2021 6th International Conference for Convergence in Technology (I2CT)
Pune, India. Apr 02-04, 2021

Stock Price Prediction Using Facebook Prophet and Arima Models

Anusha Garlapati
Dept of Electronics and Communication Engineering,
Amrita Vishwa Vidyapeetham,
Amritapuri, India.
garlapatianusha@am.students.amrita.edu

Doredla Radha Krishna
Dept of Electronics and Communication Engineering,
Amrita Vishwa Vidyapeetham,
Amritapuri, India.
doredlakrishna4@am.students.amrita.edu

Kavya Garlapati
Dept of Electronics and Communication Engineering,
Amrita Vishwa Vidyapeetham,
Amritapuri, India.
garlapatikavya@am.students.amrita.edu

Nandigama mani srikara yaswanth
Dept of Electronics and Communication Engineering,
Amrita Vishwa Vidyapeetham,
Amritapuri, India.
nmsrikarayaswanth@am.students.amrita.edu

Udayagiri Rahul
Dept of Electronics and Communication Engineering,
Amrita Vishwa Vidyapeetham,
Amritapuri, India.
udayagirirahulam@am.students.amrita.edu

Gayathri Narayanan
Dept of Electronics and Communication Engineering,
Amrita Vishwa Vidyapeetham,
Amritapuri, India.
gayathrin@am.amrita.edu

2021 6th International Conference for Convergence in Technology (I2CT) | 978-1-7281-8876-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/I2CT51068.2021.9418057

***Abstract* - With the ongoing pandemics alike COVID-19 patronage such as stock markets, textiles become subsided. Stock market prognostication is the appearance of seeking to circumscribe the anticipated marketability of stocks and different financial apparatuses patronized on an exchange. Prophesying how the retail valuation will execute is one of the ultimate back-breaking contrivances to the predicament. Retail prediction is significant for quantitative interpreters and investment organizations. Retail valuation prophecy is pre-dominant for merit expenditure in the retail market. Analyzing the valuation interrelationship of duplet assets for the anticipated period of time is essential in portfolio optimization. The recommended explication is catholic as it comprises pre-processing of the retail advertise dataset, application of various exploratory analysis procedures, collaboration beside custom-built algorithms for retail valuation bias prognosis. In this case, Facebook Prophet and Arima models are used in forecasting the retail valuation of future stocks that are used to analyze future values of stock markets and how it varied from previous stock markets. With the circumstantial architecture and consideration of conjecture premises and data pre-processing techniques, this effort commits to retail estimate analysis.**

Keywords - Stocks prediction, Facebook Prophet, Arima, stock price correlation

I. INTRODUCTION

The productive market speculation insinuates that stock finances endure a percolate of erudition besides intellectual assumption and that recently exhibited details regarding the business landscape are practically right away a review in the prevailing stock valuation. Accordingly, fluctuations in stock markets speculate the propaganda of further information. There are many factors involved in the prediction of future stock prices Like rational factors, physical factors, irrational factors. All these perspectives consolidate to compose share prices buoyant and extremely obstinate to divine with a tremendous level of efficiency. In the epoch of big data, newfangled algorithms like ARIMA for forecasting stock advertise valuation and inclination has embellished notorious than aforetime. Traditionally moreover to analyze retail evolution, traders utilized to examine the retail valuation and stock measure in enhancement to the recognition associated with those stocks.

Commercial retail is a progressive and complex scheme where the public can trade coinage, commodities, hereditary over pragmatic principles sponsored by third-party people. The retail market permits traders to the inherent division of communal associations over trading likewise by interchange or over the conflicting details. The retail market is an identity about crucial provinces that traders are devoted to, thus the retail market valuation progression is perpetually a contemporary issue for investigators for a commercial and professional domain. This retail has stated stock-holders the possibility of acquiring finances as well as having affluent life via investing a crumb of finance escorted by stunted threat contrast before the fear of perforation of contemporary enterprise. Systematic prognostication of stock market yields is an exceedingly demanding responsibility owing to the expansive and scrappy attribute of the commercial stock exchange. Amidst the enlightenment of the Time series along with enhanced ciphering inclinations, techniques referring to forecast beget demonstrated to be further productive in forecasting stock valuation.

II. LITERATURE REVIEW

The stock price may depend upon numerous circumstances determining in the contemporary division and the stock market. There are mainly two factors namely:

1. Authority of stock prices on further companies alike how the accumulative of stock prices of further companies influence the stock price of a particular company.
2. bygone production together with documentation concerning stock price predictions of the company.

Here, various techniques are utilized to enhance the historical pattern of stock marketing trade and tell future results accordingly. The stock market is distinguished as vigorous, unforeseeable, and non-linear.

While an advocate of the well-organized retail postulation believes that it is impracticable to forecast stock market prices explicitly, the formal hypothesis is exhibiting the faultless pattern and design of applicable variables that may marshal to models make use of which stock price fluctuations ornamentation can be very precisely prognosticated. Here, we use ARIMA and Facebook prophet to inspect the behavior of future stocks that examines the pattern of stocks.

III. DATA SET

Each row in the dataset pertains to a unique individual. Here, the data set comprises 8 years of data from 2012-2020. The data set has entirely 1531 rows and 6 columns. columns are namely: Date, Open, High, Low, Close, Volume.

Date- This is an Index feature.

High- Denotes highest stock value of the day.

Low- The lowest value of the stock.

Open- Opening price value of that day.

The close- Closing price for that day.

Volume- Amount of stock traded on that day.

Here, we will contemplate the 'Close' price of each stock as the confer attribute of individual equip on our destination stock.

The structure of data including sample rows is as follows:

TABLE I. DATA SET

Date	Open	High
2012-01-03	325.25	332.83
2012-01-04	331.27	333.87

TABLE II. DATA SET

Low	Close	Volume
324.97	663.59	7,380,500
329.08	666.45	5,749,400

This is specimen data. Before performing any algorithm commentary on the data, we necessitate doing fascinating pre-processing measures.

DATA ANALYSIS:

One of the best predominant actions while dispensing with data is scraping data by executing data pre-processing impressions. This furnishes a more dependable representation in the accomplishment of the model. Data interpretation impersonates a major task to get more reliable performance results.

Decentralizing data furnishes a more immeasurable perception of data. If in case data consist of large numbers of columns and rows it is difficult to analyze and plot for every distinctive parameter distinctly. To overcome this, Plotly Express is used. This comprises gatherings that can generate an entire figure at once.

The X-axis is Year, Y-axis is columns in data.

The above figure is the pattern of data on how it looks whether increasing or decreasing trend. If we observe there is some missing data in close and volume columns that is the reason for not accepting the trend of stock prices there. To overcome this, data can be filled by interpolation methods or imputing data by using mean, median, mode.

Here, the mean is used to impute NAN values in data.

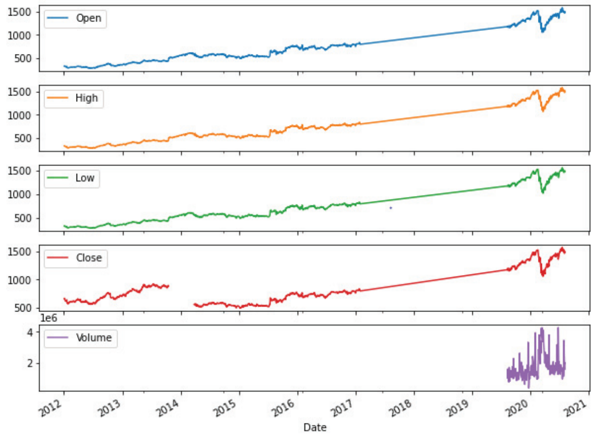

Fig. 1. Pattern of data

And here, the Volume column in data is standardized to get more dependable performance. This is done because the Volume column in data is so varied from others in terms of values, due to that we may not get better performances on data.

After imputing data with mean:

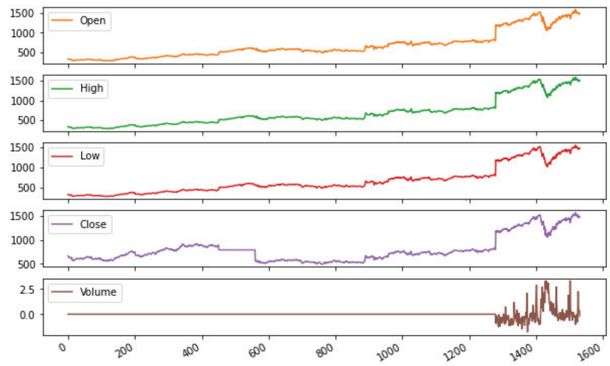

Fig. 2. Pattern of data after imputing with mean

Here, there are no missing values in the data that is shown in Figure2. This is just a pattern of data we have on how the values of stocks varied.

One more data analysis technique used here is the Slider widget. This helps to analyze data clearly as if observed in Figure1, Figure2 that is data from 2012-2020 if that need to particularly analyze a particular month in data it is difficult to analyze it. To overcome this Slider widget is used.

This is an area graph that permits an investor to select which section of the graph to see at a time. The widget consists of 2 related graphs, one positioned above the other. The bottom graph is the controller and contains a slider. The top graph is the primary graph.

Fig. 3. This is a Slider widget for 8 years of data.

The X-axis is Date and Y-axis is the Close column. Here, In the above Figure3, Data cannot be observed for a particular day for that by using the slider we can see.

Fig. 4. Data of 2 years after using a slider widget.

From Figure4 data of 2 years is shown. It is done by using the slider widget. By this, data can visualize on a particular day also.

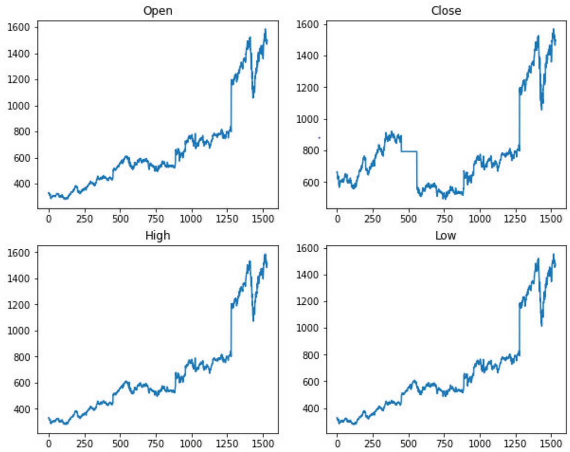

Fig. 5. Subplots of data

These are subplots of data that shows a trend or pattern of data for every column.

IV. METHODOLOGY

To get more reliable performance of ultimate forecasts of data some techniques and algorithms are used. Data can be better analyzed by using correlation plots, Moving averages, RMSE. Before performing the algorithm, techniques data must be checked for stationary test whether data is stationary or Non-stationary.

Correlation Plot:

Used to review the data interdependence between various variables. Helps us to recognize how each variable is conditioned on the other.

TABLE III. CORRELATION OUTPUT

	Open	Low	High	Close	Volume
Open	1.000	0.999	0.999	0.858	-0.027
Low	0.999	1.000	0.999	0.854	-0.037
High	0.999	0.999	1.000	0.859	-0.020
Close	0.858	0.859	0.859	1.000	-0.036
Volume	-0.027	-0.037	-0.020	-0.03	1.000

The above TABLE3 describes correlation values concerning every column, how each column relates to the other column. The table shows that almost all parameters are correlated with each other.

Heatmap for Correlation:

Graphical representation of data where values are depicted by color. It is obvious to envision complex data also.

It reveals how each variable is correlated to every individual variable. If the Correlation value is high between two variables then, it says that 2 variables are strongly correlated to each other.

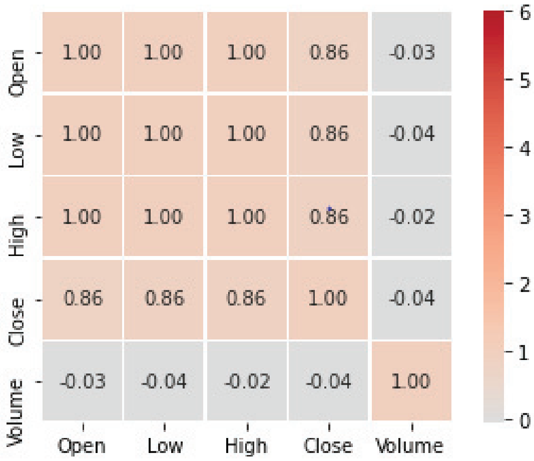

Fig. 6. Heatmap for correlated values.

Here, the above Figure6 says that almost all values are correlated with each other.

Moving Averages:

These are a simplistic and standard prototype of smoothing employed in time series analysis. It includes designing a new series where values are comprised of an aggregate of raw measurements in the initial time series. These are core foundational notions of time series. These make the baseline model for time series analysis. One of the core premises of moving averages is the time series is stationary and also it has slow ranging mean. If data has a lot of fluctuations it has a non-stationary mean and then moving average may not be the best forecasting method. If data has a constant mean, stationary data then moving average can be used. It can be used for data preparation, feature engineering, and even directions for making predictions. The below mentioned are kind of moving averages.

Simple Moving Average (SMA):

It estimates the midpoint of the last ‘n’ data points. It is a practical measure that can assist in influence if an asset price will continue or it will reverse a trend.

SMA: (t+(t-1) +(t-2) +…. (t-n))/n

It determines the proportion of an elected assortment of prices, regularly closing prices, by the number of periods in that range.

Weighted Moving Average (WMA):

It furnishes a weighted average of last ‘n’ data points where weight is assigned. It is a specialized pointer that attaches a considerable weighting to most maximum contemporary data points, and smaller weighting to data points in the outlying past.

WMA: (t*weighting factor) + ((t-1) * weighting factor-1) + …. + ((t-n) * weighting factor-n)/n

Prevailed by augmenting every number in the data set by pre-established weight and adding up emerging values.

Exponential Moving Average (EMA):

It is comparable to the weighted average but we don’t assign weights here. It takes a previous period time and going to calculate EMA and takes EMA as the next input rather than t-1 and t-2. It is a specialized chart pointer that pursues the price of an expense over time.

EMA: (Close – previous EMA) * (2/(span+1)) + previous EMA

It embraces more promptly data point changes. Here, Span is the duration we want to calculate.

Exponential Smoothing:

It expects a further parameter called ‘α’. α it is a smoothing factor. A larger value of α means the model is keeping attention to the most recent data points. Smaller α means taking history into count.

Yt+1: α [Y t+(1-α) Yt-1+(1-α2) Yt-2+(1-α3) Yt-3+…]

It is the rule of precept technique for smoothing time series data utilizing the exponential window function.

RMSE:

If RMSE is less then, the Error is less. Here, RMSE is adopted to estimate which moving average is best for data. Here, we examine the ‘Close’ column and applying RMSE for that.

RMSE: $\sqrt{\sum_{i=1}^{n}((\text{Y\^{}i} - \text{y i})2)/2}$

TABLE IV. RMSE RESULTS FOR MOVING AVERAGES

Moving Average	21.519
Weighted moving average	19.750
Exponentially weighted average	20.075
Exponential smoothing	17.664

From TABLE4 it shows that the Exponential smoothing average is the best moving average for this data set.

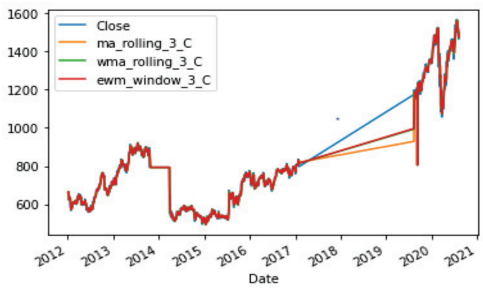

Fig. 7. Moving averages graph for Close column.

The above Figure7 shows which moving average is best.

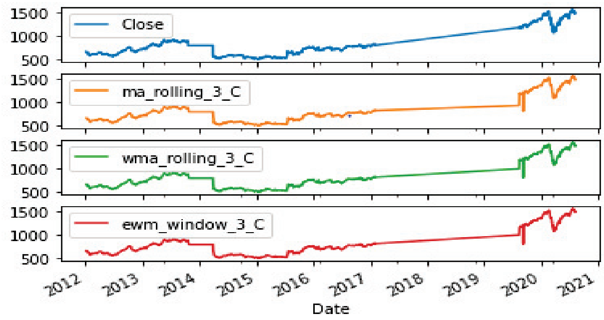

Fig. 8. Subplots showing moving averages.

One of the main important, the step before implementing algorithms to data is to check for Stationary test.

STATIONARY TEST:

There are 2 stationary tests to check whether data is stationary or not.

KPSS:

KPSS- Kwiatkowski-Phillips-Schmidt-Shin test, helps in examining if time series is stationary on their mean or on linear time. Measures KPSS test for the null hypothesis.

P-value must be greater than 0.05 for data to be stationary.

Here, P-value obtained for data is **P=0.01**

Null Hypothesis: Data is stationary

Alternate Hypothesis: Data is Non-stationary

ADF:

ADF- Augmented Dickey-Fuller, used to examine the essence of a unit root in time series, and hence in understanding if series is stationary or not.

P-value must be less than 0.05 for data to be stationary.

Here, P-value obtained for data is **P=0.98**

Null Hypothesis: Series possesses a unit root.

Alternate Hypothesis: Series is not stationary.

From the above analysis of P values Stationary test says that data is Non-Stationary.

If data is Non-stationary ARIMA is used.

Till now data is analyzed in different types like using moving averages and some plots. Now, Algorithms are applied to data.

ARIMA:

ARIMA, Auto-Regressive Integrated Moving Average

This is a familiar and broadly utilized analytical approach for time series prophecy. If data is non-stationary and has some trend component ARIMA is used. It has a further parameter called **‘I’.** This **‘I’** performs time series to be stationary out of non-stationary one.

It escorts the respective model envelope around them and escorts it as a simple interface. It is a kind of grid search. ARIMA has mainly 3 parameters namely **p, d, q.**

P – Kind of order of autoregressive part, correlation within itself.

D – Number of times that raw observations are contrast, this is considered as **'I' - Integrated** with ARIMA.

Q – Moving average parameter.

Model is developed on training data set by calling fit () function.

Finest ARIMA models have been picked by utilizing bases such as AIC, SIC, AME, RMSE, MAPE …. To determine which order of ARIMA is appropriate for series, there is a need to use AIC (or BIC) across a subset of values.

TABLE V. MODEL SUMMARY

Dep. Variable	**Y**	**No. of observations**	**1531**
Model	**SARIMAX (2, 1, 2) (1, 0, 0) [12]**	Log-Likelihood	-6597.981
Date	12-01-2021	**AIC**	13207.962
Sample	0 -1531	**BIC**	13239.961
Prob(Q)	0.32	**HQIC**	13219.871
Skew	4.85	**Kurtosis**	168.22

TABLE VI. MODEL SUMMARY

	Coef.	**Std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
ar. L1	-1.75	0.020	-89.8	0	-1.791	-1.7
ar. L2	-0.91	0.021	-44.4	0	-0.956	-0.8
ma. L1	1.67	0.027	61.7	0	1.625	1.73
ma. L2	0.83	0.028	29.5	0	0.781	0.89
ar. S. L 12	0.05	0.022	2.41	0.016	0.010	0.09
Sigma2	325.7	1.351	241.0	0	323.1	328.3

TABLE5 shows that the Best ARIMA model is ARIMA (2,1,2) (1,0,0) [12]. The above are some parameters of the ARIMA model.

Diagnostic Checking:

It is conceivable to possess various contrasting forecasting procedures for the related data set. Monitoring these resources is predominant to recognize whether a technique is utilizing outright obtainable information.

Here, we give a diagnostic plot for the residuals as follows:

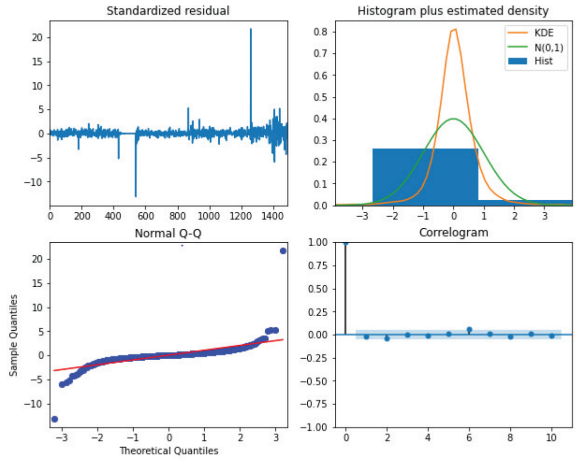

Fig. 9. Output Diagnostics

Here, above Figure9 it shows the Diagnostics output for the ARIMA model. The ideal correlogram for the residual should be flat which means all spikes should lie between a 95% confidence interval that means within a shaded region in the correlogram. Here the correlogram plot shows almost all points are within the shaded region. It lies between the 95% confidence interval region.

FACEBOOK PROPHET:

It is an algorithm developed by Facebook's Core Data Science team that is utilized in various applications of time series forecasting. It is very much utilized when there is a chance of seasonal effects.

Prophet was firstly developed to generate high-quality marketing forecasts. This tries to find points like:

Changes in trend due to various products.

Outliers

Seasonal effects like weekly, monthly, yearly cycles.

y(t): g(t) + s(t) + h(t) + ε*t

y(t) – Additive regression model

g(t) – trend

s(t) - seasonality

ε*t – Error

There are numerous techniques for time series forecasting. MAPE can be used for precise prophecy. This equips us with the expertise to make time-series predictions with good accuracy utilizing simple inherent parameters and has provision for incorporating the impression of inheritance seasonality and holidays.

Facebook Prophet is seeking to fit various linear and non-linear functions of time as elements. Modeling seasonality as an additive element is an identical strategy taken by exponential smoothing. This library is so significant that it has the potentiality of handling stationary within the data and also seasonality related elements.

But Facebook Prophet has some limitations like it presumes to be input columns with names 'ds' and 'y' where 'ds' is Date and 'y' is the target variable. Here, a trend can either be positive or negative and may be increasing or decreasing. The sample of this data is as follows:

TABLE VII. FACEBOOK PROPHET SAMPLE DATA

	ds	Y
0	2012-01-03	663.590
1	2012-01-04	666.450
2	2012-01-05	657.210

Here, ds is the Date column and Y is the target column. This is what Facebook Prophet requires to be for input commands ds and y. It does not take any other variable names other than these two. And then, the model is ready to fit for train data. After data is fit for training future predictions are done on data.

Data for future forecasting is as follows:

TABLE VIII. FUTURE DATA FRAME

ds	Y hat	Y hat_ lower	Y hat_ upper
2020-07-26	1338.352	1248.036	1430.649
2020-07-27	1404.407	1306.644	1494.335
2020-07-28	1406.271	1305.401	1494.738

This is sample data for future predictions.

Y hat- Predicted value

Y hat_ upper – upper confidence interval

Y hat_ lower – lower confidence interval

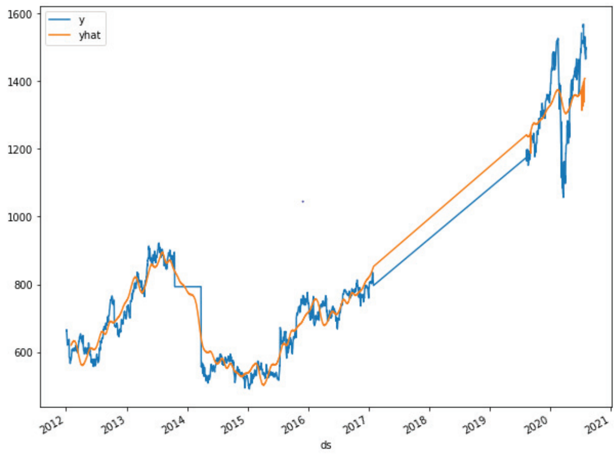

Fig. 10. Forecasting of future predictions

From Figure10, y hat is obtained based on y from previous years. Here, y is the actual prediction from data, and y hat is a future prediction based on previous data. Almost our model recognized good performance in predicting future forecasting of stock prices.

We can see it better by using a forecast plot using confidence interval also.

There is a parameter called interval _ width if that is set to 0.95 that indicates setting confidence level to 95%.

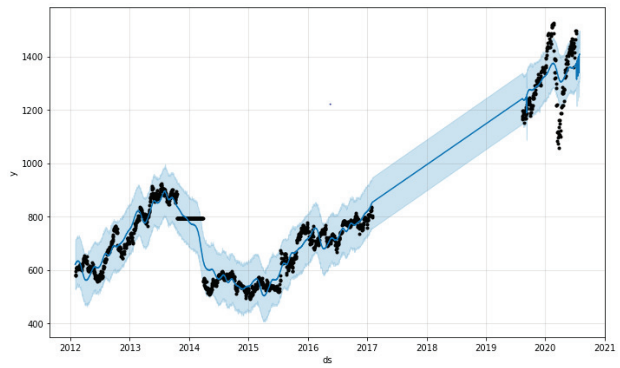

Fig. 11. Forecast future predictions

Here, X-axis is the Date and Y-axis is the target variable that is the Close column variable. Here, **Black** points indicate original data, **Blue** indicates Predicted values and **Light blue** indicates Confidence interval. Our model has done a good performance in the prediction of future stocks.

And this Figure11 can be more clearly shown by using change points.

Change Points:

These are unexpected fluctuations in time series data. Such modifications may render developments intervening variables. Detection of these change points is useful in modeling and predicting time series.

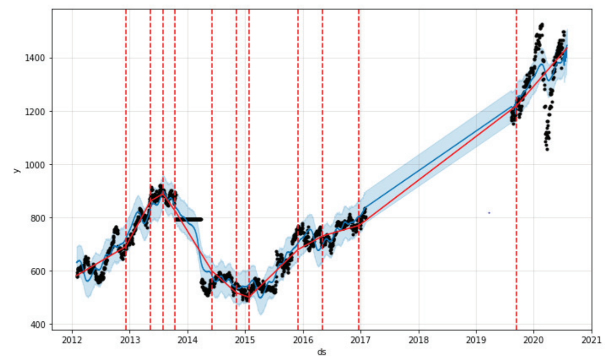

Fig. 12. Forecast future predictions using change points.

Here, It helps us to understand variations better in plotting actual and predicted values.

Some more parameters will be used to know how our model performed for predictions of future stocks. They are namely, MAPE, MSE, MAE, AME ….

MAPE: Mean Absolute Percent Error, Mean, of the quantity of outright deviation of forecasted and perceive value by dividing by perceive value is called Mean absolute error. For, illustration it is augmented by 100, which is called Mean Absolute Percent Error.

If an error is small, forecasting performance is good.

AME: Absolute mean error, Mean, of outright deviation of predicted and perceived values, is called AME.

This is utilized for the contrast of models in three periods.

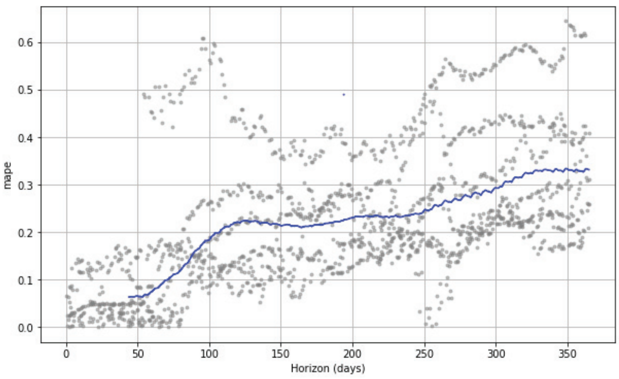

Fig. 13. Cross-validation metric using MAPE.

Here, X-axis is the Horizon of days that means if the horizon is set to 365 that indicates 50% of 365 days and the Y-axis is MAPE. It shows MAPE results concerning the horizon of days.

V. CONCLUSION

In this paper, we discussed stock market trends and analyzed different patterns of data, and done analysis for future forecasting of stock prices. For this commentary and foresight models like ARIMA and FACEBOOK PROPHET are used. To construct these models, data is deduced on Stock price predictions from 2012-2020.

This analysis has further prospective for investigation in the future.

MAPE is used as a parameter that shows that the models are adequate in forecasting retail valuation.

This empirical inquiry designated that the ARIMA (2,1,2) model is best for predicting Stock prices.

FACEBOOK PROPHET here is utilized to exhibit future forecasting of stock prices.

REFERENCES

[1] Ranjith Kumar, S. Performance divergence and financial distress of selected steel companies in India (2017) International Journal of Mechanical and Production Engineering Research and Development, 7 (5), pp. 381-392.

[2] Sri Rajini, S. An examination of the essential aspects, in the execution of enterprise resource planning in manufacturing and production industries (2017) International Journal of Mechanical and Production Engineering Research and Development, 7 (6), pp. 395-402. C

[3] Zhang, Y.; Wu, L. (2009). "Stock Market Prediction of S&P 500 via a combination of improved BCO Approach and BP Neural Network". Expert Systems with Applications. 36 (5): 8849–8854. doi:10.1016/j.eswa.2008.11.028.

[4] Mislinski, Jill (3 March 2020). "Market Cap to GDP: An Updated Look at the Buffett Valuation Indicator". www.advisorperspectives.com. Archived from the original on 14 March 2020. it is probably the best single measure of where valuations stand at any given moment".

[5] Thawornwong, S, Enke, D. Forecasting Stock Returns with Artificial Neural Networks, Chap. 3. In Neural Networks in Business Forecasting, Editor: Zhang, G.P. IRM Press, 2004.

[6] Hyeong Kyu Choi, B.A Student Dept. of Business Administration Korea University Seoul, Korea imhgchoi@korea.ac.kR Stock Price Correlation Coefficient Prediction with ARIMA-LSTM Hybrid Model.

[7] Dr. Jiban Chandra Paul, Md. Shahidul Hoque, Mohammad Morshedur Rahman University of Chittagong, Bangladesh. Selection of best ARIMA Model for forecasting Average Daily Stock Price Index of Pharmaceutical companies in Bangladesh.

[8] Xinyi Li1, Yinchuan Li2, Hong yang Yang1, Yang1, Xiao-Yang Liu1 1Columbia University, 2Beijing Institute of Technology {xl2717, yl3923, hy2500, ly2335, xl2427}@columbia.edu. DP-LSTM: Differential Privacy-inspired LSTM for Stock Prediction Using Financial News.

[9] Jaydeep Sen, Sidra Meh tab, Abhishek Dutta Statistical Finance (q-fin. ST); Machine Learning (cs. LG), Stock Price Prediction Using Machine Learning and LSTM- Based Deep learning models.

[10] Jaypee Institute of Information Technology, Noida 201304, India Department of Electronics and Communication Engineering, Jaypee Institute of Information Technology, Noida 201304, India c Department of Computer Science and Engineering, National Institute of Technology, Rourkela, 769008, India. Stock Closing Price Prediction using Machine Learning Techniques.

[11] Jingyi Shen & M. Omair Shafiq, Journal of Big data 7, Article Number: 66(2020), Short term stock market price trend prediction using a comprehensive deep learning system.

[12] Richard Waters (April 25, 2013). "Google search proves to be a new word in stock market prediction". Financial Times. Retrieved August 10, 2013.

[13] Beckmann, M. (January 24, 2017). Doctoral Thesis: Stock Price Change Prediction Using News Text Mining. COPPE/Federal University of Rio de Janeiro.