Stock Trend Forecasting Using the ARIMA Model

Can Dong

School of Mathematics, University of Birmingham, Birmingham, United Kingdom

cxd077@student.bham.ac.uk

Abstract. Stocks have always been a very important tool in the investment market. Nowadays, the stock market attracts a large number of investors as more and more people are exposed to investing their money. One of the most attractive features of equities for investors is the high returns, however, the high risks are also affecting investors' confidence. Therefore, predicting the long-term performance of the stock market can lower the risk and give investors more ideas on how to invest and help them understand the future trend of their preferred stocks, thus reducing their risk rate. In this paper, the author proposed a stock predicting method based on the closing prices of Ford Motor Company over the past 50 years. In the selection of the model, the ADF test was utilized and by analyzing the dataset, the author demonstrated that the stock closing price is a non-stationary series and therefore the ARIMA model is selected. After differencing the series, a stationary series was obtained, and the best parameter was chosen based on Auto-ARIMA analyses. Furthermore, the author obtained the long-term trend graph based on the model output and analyzed the accuracy of the prediction results by RMSE and MAPE tests. In addition, by bringing data from other stock markets into the experiment, it is evident that the ARIMA model can be effective when predicting long-term trend of stocks and is able to be used as a method to forecast stock prices.

Keywords: Stock prediction, Time Series, ARIMA model.

1. Introduction

The stock market first appeared in the 16th century as a means of spreading risk in the field of voyaging trade, and later, as society progressed, they were integrated into the financial as well as the industrial sector, since during that period, those industrial sectors all need a large amount of investment. As the economic growth and productivity developed, price increases and inflation followed, and people began to think about how to save their money as a way to combat inflation and reduce losses in order to purchase their needs and wants, thus bringing investment finance into more people's lives. Stocks, as a high-return investment vehicle, became the preferred form of finance for investors, who were attracted by the profits they could generate, but on the other hand, stocks are also relatively risky [1]. To be more specific, a fluctuation in stocks could have a huge impact on investors and even on the national economy. Especially due to COVID-19, a large number of businesses are affected, and their stock prices continue decreasing. As a result, attempts have been made to study the patterns of stock price fluctuations, as stock prices are not randomly generated values. By analyzing historical data, it is possible to predict the general trend of future fluctuations.

In recent years, stock forecasting has become a popular topic, attracting a large number of researchers from different fields to study it. From the statistical point of view, for example, there are a lot of forecasting methods such as Markov chains [2, 3], and logistic regression [4]. And also, LSTM [5], belongs to the neural network direction. In the last few years, many forecasting methods have been proposed, whereby comparing different works, the ANNs model [6] is one of the most popular techniques for artificial intelligence prediction. However, the accuracy of stock forecasting is still relatively unstable, and the current forecasting accuracy still needs to be improved, which is one of the main problems of many researchers, as accuracy has a direct impact on forecasting results if they are less accurate, for example, linear regression [7] is only effective in predicting the current day's closing price but does not perform as well in predicting future closing prices. This means such a prediction cannot offer investors a helpful result. This makes investors not able to have access to valid stock price information, therefore, a high accuracy forecasting tool can provide investors with more efficient help to cover risk and reduce losses.

By comparing several methods, it is found that time series analysis, as a method in the field of statistics, is widely used for forecasting and performs very well, especially in the direction of economic and financial forecasting. To forecast the likely future trend of stock price variations, historical stock price analysis is used. There are many useful models for time series forecasting, however, stock prices are non-stationary which means other forecasting models in the time series are not available. ARIMA models are different from other models such as ARMA, AR and MA [8]. Since ARIMA model transfer non-stationary series into stationary series by using differencing before forecasting, which makes the accuracy is higher than processing non-stationary series directly. Therefore, in this paper, the stock closing prices of Ford Motor Company for the past fifty years are mainly used as the experimental data set. Through analysis, it was found that the historical closing prices were non-stationary series, the existing non-stationary series was converted into a stationary series using differencing, and the ARIMA model was then used for forecasting. Stock data from other companies were also used at the conclusion of the trial to assess the accuracy of the prediction.

Through the news of Ford Motor Company, it was found that the share price of Ford had a very significant and sustained downward trend between 2014 and 2020 and then into a sudden rebound, this fluctuation could have been risky to both investors and business managers. Therefore, if the ARIMA model can be used to effectively predict and analyses the floating trend in this time period, investors can stop their losses instantly and thus reduce unnecessary risks, and this prediction can be a piece of trustworthy and effective evidence.

2. Methodology

The creator of Autoregressive Integrated Moving Average (ARIMA) model is Box and Jenkins [9]. The model is one of the most widely used forecasting models, and to be more precise, it is frequently used to predict financial data [10]. This model is very similar to the ARMA model used in time series analysis, however, the ARMA model can only be used for stationary series, while only the ARIMA model is available for non-stationary series, which can be written as ARIMA(p, d, q) consisting of AR(p) and MA(q) and d, where d represents the number of differences utilized to transform the non-stationary series into a stationary one. The model's equations can be expressed as follows:

$$y_t = \mu + \sum_{i=1}^{p} r_i y_{t-1} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \quad (1)$$

2.1. Dataset

The data set used for this experiment was selected from New York Stock Exchange stock market information, which comprised "Adj Close", "Opening Price", "Closing Price", "High Price", "Low Price" and "Volume". The Closing Price is chosen as the predictive feature in this paper since it reflects the data flux throughout the trading day.

The stock price of Ford Motor Company (NYSE: F) from June 1, 1972, to March 24, 2022, was chosen as the experimental data set, which is of the shape (12563, 6) and contains 12562 days of stock prices in total, which is a large capacity data set. Figure 1 depicts the Time Series Trend of the closing price of Ford shares over the preceding 50 years.



Figure 1. Ford Motor company closing prices

The first 80 percent of the time series data in the data set are picked as the training set, while the remaining 20 percent are the testing set. More specifically, the sample that was used to train the model is referred to as the training set, while the sample set that was used to test the model's ability is referred to as the testing set. Figure 2 represent the training and test sets that have been chosen.

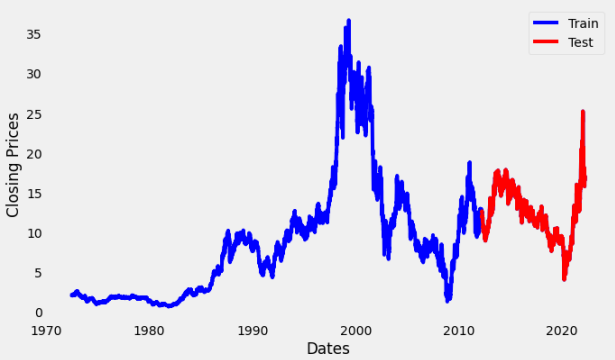


Figure 2. Training data and testing data

2.2. ARIMA Model

Stationarity is a key concept in model selection. The term "stationary series" refers to a data set in which the standard deviation and mean value stay unchanged throughout the evaluation. The term "stationary series" refers to the data in which the mean, variance, and slope remain constant over the course of the analysis. Models that are widely used in sequences such as ARMA, AR, and MA cannot be applied to non-stationary sequences, as they assume there is no unit root in the data and there is no trend in the mean and standard deviation. So, what if the series is the opposite of the above? In simple terms, when the data is non-stationary, the ARIMA model is applied, and the series is adjusted using the difference until it becomes stationary. The commonly used tests to check stationarity are Augmented Dickey-Fuller test [11], Moving Average graph and KPSS [12].

2.2.1 Moving Average graph

First and foremost, the moving average of Ford's closing price is plotted. In Figure 3 shows that the trend of the moving average of Ford's closing price varies over time and therefore can be seen as a non-stationary series due to its visible trend of change.

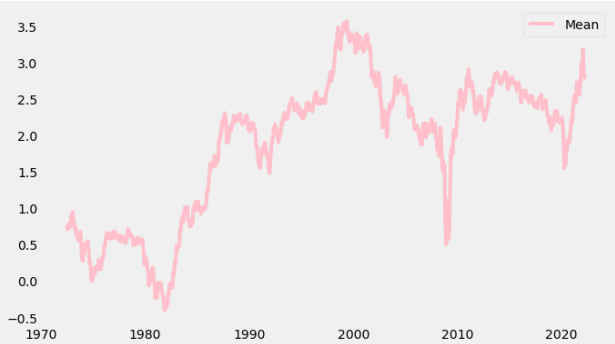


Figure 3. Moving Average graph

2.2.2 Augmented Dickey-Fuller test

Generally, the ADF test is very similar to the Dickey-Fuller test [13], where H0 assumes the existence of a unit root and is considered a non-stationary series to fail to reject H0 despite having a test statistic larger than the threshold value. Whereas the advantage of the ADF over the Dickey-Fuller test is that the Augmented Dickey-Fuller test excludes the effect of autocorrelation, the DF test can only be used as a unit root test for AR (1), and the ADF test is required for the rest of the cases.

Table 1. Augmented Dickey-Fuller test

Test Statistics	p-value	Critical Value		
		1%	5%	10%
-1.919	0.324	-3.431	-2.862	-2.567

From the Table 1, the value of test statistics is significantly greater than the Critical value at 10% and the p-value is relatively large, too. Therefore, H0 cannot be rejected. In summary, Table 1 can conclude that this series is a non-stationary series.

Following that, to convert the series into stationary format, the original one need to be processed by differencing. The parameter d in this section represents the times of differences that is needed during the completion of the data. After the first differencing, the ADF test is used once more and get the Table 2.

Table 2. Augmented Dickey-Fuller test

Test Statistics	p-value	Critical Value		
		1%	5%	10%
-18.051	2.638e-30	-3.431	-2.862	-2.567

As demonstrated in Table 2, the test statistics are significantly lower than the critical value of 1 percent, and the p-value is also significantly lower than 0.05, so H0 was successfully rejected. as a result. Henceforth, after performing a difference transformation on the series, it confirms that it becomes the stationary series.

3. Results and Discussion

3.1. Fitting model

The ARIMA model has three parameters ARIMA (p, d, q), the parameter AR(p) represents the quantity of lagged orders, which means analyzing the historical values of the data and making assumptions, and the order of the moving average is indicated by parameter MA(q), such that combining the correlations between the residuals.

In general, the parameters p and d can be obtained through ACF and PACF functions. ACF is known as the Autocorrelation Function, which is the correlation between historical and current values in a series, while PACF (Partial Autocorrelation Function) is a purer correlation obtained by removing the effects. In the previous section, it is known that there is a primary difference, so the parameter d=1. The parameters can be obtained by AIC. In essence, AIC is what determines how accurate a model is.

Table 3. Auto-ARIMA parameter

ARIMA Model	AIC	Time
(0, 1, 0) intercept	1405.253	0.76 s
(1, 1, 0) intercept	1284.776	1.57 s
(0, 1, 1) intercept	1287.083	0.88 s
(0, 1, 0)	1403.412	0.36 s
(2, 1, 0) intercept	1286.742	1.81 s
(1, 1, 1) intercept	1286.748	0.60 s
(2, 1, 1) intercept	1288.707	1.60 s
(1, 1, 0)	1282.974	0.36 s
(2, 1, 0)	1284.940	0.71 s
(1, 1, 1)	1284.946	0.44 s
(0, 1, 1)	1285.284	0.27 s
(2, 1, 1)	1286.904	0.54 s

The best model for prediction is suggested by the AIC's lowest value, but such access is relatively subjective, in order to get the best parameters, instead of conducting multiple experiments, one can

also use the auto ARIMA function. By using auto-ARIMA, the best parameter that is needed can be directly found, where the output is shown in Table 3.

According to the Table 3, there are three variables, which includes the Model, AIC and the reacting time. By comparing the value of AIC, it is clear that the AIC value of the model ARIMA (1, 1, 0) is 1282.974, which is the smallest value. Therefore, the best model is ARIMA (1, 1, 0), where the hyperparameter is p=1, d=1, q=0.

3.2. Prediction

Once the best parameters were obtained, the time-series model was built based on the selected hyperparameters. ARIMA (1, 1, 0) was then chosen as this parameter was considered to be the hyperparameter and a relatively good model could be obtained. The result is shown in Figure 4.



Figure 4. Ford's closing price prediction result

According to Figure 5, the line in green color is the prediction result, where the fluence during 2014 and 2020 is represented. Therefore, generally, this prediction has high accuracy, since it successfully predicts the decrease in stock price during that period, where it is able to provide a piece of valid information to the investor.

3.3. Accuracy

In summary, all the steps of the closing price forecast are described above, but finally, an accuracy test is required. The accuracy tests are generally MSE, RMSE, MAE and MAPE and for this test, the RMSE and MAPE tests were chosen. The RMSE is utilized in the calculation of deviation between the true and forecast data and the formula can be written as

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(x_i) - y_i)^2} \tag{2}$$

A model is regarded as a poor-quality when MAPE is greater than 1, can be written in the formula

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \cdot 100 g_0 \tag{3}$$

The answer obtained by calculation is as follows:

$$RMSE = 3.852394 \tag{4}$$

$$MAPE = 0.319954 \tag{5}$$

For a valid prediction, the smaller the value obtained for RMSE and MAPE, the better, and for this case, the RMSE was approximately equal to 3 and the MAPE was also less than 100%, so the accuracy of this experiment was relatively high, and the prediction could be considered good.

3.4. More dataset

Finally, apart from the Ford Motor Company's stock data, the experiment also brought in data from other stock markets, and most of the data sets had relatively good prediction result as Ford, but during the prediction of the stock closing prices of Google Inc. from 2004 to 2022, it was found that the predictions were very inaccurate, especially with RMSE test values is even greater than 100,

generally, this prediction is invalid. By carefully examining the data set characteristics, it was found that the difference between the stock price of google and other stocks is the size of the data set. To be more specific, Ford had over 10,000 historical data, but Google only has 4,000 closing prices, then by separating them into training and testing data, the data to train is not enough, hence, the author assume that the size of the data collection may have a direct effect on the accuracy of the predictions.

4. Conclusion

This paper is the application of time series analysis and prediction to stock forecasting. The reason for choosing this topic is that during these several years, there were many blows to the economy internationally and as a result, the stock market was very volatile, and investors were becoming more and more risk averse. Predicting stock trends gives investors a way to access visual trend data, allowing them to more efficiently select stocks that are trending better and reduce risk.

This experiment analyzed and predicted the closing price of Ford Motor on the NYSE over the last 50 years and as can be seen from Figure 6 the predicted trend is very clear, based on Ford's news between 2014 and 2021, it is clear that the share price has continued to move lower over this period, and this is shown on the graph and the experiment passed the accuracy test.

Therefore, even though there are still models with relatively inefficient forecasting results (Google Inc.), overall, the forecasts still have a good performance and the ARIMA model can be used as a reliable forecasting tool for actual stock price forecasting. In addition, time series analysis with the ARIMA model is also available in other sectors such as healthcare and criminology.

References

[1] Pindyck, and Robert S, et al. Risk, inflation, and the stock market. National Bureau of Economic Research Cambridge, Mass., USA, 1983.

[2] AChoji, Davou Nyap, Samuel Ngbede Eduno, and Gokum Titus Kassem. Markov chain model application on share price movement in stock market. Computer Engineering and Intelligent Systems, 2013, 4(10): 84-95.

[3] Svoboda M, Milan, and Ladislav Lukas. Application of Markov chain analysis to trend prediction of stock indices. Proceedings of 30th international conference mathematical methods in economics. Karviná: Silesian University, School of Business Administration, 2012, vol. 848, p. 853.

[4] Gong, Jibing, and Shengtao Sun. A new approach of stock price prediction based on logistic regression model. International Conference on New Trends in Information and Service Science, 2009, pp. 1366-1371. IEEE.

[5] Chen, Kai, Yi Zhou, and Fangyan Dai. A LSTM-based method for stock returns prediction: A case study of China stock market. IEEE international conference on big data (big data), 2015, pp. 2823-2824. IEEE.

[6] Soni, Sneha. Applications of ANNs in stock market prediction: a survey. International Journal of Computer Science & Engineering Technology, 2011, 2(3): 71-83.

[7] Bhuriya, Dinesh, Girish Kaushal, Ashish Sharma, and Upendra Singh. Stock market predication using a linear regression. International conference of electronics, communication and aerospace technology (ICECA), 2017, vol. 2, pp. 510-513. IEEE.

[8] Makridakis, Spyros, and Michele Hibon. ARMA models and the Box–Jenkins methodology. Journal of forecasting, 1997, 16(3): 147-163.

[9] Ariyo, Adebiyi A., Adewumi O. Adewumi, and Charles K. Ayo. Stock price prediction using the ARIMA model. UKSim-AMSS 16th international conference on computer modelling and simulation, 2004, pp. 106-112. IEEE.

[10] Mondal, Prapanna, Labani Shit, and Saptarsi Goswami. Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices. International Journal of Computer Science, Engineering and Applications, 2014, 4(2): 13.

[11] Mushtaq, Rizwan. Augmented dickey fuller test. Testing Time Series Data for Stationarity, 2011, pp. 1-19.

[12] Shin, Yongcheol, and Peter Schmidt. The KPSS stationarity test as a unit root test. Economics Letters, 1992, 38(4): 387-392.

[13] Baum, Christopher F. sts15: Tests for stationarity of a time series. Stata Technical Bulletin, 2000, 57: 36-39.