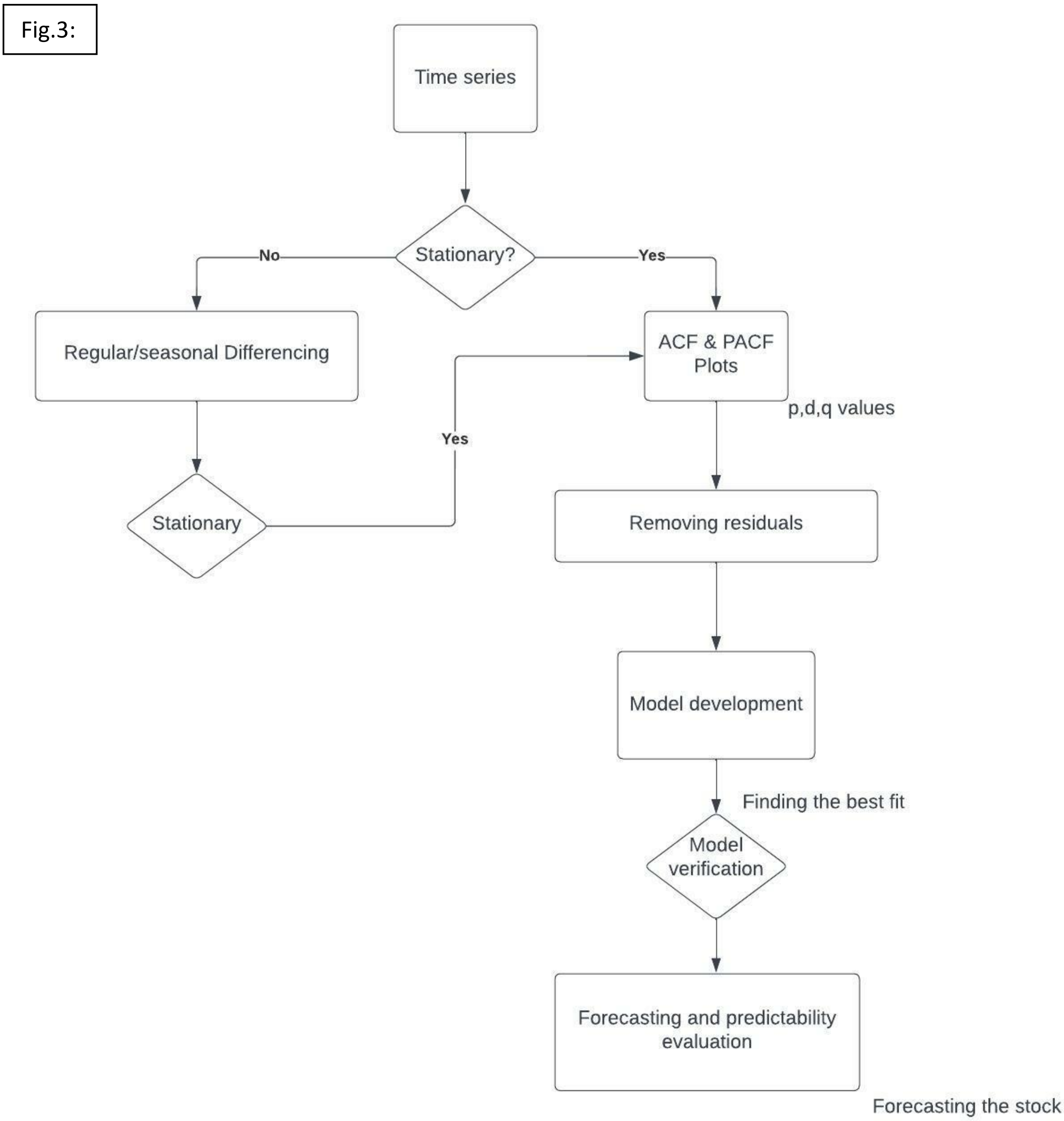
3. Plot ACF/PACF charts and find optimal parameters

4. Build the ARIMA model

5. Make Predictions

Fig.2:

Fig.3:


Time series
Stationary?
No
Yes
Regular/seasonal Differencing
ACF & PACF Plots
Yes
p,d,q values
Stationary
Removing residuals
Model development
Finding the best fit
Model verification
Forecasting and predictability evaluation
Forecasting the stock

Experimental Results and Discussion:

ARIMA MODEL:

To understand ARIMA model we'll first have to its components:

Autoregression (AR)*:* This model shows a changing variable that regresses on its own lagged, or prior values

Integrated (I): it represents the data values are replaced by the difference between the data values and the previous values.

Moving average (MA): Incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations.

The Parameters required for the ARIMA model:

- *p*: the number of lag observations in the model; also known as the lag order.
- *d*: the number of times that the raw observations are differenced; also known as the degree of differencing.
- q: the size of the moving average window; also known as the order of the moving average.

The following is the process:

1) Identify the problem statement.
2) Identify the data from different sources and acquire the relevant data.
3) Process and clean the raw data.
4) Perform the exploratory analysis.
5) Generate the model by dividing the data into training and testing data.
6) Train the training dataset with the respective data set and validate the model, apply the same on the testing dataset.
7) Visualize the results and check for the accuracy.

R Code:

i. Retrieving and cleaning the data:

```
#reading the entire dataset
df=read.csv("D:\\VIT\\SEM 5\\Data Analytics\\J component\\dataset\\all_stocks_2006-01-01_to_2018-01-01.csv\\all_stocks_2006-01-01_to_2018-01-01.csv")
df
```

There are another a total of 93613 records of differenct companies in this dataset. Out of which we are selecting 5 companies.

```
#taken five companies
c_googl<-data.frame(df[df$Name == "GOOGL",])
c_googl
C_amzn<-data.frame(df[df$Name == "AMZN",])
C_amzn
C_ibm<-data.frame(df[df$Name == "IBM",])
```

```
C_ibm
C_msft<-data.frame(df[df$Name == "MSFT",])
C_msft
c_nike<-data.frame(df[df$Name == "NKE",])
c_nike
```

ii. Data Overview:

```
#Google
summary(c_googl)
str(c_googl)
#Amazon
summary(C_amzn)
str(C_amzn)
#IBM
summary(C_ibm)
str(C_ibm)
#Microsoft
summary(C_msft)
str(C_msft)
#Nike
summary(c_nike)
str(c_nike)
```

iii. Data Cleaning

```
df[is.na(df)] <- 0
df
df$Date <- as.Date(df$Date, format = "%Y-%m-%d")
summary(df)
str(df)
```

Here so far we have taken a csv file which consists of different companies stock market values, we have selected only 5 companies such as google, amazon, Nike , IBM, Microsoft. We have cleaned the data, by removing any NA values or other discrepancies and separated each company and saved separate excel files, so that we'll have a set od datasets to work on our prediction.

STEPS WE INCORPORATED FOR THE ARIMA MODEL:

Step 1) installing all the packages and the libraries

```
install.packages('timeSeries')
install.packages('quantmod')
install.packages('tseries')
install.packages('forecast')
install.packages('xts')

library(timeDate)
library(timeSeries)
library(quantmod)
library(forecast)
library(TTR)
library(xts)
library(zoo)
```

Here the xts objects are matrix objects internally. xts objects are indexed by a formal time object. It stands for extensible time series and is an extension of zoo.

Step 2) We are converting the data which is in the form of data frame to xts to work on the data easily.

Note: that we are taking only the closing values of the companies and predicting the future values or the trends for the next couple of months.

Step 3) For the ARIMA model, we have two components which is **AR and MA models**. There are three parameters that is required to process the model **(p, d, q)** which can be obtained from the graphs:

ACF (complete auto-correlation function) it gives value of the autocorrelation of any series with lagged values it basically shows us how well the present values are related to the past values.This includes seasonality, trend, cyclic, and residual

PCF (partial auto-correlation function) it finds the correlations with the residuals (values that remain after removing the other effects).

ACF value gives the q value and PACF graph gives the p value.

We find the optimal number of features in an AR process using PACF because it removes variations explained by earlier lags.We find the optimal number of features in the MA process using ACF because the MA process doesn't have seasonal or trend components, so we only get the residual relationship with the lags in an ACF plot.

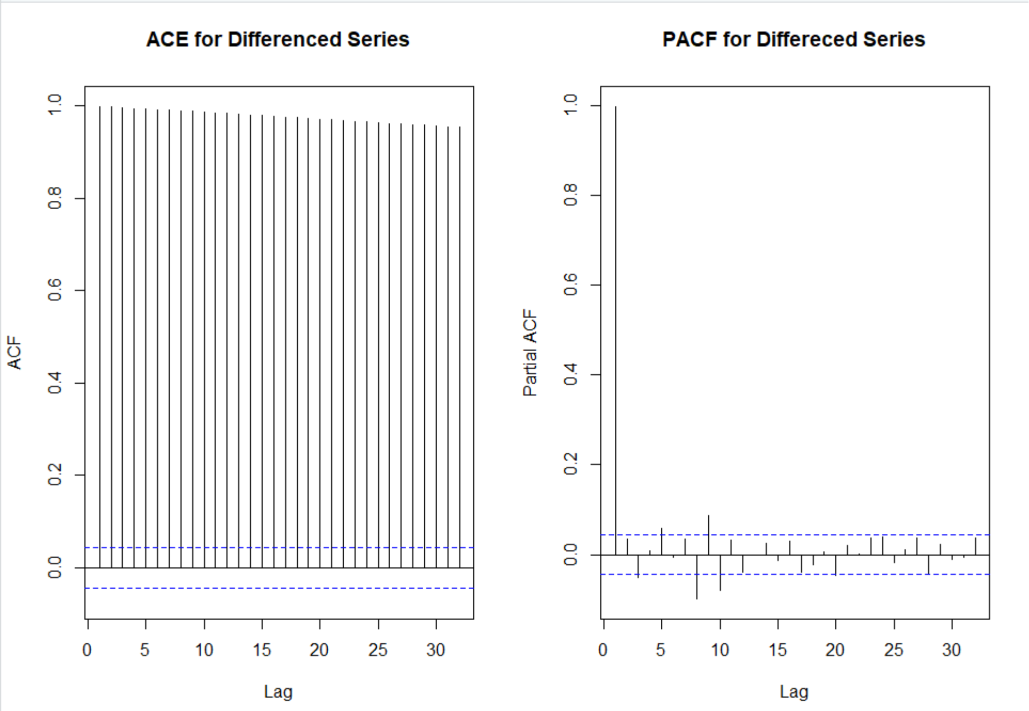


Fig.4

Figure 4 represents the ACF and the PACF plot for the whole dataset, here we observe that the first values is 1 above the blue line in both the graphs and it trends downwards in the ACF graph which gives the p value as 1.

Step 4) We also have an in built auto.arima() function in R which combines unti root tests, minimisation of the AICc and MLE to obtain an ARIMA model. It returns the best ARIMA model by searching over many models.

With a value of ARIMA(0,1,2)

Sigma^2	12.69
Log Likelihood	-5256.98
AIC	10519.95
AICc	10519
BIC	10536.69

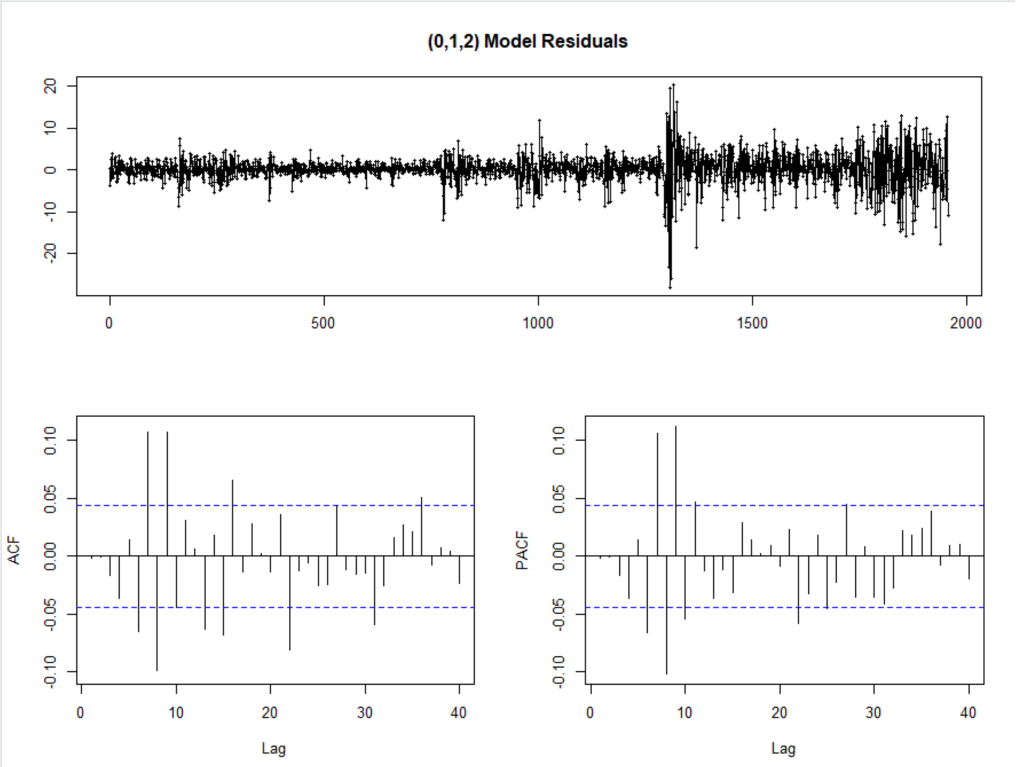


Fig.5

Figure 4 represent an auto.arima() function which gives the value of ARIMA(0,1,2) which are the respective (p, d, q) values and has been display in a plot.

Step 5) We are taking the log residual values to eliminate the areas that are non-s Stationary

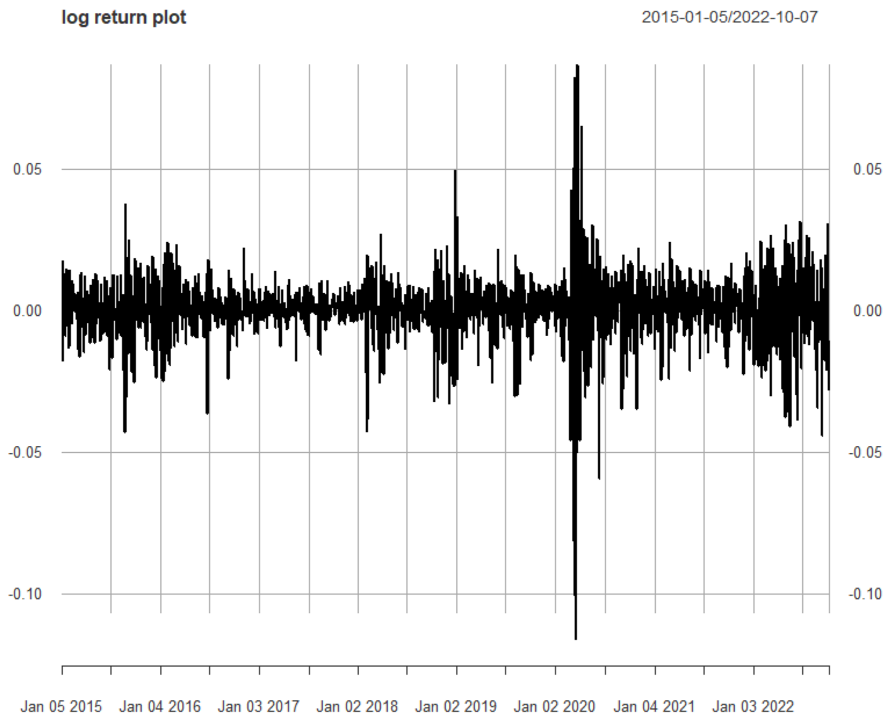


Fig.6

Figure 6 represents the log residual values after eliminating the areas that are non-stationary making the data more stable.

In an autoregressive integrated moving average model, the data are differenced in order to make it stationary. A model that shows stationarity is one that shows there is constancy to the data over time. Thus, the d value is 0.

Basically, we are making a non-stationary data into stationary. The residual values are the difference between the observed value – predicted value. Therefore, we are making prediction with the past data points that influence the future data points.

Step 6) ADF test to check whether the data is stationary or not.

A time series is said to be "stationary" if it has no trend, exhibits constant variance over time, and has a constant autocorrelation structure over time. One way to test whether a time series is stationary is to perform an augmented Dickey-Fuller test, which uses the following null and alternative hypotheses:

H0: The time series is non-stationary. In other words, it has some time-dependent structure and does not have constant variance over time.

HA: The time series is stationary.

Data	X
Dickey Fuller	-12.235
Lag order	12
p-value	0.01
Alternative hyposthesis	Stationary

Dickey fuller test for the log return values

If the p-value from the test is less than some significance level (e.g. α = .05), then we can reject the null hypothesis and conclude that the time series is stationary.

If the p-value is not less than .05, we fail to reject the null hypothesis. This means the time series is non-stationary. In other words, it has some time-dependent structure and does not have constant variance over time.

EWMA Model:

In first step we are reading the real time data of daily stock prices of from 12-08-2016 to 11-08-2017.

Next step we clean the data to remove the rows with empty variables and filter details of companies Google, Amazon, IBM, Microsoft and Nike.
the codes for the above two steps are as follows.

```
#reading the entire dataset

df=read.csv("all_stocks_1yrdata.csv")

#data cleaning

na.omit(df)

df

#taken five companies

c_googl<-subset(df,Name == "GOOGL")

C_amzn<-subset(df,Name == "AMZN")

C_ibm<-subset(df,Name == "IBM")

C_msft<-subset(df,Name == "MSFT")

c_nike<-subset(df,Name == "NKE")
```

For the next step we import the pracma and ggplot2 libraries, the pracma library is being installed be it contains the functions to compute the moving average and ggplot2 package is being installed to plot the required graphs

```
library(ggplot2)

library(pracma)
```

Next, we compute the EWMA average values of closing prices of the companies mentioned above using the movavg() function present in pracma package.

```
c_googl$EMAClose<-movavg(c_googl$Close,n=2,type='e')

C_amzn$EMAClose<-movavg(C_amzn$Close,n=2,type='e')

C_ibm$EMAClose<-movavg(C_ibm$Close,n=2,type='e')

C_msft$EMAClose<-movavg(C_msft$Close,n=2,type='e')

c_nike$EMAClose<-movavg(c_nike$Close,n=2,type='e')
```

Now, we will type the codes to plot the line graphs of EMAClose and Close variables in one graph for each of the companies.

```
ggplot(c_googl, aes(Date,group=1)) +geom_line(aes(y=Close),colour
="red")+geom_line(aes(y=EMAClose),colour ="green")+ggtitle("Actual(red colour) v/s Predicted(green colour)
Closing values for Google stocks")


ggplot(C_amzn, aes(Date,group=2)) +geom_line(aes(y=Close),colour
="red")+geom_line(aes(y=EMAClose),colour ="green")+ggtitle("Actual(red colour) v/s Predicted(green colour)
Closing values for Amazon stocks")


ggplot(C_ibm, aes(Date,group=3)) +geom_line(aes(y=Close),colour ="red")+geom_line(aes(y=EMAClose),colour
="green")+ggtitle("Actual(red colour) v/s Predicted(green colour) Closing values for IBM stocks")


ggplot(C_msft, aes(Date,group=4)) +geom_line(aes(y=Close),colour ="red")+geom_line(aes(y=EMAClose),colour
="green")+ggtitle("Actual(red colour) v/s Predicted(green colour) Closing values for Microsoft stocks")


ggplot(c_nike, aes(Date,group=5)) +geom_line(aes(y=Close),colour ="red")+geom_line(aes(y=EMAClose),colour
="green")+ggtitle("Actual(red colour) v/s Predicted(green colour) Closing values for Nike stocks")
```

STATISCAL ANALYSIS AND INTERPRETATION:

Here, we divide the whole dataset into two parts; training (80%) and test (20%)

Fig.7

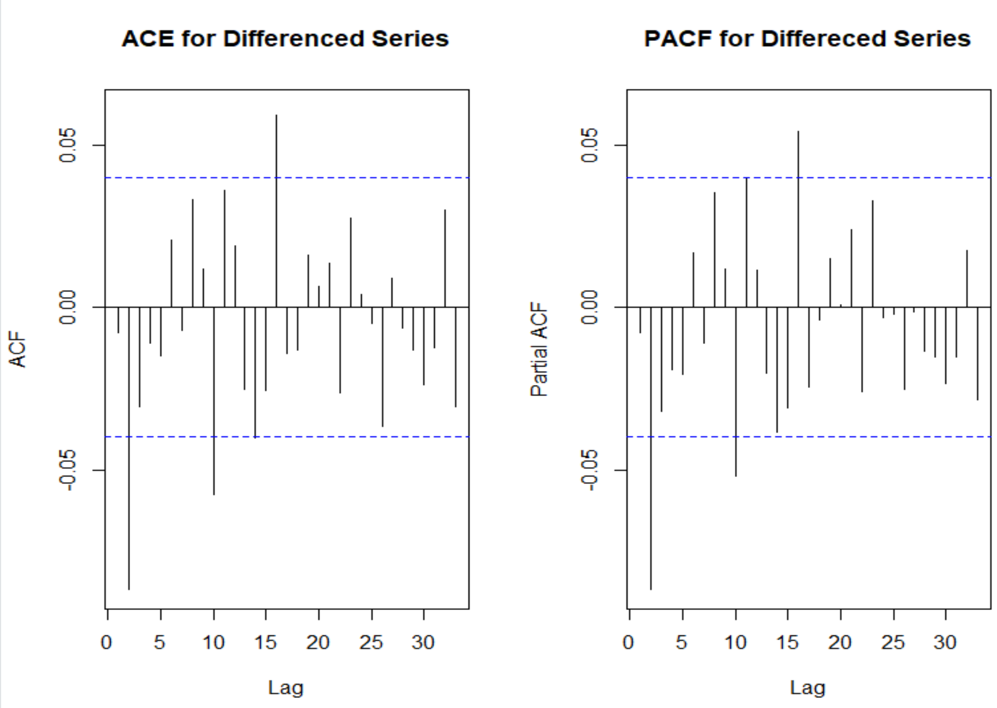


Figure 7 gives us the insights of ACF and the PACF graphs of the training data sets of the log return values.

ARIMA(0,0,3) for the training dataset

Sigma^2	0.00028817
Log Likelihood	6446.14
AIC	-12882.27
AICc	-12882.25
BIC	-12853.33

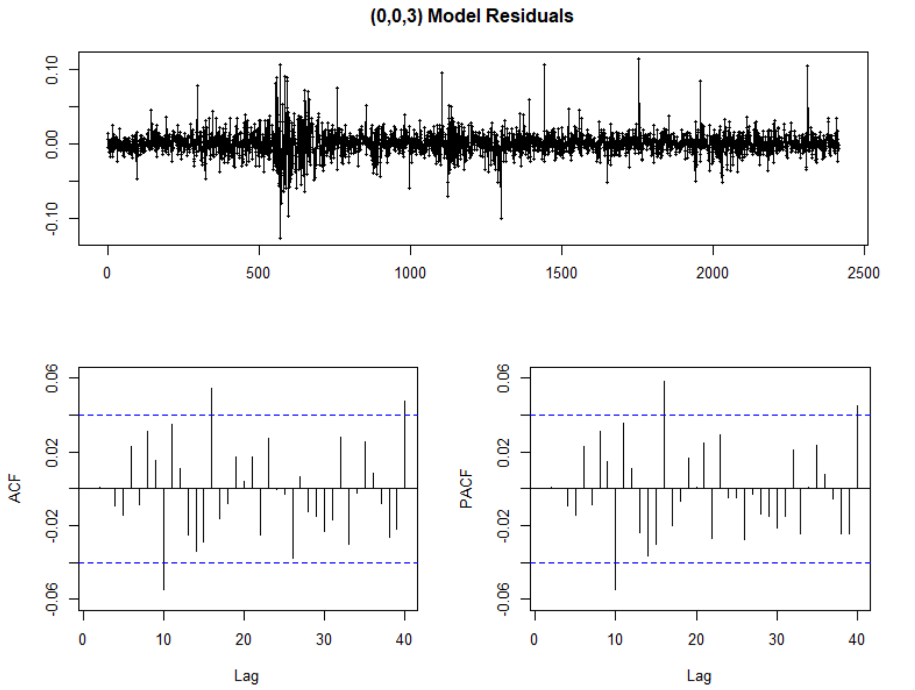


Fig.8

After trying to fit the best model with the function arima(), we can give the values of (p, d, q)

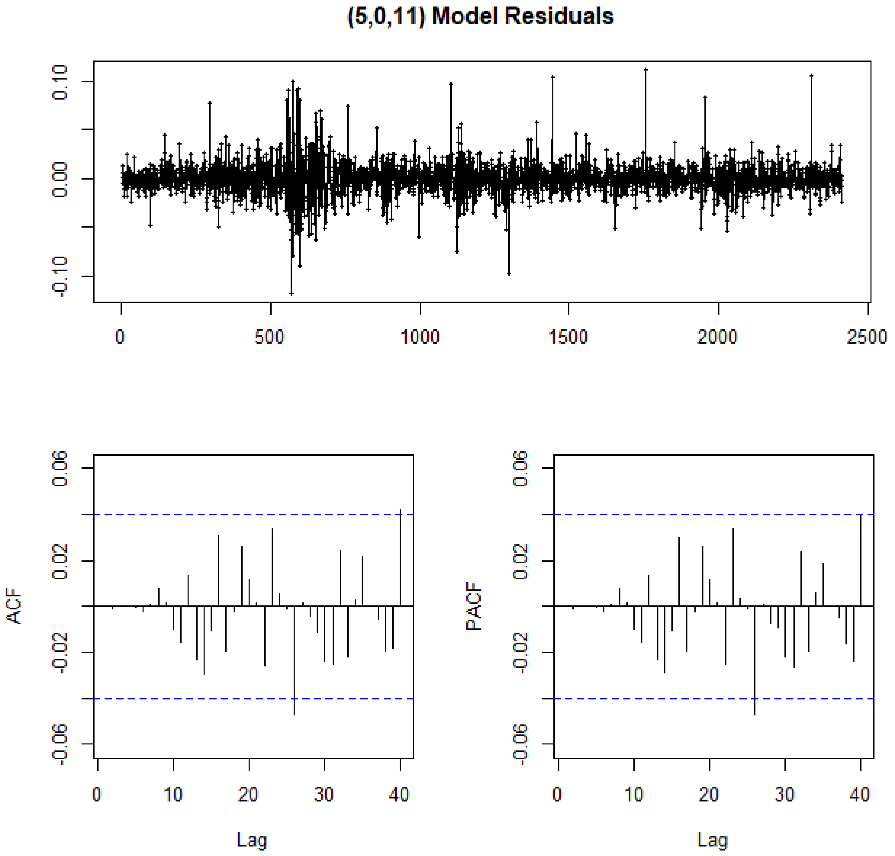


Fig.9

The lags are much less here which makes this a strong model.

Finally, we forecast the fits.

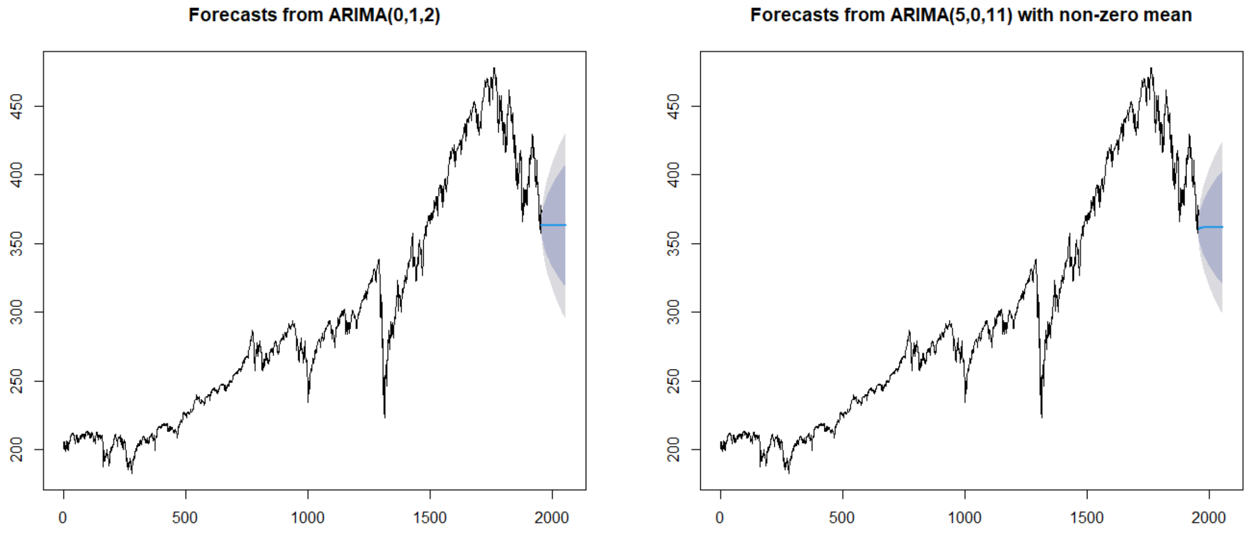


Fig.10

Followed by the accuracy test for the forecast, we have almost 99% of accuracy and we have selected the Fit 2.

Fit 1	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
	4.4291e-07	0.01677	0.011317	NaN	Inf	0.69235	0.0002119
Fit 2	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
	2.2770e-05	0.01668	0.011292	NaN	Inf	0.69082	0.0002118
Fit 3	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
	0.08272	3.56011	2.26237	0.022955	0.74932	0.99789	-0.001815
Fit 4	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
	0.086386	3.48403	2.26757	0.024358	0.75183	1.000193	0.004733

Here for the accuracy percentage, we take the value MAPE (Mean absolute percentage error) and subtract it with 100.

We follow the similar steps for all the 5 company's datasets

Google:

Fit 3	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
	0.02338	0.55998	0.37819	0.02778	1.12073	0.99832	-0.00695

Amazon:

Fit 4	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
	0.341779	6.688253	3.813708	0.067372	1.65551	1.003274	-0.003003

IBM:

Fit 1	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
	0.023656	1.861264	1.30356	0.0116001	0.936157	0.999689	0.00513

Microsoft:

Fit 1	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
	0.02338	0.5599881	0.378199	0.027781	1.12073	0.99832	-0.00169

Nike:

Fit 1	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
	-3.981e-05	0.501607	0.3141323	-0.042775	1.134891	0.999644	0.0167829

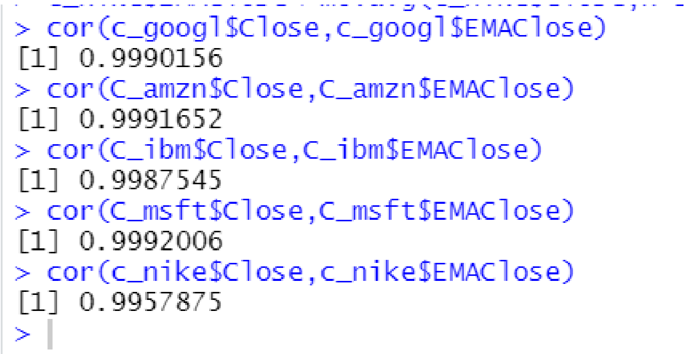
For statistical analysis, the correlation between the closing value of stock(Close column) and the predicted closing value(EMAClose) for each company was found.

If the value is closer to 1 then we can conclude than the model is correct in predicting the values.

Below is the code to check the above mentioned corelation

```
cor(c_googl$Close,c_googl$EMAClose)

cor(C_amzn$Close,C_amzn$EMAClose)

cor(C_ibm$Close,C_ibm$EMAClose)

cor(C_msft$Close,C_msft$EMAClose)

cor(c_nike$Close,c_nike$EMAClose)
```

The output of the code was as such



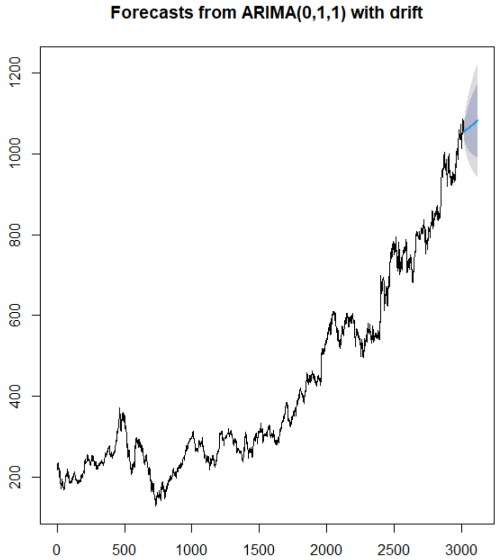
```
> cor(c_googl$Close,c_googl$EMAClose)
[1] 0.9990156
> cor(C_amzn$Close,C_amzn$EMAClose)
[1] 0.9991652
> cor(C_ibm$Close,C_ibm$EMAClose)
[1] 0.9987545
> cor(C_msft$Close,C_msft$EMAClose)
[1] 0.9992006
> cor(c_nike$Close,c_nike$EMAClose)
[1] 0.9957875
> 
```

Fig.11

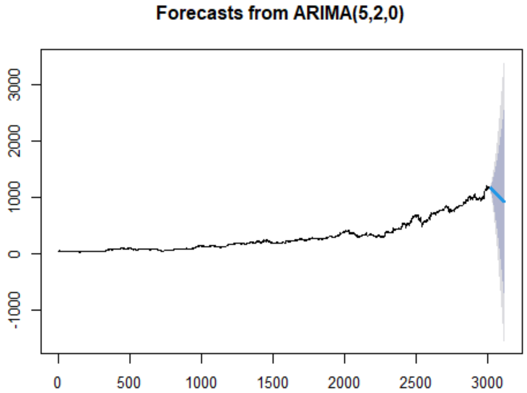
Thus, with value in each case being 0.99 we can conclude that EWMA model though may not be the model to exactly predict the outcome but is a model to give the prediction close to the actual values.

Visualization Analysis:

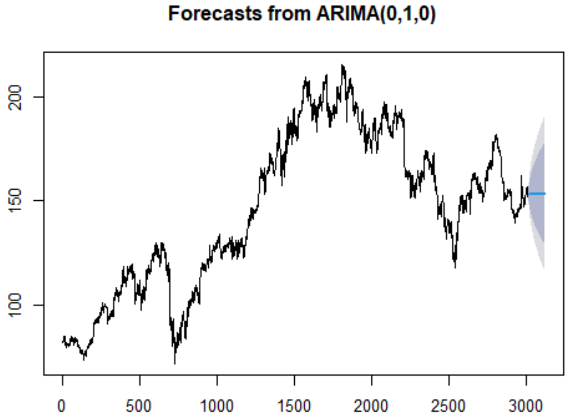
We follow the similar steps for all the 5 company's datasets



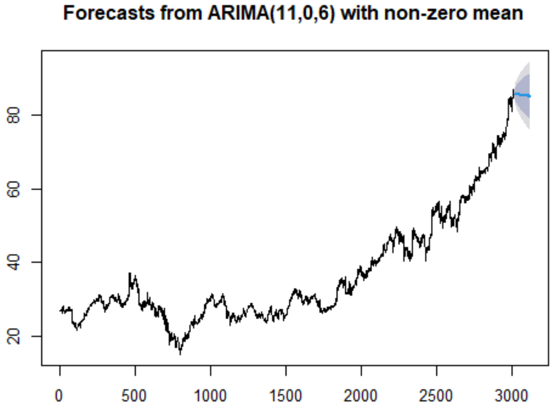
Google:10-15% upwards trend



Amazon: 20-30% downwards trend



IBM: remains the same



Microsoft: 5-7% downward trend

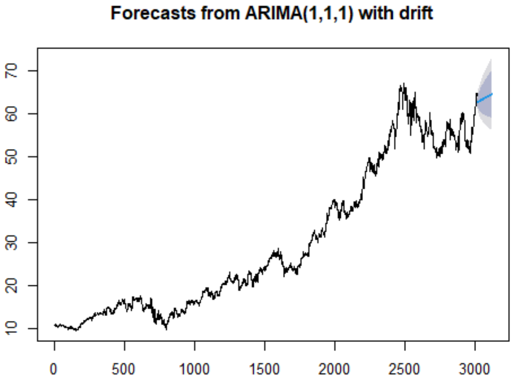


Fig.12

Nike: 10-13% upwards trend

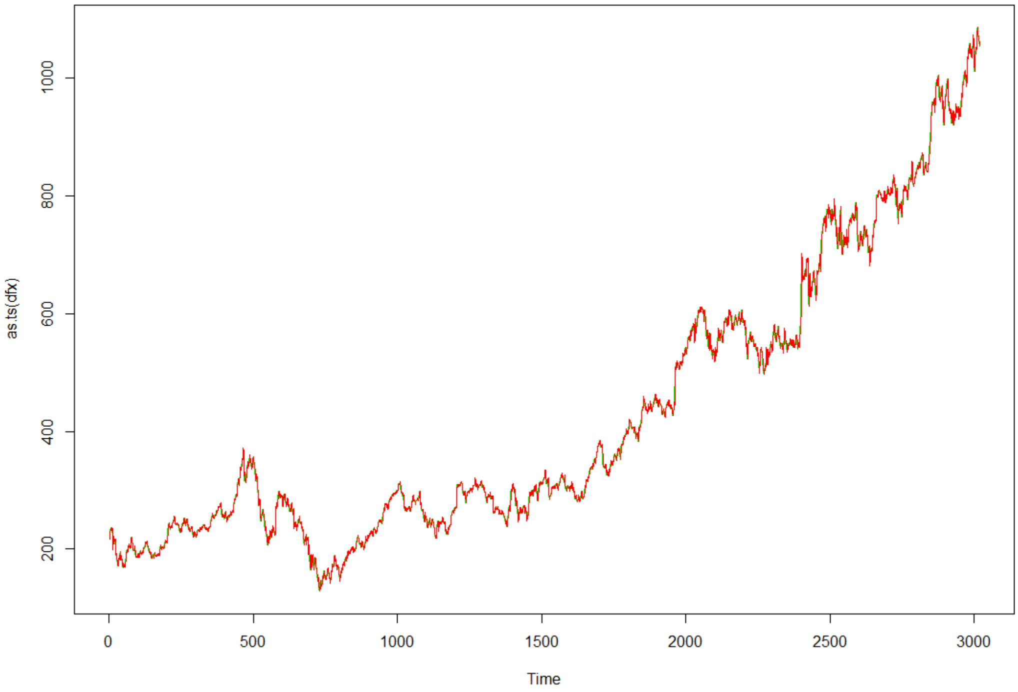


Fig.13

With a 98.78% of accuracy, here we have differentiated the predicted values and the actual values of the company Google.

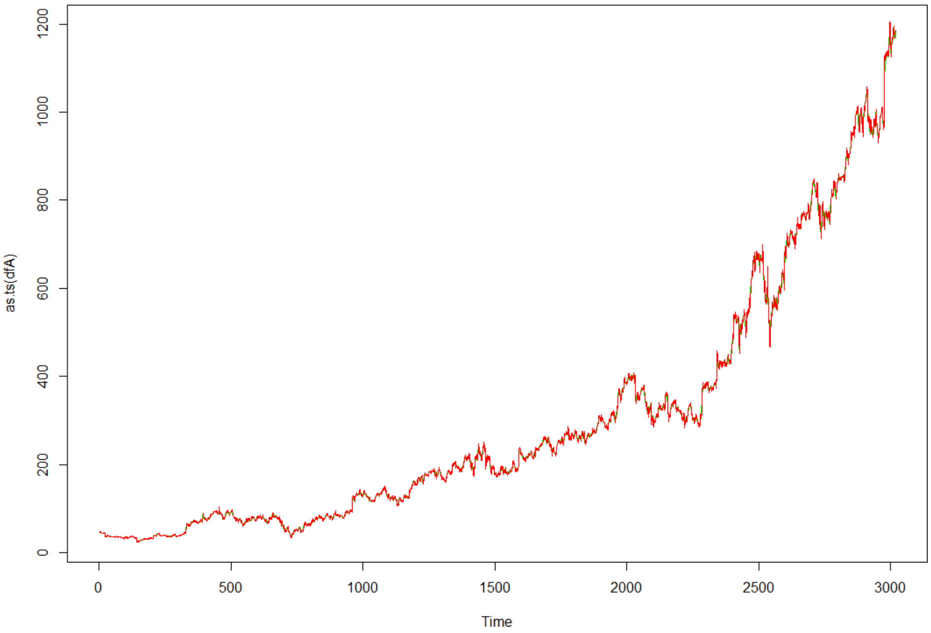


Fig.14

With a 98.34% of accuracy, here we have differentiated the predicted values and the actual values of the company Amazon.

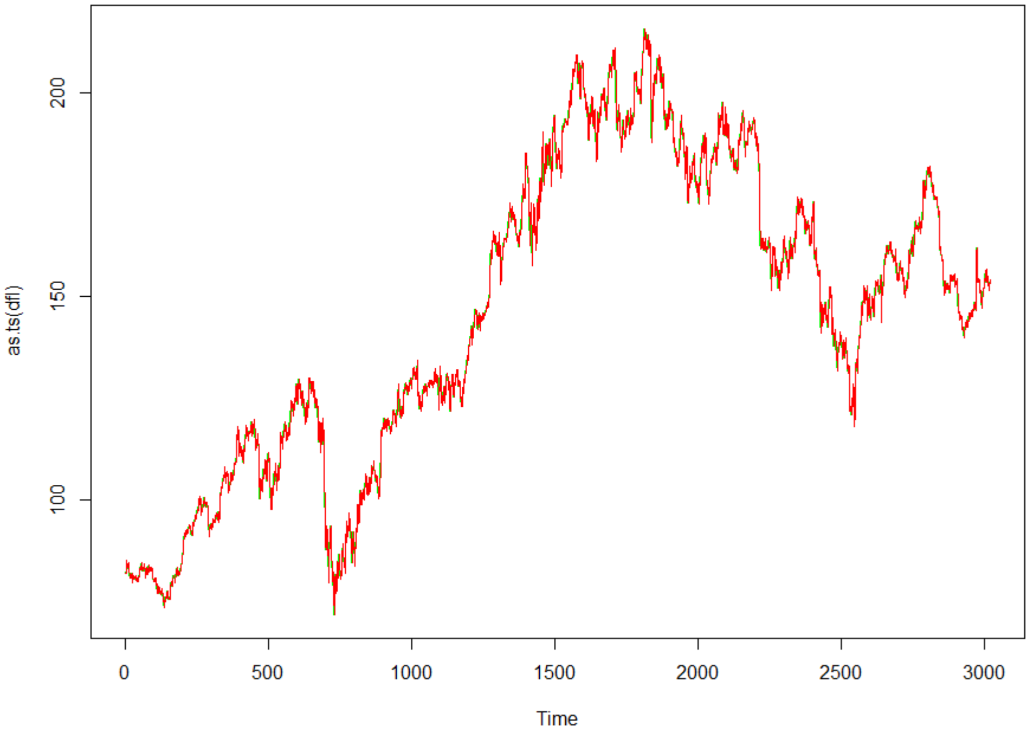


Fig.15

With a 99.07% of accuracy, here we have differentiated the predicted values and the actual values of the company IBM.

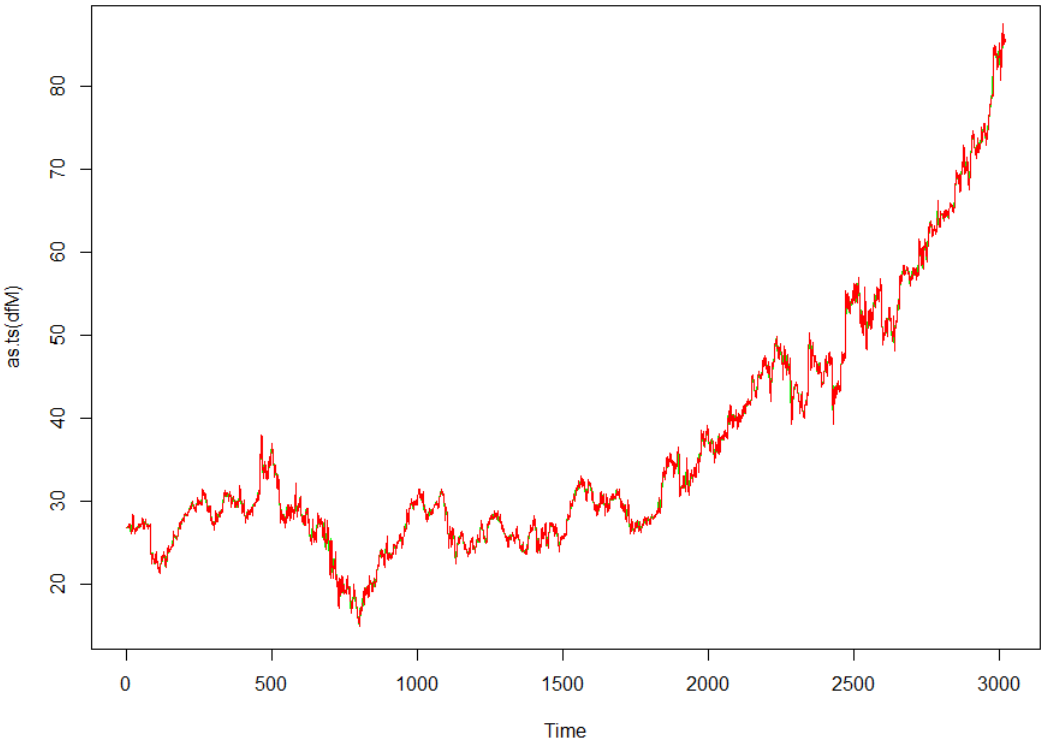


Fig.16

With a 98.88% of accuracy, here we have differentiated the predicted values and the actual values of the company Microsoft.

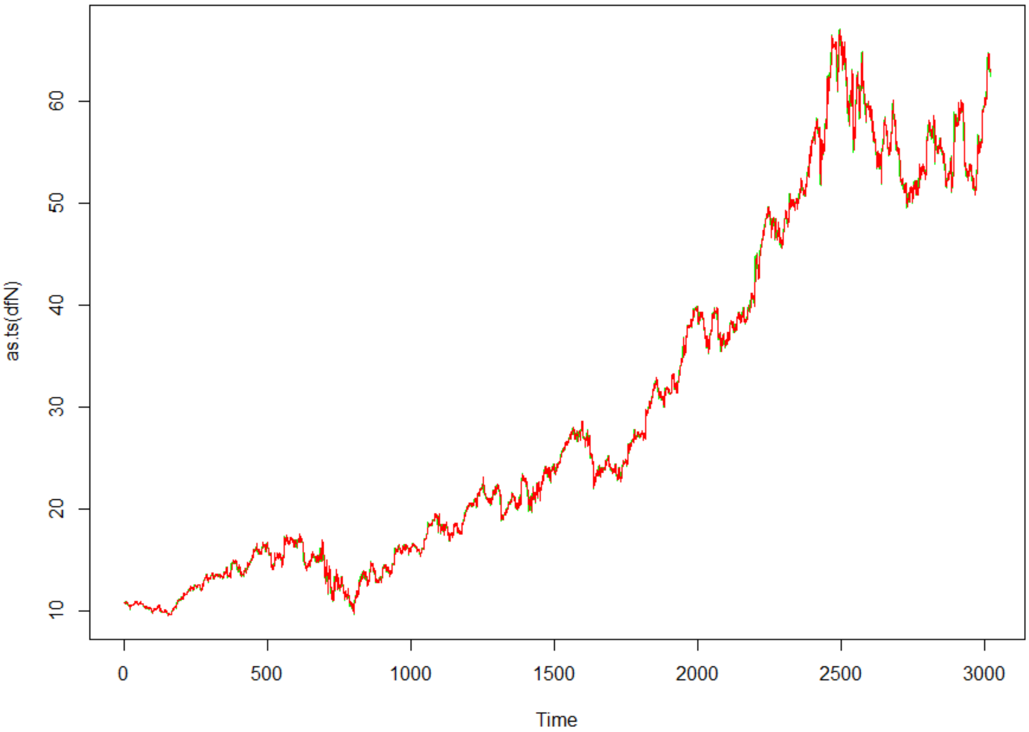
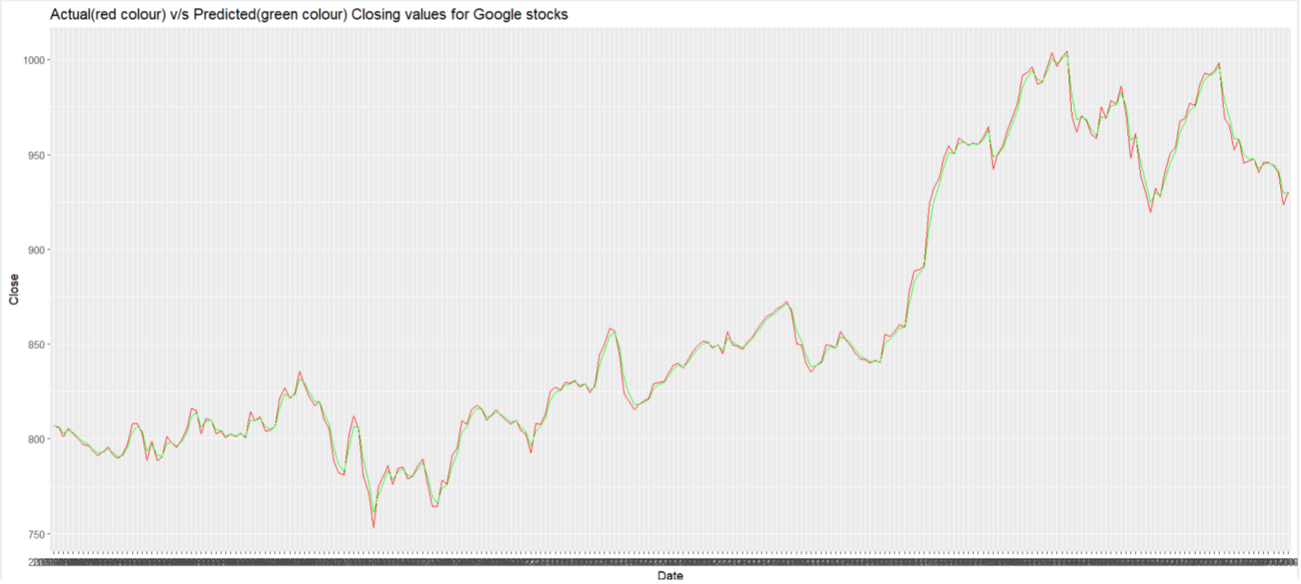


Fig.17

With a 98.76% of accuracy, here we have differentiated the predicted values and the actual values of the company Nike.

Below is the visualization analysis of the EWMA model
For Google

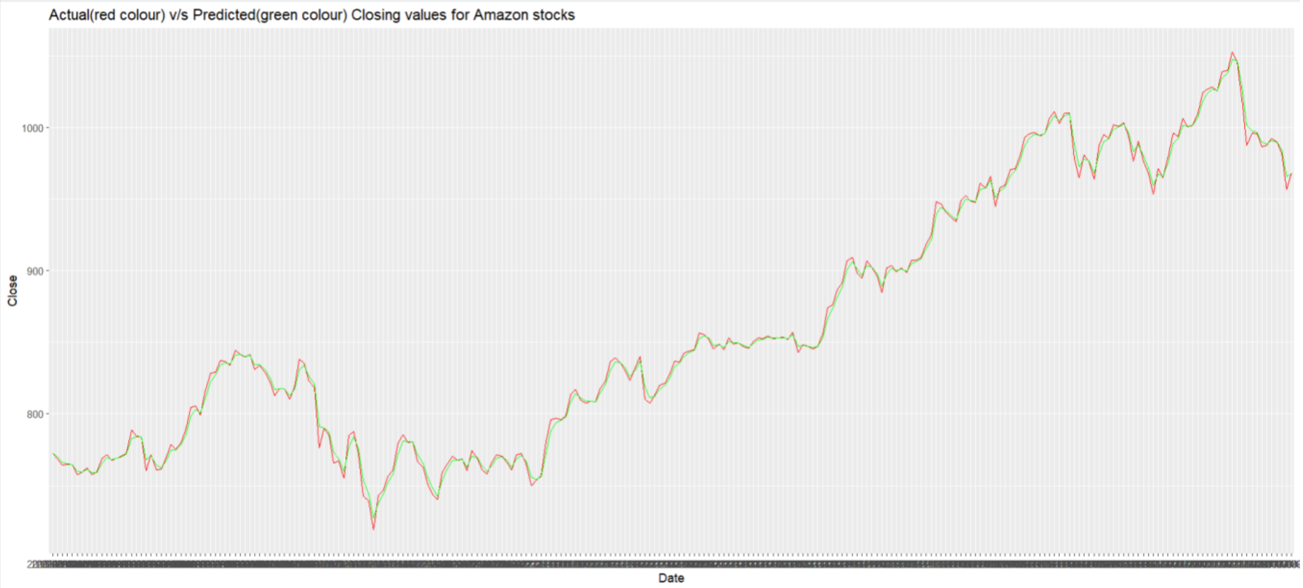
Fig.18



In the above figure we observe that the predicted values and almost equal to the actual values.

For Amazon

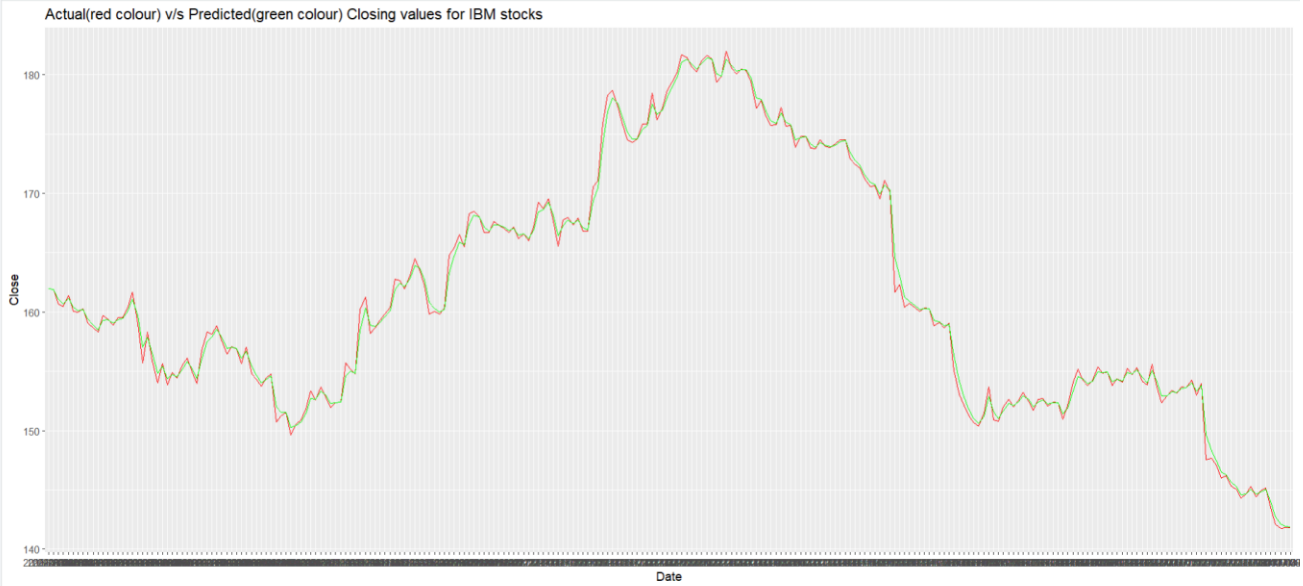
Fig.19



In the above figure we observe that the predicted values and almost equal to the actual values.

For IBM

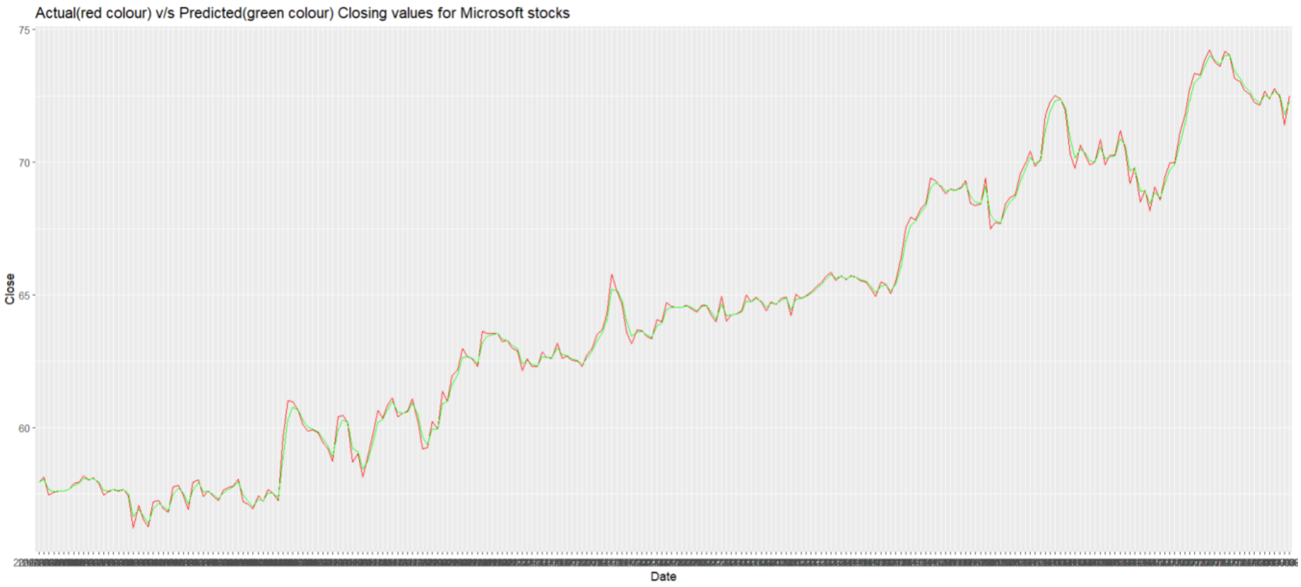
Fig.20



In the above figure we observe that the predicted values and almost equal to the actual values.

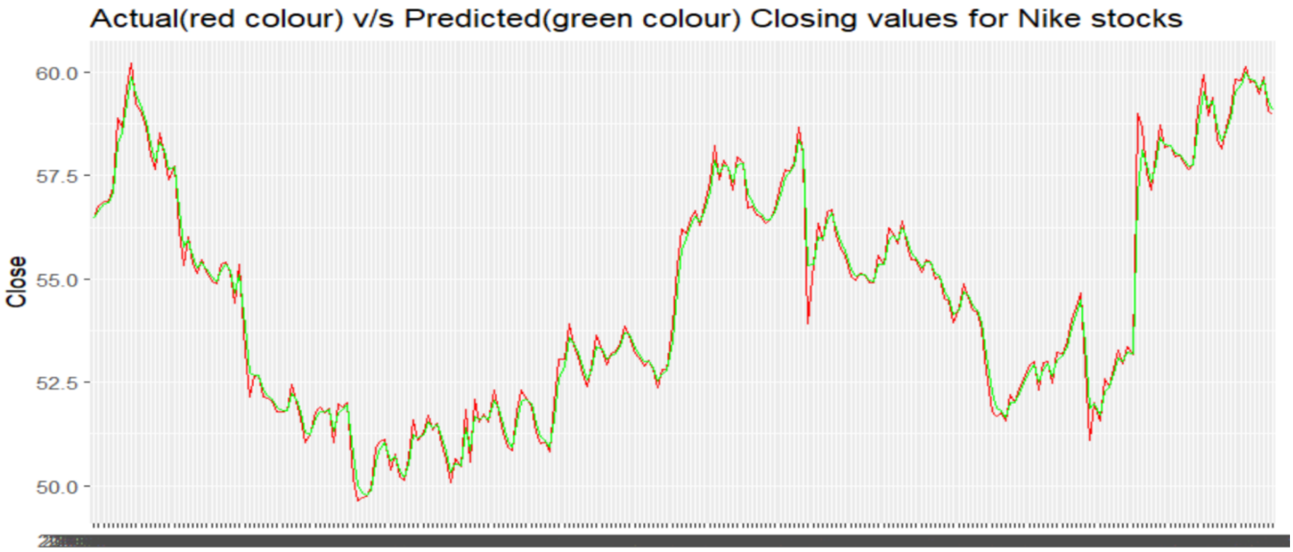
For Microsoft

Fig.21



For Nike

Fig.22



In the above figure we observe that the predicted values and almost equal to the actual values.

CONCLUSION:

On analysing the statistical and visual analysis of both the models with predicted closing values and the actual closing values of the companies Google, Amazon, IBM, Microsoft and Nike stocks we can conclude that the Auto Regressive Integrated Moving Average and Exponential Weighted Moving Average models are suited for making predictions for stock market prices.

REFERENCES:

Time Series Analysis | Time Series Modeling In R

Autoregressive Integrated Moving Average (ARIMA) Definition)

Proceedings of the 2022 7th International Conference on Social Sciences and Economic Development
Author: Ruihua Zhou, School of Mathematical Science, South China Normal University, Guangzhou, Guangdong Province, China, 510631

Research on Setting Voltage of Electrolyzer Based on LGBM-LSTM Algorithm
Author: Manshan Lin, Liangcai Ma

Time series Forecasting Using ARIMA model
Author:Ayat Ahmed Hamel, Baydaa Ismael

A Multivariate Control Chart for Monitoring Several Exponential Quality Characteristics Using EWMA By Nasrullah Khan;Muhammad Aslam;Mansour Sattam Aldosari;Chi-Hyuck Jun

Building Intelligent Moving Average-Based Stock Trading System Using Metaheuristic Algorithms

By Shu-Yu Kuo;Yao-Hsin Chou

Optimal Design of One-Sided Exponential EWMA Charts With Estimated Parameters Based on the Median Run Length By Yulong Qiao;Xuelong Hu;Jinsheng Sun;Qin Xu

Auto-Regressive Discrete Acquisition Points Transformation for Diffusion Weighted MRI Data

Emma Metcalfe-Smith;Emma M. Meeus;Jan Novak;Hamid Dehghani;Andrew C. Peet;Niloufar Zarinabad

A truncated SVD-based ARIMA model for multiple QoS prediction in mobile edge computing

By Chao Yan;Yankun Zhang;Weiyi Zhong;Can Zhang;Baogui Xin

Indonesian Stock Price Prediction including Covid19 Era Using Decision Tree Regression

By Kartika Maulida Hindrayani;Tresna Maulana Fahrudin;R. Prismahardi Aji;Eristya Maya Safitri