Design of Machine Learning Classifier for Stock Market Prediction

HIMANSHU PANDEY[1], CHIRAYU NEGI[2], PRAKHER NIGAM[3], RISHABH JAIN[4]
[1, 2, 3, 4] *Department of Computer Science and Engineering Bharati Vidyapeeth College of Engineering College, Pune, India*

Abstract— The stock market is extremely difficult to predict due to its complexity. There are various machine learning techniques are available to predict the diseases [7,11], stock market and so on. There are on predicting the stock market values. In current scenario in stock market forecasting are done using the Machine Learning and artificial Intelligence which makes the prediction process easier and based on the values of current stock rate by training on the previous values. There are different kind of model that can helps in predicting the stock market. We studied about many models like LSTM, ARIMA model etc. and at the end we seen that For short time series, ARIMA is one of the best models for predicting stock market prices.

Indexed Terms— ARIMA model, LSTM, ACF, PACF, AIC

I. INTRODUCTION

The stock market is a financial network that provides a platform for practically every large-scale economic transaction in the globe at a dynamic rate based on market equity known as the stock market value.

There are many technologies which are used to resolve this issue related to stock market prediction such as ANN, Fuzzy Logic and SVM. Recently, The ARIMA method was used for this problem in predicting the pattern. ARIMA has been done successful job in the field of analyzing and predicting the time series. ARIMA is best known for short term prediction. In this article, we use a daily fractional change in the stock value.

II. LITERATURE SURVEY

In the Korean Stock Exchange, Lee et al. offered a comparison of the forecasting technique and dependability between the BPNN model and a time-series (SARIMA) model [2].

Rafiqul et *al* presented a comparative study of three financial models ARIMA, ANN, and Geometric Brownian Motion, that helps to forecast the future prices of the stock market [3]. The ARIMA model and the stochastic model can both be used for short-term prediction utilising time series data.

Devi et *al*, shown in their paper is that inferences a new investment decision which is based on the less error percentage obtained [5]. This paper also highlighted the point on the next few years' future forecasting of each and every index.

III. PROPOSED APPROACH

- ARIMA MODEL

The ARIMA model is an acronym for (AutoRegressive Integrated Moving Average). This model is a combination of moving average and autoregressive models.It's ideal for forecasting short time series.Time series analysis typically requires stationary data, however stock market data is nonstationary.The future value of a variable is determined by a linear combination of previous errors and past values in this model.

This is represented as: -

$Yt = \varphi 0 + \varphi 1 Yt{-}1 + \varphi 2 Yt{-}2 + \ldots + \varphi p Yt{-}p + \epsilon 1 - \theta 1 \epsilon t{-}1 - \theta 2 \epsilon t{-}2 - \ldots - \theta q \epsilon t{-}q$

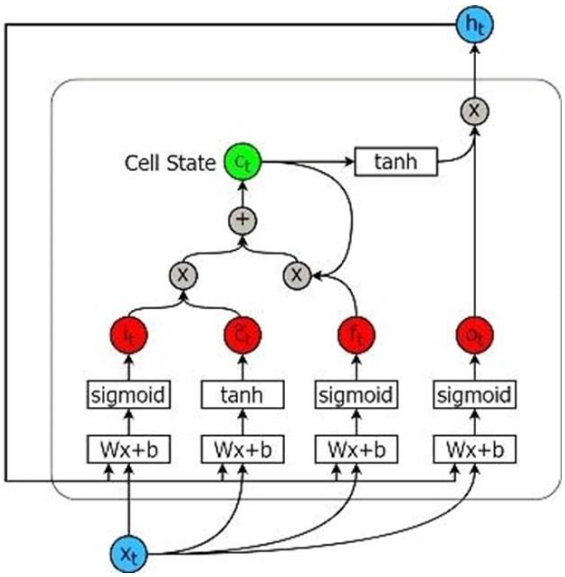
Where,

Actual value is Yt,$\varphi_x$and $\theta_y$are coefficients,$\epsilon_t$is random error at t, p and q are integers which is referred as autoregressive and moving average respectively.

- LSTM MODEL
- Hochreiter & Schmidhuber introduce the LSTM model (1997).LSTMs are designed to avoid the problem of long-term dependency.This model is capable of predicting any number of steps in the

future.It can be used to model long-term and short-term data.

- Cell state (c_t) – It represent the internal memory of the cell which stores both short term memory and long-term memories
- Hidden state (h_t) – This is the output state information which stored the previous calculated hidden state, current input, and current cell input which is eventually used to predict the future stock market values.
- Input gate (i_t) – It is used to decides how much information flows from current input to the cell state.
- Forget gate (f_t) - It is used to Decide how much information from previous cells and the current input cells flows into the current cell state
- Output gate (o_t) – It Decides how much information flows from the current cell state into the hidden state, that helps to choose the long-term memories or short-term memories and long-term memories



IV. METHODOLOGY

The basic three sections are found in the methodology section. The data that was utilised to generate the models is described in the first subsection. Then, in each subsection, the general theories and processes for building the models are described. The overall performance of each of the models was assessed using residual analysis and various error measures, including mean absolute error (MAE), mean squared error (MSE), root mean square error (RMSE), and the final R2 score, also known as 2 (coefficient of determination) regression score, which provides information about a model's goodness of fit.

$$MAE = \frac{\sum_{ni=1} |y_i - x_i|}{n}$$

$$MSE = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n}$$

$$RMSE$$

$$^2 = 1 - RSS$$

$$R$$

$$TSS$$

- Dataset

This post utilises Dell daily stock from August 17, 2016 to May 21, 2021. To acquire data directly from Yahoo Finance, we used the Pandas-Datareader library in Python software. The dataset starts out with six variables: daily open, close, high, low, volume, and adjusted Close price. All of the models were created with the goal of predicting the next day's close price based on the previous day's data.

- Arima

The DELL stock price closing is a time series that was analysed to create the model. The time series is non-stationary, as shown in the graph below. The graph in Figure 1 shows an upward trend. The Auto Correlation Function (ACF) goes down slowly, and the Partial Autocorrelation Function (PACF) function shuts off at lag 1 with correlation one, as shown in Figure 2.

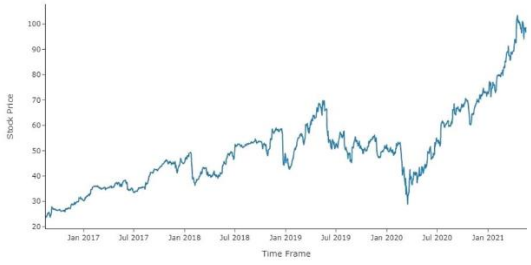


Figure 1. Time plot of the raw data

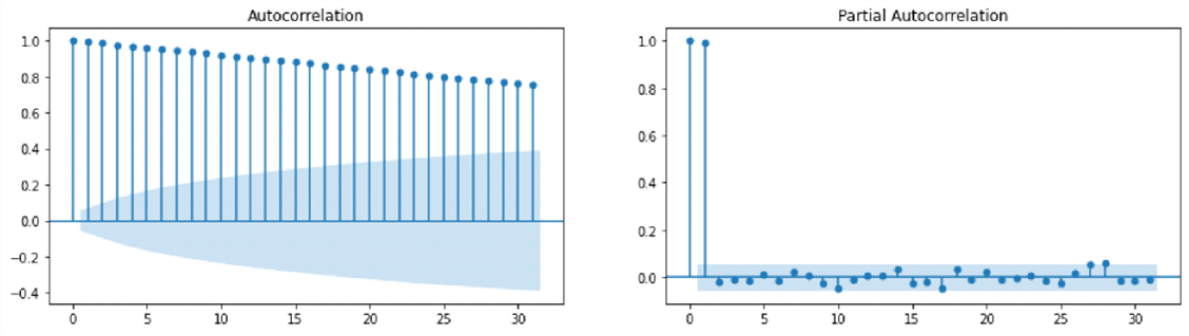


Figure 2. Sample autocorrelation plot

The ADF test's null hypothesis is that the time series is non-stationary. We can reject the null hypothesis and deduce that the time series is truly stationary if the pvalue of the test is less than the significance level (0.05). If the p-value is more than 0.05, we must determine the order of differencing.

ADF Statistic	-0.18494419386651875
n_lags	0.9402925963938609
p-value	0.9402925963938609
Critial Values: 1%	-3.435829423619109
Critial Values: 5%	-2.863959622178626
Critial Values: 10%	-2.5680582513898056

Table 1.ADF test results.

As we can see, the p-value for the data is more than 0.05, or 0.940, hence we must choose the order of differencing. We used the ndiffs function from the pmdarima python package to obtain the order of differencing value and got the value 1.

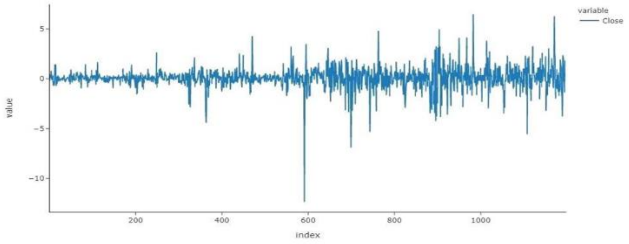


Figure 3. Plot of the first differenced log transformed stock price.

Now the autoregressive and moving average orders p and q were determined from the PACF and ACF plot of data from Figure 4.

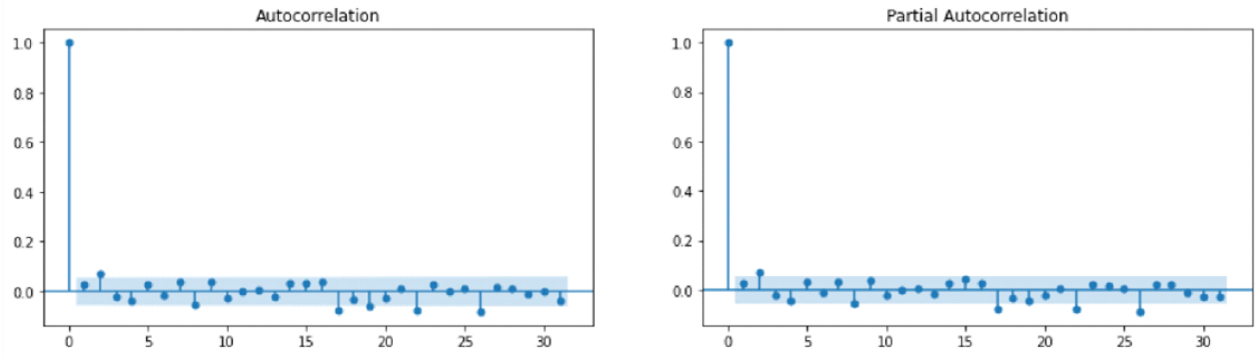


Figure 4. ACF and PACF of the first differenced log transformed stock price.

Model	AIC	Time
ARIMA (0,1,0)	2330.232	0.03 sec
ARIMA (0,1,1)	2322.378	0.06 sec
ARIMA (1,1,1)	2322.632	0.16 sec
ARIMA (0,1,2)	2324.120	0.10 sec
ARIMA (1,1,0)	2322.751	0.05 sec
ARIMA (1,1,2)	2322.900	0.46 sec

Table2. ARIMA (0,1,1) model comparison

Error Measures	MAE	MSE RMSE	R2 Score
	1.0585526723799226	2.1487177269 1582 1.4658505131 546735	0.9932782407 357383

Table 3. Error measures of ARIMA (0,1,1)

From table 1 the ARIMA (0,1,1) model has minimum AIC value. In Table 2 there are 3 error factors MAE, MSE, RMSE. R2 score shows the Goodness of fit of the model.

- LSTM

Data pre-processing is the first phase after data collecting and is used for data transformation, data cleansing, and data integration. Data normalisation is important in data transformation, and MinMaxScaler scales all of the data to be between 0 and 1. The dataset is separated into training and testing sets once it has been normalised and cleaned. The testing data makes for 30% of the overall dataset.

Table4. LSTM model summary

Error Measures	MAE Score	MSE RMSE	R2
	1.5855925139757 4.5146548497782 2.124771717092707	0.9857818665166186	

Table5. Error measures of LSTM

V. RESULT

We reviewed the aforementioned two models in this section, as well as a comparison of actual and expected prices, which is represented graphically.

- ARIMA

Calculating the error we have predited :

$$Error = \frac{Actual - Predicted}{Actual} \times 100$$

Actual	Predicted	Error
• 50	49.97661	0.046578
• 50.00259		-0.22268
• 49.62	49.89748	-0.50891
• 49.59137		1.721031
• 50.83	50.55695	0.548376
• 50.86041		2.210914
• 50.98	52.103819	-2.27885
• 50.85306		0.111842
• 51.39	50.91673	0.920906
• 1.621673		
• 50.58	52.98352	-3.50617
• 50.37357		1.382905
13 49.8	51.15498	-2.72085
14 50	49.65782	0.684369
15 49.62	50.03563	-0.83762
16	49.57673	-0.13479
17 49.83	49.50305	0.656138
18	49.86406	1.958199
19 50.6	50.96416	-0.71968

Table6. Prediction by ARIMA (0,1,1) model.

Table 5 shows that the relative errors for the daily forecast are fewer than 4, with relative errors ranging from -3.58617 to 2.210314. The graph of the actual and forecasted stock prices by the Arima model is shown in Figure 5. The actual stock price for Dell is represented by the blue line, while the expected stock price for Dell is represented by the orange line. Figure 5 also demonstrates that the ARIMA (0,1,1) forecasted prices closely track the actual price trend. The ARIMA(0,1,1) model's performance was assessed using the table 2 error measure, and table 5 displays the contrast between test and projected results. Figure 6 depicts a zoomed-in version of Figure 5.

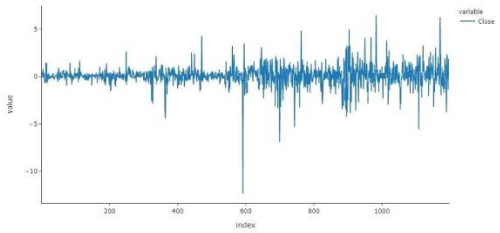
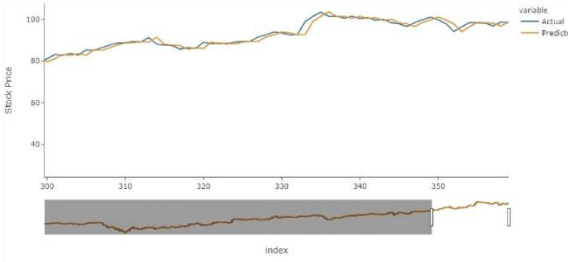


Figure 5. ARIMA (0,1,1) model prediction.

Figure 6. Zoom view ofARIMA (0,1,1) model prediction.



LSTM

$$Actual - Predicted \;\; Error = \frac{}{Actual} \times 100$$

Actual	Predicted	error
1		-1.1277
2		-1.37451
49.82	50.43788	-1.94828
50.40	50.18669	0.620805
50.83	51.00684	-0.52597
52.51	51.30624	1.18607
50.99	52.64707	-3.27904
50.94	51.33383	-0.87079
51.49	51.49783	-0.0357
52.89	51.93032	0.678492
50.57	52.85579	-4.52704
51.88	50.8366	0.480469
49.8	51.62289	-3.66042
50.15513		-0.31026
49.62	50.53177	1.83751
50.09539		-1.18236
49.83	50.03061	0.40258
50.40498		0.894646
19 50.6	51.50288	-1.78434

Table7. Prediction by LSTM model.

Table 6 shows that the relative errors for the daily forecast are less than 5, with relative errors ranging from -4.52004 to 1.18007. Figure 7 depicts a graph of the actual data and the LSTM model's predicted stock price value. The blue line in this graph depicts Dell's actual stock price, while the orange line represents its forecasted stock price. The performance of this model is assessed using the table 6 error measure and the table 4 data. It shows the difference between what happened and what was predicted. Figure 8 is a zoomed-in version of Figure 7.

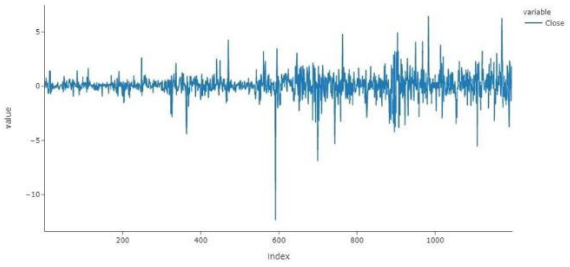


Figure 7.LSTM model prediction

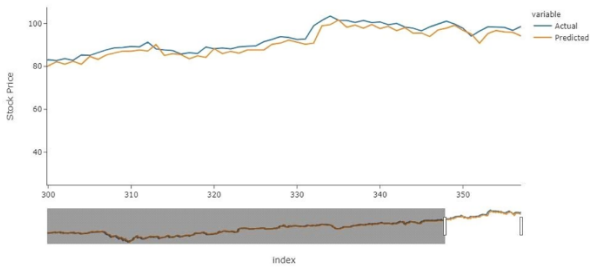


Figure 8. Zoom view of LSTM model prediction

VI. CONCLUSION

The combined outcome of the two models outlined before is shown in this section. Table 7 illustrates the experimental output generated from the supplied models, whereas Figure 9 graphically depicts the result.

	Actual	ARIMA	LSTM
1	50	49.97671	50.70328
2	49.89	50.00259	50.70702
3	49.62	49.87748	50.55855
4	50.46	49.59136	50.26105
5	50.83	50.55635	51.24133
6	52.01	50.86041	51.53958
7	50.98	52.13819	52.81548
8	50.91	50.85306	51.45399
9	51.39	50.91623	51.51614
10	52.29	51.44202	52.05922
11	50.57	52.38352	53.0074
12	51.08	50.37357	50.91423
13	49.8	51.15498	51.7465
14	50	49.65782	50.23284
15	49.62	50.03563	50.65029
16	49.51	49.57673	50.20906
17	49.83	49.50305	50.1547
18	50.86	49.86406	50.54063
19	50.6	50.96416	51.66629

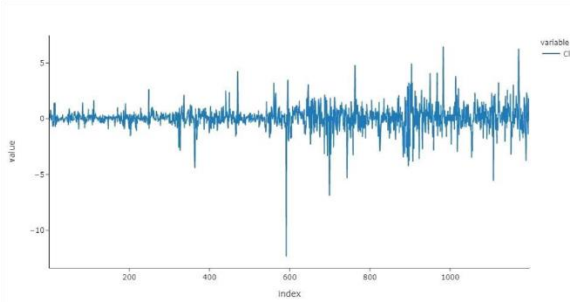


Table 8. Sample results from the models – ARIMA (0,1,1) and LSTM

Figure 9.Prediction by ARIMA (0,1,1) and LSTM against actual price.

Figure 10.Zoom view ofprediction by ARIMA (0,1,1) and LSTM against actual price.

From Figure 10it clear that that ARIMA (0,1,1) model's output and LSTM model's output are veryclose, sometimes they coincide.

Error Measures	MAE	MSE	RMSE	R2 Score
ARIMA	1.0585526723799226			0.9932782407357383
LSTM	1.5855925149757775	4.514654849757082	2.1247717170927047	0.9857816665164186

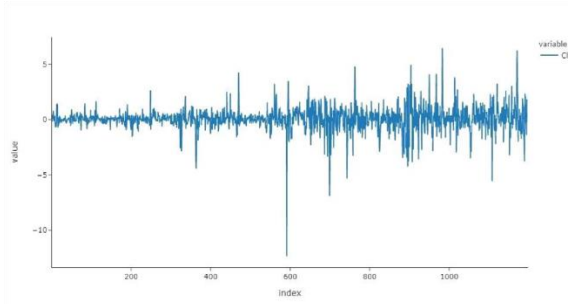


Table 9. Error measures comparison between ARIMA (0,1,1) and LSTM

When comparing the error metrics in table 8, it is evident that the ARIMA model outperforms the LSTM model when it comes to predicting the next day stock price.

The goal of this research is to compare the performance of an LSTM model and a timeseries ARIMA model in terms of predicting. We discover the following using DELLdata: For starters, the ARIMA model produces better DELL outcomes than the LSTM model. Second, the ARIMA model has a lower error rate than the LSTM model.

ACKNOWLEDGEMENT

We would like to acknowledge the support of Department of Computer Engineering, Bharati Vidyapeeth College of Engineering for their support and guidance , they have provided especially Mrs Rohini Jadhav ,Assistant Proffessor, Computer Engineering.

REFERENCES

[1] Ayodele A. Adebiyi., Aderemi O. Adewumi, Charles K. Ayo, "Stock Price Prediction Using the ARIMA Model", "2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation", 2014.

[2] Kyungjoo Lee, SehwanYoo, John JongdaeJin., "Neural Network Model vs. SARIMA Model In Forecasting Korean Stock Price Index (KOSPI), "Issues in Information System, 8(2), 372378.", 2007.

[3] Mohammad Rafiqul Islam, Nguyet Nguyen, "Comparison of Financial Models for Stock Price Prediction", "*J. Risk Financial Manag.* 2020, *13*(8), 181" ,2020.

[4] Mohammad Almasarweh & S. AL Wadi, "ARIMA Model in Predicting Banking Stock Market Data", "Modern Applied Science; Vol. 12, No. 11; 2018" ,2018.

[5] B. Uma Devi, D.Sundar and Dr. P. Alli, "An Effective Time Series Analysis for Stock Trend Prediction Using ARIMA Model for Nifty Midcap-50", "International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.1, January 2013", 2013.

[6] Aloysius Edward, Jyothi Manoj, "FORECAST MODEL USING ARIMA FOR STOCK PRICES OF AUTOMOBILE SECTOR", "International Journal of Research in Finance and Marketing (IMPACT FACTOR – 5.861)", 2016.

[7] Srivastava, A. K., Kumar, Y., & Singh, P. K. (2020). A Rule-Based Monitoring System for Accurate Prediction of Diabetes: Monitoring System for Diabetes. *International Journal of E-Health and Medical Communications (IJEHMC), 11*(3), 32-53. doi:10.4018/IJEHMC.2020070103

[8] Adil MOGHAR, Mhamed HAMICHE, "Stock Market Prediction Using LSTM Recurrent Neural Network", "International Workshop on Statistical Methods and Artificial Intelligence (IWSMAI 2020)", 2020.

[9] RangsanNochai, TitidaNochai, "ARIMA MODEL FOR FORECASTING OILPALM PRICE", "Proceedings of 2nd IMT-GT Reginal Conference on Mathematics, Statistics and Applications UniversitiSains Malaysia, Penang, June 13-15, 2006", 2006.

[10] MurtazaRoondiwala, Harshal Patel, Shraddha Varma, "Predicting Stock Prices Using LSTM", "International Journal of Science and Research (IJSR)", 2015.

[11] Srivastava A.K., Singh P.K., Kumar Y. (2019) A Taxonomy on Machine Learning Based Techniques to Identify the Heart Disease. In: Prateek M., Sharma D., Tiwari R., Sharma R., Kumar K., Kumar N. (eds) Next Generation Computing Technologies on Computational Intelligence. NGCT 2018. Communications in Computer and Information Science, vol 922. Springer, Singapore.

HYPERLINK "https://doi.org/10.1007/978-981-15-1718-1_2"https://doi.org/10.1007/97 HYPERLINK "https://doi.org/10.1007/978-981-15-1718-1_2"8 HYPERLINK "https://doi.org/10.1007/978-981-15-1718-1_2"- HYPERLINK "https://doi.org/10.1007/978-981-15-1718-1_2"98 HYPERLINK "https://doi.org/10.1007/978-981-15-1718-1_2"1 HYPERLINK "https://doi.org/10.1007/978-981-15-1718-1_2"- HYPERLINK "https://doi.org/10.1007/978-981-15-1718-1_2"1 HYPERLINK "https://doi.org/10.1007/978-981-15-1718-1_2"5 HYPERLINK

"https://doi.org/10.1007/978-981-15-1718-1_2"-
HYPERLINK "https://doi.org/10.1007/978-981-
15-1718-1_2"171 HYPERLINK
"https://doi.org/10.1007/978-981-15-1718-1_2"8
HYPERLINK
"https://doi.org/10.1007/978-981-15-1718-1_2"1_
HYPERLINK "https://doi.org/10.1007/978-
981-15-1718-1_2"2
HYPERLINK
"https://doi.org/10.1007/978-981-15-1718-
1_2"
HYPERLINK "https://doi.org/10.1007/978-981-15-
1718-1_2"