Stock Market Analysis and Prediction using Time Series Analysis

Rajdeep Singh

VIT Chennai, Computer Science Department, Chennai, India

Abstract - ***Over*** **the years stock market has been considered a very risky investment by people around the globe. This project aims to understand the historical data of stock market and derive analysis from it to reduce the gap of knowledge between the market behavior and the investor. A stock data comprises of lot of statistical terms which are difficult to understand by a normal person who wants to step into stock market investments, this project aims at reducing the gap of knowledge.**

This study aims to tell the market scenario of the future by supporting it with statistical answers. Stock market volatility, Daily returns, cumulative returns, Correlations between different stocks, Sharpe Ratio of the stocks, CAGR value, Simple Moving Average are some important statistical terms to understand the risk of the investment in the stocks. For prediction of the future behavior of stocks work on ARIMA models, Monte Carlo Method and Forecasting using Facebooks prophet library has been used here.

Keywords- **SMA, ARIMA, Monte Carlo, Fbprophet, Cumulative returns, Volatility, CAGR, Sharpe Ratio.**

I. INTRODUCTION

Time Series is the series of continuous data points index in order of date and time. Data connected through continuous series of time data. Analysis of time series data is done to extract meaningful statistics and other characteristics of data. After statistical calculations of time series data and after data analysis one can understand the behavior of the stock and deduce the amount of risk involved in it before making any investments.

Time series forecasting is a step further in making a value step towards understanding of future behavior. It refers to use of models to predict future values based on previously observed values. Models used in this study are Auto Regressive Integrated Moving Average (ARIMA) model, Augmented Dickey Fuller Test tells about Stationarity of Time Series Data, Monte Carlo Model is used to tell possible future predictions of a stock for a time period, prophet library by Facebook is very robust in processing the time series data and giving future predictions based on daily trend of data, weekly trend of data and yearly trend of data.

II. OBJECTIVE

The objective of this study is to understand and predict the stock behavior through statistical calculations and visualizations of historical data analysis. These objectives are:

a. Analysis of change in prices of the stock overtime.
b. Comparative analysis of the daily and cumulative return of the stocks.
c. Analysis using the Simple Moving Average of various stocks.
d. To find the correlation between different stocks' closing prices and daily returns.
e. To find the Sharpe Ratio of the stocks and to learn how it can be a helpful parameter while making investments.
f. To find the compounded annual growth rate of the stocks over the last 10 years.
g. To predict the future stock behavior and future prices using algorithms.

III. BACKGROUND

Stock market data over the years was considered to be very unpredictable and investment in the stock market was not very difficult for newcomers. Moreover, so much data and analytics was also not so easily available to the people. But Data Scientists, statisticians and tried to reduce the gap bit by bit over the years. Now a day's mutual fund applications use AI models, use certain statistic benchmarks to make it easy for a new comer to understand it in some extent.

Moreover, there has been a lot study done by people around the globe to predict the future prices of the stocks but stock market does not solely depend on the historical data. It is also affected by the sentiments of the people, which depend on some future events and so one cannot predict future events with 100% accuracy.

IV. MOTIVATION

Motivation behind this topic of study was the gap of knowledge most people have before they start investing in the stock market. Every year smart investors make good chunk of money out the stock market. Self-made billionaire Warren Buffet is one such example who earned most of his wealth through smart investing. Prediction of future of

stock behavior is another motivation for me. This gap of investment knowledge needs to be reduced for a common man.

V. LITERATURE SURVEY

The author Banarjee D. [1] has done his study on forecasting of National Stock Exchange data using ARIMA model and has done a comparative study on different values of p, d and q on ARIMA and did validation check of forecasted stock price with actual stock price. In his study ARIMA (1,0,1) gave the best fit compared to other models.

The authors Viswam, N., & Reddy, G. S. [2] did study on historical stock market data predicted future prices using ARIMA model and they used MACD model to better analysis of the data.

The authors Angadi, M. C., & Kulkarni, A. P. [3] used ARIMA model for prediction of stock prices, for p, d, q values they used auto ARIMA to get the best fit for the model. Obtained results reveal that ARIMA model has strong tendency to make short time predictions.

The authors Devi, B. U., Sundar, D., & Alli, P. [4] used NIFTY MIDCAP 50 as the index and selected top four MIDCAP companies. They used ARMA and ARIMA models to predict the future stock prices and used AIC and BIC criteria to get the best fit for the model.

The authors Varghese, A., Tarhen, H., Shaikh, A., Banik, P., & Ramadasi, A. [5] used ARMA, Moving average, and ANN model to predict the future prices of the stock and used maximum likelihood estimator and Yule-Walker estimation to check the validation of their models.

The authors Sharma, A., Modak, S., & Sridhar, E. [6] using LSTM model from RNN to predict the random nature of stock prices in future. MSE value of the model came out to be significant and improved.

The authors Selvin, S., Vinayakumar, R., Gopalakrishnan, E. A., Menon, V. K., & Soman, K. P. [7] used NSE listed companies as the stock price data. They used sliding window approach on non-linear model RNN, LSTM and CNN and linear model ARIMA. Non-linear model outperformed the linear model during error calculation.

The authors Mondal, P., Shit, L., & Goswami, S. [8] did a study on effectiveness of ARIMA model in forecasting security values. They used Indian Stock market data from NSE for the analysis. AIC has been used for selecting the best ARIMA model.

Th authors Wang, J., & Wang, J. [9] used principle component analysis and Stochastic time effective neural networks for forecasting the future and used MAE, MAPE, MSE and RMSE to calculate the performance of the model.

VI. TECHNICAL SPECIFICATIONS

6.1 Software Specification

Anaconda Distribution with software version Anaconda3 2020.02(version).

Jupyter Notebook framework (version 6.0.2) for implementation of the code.

6.2 Packages Installed

All the basic packages such as Numpy, Pandas, Matplotlib, Seaborn, Scipy, Statsmodels and Sklearn.

Some special libraries requied for this project are: pandas_datareader, datetime, pmdarima and fbprophet.

VII. DATASET DESCRIPTION

Data used for this project is day wise historical time series data of stock of past 10 years in numerical form.

Dataset size – 10 Business years

Data Source used – Yahoo finance

Dataset imported from yahoo finance consists of 7 columns of consisting of Open, High, Low, Close, Adj. Close, Volume and Date as index.

Date (Index of the Dataset): Dates of all Business Days in a year.

Open: Depicts the Opening price of the Security.

Close: Depicts the Closing price of the Security.

High: Depicts the Highest value gained by the stocks in a particular day.

Low: Depicts the Lowest value gained by the stocks in a particular day.

Adj. Close: The adjusted closing price is calculated after analyses of the stock's dividends, stock splits and the new stock offerings which determines a new value of the stock know as adjusted price.

Volume: Volume, or trading volume, is the amount of a security that was traded during a given period of time.

VIII. STATISTICAL PARAMETERS

A. Daily Stock Return

This parameter tells us about how much the stocks gained or lost per day per share. It is calculated by subtracting previous day's closing price from todays' closing price.

B. Stock Volatility

Volatility is a measure of the dispersion of returns for a given stock or the market index. Mostly, the higher the volatility, the riskier is the stock. It is often measured as either the standard deviation or variance between returns from that same stock or the market index.

In the stock markets, volatility is often associated with big swings in the price in either direction of the trend. For ex., when the market gains and loses value more than one percent over a specific time period, this is known as volatility of a market.

$$Daily\ Volatility\ Formula = \sqrt{Variance}$$

$$Annual\ Volatility\ Formula = \sqrt{252} \times \sqrt{Variance}$$

C. Cumulative Stock Return

This parameter tells us about how much the stocks gained or lost per share over time, independent of the time period. Cumulative Return is equal to:

$$\frac{(Current\ Price) - (Original\ Price)}{Original\ Price}$$

D. Compounded Annual Growth Rate(CAGR)

CAGR is the rate of return by which tell us that this rate would be required for a company to grow from its starting value to its ending value, it is assumed that the profits gained were again invested at the end of each business year of an investment firm.

$$CAGR = \left(\frac{V_{final}}{V_{Begin}}\right)^{1/t} - 1$$

Where:

CAGR = compound annual growth rate
V_{begin} = beginning value
V_{final} = final value
T = time in years

E. Correlation of Stocks

Correlation(r) is a statistical measure which tells us the amount to which two variables move in relation with each other. In finance, the correlation can measure the movement of a stock price with stock markets benchmark index.

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2}\sqrt{(Y - \bar{Y})^2}}$$

Where:

r = the correlation coefficient
$\bar{X}$ = the average of the observations of variable X
$\bar{Y}$ = the average of the observations of variable Y

F. Sharpe Ratio

Sharpe ratio is a measure of risk adjusted returns of a financial portfolio. Sharpe ratio is a measure of excess return one gets over the risk-free rate, relative to its standard deviation.

$$Sharpe\ Ratio = \frac{R_p - R_f}{\sigma_p}$$

Where:

R_p = return of portfolio
R_f = risk free return rate
σ_p = standard deviation of the portfolio excess return

G. Simple Moving Average

Simple moving average (SMA) is a simple method for technical analysis that smooths out price data of stock market by updating average price of stocks constantly. This average is calculated over a specific period of time, like 10 days, 30 minutes, 20 weeks or any time period the investor chooses. Moving average techniques are very popular and can be calculated for any time frame, suiting both long-term investments and short-term investments.

$$SMA = \frac{A_1 + A_2 + \cdots + A_n}{n}$$

Where:

A_n= The price of a stock during a period n
n = the number of total periods

IX. DATA ANALYSIS

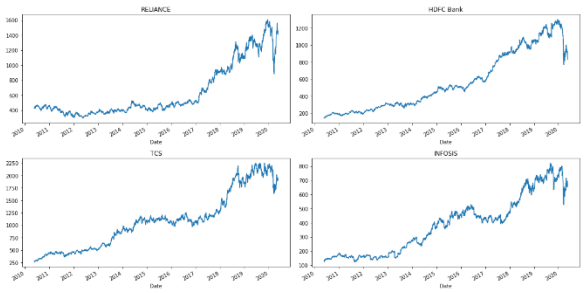

Fig 1: Plots of Stock Prices overtime

Prices of stocks can be seen increasing with an upward trend in plots for the 4 companies we selected for the selected time period.

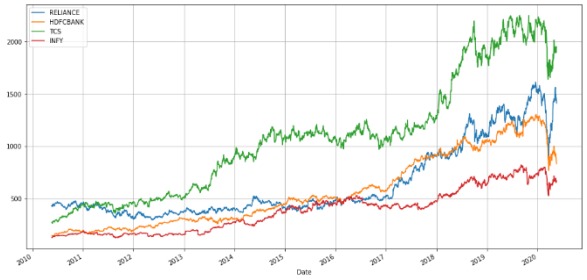

Fig 2: Combined plot of stock for price Comparisons

We can observe two Tech companies TCS and INFOSIS have a close similarity in the plots (fig 1) that there must some correlation between these stocks because of their behavior whereas both of them have a big difference in the Stock prices which can be observed on the y axis.

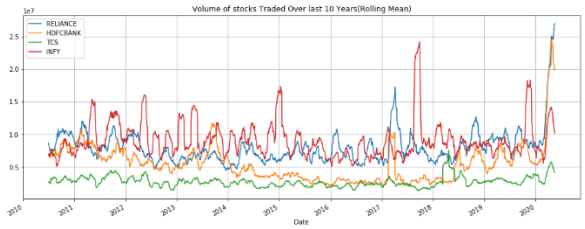

Fig 3: Comparative analysis of Volume Traded with rolling mean of 30.

Combined plot of Volume traded of all the four stocks made the plot look very cluttered. To get a clear understanding between the Values we will take rolling mean of volume traded of 30 days. By observing this plot, we can analyze that higher is stock price lesser will be the volume traded. For ex. If a person invests 10000 Rs in each of them, higher is the price of the stock lesser will be the number of stocks bought.

Calculating Daily Returns:

Date	Reliance	HDFC	TCS	INFY
2010-05-21	-0.004400	-0.013183	-0.015480	-0.007306
2010-05-24	0.027221	-0.001095	-0.001530	0.007223
2010-05-25	-0.036181	-0.011400	-0.025153	-0.025765
2010-05-26	0.022726	0.013306	0.054964	0.086040
2010-05-27	0.014087	0.037669	0.005217	0.008510
...	...	...	...	...
2020-05-13	0.021217	0.028950	0.000077	0.009452
2020-05-14	-0.040429	-0.036598	-0.024261	-0.051862
2020-05-15	0.016331	-0.006210	-0.004968	-0.008889
2020-05-18	-0.012779	-0.057986	0.027841	0.017783
2020-05-19	-0.022107	-0.007171	0.001568	0.007079

2461 rows × 4 columns

Tab 1: Daily returns of stocks

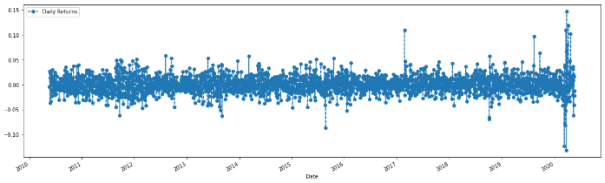

Fig 4: Daily returns when plotted

Daily returns can be considered as the value gained or lost by stock with respect to the previous day. We can see a graph of daily returns with mean close to zero.

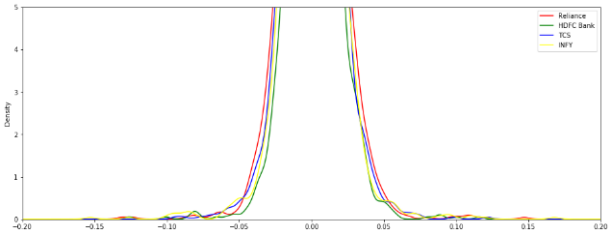

Fig 5: Kernel Density plot to compare Volatility

From here we analyzed and tell that Reliance stocks are most volatile, and HDFC Bank Stocks are least volatile.

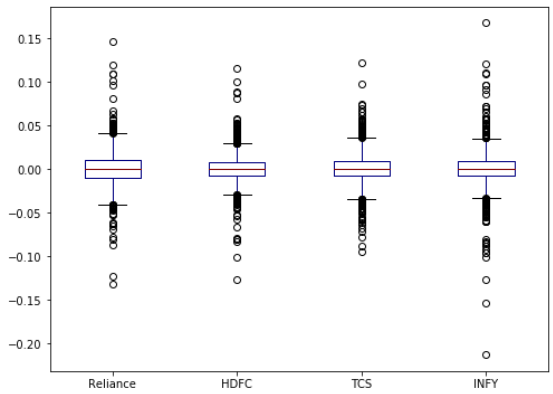

Fig 6: Box plot to compare volatility of stocks

Through box plot we can observe that INFOSIS has wider graph due to some outliers, neglecting those above analysis is completely supported.

Calculating Cumulative Return of the stocks:

Date	Reliance	HDFC	TCS	INFY
2010-05-20	1.000000	1.000000	1.000000	1.000000
2010-05-21	0.995600	0.986817	0.984520	0.992694
2010-05-24	1.022701	0.985737	0.983014	0.999865
2010-05-25	0.985699	0.974500	0.958288	0.974103
2010-05-26	1.008100	0.987466	1.010959	1.057914
...	...	...	...	...
2020-05-13	3.485026	6.451662	7.129287	5.389721
2020-05-14	3.344130	6.215545	6.956325	5.110199
2020-05-15	3.398742	6.176945	6.921769	5.064777
2020-05-18	3.355308	5.818771	7.114477	5.154845
2020-05-19	3.281134	5.777042	7.125630	5.191338

2462 rows × 4 columns

Tab 2: Cumulative return of the stocks over the years

In the last 10 years, TCS gave us the maximum return, amount invested in TCS stocks have surged up to 7 times in last 10 years. while Reliance gave us minimum return where prices surge up to 3.3 times approx. With least volatility among the stocks HDFC has Stock seems to very promising with its returns.

Calculation of Compounded Annual Growth Rate of stocks and the market index.

```
Reliance      12.616596
HDFC Bank     19.170991
TCS           21.697688
Infosis       17.903835
Nifty          6.022352
dtype: float64
```

In the past 10 years, TCS attained really excellent CAGR value and Reliance have the least Annual growth rate. All the four stocks have tended to perform better than what the market have received over last 10 years.

Finding correlation:

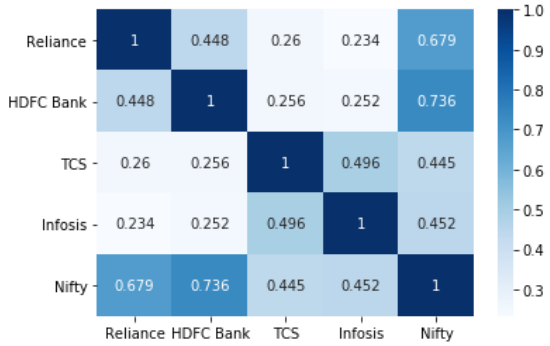

Fig 7: Correlation of Daily returns using Heat Map

We can observe market index NIFTY has good correlation with the stocks. Moreover, both the tech companies seem to have a correlation which tends to slightly weak.

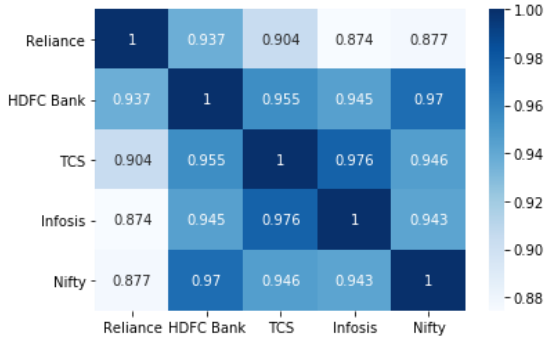

Fig 8: Correlation of Closing Price using Heat Map

Market index is closely related with all the four companies. Moreover, we can observe highest correlation between the tech companies.

Fig 9: Simple Moving Average for 10, 50 and 200 days

Moving averages tells us that this is average or minimum possible return that we are likely to get in the near future. Moving averages change slowly in comparison to the actual price.

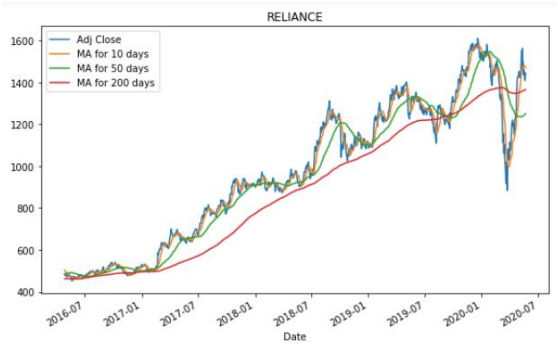

Fig 10: SMA of Reliance

From the SMA plot of Reliance stock over 10, 50 and 200 days we can analyze that 10 MA removes the noise from the plot while 50 days MA smoothens it over a period of time and help understand market behavior over next 50 days and the 200 days MA gives us the trend.

As per mean reversion strategy it tells us that whenever the actual price plot line crosses the MA line it states that market is going to rebound again towards the mean. So, in the above plot we can observe that in March-April 2020 when the prices fell below the MA200 they are bound to rise up in future towards the mean.

SMA have been very helpful over the years for making an investment.

Calculating Sharpe Ratio:

Date	Reliance	HDFC Bank	TCS	Infosis
2011-12-31	-1.290702	-0.392527	-0.124830	-0.590615
2012-12-31	0.576007	2.691502	0.098416	-0.444074
2013-12-31	0.068441	-0.272582	2.776967	1.834415
2014-12-31	-0.207587	1.953733	0.590855	1.130394
2015-12-31	0.353913	0.452002	-0.383376	0.647585
2016-12-31	0.084834	0.356861	-0.238275	-0.484872
2017-12-31	2.624157	3.835127	0.588217	0.148690
2018-12-31	0.637510	0.527094	1.489296	1.083440
2019-12-31	1.063034	0.629196	0.364209	0.258321

Tab 3: Annual Sharpe Ratio of last 9 years from 2011 to 2019

Sharpe ratio also known as Risk Adjusted Return is the measure of excess return of a portfolio of stocks over the risk-free return. It tells us whether the risk taken to earn the return over the risk-free possible return, from an Investment security is worth the taking a risk or not. It helps us in choosing right stocks with minimum risk good amount of return. Sharpe Ratio is a term which is calculated annually, and it changes overtime.

What Sharpe Ratio is considered Good?

ASR< 1 – Bad (Not Worth taking the risk)

1<ASR<2 – Acceptable

2<ASR<3 – Good

ASR>3 – Excellent

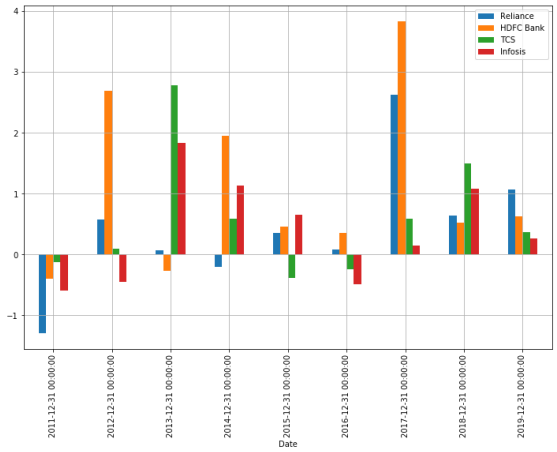

Fig 11: Annual Sharpe Ratio Bar Plot of last 9 years from 2011 to 2019

X. FUTURE PRICE PREDICTION

1. *Augmented Dickey Fuller Test:*

In AD Fuller test, time series data is tested for the null hypothesis. Null hypothesis states that a unit root is found in a time series data. Alternate hypothesis depends on whether which type of the test is being used on the data, which is usually stationarity or trend-stationarity. For large and complex set of time series models AD Fuller test is used.

The ADF statistic measure which is used for the test is a negative number and more the number is negative, hypothesis will be rejected more strongly which means that there is a unit root with some level of confidence. Results of AD fuller test are:

```
Results of dickey fuller test
Test Statistics                  -2.036163
p-value                           0.581605
No. of lags used                 21.000000
Number of Observations used    2440.000000
Critical Value (1%)              -3.962485
Critical Value (5%)              -3.412291
Critical Value (10%)             -3.128110
dtype: float64
```

p-value>0.05 implies data is non stationary.

2. *Finding ACF & PACF:*

The plot of ACF is a bar graph of coefficients of correlation which is plotted between a time series and lags with itself. The PACF plot is a plot between the series and the lags of itself which gives partial correlation coefficients. ACF and PACF plots are used to identify the number of AR and MA terms.

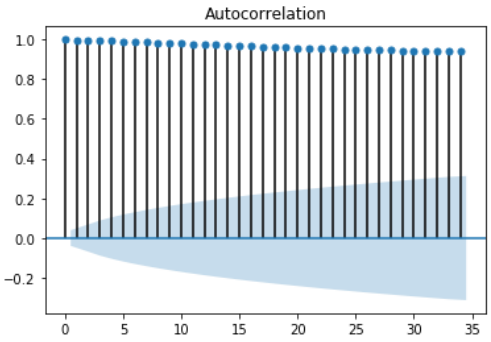

Fig 12: ACF Plot of Reliance Stock

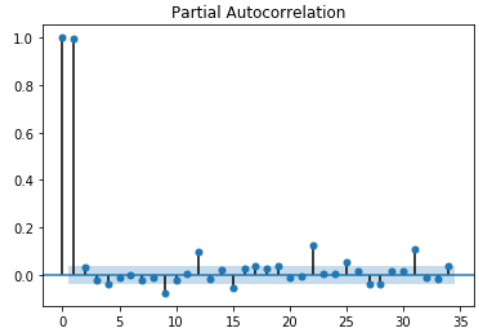

Fig 13: PACF Plot of Reliance Stock

3. Seasonal decomposition of Time Series:

Decomposition of Time series data into Seasonal, Trend and Residual component is known as the seasonal decomposition.

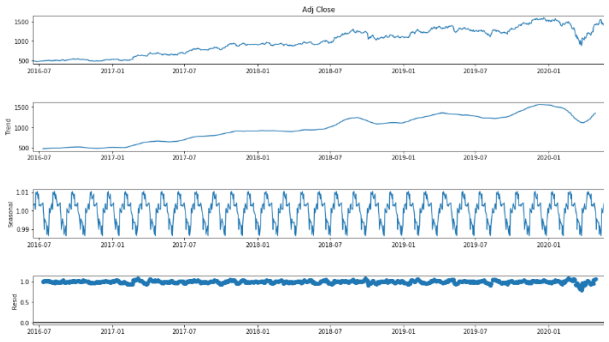

Fig 14: Seasonal Decomposition of Reliance Stock into Trend, Seasonal and Residual Components.

4. ARIMA model:

Auto-Arima function:
Auto Arima function does the job of finding best possible values of p, d, q for our model. Model with lowest AIC value is selected and gives us the required values of p, d, q for our ARIMA model. The results gave p=0, d=1 and q=1 as the best fit for the model.

ARIMA stands for Auto-Regressive Integrated Moving Average. Here, "Auto-Regressive term" means the lags in a stationary series of a forecasting equation, "moving average term" means the lags in the forecast errors, and "integrated" means the time series data is made stationary by using differencing method on the series. Special cases of an ARIMA model are autoregressive models, random-walk, exponential smoothing models, and random-trend models.

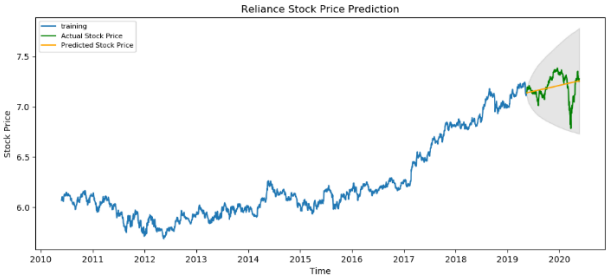

Fig 15: ARIMA forecasting plot of future values with 95% confidence

Results of the ARIMA model are:

```
MSE: 0.013511206582547987
MAE: 0.0852704149709506
RMSE: 0.11623771583504205
MAPE: 0.011900728052480233
```

5. Monte Carlo Simulation:

There are situations when problem comprises of random variables which cannot be predicted easily then probability of different possible outcomes are modeled, this is known as monte carlo simulation. This method is useful in understanding the impact of risk and uncertainty of prediction and forecasting models during analysis. Monte Carlo model can also be used to solve various problems such as in fields of engineering, supply chain, science and finance. This model is also referred as multiple probability simulation model.

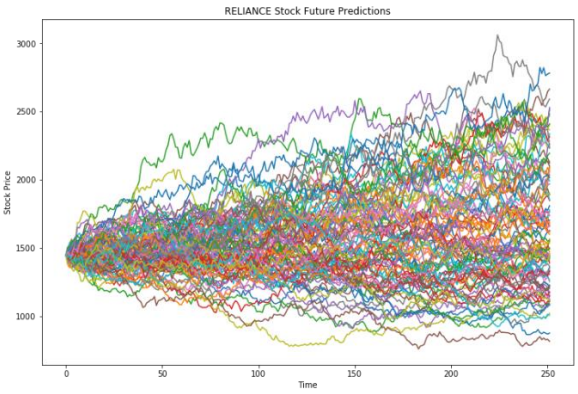

Fig 16: Monte Carlo Simulation of Reliance of 1 Business year

6. Fbprophet for forecasting:

Fbprophet is a very handy library made by Facebook for Time Series data. Here, The Blue line in center

implies the future trend of average stock prices while the actual value can vary within the blue band. The blue band helps in tracking in what range can be the future stock prices can lie. Black scattered points are the actual stock data over the time period.

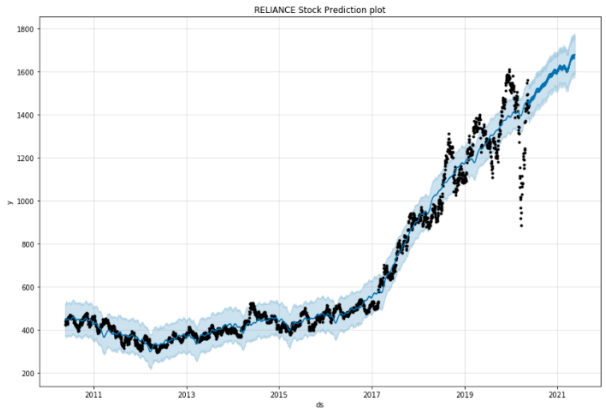

Fig 17: Prophet Forecast of Reliance of Next one year

Using prophet data can be properly distributed into overall trend, weekly trend, and yearly trend. Here we can observe yearly trend have some seasonality.

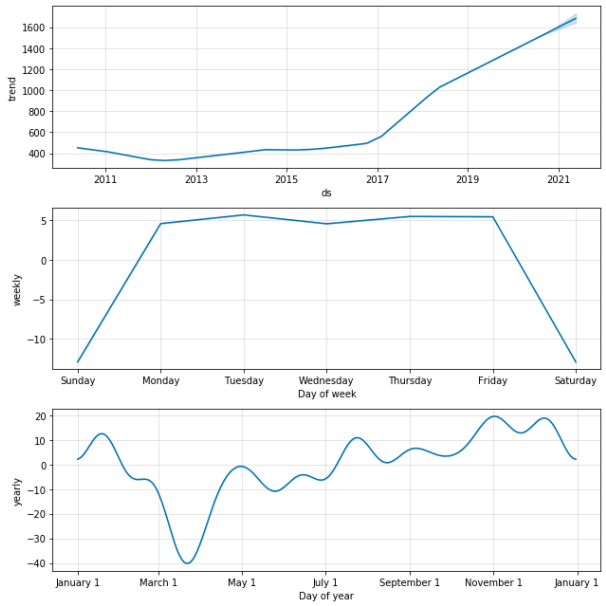

Fig 18: Decomposition into daily trend, weekly trend and yearly trend.

XI. CONCLUSION

In this study, we could analyze and understand the risk involved in the stock price and are well accounted by statistical terms such as CAGR, Sharpe Ratio volatility and the cumulative return. This study has proved to be successful in comparison analysis between stocks and get stock with lesser risk and good return.

We also found out that ARIMA model performed really well in predicting the stock prices. Mont Carlo also gave us an overview of possible future trends. Prophet library by Facebook worked well gave us fulfilling results of future prices.

XII. FUTURE WORK

In this study we worked with linear model for the forecasting of the future price. But in there no evidence that stock data was linear, this motivates to work on forecasting using the non-linear data in the near future. Maybe with some statistical measurements we can develop a model which can help in choosing the right stock.

REFERENCES

[1] Banerjee, D. (2014, January). Forecasting of Indian stock market using time-series ARIMA model. In *2014 2nd International Conference on Business and Information Management (ICBIM)* (pp. 131-135). IEEE.

[2] Viswam, N., & Reddy, G. S. (2018). Stock market prediction using time series analysis.

[3] Angadi, M. C., & Kulkarni, A. P. (2015). Time Series Data Analysis for Stock Market Prediction using Data Mining Techniques with R. International Journal of Advanced Research in Computer Science, 6(6).

[4] Devi, B. U., Sundar, D., & Alli, P. (2013). An effective time series analysis for stock trend prediction using ARIMA model for nifty midcap-50. *International Journal of Data Mining & Knowledge Management Process*, *3*(1), 65.

[5] Varghese, A., Tarhen, H., Shaikh, A., Banik, P., & Ramadasi, A. (2016). Stock Market Prediction Using Time Series. International Journal on Recent and Innovation Trends in Computing and Communication, 4(5), 427-430.

[6] Sharma, A., Modak, S., & Sridhar, E. (2019). Data Visualization and Stock Market and Prediction.

[7] Selvin, S., Vinayakumar, R., Gopalakrishnan, E. A., Menon, V. K., & Soman, K. P. (2017, September). Stock price prediction using LSTM, RNN and CNN-sliding window model. In 2017 international conference on advances in computing, communications and informatics (icacci) (pp. 1643-1647). IEEE.

[8] Mondal, P., Shit, L., & Goswami, S. (2014).

Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices. International Journal of Computer Science, Engineering and Applications, 4(2), 13.

[9] Wang, J., & Wang, J. (2015). Forecasting stock market indexes using principle component analysis and stochastic time effective neural networks. Neurocomputing, 156, 68-78.

[10] How to Use the Sharpe Ratio to Analyze Portfolio Risk and Return. (2020). Retrieved 22 May 2020, from https://www.investopedia.com/terms/s/sharperatio.asp

[11] Brownlee, J. (2020). What Is Time Series Forecasting? Retrieved 22 May 2020, from https://machinelearningmastery.com/time-series-forecasting/