Materials Today: Proceedings xxx (xxxx) xxx

Contents lists available at ScienceDirect

Materials Today: Proceedings

journal homepage: www.elsevier.com/locate/matpr

Stock market analysis and prediction using time series analysis

Christy Jackson J. [a,*], Prassanna J. [a], Abdul Quadir Md. [a], Sivakumar V. [b]

[a] Vellore Institute of Technology, Chennai, Tamil Nadu 600127, India
[b] Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai 600062, India

ARTICLE INFO

Article history:
Received 3 November 2020
Accepted 12 November 2020
Available online xxxx

Keywords:
ARIMA
Monte-Carlo
Fbprophet
Cumulative return
Volatility
CAGR
Sharpe ratio

ABSTRACT

Over the years the stock market has been considered a very risky investment by people around the globe. This project aims to understand the historical data of the 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 a 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 the prediction of the future behavior of stocks work on ARIMA models, Monte Carlo Method and Forecasting using Facebook's prophet library have been used here.

Selection and peer-review under responsibility of the scientific committee of the Emerging Trends in Materials Science, Technology and Engineering.

1. Introduction

Time Series is a series of continuous data point's index in order of date and time. Data connected through a 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 valuable step towards the understanding of future behavior. It refers to the 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 stock for some time, prophet library by Facebook is very robust in processing the time series data and giving future predictions based on a daily trend of data, a weekly trend of data and yearly trend of data.

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 were 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 newcomer to understand it to some extent.

Moreover, there has been a lot of studies done by people around the globe to predict the future prices of the stocks but the 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.

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 over time.
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.

* Corresponding author.
E-mail address: christyjackson.j@vit.ac.in (J. Christy Jackson).

https://doi.org/10.1016/j.matpr.2020.11.364
2214-7853/
Selection and peer-review under responsibility of the scientific committee of the Emerging Trends in Materials Science, Technology and Engineering.

g) To predict future stock behavior and future prices using algorithms.

2. Related work

The 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 a good chunk of money out of the stock market. Self-made billionaire Warren Buffet is one such example who earned most of his wealth through smart investing. Prediction of future stock behavior is another motivation. This gap of investment knowledge needs to be reduced for a common man.

The author Banarjee D. [1], has done his study on forecasting of National Stock Exchange data using the ARIMA model and has done a comparative study on different values of p, d, and q on ARIMA and did a validation check of the forecasted stock price with the actual stock price. In this study, ARIMA (1,0,1) gave the best fit compared to other models. The authors Viswam, N., & Reddy, G. S. [2], did a study on historical stock market data predicted future prices using the ARIMA model and they used the MACD model to better analysis of the data.

The authors Angadi, M. C., & Kulkarni, A. P. [3], used the 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 the ARIMA model has a 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 the top four MIDCAP companies. They used ARMA and ARIMA models to predict 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, an 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 the LSTM model from RNN to predict the random nature of stock prices in the future. The 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 a sliding window approach on non-linear model RNN, LSTM, and CNN and linear model ARIMA. The non-linear model outperformed the linear model during error calculation. The authors Mondal, P., Shit, L., & Goswami, S. [8], did a study on the effectiveness of the 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. The author's 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.

3. Dataset description

Data used for this project is day-wise historical time series data of stock of the 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 consisting of Open, High, Low, Close, Adj. Close, Volume, and Date as the 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 on 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 determine 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 time.

4. Statistical parameters

4.1. Daily stock return

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

4.2. 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, this is known as volatility of a market.

$$DailyVolatilityFormula = \sqrt{Variance} \tag{1}$$

$$AnnualVolatilityFormula = \sqrt{252} \times \sqrt{Variance} \tag{2}$$

4.3. Cumulative stock return

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

$$\frac{(CurrentPrice) - (OriginalPrice)}{OriginalPrice} \tag{3}$$

4.4. 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.

$$CAGR = \left(\frac{V_{final}}{V_{Begin}}\right)^{1/t} - 1 \tag{4}$$

Where:

CAGR = compound annual growth rate

V_{begin} = beginning value

V_{final} = final value

T = time in years

4.5. Correlation of stocks

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

$$r = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum\left(X-\bar{X}\right)^2}\sqrt{\left(Y-\bar{Y}\right)^2}} \quad (5)$$

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

4.6. 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 [10].

$$SharpeRatio = \frac{R_p - R_f}{\sigma_p} \quad (6)$$

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

4.7. Simple moving average

The simple moving average (SMA) is a simple method for technical analysis that smooths out the price data of the stock market by updating the average price of stocks constantly. This average is calculated over a specific period, like 10 days, 30 min, 20 weeks, or any 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.

$$SMA = \frac{A_1 + A_2 + \cdots + A_n}{n} \quad (7)$$

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

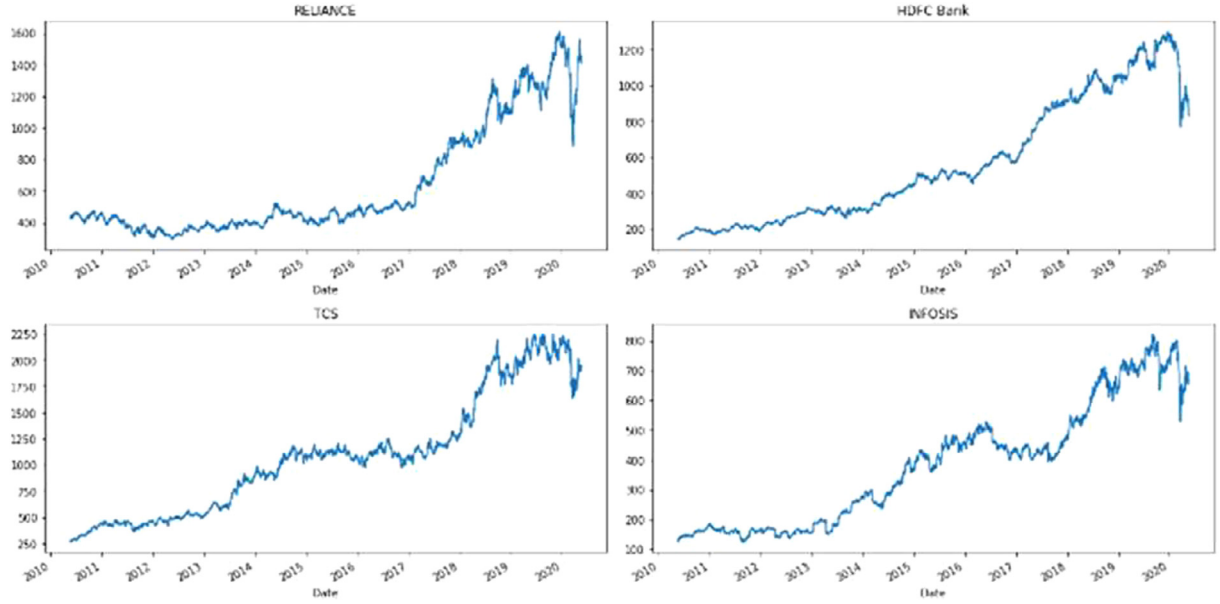

Fig. 1. Plots of stock prices overtime.

Fig. 2. Combined plot of stock for price comparisons.

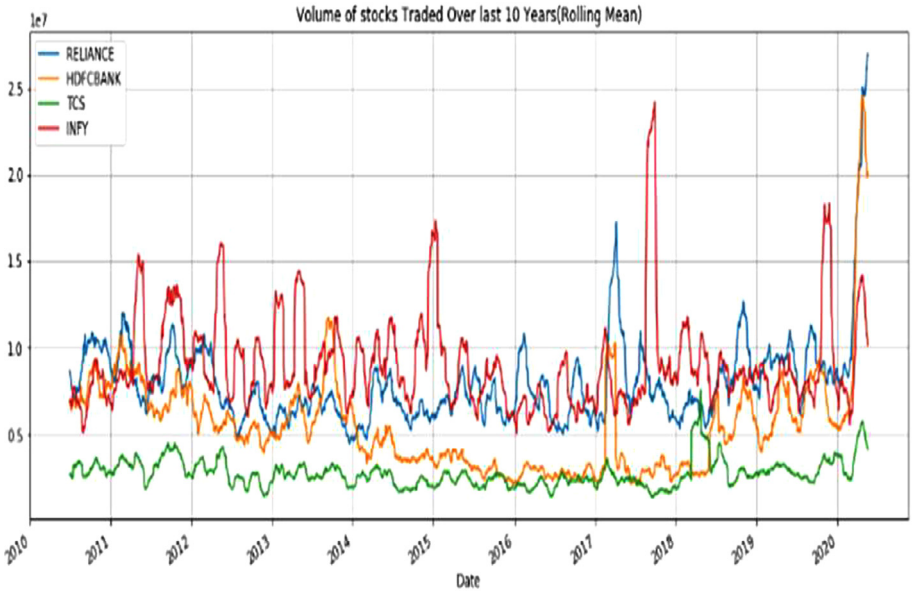

Fig. 3. Comparative analysis of volume traded with a rolling mean of 30.

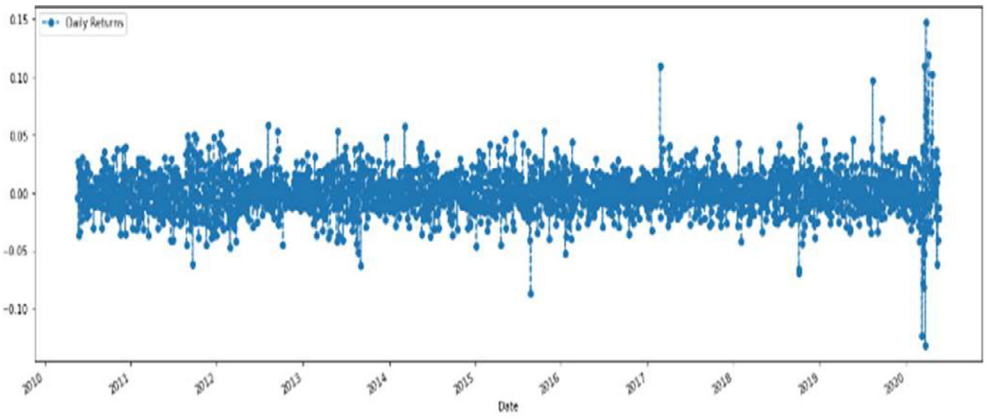

Fig. 4. Daily returns when plotted.

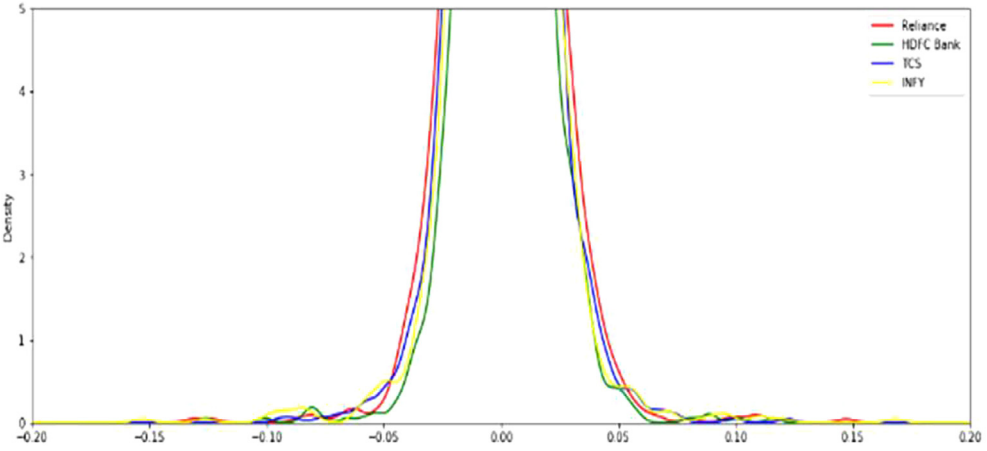

Fig. 5. Kernel density plot to compare volatility.

Table 1
Daily returns of stocks.

Date	Reliance	HDFC	TCS	INFY
21-05-2010	−0.0044	−0.013183	−0.01548	−0.007306
24-05-2010	0.027221	−0.001095	−0.00153	0.007223
25–05-2010	−0.036181	−0.0114	−0.025153	−0.025765
26–05-2010	0.022726	0.013306	0.054964	0.08604
27–05-2010	0.014087	0.037669	0.005217	0.00851
...	...	...	...	...
13-05-2020	0.021217	0.02895	0.000077	0.009452
14-05-2020	−0.040429	−0.036598	−0.024261	−0.051862
15-05-2020	0.016331	−0.00621	−0.004968	−0.008889
18-05-2020	−0.012779	−0.057986	0.027841	0.017783
19-05-2020	−0.022107	−0.007171	0.001568	0.007079
2461 rows × 4 columns				

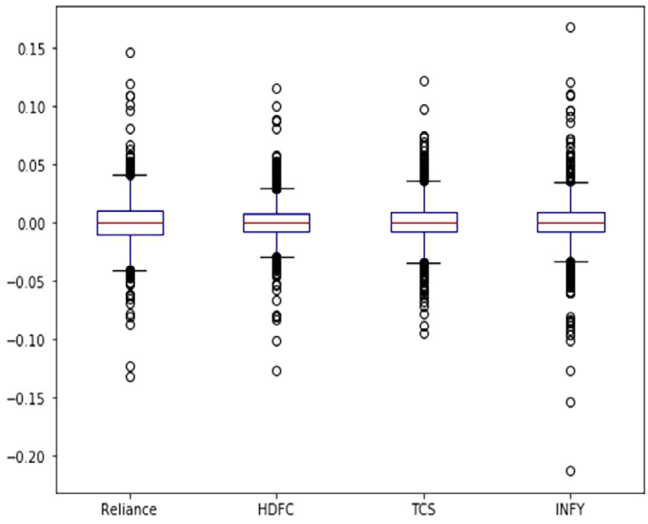

Fig. 6. Box plot to compare the volatility of stocks.

5. Analysis

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

We can observe two Tech companies TCS and INFOSYS have close similarities 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. 2. Fig. 3).

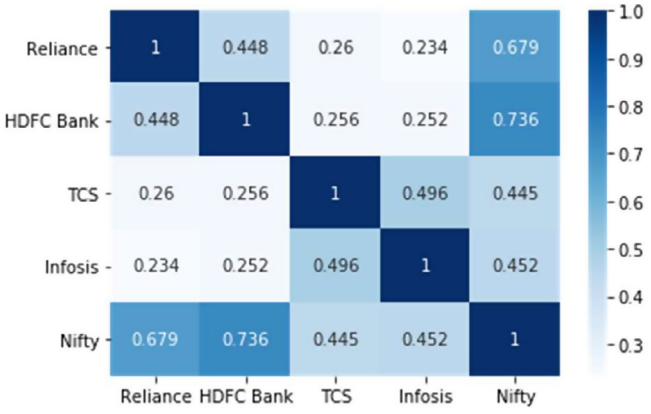

Fig. 7. Correlation of daily returns using heat map.

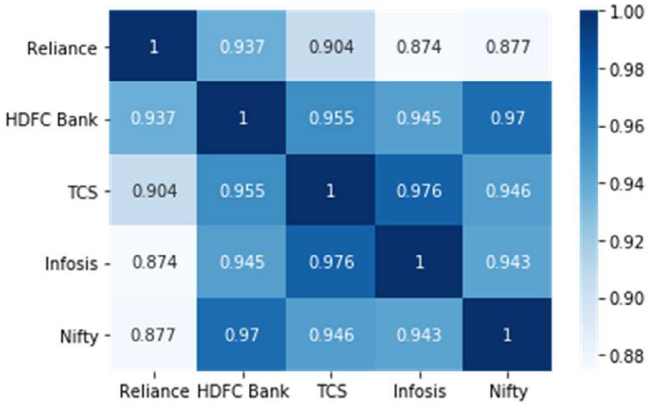

Fig. 8. Correlation of closing price using heat map.

Table 2
Cumulative return of the stocks over the years.

Date	Reliance	HDFC	TCS	INFY
20-05-2010	1.000000	1.000000	1.000000	1.000000
21-05-2010	0.9956	0.986817	0.98452	0.992694
24-05-2010	1.022701	0.985737	0.983014	0.999865
25-05-2010	0.985699	0.9745	0.958288	0.974103
26-05-2010	1.0081	0.987466	1.010959	1.057914
...	...	...	...	...
13-05-2020	3.485026	6.451662	7.129287	5.389721
14-05-2020	3.34413	6.215545	6.956325	5.110199
15-05-2020	3.398742	6.176945	6.921769	5.064777
18-05-2020	3.355308	5.818771	7.114477	5.154845
19-05-2020	3.281134	5.777042	7.12563	5.191338
2462 rows × 4 columns				

Fig. 9. Simple moving average for 10, 50, and 200 days.

5.1. Calculating daily returns

Daily returns can be considered as the value gained or lost by stock to the previous day. We can see a graph of daily returns with a mean close to zero (Fig. 4. Fig. 5. Table 1).

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

Through the box plot, we can observe that INFOSYS has a wider graph due to some outliers, neglecting those above analysis is completely supported (Fig. 6).

5.2. Calculating cumulative return of the stocks

In the last 10 years, TCS gave us the maximum return, the amount invested in TCS stocks has surged up to 7 times in the last 10 years. While Reliance gave us minimum return where prices surge up to 3.3 times approx. With the least volatility among the stocks HDFC has Stock seems to very promising with its returns (Table 2).

5.3. Calculation of compounded annual growth rate of stocks and the market index

In the past 10 years, TCS attained an excellent CAGR value and Reliance have the least Annual growth rate. All four stocks have

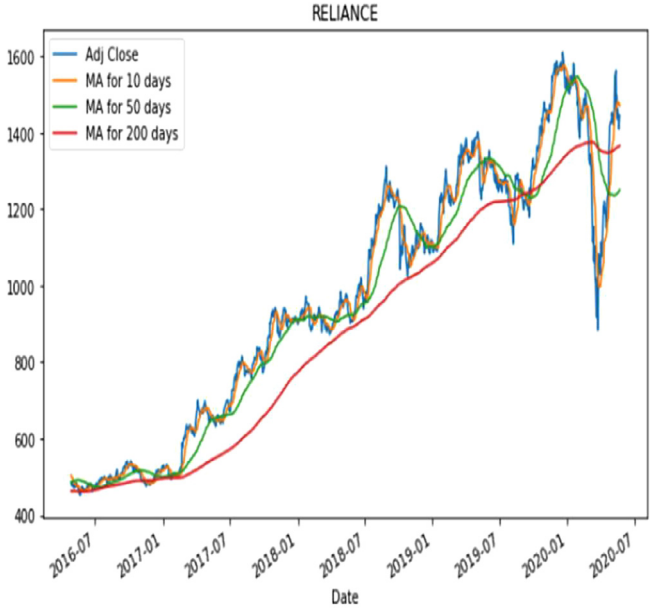

Fig. 10. SMA of reliance.

Table 3
Annual Sharpe Ratio of last 9 years from 2011 to 2019.

Date	Reliance	HDFC	TCS	INFY
31-12-2011	−1.290702	−0.392527	−0.124830	−0.590615
31-12-2012	0.576007	2.691502	0.098416	−0.444074
31-12-2013	0.068441	−0.272582	2.776967	1.834415
31-12-2014	−0.207587	1.953733	0.590855	1.130394
31-12-2015	0.353913	0.452002	−0.383376	0.647585
31-12-2016	0.084834	0.356861	−0.238275	−0.484872
31-12-2017	2.624157	3.835127	0.588217	0.148690
31-12-2018	0.63751	0.527094	1.489296	1.08344
31-12-2019	1.063034	0.629196	0.364209	0.258321

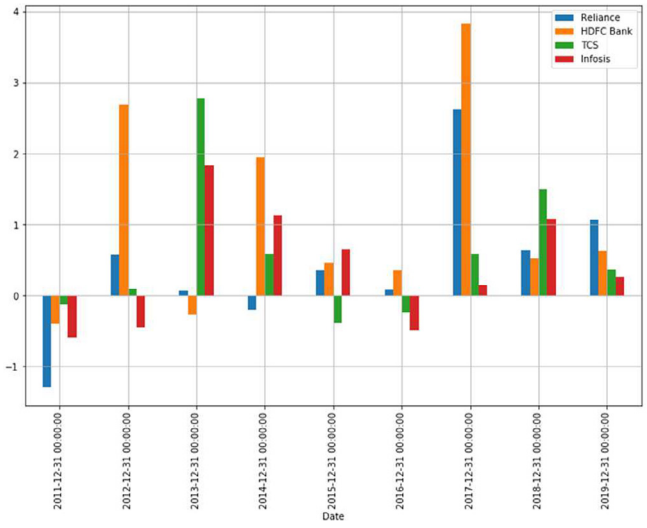

Fig. 11. Annual sharpe ratio bar plot of last 9 years from 2011 to 2019.

Table 4
Results of Dickey-Fuller test.

Parameters	Values
Test Statistics	−2.036163
p - value	0.581605
No. of lags used	21
Number of Observations used	2440
Critical Value (1%)	−3.962485
Critical Value (5%)	−3.412291
Critical Value (10%)	−3.12811

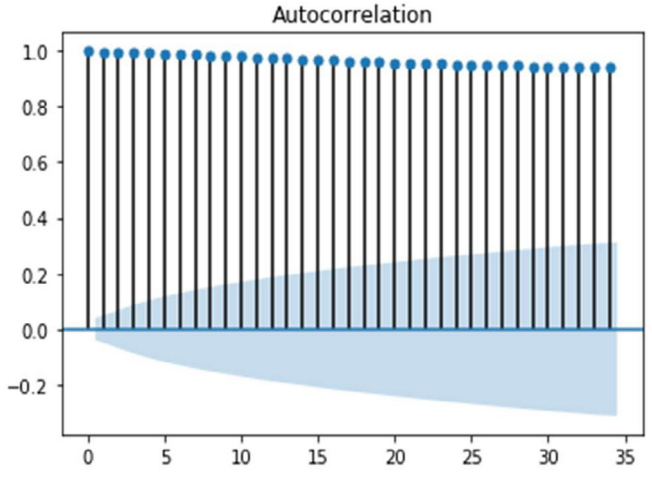

Fig. 12. ACF plot of reliance stock.

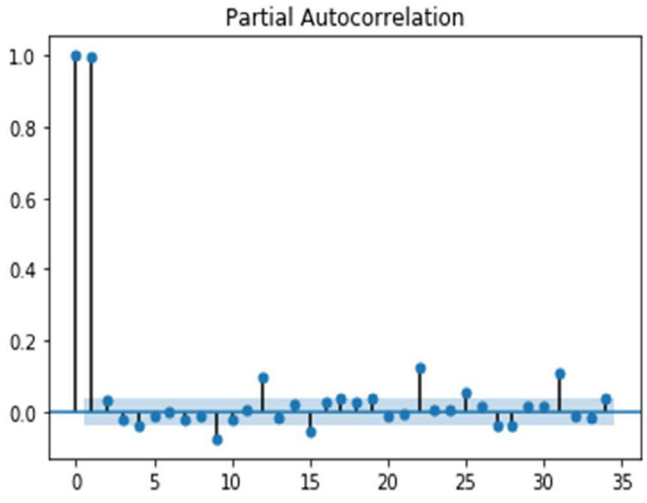

Fig. 13. PACF plot of reliance stock.

tended to perform better than what the market has received over the last 10 years.

5.3.1. Finding correlation

It can be observed, market index NIFTY has a good correlation with the stocks. Moreover, both the tech companies seem to have a correlation which tends to slightly weak (Fig. 7. Fig. 8).

The market index is closely related to all four companies. Moreover, we can observe the highest correlation between tech companies (Fig. 9).

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

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 some time and help understand market behavior over the next 50 days and the 200 days MA gives us the trend.

As per the mean reversion strategy, it tells us that whenever the actual price plot line crosses the MA line it states that the 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 arise in the future towards the mean. SMA has been very helpful over the years for investing (Fig. 10).

5.4. Calculating sharpe ratio

Sharpe ratio is also known as Risk-Adjusted Return is the measure of the 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 Investment security is worth taking a risk or not. It helps us in choosing the right stocks with minimum

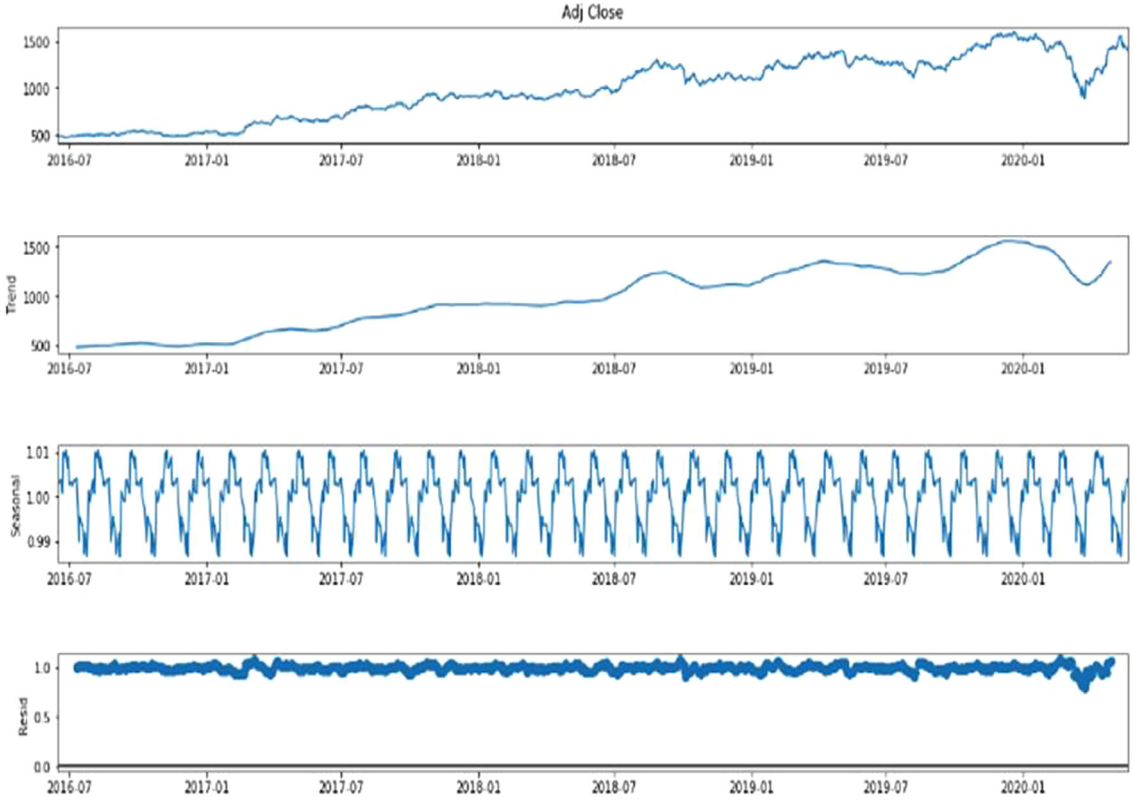

Fig. 14. Seasonal decomposition of reliance stock into trend, seasonal, and residual components.

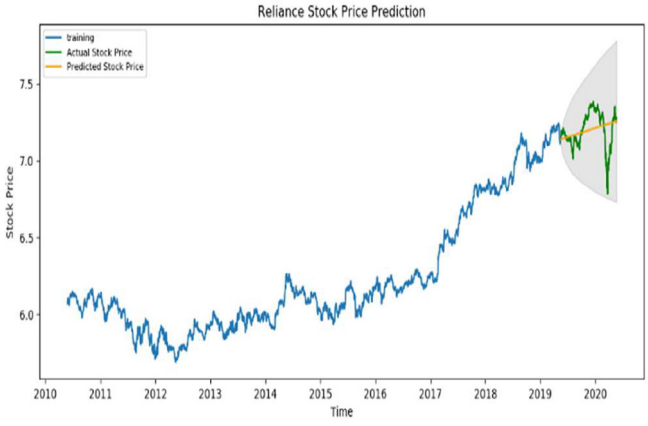

Fig. 15. ARIMA forecasting plot of future values with 95% confidence.

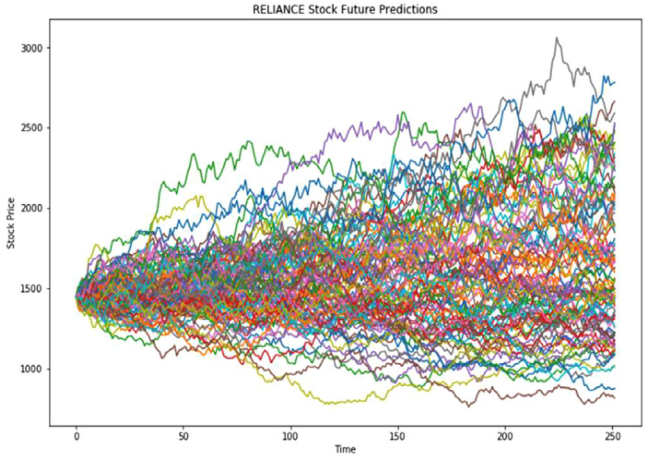

Fig. 16. Monte Carlo simulation of reliance of 1 business year.

risk good amount of return. Sharpe Ratio is a term that is calculated annually, and it changes over time (Table 3).

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.

Table 5
Results of the ARIMA model.

Parameters	Values
MSE	0.013511207
MAE	0.085270415
RMSE	0.116237716
MAPE	0.011900728

6. Price prediction

6.1. Augmented dickey-fuller test

In the AD Fuller test, time-series data is tested for the null hypothesis. The null hypothesis states that a unit root is found in time-series data. An alternate hypothesis depends on whether which type of test is being used on the data, which is usually stationarity or trend-stationarity. For a 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 the more the number is negative, the hypothesis

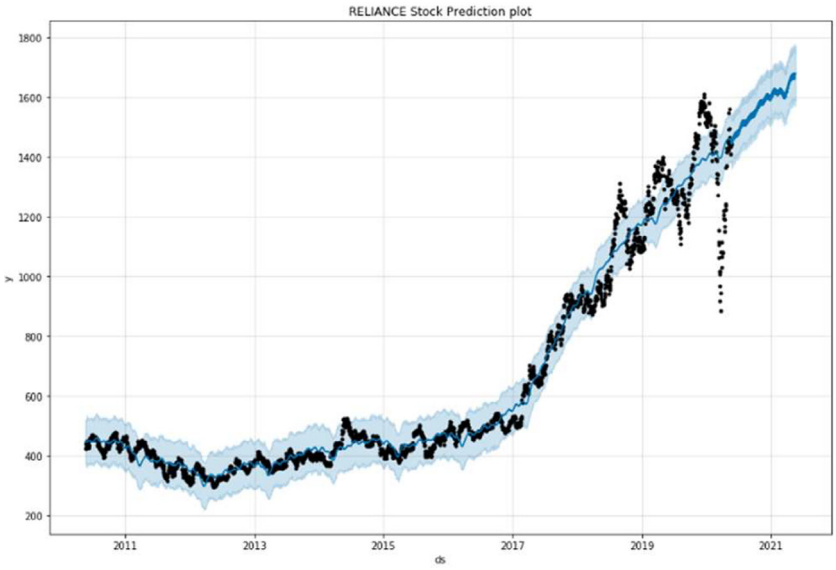

Fig. 17. Prophet forecast of reliance of next one year.

will be rejected more strongly which means that there is a unit root with some level of confidence. Results of the AD fuller test are given in Table 4. p-value > 0.05 implies data is non stationary.

6.2. 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. Fig. 13).

6.3. Seasonal decomposition of time series

The decomposition of Time series data into Seasonal, Trend, and Residual component is known as the seasonal decomposition (Fig. 14) [11].

6.4. ARIMA model

6.4.1. Auto-Arima function

Auto Arima function does the job of finding the best possible values of p, d, q for our model. The model with the 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, the "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 a 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. Fig. 16. Table 5).

6.5. Monte carlo simulation

There are situations when a problem comprises of random variables which cannot be predicted easily then the probability of different possible outcomes are modeled, this is known as Monte Carlo simulation. This method is useful in understanding the

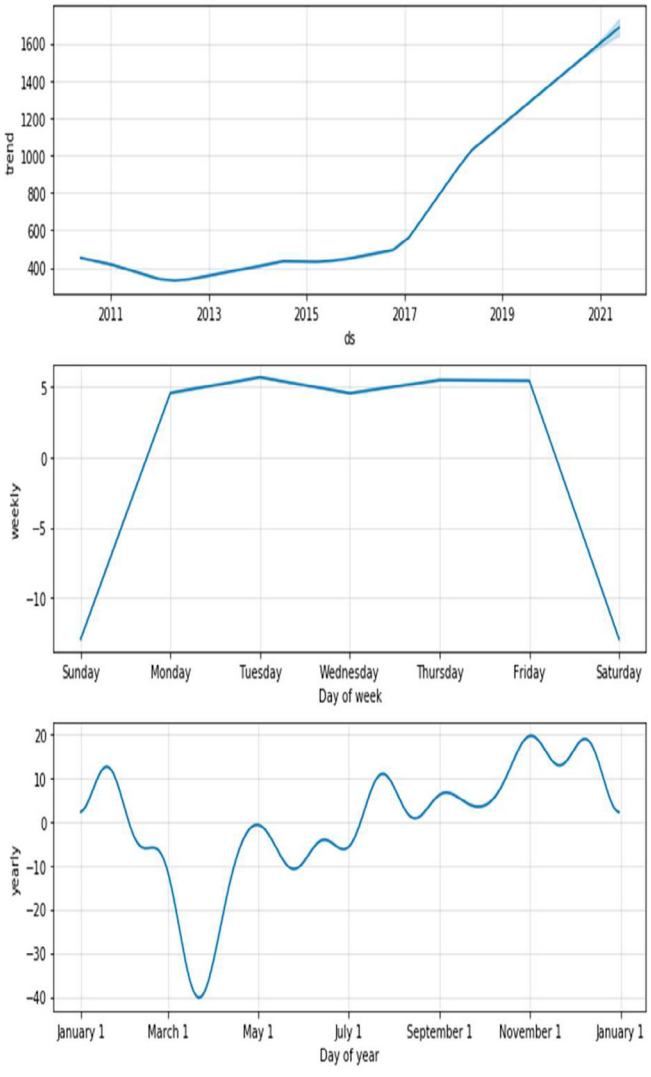

Fig. 18. Decomposition into a daily trend, weekly trend, and yearly trend.

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 to as multiple probability simulation models.

6.6. Fbprophet for forecasting

Fbprophet is a very handy library made by Facebook for Time Series data. Here, The Blue line in the 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 period. Using prophet data can be properly distributed into an overall trend, weekly trend, and yearly trend. Here we can observe a yearly trend that has some seasonality (Fig. 17. Fig. 18).

7. Conclusion

In this study, we could analyze and understand the risk involved in the stock price and are well accounted for 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.

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

CRediT authorship contribution statement

Christy Jackson J.: Conceptualization, Data Curation, Writing Original Draft. **Prassanna J.:** Formal Analysis. **Abdul Quadir Md.:** Methodology. **Sivakumar V.:** Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] D. Banerjee. (2014, January). Forecasting of Indian stock market using the time-series ARIMA model. In 2014 2nd International Conference on Business and Information Management (ICBIM) (pp. 131-135). IEEE.
[2] N. Viswam, G.S. Reddy. (2018). Stock market prediction using time series analysis.
[3] M.C. Angadi, A.P. Kulkarni, Time series data analysis for stock market prediction using data mining techniques with R, Int. J. Adv. Res. Comput. Sci. 6 (6) (2015).
[4] B.U. Devi, D. Sundar, P. Alli, An effective time series analysis for stock trend prediction using the ARIMA model for nifty midcap-50, Int. J. Data Min. Knowl. Manage. Process 3 (1) (2013) 65.
[5] A. Varghese, H. Tarhen, A. Shaikh, P. Banik, A. Ramadasi, Stock market prediction using time series, Int. J. Recent Innov. Trends Comput. Commun. 4 (5) (2016) 427–430.
[6] A. Sharma, S. Modak, E. Sridhar. (2019). Data Visualization and Stock Market and Prediction.
[7] S. Selvin, R. Vinayakumar, E.A. Gopalakrishnan, V.K. Menon, K.P. Soman, in: Stock price prediction using LSTM, RNN, and CNN-sliding window model, IEEE, 2017, pp. 1643–1647.
[8] P. Mondal, L. Shit, S. Goswami, Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices, Int. J. Comput. Sci. Eng. Appl. 4 (2) (2014) 13–29.
[9] J. Wang, J. Wang, Forecasting stock market indexes using principle component analysis and stochastic time effective neural networks, Neurocomputing 156 (2015) 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] J. Brownlee. (2020). What Is Time Series Forecasting? Retrieved 22 May 2020, from https://machinelearningmastery.com/time-series-forecasting/.