PDD Stock Price Prediction Using ARIMA Model

Yutong Ge[1, a]

[1]China Agricultural University, Beijing, 100089, China
[a]E-mail: yutong.ge@ucdenver.edu

Abstract
In order to study and predict the short-term stock price of pinduoduo, an emerging e-commerce company in China, the sample data of its closing prices of January 1, 2020, to January 1, 2021, are selected as the research object. Firstly, ADF stationarity test is used to judge whether the time series is stable, and then ARIMA model is established by R language. Based on the model, the prediction results of the last 13 data are tested. The results show that the ARIMA model can accurately predict pinduoduo's short-term stock price.

Keywords: *Pingduoduo, stock price, time series, ARIMA model, adf test.*

1. INTRODUCTION

Pinduoduo is one of the most popular mobile online e-commerce applications in China. It is a third-party social e-commerce platform focusing on C2M (customer-to-manufacturer) shopping. By organizing group purchasing with your friends, classmates, family members, neighbours, net friends, as well as even the people you do not know about, as long as users are willing to purchase the same product, they are able to consume productions that are high in quality and low in price. Communication and sharing have shaped Pinduoduo's unique new social e-commerce ideology. On the one hand, research on Pinduoduo's stock price can help us to assess whether it has investment value, so that investors will be able to find and catch the best time of investment. On the other hand, this research will be effective in helping us to explore the way out of Pinduoduo's new business model for the lower income group. Pinduoduo officially listed in the US capital market on July 26, 2018, with an issue price of $19 and a market value of $24 billion. This paper will use its historical stock price data to predict the trend of its stock price. Previously, on December 29, 2020, Pinduoduo's share price rose 15.57% in the U.S. stock market trading on that day, reaching a record high. Pinduoduo shares closed at $166.19, up $22.39. It is not only because of Pinduoduo's special shopping mode, but also attributed to the unexpected trend of the market in 2020.

2. LITERATURE REVIEW

Time series forecasting is based on the understanding of the economic process, and then uses the historical experience to make the future forecast. J. H. stock (1996) has made great progress in economic forecasting of time series [1]. Liu and Morley (2009) think that the traditional econometric model has the possibility of serious deviation [2]. Based on this problem, scholars try to constantly improve the traditional econometric models, including ARIMA based stock price forecasting model. The results show that the model can predict the trend of stock price.

Paul et al. (2013) tried to determine the best ARIMA model for predicting the average daily stock price index of a pharmaceutical company in Bangladesh [3]. It is found that ARIMA (2,1,2) is the best prediction model. Similarly, wahyudi (2017) attempted to predict stock prices using ARIMA's price fluctuation model for stocks listed on the Indonesian stock exchange [4]. At the same time, we use Indonesia composite stock price index to analyze the empirical results, and prove that the best ARIMA model is (0,0,1). This model is based on the standards established by AIC. Zhang Yingchao et al. (2019) established ARIMA (4,1,4) model [5]. The results show that the prediction effect is related to the time range of the prediction, and the prediction is accurate in the short term. However, the long-term prediction effect does not seem optimistic.

The ARIMA model has been applied to COVID-19's research recently. Prior to the covid-19 epidemic [6], alzahrani et al. (2020) attempted to predict the daily increase in Saudi Arabia cases. Singh et al. (2020) used ARIMA model to predict the transmission[7]. As of April 2020, the transmission trajectory and mortality of covid-19 in the top 15 countries. The mortality rate of the

A. El-Hashash et al. (Eds.): IEIT 2022, ACSR 100, pp. 83–87, 2023.
https://doi.org/10.2991/978-94-6463-058-9_15

epidemic will decrease in China, Switzerland and Germany. However, the United States, Spain, Italy, France and the United Kingdom will witness an increase in the spread of the virus.

3. METHOD

3.1.DATA SOURCE

We will use the stock price data of Pinduoduo in Yahoo Finance from January 1, 2020, to January 1, 2021, and take the logarithmic difference of its closing price to explore logarithmic stock price return. We use R language programming and ARIMA model to analyze and process the data and based on the data of this year to get the prediction model, to predict the change trend of Pinduoduo stock price in the first month of 2021.

3.2. TOOLS FOR PREDICTION

ARIMA model is called autoregressive integrated moving average model. Also known as ARIMA (p, d, q), is one of the most common statistical models used for time series prediction. The advantage of ARIMA is that the model is relatively simple, only endogenous variables are needed, and other exogenous variables are not needed. But it requires time series data to be stationary, or stationary after differentiation.

ARIMA model combines three basic methods

Autoregressive (p): In autoregression, a given time series data has their own lag value, which is represented by the "p" value in the model. Using the weighted sum of past values, a regression equation is created to describe the data points of time series data regressed according to their lag values. Unlike multiple regression, which is based on linear combination of exogenous variables, autoregression uses a single predictor.

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} + \varepsilon_t \qquad (1)$$

Integration via Differences(d): This involves differencing time series data to eliminate trends and converting non-stationary time series to static time series. This is represented by the "d" value in the model. The difference subtracts the original data observations for the current period from the previous observations until the data does not grow at an increasing or decreasing rate.

$$\boldsymbol{Y_t - Y_{t-1}} \qquad (2)$$

Moving average model:(q): The model creates trend tracking or lag indicators based on variance data to help determine the probability of an upward or downward trend. The longer the time period of the moving average, the greater the lag and the possibility of change. The moving average property of the model is represented by the "q" value, which is the number of lag values of the error term.

Therefore, ARIMA model can be represented by the combination of AR and MA model, ARMA model, and on this basis, d-order difference must be made to satisfy the assumption of stationarity.

$$\boldsymbol{Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}} \qquad (3)$$

4. ANALYSIS

First of all, we consider the value of D in ARIMA (p, d, q), that is, how many differences we need to make to satisfy the smoothness assumption of ARIMA model. After drawing the time series graph of log return (figure 1), viewing the Y axis of the figure, the data oscillates around 0. This means that the data can be roughly estimated to be smooth. The fluctuation of the chart can describe the trend of stock value. In order to be more accurate, we use Augmented Dickey Fuller Test test to determine the unit root.

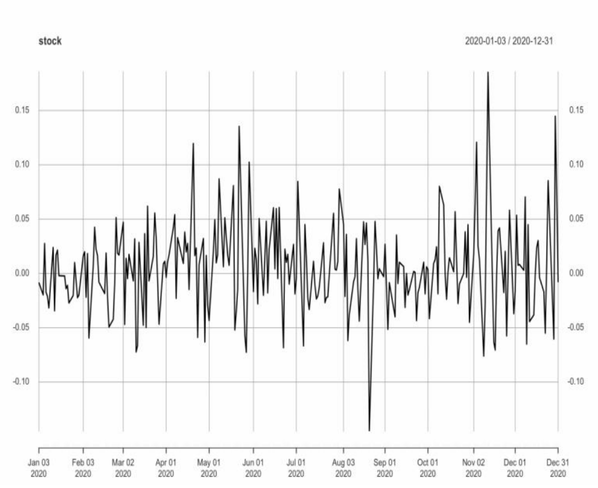


Figure 1. Time series diagram of logarithmic regression

Ho: time series data are non-stationary. The average of the data will change over time.

Ha: time series data is fixed. The average of the data does not change over time.

We get a value of P less than 0.01, which means that we can reject Ho and get the results that the data is stationary, and no further difference is needed. Thus, the value of D is equal to 0.

Secondly, we start to think about the value of p, q.

ACF(p): complete auto-correlation function in order to find values of autocorrelation of any series with its lagged values.

PACF(q): partial auto-correlation function to get the correlation of the residuals with the next lag value.

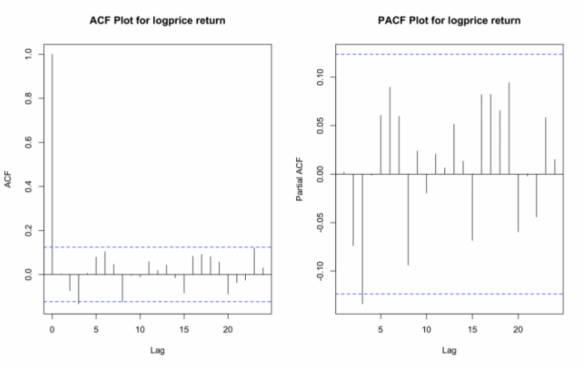


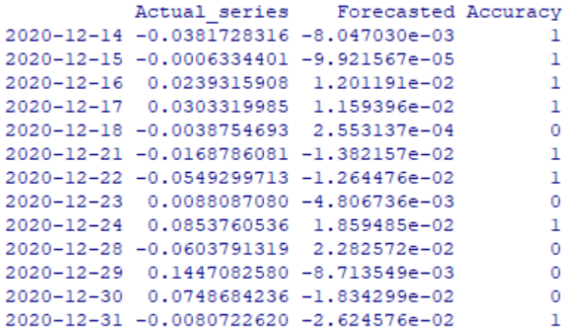
Figure 2. ACF & PACF plot for log price return

People usually use ACF and PACF images to determine the value of p, q. But we can see from Figure 2 that we can't get the exact p and q values. However, we can guess that the values of p and q can be equal to 1, 2 or 3, because after lag is greater than 3, there is no line exceeds the confidence bond, which means that there is no strong autocorrelation between lag values greater than 3.

Therefore, we tried different combinations of p, q values equal to 1, 2, 3 and compare the AIC of these models. The smaller the AIC value is, the stronger the model fitting is. So we finally chose the ARIMA (2, 0, 2) model, which is also the R command “auto.arima” returned.

We want to test on the prediction power of our model by compare the real log returns and the forecasted log returns. The way we achieve this test is followed by the four steps: Firstly, we split the dataset into a training and testing set by introducing a breakpoint. Secondly, using training set and ARIMA (2,0,2) model to build a regression equation. Thirdly, we generate predicted values for the next trading day after the training set based on the regression equation built. Finally, comparing real log returns with forecast log returns and calculate the proportion of accurate predictions to see how accurate our prediction is.

Our results are shown in figure 3 and figure 4. In figure 4, the red line is our forecasted log returns, and the black line is the real returns. By comparing the two lines, it shows that the general trend is very similar. Because we are using the log returns, so when the sign of the predicted value is consistent with the actual value, that is, when we accurately predict the increase or decrease trend of the stock price, we count it as an accurate prediction. Calculating the percentage of the accurate prediction, it shows 61.54% of times in the testing set, we have a correct prediction for the stock price changes.



```
           Actual_series    Forecasted Accuracy
2020-12-14 -0.0381728316 -8.047030e-03        1
2020-12-15 -0.0006334401 -9.921567e-05        1
2020-12-16  0.0239315908  1.201191e-02        1
2020-12-17  0.0303319985  1.159396e-02        1
2020-12-18 -0.0038754693  2.553137e-04        0
2020-12-21 -0.0168786081 -1.382157e-02        1
2020-12-22 -0.0549299713 -1.264476e-02        1
2020-12-23  0.0088087080 -4.806736e-03        0
2020-12-24  0.0853760536  1.859485e-02        1
2020-12-28 -0.0603791319  2.282572e-02        0
2020-12-29  0.1447082580 -8.713549e-03        0
2020-12-30  0.0748684236 -1.834299e-02        0
2020-12-31 -0.0080722620 -2.624576e-02        1
```

Figure 3. The comparison values of returns results

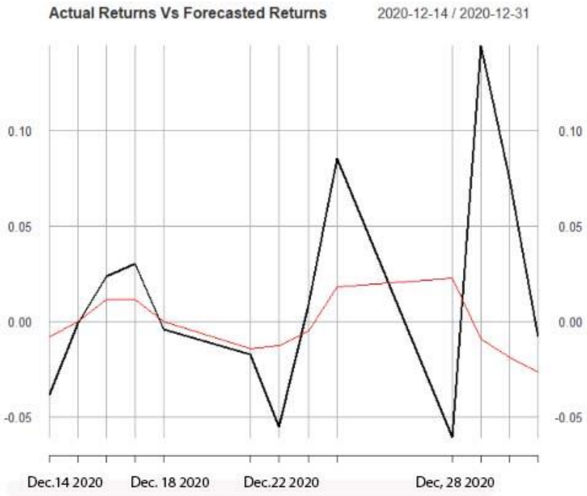


Figure 4. The comparison charts of returns results

Therefore, we believe that the model can help us correctly predict the change of stock price. Then we generated our final ARIMA (2,0,2) model by using the log returns in all trading days or year 2020.

$$\mathbf{Y_t = 1.0104Y_{t-1} - 0.9738Y_{t-2} - 1.0353\varepsilon_{t-1} + 0.9424\varepsilon_{t-2} + 0.0058} \quad (4)$$

When we are testing the prediction power, a for loop is used to generate the prediction values for all trading days in the testing set. For each cycle, we need to test on the residuals of the model to ensure the applicability of the models. Here, we use the final model to see the process of testing on residuals.

Figure 5 and figure 6 displays the lagged scatterplot, Q-Q plot, ACF and PACF plot. From these we can see that the expected value of the residuals is equal to 0. The quantile of the residual is roughly in a line, which means that the residuals is normal distributed. There is no value that exceeds the bound line, which means that there is no autocorrelation in the residuals. To further prove this, we need to do the Ljung–Box test. Ljung–Box test is always used in testing the residuals in ARIMA model to see they are independently distributed or not. And the H_0 is that the data are independently distributed. From figure 7, running this test, we get a very big p-value 0.9829, which means that we fail to reject H_0, so that the residuals are independently distributed and the ARIMA model can be used. For all the models generated based on our training set, we get very similar results, which means the ARIMA models can be used to generate our predictions.

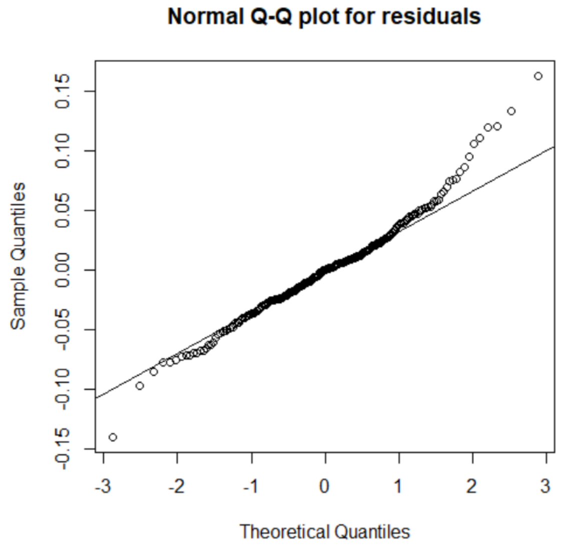


Figure 5. Normal Q-Q plot for residual

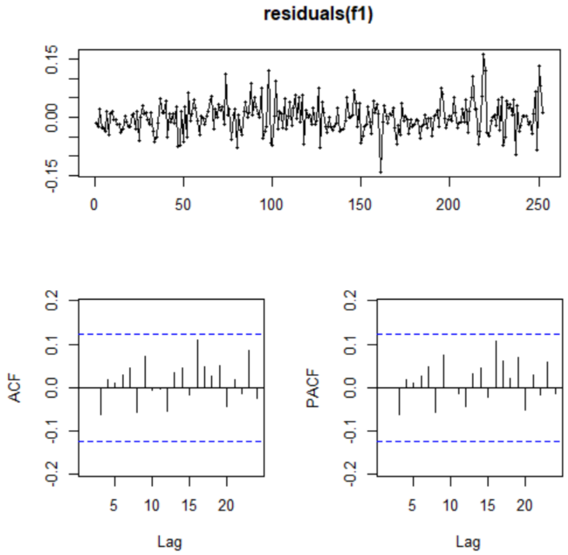
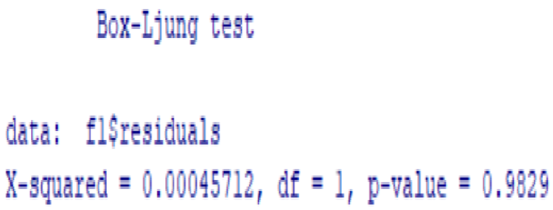


Figure 6 ACF and PACF plot



```
Box-Ljung test

data:  f1$residuals
X-squared = 0.00045712, df = 1, p-value = 0.9829
```

Figure 7 Box-Ljung test

5. CONCLUSION AND PREDICTION

The ARIMA (2,0,2) model is applicable to predict log returns, and has a 61.54% accuracy rate in our test. Therefore, we are using ARIMA (2,0,2) to forecasts for the next 30 trading days based on the log returns in 2020. From figure 8， we can see that the predicted log returns are basically fluctuating around zero. Using the results to predict the PDD stock price changes, we calculate the mean value of the forecasted log returns, which is a positive number. It means the expected log returns is positive in our prediction, which is a signal of increasing trend of PDD stock price. Furthermore, we calculate the proportion of the positive value in our predictions, which is 63.33%. It means there are around 19 days in the next 30 trading days that the stock price goes up.

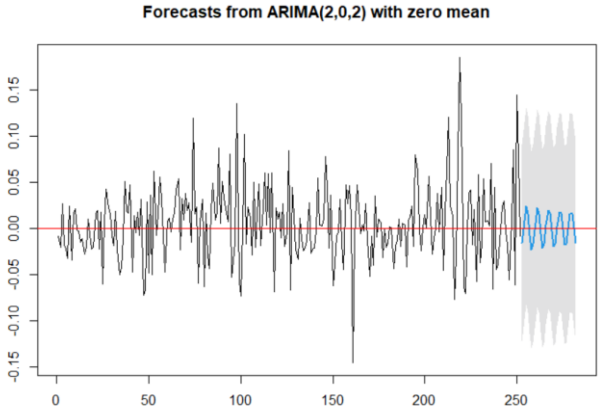


Figure 8 forecasts for the next 30 trading days

According to the model, the PDD stock price tends to inhence, since the forecast log returns are more frequent over 0 than below 0. This model is not effective enough to predict the stock price changes, but it indeed predicts the trend of the stock price in short-term intervals more than half of the times. Thus, our recommended investment strategy for PDD stock is to buy it, especially for short-term investors.

ARIMA model for short-term stock price forecast, has a great reference value, but in the long run, its standard error increases with the number of forecast steps, so once the number of forecast steps is too long, the forecast will become very inaccurate, and the forecast result may have a large deviation. When investors make investment decisions, they can use ARMA model to predict the short-term stock price and make short-term investment plans. In the long run, we need to explore more accurate stock price forecasting model.

REFERENCES

[1] J. H. Stock. Timeseries: Economic Forecasting, 1996:15721-15724.

[2] Liu W,Morley B.Volatility forecasting in the Hang Seng Index using the GARCH approach［J］.Asia-Pacific Financial Markets,2009,16(01).

[3] Paul, J. C., Hoque, M. S., & Rahman, M. M. Selection of Best ARIMA Model for Forecasting Average Daily Share Price Index of Pharmaceutical Companies in Bangladesh: A Case Study on Square

Pharmaceutical Ltd. Global Journal of Management and Business Research, 13(3), 14-26

[4] Wahyudi, S. T. The ARIMA Model for the Indonesia Stock Price. International Journal of Economics & Management, 11(1), 223-236.

[5] Yingchao,Z. &Yingjun.S. An Empirical Study on the analysis and prediction of Shanghai stock index based on ARIMA model. Economic Research Guide, 2019 (11): 131-135

[6] Alzahrani, S. I., Aljamaan, I. A., & Al-Fakih, E. A. Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions. Journal of infection and public health, 13(7), 914-919.

[7] Singh, R. K., Rani, M., Bhagavathula, A. S., Sah, R., Rodriguez-Morales A. J., & Kalita, H., Nanda, C., Sharma, S., Sharma, Y. D., Rabaan, A. A., Rahmani, J., & Kumar, P. Prediction of the COVID-19 pandemic for the top 15 affected countries: Advanced autoregressive integrated moving average (ARIMA) model. JMIR public health and surveillance, 6(2), e19115.