Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 45 (2024), No. 4, pp. 1129–1139
DOI : 10.47974/JIOS-1697

Stock market prediction using ARIMA-LSTM hybrid

Aayushi Pandya *
Vivek Kapoor †
Department of Information Technology
Institute of Engineering & Technology
Devi Ahilya Vishwavidyalaya
Indore
Madhya Pradesh
India

Apash Joshi §
Business Information Systems
University of Applied Sciences and Arts Northwestern Switzerland FHNW
Windisch
Switzerland

Abstract

The nation's economy greatly depends on the stock market. A healthy stock market can boost the economy, but if not going well, it can also trigger a recession. It is one of the most volatile markets, it is subjected to sudden changes due to variety of factors including interest rates, supply and demand, inflation, political issues, natural disasters, etc. Therefore, accurate stock market forecasting is crucial in order to help stock customers. Many models which have already been worked on for predicting stock prices are using techniques like ARIMA and LSTM. But hybrid models are now being developed for better outcomes. This paper proposes composition of ARIMA and LSTM for stock prediction and compares the results of ARIMA, LSTM and ARIMA-LSTM hybrid. To test the effectiveness of the hybrid model to others, the model is applied to 5 businesses from various industries. For comparison, different errors are taken into account.

Subject Classification: *Primary 93A30, Secondary 49K15.*

Keywords: *Stock market, Stock prediction, Hybrid model, ARIMA, LSTM, Neural networks.*

* *E-mail:* **aayushipandya333@gmail.com (Corresponding Author)**
† *E-mail:* **vkapoor@ietdavv.edu.in**
§ *E-mail:* **joshiapash017@gmail.com**

1. Introduction

Humans have been attempting to make life simpler since the beginning of life on this planet. And now we are expecting machines to solve complex real world problems like a human brain would do. We are attempting to have machines which can learn enormous amounts of prior data and deliver relevant insights from that data after meticulously analyzing it because an average human can only learn up to a certain limit. These insights, which machines produce after studying a lot of past data, are used to forecast future data. Predictions are needed in many areas of life in order to leave a comfortable and luxurious existence and make increasing amounts of profits. One such area that demands future forecasting is the stock market. Although lot of effort is put into forecasting stock prices, but still doing it as accurately as possible is still exceedingly difficult. As previously mentioned, stock prices are highly unstable and dependent on a variety of factors, so even if a lot of work has already been done, the accuracy still falls short of expectations. Therefore, there is still a lot of space for advancement in this field of prediction. Future stock prices are exceedingly difficult to anticipate because of how widely stock prices fluctuate, which makes the market unstable. Online data availability has increased dramatically over the past few years, creating data explosion. It is impossible to manually analyze the vast amount of data available on web. Additionally, data for each and every company is accessible, so if an investor wants to buy a share that would provide the most return, he must examine the data for all the businesses he is considering. If analyzing from vast amounts of data for a single company is challenging, so it will multiply number of times for number of companies. And because of this, several changes to currently used techniques are needed to produce the best predictions.

For time series analysis, there are essentially two perspectives: the statistical perspective and the artificially intelligent perspective. In statistical perspective, statistics is applied on previous data whereas in artificially intelligent perspective different machine learning algorithms are applied to historical data in order to uncover some insights or trends. The future prices of equities are then predicted using these findings.

2. Literature Survey

Given the significance of the stock market, numerous investigations have been carried out by various scholars in this field. Machine learning has been extensively used to obtain accurate findings that can generate large profits for stockholders through wise decision-making. The following list includes a few of these studies. It is reported a method for developing astock price prediction model based on ARIMA [1 - 4]. The graphed data indicated that the performance was satisfactory. ARIMA competed fairly with then-current methods for predicting short-term stock prices. This is also a disadvantage because the ARIMA model is only appropriate for making short-term forecasts with high accuracy. Researchers [5, 6] applied RNN on stock data of a decade of Apple (AAPL). The findings demonstrated a good fit between the predicted and actual stock price curves. The calculated MAE fell into an acceptable range. RNN can accurately predict stock values when time steps (the length of the preceding input) is small .Since RNN cannot hold long time memory, the MAE grows higher as the timesteps rise. In order to evaluate the effectiveness of LSTM in stock prediction, Few Researchers [7-9] picked stock data of businesses from a variety of industries. The calculated RMSE fell into a range that is acceptable. The ideal configuration seems to be one hidden layer. If the number of hidden layers were to expand, over fitting problems might readily arise and result in high RMSE values. By taking into account numerous hyper-parameters, error can be reduced. More room exists for mistake reduction. Using genetic algorithms and Neural Networks people [10,11,12] attempted to forecast whether stock values would rise or fall. To obtain variation in data, they took into account the stock prices of eight different companies. They considered Adobe, Apple, Google, IBM, Microsoft, Oracle, Sony, and Symantec. The genetic algorithm yielded maximum accuracy of 73.87%, and evolutionary techniques yielded maximum accuracy of 71.77%. Certainly, accuracy needs to be increased. Other than these, all other algorithms, including Logistic Regression, Naive Bayes, Random Forest, etc., only offer accuracy ranging from 45.1% to 86.2%.

3. Background

*A. ARIMA (Auto Regressive Integrated Moving Average)*has 3 components :AR (Autoregression) :An AR model determinesthe relationship between its lag data from the past and uses that relationship to forecast the future. p value(order of AR model) - Indicates how many prior values were taken into account.

According to AR model,

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \text{-------} + \alpha_p y_{t-p} + \varepsilon_t$$

Where,

y_t = Actual value at time t

α_i = coefficients

p = order of AR

ε_t = Error value (Difference between actual value and calculated value of y_t)

MA (Moving Average) :According to the MA model, Presentvalue is not only influenced by the past Values but also by the errors of past values.

q value(order of MA) – shows number of error terms considered.

According to MA model,

$$y_t = \beta_0 + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \text{-------} + \beta_q \varepsilon_{t-q} + \varepsilon_t$$

Where,

y_t = Actual value at time t

β_i = coefficients

q = order of MA

ε_t = Error value at time t

I (Integrated) **:**'Integrated' is a concept that derives from 'difference'. In order to remove trends and seasonality and make data stationary, we difference the previous data from current data.d value(order of I) – Number of times we difference the data to make it stable.

$$z_t = y_t - y_{t-1}, \text{ when d = 1}$$

Final Equation of ARIMA model is :

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \text{-------} + \alpha_p y_{t-p}$$
$$+ \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \text{-------} + \beta_q \varepsilon_{t-q} + \varepsilon$$

p, d and q are all hyperparameters used in ARIMAmodel and found by various methods.

B. *LSTM(Long Short Term Memory)*

LSTM is a Neural Network. More specifically, it is a type of RNN. RNN has a limited capacity for storing historical data, while LSTM has the ability to remember crucial information while forgetting less crucial information. LSTM can store more significant data than a simple RNN. Each LSTM block includes:

A basic RNN cell : **that serves as the foundation**

for **LSTM** cellCell state: Used for long term memory. Contains memoryfrom previous cell to next cell. Cell state is continuously updated. If any information is not necessary, it will be removed, and the information that is necessary will be added. The following 3 gates are used to accomplish this.Forget gate – The forget gate outputs either 0 or 1. The output value of '0' indicates that the data is notimportant and will be disregarded, whereas '1' indicates that the data is important for prediction and will be remembered.Input gate – Provides current input to the cell. Output gate – Provides output of the cell. The output given by output gate will be based on cell state andthe current input.s

4. Methodology

Firstly, data of previous 8 years(from 2015 to 2023) of 5 companieswhich are Apple, Amazon, Coal India, ITC and Cognizant, each of them belonging to different sectorsis collected from yahoo finance. Selection of companies from different sectors is done so that we can get data of different trends, seasonality and residuals.Data collected is in OHLCV (Open, High, Low, Close and Volume) format and 'Closing price'is estimated. In this experiment, two methodologies—ARIMA and LSTM—that could capture both the linear and non-linear portions of time series—were applied. Initially ARIMA and LSTM are applied individually on the identical data of all 5 organizations and then on the same data ARIMA-LSTM HYBRID is applied to compare the efficiency of the hybrid model to the other two.

4.1 *Develop ARIMA model*

It is crucial to verify that the data is stationary before using the ARIMA model on time series data. The more stationary the data, the better the findings ARIMA will produce. Thus, the Augmented Dickey Fuller (ADF) test is run to determine stationarity. ADF test results, or p value, indicate whether or not data is steady. P value > 0.05 indicates a non-stationary series. Aggregation of data on monthly basis is done.Since the ARIMA model is defined by three parameters—the p value (order of AR), the q value (order of MA), and the d value (order of I)—getting the best value for each is crucial. ACF (Auto Correlation Function) and PACF

(Partial Auto Correlation Function) graphs are plotted for data from all the organizations in order to obtain ideal values of p,d and q. ACF yields "q value" while PACF yields "p value". Thus optimal p,d,q values of all 5 companies are determined. An 80:20 ratio is used to split the data into training and testing sets, and training data is fitted into ARIMA model of required order. Predictions are made of testing data and then different errors, including MAE, RMSE, and R^2, are computed for effective result comparison. The output graph is also shown between the train, test, and predicted data for visualization.

4.2 *Develop LSTM model*

Before anticipating decent results from LSTM, a variety of hyperparameters must be carefully chosen. To speed up model training, original data must be scaled in a specific range, in this case, (0,1). Scaled data splits into training and testing set in 80:20 ratio. Further train and test data is divided into the subsets xtrain, ytrain, xtest, and ytest considering a fixed timestep (60 in this case). The model is then trained using the training data once the data has been reshaped to meet LSTM standards. For LSTM training, the "adam" optimizer is utilized, while "Mean Squared Error" is employed to determine loss. After the training data has been fitted into the model, the test data is predicted, and the findings are then rescaled to it's **original form. Test data and predictions are compared, and MAE, RMSE, and R2 errors are determined. To examine how well the model performs, graph is plotted between the train, test, and forecast series.**

4.3 *Develop ARIMA-LSTM HYBRID model*

For the HYBRID model, the work that has been done so far on ARIMA and LSTM independently needs to be expanded. The same parameters (p, d, and q) from the ARIMA model are employed, however this time the emphasis is on the "Residuals" of the prediction rather than the actual prediction. Residual is essentially the difference between actual and predicted data. LSTM would receive residuals from both the train and test sets as input. Since ARIMA is good at making 'short term' forecasts and because it can handle 'Linear portion' of time series, ARIMA has been handling this section up till now.

The "Non-Linear" component of time series, or residuals, will now be predicted using the LSTM.Remainders from the training and test sets generated by ARIMA inputs as training and test sets for LSTM. Similar training methods are used to train further and predictions from LSTM are stored.Predictions from ARIMA are combined with LSTM residual predictions to produce the final predicted values. Graphs are plotted and errors are calculated.

5. Result

MAE :

Table 1

MAE of all companies

Company	ARIMA	LSTM	HYBRID
Apple	449.521	7.31	2.497
Amazon	267.170	6.029	2.906
Coal India	302.821	12.549	2.7479
ITC	465.789	4.834	3.155
Cognizant	220.542	2.380	0.995

RMSE :

Table 2

RMSE of all companies

Company	ARIMA	LSTM	HYBRID
Apple	667.155	8.589	3.204
Amazon	316.766	7.870	3.931
Coal India	380.823	13.813	3.770
ITC	572.8008	6.242	4.219
Cognizant	331.174	2.966	1.428

R-Squared :

Table 3

R-Square value of all companies

Company	ARIMA	LSTM	HYBRID
Apple	-0.029	0.524	0.933
Amazon	0.77	0.903	0.975
Coal India	0.582	0.738	0.980
ITC	0.776	0.989	0.995
Cognizant	0.145	0.933	0.984

Graphs were plotted for all 5 companies for all the three models(ARIMA, LSTM and Hybrid), which are presented below:

APPLE : ARIMA -

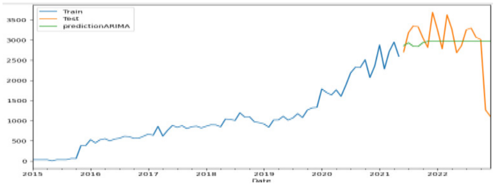

Figure 1

ARIMA model output of apple

LSTM–

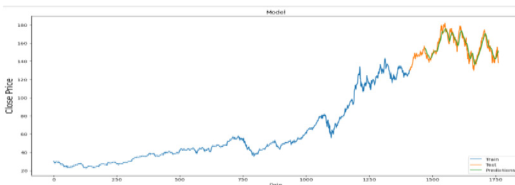

Figure 2

LSTM model output of apple

Hybrid-

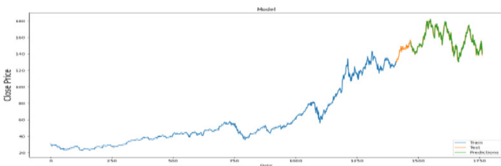

Figure 3

Hybrid model output of apple

AMAZON : ARIMA –

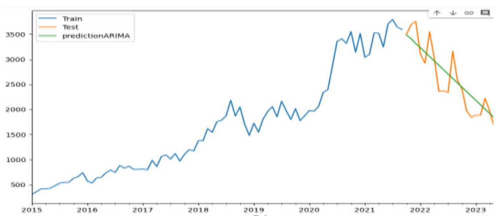

Figure 4

ARIMA model output of amazon

LSTM –

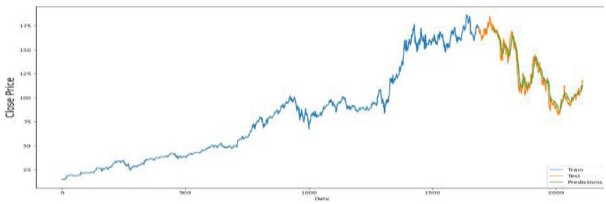

Figure 5

LSTM model output of amazon

Hybrid –

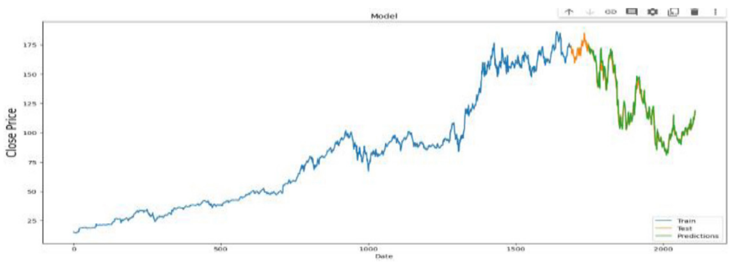

Figure 6

Hybrid model output of amazon

6. Conclusion and Future Work

It is clear from the graphs drawn that the ARIMA model does a good job of forecasting the linear section of time series see Figure 1, Figure 2, Figure 3, Figure 4, Figure 5 and Figure 6. It is capable of predicting trends but not seasonality. The best prediction method for the short term is ARIMA, but not for the long term prediction. However, LSTM can forecast non-linear portion of data. It is effective in forecasting seasonality. Although LSTM results are superior, they can still be enhanced. Hybrid outperformed than ARIMA and LSTM models separately see Table 1, 2 & 3. Graphs, as well as MAE, RMSE, and R-Square values, makes it very clear. Hybrid models are though more difficult to code, train, comprehend, and implement than separate models, but they also produce superior outcomes. The reason for this is that it builds on the strengths of both models that are being hybridized, and the hybridized model can overcome the limitations of both models when used alone. Choosing many businesses from various industries allowed us to test our model on a wide range of data. The results unambiguously demonstrated that hybrid can function very effectively on a variety of data with varying trends, seasonality, and noise.More hybrid models may be taken into consideration in the future, such as the SARIMA-ANN hybrid, and they may be trained to observe outcomes. Moreover, It is possible to consider how news and events affect stock market analysis.

References

[1] Zhu, Yongqiong. (2020). Stock price prediction using the RNN model. *Journal of Physics: Conference Series* (1650). 032103. 10.1088/1742-6596/1650/3/032103.

[2] Altameem, Ayman, et al. "P-ROCK: a sustainable clustering algorithm for large categorical datasets." *Intell. Autom. Soft Comput* 35.1 : 553-566 (2023).

[3] Singh, Vijander, et al. "Prediction of COVID-19 corona virus pandemic based on time series data using Support Vector Machine." *Journal of Discrete Mathematical Sciences and Cryptography* 23.8 : 1583-1597 (2020).

[4] Bonde, G. & Khaled, Rahma. Stock price prediction using genetic algorithms and evolution strategies. Proceedings of the 2012 International Conference on Genetic and Evolutionary Methods. Pp 10-15 (2012).

[5] Sakshi Kulshreshtha and Vijayalakshmi A, An ARIMA- LSTM Hybrid Model for Stock Market Prediction Using Live Data. *Journal of Engineering Science and Technology Review* 13 (4) pp 117 - 123 (2020).

[6] Umar Farouk Ibn Abdulrahman, Najim Ussiph, Benjamin Hayfron-Acquah, A Hybrid Arima-Lstm Model for Stock Price Prediction, *International Journal of Computer Engineering and Information Technology,* 12 (8), pp 48-51, (2020).

[7] Moghar, Adil & Hamiche, Mhamed. Stock Market Prediction Using LSTM Recurrent Neural Network. Procedia Computer Science. 170. 1168-1173 (2020). 10.1016/j.procs.2020.03.049.

[8] M, Hiransha & Gopalakrishnan, E. A & Menon, Vijay & Kp, Soman. NSE Stock Market Prediction Using Deep-Learning Models. Procedia Computer Science. 132. 1351-1362 (2018). 10.1016/j.procs.2018.05.050.

[9] Pakdaman Naeini, Mahdi & Taremian, Hamidreza & B. Hashemi, Homa, Stock market value prediction using neural networks. Pp 132 - 136 (2010). 10.1109/CISIM.2010.5643675.

[10] K.Sai Sravani, Dr. P. Raja Rajeswari, Prediction Of Stock Market Exchange Using LSTM Algorithm. *International Journal of Scientific & Technology Research,* 9. 3, pp 417-421, (2020).

[11] Vijh Mehar, Chandola Deeksha, Tikkiwal Vinay, Kumar Arun, Stock Closing Price Prediction using Machine Learning Techniques. *Procedia Computer Science*. 167. Pp 599-606. 10.1016/j.procs.2020.03.326.

[12] Jawad Nadia, Kurdy Mohamad-Bassam, Stock Market Price Prediction System using Neural Networks and Genetic Algorithm, *Journal of Theoretical and Applied Information Technology.* 97. 15, pp 4175-4187. (2019).