Dean&Francis
ISSN 2959-6157

Tesla Stock Price Forecast Based on ARIMA Model and Machine Learning Techniques

Haowen Shi

School of Mathematics, Nanjing University, Nanjing, 210093, China

231840024@smail.nju.edu.cn

Abstract:
One of the most significant aspects of the world economy is the stock market, and the forecasting of stock prices is gaining more and more significant attraction. Under the current international situation, Tesla's stock price was selected as the research object, and Tesla stock's open price from July 15, 2020, to July 15, 2024, was chosen as the data set. Auto-Regressive Integrated Moving Average (ARIMA) model and machine learning techniques were taken to forecast the stock price in this paper. Both methods give essentially correct predictions, but there are also some differences between those two methods. The ARIMA model provides statistical trends within a time range, while machine learning techniques provide a stock price prediction that is accurate to the day but has weaker interpretability. By combining the two methods people can get a more accurate forecast of the stock price and then make more informed decisions in the stock market.

Keywords: ARIMA model, machine learning techniques, Random Forest, stock price forecasting

1. Introduction

An increasing number of nations and areas are pursuing their aims of "carbon neutrality" in light of the mounting pressure on the planet's climate and environment as well as the unsustainable nature of conventional energy [1]. In such circumstances, the global popularity of electric vehicles is growing. Founded in 2003, Tesla Inc. is a transnational American firm focused on clean energy and automotive technology. It has become one of the world's most valuable companies in terms of market capitalization. As a prominent company in the electric vehicle sector, Tesla holds a significant market share, with its market size continuing to expand [2]. It has attracted many investors and has made many people predict its stock price. Since the stock market is constantly changing, it is of great significance for investors to predict stock price changes correctly [3]. Although the stock market cannot always be accurately predicted, attempts to predict stock price changes still have significant advantages in many aspects. Therefore, people continue to explore various methods to analyze and predict stock price trends in order to improve the success rate of investment.

There are a great number of time series models be-

ISSN 2959-6157

ing used to forecast the trend of stocks. Lots of emerging technologies are also applied in such fields. As one of the hottest technologies in the world, machine learning techniques have contributed a lot to stock forecasting. According to Tsai's analysis of stock price forecasting using Artificial Neural Networks (ANN) and decision trees (DT), The combined model has greater accuracy than the individual ANN and DT models in the electronic industry [4]. According to Shahi's research, combining financial news feelings with stock attributes as the input can significantly improve the performance of LSTM and GRU models for stock price forecasting [5]. Sen offers a detailed strategy for predicting stock prices by fusing machine learning and statistical techniques with some ideas that have been developed in technical analysis literature. For both of the equities they studied, they obtained comprehensive results regarding the performance of various forecasting systems [6]. A popular forecasting method in addition to these is the ARIMA model. It has been applied in several research to analyze time series data and forecast stock price patterns, such as those during the COVID-19 pandemic, as well as for modeling long-term behaviors and various types of price forecasting [7]. In Chen Yang's research, it was also used to provide a short-term prediction of Tesla stock price and a satisfactory result has been achieved [8]. There are still many questions like whether the model can give an accurate prediction or which model can best fit the stock trend at present.

In this paper, the latest stock market data will be collected and processed with ARIMA model and Random Forest model. This paper aims to provide valuable insights for investors and researchers in selecting models.

2. Methodology

2.1 Data Collection

This study used data from Kaggle.com to compile daily stock price data for Tesla over four years, from July 19, 2020, to July 15, 2024. The opening price is chosen in this study to reflect the index's anticipated price.

2.2 ARIMA Model

2.2.1 Explanation of ARIMA components: p, d, q

A statistical model called ARIMA uses time series data to predict future trends or improve comprehension of the data set. It is a continuation of the ARMA model. ARIMA model has three components, where p represents the order of the AR model, d represents the degree of differencing, and q represents the order of the MA model. All parameters in this formula are non-negative integers [9, 10]. Any time series can exhibit patterns free of random white noise when ARIMA models are used [11].

2.2.2 Data preprocessing

Firstly, the opening price data collected was used to create the time series. Secondly, to find out whether the sequence is stable, doing the stationarity test is essential. Since the Augmented Dickey-Fuller (ADF) Test is one of the most often used statistical tests to determine if a time series is stational, the author decided to employ it. The results are shown in Table 1. The sequence appears to lack the characteristics of a stable sequence, as the p-value is greater than 0.05 and shows a distinct temporal trend. Determining the amount of different orders necessary to transform the initial non-stationary time series into a stationary one is also crucial. The results in Table 2 indicate that the time series after the first-order difference is stationary because the p-value is less than 0.05, as determined by the ADF Tests. The Ljung-Box test is also needed to find out whether the values in the time series are highly random and have no regularity. It is found that the p-value is less than 0.05 so there is no white noise in the time series, and it is suitable to use models to fit the trend.

Table 1. ADF Test of the opening price

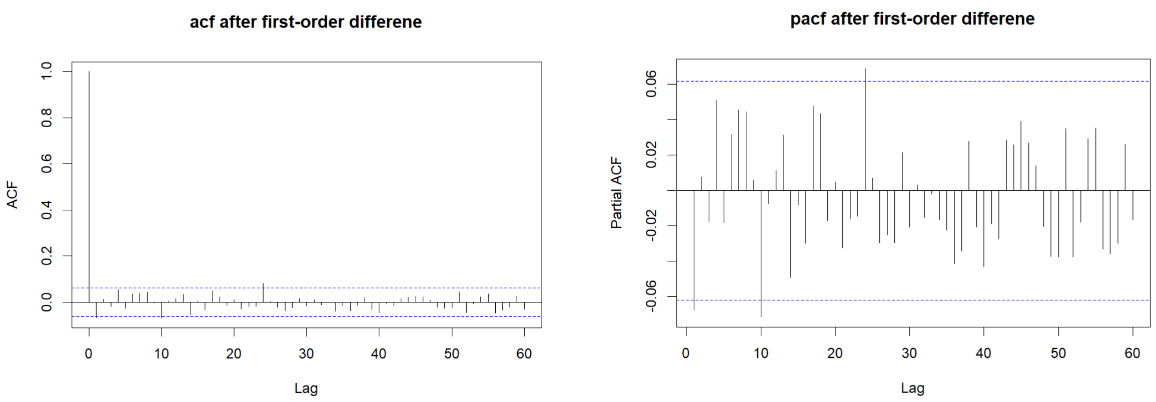
Dickey-Fuller	-3.0034
Lag order	10
P-value	0.1535

Table 2. ADF Test of the opening price after first-order difference

Dickey-Fuller	-9.2892
Lag order	10
P-value	0.01

2.2.3 Steps to identify the best ARIMA parameters using ACF/PACF plots.

According to the sequence stabilization process, d=1. To determine the values of p and q, the auto-correlation coefficient (ACF) and partial auto-correlation coefficient (PACF) of the first-order difference sequence are plotted using the functions acf and pacf. According to Fig 1(A) and Fig 1(B), the ARIMA model (0,1,1) is initially selected. Then Q-Q Plot and the fit of the line are drawn to judge whether the residuals follow a normal distribution. A test for residual white noise is conducted on the model to assess if the residuals exhibit any correlation. The findings indicate that the residuals are normally distributed and uncorrelated. So, it is reasonable to apply the model in further forecasting.



(A) (B)

Fig. 1 After first-order difference, (A) acf; (B) pacf. (Photo/Picture credit: Original)

2.3 Machine Learning Techniques

2.3.1 Selection of algorithms

Machine learning technology has important applications in various industries in recent years. It includes many algorithms like Linear Regression, Decision Trees, Random Forest, LSTM, and so on. These algorithms have all made great contributions to forecasting. Due to space and time constraints, the Random Forest model is used to detect accuracy in this paper.

2.3.2 Explanation of random forest

Random Forest is an emerging learning technique that solves issues like regression and classification by constructing a great number of decision trees during the training phase [12, 13]. Random Forests are renowned for their performance across a variety of applications, simplicity, and ease of usage. They require relatively few hyper-parameters and can handle both numerical and categorical data, making them a well-liked option for a variety of real-world machine-learning issues.

2.3.3 Train the model on the training set

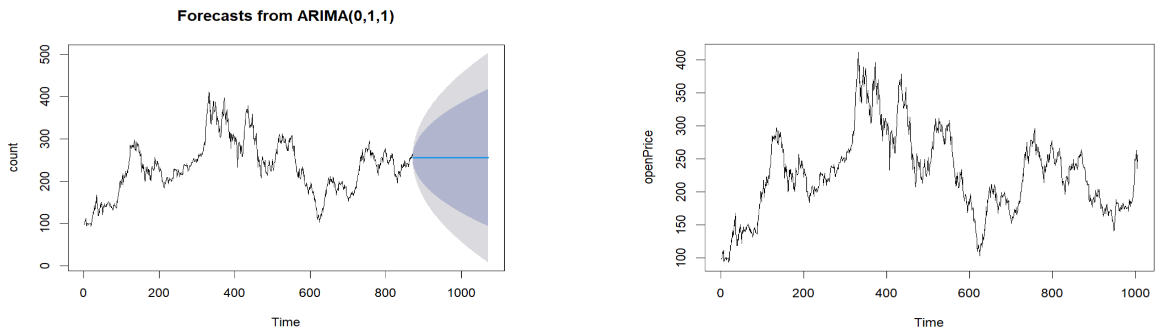
The data is split into training and test sets, with lag prices serving as feature variables for random forests. The training set's data is used to generate and train the random forest model. The random seeds are set to obtain reproducible results.

3. Results and Discussion

3.1 Implementation of the ARIMA Model on the Training Set

The training set consists of data before January 1, 2024. By using the ARIMA model (0,1,1) and the data in the training set, the forecast result is in Fig 2(A) compared with actual prices in Fig 2(B). The ordinate denotes the change in the opening price, while the horizontal coordinate displays changes in the number of days. The anticipated value is shown by the blue line, while the predicted value's confidence interval is shown by the shaded area. By calculation, the Root Mean Squared Error (RMSE) value is 412.

ISSN 2959-6157



(A) (B)

Fig. 2 Tesla stock price, (A) forecasts from ARIMA(0,1,1) (B) actual prices (Photo/Picture credit: Original)

3.2 Results from the Random Forest Model

By using the trained model, the result is displayed in Fig 3. The horizontal and vertical coordinates of each point represent the actual and predicted price of the day respectively. The shade of the dot represents the predicted value. The blue line shows the moving average of the stock price and the shaded part represents the confidence interval for the predicted moving average. By calculation, the Root Mean Squared Error (RMSE) value is 110.

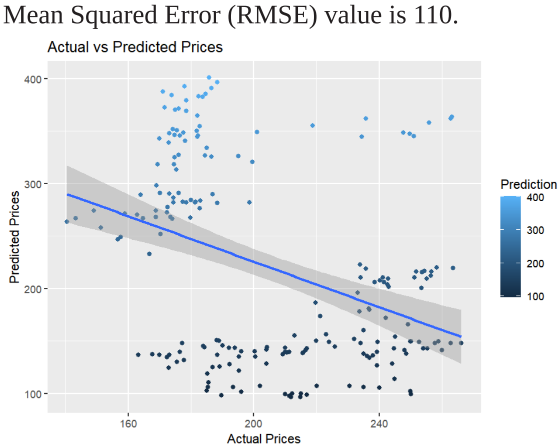


Fig 3. Predicted price from random forest model and actual price from the testing set (Photo/Picture credit: Original)

3.3 Comparison of the Results

From the forecasts shown above, it can be seen that both methods have their advantages and shortages. ARIMA transforms time series prediction problems into regression problems, using the past values of data, past prediction errors, and differences to predict future values. Random forest constructs multiple decision trees, runs input data on these trees, and finally obtains prediction results through voting or averaging methods. The ARIMA model can provide a relatively easy-to-explain statistical prediction with lower computational complexity. By calculating and comparing the predicted RMSE values given by these two models, we can conclude that the Random Forest model can provide more accurate predictions. However, according to the figure, it has weaker interpretability than a black box model. The combination of the two methods allows people to anticipate possible daily stock price changes based on the understanding of the overall situation.

3.4 Discussion

From the results shown above, both models have limitations in their predictions. Due to the complex nonlinear trend of the data, there is a certain gap between the prediction results of the ARIMA model and the prediction. When using the Random Forest model to predict, the result is not perfect because of insufficient feature processing and too little feature quantity. A lot of work has been done to improve these models. The Data Envelopment Analysis model is employed in Narimani's work to determine the optimal lags of the AR and MA processes to enhance the prediction from the conventional ARIMA model [14]. The result of the new method has great advantages over that of the traditional ARIMA model. For the Random Forest model, by using the Partial Dependence Plot, the influence of features on model output can be seen intuitively, to improve the interpretability of the model. Various other features will be analyzed and processed to train the model in follow-up studies to improve accuracy.

4. Conclusion

In this paper, the traditional forecasting model ARIMA and emerging machine learning techniques are both used to predict the Tesla stock price trend. Judging from the results presented by these two methods, both models can fit into the stock price trend to some extent. From the research, it can be seen that the ARIMA model can give statistical trends within a time range, while the Random Forest model can give a more accurate day-to-day stock price change. In general, the ARIMA model is suitable for predicting time series data that is relatively stable in the short term and without significant external influences, while the Random Forest model is more suitable for datasets with complex interactions and nonlinear relationships between variables. Mutual testing of data predicted by these two models generally makes stock price predictions applicable to various situations. Therefore, combining the ARIMA model with machine learning techniques such as the Random Forest model can help investors have a greater possibility to benefit more in the stock market. However, the research also has many limits. Only Tesla's open price data is used in the model prediction process so the prediction may not fully reflect the market situation. While using machine learning techniques to make predictions, due to the time limit and the complexity of conducting tests, the Random Forest model was used as a representative of machine learning techniques, so machine learning techniques cannot be able to be taken full advantage of. In future research, more stock market data will be collected and analyzed while other machine learning techniques will be attempted to take advantage of while making predictions.

References

[1] Chen F. Analysis of key factors for tesla's success. 2022 2nd International Conference on Enterprise Management and Economic Development (ICEMED 2022). Atlantis Press, 2022.

[2] Moritz T. R., Redlich P., Krenz S., Buxbaum-Conradi J. P., Wulfsberg J. P. Tesla Motors, Inc.: Pioneer towards a new strategic approach in the automobile industry along the open source movement?. 2015 Portland International Conference on Management of Engineering and Technology (PICMET). Portland, OR, USA, 2015: 85-92.

[3] Zhou H. Research on stock price prediction based on time series analysis. 2024 Guangdong-Hong Kong-Macao Greater Bay Area International Conference on Digital Economy and Artificial Intelligence (DEAI2024). Hong Kong, China: ACM, 2024: 7 pages.

[4] Tsai C.-F., Wang S.-P. Stock price forecasting by hybrid machine learning techniques. Proceedings of the International MultiConference of Engineers and Computer Scientists 2009. Hong Kong, 2009, I, 18-20.

[5] Shahi T.B., Shrestha A., Neupane A., Guo W. Stock price forecasting with deep learning: a comparative study. Mathematics, 2020, 8(9): 1441.

[6] Sen J., Chaudhuri T. A robust predictive model for stock price forecasting. Proceedings of the 5th International Conference on Business Analytics and Intelligence (ICBAI 2017). Indian Institute of Management, Bangalore, INDIA, 2017: 11-13.

[7] Benvenuto D., Giovanetti M., Vassallo L., Angeletti S., Ciccozzi M. Application of the ARIMA model on the COVID-2019 epidemic dataset. Data Brief, 2020, 29: Article 105340.

[8] Yang C. Tesla stock price timeseries analysis and forecasting. Dang C.T., Cifuentes-Faura J., Li X. (eds). Proceedings of the 2nd International Conference on Business and Policy Studies. Singapore: Springer, 2023.

[9] Notation for ARIMA Models. Time series forecasting system. SAS Institute, 2015.

[10] Hyndman R. J., Athanasopoulos G. 8.9 Seasonal ARIMA models. 2015.

[11] Filder T.N., Muraya M.M., Mutwiri R.M. Application of seasonal autoregressive moving average models to analysis and forecasting of time series monthly rainfall patterns in embu county, kenya. Asian Journal of Probability and Statistics, 2019: 1-15.

[12] Ho T.K. Random decision forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition. Montreal, QC, 1995: 278-282.

[13] Ho T.K. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(8): 832-844.

[14] Narimani R., Narimani A. A new hybrid model for improvement of ARIMA by DEA. Decision Science Letters, 2012, 1(2): 59-68.