Stock Market Index Prediction of SBI in India using ARIMA Models

T. Sartitha [1]

Under Supervision of

Dr. M. Raghavender Sharma [2]

[1] Ph.D. Research Scholar, Career Point University, Kota, Rajasthan
[2] Department of Statistics, Osmania University

Abstract

The work presented in this paper constitutes a contribution to modelling and forecasting or predicting the Stock prices of state bank of India using Box-Jenkins Auto Regressive integrated moving average models. The adequate model is selected according to the performance criterion such as Akaike information criterion (AIC), Schwartz Bayesian criterion (SBC). The final selected model is ARIMA (0,1,0) × (1,0,0)5 and it is validated by another historical stock prices data under the same conditions. The model performance is on training and test samples measured using mean Absolute Error (MAE), mean square error (MSE), mean absolute and percentage error (MAPE), Root Mean square error (RMSE).

Keywords: FFNN, Feed Forward Neural Network, MAPE, Mean Absolute Percentage Error, RMSE, Root Mean Square Error, MAE, Mean Absolute Error

1. Introduction

State Bank of India (SBI) is formed as an Indian multinational; public sector bank and it headquarter is located in Mumbai. The stock prices observed of SBI in India on 2-January-2017 is Rs. 243.712, and on 17-July-2019 is Rs. 371.92. The decreasing trend observed data from 2-January-2019 is [338.91], to 22-May-2020 is 151.05, and later continuously increasing till 29-October-2021 is 502.14.

Objectives

- Forecasting stock prices in India using Box-Jenkins methodology in time series.
- Compare forecasting and prediction models using error measures (MAPE, MAE, RMSE) etc.
- Develop forecasting models for different time series data
- Estimate the effect of volatility and forecast of banking sector with best suitable methods.
- Develop neural network models for forecasting time series.

2. Materials and Methods

A historical data of daily stock prices is collected from Bombay Stock Exchange (2-January-2017 to 29-October-2020). The Box-Jenkins ARIMA models developed and discussed in this paper for modelling and forecasting the daily stock prices. R-software, Ms Office Excel are used for statistical analysis, modelling and forecasting and predicting stock prices.

https://doi.org/10.5281/zenodo.7466293

2.1. Methodology

The Box-Jenkins is the most popular method for to get the accurate model by build the auto regressive integrated model for historical data sets. Box-Jenkins method consists of multiple advantages, for obtaining the minimum number of parameters in a model like p and q and to verify the data pattern and identify the data is stationary or non-stationary. The Box-Jenkins method consists of mainly four steps to build the model. They are identification, estimation, diagnostic checking, and forecasting. The model identification is the first step to check the parameters in a model p and q by using the auto correlation that is ACF and partial auto correlation plots that is (PACF). The auto correlation function dies out for more lags and q spikes in the auto correlation plots, then this is parameters q and PAC. F dies out more lags and p spikes in the PACF plot, then this is p parameter. By using ACF, we have to identify p value, by using PACF we have to identify q values. Once the parameter estimation is done, we will start checking the accuracy and adequacy that is diagnostic checking by using SBC and AIC and L-Ljungberg statistic. The main use of diagnostic checking is to get the model adequacy using the Ljung-Box Q statistics test and verify the assumptions with reference to error are random. Repeat the process until model parameters should be significant and they are adequate. Test the several tentative models for identifying the parameter significant and adequate for the given data set and selected model for forecasting the daily stock prices in India according to minimum error measures like and MAE (mean absolute error). MAPE (mean absolute percentage error) and RMSE (root mean square error).

3. Literature on Forecasting Stock Price of SBI

R. Tan, Q. Tan, P. Zhang and Z. Li (2021)

They provided a technique for mining Ethereum-based transaction records to find Ethereum scams. In particular, web crawlers are utilised to collect addresses that have been flagged as fraudulent, after which a transaction network is formed using the public transaction book. Then, a network embedding approach based on quantity is suggested in order to extract node attributes for detecting fraudulent transactions. Addresses are finally divided between legitimate and fraudulent addresses using the graph convolutional network model. The experimental findings demonstrate that the system for identifying fraudulent transactions may reach an accuracy of 95%, demonstrating the system's high performance in identifying fraudulent Ethereum transactions.

Chaleampong Kongcharoen and Tapanee Kruangpradit (2013)

They forecasted the Thailand ports to major trade partners. This paper compares the autoregressive integrated moving average (ARIMA) and ARIMA with explanatory variable (ARIMAX). The study concluded that the ARIMA model with leading indicator outperforms the ARIMA model.

Matheus Henrique Dal Molin Ribeiroa, Leandrodos Santos Coelho (2020)

Their study aimed to compare the predictive performance of the GBM, XGB, RF and STACK regression ensembles, as well as the MLP and VAK models. Two case studies related agri-business namely case 1 - the soybean price, and case 2 - the wheat price, paid to the producer from the state of Parana, for short-term forecasting. The results, regarding the test set, allows to ensure that the ensemble approaches perform better than single models, especially the XGB model, to forecast agricultural commodities in the short-term.

Saha, D. Sarma, R.J. Chakma, M.N. Alam, A. Sultana, S. Hossain (2020)

To entice consumers, phishers create bogus websites that seem just like the real thing and send spam emails. When a person accesses the fake websites via spam, phishers steal their login information. To identify phishing websites, researchers have developed powerful technologies like antivirus software, whitelists, and

https://doi.org/10.5281/zenodo.7466293

blacklists. Attackers always come up with inventive strategies to exploit network and human vulnerabilities and get past stock market protection. Additionally, a data-driven system for phishing webpage detection with a deep learning technique was demonstrated. To be more explicit, the phishing webpages are predicted using a multilayer perceptron, commonly known as a feed-forward neural network. The dataset, which includes data from 10,000 websites, was gathered through Kaggle. It has 10 different qualities. The suggested model exhibits 93% test accuracy and a 95% training accuracy.

Ray, B. Ganguli, A. Chakrabarti (2021)

They offered an application of the Bayesian structural time (BST) series model that is more transparent and facilitates better handling of uncertainty than the autoregressive integrated moving average (ARIMA) model and the vector autoregression (VAR) method by using prior information about the structure of the model. One of the main pitfalls of this model is the presumption of linearity. The long short-term memory (LSTM) model is a nonlinear model that can capture various nonlinear structures present in the data set. We propose a hybrid model, which combines the LSTM model with the BST model along with the regression component that captures information from different news sources to identify market predictors. The proposed model detects unusual behaviour or anomalous pattern of the stock price movement, which makes our model superior compared to the traditional methods. Our new hybrid model accumulates error with lower rates (3.5%) and shows a remarkable performance over some of the other existing hybrid models, such as AR-MLP, ARIMA-LSTM, and VAR-LSTM model.

K. Iqbal, A. Hassan, S.S.M.U. Hassan, S. Iqbal, F. Aslam, K.S. Mughal (2021)

They employed historical stock price data S&P500 (Daily Prices) from Yahoo's financial website and six months dataset of State Bank of Pakistan (Hourly values). Different prediction models have been tested for the S&P500 dataset which is publicly available and after finding out that the proposed model performed well it has been applied to the SBP dataset as well. The effectiveness of the proposed model has been calculated based on the following performance metrics, root means square error (RMSE), mean absolute error (MAE), mean square error (MSE), and mean absolute percentage error (MAPE). When compared to other comparable studies, the experimental findings indicate that the proposed model has the best performance metrics values. As a result, we can infer that our model is appropriate for accurate stock market time series prediction.

K. Tamersit, F. Djeffal (2019)

Numerical simulation of quantum-mechanical approaches with graphene nanoribbon field-effect transistors, sometimes synonymous with severe computational burdens, indicates the immediate need for new methods to solve the problem. A hybrid methodology for ballistic GNRFET quantic simulations is built in this context. "The suggested simulation approach is based on the resolution, in a two-dimensional (2D) Poisson equation, of the non-balances Green's function (NEGF) producing load density, utilizing both adaptive meshing based on a wavelet and matrix-based compression methods. The findings of the hybrid method are compatible with the NEGF simulations. Numerical tests show that the evolved simulation method will accelerate the traditional MS NEGF approach by approximately one order of magnitude. The encouraging results obtained in this work indicate that the built numerical model especially fits the integration of future GNRFET-based systems in nano-electronic computer simulators.

Wide Li (2017)

In energy markets and electricity networks, accurate electricity demand forecasts play a key role. For different unexplained causes, the need for energy is typically a non-linear challenge, which render

mitigation by conventional approaches more complicated. The goal of this paper is to propose a new hybrid prediction approach for energy management and preparation. Ensemble empire mode decomposition (EEMD), seasonal change (S), cross validation (C), general neural regression (GRNN) and supporting vector regression machine, the EEMD-SCGRNN-PSVR, the latest approach suggested (PSVR). The EEMD-SCGRNN-PSVR's concept is to anticipate waveforms and patterns that are concealed in demand sequence to replace original electrical demand directly. EEMD-SCGRNN-PSVR is used in two data sets (New South Wales (NSW) and Victorian State (VIC) in Austria) to forecast a one week ahead of half-hour energy need. Experimental findings reveal that the current hybrid model is beyond predictive precision and model robustness in the other three versions.

Mohamed Ashik (2017)

He was done analysis on national Stock exchange deals in securities like shares or bonds issued by the companies or corporations in the private and public sectors. National Stock Exchange is widest and fully automatic trading system in India. Nifty 50 is one of the stock price investors and 50 companies were invested in the traders. Autoregressive Integrated Moving Average model is one of the most accepted forecasting models and a vital area of the Box-Jenkins approach to time series modelling. In this paper, the Nifty 50 stock market prices were evaluated and predicted the trend of upcoming trading days stock market fluctuations using Box-Jenkins methodology. From the results, it can be observed that influence R-Square value is (94%) high and Mean Absolute Percentage Error is very small for the fitted model. Thus, the prediction accuracy is more suitable of Nifty 50 closing stock price. It is concluded that closing stock price of Nifty 50 taken in the present study shows slow decreasing fluctuations trend for upcoming trading days.

C. Narendra Babu (2014)

Accurate long-term prediction of time series data (TSD) is a very useful research challenge in diversified fields. As financial TSD are highly volatile, multi-step prediction of financial TSD is a major research problem in TSD mining. The two challenges encountered are, maintaining high prediction accuracy and preserving the data trend across the forecast horizon. The linear traditional models such as autoregressive integrated moving average (ARIMA) and generalized autoregressive conditional heteroscedastic (GARCH) preserve data trend to some extent, at the cost of prediction accuracy. Non-linear models like ANN maintain prediction accuracy by sacrificing data trend. In this paper, a linear hybrid model, which maintains prediction accuracy while preserving data trend, is proposed. A quantitative reasoning analysis justifying the accuracy of proposed model is also presented. A moving-average (MA) filter-based pre-processing; partitioning and interpolation (PI) technique are incorporated by the proposed model. Some existing models and the proposed model are applied on selected NSE India stock market data.

4. Results and Discussion

The Figure 4.1 shows the time series plot of daily stock prices of SBI from 2-Jan-2017 to 29-Oct-2021. The data shows continuous increment in the stock price per from 2-Jan-2017 (Rs. 243.70) to 17-July-2019 (Rs. 371.92), the decreasing trend observed from 02-Jan-2019 (Rs. 338.91) to 22-May-2020 (Rs. 151.05), and later continuously increased till 29-Oct-2021 (Rs. 502.14).

https://doi.org/10.5281/zenodo.7466293

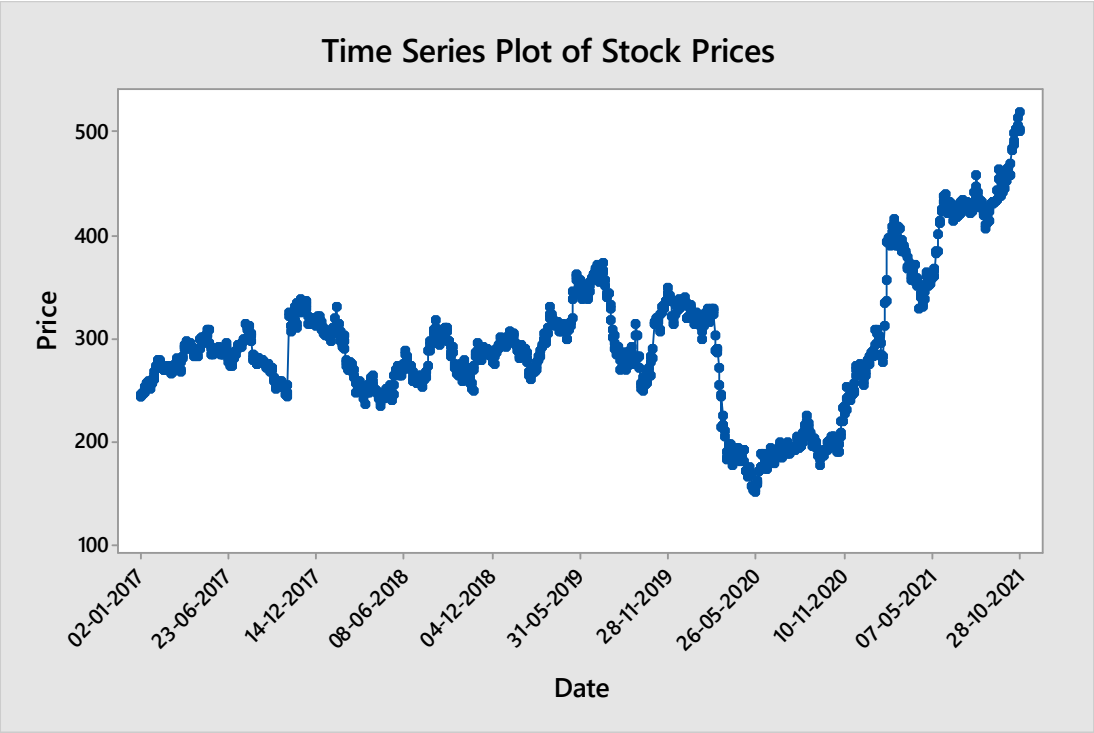


Figure 4.1: Time Series Plot for Daily Stock Prices of SBI

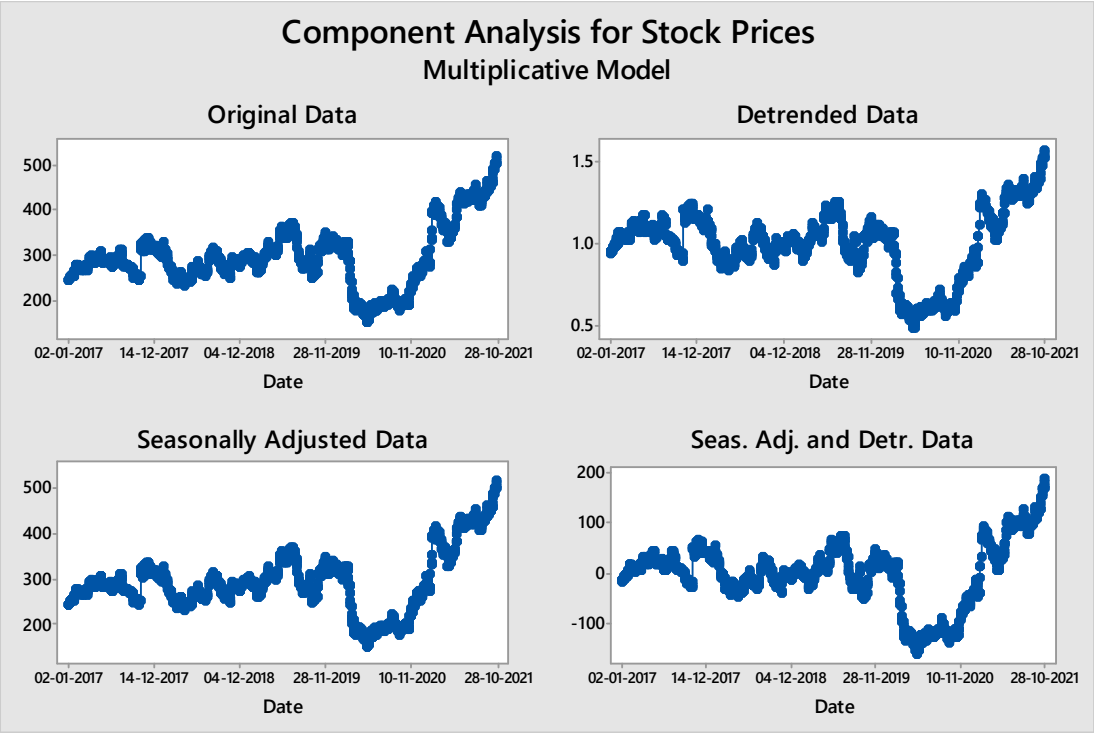
Table 4.1: Descriptive Statistics of the Stock Prices

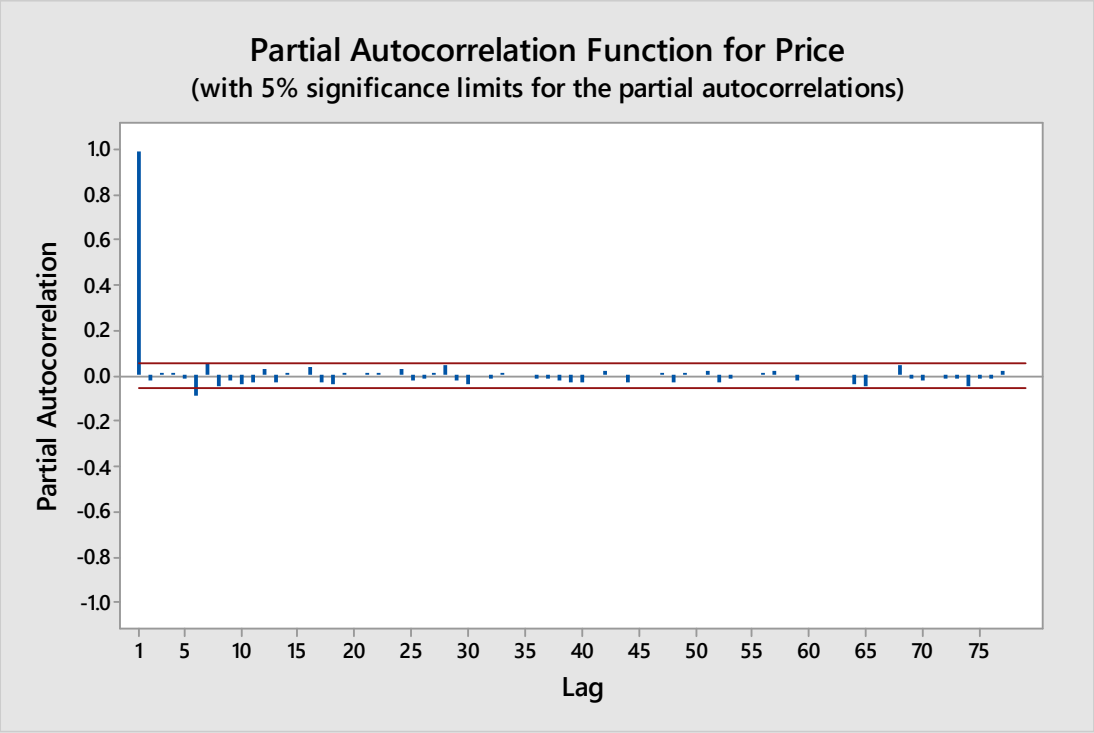
Statistics	Values
N	1191
Mean	295.53
SE Mean	1.9
Std. Dev.	66.98
Variance	4579.42
Minimum	151.55
Q1	258.99
Median	286.95
Q3	323.05
Maximum	518.95

From Table 4.1, it is observed that, stock prices range from Rs. 151.55 to Rs. 518.95, mean is Rs. 294.53, standard deviation is 66.98.

4.1. Seasonal ARIMA Model

The time series data can be decomposed in to trend, seasonal, and random or irregular components. The multiplicative decomposition of stock prices data and ACF and PACF shows in Figure 2.





https://doi.org/10.5281/zenodo.7466293

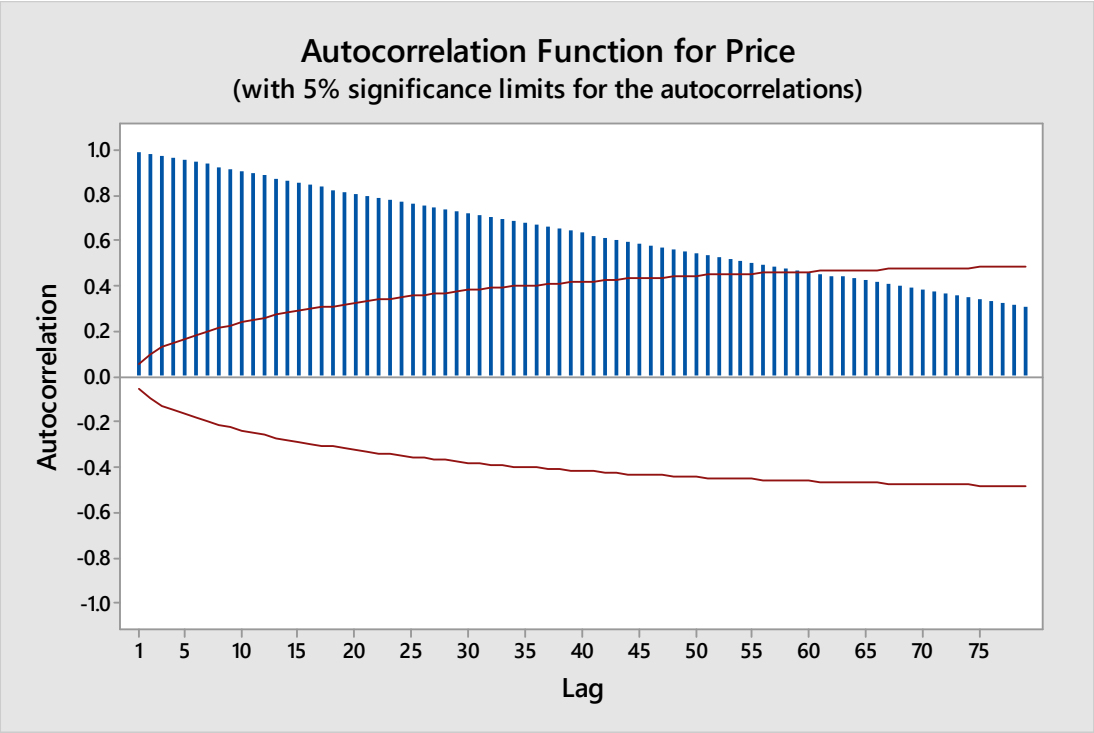


Figure 4.1.1: Daily Stock Prices Data and its Three Additive Components with ACT and PACF Plots

The above Figure 4.1.1 shows the original data, detrended data, seasonally adjusted data, seasonally adjusted, detrended data are shown separately in Figure 2 and it's observed that no seasonality component is present and trend component is present the stock prices data. The whole data was divided as training data (January-2017 to September-2011) and test data (October-2021). The model is to be developed on training data and validate on the test data.

The p-value is 0.01 which is lesser than significance level 0.05, so reject null hypothesis. Augmented Dicky-Fuller test results, the first order non seasonal differenced data is stationary.

Table 4.1.1: Augmented Dicky-Fuller Test

P value	0.01
Lag Order	10
ADF	-9.52

An appropriate ARIMA model should be selected based on ACF and PACF showed in Figure 4.1.1. From ACF and PACF graphs its observed that the order of non-seasonal moving average terms q is '0', the order of seasonal moving average terms Q is at most '1', the order of non-seasonal autocorrelation terms p is '0', the order of seasonal autocorrelation terms P is at most '1', the number of non-seasonal differences 'd' is 1 and the number of seasonal differences 'D' is 0 to be included in the model. The error measures of models are computed for training and test data of all the possible combinations of the parameters p, d, q, P, D, and Q. Select the best model based on Akaike's information criterion (AIC) and Schwartz-Bayesian criterion (SBC). The Ljung-Box test was considered for verifying the adequacy of the model. The results are shown in table 4.1.2.

Table 4.1.2: ARIMA Models and their Significance (Training Data)

Sr. No.	Model	AIC	SBC	Significance of the Parameters	Ljung-Box Statistic	Significance Value (p)	Adequacy
1	ARIMA (0,1,0) × (1,0,0)5	7743.74	7754.18	Significant	244.6	0.2804	Adequate
2	ARIMA (0,1,0) × (1,0,1)5	7744.74	7760.66	Insignificant	244.62	0.2657	Adequate
3	ARIMA (0,1,0) × (0,0,1)5	7744.16	7755.29	Significant	246.1	0.263	Adequate

From Table 4.1.2, it is observed that ARIMA (0,1,0) × (1,0,0)5 model is significant with respect top, d, q parameters as well as adequacy of the model comparing and has smaller AIC (7743.74) and SBC (7754.18) values as compared with other models. So, the most suitable model is ARIMA (0,1,0) × (1,0,0)5. The estimated parameters of the model are presented in the table 4.1.3.

Table 4.1.3: Parameters of ARIMA (0,1,0) × (1,0,0)5 Model

Parameter	Estimated	S.E.	Z value	Significance Value (p)	Remarks
SAR1	0.1083	0.0290	3.72	0.0001	Significant

The ARIMA model (0,1,0) × (1,0,0)5 model equation is:

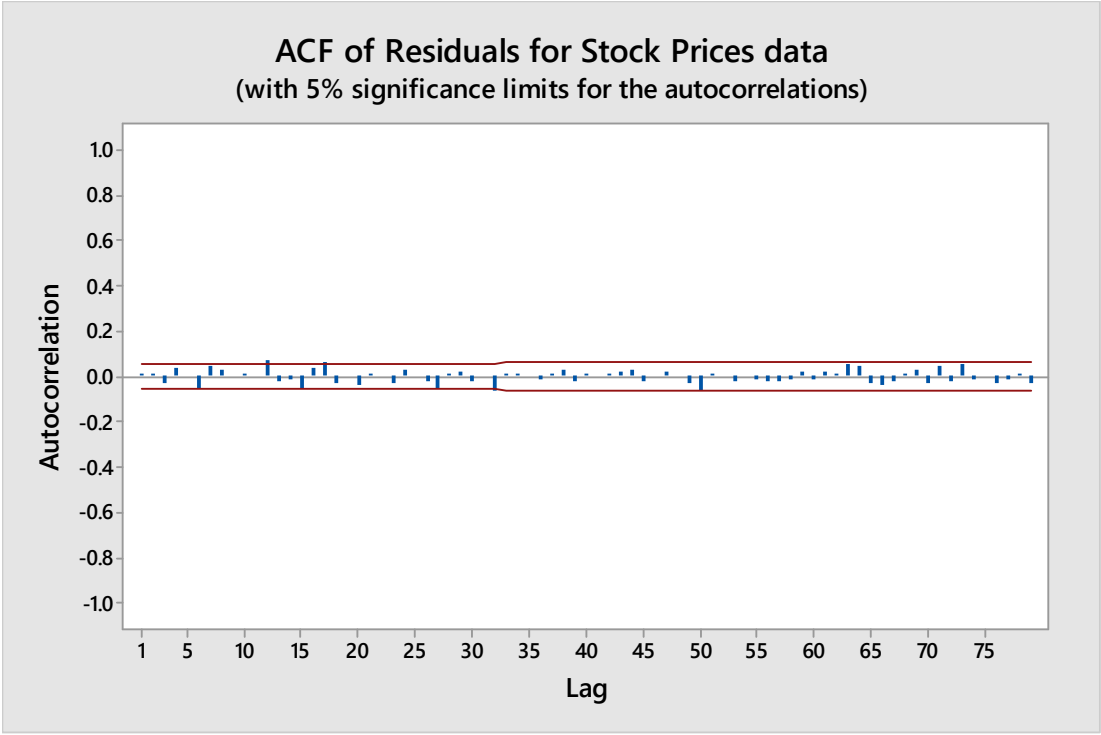
$$(1 - \Phi 1B5)(1 - B)\, Yt = \varepsilon t$$

The Fitted ARIMA model (0,1,0) × (1,0,0)5 model equation is

$$(1 - 0.1084 \times B5)(1-B)\, Yt = \varepsilon t$$

The diagnostics plots are verified using of ACF and PACF of the multiple views.

https://doi.org/10.5281/zenodo.7466293



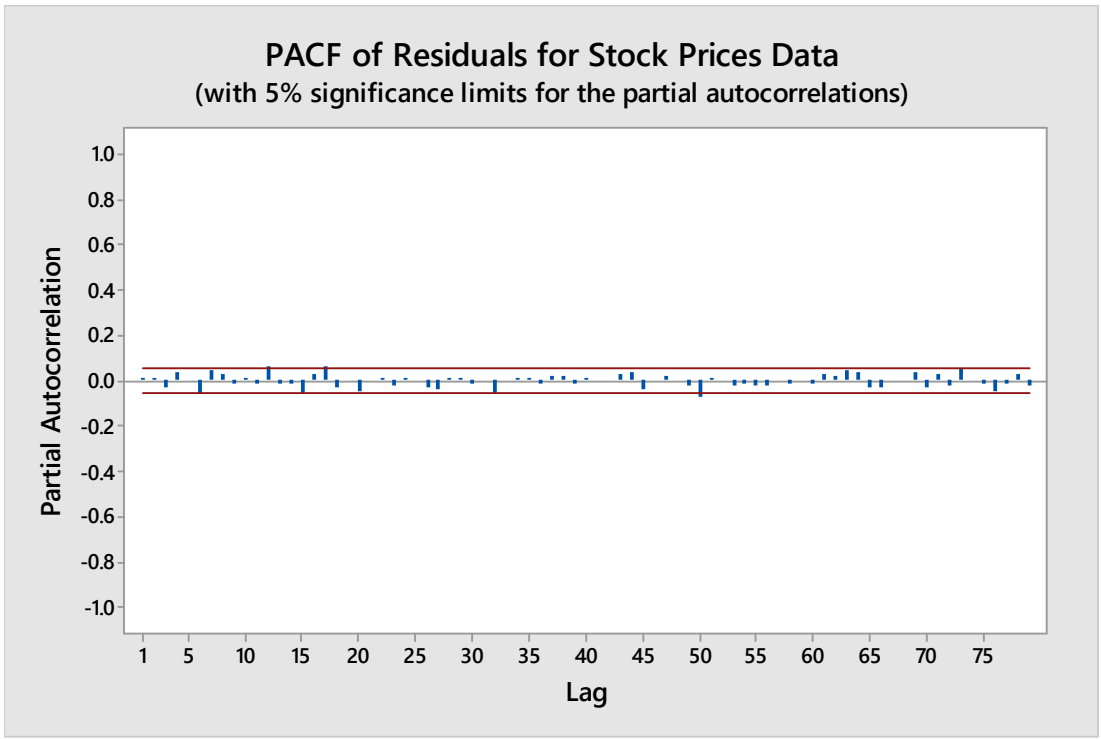


Figure 4.1.2: ACF and PACF Residuals Series for Stock Prices

Diagnostic or checking is done by analysing ACF and PACF (auto correlation, partial auto correlation) graphs of the residuals. From above figures, it's observed that none of the autocorrelations are significant at 0.05 level. Therefore, the model is appropriate.

The Daily Forecasts of Stock Prices of SBI from 01-October-2021 to 29-October-2021

The point forecasts and 95% confidence interval values are obtained using ARIMA (0,1,0) × (1,0,0)5 models are presented in below Figure 4.1.3.

Table 4.1.4: Forecasts of Stock Prices using ARIMA (0,1,0) × (1,0,0)5 Model

Date	Original Values	Point Forecasted Values	95% Confidence Intervals
01/10/2021	451.63	451.52	439.05-464.98
04/10/2021	463.12	451.65	434.32-470.99
05/10/2021	463.71	451.47	430.01-474.92
06/10/2021	458.20	455.11	428.18-480.04
07/10/2021	458.89	454.35	424.36-482.34
08/10/2021	459.10	456.24	420.88-485.60
10/10/2021	470.25	455.31	417.90-488.72
12/10/2021	482.90	455.29	415.08-491.51
13/10/2021	480.90	455.47	412.64-494.30
14/10/2021	491.60	457.39	410.10-496.68
18/10/2021	496.95	453.48	407.71-499.04
19/10/2021	489.20	453.38	405.47-501.30
20/10/2021	498.89	453.38	403.31-503.45
21/10/2021	501.95	453.40	401.26-505.54
23/10/2021	503.95	453.39	399.27-507.52
25/10/2021	505.80	453.39	397.34-509.44
26/10/2021	513.54	453.39	395.49-511.29
27/10/2021	518.95	453.39	393.69-513.09
28/10/2021	502.34	453.39	391.94-514.84
29/10/2021	502.94	453.39	390.24-516.54

https://doi.org/10.5281/zenodo.7466293

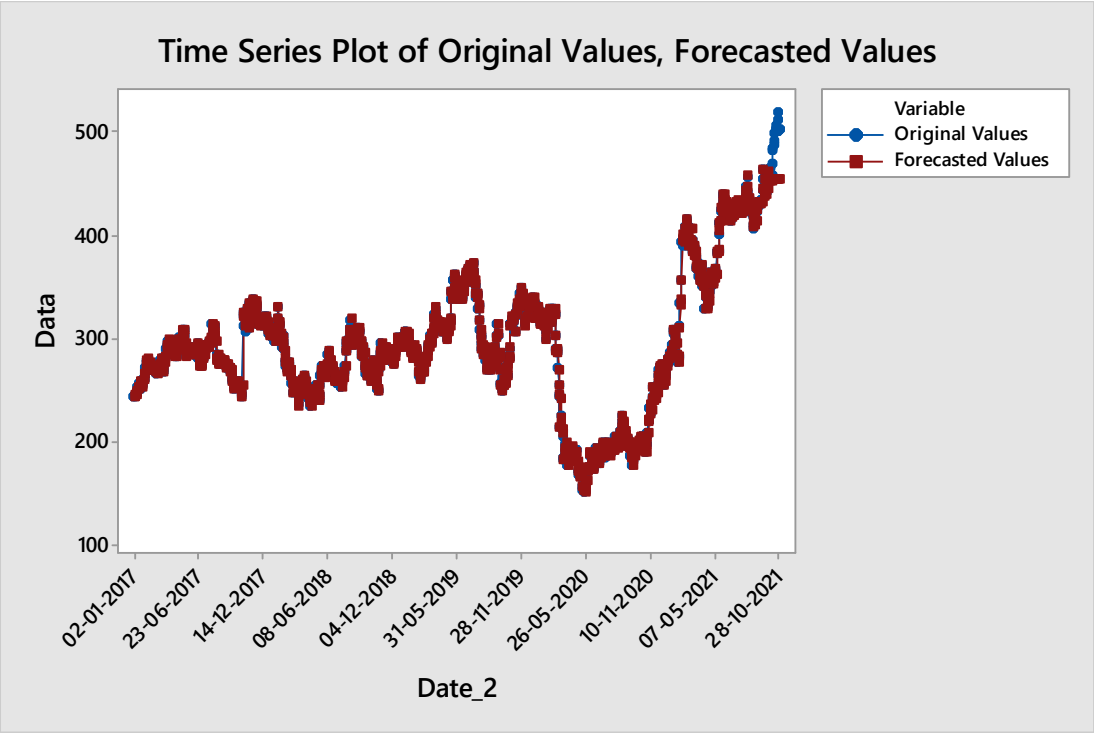


Figure 4.1.3: ARIMA Model with Actual and Forecasted Values

Table 4.1.5: Error Measures of ARIMA (0,1,0) × (1,0,0)5 Model

Sample	RMSE	MAE	MAPE
Training	6.59	4.49	1.59
Test	37.92	32.11	7.02

From Table 4.1.5, It is observed that ARIMA (0,1,0) × (1,0,0)5 model has insignificant error measures on training and test data. Therefore ARIMA (0,1,0) × (1,0,0)5 is the best model for forecasting the daily stock prices.

5. Conclusions

The above results give the stock prices in India and Box-Jenkins provides a good approach for predicting the future stock prices. The suggested technique is used on stock prices in India data to check the prediction accuracy based on mean absolute percentage error and root mean square error. In this analysis, ARIMA (0,1,0) (1,0,0)5 model is built for analysing the prediction for stock prices in India among all tentative approaches as it have minimum AIC and SBC and also verified parameter significance along with adequacy. From all the above results, it can be concluded that the stock prices take present analysis gives slow increasing trend for future trading days.

6. References

1. Batten, J.A., Ciner, C. and Lucey, B.M. (2010). The macroeconomic determinants of volatility in precious metals markets. Resources Policy, 35, pp. 65–71.
2. Brown, R.G. (1959). Statistical forecasting for inventory control. New York: McGraw Hill.
3. Sjaastad, L.A. and Scacciavillani, F. (1996). The price of gold and the exchange rate. Journal of International Money and Finance, 15(6), pp. 879-897.

4. Box, G.E.P., Jenkins, G.M. and Reinsel, G.C. (1994). Time Series Analysis Forecasting and Control. 3rd ed., Englewood Cliffs, N.J. Prentice Hall.
5. Ramakrishna, R., Naveen Kumar, B. and Krishna Reddy, M. (2011). An International Journal of Business & Economics, 10(1) (2015), pp. 22-47.
6. Anderson, B. and Ledolder, J. (1983). Statistical Methods for Forecasting, John Wiley & Sons, New York.
7. Pai, P.F. and Lin, C.S., 2005. A hybrid ARIMA and support vector machines model in stock price forecasting. Omega, 33(6), pp.497-505.
8. Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31(3), pp. 309-317.
9. Box, G.E.P. and Jenkins, G. M. (1970). Time Series Analysis: Forecasting and Control. Holden Day, San Francisco.
10. Zazzaro, G., Romano, G. and Mercogliano, P. (2017). Data Mining for Forecasting Fog Events and Comparing Geographical Sites. Designing a novel method for predictive models' portability. International Journal on Advances in Networks and Services, 10, pp. 160-171.
11. Box, G.E.P. and Pierce, D. (1970). Distribution of Residual Autocorrelations in Autoregressive Integrated Moving Average Time Series Models. Journal of the American Statistical Association, 65, pp. 1509-1526.
12. Brockwell, P.J. and Davis, R.A. (2002). Introduction to time series and forecasting. 2nd edition, Springer-Verlag, New York.
13. Javier, C., Rosario, E., Francisco J.N. and Antonio, J.C. (2003). ARIMA Model to predict Next Electricity Price. IEEE Transactions on Power Systems, 18(3), pp. 1014-1020.
14. Jawahar Farook, A. and Senthamarai Kannan, K. (2014). Stochastic Modelling for Carbon Dioxide Emissions. Journal of Statistics & Management Systems, 17(1), pp. 92-117.
15. Devi, B.U., Sundar, D. and 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), p. 65.
16. Jun Zhang, Rui Shan and Wenfang Su. (2009). Applying Time Series Analysis Builds Stock Price Forecast Model. Modern Applied Science, 3(5), pp. 152-157.
17. Nail, P.E. and Momani, M. (2009). Time Series Analysis for Rainfall Data in Jordan: Case Study for Using Time Series Analysis. American Journal of Environmental Science, 5(5), pp. 599-604.
18. Renhao Jinn, Sha Wang, Fang Yan and Jie Zhu. (2015). The Application of ARIMA Model in 2014 Shanghai Composite Stock Price Index. Journal of Applied Mathematics and Statistics, 3(4), pp. 199-203.
19. Walder Enders. (2014). Applied Econometric Time Series. 4th edition, Wiley.
20. Young H. Kim, Edward L. Davis and Charles T. Moses. An ARIMA Model Approach to the Behaviour of Weekly Stock Prices of Fortune 500 Firms and S&P Small Cap 600 Firms, Oxford.
21. Si, W., Li, J., Ding, P., Rao, R. A multi-objective deep reinforcement learning approach for stock index future's intraday trading. In Proceedings of the 2017 10th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 9–10 December 2017; Volume 2, pp. 431–436.
22. Song, Y., Lee, J. (2020). Importance of Event Binary Features in Stock Price Prediction. Appl. Sci., 10, p. 1597.
23. Thakkar, A., Chaudhari, K. (2021). Fusion in stock market prediction: A decade survey on the necessity, recent developments, and potential future directions. Inf. Fusion, 65, pp. 95–107.
24. Zhang, M., Jiang, X., Fang, Z., Zeng, Y., Xu, K. (2019). High-order Hidden Markov Model for trend prediction in financial time series. Phys. A Stat. Mech. Its Appl., 517, pp. 1–12.

https://doi.org/10.5281/zenodo.7466293