Journal Title: Journal of computational and theoretical nanoscience

Volume: 16
Issue: 8
Month/Year: 1 2019

Pages: 3519-

Article Author: Jiang, Loh Chi,

Article Title: Forecasting of Stock Price Using Autoregressive Integrated Moving Average Model

Imprint:

ISSN: 1546-1955

Lender String:

Notes

Special Instructions:

Note to Scanner: SCAN THIS SHEET!!

Odyssey

Trans. # 3348229

LENDING ARTICLE

****

RAPID

Copy To:

RAPID:UOP (1) - NEW: Holt Memorial Library

Borrowed From: EYM / MIUG

Interlibrary Loan
University of Michigan
Phone: 734-764-0295
Email: ill-lending@umich.edu

Courier Reply

	1st Searched	2nd Searched
☐ NOS ☐ Volume ☐ Call #		
☐ NFC ☐ Vol/Year don't agree ☐ Other		
☐ Tightly Bound		
☐ Missing Pages		
☐ Bound w/o issue		
☐ Non-circ		
☐ Other		

Journal of
Computational and Theoretical Nanoscience
Vol. 16, 3519–3524, 2019

RESEARCH ARTICLE

Forecasting of Stock Price Using Autoregressive Integrated Moving Average Model

Loh Chi Jiang and Preethi Subramanian*

Faculty of Computing, Engineering and Technology, Asia Pacific University of Technology and Innovation, 57000, Malaysia

Finance sector is highly volatile where the stock prices fluctuate rapidly and it is usually challenging to forecast. The unstable conditions and rapid changes can drastically modify the monetary value of an organization or an individual. Hence, the prediction of stock prices continues to remain as one of the sizzling and vital topics in the applications of data mining in the finance sector. This forecasting is significant as it has the potential to reduce the losses that happen mainly due to erroneous intuitions and blind investment. Moreover, the prediction of stock prices endure to increase in complexity with accumulation of more and more historical data. This paper focuses on American Stock Market (New York Stock Exchange and NASDAQ Stock Exchange). Taking into account the complexity of the prediction, this research proposes Autoregressive Integrated Moving Average (ARIMA) model for estimating the value of future stock prices. ARIMA demonstrated better results for prediction as it can handle the time series data very well which is suitable for forecasting the future stock index.

Keywords: ARIMA, Machine Learning, NYSE, NASDAQ, Stock Price.

1. INTRODUCTION

The prediction of stock prices continues to be one of the hot and very important topics in the finance area. Stock market reflects the economy and attracts the attention of millions of investors. Stock market can be considered as a high risk as well as a high yield market. Due to this reason, many investors are focused on the analysis of the stock prices and plan to predict the trend of the stock market [1]. The benefits of predicting future stock prices enable better investment decisions and develops effective strategy about their future endeavors. Moreover, the investors profit out of the venture at less risk [2]. However, due the noisy, non-stationary and dynamic nature of stock prices, it is undeniably challenging to forecast stock market prices accurately [3]. This is because stock prices are a random walk behavior and the impact makes predictions highly difficult [6]. In addition to the dynamic nature of the stock market, the collapse of the Lehman Brothers has been one of the additional problems to cause instability since 2008. After the collapse of Lehman Brothers, the global economy incurred recession and many companies and investors went bankrupt. It was evident that stock indices dropped and downturn of economies was witnessed. The stock indices showed huge fluctuations in the new emerging market one year later after the collapse and this made the stock prices prediction even more difficult [4]. Hence, the prediction of future stock prices is vital in this industry. Due to the advancements in technology, the prediction of the stock market has become more advanced. The strong competition between organizations also makes the forecasting process more important in today's business [5]. Several researchers are working on how to forecast the stock prices by using different prediction models but there are dissimilarities in the reliability of the predictions [7]. In order to forecast the future stock prices, Autoregressive Integrated Moving Average (ARIMA) model is most suitable due to its ability to handle time series transactions. ARIMA models are known to be robust and efficient in financial time series forecasting, especially in short-term prediction [2].

This paper focuses on the American Stock Market which is the world's largest stock market. Hence, this market plays a significant role and the stock exchange of it which is New York Stock Exchange (NYSE) and NASDAQ Stock Exchange must be managed carefully as any small changes will affect the other stock markets in the world. Moreover, the fluctuation of the stock prices is closely related to the nonlinearities, volatility, discontinuity, movement of other stock markets, political and economic factors and individual psychology [8]. In addition to the above, the prediction of stock prices also aids the common people and small investors who pursue trading in stocks for income or investment.

*Author to whom correspondence should be addressed.

 1546-1955/2019/16/3519/006 doi:10.1166/jctn.2019.8317

2. RELATED WORKS

ARIMA model also can be called Box-Jenkins method and it was introduced by Box and Jenkins in 1970. ARIMA model composed of set of activities for diagnosing, estimating and identifying with time series data [2]. MA is a function of white noise to model the current price while AR is the function of previous prices to model current price. ARMA combines the functions of white noise and previous prices to model the current price [9]. According to the Mohamed Ashik and Senthamarai Kannan, Box-Jenkins approach had been used by researchers to forecast the closing Nifty 50 stock market prices from National Stock Exchange (NSE) which is the stock exchange in India. Time series analysis was used to analyze the stock prices. The original Nifty 50 stock prices were found to be non-stationary while it was stationary after the first order differencing. The influence R-Square value was found to be 94% high and the Mean Absolute Percentage Error (MAPE) was very small. Hence, the prediction was more fitting with Nifty 50 stock prices and it showed the slow decreasing fluctuations trend for the upcoming days [10]. The same technique had also been used to forecast the Zenith Bank and Nokia stock prices from Nigeria Stock Exchange (NSE) and New York Stock Exchange (NYSE) respectively by Adebiyi et al. [2]. In order to determine the best ARIMA model, some experiments and criteria had been checked such as Bayesian or Schwarz Information Criterion (BIC), standard error of regression (SER) must be relatively small while the adjusted R-square must be high and the residual of the model must be white noise which means that Q-statistics and correlogram shows that there is not significant pattern left in the autocorrelation functions (ACFs) and partial autocorrelation function (PACFs). From the results, both Nokia and Zenith were showed non-stationary but it became stationary after first differencing. In order to confirm that the stock data became stationary, Augmented Dickey Fuller (ADF) unit root test on "DCLOSE" was used to test both the stocks.

Kuttichira and Vidyapeetham also used the Box-Jenkins method to forecast the stock prices of Abbot India and Cipla [9]. The applications of ARIMA had been used to forecast the gold price in the Multi Commodity Exchange of India Ltd. Durbin-Watson Test was used to detect the presence of autocorrelation of regression analysis [11]. ARIMA model yielded significantly better results in the prediction of stock prices of Square Pharmaceuticals Limited (SPL) for the year 2012 [12]. According to the Devi and Alli, the researchers proposed ARIMA model to forecast the Nifty Midcap-50, Reliance, OFSS, ABB, JSW-STEEL in the National Stock Exchange [13]. This model also predicted fifty-six Indian stocks from various sectors in the National Stock Exchange [14]. From the articles, all the researchers used the Box-Jenkins technique for the ARIMA model. This model need to go through five steps for forecasting the future stock prices which is data preparation, model selection, estimation, diagnostics and forecast. Each stage must be converted to stationary first before moving to the next stage. Hence, every researcher used different kinds of evaluation methods to check the accuracy such as BIC, MAE, RMSE and more. In summary, ARIMA is very useful in prediction with short-term periods but not as effective in the long-term. Nevertheless, external factors such as politics, inflation, and internal management can modify the forecasts in the long run.

3. METHODOLOGY

Cross Industry standard practices for data mining (CRISP-DM) is an idealized sequence of events in which the tasks can be performed in a varied manner as it's an iterative methodology. The following are the main tasks implemented keeping in mind that the deployment stage has been excluded in this paper.

3.1. Data Understanding and Preparation

The daily historical stock prices of Standard and Poor 500 (S&P500) was collected from Yahoo Finance. There are 1260 stock data recorded from 1 January 2013 to 31 December 2017. The stock data consists of price, open, high, low, volume and change variables as shown in Table I. The stock data is split into 80% for training and 20% for testing.

For the data preparation stage, the raw data will be fed into the modelling tool to construct the final dataset. Prescribed order for the data preparation tasks is not required. However, there are some data pre-processing tasks that need to be performed repeatedly [15]. All the data preparation steps will be processed in R Studio.

3.2. Modeling

ARMA model always assumes the time series data is stationary but time series data like stock data is non-stationary and it is very difficult to analyze. In order to make the stock become stationary, ARIMA can does this by differentiating the series to become stationary. The first order differencing process of the time series Xt is defined as $X't = Xt - Xt - 1$. Hence, this differencing process known as Integrated Autoregressive Moving Average (ARIMA) model. This model consists of three parameters which are p, d and q. P is order of autoregressive

Table I. Meta data.

Variable	Description
Date	Date for the stock data
Close	Close price of the day
Open	Open price of the day
High	Highest price of the day
Low	Lowest price of the day
Volume	Volume traded of the day
Adj. close	Adjusted close price of the day

model while d is order of differencing and q is order of moving average model. The historical stock data will be decomposed into an autoregressive process which is AR and it maintains the memory of past events. Next, the Integrated (I) process will be taking part after AR process to make the data become stationary by differencing to make the forecast more easily while the last one is Moving Average (MA) process which is forecast errors and it does not suffer from existence of serial correlation between the error residuals and their own lagged values. The formula od ARIMA model is as per Eq. (1) where Xt is stationary, and θ, ϕ are coefficients, prices are labelled as x at times t, $t-1$, $t-2$ etc. [9, 16]

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p + w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q} \quad (1)$$

3.2.1. Identification Stage

In the identification stage, the time series data will be checked for stationary and compared with the estimated Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) to find a match [10]. The graph of ACF will be used to determine whether the series is stationary. The time series will be considered as stationary if the time series values cuts off quickly or dies down fairly quickly. The differencing will be needed if the data are non-stationary [16].

3.2.2. Estimation Stage

Estimating the parameters for ARIMA is very complicated and the main approaches for fitting ARIMA are nonlinear least squares and maximum likelihood estimation. The maximum likelihood will be obtained by the parameter estimates which act as asymptotically correct for time series. Gaussian distribution is always efficient and consistent with estimators. However, some non-Gaussian distribution will be asymptotically normal with it [10].

3.2.3. Diagnostic Checking Stage

The model is checked if it fit the data correctly by using residuals. Gaussian white noise will be formed by using the residuals if the model is suitable for the data [9]. The residuals must be checked whether has a normal distribution and random. The randomness can be checked by using Ljung-Box (Q) test. The properties of the residual can be checked with the normality by considering the normal propability plot or the p-value from the One-Sample Kolmogorox–Smimov Test and the graph of ACF and PACF. The residual autocorrelation should be small and within [16]. On the other hand, the model selection will be based on criteria like log likelihood, Akaike Information Criteria (AIC)/Bayesian Information Criteria (BIC)/Schwarz-Bayesian Information Criteria (SBC). The verification of the satisfactoriness of the estimated model will be done after the model selection. The verification will be studying the pattern among the residuals and this can be computed by $\hat{e} = Yt - \hat{Y}t$, where the $\hat{Y}$ is the estimated observations at time t. The value of $\hat{e}t$ will range from -3 to 3 which indicates that the residuals are outliers. The ACF verifies if the residual is a white noise. The diagnostic checks reveal the satisfactoriness of the model after the tentative model is fit to the data. The overall adequacy of the model will be examining Q which is called Ljung-Box statistic and its approximate distribution is chi-square. The Ljung-Box (Q) statistic will be compared to the critical values of the chi-square distribution. The fitted model can be used for forecasting when the diagnostic checking found to be effectively and adequate [10].

3.2.4. Forecasting Stage

Forecasting is predicting the stock prices in the future based on the historical values of the selected sample period and variables.

3.2.5. Evaluation

Model is evaluated using the following parameters: Mean Absolute Error (MAE) is the "mean of the sum of absolute deviation of predicted and observed value dividing by the observed value." Mean Absolute Percentage Error (MAPE) is the "mean of the sum of absolute deviation of predicted and observed value dividing by the observed value" then multiplied by 100 [12]. Root Mean Square Error (RMSE) is a "frequently used measure of the differences between values predicted by a model or an estimator and the values actually observed from the thing being modeled or estimated" [1].

3.3. R Programming

R is a language and environment for statistical computing and graphics as developed by John Chambers and colleagues [17]. By using R Programming to implement the results, it will be comparable to other environments such as Eviews and MATLAB as R has the ability to handle huge sets of data with many mathematical functions. Furthermore, the R packages and libraries that have been used in this paper are MASS, tseries, forecast and metrics. The MASS package is used for plotting the graph and tseries package is used for dealing with the time series analysis like ACF and PACF plots. The forecast package is used for predicting the future stock prices and metrics is used for the function of evaluation like MAE, MAPE and RMSE.

4. RESULTS AND DISCUSSION

Initial exploration revealed that the closing price of S&P 500 is random in nature as seen in Figure 1. The close price was standardized by the log function in order to make it more stable as in Figure 2. The ACF plot was

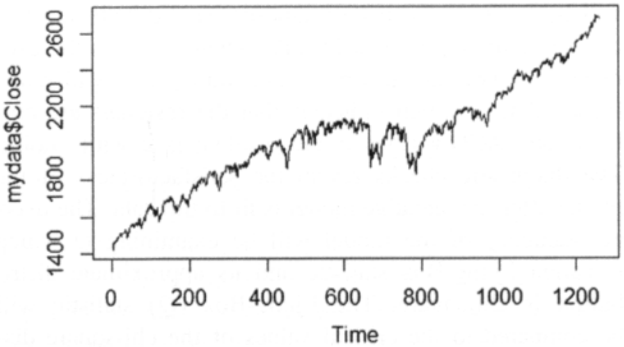

Fig. 1. Close price of the S&P 500 index.

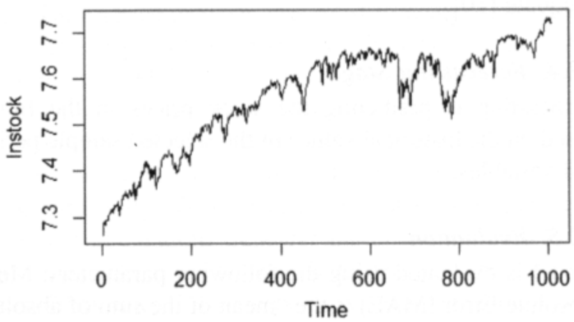

Fig. 2. Close price of S&P 500 after log function.

observed to be decreasing gradually and it remains well above the significance range (dotted blue lines) which indicated that the stock data is non-stationary as in Figure 3. Likewise, the PACF plot also showed that the stock data is non-stationary as in Figure 4. The reason is because the plot did not show significance lag as most of the lines are above zero. In order to confirm that the stock data is non-stationary, the Augmented Dickey-Fuller Test was used to compare the p-value at 5%. The p-value of Augmented Dickey-Fuller Test showed 0.2653 which higher than 5% as shown in Table II and the null hypothesis are accepted which is the stock data are non-stationary.

As the stock data was non-stationary, the first differencing of the stock data was done to make it stationary

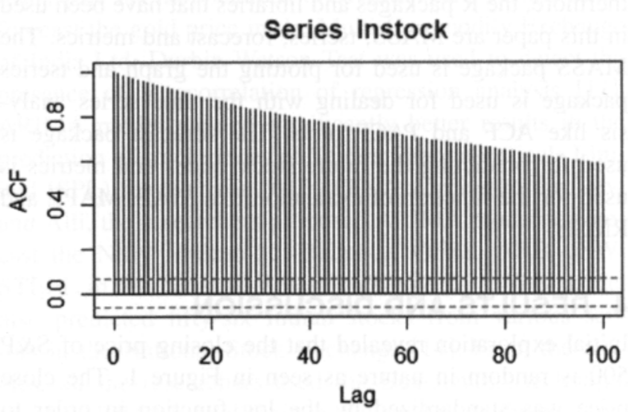

Fig. 3. ACF plot before first differencing.

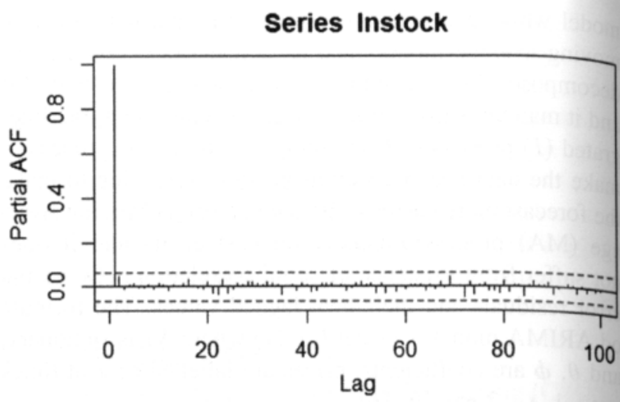

Fig. 4. PACF plot before first differencing.

Table II. Dickey fuller test.

	Dickey fuller test (5%)	Lag order	P-value	Null hypothesis
S&P 500 index	−2.7394	10	0.2653	Non-stationary
S&P 500 index after first differencing	−10.37	10	0.01	Non-stationary (rejected)

by using the "diff(x, 1)" function. The first differencing is using the current value minus the previous value. Augmented Dickey-Fuller Test had been use to check the stationary after first differencing and the p-value showed 0.01 which is smaller than p-value of 5%. This indicated that the null hypothesis was rejected and the stock data became stationary. The ACF and PACF plots for the stock data after first differencing showed the stock data as stationary in Figures 5 and 6. The ACF plots that after differencing showed that all the lines are in the significance range as it did not over the blue dotted lines. The PACF plot after differencing showed that the plot has significant lags as the lines appeared more balance between 0.

The best ARIMA model showed that ARIMA $(1, 1, 0)$ $(0, 0, 1)$ with drift. The best parameters of p, d, q had been estimated which is 1, 1 and 0 or 0, 0 and 1

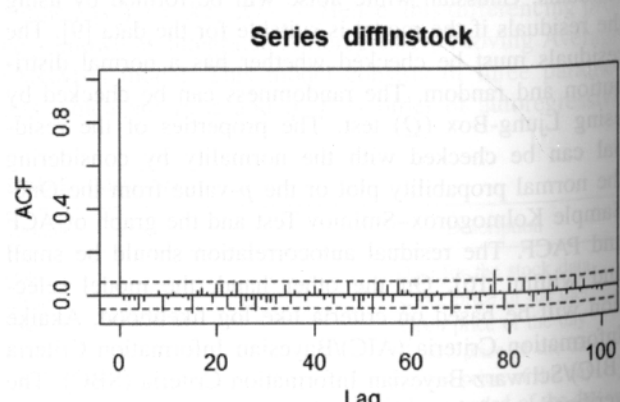

Fig. 5. ACF after first differencing.

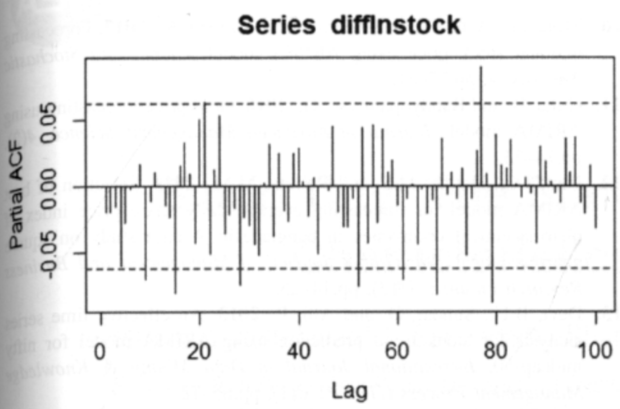

Fig. 6. PACF plot after first differencing.

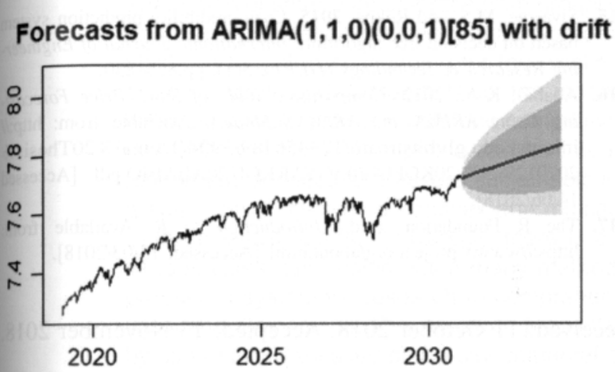

Fig. 7. Forecasted stock of S&P 500 with the ARIMA model.

respectively as shown in Figure 7. This ARIMA (1, 1, 0) (0, 0, 1) with drift is chosen by R programming due to the model performance. The stock data from year 2019 to 2030 had been forecasted by using the testing set and is shown in Figure 8. The stock index showed an increasing trend from 2020 to 2026 but there is some fluctuation from 2027 to 2028. The stock index is forecasted to drop in 2027 and 2028 with a slight peak in early 2028. Commencing 2029, the trend is stable showing that the stock index continues to increase. The prediction of stock data is generally stable with drops at one or two instances.

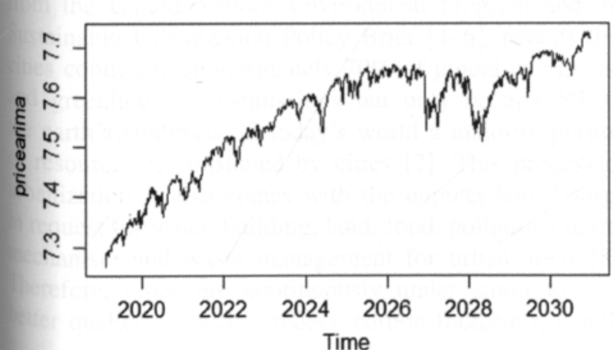

Fig. 8. Forecasted S&P 500 index from 2019 to 2031.

Table III. Evaluation results.

Percentage error	0.33%
MAE	3.01%
MAPE	0.33%
RMSE	2.60%

Table IV. Ljung box Q statistic (5%).

	X-squared	DF	P-value
Lag = 25	33.092	25	0.1288
Lag = 50	56.333	50	0.25
Lag = 75	90.728	75	0.1043

A code snippet for the ARIMA model and forecast are shown as below:

```
pricearima<-ts(lnstock, start = c(2019,10),
frequency = 85)
fitlnstock<-auto.arima(pricearima)
fitlnstock
plot(pricearima, type = "1")
title('Forecasted S&P 500 index from 2019 to 2031')
exp(lnstock)
forecastedvalues_ln = forecast(fitlnstock, h = 253)
forecastedvalues_ln
plot(forecastedvalues_ln).
```

4.1. Evaluation

The forecasted values had been compared with the actual values and the percentage error by using the actual value minus predicted value and divided by actual value. Error values were as in Table III.

Besides that, the residuals had been tested by using Ljung Box test with lag of 25, 50 and 75. The results showed that the p-value is higher than 5% of significance for all lags which indicated that the residuals are no auto-correlation with the lag which is a good model as in Table IV.

5. CONCLUSIONS

In conclusion, ARIMA (1, 1, 0) (0, 0, 1) had been chosen for building the ARIMA model to predict the future stock index of S&P 500. The evaluation result showed that ARIMA model has been fit with the percentage error, MAE, MAPE and RMSE of 0.33%, 3.01%, 0.33% and 2.60% respectively. The model is significantly better as the error is found to be quite small and better than the previous researches. Nevertheless, more adjustments can be done in the future to further reduce the error and increase the prediction accuracy.

References

1. Sureshkumar, K.K. and Elango, N.M., **2011**. An efficient approach to forecast Indian stock market price and their performance analysis. *International Journal of Computer Applications, 34*(5), pp.44–49.

2. Adebiyi, A.A., Adewumi, A.O. and Ayo, C.K., **2014**. Stock Price Prediction Using the ARIMA Model. *UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*, Vol. 6, pp.105–111.

3. Wang, L., Zou, H., Su, J., Li, L. and Chaudhry, S., **2013**. An ARIMA-ANN hybrid model for time series forecasting. *Systems Research and Behavioral Science, 30*(3), pp.244–259.

4. Tsung-Sheng, C., **2011**. *A Comparative Study of Artificial Neural Networks, and Decision Trees for Digital Game Content Stocks Price Prediction*. Elsevier Ltd., Vol. 38, pp.14846–14851.

5. El-Bakry, H.M. and Mastorakis, N., **2010**. Fast forecasting of stock market prices by using new high speed time delay neural networks. *International Journal of Computer and Information Engineering, 4*(2), pp.138–144.

6. Hegazy, O., Soliman, O.S. and Salam, M.A., **2013**. A machine learning model for stock market prediction. *International Journal of Computer Science and Telecommunications, 4*(12), pp.17–23.

7. Heo, J. and Yang, J.Y., **2016**. Stock price prediction based on financial statement using SVM. *International Journal of Hybrid Information Technology, 9*(2), pp.57–66.

8. Nayak, S.C., Misra, B.B. and Behera, H.S., **2015**. Artificial chemical reaction optimization of neural networks for efficient prediction of stock market indices. *Ain Shams Engineering Journal, 8*(3), pp.371–390.

9. Kuttichira, D.P. and Vidyapeetham, A.V., **2017**. Stock price prediction using ARIMA. *International Journal of Engineering Technology Science and Research, 4*(8), pp.676–678.

10. Mohamed Ashik, A. and Senthamarai Kannan, K., **2017**. Forecasting national stock price using ARIMA model. *Global and Stochastic Analysis, 4*, pp.77–81.

11. Guha, B. and Bandyopadhyay, G., **2016**. Gold price forecasting using ARIMA model. *Journal of Advanced Management Science, 4*(2), pp.14–33.

12. Paul, J.C., Hoque, M.S. and Rahma, M.M., **2013**. Selection of best ARIMA model for forecasting average daily share price index of pharmaceutical companies in bangladesh: A case study on square pharmaceutical *Ltd. Global Journal of Management and Business Research Finance, 13*(3), pp.14–25.

13. Devi, B.U., Sundar, D. and Alli, P., **2013**. 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), 3*(1), pp.65–78.

14. Mondal, P., Shit, L. and Goswami, S., **2014**. Study of effectiveness of time series modeling (Arima) in forecasting stock prices. *International Journal of Computer Science, Engineering and Applications (IJCSEA), 4*(2), p.13.

15. Sivasree, M.S. and Rekha, **2015**. Loan credibility prediction system based on decision tree algorithm. *International Journal of Engineering Research & Technology (IJERT), 3*(1), pp.825–830.

16. Ababio, K.A., **2012**. *Comparative Study of Stock Price Forecasting Using ARIMA and ARIMAX Models*. Available from: http://ir.knust.edu.gh/bitstream/123456789/3904/1/Final%20Thesis%202012%2C%20KOFI%20AGYARKO%20ABABIO.pdf [Accessed 14/06/2018].

17. The R Foundation. (n.d.) *Introduction to R*. Available from: https://www.r-project.org/about.html [Accessed: 13/03/2018].

Received: 11 October 2018. Accepted: 13 November 2018.