Stock Change Predictions with ARIMA Model

Zihao Guo[1]

EMAIL: 20202333057@m.scnu.edu.cn (Zihao Guo)

[1] South China Normal University, No. 55, Zhongshan Avenue West, Tianhe District, Guangzhou, Tianhe District, Guangdong, China

Abstract: The Stock investment is a wealth management model, based on the customer's financial situation, financial goals and risk preferences and other factors for customers to generate intelligent and personalized investment solutions. However, algorithm for predicting stock changes is complex. And wide range of equity investment options based on different algorithms and factor indicators. Therefore, choosing options for investors is difficult. To address the wide disparity between different equity investment options, the present study proposes that forecasting analysis of stocks based on ARIMA time series models. First, the previous stock data will be filtered, and the ARIMA time series model will be used to forecast the changes The results of this study showed that compared to the multiple linear regression prediction model that is based on ARIMA time series model of the Stock prediction, the prediction root mean square error decreased by 25%, the correlation coefficient increased by 15%. Therefore, ARIMA time series model can help the investors to choose the Stock exactly, and it has important role in stock price forecasting.

Keywords: Stock prediction; the ARIMA time series model; the multiple linear regression model

1. Introduction

In the global financialization of the background, the number of companies is increasing, and the stock types is various, the capital market has been rapid development. Individual investors have the analytical capacity restricted and access to information channels for a single, as a result, individual investors lack rational analysis and a long-term perspective [1]. Therefore, how to accurately predict Stock of the changes, for investors, is also very important.

The Stock forecast is a popular area, the accuracy of the Stock forecast for investors and institutions are very important. Mehtab S use CNN and LSTM model for the stock price time series for the robustness analysis [2]; and Emma Katariina Kontsas for the new stock market under the volatility forecast research it [3].

This study analyses and forecasts stock data based on the ARIMA time series model. Firstly, the parameters and order of the ARIMA time series model are determined, and then the ADF test is used to verify whether the time series model is stable, if it is stable then it is steady state, otherwise it is differenced to obtain a smooth time series model. This study uses this model to forecast the changes in individual stocks and to verify the accuracy of this time series model, and to compare and analyze the results with those of other models. The results of the study show

FFIT 2023, July 07-09, Chongqing, People's Republic of China

DOI 10.4108/eai.7-7-2023.2338052

that the ARIMA time series forecasting model reduces the root mean square error of forecasts by 25% and increases the correlation coefficient by 15% compared to the multiple linear regression forecasting model.

2. Building the Model [4]

The present study adopted a time series analysis method to predict a single stock for the quality of the performance, where ARIMA time series model is commonly used in time series analysis method. The ARIMA model is an autoregressive model to describe the current value and historical value of the relationship, with the variable itself of historical data on their own to predict, contain autoregressive （AR） , differential(I) and the moving average MA three parts, and therefore also referred to as $\boldsymbol{ARIMA}\ (\boldsymbol{p, d, q}\ )$model.

When using changes in time series stocks, it is first determined whether the time series is stable or not, and if it is not smooth, it is differenced. Then, determine the time series p and q. After obtaining the ARIMA model, compare it with a multiple linear regression model. [5]. On this study, the raw data was first pre-processed and examined to find that there were no outliers in the study data, but there were missing values, so an average fill was used to fill in the missing values.

2.1. Inspection of the time series stationarity

This study uses the ADF test to test the smoothness of a time series, which is based on the principle of the presence or absence of a unit root in a time series to determine whether it is smooth: If there is no unit root in the series, the series is considered to be smooth, and vice versa, the series is considered to be non-smooth [6]. Based on this, this study proposed the original assumptions is $\boldsymbol{H}_{\boldsymbol{0}}$ and alternative assumptions $\boldsymbol{H}_{\boldsymbol{1}}$:

$\boldsymbol{H}_{\boldsymbol{0}}$ assumptions: the sequence is not the presence of unit root and is non -stationary sequences

$\boldsymbol{H}_{\boldsymbol{1}}$ assumptions: the sequence is the presence of unit root and is stationary sequences

Assume that the time series of $\boldsymbol{X}_{\boldsymbol{t}}$, and that the $\boldsymbol{Xt}$ of the changes is with a drift term and trend term of the autoregressive process, then $\boldsymbol{X}_{\boldsymbol{t}}$ satisfies the following equation:

$$\boldsymbol{X}_{\boldsymbol{t}} = \boldsymbol{\mu} + \boldsymbol{\rho}\,\boldsymbol{X}_{\boldsymbol{t-1}} + \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} \boldsymbol{C}_{\boldsymbol{i}}\,\boldsymbol{X}_{\boldsymbol{t-i}} + \boldsymbol{\varepsilon}_{\boldsymbol{t}} \qquad (\mathbf{1})$$

After deriving the constraints and assumptions, this study tested the significance of the constraints and assumptions. Assuming a significance level of α with values of 1% and 5% and 10% and a probability of $\boldsymbol{p}$ for the test statistic, the series is very smooth if $\boldsymbol{p}$ > α and the null hypothesis is not rejected.

2.2. Determine the order of the time series

In this study, the $\boldsymbol{X}_{\boldsymbol{t}}$ is first differenced. Assuming that the post-processed kth order differential series is $\boldsymbol{\Delta}\,\boldsymbol{y}_{\boldsymbol{t}}$, the following equation is given according to the definition of kth order differential equation:

$$\Delta y_t = \sum_{i=0}^{k} (-1)^i C_k^i X_{t+k-i} \quad (2)$$

The differential order is determined by performing a smoothness test for each k value corresponding to the differential series Δy_t such that Δy_t satisfies the smoothness condition and the smallest k value is chosen.

This study is based on the red pool information criterion (AIC) to determine the parameters of the ARIMA model. The AIC is a criterion to assess the complexity of the model and to measure the quality of the model fit. The AIC criterion is expressed as:

$$AIC = 2k - \ln(L) \quad (3)$$

k is the number of parameters, and L is the likelihood function.

The AIC criterion balances the complexity and accuracy of the model, so when selecting a model, the one with the smallest AIC value should be preferred.

2.3. Testing the reasonableness of the model

The present study used the residual sequence in order to test the model of rationality, the if test to obtain the residual sequence is a White noise sequence, the description of ARIMA model is suitable.

Assuming White noise for ε_t, and then ε_t satisfies the following conditions:

$$E(\varepsilon_t) = \mu \quad (4)$$

$$var(\varepsilon_t) = \sigma^2 \quad (5)$$

$$\text{cov}(\varepsilon_t, \varepsilon_s) = 0, t \neq s \quad (6)$$

To simplify the study, the following assumptions are made: the delay of the kth order autocorrelation coefficient of the sample series is around 0, satisfying the condition of white noise. A residual series that is approximated as a white noise series indicates that the series is not affected by time and can be used as a predictive model.

2.4. Predicting changes in stocks

In get to build ARIMA time series models, can be achieved fora certain stock future changes in the forecast. Based on $ARIMA(p, q, d)$ order number and parameters, give the future the first h of days the predicted value of y_{t+h}, the prediction data before i day of history data of y_{t+h-i} relationship is as follows:

$$y_{t+h} = a + \sum_{i=1}^{h} b_i y_{t+h-i} + c_i \epsilon(t + h - i) \quad (7)$$

Where $\epsilon(t + h - i)$ is an unknown random variable, the need for simulation or using the history of the residuals of the mean estimate, the present study used historical residuals of the mean value to estimate it.

This article uses the root mean square error （$RMSE$） , the average absolute error(MAE), the decision coefficient(R^2)as the evaluation index of. They are defined as follows:

$$RMSE = \sqrt{\sum \frac{1}{n}(\hat{y}_i - y_i)^2} \quad (8)$$

Where in the number of samples is n, the predicted value of $\hat{y}_i$, the true value of y_i on.

$$MAE = \sum |\hat{y}_i - y_i| \quad (9)$$

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (10)$$

3. Experimental results and analysis

3.1. The ADF test

Table 1 ADF test table

variable	d	t	P	AIC	critical value		
					1%	5%	10%
Stock price	0	-1.623	0.471	792.62	-3.44	-2.866	-2.569
	1	-8.461	0.000***	777.06	-3.44	-2.866	-2.569
	2	-11.152	0.000***	781.25	-3.44	-2.866	-2.569

Note: ***, **, * represent 1%, 5%, 10% level of significance respectively

As shown in Table 1, on the table of the ADF test results, including the variables, the order of differencing d, the test results of the t, the AIC value, etc. The sequence of the test results displays that: In the differential to the order 0, the significant P-value of 0. 471, do not exhibit significant, cannot reject the null hypothesis that the sequence is non-smooth time series. In the differential of order 1, the significant P-value of 0. 000***, presenting significant, reject the null hypothesis, the sequence of stationary time series. The present study is based on AIC guidelines, select the time series of the difference of order1, d=1.

3.2. The ARIMA model parameters table

Table 2 ARIMA (1,1,1) inspection table

item	symbol	value
Q Statistic	Q6(P value)	0.013(0.910)
	Q12(P value)	8.374(0.212)
goodness-of-fit	R2	0.782
Information criteria	AIC	777.06

Note: ***, **, * represent 1%, 5%, 10% level of significance respectively

As shown in Table 2, the system is based on the AIC information criterion to automatically find the optimal parameters, the model results for the ARIMA (1,1,1) inspection table, based on the variables of the Stock's overall performance on. From the Q Statistic analysis of the results can be obtained: Q6 on the level does not exhibit significant, cannot reject the model of the residual is white noise sequence of the assumption, while the model's goodness-of-fit R2 of 0.782, the

model showed a relatively good model to meet the basic requirements. The parameters of the model are shown in Table 3.

Table 3 Model Test table

	coefficient	standard deviation	t	P>|t|	0.025	0.975
a	0.524	0.395	1.330	0.184	-0.249	1.229
b_1	-0.325	0.433	-0.745	0.456	-1.170	0.526
c_1	458.06	5.652	8.101	0	37.42	56.80

Note: ***, **, * represent 1%, 5%, 10% level of significance respectively

3.3. Prediction results of the ARIMA model

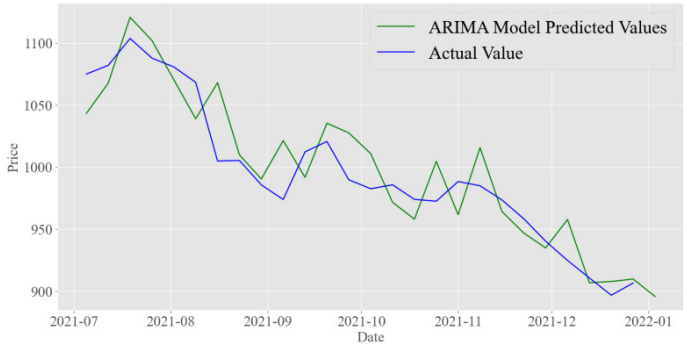

Figure 1: Predicted and actual values from the ARIMA model

As shown in Figure 1, contrast ARIMA the predicted value of the result and real data results can be drawn: ARIMA model predicted trends and stock trends the basic agreement on. But due to the price of the Stock's volatility is strong, not absolutely accurately forecast the Stock price.

3.4. The prediction results comparison

As a comparison, this study used a multiple linear regression model to fit the predictions, and the predictions are shown in Figure 2:

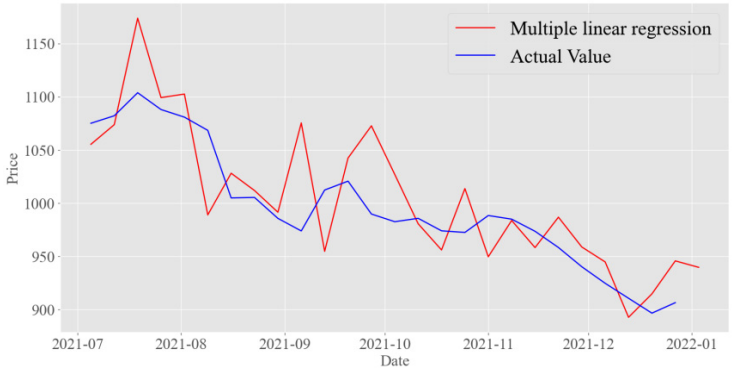

Figure 2: Multiple linear regression predicted value and the true value

The analysis of two model results, as can be seen both predictions are having lag on. Andin the short term, the two models fitting accuracy is not high; however, in the long term, the two models predict the results can reflect the Stock's changing trends in.

A comparative analysis shows that, compared to the multiple linear regression model, the ARIMA time series has good predictive stability and is still very useful in predicting stock movements for a more volatile type of data such as stocks.

3.5. Model evaluation index comparison

Table 4 evaluation results table

	ARIMA	**Multiple Linear Regression**
$RMSE$	**19.2**	**25.6**
MAE	**1425**	**2041**
R^2	**0.867**	**0.732**

Table 4 shows that ARIMA model and the multiple linear regression model corresponding evaluation indicators, the analysis concluded that ARIMA model of evaluation indicators are better than the multiple linear regression model, wherein the root mean square error$RMSE$ decreased by 25%, the correlation coefficient R^2 increased by 15%. Indicates that ARIMA (1,1,1) model for the Stock changes the prediction effect is better in.

4. Conclusion and Discussion

Forecasting of stocks is a complex problem. This study proposes a forecasting analysis of stocks based on the ARIMA time series model and compares it with a multiple linear regression model. The results of the study show that the root mean square error of forecasting stocks based on the ARIMA time series model is reduced by 25% and the correlation coefficient is increased by 15%.

The ARIMA model is a classical time series model that can be used not only to forecast changes in stock prices, but also for predictive analysis of population size and crop yields. In the future, this study will optimize the parameters of the ARIMA model in order to predict changes in stock prices more accurately.

References

[1] Sun Chen-Hao Wang Lin. Based on ARIMA and LSTM Shanghai Composite Index forecast and analysis[J]. Computer engineering application technology, 2023(2):29-31.

[2] Kontsas E K. Volatility Forecasting in Emerging Markets[D]. The University of Vaasa, 2020.

[3] Mehtab S, Sen J, Dasgupta S. Robust Analysis of Stock Price Time Series Using CNN and LSTM-Based Deep Learning Models[C]: ICECA'2020, Cornell University, 2020.

[4] Chen Dengjian, Du Feixia, Xia Chang. Stock forecasting based on a combination of ARIMA and SVR rolling residual models[J]. Computer Age,2022, No.359(05):76-81.
[5] Zhou WJ, Zhou QD, Zhuge B. LSTM-based stock prediction model [J]. 2023.
[6] Song Rui-biao, Liu Hai-hong. A study of sliding GA-BP-GRACH model based on wavelet analysis for stock prediction[J]. Mathematical Practice and Understanding, 2023, 53(3).