Hikvision Price Prediction based on ARIMA, OLS Multiple Regression and Random Forest

Jinyang Xia*

Shanghai Datong High School, Shanghai, China

*Corresponding author: yili@ldy.edu.rs

Abstract. As a matter of fact, stock price prediction is quite crucial for investors to gain extra return in financial market. Thanks to the rapid development of the machine learning techniques. With this in mind, this paper takes Hikvision as sample data to study the stock forecast situation based on ARIMA time series model, OLS multiple regression, random forest machine learning these three models, Hikvision as a sample is also fancy the high growth of the company, and the stock forecast meets the current needs, which is of great significance, and finally the random forest machine learning and ARIMA time series model are obtained by comparison finally predict better effects relatively, through the computer model to predict, compared with the manual, has a higher running speed, in order to achieve better prediction results, in line with the current needs, is of great significance. Overall, these results shed light on guiding further exploration of price prediction.

Keywords: ARIMA; OLS; prediction; random forest; Hikvision.

1. Introduction

Hangzhou Hikvision Digital Technology Co., Ltd. (abbreviation: Hikvision, stock code: SZ002415) Founded in 2001, with scientific and technological innovation as the core, in A&S Safety Automation publishes "Global Security 50 Strong" list, it has ranked first in the world for several consecutive years [1]. The main business of the enterprise is security business, visual management business, and big data server [2]. Since its listing, Hikvision has undergone continuous development, changing its scale from small to large, operating efficiency from low to high, and competitiveness from weak to strong [3]. Over the years, Hikvision has maintained a huge scale of R&D investment, and has greater development potential in the future, although individual indicators have been affected by the external environment in the past two years to slow down the growth rate, but the overall development is good, so overall it has good growth [1]. With the increasingly fierce competition in the security industry, technological changes and changes in market supply, Hikvision's stock price fluctuations are more worthy of study [4-8]. At the same time, because the algorithm can conduct in-depth analysis of the prediction problem, the algorithm with the help of the computer is more efficient than the manual prediction and can achieve better prediction effect [4]. Thanks to long-term sedimentation, efficient and diverse stock prediction methods have emerged one after another [6]. Therefore, it is important to meet the needs of the present as a stock forecast.

By sorting out and analyzing the literature on stock forecasting, it can be found that it is mainly divided into time series models and machine learning prediction models, and the underlying logic of the time series model is to describe the past, analyze laws, and predict the future, such as: exponential smoothing model, seasonal decomposition, ARIMA model, etc. Back in 1982, the ARCH model (Autoregressive conditional heteroskedasticity model) was first proposed, laying the foundation for subsequent progress, and later the GARCH model (Generalized ARCH) was proposed [6]. Among them, when using the ARIMA model for prediction, it is necessary to pay attention to the fluctuation degree of the real value. If the fluctuation is too large, the output effect will not be as good as LSTM and RNN, and if the data is mostly high-frequency, the ARIMA-GARCH model has relatively better prediction performance [5]. In the current era of information explosion, the benefits of machine learning models are increasingly evident, which is of great practical significance for stock prediction [9]. The nonlinear relationship of the stock market can be well captured by the machine learning algorithm model, so as to realize the prediction task. The combination of machine learning and

traditional economics can promote the development of stock market research [10], among which machine learning prediction models mainly include decision trees, support vector machine regression (SVR), etc., SVM, ARIMA methods have been applied to financial stock price prediction research in the past few decades, and have achieved certain prediction effects [7].

This paper hopes to use the three models of ARIMA time series model, OLS multiple regression, and random forest machine learning to predict stock prices, and use samples for training and testing, and finally evaluate with evaluation models to compare the results, so as to compare the similarities and differences and good and bad of the three.

2. Data and Method

The sample data used in this study comes from the stock prices of Choice financial terminal Hikvision (SZ002415), which selects its daily data from May 28, 2010 to August 18, 2023, and the main data is the closing price, aiming to predict future stock data. The models selected in this paper are ARIMA time series model, OLS multiple regression, and random forest machine learning.

The ARIMA model consists of an autoregressive model (AR model), a differential integration model (I model), and a moving average model (MA model). where p represents the autoregressive order, d represents the difference order, and q represents the moving average order. The essence is to convert a nonstationary sequence into a stationary sequence, and the formula is expressed as:

$$X_t = a_1 X_{t-1} + a_2 X_{t-2} + \ldots + a_p X_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \ldots + \beta_q \varepsilon_{t-q} \quad (1)$$

The full name of OLS is Ordinarily Least Square, also known as the Ordinary Least Squares method.

As a common mathematical optimization method, the core idea of least squares is to estimate by minimizing the sum of squares of the residuals. It is to use scatter to fit an optimal line, and when the sum of squares of the distance between scatter and the fitted line is the smallest, such an optimal line is obtained, that is, the sum of squared residuals is the smallest. The formula is expressed as:

$$\beta = (\mathrm{X}^{\mathrm{T}} \cdot \mathrm{X})^{-1} \cdot \mathrm{X}^{\mathrm{T}} \cdot \mathrm{Y} \quad (2)$$

Random forest through the idea of ensemble learning is an algorithm that integrates multiple trees, its basic unit is the decision tree, using the ensemble idea, when the data sample is predicted, each tree in the random forest predicts the result of the sample, and then selects the result from these prediction results by voting.

The R-squared value is a measure of how well the predictor variable explains the variation in the outcome variable. It ranges from 0 to 1, with 1 indicating a perfect fit. It is calculated as the square of the correlation coefficient between the observed actual values and the predicted values generated by the model. A higher R-squared value indicates that the model has a better fit to the data. The RMSE is a measure of the average squared difference between the predicted values and the actual observations. It is calculated as the square root of the mean of the squared differences between the observed actual values and the predicted values. Thus, RMSE is equal to the square root of the mean squared error. A lower RMSE indicates that the model has better predictive accuracy.

Fig. 1 Visualized data from August 18, 2016 to August 18, 2023

3. Results and Discussion

First, the data of the closing price of Hikvision daily is taken out for resampling, and then the data from August 18, 2016 to August 18, 2023 is designated as the training data, and then the training data is visualized. As can be seen from Fig. 1, the amplitude of the original data fluctuation is large, the sequence is not stable, and it needs to be converted into a stationary sequence differentially, and the data is visualized here by first-order differential and second-order difference. From the above figure, it can be found that the first-order difference already has a stable trend (Fig. 2), and to the second-order sum, the amplitude of fluctuation is larger, so the parameter d is directly determined here as 1.

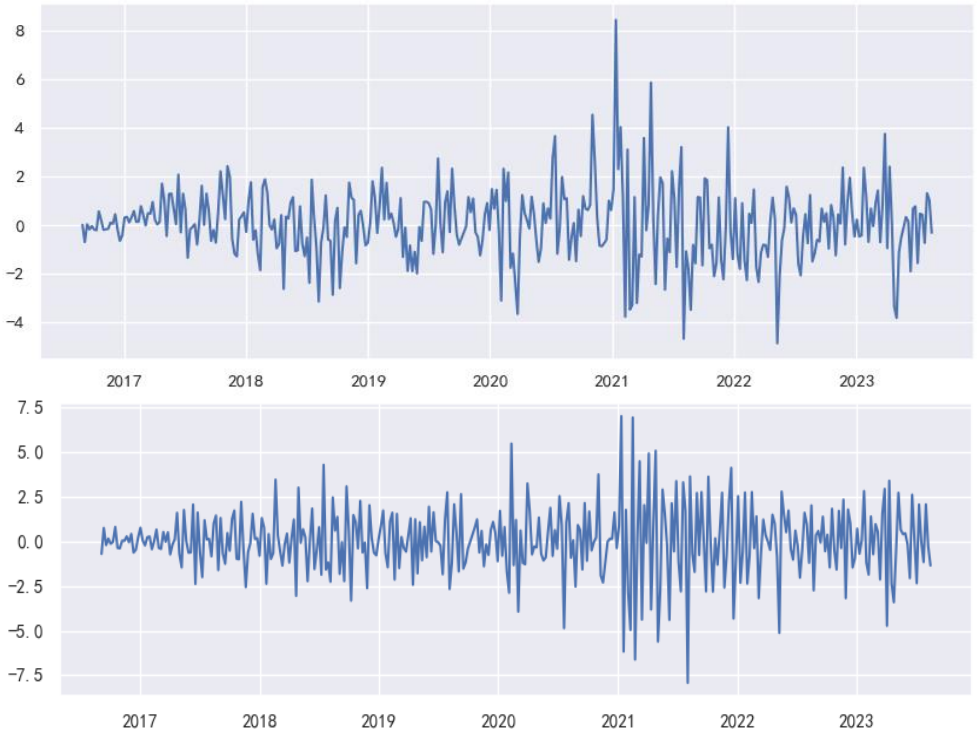

Fig. 2 The first-order differential (upper) and second-order difference (lower).

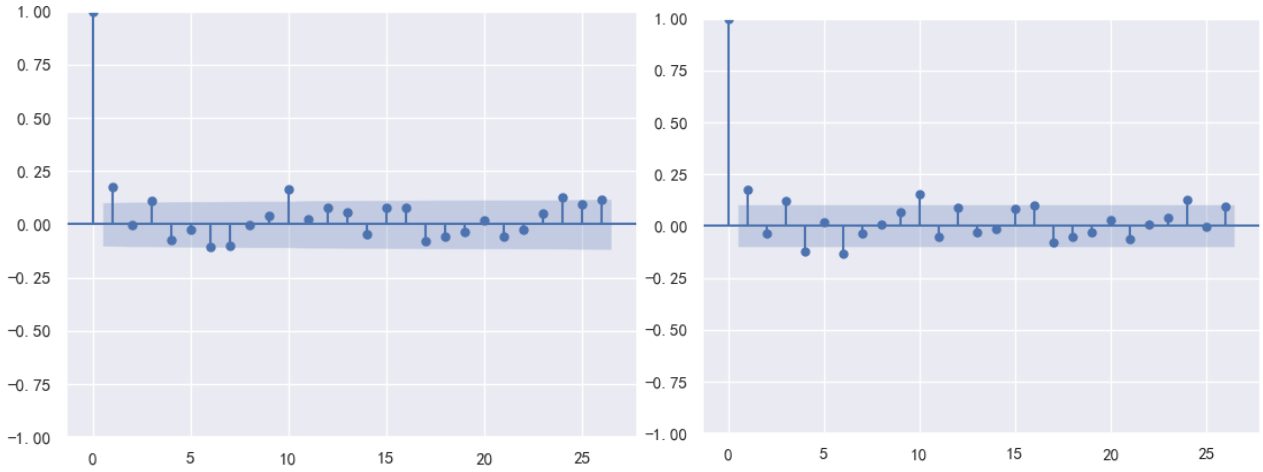

Fig. 3 The ACF (left) and PACF (right).

Then this study makes ACF, PACF plots to determine the parameters q and p. By looking at the Fig. 3, it can be determined that it is best for the parameters p and q to be both 1. Then it's time to train the model and make predictions, fit the ARIMA model, and visualize the predicted and true values as given in Fig. 4.

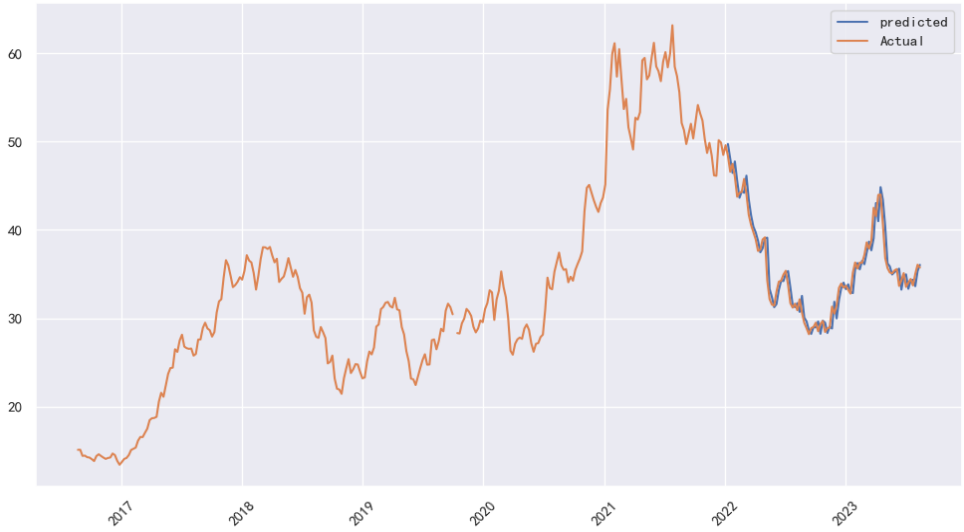

Fig. 4 The comparison of prediction and actuality

As for OLS multiple regression, first, one needs to normalize the data, first consider the most general form (here the form y = kx + b), establish a regression model, and then extract the independent and dependent variables respectively, train the input data, substitute it into the general form, and find the values of k and b of the function, where k=0.00024622396829375773 and b=-0.0904971963816601. Then, this study will make predictions based on this linear equation, denormalize the results of the predicted output, handle the errors, and finally visualize the results in Fig. 5 with R square 0.79.

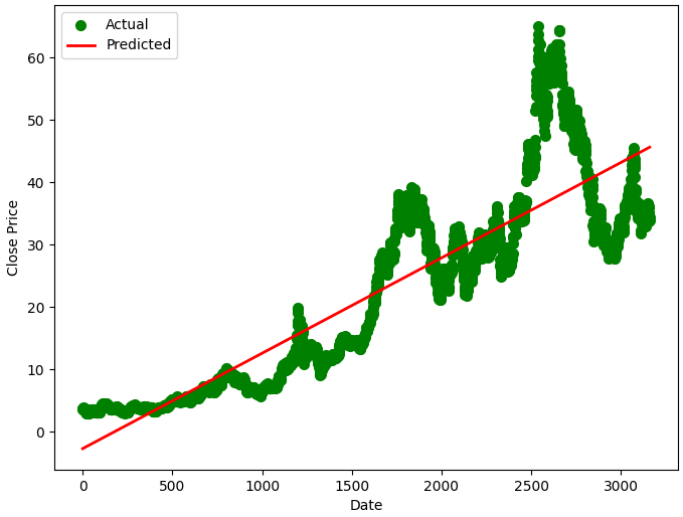

Fig. 5 the comparison of prediction and actuality.

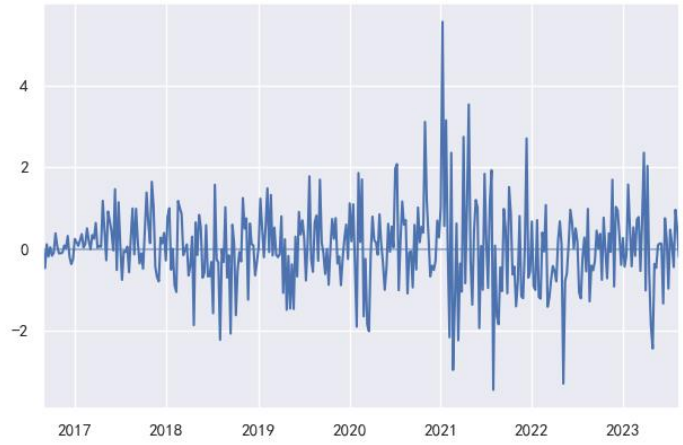

Fig. 6 The residual analysis plot, which can find the model residuals.

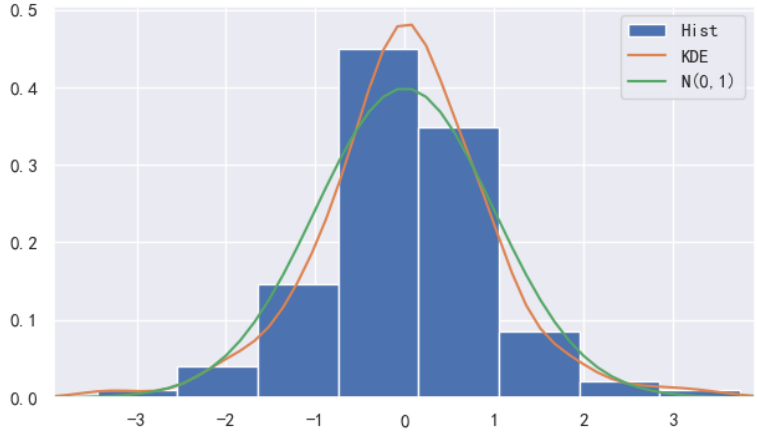

Fig. 7 The histogram and the orthostatic distribution plot.

For random forest machine learning, one imports the dataset and convert it into a time series, then create a target column, split the dataset into features (X) and target (Y), divide the data into training and test sets, secondly, create and train a random forest regression model, make predictions on the training set, fit the random forest model, and visualize the predicted value and the real value.

As for ARIMA Time Series Model, it visualizes the model evaluation results (Fig. 6) and it can be found that the model is approximately the orthostatic distribution (seen from Fig. 7). Seen from Fig. 8, one can find that except for a few poles at both ends, most of the data can be fitted with a straight line. According to the assessment, for ARIMA, RMSE is 1.3508737952195569, R-squared is 0.9341122283637149; for OLS Multiple Regression, RMSE is 20.993080118549802, R2 is

0.7338290567639212; regarding to random forest machine learning, the RMSE is 2.9702812933440947; R-squared is 0.964925966617636. According to the comparison of the results of the above three models, and according to the higher the R-squared introduced in 3.2, that the better the model, and the lower the RMSE, the better the model. The Random forest machine learning model and ARIMA time series model both have better predictions relatively.

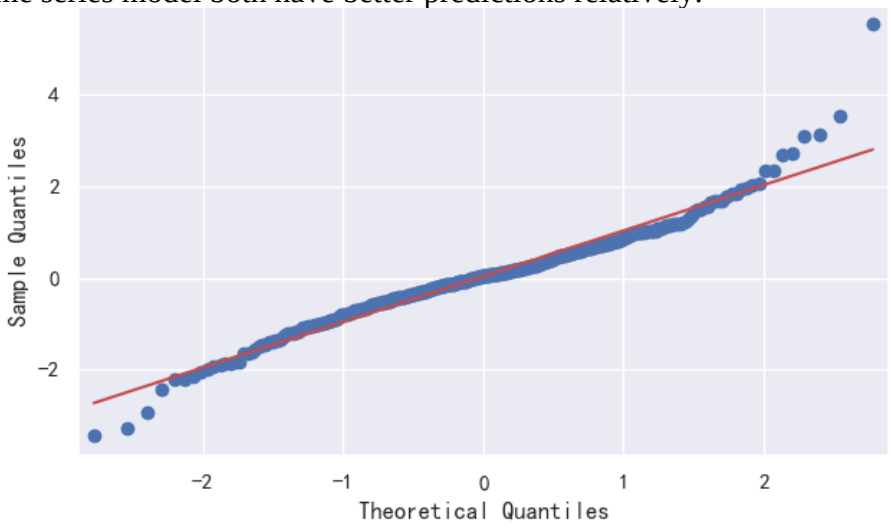

Fig. 8 The QQ chart.

4. Limitations and Prospects

The forecasts for the stocks studied in this paper are based on only three traditional computer models: the ARIMA time series model, OLS multiple regression, and random Forest machine learning, and the residual analysis needs to be improved, and it does not consider the impact of financial crises, major policy changes, etc. on the forecast, does not study the stock forecasts of other companies in the same industry in the same period, and hopes to compare the differences of other companies in the same period, and can also combine the combination of stock price forecasting and investment strategies, and further improve the use of models for stock forecasts to improve their practical value.

5. Conclusion

To sum up, Hikvision is used as sample data to study the stock prediction situation under three models, i.e., ARIMA time series model, OLS multiple regression, and random forest machine learning. According to the analysis, the results show that random forest machine learning prediction and ARIMA time series model prediction both have better effects relatively, can also further compare the differences of other companies in the same period, can also be combined with the combination of stock price prediction and investment strategy, and further improve the use of stock forecasting models, to achieve long-term stock forecasting, according to this research can better compare the advantages and disadvantages of the three models, and the importance of the use of stock forecasting methods and its practical value can be displayed.

References

[1] Dai W. A case study of Hikvision's growth. Shenyang University of Technology, 2021.

[2] Cui D. Research on the implementation effect of Hikvision "business co-investment + restricted stock" equity incentive. Hebei University of Economics and Business, 2020.

[3] Zou Z. Analysis of the impact of executive incentive on enterprise R&D innovation. Jiangxi University of Finance and Economics, 2019.

[4] Yang C. Research on financial stock prediction based on machine learning method. Hunan University, 2020.

[5] Xiong Z, Chi W. Application of ARIMA-GARCH-M model in short-term stock forecasting. Journal of Shaanxi University of Technology (Natural Science Edition), 2022, 38(04): 69-74.

[6] Chen X. Research and implementation of stock analysis and prediction technology. University of Electronic Science and Technology of China, 2022.

[7] Wang Y. Research on in-depth feature extraction and prediction of multivariate time series stock information. Shenyang University of Science and Technology, 2020.

[8] Zeng W, Chang L, Tang Y, et al. financial performance evaluation of Hikvision's asset-light operation model transformation. Finance and Finance, 2021, 3: 29-34.

[9] Xu H, Xu B, Xu K. Review of the application of machine learning in stock forecasting. Computer Engineering and Applications, 2020, 56(12): 19-24.

[10] Yang Y. Research on liquor stock prediction based on machine learning algorithm model. Shandong University, 2021.