Stock Price Prediction Using LSTM, ARIMA and UCM

Journal of Development Economics and MANAGEMENT RESEARCH STUDIES (JDMS) ***A PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL***
ISSN: 2582 5119 (ONLINE)
Crossref Prefix No: 10.53422
09(11), 55-66, January-March, 2022
@Center for Development Economic Studies (CDES)
Reprints and permissions
https://www.cdes.org.in/
https://www.cdes.org.in/about-journal/

Stock Price Prediction Using LSTM, ARIMA and UCM

Dr. Poorna Shankar[1], Dr. Neha Sharma[2], Mr. Kota Nagarohith[3], and Mr. Ashish Ghosh[4]

Abstract

The stock market is growing abundantly due to the rise of investors for their passive income is This article aims to develop an innovative artificial recurrent neural network approach for better stock market forecasts. The stock 611market is receiving a lot of attention from investors. Capturing the regularity of stock market changes has always been a key point for investors and investment firms. Investors are very interested in the field of stock price forecasting research. To make a successful investment, many investors want to know the future of the stock market. The data is pulled from the livestock market for analysis and visualization and results in analysis in real-time and offline. Predictive methods can be divided into two broad categories: statistical methods and artificial intelligence methods. Statistical methods include the logistic regression model and ARIMA, UCM model. Artificial intelligence techniques include multi-layer perceptron's, accumulative neural networks, Naïve Bayes, back-propagation networks, single-layer LSTMs, vector bearers, cyclic neural networks, and more. From this research, LSTM is achieving less Error percentage than any other model. LSTM helps investors, analysts, or anyone interested in investing in the stock market to get a better understanding of the future state of the stock market.

Keywords: stock market, investment, LSTM, RNN, ARIMA, UCM.

[1] Professor, Keystone Global, Ahmedabad, Gujarat.
[2] Director, Shanti Business School, Ahmedabad, Gujarat.
[3] Pursuing MSC in Big Data, FOM University, Germany.
[4] Pursuing MSC in Big Data, FOM University, Germany.

1. Introduction

The stock market will always be a popular investment for investors due to the high profit and returns it generate [4]. As the level of investing and trading grew, people searched for tools and methods that would increase their gains while minimizing the risk [14]. Predicting the stock market will be complicated and difficult for investors [2]. A lot of research is going on to predict the future stock market movement. To determine the set of relevant factors for making accurate predictions is a complicated task and so regular stock market analysis is very essential [1]. Trading in the stock market has gained huge popularity and it has become a part of the daily routine for many people to reap handsome profits. The prediction of the stock price is becoming difficult due to the complexity of the stock data, and the analysis of stock movement will be on the analysis of previous stock prices [3]. Before the use of computers, people used to do trading based on their gut feelings [12]. As investment increases and the stock market grew, people are searching for tools and methods that would increase their profits while minimizing the risk [7]. A successful stock price prediction will achieve with minimum data and the least complex stock market model [14]. To predict stock prices accurately, we must consider various factors which can affect the stock price. Only by fully understanding these factors' change trend and effect, can a reasonable and effective judgment be made. Despite having different time series models, the effectiveness for improving the predictions was never stopped [1].

In this paper, we use LSTM (Long Short-Term Memory) and ARIMA (Autoregressive Integrated Moving Average) and UCM (Unobserved Component Model). LSTM is an artificial recurrent neural network architecture used in the field of deep learning, which has feedback connections. It can process not only single data points but also entire sequences of data [14]. ARIMA is mainly used in statistics, and particularly in time series analysis, it is a generalization of an integrated autoregressive moving average model [13]. The UCM procedure analyses and predicts similarly spaced univariate time series data by using an unobserved components model (UCM) [11,15]. The UCMs procedure are also known as structural models in the time series literature [11]. A UCM decomposes the response series into components such as trend, seasonal, cycles, and the regression effects due to predictor series [15]. The components in the model are supposed to capture the salient features of the series that are useful in explaining and predicting its behaviour [11]. Harvey (1989) is a good reference for time series modelling that uses the UCMs. In recent years, some researchers have proposed that long short-term memory (LSTM) networks have higher prediction accuracy [7]. These models are applied with time-series data to predict future points in the series. A considerable number of researchers claim that some models are acceptable as long as they can produce predictions with significant accuracy. In the following sections, the differences in the process and results of various models will be further discussed.

2. Methodology

2.1 Description of Data

The historical data for the two companies has been extracted from NSE India. The data set includes 5 years of data from Tata Motors India Ltd. The data contains information about stocks such as High, Low, Open, Close, Adj. Close and Volume.

2.2 ARIMA

The ARIMA method was first promoted by Box and Jenkins, and it is often referred to as the Box Jenkins model [13]. Box and Tiao (1975) analyse the general transfer function model used in the ARIMA process. This model is sometimes called an ARIMAX model if it contains other time series as input variables. The ARIMA process uses an integrated auto-regressive moving average (ARIMA) or an auto-regressive integrated moving average (ARMA) to analyse and predict evenly spaced univariate time series [4]. As the transfer function affects the data, the ARIMA model combines the values of the response time series with its past values, past errors, the linear combination of the present and the past. The ARIMA process provides a comprehensive set of tools for univariate time series modelling, parametric estimation, and prediction, providing great flexibility for the types of ARIMA or ARIMAX models that can be analyzed [10]. The time series model is confusing or failing. The ARIMA procedure supports seasonal, subset, and factored ARIMA models; intervention or interrupted time series models; multiple regression analysis with ARMA errors; and rational transfer function models of any complexity [10]. In this model, the future value of a variable is a linear combination of past values and past errors, expressed as follows:

$$W_t = \mu + \frac{\theta(B)}{\phi(B)} a_t$$

Equation (1). ARIMA Equation

Where

t index times

W_t is the response series Y_t or a difference of the response series

μ is the mean term

B is the backshift operator; that is $BX_t = X_{t-1}$

$\phi(B)$ is the autoregressive operator, represented as a polynomial in the backshift operator: $\phi(B) = 1 - \phi_1 B - \ldots - \phi_P B^P$

$\theta(B)$ is the autoregressive operator, represented as a polynomial in the backshift operator: $\theta(B) = 1 - \theta_1 B - \ldots - \theta_P B^P$

a_t is the independent disturbance, also called the random error

The analysis performed by the ARIMA model is divided into three stages, corresponding to the stages described by Box and Jenkins (1976) [10].

1. In the identification stage, it specifies the response series and identifies candidate ARIMA models for it. And it reads time-series data that are to be used in later statements, possibly differencing them, and computes autocorrelations, inverse autocorrelations, partial autocorrelations, and cross-correlations. Stationarity tests can be performed to determine if differencing is necessary [10].
2. In the estimation and diagnostic checking stage, it specifies the ARIMA model to fit the variable specified in the identification stage and also estimates the parameters of the model. It also produces diagnostic statistics to help you judge the adequacy of the model. Significance tests for parameter estimates indicate whether some terms in the model might be unnecessary.

If the diagnostic tests indicate problems with the model, you try another model and then repeat the estimation and diagnostic checking stage [10].

3. In the forecasting stage, future values of time series are forecasted and to generate confidence intervals for these forecasts from the ARIMA model produced by the preceding estimating stage [10].

2.3 UCM:

The UCM procedure analyses and forecasts equally spaced univariate time series data by using an unobserved components model (UCM) [11]. The UCMs are also known as structural models in the time series literature. A UCM decomposes the response series into components such as trend, seasonal, cycles, and the regression effects due to predictor series [15]. The components in the model are supposed to capture the salient features of the series that are useful in explaining and predicting its behaviour. The UCMs capture the versatility of the ARIMA models while possessing the interpretability of the smoothing models [15].

The fully specified UCM is represented as:

$$y_t = T_t + S_t + C_t + I_t$$

Equation (2). UCM

Where

T_t trend in y_t

S_t seasonal effect at time t

C_t cyclic effect at time t

I_t Irregular effect at time t

All these components are assumed to be unobserved so that’s why the model is Unobserved Component Model.

If a suitable UCM is found, it can be used for various purposes. It used for the following:

• forecasting the values of the response series and the component series in the model

• obtaining a model-based seasonal decomposition of the series

• obtaining a “denoised” version and interpolating the missing values of the response series in the historical period

• obtaining the full sample or “smoothed” estimates of the component series in the model

2.4 LSTM:

Artificial Neural Networks are good at recognizing hidden patterns in time series data [8]. Artificial Neural Network will undergo a process of training in which it records previous time-series data points and tune them using the hidden layer and produce the output [9]. LSTM is a subset of Recurrent Neural Network (RNN), in which it is a subset of ANN. Long Short-Term Memory (LSTM) is an advanced Recurrent Neural Network (RNN). It mainly solves the problem of gradient disappearance that frequently occurs in conventional RNNs and enables data analysis from a longer time series [2]. Compared with RNN, the LSTM model adds three memory modules, the input information at time t, select useful information with a certain probability, and finally extract useful information through the output gate as the state of the final retention layer, and then participate in the calculation of the next time.

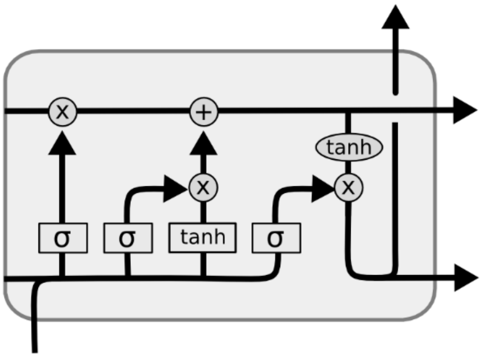

Figure 1 Long Short Memory Network Cell State

The first step involved in LSTM is to decide which information will go through from the cell state. It will be made by a sigmoidal layer called "Forget gate Layer". It looks at h_{t-1} and x_t and outputs are between o and 1 for each number in cell state.

$$f_t = \sigma(W_f . [h_{t-1}, x_t] + b_f$$

Equation (3) Forgot Layer

The next step is to decide, the new information will be stored in the cell state. It has two steps. First, the sigmoid layer called the "input gate layer" will decide the value which was updated. Next, a tan layer creates a new layer which creates a new candidate value. After this, these two values will be updated to the state.

$$i_t = \sigma(W_i . [h_{t-1}, x_t] + b_i$$
$$\tilde{c}_t = \tanh(W_C . [h_{t-1}, x_t] + b_C$$

Equation (4) Input layer Gate

After this C_{t-1} old cell state will be updated into the new cell state C_t, by multiplying with the Forget state value and then adding it to $i_t * \tilde{c}_t$. This will be the new candidate value.

$$C_t = f_t * C_{t-1} + i_t * \tilde{c}_t$$

Equation (5) New Candidate Equation

The final output will be the filtered version of the cell state. First, we run with the sigmoid layer which decides the output we need based on the cell state, then we put the cell state through tan and multiply with the output of the sigmoid gate, to decide the output.

$$o_t = \sigma(W_o . [h_{t-1}, x_t] + b_o$$
$$h_t = o_t * \tanh (C_t)$$

Equation (6) Output of sigmoid gate

2.5. Data Pre-processing:

Real-Time stock data will be collected from NSE India. Variables will be removed which are irrelevant for the model. Data cleaning should have happened and if data is missing then imputation methods will be used to get those data. The model will be built using Python and SAS tools i.e., ARIMA, UCM in SAS and LSTM in python.

3. Results

3.1. LSTM

There are relatively few studies on stock price prediction using LSTM models. In this paper, we will discuss a comparison of LSTM with UCM and ARIMA. In LSTM we will check the accuracy of prediction and epoch loss. In both LSTM with ANN and LSTM with regression, the models are a good fit, the figures are given below.

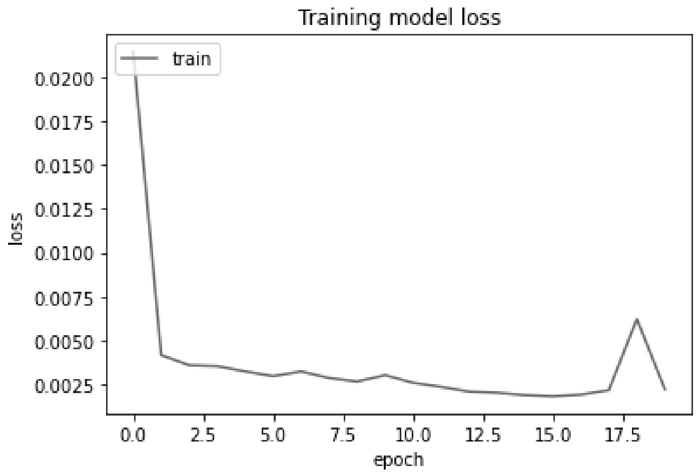

Figure 2. Loss vs. Epoch in ANN

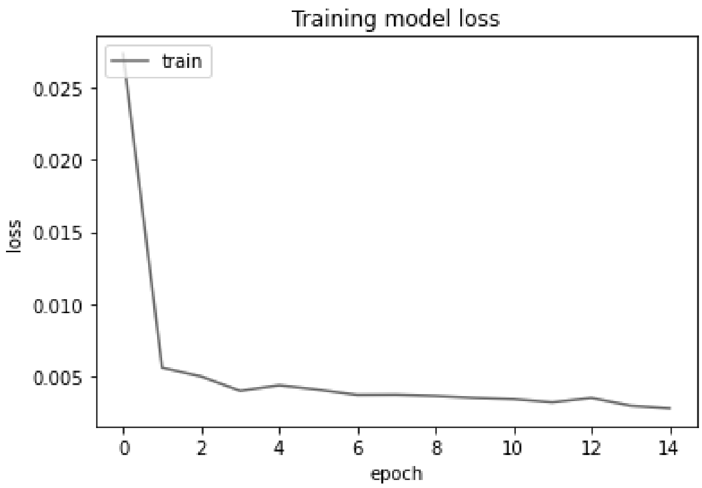

Figure 3. Loss vs Epoch in LSTM

Both the models are good fit, if we see the above diagram as the epoch are increasing the losses are decreasing and the loss value is becoming stable.

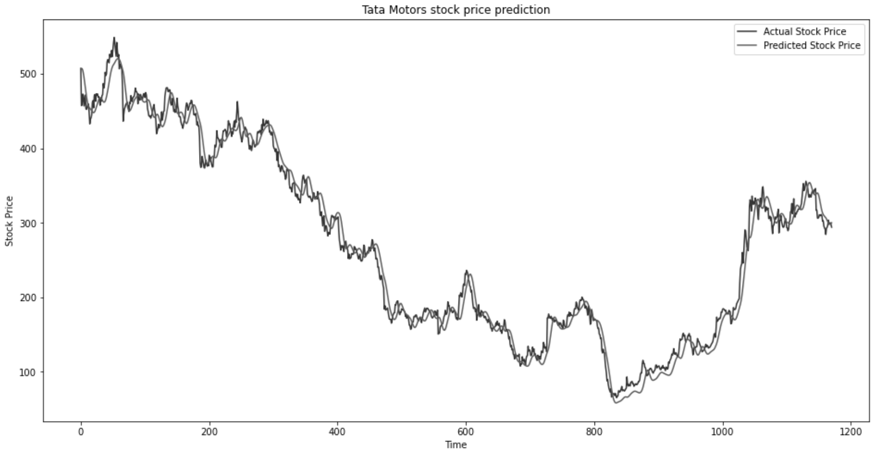

Figure 4. ANN model Tata Motors Stock Price vs Time

From the above graph, we can check that the values predicted by the model closely resemble the actual value and this also shows that the model is a good fit. It is an LSTM ANN model.

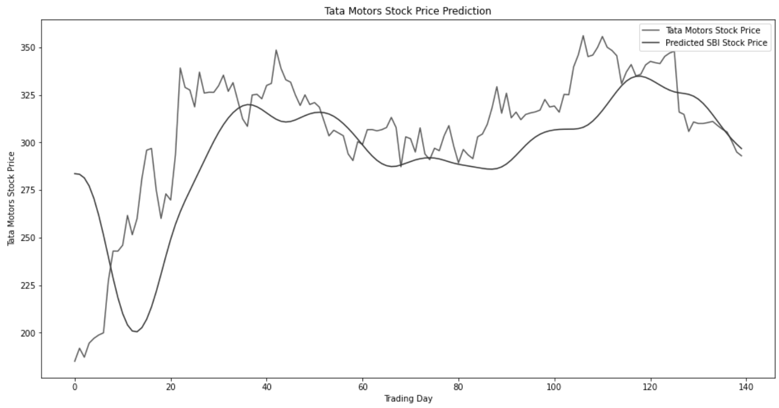

Figure 5. LSTM Regression Stock Price Prediction graph

From the above graph, in the first, the loss of the model is so much that the prediction error is more, but after some epoch, the predicted values is similar to the actual stock values. For this model, Adam optimization is used. Adam optimization is a stochastic gradient descent method that is based on the adaptive estimation of first-order and second-order moments [4]. The sparse implementation of this algorithm does apply momentum to variable slices even if they were not used in the forward pass. Momentum decay is also applied to the entire momentum accumulator. This means that the sparse behaviour is equivalent to the dense behaviour.

3.2. UCM:

In UCM, the model was specified and the season variables and slope variables are mentioned as a dummy to calculate. The model was estimated for the past 365 days to predict the future and the forecast was done for the past 30 days and the future 30 days. In the estimation phase, many useful parameters are estimated and a variety of goodness of fit statistics are displayed. This paper contains statistics, prediction error for the model, residual vs percentage graph and the forecasting graph.

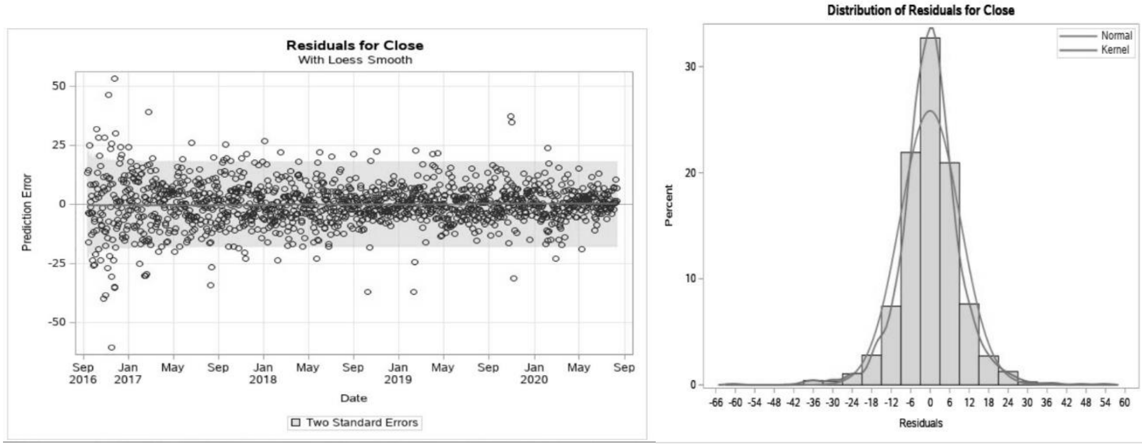

Figure 6. Data vs Prediction Error and Residual vs Percentage

Fit Statistics Based on Residuals	
Mean Squared Error	85.76165
Root Mean Squared Error	9.26076
Mean Absolute Percentage Error	2.69430
Maximum Percent Error	21.60754
R-Square	0.99593
Adjusted R-Square	0.99591
Random Walk R-Square	-1.55174
Amemiya's Adjusted R-Square	0.99588
Number of non-missing residuals used for computing the fit statistics = 1428	

Figure 7. Statistics of UCM

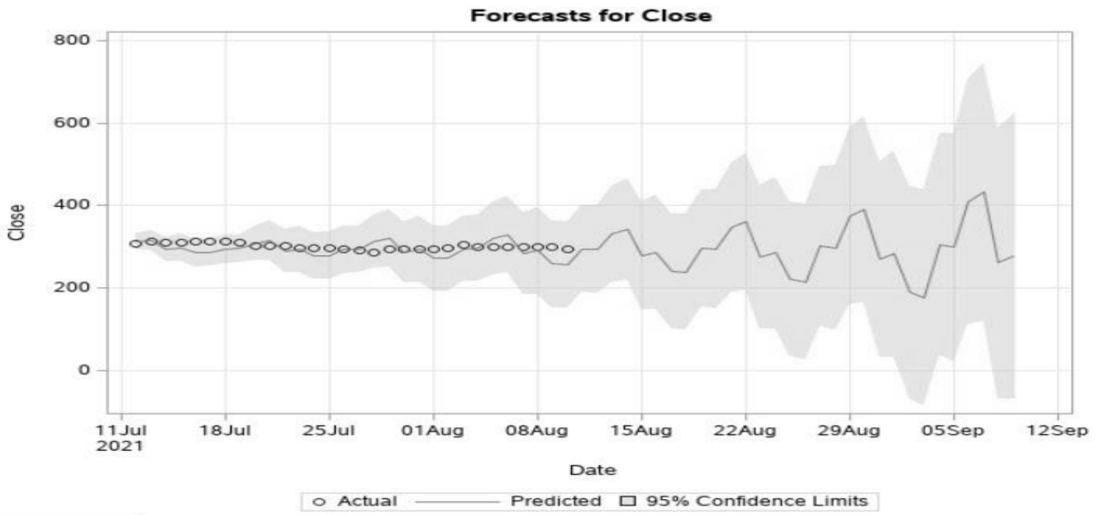

Figure 8. Forecasting of Close vs Date

3.3. ARIMA:

ARIMA models are denoted with the notation ARIMA (p, d, q). These three parameters account for seasonality, trend, and noise in data: `

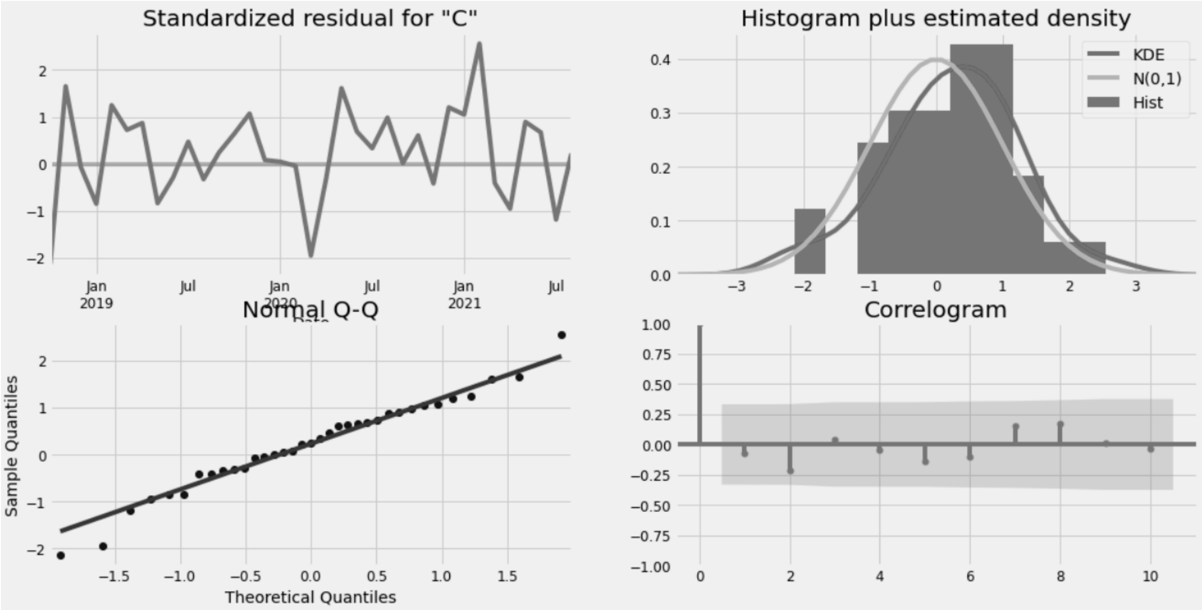

Figure 9. Residual and Correlation

Validating forecasts

To help us understand the accuracy of forecasts, the model was compared to the predicted close price to the real close price in time-series, and the forecast was started at 31-12-2017 to the end of the data.

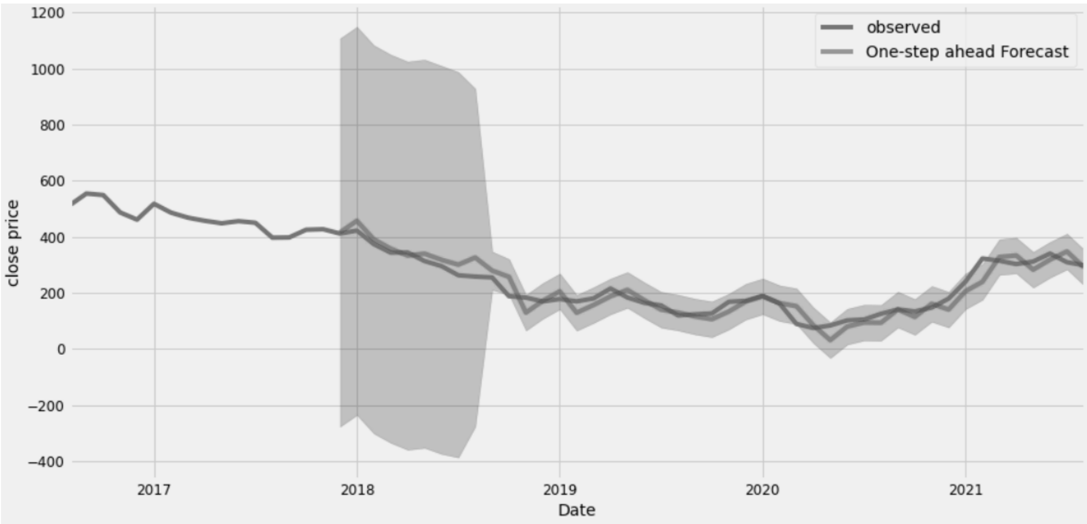

Figure 10. Prediction of close Price

The line plot shows the observed values compared to the rolling forecast predictions. Overall, forecasts align with the true values very well, showing an upward trend that starts from the beginning of the year.

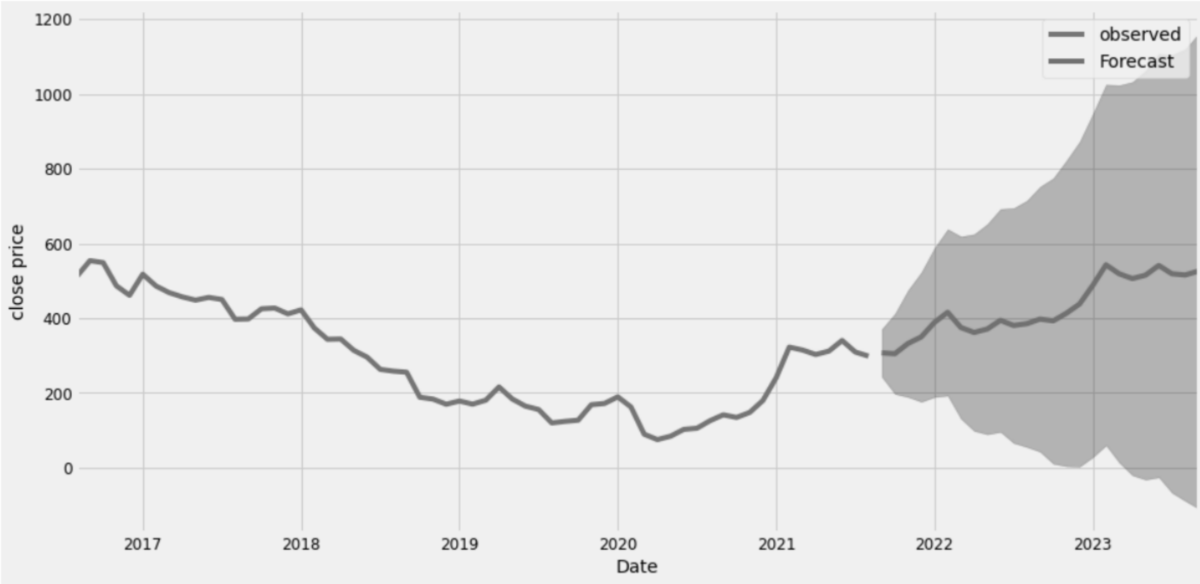

Figure 11. Prediction of Future Stock Close Price

the model captured close price seasonality. As we forecast further out into the future, it is natural for us to become less confident in values. This is reflected by the confidence intervals generated by our model, which grow larger as we move further out into the future.
The standard metrics that are used to compare the time series data will be Mean Squared Error, Root Mean Squared Error and Mean Absolute Percentage Error. The lower the value of MSE and RMSE the better the fit of the model.

Mean Squared Error: it measures the average of squares of errors i.e., the average squared difference between the actual values and the predicted values.

$$\mathrm{MSE} = \left(\frac{1}{\mathrm{n}}\right) * \Sigma(\mathrm{actual} - \mathrm{predicted})^2$$

Equation (7) Mean Squared Error Equation

Where,

n	=	number of total items
actual	=	actual or original value
predicted	=	forecasted value or value from regression

Root Mean Squared Error: it measures the difference between values predicted by a model and the values measured.it represents the square root of the difference between predicted values and the observed values of these differences. These deviations are called Residuals.

$$\mathrm{RMSE} = \left(\sqrt{\frac{\Sigma\,(\mathrm{predicted}-\mathrm{actual})^2}{\mathrm{total\ predictions}}}\right)$$

Equation (8) Root Mean Squared Error Equation

Mean Absolute Percentage Error: it is a measure of prediction accuracy of a forecasting method in statistics.it is commonly used as a loss function for regression problems and in model evaluation.

$$\mathrm{MAPE} = \mathrm{sum}\left(\frac{\mathrm{abs(predicted - actual)}}{\mathrm{total\ predictions}}\right) * 100$$

Equation (9) Mean Absolute Percentage Error Equation

Table1 Comparison of Techniques with Statistics

Model	MSE	RMSE	MAPE
ARIMA	1032.73	32.14	25.69
ANN	249.53	15.79	12.35
LSTM Regression	993.72	31.523	22.65
UCM	86.7	9.31	26.65

From the above table, we can clearly see that MSE and RMSE values are lower to UCM than any other model.

4. Conclusion

As researchers and investors strive to outperform the stock market for investing, the need for the prediction of stock price will be a progressing area in research [5,6]. The main motive of the investors is to increase their growth and investment in the stock market. Through this research work, the process and the results of ARIMA, ANN and UCM one can conclude that in MSE UCM is performing well due to the seasonality check. After UCM, LSTM is performing well due to ANN. The accuracy of ARIMA needs to be improved by improving the white noise test. While the percentage error of LSTM is very low when compared to ARIMA and UCM.

The main disadvantage in stock price prediction is that it will not change with time, but the stock price will also be affected by economic factors, socio-political factors, investments in companies, current events and the listings of rival companies. The above three mentioned models are not considering these events while predicting the stock price. To predict the stock price more accurately we need to involve these events and also adding the annual budget of the company while forecasting using the sentimental analysis. Like a new model can be made by training the model with ANN and predicting the stock price with ARIMA while using the sentiments, this can be a future research work for Stock price prediction.

References

1. Kumar, D. and Murugan, S., (2013). Performance analysis of Indian stock market index using neural network time series model. 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering.
2. Ma, Q., (2020). Comparison of ARIMA, ANN and LSTM for Stock Price Prediction. E3S Web of Conferences, 218, p.01026.
3. Charkha, P., (2008). Stock Price Prediction and Trend Prediction Using Neural Networks. 2008 First International Conference on Emerging Trends in Engineering and Technology,

4. Deorukhkar, O., Lokhande, S., Nayak, V. and Chougule, A., (2019). Stock Price Prediction using combination of LSTM Neural Networks, ARIMA and Sentiment Analysis. International Research Journal of Engineering and Technology (IRJET), 06(03), pp.3497-35003.
5. Agrawal, N., (2019). Stock Market Prediction Approach: An Analysis. International Journal of Engineering Research & Technology (IJERT), 06(03), pp.847-849.
6. Haider Khan, Z., Sharmin Alin, T. and Hussain, A., (2011). Price Prediction of Share Market Using Artificial Neural Network 'ANN'. International Journal of Computer Applications, 22(2), pp.42-47.
7. Selvamuthu, D., Kumar, V. and Mishra, A., (2019). Indian stock market prediction using artificial neural networks on tick data. Financial Innovation, 5(1).
8. Poorna Shankar and Pawar, (2011). Classification of Global Carbon Emissions using Artificial Neural Networks. International Journal of Computer Applications, [online] Available at: <https://www.ijcaonline.org/archives/volume29/number3/3544-4859> [Accessed 20 August 2021].
9. Poornashankar., (2013). Performance Analysis of Different Feed Forward Networks in Non-Linear Classification. International Journal of Soft Computing and Engineering (IJSCE), [online] Available at: <https://www.ijsce.org/wp-content/uploads/papers/v3i2/B1532053213.pdf> [Accessed 20 August 2021].
10. Support.sas.com. (2021). PROC ARIMA. [online] Available at: <https://support.sas.com/documentation/onlinedoc/ets/132/arima.pdf> [Accessed 20 August 2021].
11. Gharibvand, S. and Gharibvand, L., (2010). Financial Analysis Using SAS ® PROCS. SAS Global Forum 2010.
12. Fister, D., Perc, M. and Jagrič, T., (2021). Two robust long short-term memory frameworks for trading stocks. Applied Intelligence.
13. Ariyo, A., Adewumi, A. and Ayo, C., (2014). Stock Price Prediction Using the ARIMA Model. 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation.
14. Vijh, M., Chandola, D., Tikkiwal, V. and Kumar, A., (2020). Stock Closing Price Prediction using Machine Learning Techniques. Procedia Computer Science, 167, pp.599-606.
15. Support.sas.com. (2021). PROC UCM. [online] Available at: <https://support.sas.com/documentation/onlinedoc/ets/132/ucm.pdf> [Accessed 20 August 2021].

***