Predicting Long Term Prices of Nifty Index using Linear Regression and ARIMA: A comparative study

Mohit Beniwal
Delhi School of Management, Delhi Technological University, Delhi, India
Email: mohitbeniwal@dtu.ac.in

Archana Singh
Delhi School of Management, Delhi Technological University, Delhi, India
Email: archanasingh@dtu.ac.in

Nand Kumar
Department of Humanities, Delhi Technological University, Delhi, India
Email: nandkumar@dce.ac.in

Abstract

Stocks are traded continuously in the financial market, creating huge time series data. Stock market time series is very volatile and highly complex to model. There are many methods to forecast a time series. This study predicts and compares the performance of two statical methods, linear regression, and ARIMA also referred to as the Box Jenkins method, which stands for Autoregressive Integrated Moving Average, is a robust time series forecasting technique used frequently by researchers. Simple linear regression is generally assumed unsuitable for non-linear stock time series data. However, a literature gap exists in comparing linear regression and the ARIMA method for forecasting stock prices. Hence, this study compares the ARIMA and linear regression methods to forecast stock prices using daily and weekly NIFTY data. Further, this study also experiments using a different length of time series, namely 1-year and 2-year data. The result shows that ARIMA outperformed regression on daily and weekly data when the test for 1 year of data. When 2-year data is taken, linear regression outperformed ARIMA on both daily and weekly data. Hence, the linear regression model and ARIMA are sensitive to input parameters such as the number of days for training and forecasting. An automated method or algorithm could improve the robustness of the model. Further, as stock price data is non-linear, machine learning algorithms such as neural networks and support vector regression can be more suitable for prediction.

Keywords*: Linear Regression; ARIMA; Nifty; Prediction; Forecasting.*

Suggested Citation:
Beniwal, M., Singh, A. & Kumar, N. (2022). Predicting Long Term Prices of Nifty Index using Linear Regression and ARIMA: A comparative study. *International Journal of Accounting, Business and Finance,* 2 (1), 31-41.

1. Introduction

A stock market is a voting machine in the short term, but in the long term, it is a weighing machine (Graham & David, 1965). Time series analysis has an essential role in forecasting in various domains, including sales forecasting, inventory management, and financial market analysis. Stock market forecasting is a branch of economic forecasting (Du, 2018). Effective forecasting can be very fruitful for traders and investors. Many investors rely on fundamental and technical analysis for stock price prediction. However, the stock market is influenced by various factors, including political stability, macroeconomic data, and financial news. It makes the stock market noisy, complex, volatile, and risky. Widely accepted hypotheses such as the efficient market hypothesis (Fama, 1970; Malkiel, 2003; Timmermann & Granger, 2004) and the random walk hypothesis (Fama, 1995; Malkiel, 1973) argues that stock prices cannot be predicted. According to these theories, prices move randomly and immediately include all public and private information. Hence, it is futile to predict the stock market. On the other hand, some studies (Fama & French, 1988; Jegadeesh & Titman, 1993;

Lo & MacKinlay, 1998; Poterba & Summers, 1988) argue that there are some predictabilities in the stock market.

There are many forecasting techniques, which can be categorized into statistical and machine learning. This study investigates two statistical techniques, simple linear regression and Autoregressive integrated moving average (ARIMA). Both these techniques are widely used for forecasting purposes. Linear regression analyzes the relation between two variables. In the case of the stock market, these variables are prices and time. Linear regression provides prediction in terms of continuous values. ARIMA, also known as the Box-Jenkins model or methodology, is a widely applied statistical technique. ARIMA is more advance than linear regression in analyzing time series.

The study's objectives are to evaluate the performance of these two statistical techniques by using Nifty index data in different time frames, namely daily and weekly time frames, and different sample sizes, i.e., 1-year and 2-year data. The findings of the study indicate that linear regression effectively captures the overall trend when using a longer time frame. However, it falls short in capturing the price fluctuations. On the other hand, ARIMA, which is widely recognized as a robust model for time series analysis, also struggles to capture both aspects.

The structure of the paper is as follows: Section 2 provides a literature review. The methodologies employed are described in Section 3, while Section 4 presents the obtained results. In Section 5, we present our concluding remarks. Finally, limitations and directions for future research are discussed in Section 6.

2. Related Works

Generally, a linear model of forecasting is assumed to be unsuitable for non-linear time series such as the stock market. Bhuriya et al. (2017) predicted the Tata Consultancy Services (TCS) share listed on National Stock Exchange (NSE) using linear regression. The independent variables in the study consisted of Open price, High price, Low price, and the number of trends, while the dependent variable was the Close Price. They compared their prediction result with the prediction result of Polynomial regression and Radial Bias Function (RBF) regression. They found that linear regression was better than other compared regression models. This study did not compare the performance of linear regression with ARIMA.

In a study conducted by Roy et al. (2015), a Least Absolute Shrinkage and Selection Operator (LASSO) method based on a linear regression model was employed to predict the stock of Goldman Sachs Group Inc. The study used a sample that included Open, High, Low, and Close prices as features. The performance of the LASSO method was compared with that of the Ridge method and a Bayesian regularized artificial neural network using Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). The results indicated that the LASSO method outperformed the Ridge method; however, the study did not compare the results with those of simple linear regression.

Linear Regression and ARIMA are both linear models. Du (2018) predicted the Shanghai Securities Composition stock index using hybrid ARIMA and Back Propagation Neural Network (BPNN). BPNN is a capable machine learning algorithm to model non-linear time series. They compared the accuracy of a single ARIMA and single BP neural network method. The result showed that the hybrid ARIMA-BP neural network performed better than the single BPNN. Moreover, the single BPNN performed better than the single ARIMA model. This result was expected as machine learning algorithms such as neural networks or Support Vector Machines (SVM) are robust algorithms for non-linear model data.

ARIMA model is a popular linear modeling method for time series data. (Khan & Alghulaiakh, 2020) experimented with multiple ARIMA parameters and evaluated the models using Mean Absolute Percentage Error (MAPE). They conclude that ARIMA has the potential for accurate time series stock forecasting. Combining the ARIMA model with other models, such as support vector machine, has the potential to enhance its performance. (Rubio & Alba, 2022) forecasted the selected Colombian shares using a hybrid ARIMA-SVR model. They evaluated the performance using MAE, MSE, and RMSE. Their result confirmed that ARIMA-SVR generated the smallest error for most of the shares.

To the best of our knowledge, there is no study comparing the performance of simple linear regression and ARIMA to forecast stock market time series. This study also investigates the impact of changing the timeframe of the stock time series. Further, this study compares the performance of the model in different time frames by measuring RMSE and MAPE.

3. Methodologies

The data for this study is obtained from Yahoo Finance. The NIFTY 50 index is used for analysis. The Nifty 50 is a stock market index in India comprising 50 of the top publicly traded companies listed on the National Stock Exchange of India (NSE). The index is designed to provide a benchmark for investors to gauge the performance of the Indian stock market.

3.1. Data Description

The daily data consists of Open, High, Low, and Close prices. There are four sets of data: 1 year of daily and weekly data from 1-Jan-2022 to 31-Dec-2022 and 2 years of daily and weekly from 1-Jan-2021 to 31-Dec-2022. For both models, the close price is chosen as the dependent variable, and the date time index is the independent variable. **Error! Reference source not found.** and **Error! Reference source not found.** show the chart of 1-year daily and weekly close prices with respect to date. Similarly, **Error! Reference source not found.** and **Error! Reference source not found.** show 2-year daily and weekly closing prices.

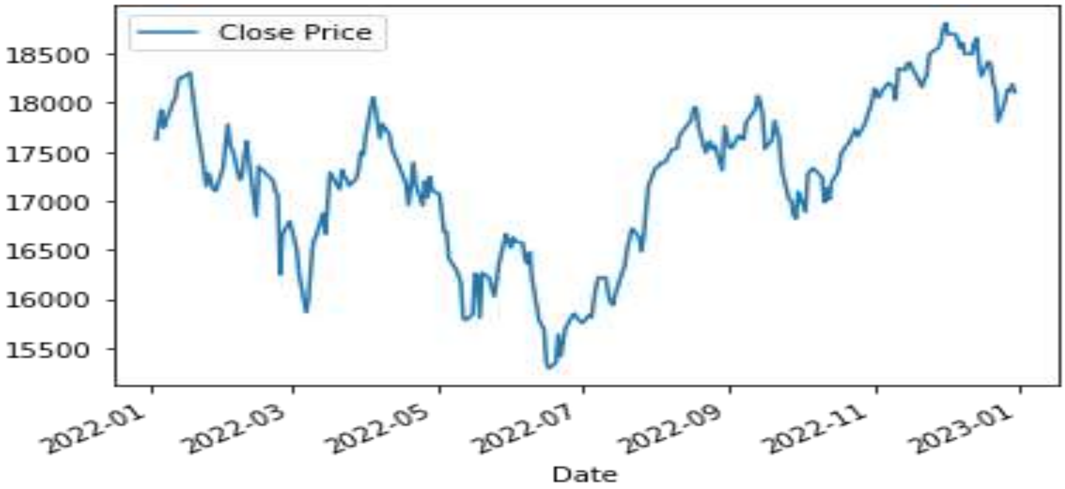

Figure 1. One-year daily close price

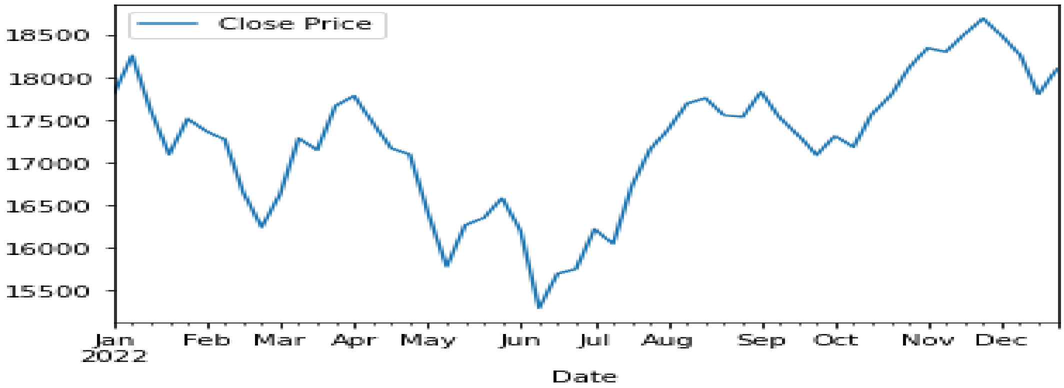

Figure 2. One-year weekly close price

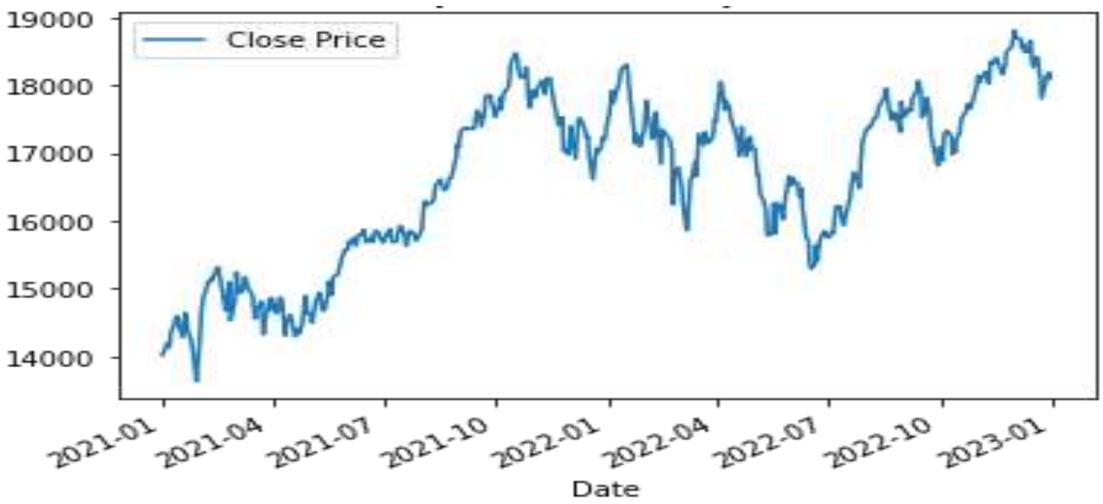

Figure 3. Two-year daily close price

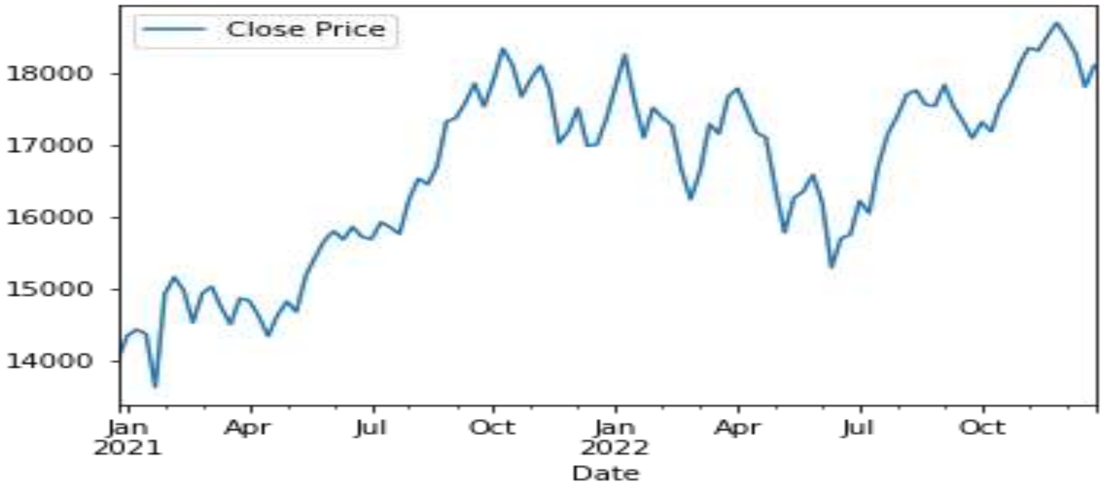

Figure 4. Two-year weekly close price

The parameter of ARIMA, namely, p, d, and q, are estimated using minimum AIC criteria. The Augmented Dickey-Fuller test is used to check the time series stationarity for the ARIMA model. Finally, models are evaluated based on RMSE and MAPE.

3.2. Linear Regression

Regression forecast continuous numerical values. Linear regression is the simplest model, which establishes a linear equation to model the relationship between two variables. It is called "linear" because the model makes a prediction based on a linear combination of the input variables. The variable that is subject to change based on another variable is referred to as the dependent variable, while the variable that influences the dependent variable is known as the independent variable. In a stock market time series, the dependent variable is the current price, and the independent variable can be either time/date or past prices. In case of the independent variable is past price, regression is also known as autoregression. In this study, the independent variable is the date time index.

3.3. ARIMA

The Box-Jenkins, also known as the ARIMA model, is divided into three stages: identification, estimation, and diagnostic checking (Box et al., 1970). The ARIMA model comprises three key parameters, namely "p," "d," and "q." These parameters represent the orders of the autoregressive (AR), integrated (I), and moving average (MA) components of the model. Identifying these parameters is essential in determining the optimal fitting model. The following is a brief description of each component:

Auto-Regressive AR(p): The dependent variables are lagged observations in this model. The significant cut-off number of lagged observations is p. The value of p can be identified from the partial autocorrelation function (PACF) chart.

Integrated I(d): The time series needs to be stationary to implement the ARIMA model. The make the time series stationary, the observations are subtracted from their predecessor observations. The number of times the observations are subtracted is denoted by d.

Moving Average MA(q): The number of important lag of forecasting error is denoted by the letter "q," which stands for the order of moving average. To forecast future values, this model regresses over prior forecasting errors.

The ARIMA model is expressed in Equation (1).

$$\widehat{Y_t} = c + \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + .. + \theta_p Y_{t-p} - \phi_1 e_{t-1} - \phi_2 e_{t-2} - .. - \phi_q e_{t-q} \quad (1)$$

$\widehat{Y_t}$ denotes the predicted target, c is a constant, e is an error term, Y_t is past observation at time t after differencing d times.

3.4. Framework

All set of data is split into train and test data in a 75:25 ratio. 75% of the data is used to estimate the model fit, and 25% of the data is held back for testing. After prediction, test data is compared with forecasted values using RMSE and MAPE. **Error! Reference source not found.** shows the frame of the study.

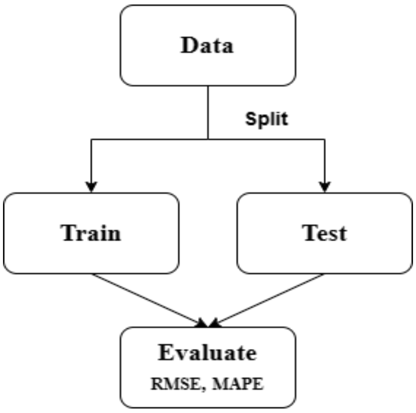

Figure 5. Framework

4. Results

Figure 6 and Figure 7 show the linear regression model fit line, prediction line, and train and test data for daily and weekly data for 1 year. Similarly, Figure 8 and Figure 9 show the charts of 2-year data.

The utilization of the Augmented Dickey-Fuller (ADF) test helps to ascertain if a time series is stationary or not. Further, the parameters p, d, and q of the ARIMA model are chosen using the Akaike Information Criterion (AIC). The results of the ADF test on NIFTY 50 1-year and 2-year data are shown in **Error! Reference source not found.** and **Error! Reference source not found.**, respectively.

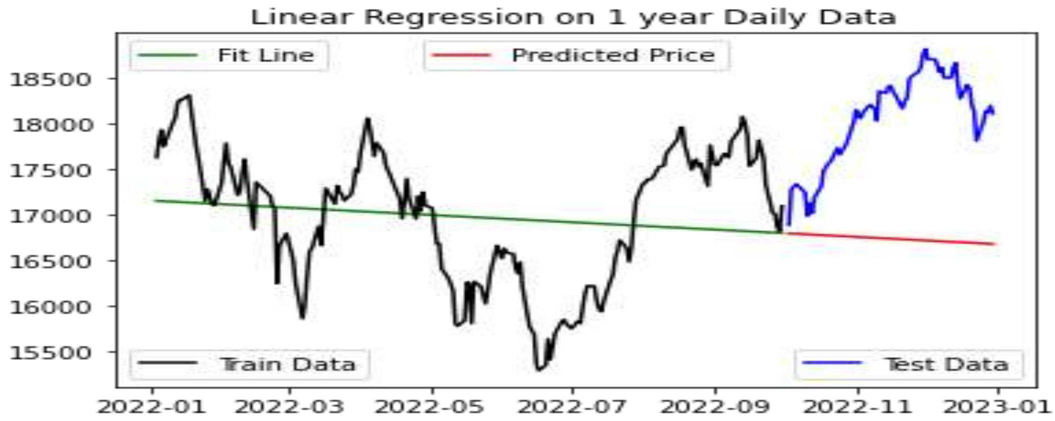

Figure 6. Linear Regression on 1-year daily data

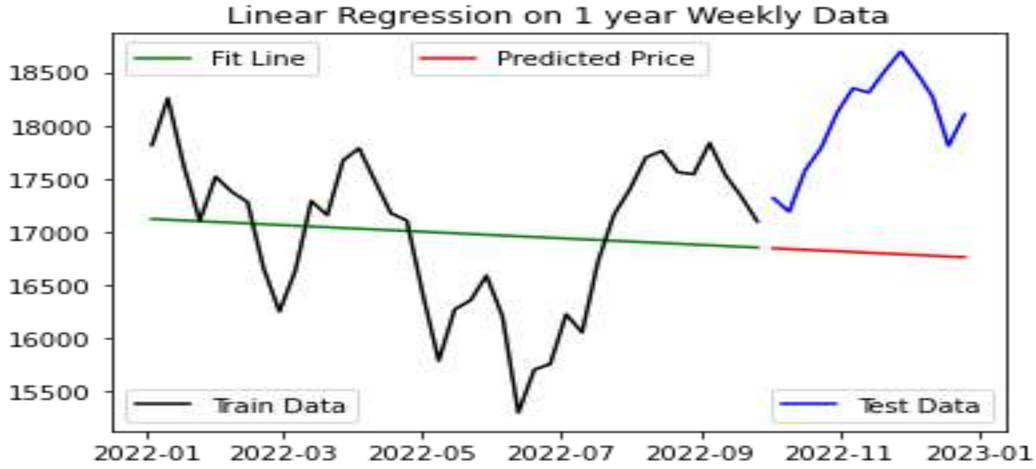

Figure 7. Linear Regression on 1-year weekly data

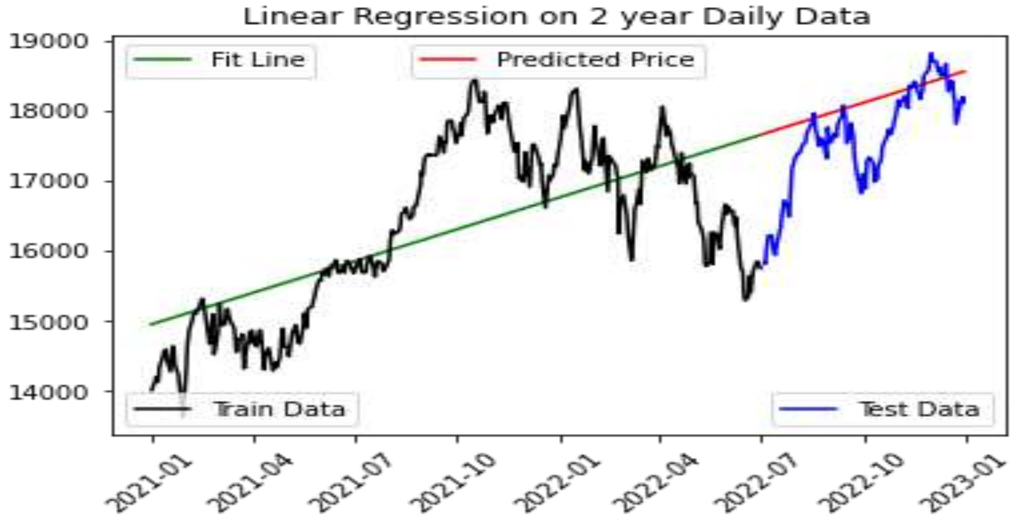

Figure 8. Linear Regression on 2-year daily data

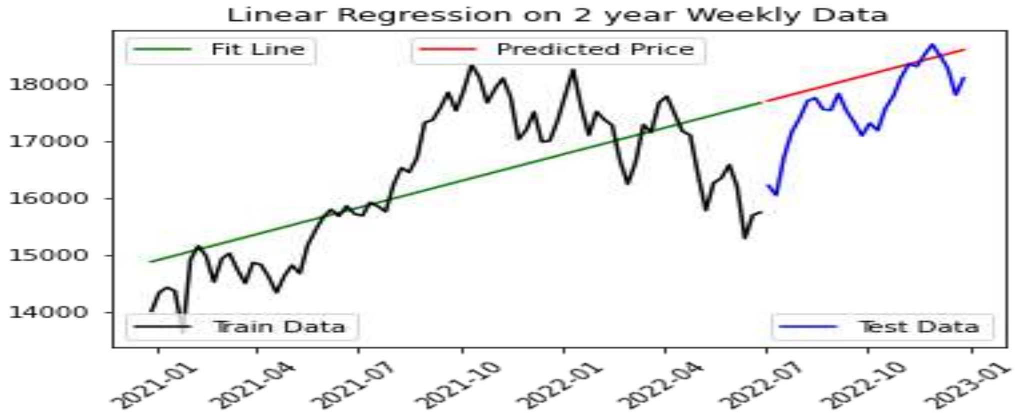

Figure 9. Linear Regression on 2-year weekly data

Table 1. NIFTY ADF Test on 1-year data

	Daily	Weekly
ADF Statistic:	-1.61	-1.59
p-value:	0.47	0.48
Used Lags:	0.00	0.00
The number of observations used	247.00	51.00
Critical Values: 1%	-3.45	-3.56
Critical Values: 5%	-2.87	-2.92
Critical Values: 10%	-2.57	-2.59

Table 2. NIFTY ADF Test 2-year Data

	Daily	Weekly
ADF Statistic:	-2.02	-1.67
p-value:	0.27	0.44
Used Lags:	0.00	2.00
The number of observations used	495.00	102.00
Critical Values: 1%	-3.44	3.49
Critical Values: 5%	-2.86	-2.89
Critical Values: 10%	-2.56	-2.58

Table 3. AIC test on NIFTY (1-year Data)

ARIMA(p,d,q)	AIC Daily Value	AIC Weekly Value
ARIMA(0,1,0)	2486.86	568.80
ARIMA(0,1,1)	2488.62	568.51
ARIMA(0,1,2)	2490.49	569.54
ARIMA(1,1,0)	2488.63	569.23
ARIMA(1,1,1)	2490.53	570.16
ARIMA(1,1,2)	2492.49	inf
ARIMA(2,1,0)	2490.50	568.36
ARIMA(2,1,1)	2492.49	570.10
ARIMA(2,1,2)	2493.74	569.63
ARIMA(3,1,0)	2492.48	569.93
ARIMA(3,1,1)	2494.49	571.38
ARIMA(3,1,2)	inf	inf
ARIMA(4,1,0)	2494.38	571.48
ARIMA(4,1,1)	2496.09	573.27
ARIMA(4,1,2)	2498.21	inf
ARIMA(5,1,0)	2496.04	573.21
ARIMA(5,1,1)	2497.87	571.86
ARIMA(5,1,2)	2499.97	574.13

Table 4. AIC test on NIFTY (2-year Data)

ARIMA(p,d,q)	AIC Daily Value	AIC Weekly Value
ARIMA(0,1,0)	4900.20	1156.10
ARIMA(0,1,1)	4900.71	1157.68
ARIMA(0,1,2)	4899.51	1153.75
ARIMA(1,1,0)	4901.02	1157.88
ARIMA(1,1,1)	4901.90	1155.61
ARIMA(1,1,2)	4900.74	1155.43
ARIMA(2,1,0)	4898.31	1154.06
ARIMA(2,1,1)	4899.60	1155.96
ARIMA(2,1,2)	4898.12	1156.44
ARIMA(3,1,0)	4898.46	1156.03
ARIMA(3,1,1)	4899.31	1157.91
ARIMA(3,1,2)	4899.85	1157.48
ARIMA(4,1,0)	4898.09	1157.99
ARIMA(4,1,1)	4899.36	1158.82
ARIMA(4,1,2)	4899.19	1158.76
ARIMA(5,1,0)	4899.32	1156.12
ARIMA(5,1,1)	4901.22	1157.72
ARIMA(5,1,2)	4894.10	1158.51

The p-value of all series shows that all-time series are non-stationary. Hence, the time series needs to be differenced to fit into the ARIMA model. The library auto_arima from pmdarima is used to find the best-fit parameter of the ARIMA model. **Error! Reference s ource not found.** and **Error! Reference source not found.** show the AIC value up to the first few combinations.

According to minimum AIC values from the tables, ARIMA(0,1,0) and ARIMA(2,1,0) are the best-fit models for 1 year of daily and weekly data, respectively. Similarly, ARIMA(5,1,2) and ARIMA(0,1,2) are best fit for 2-year daily and weekly data. Figure 10 and Figure 11 show the train data, test data, and predicted prices for the best ARIMA model on 1-year daily and weekly data. Similarly, Figure 12 and Figure 13 show the charts of 2-year daily and weekly data.

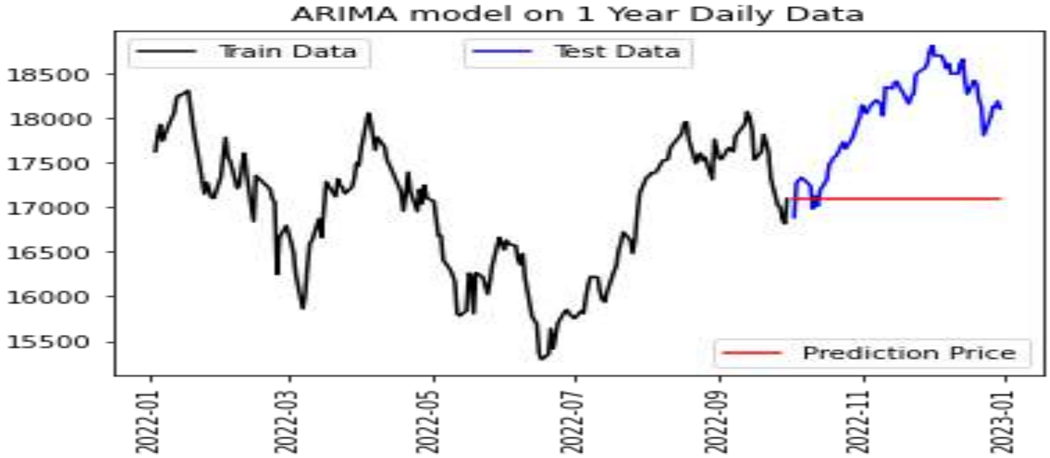

Figure 10. ARIMA model on 1-year daily data

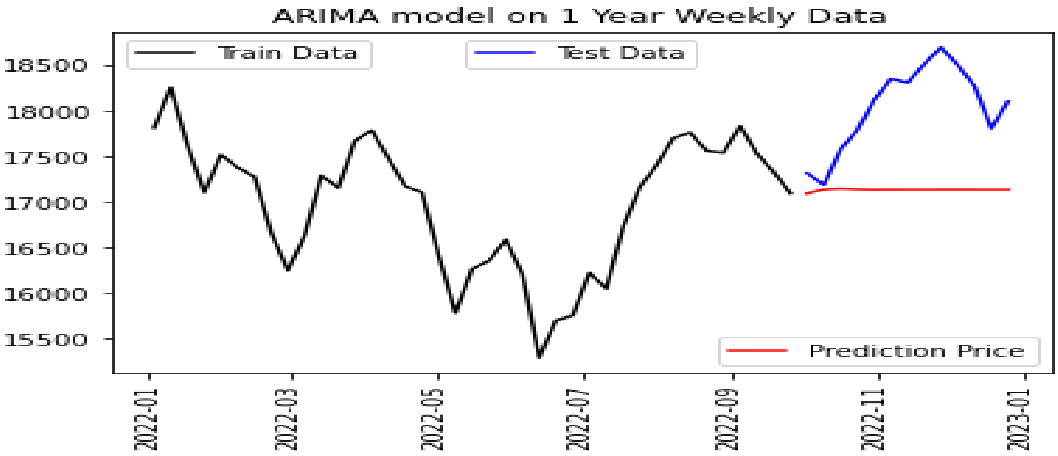

Figure 11. ARIMA model on 1-year weekly data

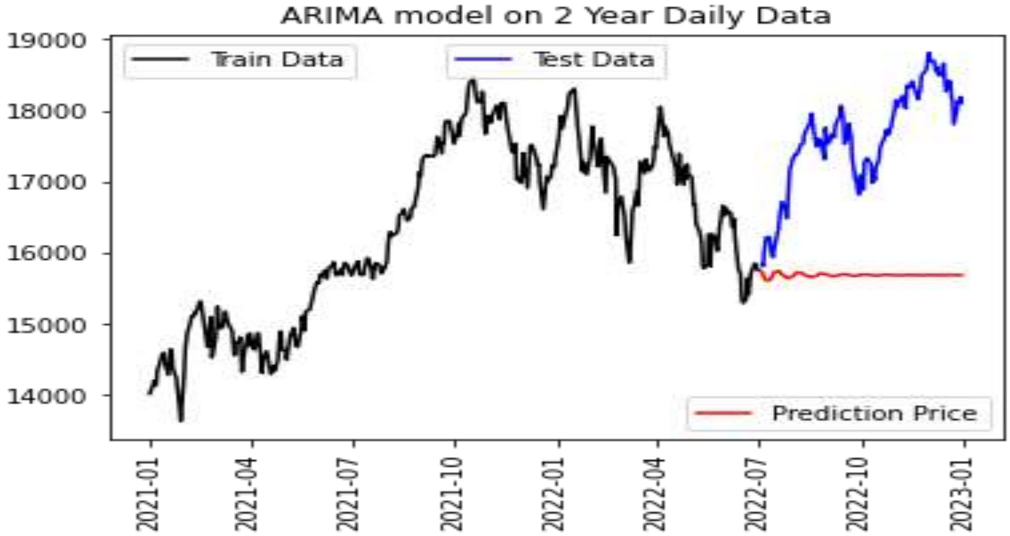

Figure 12. ARIMA model on 2-year daily data

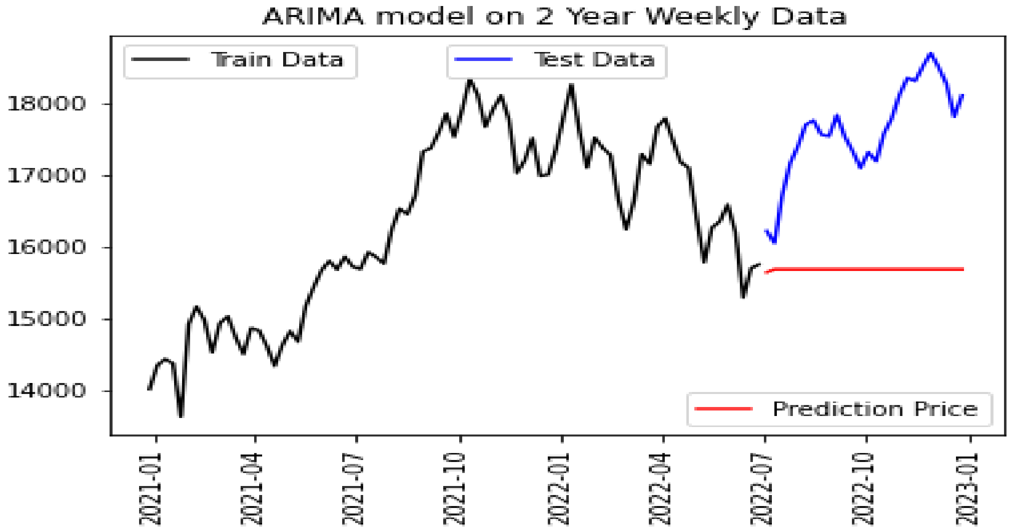

Figure 13. ARIMA model on 2-year weekly data

All ARIMA models rendered nearly flat lines. Notably, ARIMA is predicting the last price of the trained period to be the future price of all day. RMSE and MAPE are captured to evaluate and compare all models. **Error! Reference source not found.** shows the RMSE and M APE.

Table 5. RMSE and MAPE of model

Model/Timeframe		1 year daily	1 year weekly	2-year daily	2-year weekly
LR	RMSE	1427.39	1327.3	740.59	699.01
	MAPE	7.26	6.82	3.26	3.2
ARIMA	RMSE	1089.25	1011.37	2062.28	2050.48
	MAPE	5.35	4.96	10.78	10.92

5. Conclusions

The predictability of the stock market is always questionable, considering the efficient market hypothesis and random walk theory. Still, there is great interest from investors and traders who generally use fundamental and technical analysis. Many researchers have experimented with methods like statistical techniques and machine learning, but no established method exists to predict stock market time series. This paper attempts to predict the NIFTY 50 index using simple linear regression and ARIMA using different samples of data at different time frames.

The linear regression clearly predicts the trend in the time series. On the 1-year daily and weekly data, linear regression predicts a downward movement in stock price, but the market actually moves upward. On the other hand, linear regression predicted upward movement on 2-year daily and weekly data, and the market does go upward. Hence, it confirms stock market does not have fixed behavior. The linear regression prediction performance, based on RMSE and MAPE, is best on 2-year weekly data. There is also a small improvement when 2 years of data are taken in place of 1 year.

ARIMA models mostly predicted the last price of the training period to continue in the test period. The ARIMA models do not find any downward or upward trend in the NIFTY 50 time series. In contrast to the linear regression model, ARIMA performance is better on 1-year data than on 2-year data. However, there is a small improvement in weekly data when compared to daily data. The ARIMA(2,1,0) has the best RMSE and MAPE on 1-year weekly data.

The linear regression is able to capture the bigger trend using a bigger time frame. However, it fails to capture the price fluctuation. ARIMA, considered a more robust model for time series analysis, fails to capture both. Although the best RMSE and MAPE are the lowest for linear regression, further work is needed to find the best time frame. It can be inferred from the result that the bigger the time frame, the better the linear regression performance. Using linear regression, traders and investors might gauge the overall trend, if any, in the stock index. They can expect that, at some point, the price should catch up with the predicted price of linear regression. ARIMA might need a further transformation of prices to improve performance.

6. Limitations and Future Work

Researchers can further study what frequency of data, like daily, weekly, monthly, or quarterly, to select for linear regression. Further, it will be interesting to study how to adjust to different conditions, such as volatile/non-volatile, bear, or bull markets.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding information

The author declared no financial support for the research, authorship, and/or publication of this article.

Acknowledgments: None

References

Bhuriya, D., Kaushal, G., Sharma, A., & Singh, U. (2017). Stock market predication using a linear regression. *International Conference of Electronics, Communication and Aerospace Technology (ICECA)*, 510–513.

Box, G. E., Gwilym M. Jenkins, & G. Reinsel. (1970). Time series analysis: forecasting and control Holden-day San Francisco. *BoxTime Series Analysis: Forecasting and Control Holden Day*.

Du, Y. (2018). Application and analysis of forecasting stock price index based on combination of ARIMA model and BP neural network. *Chinese Control and Decision Conference (CCDC). IEEE*, 2854–2857.

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, *25*(2), 383–417. http://www.jstor.orgURL:http://www.jstor.org/stable/2325486Accessed:30/04/200806:47

Fama, E. F. (1995). Random Walks in Stock Market Prices. *The Journal of Finance*, 75–80.

Fama, E. F., & French, K. R. (1988). Permanent and temporary components of stock prices. *Journal of Political Economy*, *96*(2), 246–273.

Graham, B., & David, L. Dodd. (1965). The Intelligent Investor. *New York: Harper & Row*.

Jegadeesh, N., & Titman, S. (1993). Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency. *The Journal of Finance*, *48*(1), 65–91. https://doi.org/10.1111/j.1540-6261.1993.tb04702.x

Khan, S., & Alghulaiakh, H. (2020). ARIMA Model for Accurate Time Series Stocks Forecasting. *IJACSA) International Journal of Advanced Computer Science and Applications*, *11*(7). www.ijacsa.thesai.org

Lo, A. W., & MacKinlay, A. C. (1998). Stock market prices do not follow random walks: Evidence from a simple specification test. *The Review of Financial Studies*, *1*(1), 41–61.

Malkiel, B. G. (1973). A Random Walk Down Wall Street. *Norton & Co, New York*.

Malkiel, B. G. (2003). *The Efficient Market Hypothesis and Its Critics*.

Poterba, J. M., & Summers, L. H. (1988). Mean reversion in stock prices: Evidence and implications. *Journal of Financial Economics*, *22*(1), 27–59.

Roy, S. S., Mittal, D., Basu, A., & Abraham, A. (2015). Stock market forecasting using LASSO linear regression model. *Advances in Intelligent Systems and Computing*, *334*, 371–381. https://doi.org/10.1007/978-3-319-13572-4_31

Rubio, L., & Alba, K. (2022). Forecasting Selected Colombian Shares Using a Hybrid ARIMA-SVR Model. *Mathematics*, *10*(13). https://doi.org/10.3390/math10132181

Timmermann, A., & Granger, C. W. J. (2004). Efficient market hypothesis and forecasting. *International Journal of Forecasting*, *20*(1), 15–27. https://doi.org/10.1016/S0169-2070(03)00012-8