International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V68I11P202

Volume 68 Issue 11, 9-15, November 2020
© 2020 Seventh Sense Research Group®

Comparative Analysis of Companies Stock Price Prediction Using Time Series Algorithm

Lakshmana Phaneendra Maguluri[#], R. Ragupathy*

[#,*] *Department of Computer Science and Engineering*
Annamalai University, Annamalai Nagar,
Chidambaram, Tamil Nadu 608002, INDIA

[#]phanendra51@gmail.com.
[*]cse_ragu@yahoo.com

Abstract — *As the Company's financial values change day-by-day with uncertainty, forecasting of the stock market prices is a challenging task. The motive to forecast the stock market prices is to ensure that the investor's gains profit in what they are investing in. One of the core areas in stock market prediction is time series forecasting by using machine learning algorithms. This paper uses one of the time series prediction models called Auto-Regressive Integrated Moving Average (ARIMA) for future price prediction and forecasting.*

Keywords — *Time Series Forecasting, Formatting, ARIMA, Stock Price Prediction, Trend.*

I. INTRODUCTION

Typically, through this paper, we would like to predict something in the future and it could be stock prices [1] or it could be sales or anything that needs to be predicted into the future that is when we use time series analysis. So as the name suggests it is forecasting and typically when we say predict it need not be into the future in machine learning and data analysis [2], when we talk about predicting we are not necessarily talking about the future but in time series analysis we typically predict the future, so we have some fast data and we want to predict the future that is when we perform time series analysis. A few daily business examples are that it could be daily stock price of the shares we talk about or it could be interest rates [3] weekly interest rates or sales figures of a company so these are some of the examples where we use time series data, we have historical data which is dependent on time and then based on that we create a model that will make sure to predict the future [4]. The data time interval of time series can be categorized as daily, weekly, or hourly. In advanced cases there is data sensor where the interval could be for every few milliseconds or microseconds as well [6].So, we can conclude this point that the size of the time intervals can vary but they are fixed. In the case of daily data, then the interval is fixed as daily. In the case where the data is an hourly data, the data is captured every year recursively [7]. So, when the time intervals are fixed, we can know what kind of data we are capturing by knowing the interval itself.

As the name suggests, time series data is basically a sequence of data that is recorded over a specific interval and based on past values. So if we want to make an analysis of time series fast data [8], we try to forecast a future and as the name suggests it is time series data which means one of its components are time dependent.

Time series data consists of primarily four components, one is the trend and next comes the seasonality and then cyclicity and at last the irregularity. Sometimes the irregularity is also referred as the random component [10]. Trend is overall change or the pattern of the data which means that the data may increase or decrease in the series over a period of time. Then we have the next component seasonality. Seasonality is, as the name suggests, is a short-term variation occurring due to the season factors. Coming to cyclist, it is somewhat similar to seasonality but here the duration between two cycles is much longer [11]. Irregularity is like random component of the time series data which occur due to unpredictable factors and do not repeat in particular pattern. These are the four components of time series data. There are some conditions where we cannot use time series analysis right. So, we cannot perform time series analysis with all kinds of data. So, what are the situations where we are cannot do time series analysis. There will be some data which is collected over a period of time but it's really not changing in terms of any factor [12]. So, it will not really make sense to perform any time series analysis on that data. Another possibility is that there is change in the data, but it is changing as per fixed functions like a sine wave or a cos wave. Again, time series analysis will not make any sense in this kind of situation because there is a definite pattern here. There is a definite function that the data is following so it will not make any sense to do a time series analysis on that kind of data [13]. One of the important benchmarks is that before performing time series analysis, the raw data must be stationary.

The Stock Market is one field where it can make huge amount of profits and losses and it keeps on changing every day [14]. But many businesspeople, investors, and common people invest their money in the stocks for higher returns knowing it is a risk. What if, we can predict stock market prices up to a maximum extent and forecast the next day prices? This can help the investors and traders who have no idea about the quality of the stock and end up in loss at the end of the day [15]. The term forecasting means a way of predicting something earlier before it happens based on the previous and present events. Many predictive models using machine learning are developed to forecast the market prices that help the investors and traders to buy a good quality of stock and to get an idea of the moving of market prices. Forecasting can be used in different fields like business, finance, politics, industry, and others.

The determination of stock market prices which are never linear and unstable which makes it a critical field in the financial economy to forecast market values of companies. Back in time, there are many models like the SVM which is abbreviated as Support Vector Machine [16], Sentiment Analysis, LSTM which is abbreviated as Long – Short Term Memory [17] and others, but among all the Artificial Neural Network's (ANNs) [18] is the most popular technique due to its capability to read and understand the patterns in the data which is available, later predict the unknown or the future values. Another area in machine learning which can be used to forecast market prices is the time series forecasting.

ARIMA model, which is also known as Auto-Regressive Integrated Moving Average is the model in the time series forecasting which we demonstrate in this paper. It is an example of statistics and it is considered as an efficient model in the time series forecasting. This model can be used for short-term forecasting and more efficient than the ANNs.

In the Section II, of this paper, it describes the Time Series Forecasting, the ARIMA model is described in Section III, the methodologies of the model are discussed in Section IV, Section V, deals with the experimental results obtained from the model and Section VI, this paper is concluded. (Size 10 & Normal)This document is a template.

II. TIME SERIES FORECASTING

This area of machine learning is important to predict the data which is dependent on time. This can be used in many fields like weather forecasting, stock market prediction, statistics, and a few others. It is mostly used in the financial field. It collects the data points of regular intervals and analyses them to predict future values. The time series is broken down into systematic and non - systematic components to select the forecast model.

Systematic – The time series which can be specifically modelled and have consistency in the data.

Non-Systematic – The time series which is not modelled directly with the data.

The main components of the time series are,

- Long term movements or Trend
- Short term movements – Seasonal and Cyclic
- Irregular movements.

A. Long term movements or Trend (T)

The pattern of the word signifies a trend. The secular tendency is, therefore, that dimension of time series, which offers data the general trend for a long time. It is a series of smooth, normal, and long-term motion. A secular trend can be used to research the same status steady growth for a certain company product or the decrease in a request for a specific article over many years. Population growth in an area over the decades is considered as an example of long-term movement.

B. Seasonal variation(S)

This comes under the category of short-term movement. We will find that the sale curve is not uniform throughout the year if the market framework is observed closely, off the clothes. Different patterns will be observed in different seasons. This will completely rely on the people who live in that locality. Every year, the revenue structure in this period is approximately the same as the year before. So, this component occurs uniformly and regularly. This transaction is routine and periodic. So, the aspect occurs regularly and uniformly. This variation in nature is intermittent and in character routine.

C. Cyclic variation(C)

The variation also falls within the short-term movement group. In addition to the seasonal variations, there is another type of fluctuation typically lasting over a year. The Business Cycle effects the fluctuation. Each business has four essential phases.

i) Growth
ii) Decline
iii) Depression
iv) Change or Recovery.

The time to come back from abundance is a full process. But the loop never shows normal periodicity. A cycle length may vary but the shift sequence should be constant, critically, and it is this aspect of regularity that helps us to investigate cyclical fluctuation.

D. Irregular movements (I)

These are completely unpredictable as the name suggests. The consequences of floods, droughts, famines, earthquakes, and so on are known as periodic variability. All variations are normal, excluding pattern, seasonal and cyclical variances. However, cyclical fluctuations can also sometimes be caused by natural calamities. Essentially, what we need to do is to isolate the variables that are responsible for the ups and downs, that is, we need to take care of the seasonality, the stationary and the interactions between the components before we can get a reasonable set of data that can be used to predict future trends and variations.

It is possible to combine all these elements in many ways. It is commonly thought, however, that they are multiplied or added, as shown in "Eq. (1)" and "Eq. (2)".

$$(t) = T(t) * C(t) * S(t) * I(t) \qquad (1)$$

$$X(t) = T(t) + C(t) + S(t) + I(t) \qquad (2)$$

Where,

X (t) = Time-series observation; T = Trend component; C = Cyclic component; S = Seasonal component; and I = Irregular component

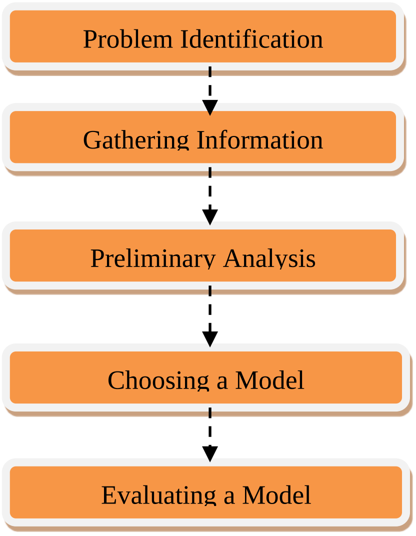


Fig 1: Steps in forecasting

The above-mentioned multiplication model assumes that all those elements of a time series are not inherently separate, interlinked, and mutually affecting. The additive model, however, accepts that all four elements are mutually independent. There are 5 steps to be followed in a time series forecasting problems shown in the below Fig.1

E. *Problem Identification*

The agenda is who requires this forecast and in what way it is needed to be done that can benefit them who use these predictions made by the model.

F. *Gathering Information*

In this step, the data required to analyse and model to forecast the predictions are gathered from several sources. This also includes getting access to domain experts and gathering statistics that can help to first-class interpret the historical statistics, and in the end the forecasts to be made.

G. *Preliminary Exploratory Analysis*

In this step, some of the tools like graphs, charts, and summary to reveal the trends and seasonality in the data to make it useful for the forecasting to have an accurate impact on the prediction. This also checks for the redundant data that may cause a disturbance in the forecasting.

H. *Choosing a model*

When all the data is gathered, this calls for a model of expectations that can offer the best forecast results. The choice of a model is to a great extent represented by the accessibility and nature of the information. Models are designed and fitted to the historical data.

I. *Evaluating the model*

When all the data is gathered, and the model is selected for forecasting the trained model and the validation of the data is done for the outcomes required. The data is trained and tested with many observations and forecasts the expected outcome.

III. ARIMA MODEL

ARIMA means Auto Regression Integrated Moving Average, which is a time series model for statistical analysis. In this, the time series can be trained and forecast future trends. It has three components in it (p, d, q). The standard form of this model is shown in “Eq. (3)”.

$$i_t = C + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + S_p i_{t-p} + \theta_1 \in_{t-1} + \theta_2 \in_{t-2} + \cdots + \theta_t \epsilon_t - q + \epsilon_t \quad (3)$$

This will capture complicated relationships because error states and lagged words findings. These models depend on a vector of previous values. Every term in ARIMA has its definition as stated below.

A. *AR (Auto Regression)*

This is a stochastic process and reveals a contingent relationship between the observations and their lagging values, that is, this regresses their prior values.

It can be referred as (p) in the ARIMA model. Time series data from previous years will influence current and future time series. Current and potential values are estimated in the ARIMA models when accounted for. AR(x) means x failure words lagging and this can be used in the ARIMA model. Auto-regression is a tool used to predict the variable on past values by it.

B. *Integrated (I)*

It is used to decrease the seasonality occurs when the data follows a certain pattern, and the time series is called stationary. It is referred with (d) in the ARIMA model which means the gaps in the number of findings. This can be put mathematically as shown in “Eq. (4)”.

$$d(a) = data(i) - data(i-1) \qquad (4)$$

C. *Moving Average (MA)*

It removes the random step in the time series and shows the amount of lagged error value. It is referred to as (q) in the ARIMA model. MA(x) where current observations are calculated using previous x observations. For example, if the value of q=1 then the error term exists in the data.

D. *Data Collection*

There are various websites from which you can download or fetch stock historical data or fundamental data. Few of them are Yahoo, Google, alpha vantage, quandl and other sites. Some are free to use and some are premium. In few of the sites you need to sign up and the rest you need not. Most commonly we use yahoo finance for getting Historical Data. For alpha vantage website, you will have to sign up and then it is completely free. You will get a free API key for quandl, you need to sign up and is very limited in its content. Eoddata-client is not a free to use website. You will have to pay for the service offered by them. The data collected in this paper relate to the stock market which continues to change with every second on the stock exchange's working days. The data is taken from the publicly accessible yahoo finance. The dataset contains the ‘opening price’, ‘closing price’, ‘highest price’, ‘lowest price’, and ‘volume’ for each date of the year. The historical data is gathered from various companies over three years from May 2016 to May 2020 through the model. The data is split into two categories: one is data from preparation, and the other is data from tests. The training data is used by the algorithm to train and understand the results, and the test data are used to verify the outcome algorithm.

E. Analysing the data

The time series must be stationary in terms of pattern and seasonality. If we use stable data that follows a certain trend and seasonality it may lead to incorrect predictions. So, the first thing to do after the collection of data will test whether the time series is set. If it is not stationary, we need to change it into stationary series by differencing the series or by finding the value of ‘p’. This needs to be done as the ARIMA model works on stationary data. The time series’ factor of stationary or not stationary is calculated by many methods such as ‘Auto-Correlation Function (ACF)’, ‘Partial auto-correlation Function (PACF)’, ‘W-D test ‘, ‘t-statistic test’, ’Augmented Dickey-Fuller test (ADF)’.

In this model, we perform a unit root test for stationary called ‘Augmented Dickey-Fuller test (ADF)’. It is the most common statistical test used to assess whether the time series is set. The difference between stationary and non-stationary should be understood to perform this test. As shown in Fig.2 and Fig.3 graphs are plotted for stationary and non-stationary series

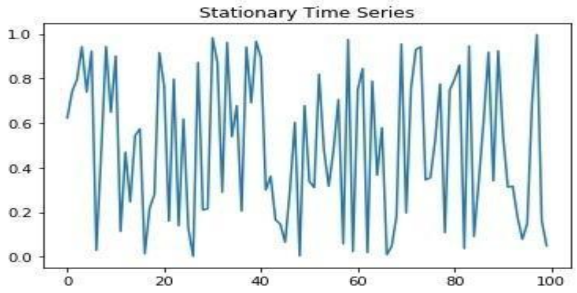


Fig 2: Stationary Time-Series

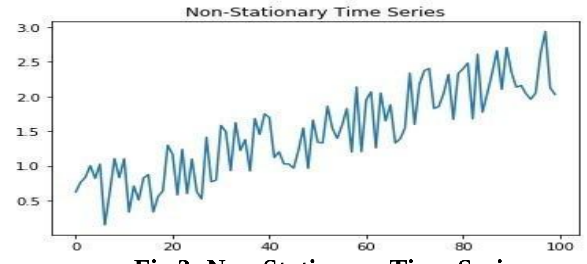


Fig 3: Non-Stationary Time-Series

ADF makes use of an autoregressive model and improves information standards across many different lag values. The two hypotheses named Null and Alternate Hypothesis are used in this. The unit-roots cause for the non-stationary of the data. To define this mathematically, when the value of alpha = 1 in the given “Eq. (5)” it is implied that in the time-series, the unit root remains.

$$Y_d = \alpha Y_{t-1} + \beta E_{xt} + \in \qquad (5)$$

F. PROCESS

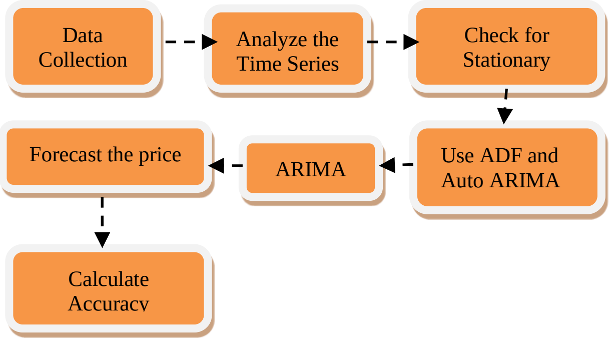
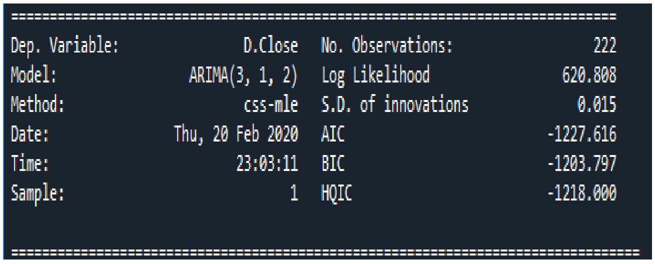


Fig 4: Workflow of the process

The completion of the first three steps i.e., finds whether is stationary or not we use the ARIMA model for the values of AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) and compares the values of both the lower and the higher the model used are. The rest of the process of the model is explained in ARIMA model earlier as shown in Fig.4.

Fig 5: ARIMA Results



```
==============================================================================
Dep. Variable:              D.Close   No. Observations:                  222
Model:               ARIMA(3, 1, 2)   Log Likelihood                 620.808
Method:                     css-mle   S.D. of innovations              0.015
Date:              Thu, 20 Feb 2020   AIC                          -1227.616
Time:                      23:03:11   BIC                          -1203.797
Sample:                           1   HQIC                         -1218.000

==============================================================================
```

From the Fig. 5 deals with the results of the values of p, q, and d and with the AIC and BIC values along with the number of observations used in the model.

IV. RESULTS AND DISCUSSIONS

The results have been executed with the model with the values of p, q, and d as ARIMA (3, 1, 2) the prediction of stock prices has been successfully predicted in form of the graph shown in Fig. 6 clearly explains about the statistics of Rolling Mean and Standard Devotion plots of INFOSYS, TCS, WIPRO and TECH M from the Fig.6, Fig.7, Fig.8 its clearly observed that due to this pandemic Situation the stock price of TECH M as observed downfall in share prices.

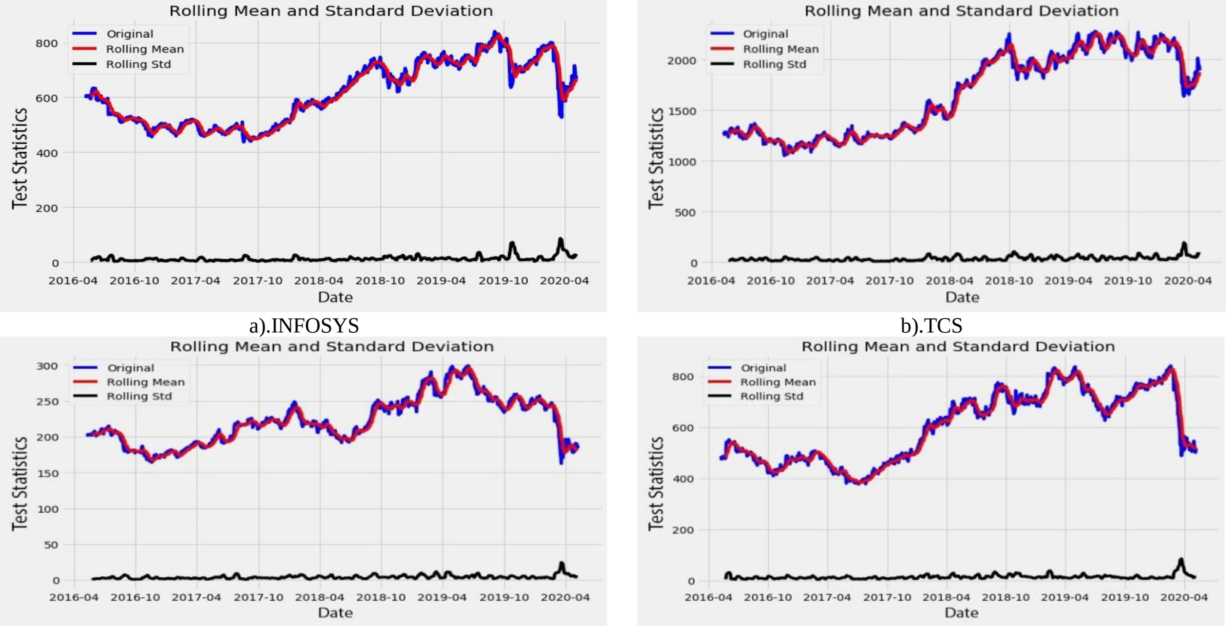
Table 1. Shows the results of Dickey Fuller Test of 4 companies

	INFOSYS	TCS	WIPRO	TECHM
Test Statistics	-1.3207	-1.0463	-1.5475	-1.6432
p-value	0.6196	0.7359	0.5098	0.4605
No. of lags used	8.0000	20.0000	14.0000	11.0000
No. of Observations Used	974.0000	962.0000	968.0000	971.0000
Critical Value (1%)	-3.4370	-3.4371	-3.4371	-3.4371
Critical Value (5%)	-2.8645	-2.8645489	-2.864530	-2.8645
Critical Value (10%)	-2.5683	-2.5683	-2.5383	-2.5683

From Table 1 and Table 2. it was compared with other 3 companies RMSE value for TECH M is high

Table 2. Shows the results of Error ratio of 4 companies

	INFOSYS	TCS	WIPRO	TECHM
MSE	0.0117	0.0128	0.0298	0.0536
MAE	0.0868	0.0952	0.1241	0.0136
RMSE	0.1085	0.1133	0.1726	0.2317
MAPE	0.0133	0.0126	0.0236	0.0259



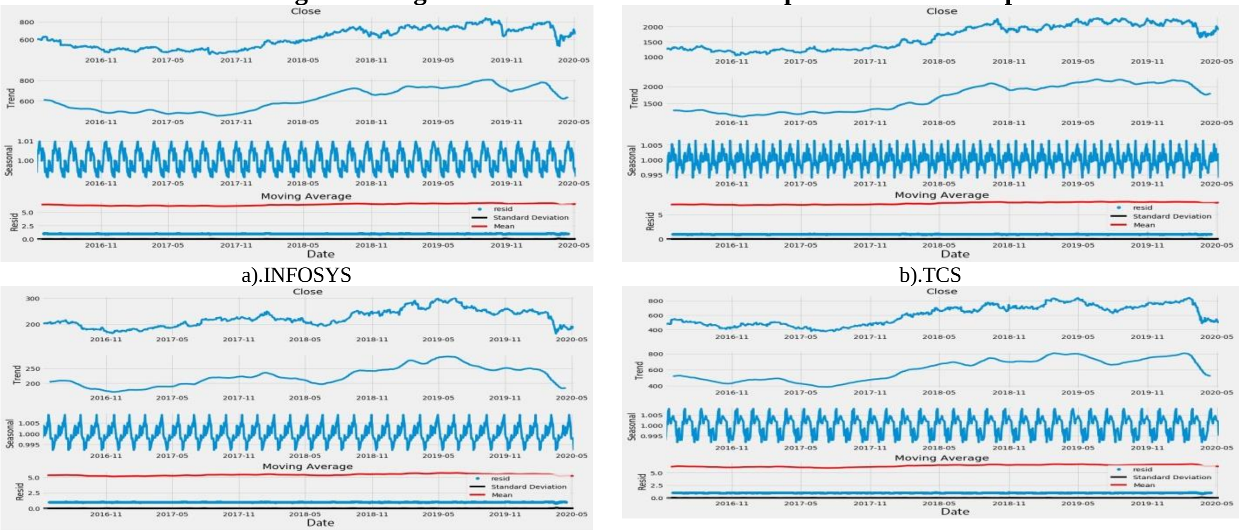
a).INFOSYS

b).TCS

c). WIPRO

d). TECH M

Fig 6: Rolling Mean and Standard Deviation plot of above 4 companies



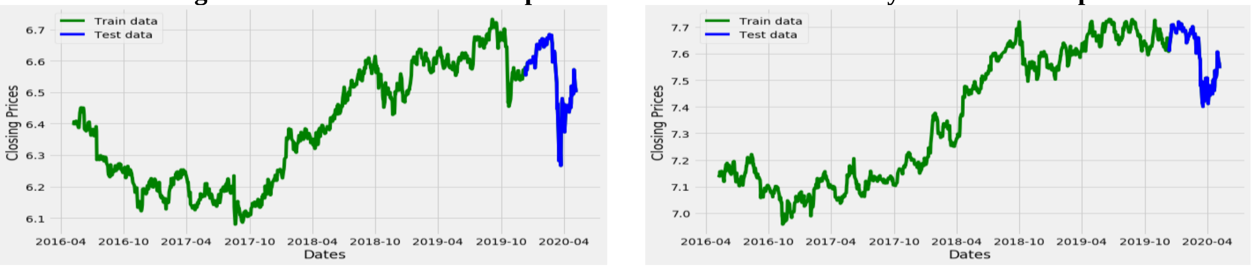
a).INFOSYS

b).TCS

c). WIPRO

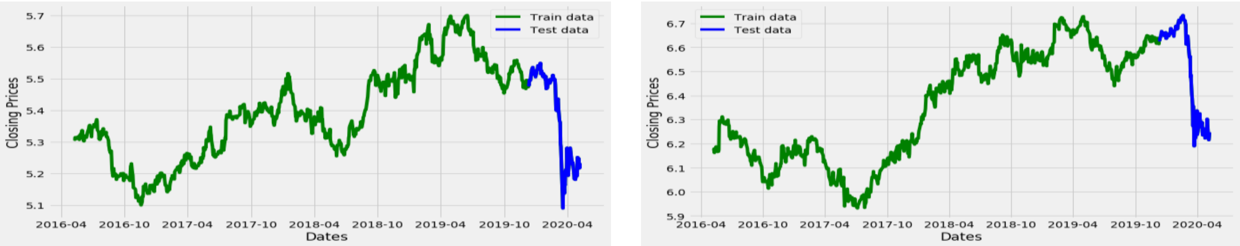
d). TECH M

Fig 7: Factors that affect stock prices like trends and seasonality of above 4 companies



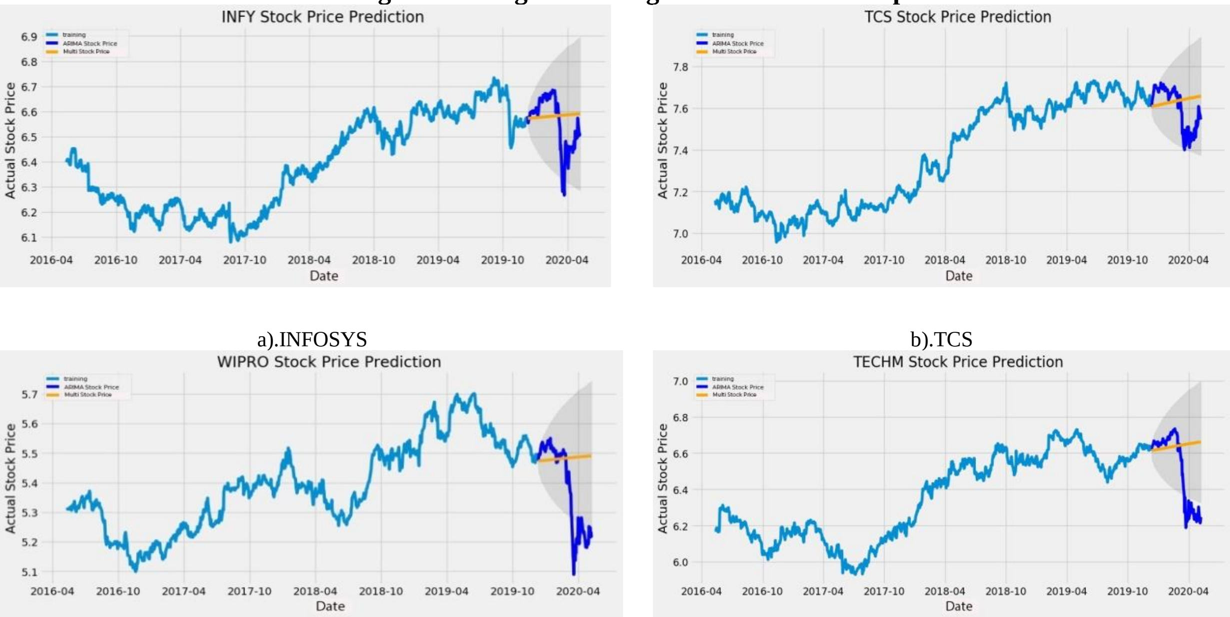
a).INFOSYS

b).TCS



c). WIPRO d). TECH M

Fig 8: Training and testing data of above 4 companies



a).INFOSYS b).TCS

c). WIPRO d). TECH M

Fig 9: Stock Price Prediction of above 4 companies

From the Fig.9, the blue coloured line indicates the stock prices prediction using ARIMA Model and the yellow coloured line indicates the predicted stock prices by using Multi Stock Trend Prediction Model.

V. CONCLUSION

This paper shows the results of a company by considering the uncertainties in the financial market as the values are varying over a particular interval of time. Compared with model implemented by L. P. Maguluri, and R. Ragupathy, this model is applicable only for a short period or short interval of time. In accuracy point of view, Multi Stock Trend Prediction Model produces 11% high accuracy when compared with ARIMA model for stock trend prediction.

REFERENCES

[1] B. Krollner, B. Vanstone, and G. Finnie., Financial Time Series Forecasting with Machine Learning Techniques: A Survey, in Proc. ESANN 2010. (2010) 1-7, 28-30.

[2] G. Bontempi, S. B. Taieb, and Y. Borgne., Machine Learning Strategies for Time Series Forecasting, in Proc. eBISS 2012: Business Intelligence,. (2013) 62-77.

[3] N. K. Ahmed, A. F. Atiya, N. El. Gayar, and H. El-Shishiny, An Empirical Comparison of Machine Learning Models for Time Series Forecasting, Journal Econometric Reviews, 29(5-6) (2010) 594-621.

[4] S.Panigrahi, and H.S. Behera., A study on leading machine learning techniques for high order fuzzy time series forecasting Engineering Applications of Artificial Intelligence Elsevier, 87 (2010) 1-10 .

[5] S.Kaushik, A.Choudhury, N.Dasgupta, S.Natarajan, L. A. Pickett, and V. Dutt., Ensemble of multi-headed machine learning architectures for time-series forecasting of healthcare expenditures, Algorithms for Intelligent Systems, Applications of Machine Learning. chapter.14 (2020) 199-216.

[6] J.A.Fischer, P.Pohl, and D. Ratz., A machine learning approach to univariate time series forecasting of quarterly earnings, Review of Quantitative Finance and Accounting. (2020) 1163-1179.

[7] D. S. De, J.F.L. de Oliveira, and P. S.G. de M. Neto., An intelligent hybridization of ARIMA with machine learning models for time series forecasting, Knowledge-Based Systems. 175 (2019) 72-86.

[8] B. M. Pavlyshenko., Machine-Learning Models for Sales Time Series Forecasting In Proceedings of the 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine. 4(10) (2018) 1-11.

[9] V. Cerqueira, L. Torgo, and C. Soares., Machine learning vs statistical methods for time series forecasting: Size matters, ,arXiv preprint arXiv:1909.13316. (2019).

[10] C. Fan, Y. Zhang,Y.Pan, X. Li, C.Zhang, R. Yuan, D.Wu, W.Wang, J. Pei, and H.Huan., Multi-Horizon Time Series Forecasting with Temporal Attention Learning, KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. (2019) 2527–2535.

[11] H. Shih, and S. Rajendran., Comparison of time series methods and machine learning algorithms for forecasting Taiwan Blood Services Foundation's blood supply, Journal of healthcare engineering. (2019) 1-6.

[12] G. Papacharalampous, H.Tyralis, and D. Koutsoyiannis., Univariate time series forecasting of temperature and precipitation with a focus on machine learning algorithms: A multiple-case study from Greece, Water Resources Management. 32 (2018) 5207–5239.

[13] C. Deb, F. Zhang, J. Yang, SE. Lee, and KW. Shah., A review on time series forecasting techniques for building energy consumption, Renewable and Sustainable Energy Reviews. 74 (2017) 902-924.

[14] L. P. Maguluri, and R. Ragupathy., A New Sentiment Score Based Improved Bayesian Networks For Real-Time Intraday Stock Trend Classification, International Journal of Advanced Trends in Computer Science and Engineering. 8(4) (2019) 1045-1055.

[15] L. P. Maguluri, and R. Ragupathy., An Efficient Stock Market Trend Prediction Using the Real-Time Stock Technical Data and Stock Social Media Data, International Journal of Intelligent Engineering and Systems. 13(4) (2020) 316-332.

[16] L. P. Maguluri, and R. Ragupathy., A Cluster Based Non-Linear Regression Framework for Periodic Multi-Stock Trend Prediction on Real Time Stock Market Data, International Journal of Advanced Computer Science and Applications. 11(9) (2020) 537-551.

[17] M. Syamala, and N. J. Nalini., A deep analysis on aspect based sentiment text classification approaches, International Journal of Advanced Trends in Computer Science and Engineering. 85 (2019) 1795-1801.

[18] M. Syamala, and N.J. Nalini., A filter based improved decision tree sentiment classification model for real-time amazon product review data, International Journal of Intelligent Engineering and Systems. 13(1) (2020) 191-201.