Stock Price Prediction using the ARIMA Model

Ziyue Yu*

Beijing Huijia Private School, Beijing, China

*Corresponding author: 25yuziyue@huijia.edu.cn

Abstract. Stock price prediction is a common topic in finance and economics that investors are interested in. Successful forecasts in stock price can bring significant profits to individuals and communities, thus, it is to be regarded as an important factor that is related to the whole economic society. Over the years, researchers have developed myriad models to figure out the pattern of price change. The autoregressive integrated moving average (ARIMA) model is one of the well-known statistical models. This paper is going to demonstrate the application of the ARIMA model using real stock data from the China Stock Market & Accounting Research Database (CSMAR). The research would be conducted with the assistance of SPSS. In this paper, results obtained not only refer the potential usefulness of the ARIMA model to identify and predict trends in the stock market, but also imply the oversimplification of the model based on a particular stock's performance in a real financial market.

Keywords: Stock price; prediction; ARIMA model.

1. Introduction

Prediction and forecasting play crucial roles in businesses and investments. Sales forecasting, for example, helps companies to formulate better financial planning, resource allocation, and decision-making [1]. Stock price prediction is one of the most popular themes in financial forecasting since investors can earn plenty of money with low risk if they predict the share market accurately. Although the random walk theory and efficient market hypothesis suggest that changes in stock prices are random and unpredictable, a great many economists still develop and use models to forecast the future values of stocks using historical data [2].

Over the years, researchers have conducted several methods of stock prediction. They can be mainly divided into 2 categories, statistical and soft computing techniques. Statistical techniques involve time series forecasting models, such as generalized autoregressive conditional heteroskedasticity (GARCH) and autoregressive integrated moving average (ARIMA) [3]. Separately speaking, the GARCH model is a great extension of the ARCH model that estimates future volatility by analyzing volatility in time series [4]. Economists utilize it as a technique to forecast the volatility of returns on financial assets when the variances are random. In comparison, ARIMA is a more popular model for most people. It is a linear regression model based on the statistical concept of serial correlation, where "AR" stands for autoregressive model and "MA" refers to the moving average model [5]. The AR model can determine the linear regression of present fitted values using past ones; the MA model presents the errors from the past while generating current values; and the Integrated(I) represents the times differencing that makes the time series stationary, thus, this model is also useful in real life situations that are non-stationary [6]. As a mixture of these, ARIMA is capable of predicting short-term market movements and highly used among economists [7]. Moreover, studies show that the 2 statistical techniques can be combined as a whole, which is called the ARIMA-GARCH model. It is found to be feasible because it has performed a good prediction effect through testing [8]. However, the complexity of financial markets makes stock price trends nonlinear. Then, some people argue that the results generated from traditional models may have inevitable errors [9]. By contrast, artificial neural network (ANN), a soft computing technique as a representation of human biological neurons, is widely used for estimating stock prices due to its efficiency in solving complex nonlinear problems by finding non-linear relationships within given data sets [10]. This model can provide competitive results compared to traditional time series models (e.g., ARIMA). Nevertheless, the most obvious challenge of the ANN model is thinking precisely when one determines the selection

of the optimum number of units (number of input and output units). Furthermore, the ANN is relatively difficult to conduct because researchers need to consider the training. Therefore, in dealing with short-term predictions, the ARIMA model is found to be more robust and efficient in financial time series forecasting than the ANNs techniques.

This paper is going to demonstrate the use of the ARIMA model in the prediction of stock price by analyzing real-life data and talking about its potential strength which may provide investors with information to aid their decision-making process. It's crucial to do this research as long as stock markets have significant influences on individuals and the economy as a whole. Finally, the weaknesses and limitations of the ARIMA model will be discussed, to inform investors about its advantages and imperfections objectively.

2. Methods

2.1. Data Sources

The data that will be used in this paper is acquired from the China Stock Market & Accounting Research Database (CSMAR). Its professional standards are learned and similar to those of CRSP, Compustat, TAQ, and some other databases with international awareness. The historical close share prices of Ping An Bank (a public company in China) are going to be analyzed using the ARIMA model. This paper only pursues the daily price changes in 3 years (2020/9/15 – 2023/9/15) for this company.

2.2. Model Introduction

In the ARIMA model, the future values can be forecasted by a linear combination of past values and errors, which can be expressed as:

$$y_t = I + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots + \alpha_p y_{t-p} + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \ldots + \theta_q e_{t-q} \quad (1)$$

where y_t is the actual value, α and θ are coefficients, p, and q are orders of AR and MA models, and I refer to the data values that have been replaced by different values d.

2.3. Research Protocol

To conduct this research, software tools such as Microsoft Excel and SPSSAU are needed. SPAASU is a data analysis platform that is used for lots of mathematical models, including the ARIMA model. In this paper, the degree of differencing d can be determined by checking the augmented dickey-fuller test (ADF); the values of p and q are estimated from the maximum lag points of the partial autocorrelation function (PACF) and autocorrelation function (ACF).

3. Results and Discussion

3.1. Stationary Analysis

For a time-series problem, the stationarity of data needs to be considered first. ARIMA model only takes place in analyzing stationary data, which have no changes in mean and variance over time. As an icon tool, the sequence diagram (which can be a scatter diagram) can immediately indicate whether the time series is stationary.

Specifically to the research, the trading dates and close prices of Ping An Bank are organized by a chart in chronological order using Microsoft Excel. The first 18 rows are shown in Table 1.

Table 1. Trading Date and Close Price of Ping An Bank

Trading Date	Symbol	Close Price
2020-09-01	000001	2276.166
2020-09-02	000001	2303.228
2020-09-03	000001	2240.085
2020-09-04	000001	2249.105
2020-09-07	000001	2246.098
2020-09-08	000001	2319.765
2020-09-09	000001	2286.69
2020-09-10	000001	2306.235
2020-09-11	000001	2256.622
2020-09-14	000001	2300.221
2020-09-15	000001	2307.738
2020-09-16	000001	2321.269
2020-09-17	000001	2340.813
2020-09-18	000001	2415.984
2020-09-21	000001	2384.412
2020-09-22	000001	2340.813
2020-09-23	000001	2349.834
2020-09-24	000001	2273.16

Then, a scatter diagram is made with the trading date as the x-axis, and the close price as the y-axis.

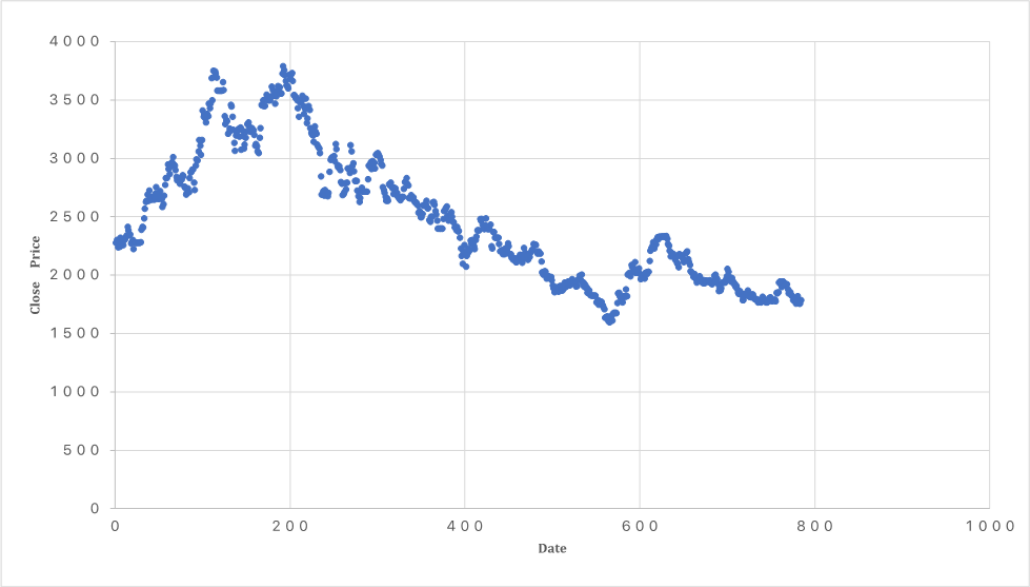

Fig. 1 Scatter Plot of Trading Date and Close Price

The trend shown in Figure 1 is non-stationary because the scatters in the plot show a significant decline. Thus, the data needs to be processed while taking it into an ARIMA model.

3.2. Determine d, p and q of ARIMA Model

3.2.1 ADF test

Besides looking at the sequence diagram, an ADF gives a more accurate result of stationarity by checking the presence of a unit root. The original hypothesis of the ADF test is that the series is non-stationary. Generally, while p is smaller than 0.5, the hypothesis can be rejected, that is, the time series is stationary.

Table 2. Close Price ADF Test

Differencing Degree	t	p	Critical Value		
			1%	5%	10%
0	-1.099	0.716	-3.439	-2.865	-2.569
1	-28.118	0.000	-3.439	-2.865	-2.569

In Table 2, since the p of raw data equals 0.716, which is larger than 0.5, the time series is non-stationary such as what is indicated in the scatter diagram. When the difference degree becomes 1, p equals 0.000 and it is much smaller than 0.5, which has a 99% percent possibility to refuse the hypothesis. Thus, set d equals 1 is suitable.

3.2.2 ACF and PACF

An ACF measures the correlation of the same event in different time period; a PACF determines the impact that one factor brings to another without considering other factors.

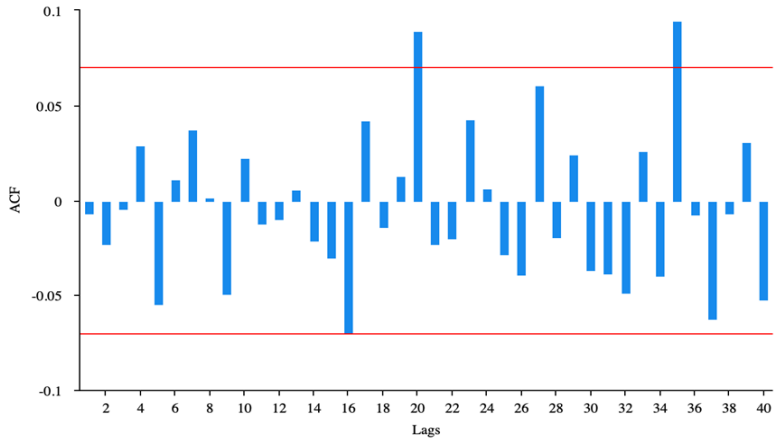

Fig. 2 ACF plot

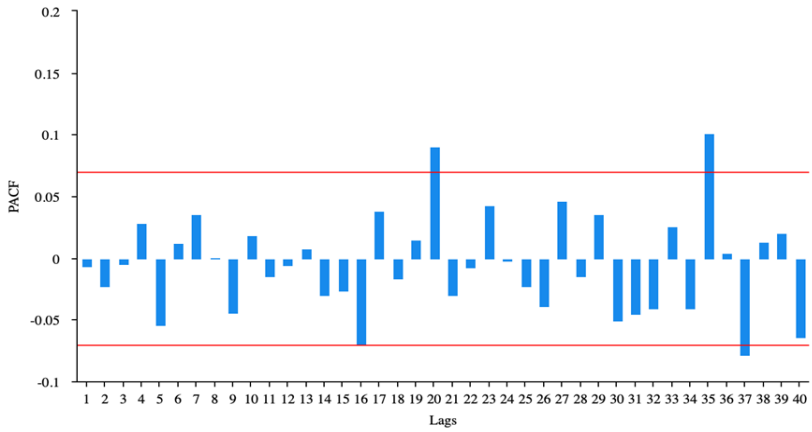

Fig. 3 PACF plot

According to Figure 2 and Figure 3, the SPSS automatically identifies and suggests the values of autoregressive order p value is 1, and the moving average order q value is 1.

3.3. ARIMA Model Prediction

While the degree of differencing (d), the order of AR (p), and MA models (q) are all deduced, the ARIMA(1,1,1) is considered to be the best model for this case, which can be constructed by SPSSAU.

Table 3. ARIMA (1,1,1) Model Parameter

Term	Symbol	Coefficient	S. E.	z Value	p Value	95% CI
Constant Term	c	-0.039	0.148	-0.264	0.792	-0.329 ~ 0.251
AR Parameter	α1	0.939	0.128	7.344	0	0.688 ~ 1.189
MA Parameter	β1	-0.947	0.119	-7.981	0	-1.180 ~ -0.715

Table 3 shows the model construction results. According to the table, the formula for the model can be written as:

$$y_t = -0.039 + 0.939y_{t-1} - 9.47\varepsilon_{t-1} \quad (2)$$

Table 4. Model Q Statistics

Term	Statistical Magnitude	p Value
Q_6	0.001	0.975
Q_{12}	0.187	0.911
Q_{18}	0.191	0.979
Q_{24}	1.183	0.881
Q_{30}	2.995	0.701
Q_{36}	3.224	0.78

After construction, the ARIMA model is expected to be white noise. For Table 4, generally, only Q_6 needs to be looked at. While the p-value is bigger than 0.1, the model passes the white noise test. The p value is 0.975 at Q_6 in this research, thus, the ARIMA model can be normally used.

Fig. 4 Close Price Model Fitting and Prediction

Figure 4 presents the actual data graph, model fitting value, and model predicted value. It is not hard to see that the actual one and fitting one are highly similar or even coincident, which declares that the model has a good fitting for the real-life situation.

Table 5. Predicted Value (7 phases)

Lag1	Lag2	Lag3	Lag4	Lag5	Lag6	Lag7
1788.3	1787.8	1787.4	1786.9	1786.5	1786	1785.5

The most important part that researchers expected to see is the predicted value chart. The ARIMA model forecasts the close price of Ping An Bank in the next 7 days (Table 5). A comparing table is being shown to identify the model's precision.

Table 6. Predicted Value and True Value

Date	Predicted Value	True Value
2023/9/4	1788.3	1826.589
2023/9/5	1787.8	1799.727
2023/9/6	1787.4	1806.048
2023/9/7	1786.9	1790.247
2023/9/8	1786.5	1780.766
2023/9/11	1786	1791.827
2023/9/12	1785.5	1782.346

Table 6 indicates that the ARIMA model shows an impressive effect in this case as its general trend is fit and some values are pretty related to the actual ones, especially on 2023/9/7 and 2023/9/12. However, some deviations occur, too. The real close price of the business is more volatile than the data generated from the model.

4. Conclusion

This paper presents the operated process of selecting and structuring the best ARIMA model to predict a particular stock's performance in a real financial market. The procedures show how to select the most suitable parameters for a given data. and the experimental results imply the potential usefulness of the ARIMA model in stock price forecasting. It can provide investors a general direction about the fluctuations in the stock market that may help them to make a more informed judgment during purchasing. Nevertheless, the erroneousness of the model is also non-negligible. In the case of Ping An Bank, the real share price experiences a more turbulent trend, which reveals that financial markets are much more complex than what models predict. In addition, the ARIMA model for this research does not render its accuracy in predicting short-term data. Thus, whether the stock price trends are random is still questionable. Economists should continue to work to reduce the uncertainties of using the ARIMA model in stock price prediction so that a more secure decision can be ensured.

References

[1] Smith Tim. Random Walk Theory. Investopedia, 2019.

[2] Adebiyi, Ayodele Ariyo, et al. Comparison of ARIMA and Artificial Neural Networks Models for Stock Price Prediction. Journal of Applied Mathematics, 2014.

[3] Hayes Adam. Autoregressive Integrated Moving Average (ARIMA). Investopedia, 2022.

[4] Gao Jie. Research on Stock Price Forecast Based on ARIMA-GARCH Model. Web of Conferences, 2021, 292.

[5] Ma Qihang. Comparison of ARIMA, ANN and LSTM for Stock Price Prediction. Web of Conferences, 2020.

[6] Ariyo A A, et al. Stock Price Prediction Using the ARIMA Model. 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, 2014.

[7] Dhaduk, Hardikkumar. Stock Market Forecasting Using Time Series Analysis with ARIMA Model. Analytics Vidhya, 2021.

[8] Chaudhary Mukesh. Why Is Augmented Dickey–Fuller Test (ADF Test) so Important in Time Series Analysis. Medium, 2020.

[9] Wu Songhao. Stationarity Assumption in Time Series Data. Medium, 2021.

[10] Lin Chunyan, Zhu Donghua. Research on Stock Price Prediction Based on Elman Neural Network. Computer Application, 2006, 26(2): 3.