Stock Market Prediction

Aswathkrishna
2nd year
Department of Computing-Data Science
Coimbatore Institute of Technology
Coimbatore, India
aswathkrishna0922@gmail.com

Sujith
2nd year
Department of Computing-Data Science
Coimbatore Institute of Technology
Coimbatore, India
sujithvl109@gmail.com

Dr.D.Sudhadevi
Associate Professor
Department of Computing-Data Science
Coimbatore Institute of Technology Coimbatore, India
sudhadevi@cit.edu.in

Dr.M.Srividya
Assistant Professor
Department of Computing-Data Science
Coimbatore Institute of Technology
Coimbatore, India
srividhya@cit.edu.in

***ABSTRACT--* The stock market is highly volatile and unpredictable by nature. The use of statistical analysis has played a major role in assisting stock market prediction. This research paper aims at using various statistical methods to analyze historical data of the Nifty 50(NSE) as well as forecast its direction for the next twelve months. This study aims to tell the market scenario of the future by supporting it with statistical answers. Specifically, Auto Regressive Integrated Moving Average (ARIMA) Modelling, Exponential Smoothing, and Mean Squared Error analysis are the foci of interest. Nifty 50 dataset is publicly available on Yahoo Finance. It consists of historical daily opening and closing stock prices along with a few other additional market parameters.**

KEYWORDS-- Nifty50, ARIMA, Exponential Smoothing, Stock Price Forecasting

I. INTRODUCTION

Stock price prediction is regarded as one of most difficult task to accomplish in financial forecasting due to complex nature of stock market. The desire of many investors is to lay hold of any forecasting method that could guarantee easy profiting and minimize investment risk from the stock market. This remains a motivating factor for researchers to evolve and develop new predictive models. In the past years several models and techniques had been developed to stock price prediction. This paper is concentrated on using machine learning models for stock market analysis and prediction. Specifically, Auto Regressive Integrated Moving Average (ARIMA) and Exponential Smoothing are discussed.

A popular and widely used statistical method for time series forecasting is the ARIMA model. It is one of the most popular models to predict linear time series data. This model has been used extensively in the field of finance and economics as it is known to be robust, efficient, and has a strong potential for short-term share market prediction. Exponential smoothing and ARIMA models are the two most widely used approaches to time series forecasting and provide complementary approaches to the problem. While exponential smoothing models are based on a description of the trend and seasonality in the data, ARIMA models aim to describe the auto-correlation(Autocorrelation is the degree of similarity between a given time series and lagged version of itself over successive time intervals) in the data.

II. NIFTY 50

The NIFTY50 is a benchmark Indian stock market index that represents the weighted average of 50 of the largest Indian companies listed on the National Stock Exchange. It is one of the two main stock indices used in India, the other being the BSE SENSEX.The Nifty 50 index was launched on 22 April 1996, and is one of the many stock indices of Nifty.

The Nifty 50 is an indicator of the top 50 major companies on the NSE. If the Nifty 50 goes upward, it means that most of the stocks in India went up during the period and the Nifty 50 goes downward, that the stock price of most of the major stocks on down.

III. ARIMA MODEL

Box and Jenkins in 1970 introduced the ARIMA model.It also referred to as Box-Jenkins methodology composed of set of activities for identifying, estimating and diagnosing ARIMA models with time series data. The model is most prominent methods in financial forecasting. ARIMA models have shown efficient capability to generate short- term forecasts. It constantly outperformed complex structural models in short-term prediction. In ARIMA model, the future value of a variable is a linear combination of past values and past errors, expressed as follows:

$$Y_t = \varphi_0 + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + .. + \varphi_p Y_{t-2} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta 1_q \varepsilon_{t-q} \quad (1)$$

where,Y_t is the actual value and t ε is the random error at t; φ_i and θ_j are the coefficients, p and q are integers that are often referred to as orders of autoregressive and moving average,respectively.

Building an ARIMA predictive model consists of many steps, which are identification of model, estimation of parameter and checking of diagnostic

An ARIMA (p, d, q) model requires three input parameters:

- p is used for autoregressive.
- d is used for to make the time series data stationary.
- q is used for moving average.

IV. Exponential Smoothing

Exponential smoothing is a set of methods for smoothing time series in order to forecast the immediate future. The main idea behind exponential smoothing is to forecast future values using a weighted average of all the previous values in our time series data. Exponential smoothing can be used for forecasting the time series data based on the three aspects: the level, the trend and the seasonal component. Based on these three aspects we have three types of exponential smoothing. Exponential smoothing is very popular because it is simple, it's adaptive and it's in expensive to compute.

A. Holt Winter's Exponential Smoothing

Holt Winter's Exponential Smoothing (HWES) also called Triple Exponential Smoothing. This method takes the idea of Holt's method and adds a seasonal component to create the even more complex system. We assume here, that the series has a level, trend, seasonality with M seasons and noise. Therefore, forecast equation for Winter's Exponential Smoothing is:

Forecast = Estimated level + Trend + Seasonality at most recent time point

$$F_{t+k} = L_t + kT_t + S_{t+k-M} \tag{1}$$

where L_t is the level estimate for time t, k is the number of forecasts into the future, T_t is the trend estimate at time t, S_t is the seasonal estimate at time t, and M is the number of seasons.

V. Objective

In the present scenario, investment in financial markets is very important for the growth of the economy. Limited analysis of an investment can result in loss.Currently, investors mainly focus on the fundamental data of the company for making an investment. So the objective of this study is to forecast stock prices of the Indian market in order to make better investment decisions.

This paper aims to gain insight into market behavior over time with effective stock prediction models. The main objectives of the research are as follows: 1. To identify the impact of feature selection process and hyper-parameter optimization on prediction quality and metrics used in the prediction of stock market performance and prices. 2. To analyze historical data of NIFTY 50 and used it for training and validation purposes. 3. The use of machine learning models to forecast the Price of NIFTY 50 index. 4. To examine the results and analyze the efficiency of each model evaluation metrics.

Those objectives investigate the best forecasting models. Thus, providing a comprehensive analysis of the NIFTY 50 index.

VI. Literature Review

Nayak et al. (2016) in the paper titled 'Prediction models for Indian stock market'stated that, for the past few years, many people have been showing interest in investing in the stock market. When investing, the investor may lose all the money they invested. For this reason, an efficient predictive model is required to understand the future behaviour of the stock market. Many forecasting models have been developed about the market trend but very few give good results.

Babai et al. (2011) in the paper titled 'Forecasting and inventory performance in a two-stages supply chain with ARIMA (0,1,1) demand' stated that the demand model for ARIMA (0,1,1) was analysed extensively by researchers. Forecasting practitioners use ARIMA widely, as it has promising theoretical features. They analyse the correlation between the accuracy of forecasting and the performance ofthe inventory in order to investigate if there are any benefits of sharing the forecast data with retailers and manufacturers.

Ertekin and Büyükşahin (2019) in the paper titled 'Improving forecasting accuracy of time series data using a new ARIMA-ANN hybrid method' stated that itis important to forecast time series data as it is also a very challenging task. It is usedin a lot of other fields of application. Studies have been done on linear dataindividually or in a combination with non-linear data. To forecast stationary timeseries data, a linear model like ARIMA has a good forecasting accuracy.The past studies also classify forecasting models according to their perspective:statistical approach and artificial intelligence approaches.

An ARIMA model relates to the statistical perspective (Wang et al., 2012). It is considered as being efficient as well as predominant for time series forecasting. Many researchers showed that theshort-term predictions of ARIMA technique perform better than those of ANNsmodels (Lee et al., 2007; Merh et al., 2010; Sterba & Hilovska, 2010).Adebiyi et al. (2014) in their study demonstrated the ability of an ARIMA modelto provide relatively accurate short-term predictions about stock prices.

Over the years, traders at the Nairobi stock exchange have made significant losses buying and trading stocks, but with Holt Winters model such losses will be eliminated, as stock investors will be able to predict stock movement and avoid losses in advance should the stock price head south. Such a stock prediction model, the Holt Winters Exponential Smoothening model, has never been used at the Kenyan stock market in the prediction of the stock prices, compared to other prediction models like neural networks that gas been used at the Nairobi stock exchange.

VII. Dataset Description

For this research,we have collected data on the monthly closing stock indices of NIFTY50 for fourteen years(2007-2022). It consists of historical daily opening and closing stock prices along with a few other additional market parameters. Fig.1 shows the monthly closing price of Nifty50 from October 1, 2007 to October 1, 2022

VIII. Experimental Environment

The following libraries were used in conducting the statistical analysis:

- Matplotlib was used for plotting the graphs for the actual time-series aswell as the predicted trends.
- Pandas was used for reading value from csv files as DataFrames.
- Numpy was used to perform matrix operations like flip, reshape, and create random matrices.
- Statsmodels was used for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration.

IX. METHODOLOGY :

The machine learning models used in this study were Auto Regressive Integrated Moving Average(ARIMA) and Exponential Smoothing. The tool used for implementation of model is Python.

A univariate analysis was done using ARIMA, applying it on the close column of stock prices of NIFTY50. The data from October 2007 to September 2019 (144 records) was used as a training dataset, and the data from October 2019 to October 2022 (37 records) was used as a testing dataset. First, for ARIMA, tests were performed to check stationarity, which is particularly useful indetermining whether the time-series data was stationary or not. First, rolling statistics were examined, which included rolling mean and rolling standard deviation to plot a moving average. Second, the Augmented Dickey–Fuller(ADF) test was conducted. Furthermore, a statistical trend was also calculated by taking the logarithmic approach. Non-stationarity was also proven with moving average calculation derived from successive segments of the series values The ARIMA() function was imported from the arima_model sub-package of statsmodel.tsa package. The auto_arima function was run to determine the optimum values of p, d and q. The values of p, d, and q were obtained as 0, 1, 0 with the lowest value of AIC = 2089.635. After obtaining the optimum value of p, d, and q for a particular sector, it was fed as an input to the ARIMA() function to get the forecasted stock price value.

Secondly, a univariate analysis was done using Holt Winters Exponential Smoothing, applying it on the close column of stock prices of NIFTY50. The data from October 2007 to September 2019 (144 records) was used as a training dataset, and the data from October 2019 to October 2022 (37 records) was used as a testing dataset. An Exponential Smoothing () function was imported from the statsmodels package. The seasonal and trend parameter was set as 'multiplicative' as seasonality was not constant in the dataset. The seasonal periods were chosen as 12, which means the data will exhibit seasonality after every twelve months.

To check the statistical accuracy of the forecasted result, Root Mean Square Error (RMSE), Mean Absolute Deviation(MAD) and Mean Squared Error(MSE) are computed.

X. Results And Discussion

A time series data is considered to be stationary when the mean and standard deviation remain constant over time. In order to check the stationarity of data, the rolling mean and standard deviation are computed and plotted on the original time series data(Fig.3). To check the stationarity of the time series, we will also use the Augmented Dickey-Fuller test (ADF).

Fig.1. NIFTY50 Closing Prices Graph from 10/01/2007 -- 10/01/2022

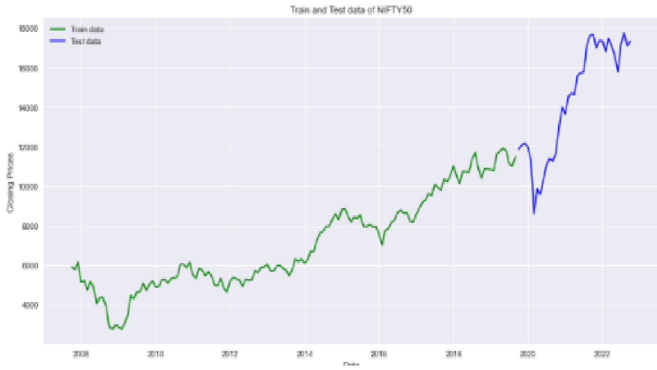

Fig.2. Data Distribution based on Train and Test Datasets

Fig.3. Rolling Mean,Standard Deviation and Original Data

Results of Agumented Dickey Fuller Test:	
Test Statistic	0.613947
p-value	0.987954
#Lags used	0.000000
Number of observations used	180.0000
Critical value (1%)	-3.467211
Critical value (5%)	-2.877735
Critical value (10%)	-2.575403

It is evident from the graph in “Fig.3” that the mean and standard deviation is not constant over time. For a more accurate assessment the Agumented Dickey-Fuller Test (ADF) was performed and the results confirm the same i.e. the data is not stationary. So the data needs to be made stationary. To do that, the first-order difference of the data was taken. In simple words, subtract this months price from previous months price and plot it again (Fig.5).

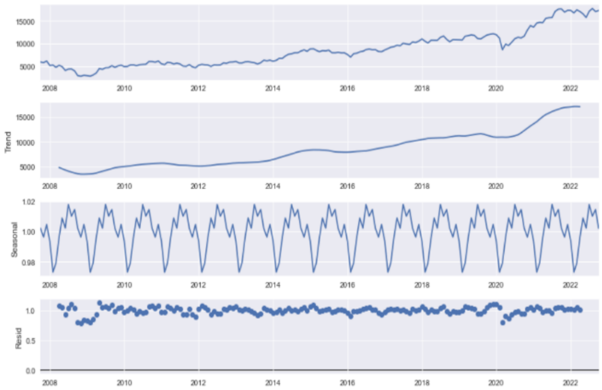

Fig.4. Decomposition of Time Series Data

The original time series data is decomposed and plotted in (Fig.4). It contains trend, seasonality and the residual part of the original time series data.


Fig.5. Rolling Mean, Standard Deviation and Differenced Data

From the graph in Fig.5, it is evident that mean and standard deviation remain constant. Since stationarity is achieved by differencing once, the d term for ARIMA will be 1.

Autocorrelation function (ACF) and Partial Autocorrelation Function (PACF, also called Partial ACF) are important functions in analyzing a time series. They generally produce plots that are very important in finding the values p and q for Autoregressive (AR) and Moving Average (MA) models.

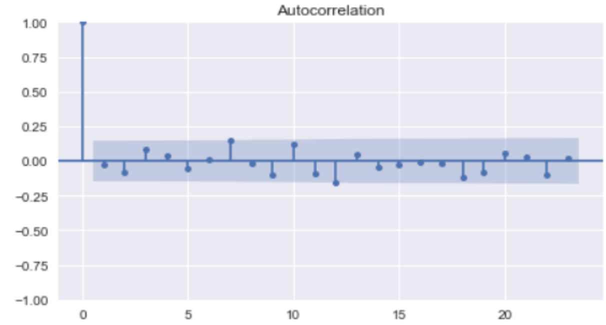

Fig.7. Auto Correlation Plot(ACF)

The p term for ARIMA is computed from the autocorrelation plot (Fig.7).

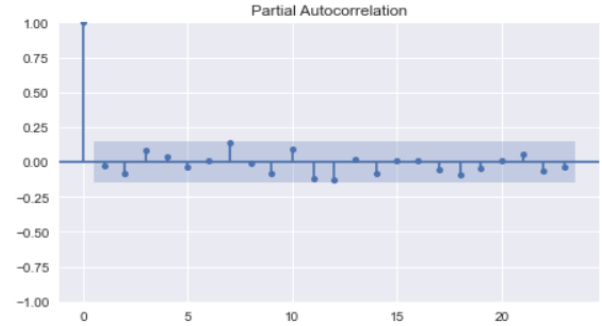

Fig.8. Partial Auto Correlation Plot(PACF)

The q term for ARIMA is computed from the partial autocorrelation plot(Fig.8).

```
Performing stepwise search to minimize aic
 ARIMA(1,1,1)(0,0,0)[0] intercept   : AIC=2093.624, Time=0.83 sec
 ARIMA(0,1,0)(0,0,0)[0] intercept   : AIC=2089.929, Time=0.02 sec
 ARIMA(1,1,0)(0,0,0)[0] intercept   : AIC=2091.549, Time=0.03 sec
 ARIMA(0,1,1)(0,0,0)[0] intercept   : AIC=2091.426, Time=0.09 sec
 ARIMA(0,1,0)(0,0,0)[0]             : AIC=2089.635, Time=0.02 sec

Best model:  ARIMA(0,1,0)(0,0,0)[0]
Total fit time: 1.098 seconds
```

Fig.9. Results of auto_arima function

The auto_arima function was run to determine the optimum values of p, d and q. The values of p, d, and q were obtained as 0, 1, 0 with the lowest value of AIC = 2089.635(Fig.9).

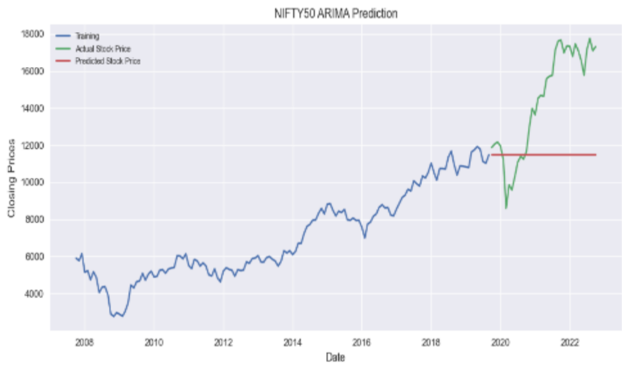

Fig.10. Actual vs Predicted Data using ARIMA

For ARIMA (p,d,q) model, the values obtained are p = 0, d = 1 and q = 0. So the model is fitted for ARIMA (0,1,0). The result obtained and plotted in (Fig.10).

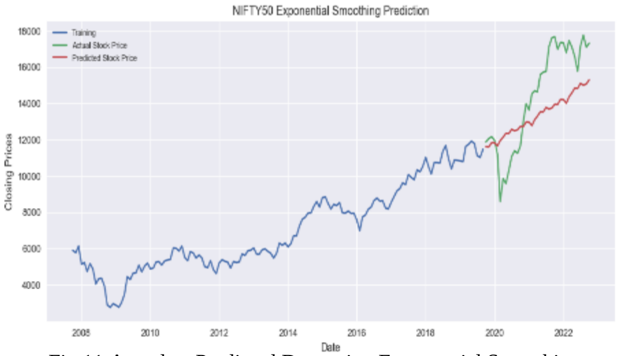

Fig.11.Actual vs Predicted Data using Exponential Smoothing

TABLE 1.ARIMA FORECASTED VALUES

Date	Forecasted Values
01-11-2022	17314.65
01-12-2022	17314.65
01-01-2023	17314.65
01-02-2023	17314.65
01-03-2023	17314.65
01-04-2023	17314.65
01-05-2023	17314.65
01-06-2023	17314.65
01-07-2023	17314.65
01-08-2023	17314.65
01-09-2023	17314.65
01-10-2023	17314.65

Fig.12. Original and ARIMA 12 Months Forecasted Data

TABLE 2. EXPONENTIAL SMOOTHING FORECASTED VALUES

Date	Forecasted Values
01-11-2022	16783.66
01-12-2022	16995.9
01-01-2023	16280.44
01-02-2023	16700.72
01-03-2023	17378.78
01-04-2023	18357.21
01-05-2023	19083.57
01-06-2023	19041.04
01-07-2023	19625.77
01-08-2023	19678.71
01-09-2023	20006.31
01-10-2023	19676.6

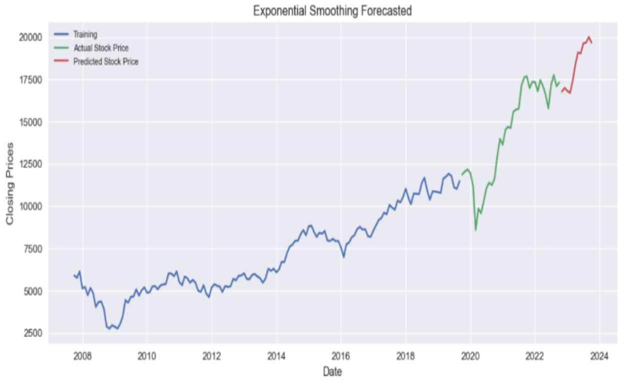

Fig.13. Original and Exponential Smoothing 12 Months Forecasted Data

TABLE 3.ERROR METRICS

MODEL	MSE	MAD	RMSE
ARIMA	**16189338.20**	**3355.35**	**4023.60**
Exponential Smoothing	**4710856.92**	**1897.19**	**2170.45**

XI. CONCLUSION

Predictions of the NIFTY50 closing prices that are generated using Auto Regressive Integrated Moving Average(ARIMA) and Exponential Smoothing are easy to implement, reliable and justifiable. The comparative analysis based on MSE, MAE, RMSE and MAPE values clearly indicate that Exponential Smoothing gives better prediction of stock prices as compared to ARIMA. Results show that the best values obtained by Exponential Smoothing model gives MSE(4710856.92), MAD(1897.19) and RMSE(2170.45). TABLE 3 shows the evaluation metrics uesd for the comparative analysis of predictive models.

XII. REFERENCES

[1] Adebiyi, A. A., Adewumi, A., & Ayo, C. (2014). Comparison of ARIMA and artificialneural networks models for stock price prediction. Journal of Applied Mathematics,2014, 1–7.

[2] Kahkashan Khan and Dileep Singh.(2022).Stock Price Forecasting of Maruti Suzuki using Auto Regressive Integrated Moving Average (ARIMA) Model.

[3] Babai, M. Z., Ali, M. M., Boylan, J. E., & Syntetos, A. A. (2013). Forecasting and inventory performance in a two-stages supply chain with ARIMA (0,1,1) demand. ProductionEconomics, 143, 463–471

[4] Ertekin, S., & Büyükşahin, Ü. Ç. (2019). Improving forecasting accuracy of time seriesdata using a new ARIMA-ANN hybrid method. Neurocomputing, 361, 151–163.

[5] Lee, C., & Ko, C. (2011). Short-term load forecasting using lifting scheme and ARIMAmodel. Expert System with Applications, 38(5), 5902–5911.

[6] Lee, K., Sehwan, Y., & Jin, J. J. (2007). Neural network model vs. SARIMA model inforecasting Korean stock price index (KOSPI). Issues in Information Systems, VIII(2).

[7] Meyler, A., Kenny, G., & Quinn, T. (1998). Forecasting Irish Inflation using ARIMA

Models.Paper presented at Central Bank of Ireland Research Department.
[8] Nayak, A., Pai, M. M. M., & Pai, R. M. (2016). Prediction models for Indian stock market.Procedia Computer Science, 89, 441 –449.
[9] Shastri, Sourabh& Sharma, Amardeep&Mansotra, Vibhakar& Sharma, Anand&Bhadwal, Arun&Kumari, Monika. (2018). A Study on Exponential Smoothing Method for Forecasting.International Journal of Computer Sciences and Engineering. 6. 482-485. 10.26438/ijcse/v6i4.482485.
[10] Ostertagova, Eva &Ostertag, Oskar. (2011). The Simple Exponential Smoothing Model.
[11] Bhattacharjee, Indronil&Bhattacharja, Pryonti. (2019). Stock Price Prediction: A Comparative Study between Traditional Statistical Approach and Machine Learning Approach.
[12] PAUL, Dr.Sanjoy Kumar. Determination of Exponential Smoothing Constant to Minimize Mean Square Error and Mean Absolute Deviation. Global Journal of Research In Engineering, [S.l.], v. 11, n. 3, apr. 2011. ISSN 2249-4596.
[13] Lima, Susana &Gonçalves, A. Manuela & Costa, Marco. (2019). Time series forecasting using Holt-Winters exponential smoothing: An application to economic data. AIP Conference Proceedings. 2186. 090003. 10.1063/1.5137999.
[14] Pan, Rong. (2010). Holt–Winters Exponential Smoothing. 10.1002/9780470400531.eorms0385.
[15] Sharif, Omar &Hasan, Md. (2019). Forecasting the Stock Price by using Holt’s Method.Indonesian Journal of Contemporary Management Research. 1. 15. 10.33455/ijcmr.v1i1.83.