Using SARIMA Model for the Prediction of Stock Price of TESLA, Inc.

Cao Kaiyan, Fu Tianchi, Zhang Junjie
University of Michigan-Shanghai Jiaotong University Joint Institute

Abstract

As the prices of a certain stock can be considered as time series, our project aims to take close look at the stock price of Tesla, Inc. and use Seasonal ARIMA (SARIMA) model to predict its prices at close from Dec.9 2019 to Dec.13 2019. Inside the SARIMA, according to seasonal properties and non-stationary pattern detected in the time plot as well as plot of Autocorrelation Function (ACF) and Partial ACF(PACF), the order of autoregression p, difference d and lagging error term q, together with the seasonal parameters are decided . Based on the errors on validation set, SARIMA$(0,1,0)(2,1,1)_{100}$ is chosen.

Key word: Stock Price, Seasonality, ARIMA, SARIMA, TESLA, Inc.

Introduction

Tesla launched its IPO on June 29, 2010. Trading on the NASDAQ, Tesla offered 13.3 million shares at a price of $17 per share. It raised a total of just over $226 million. This project was designed to explore the behavior of the stock price of Tesla, Inc. with passage time by fitting SARIMA model to it, and to predict its closing price on the five trading days from Dec.9 2019 to Dec.13 2019. The data 's period is from January 1st, 2017 to present and its source is Yahoo Finance.

Mathematic Background

- **ARIMA**

Commonly used in time series, an autoregressive integrated moving average (ARIMA) model is classified as an ARIMA(p,d,q) model, where p is the number of autoregressive terms, d is the number of differences needed for stationarity, and q is the number of lagged forecast errors. It is a generalization of Autoregressive moving average (ARMA) model, in that ARIMA introduces the extra parameter d to first eliminate non-stationarity.

For difference operator:

$$\Delta x_t = x_t - x_{t-1} = x_t - Lx_t = (1-L)x_t$$
$$\Delta^2 x_t = \Delta x_t - \Delta x_{t-1} = (1-L)x_t - (1-L)x_{t-1} = (1-L)^2 x_t$$
$$\Delta^d x_t = (1-L)^d x_t$$

For order d:

$$w_t = \Delta^d x_t = (1-L)^d x_t$$

Now that w_t is stationary, we can use ARMA model for it and the resulting model is $x_t \sim$ ARIMA (p,d,q), which takes the form

$$x_t = \phi_1 w_{t-1} + \phi_2 w_{t-2} \ldots + \phi_p w_{t-p}$$
$$+\delta + u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \cdots + \theta_q u_{t-q}$$

$$\Phi(L)\Delta^d x_t = \delta + \Theta(L) u_t$$

- **SARIMA**

SARIMA or Seasonal ARIMA, is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component. It takes the form SARIMA$(p,d,q)(P,D,Q)_s$, with P, Seasonal autoregressive order; D, Seasonal difference order; Q, Seasonal moving average order and m, the number of time steps for a single seasonal period.

Analysis

We first look at the original data points from 2016.10.27 to 2019.11.27. According to Fig 1, the series seems not to be stationary. We verified our guess with the Augmented Dickey–Fuller (ADF) test, which has the p-value of 0.09279, indicating that we have no strong evidence to reject the null hypothesis that the series is not stationary.

Fig 1. Stock price of Tesla, Inc.

We take the first differentiation of the data and check again if it is stationary with ADF test. This time, with the extremely small p-value, we reject the null hypothesis and accept the stationarity of the differentiated data. Therefore, the order of d we choose should be 1.

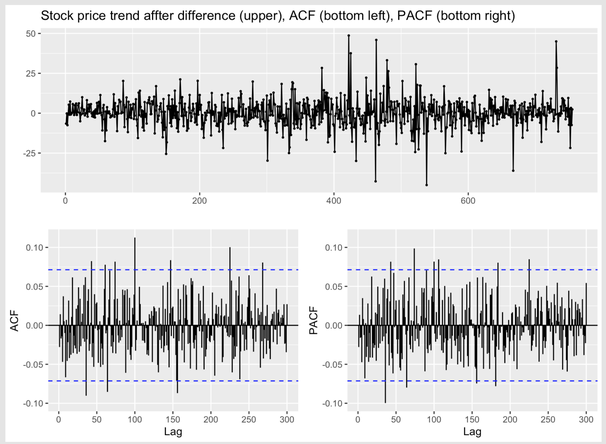

Fig 2. Trend, ACF, PCF of first differentiated data

- **Determination of the non-seasonal parameters:**

From the ACF plot of figure 2, we can see that there is no particular spikes at the start and the data generally shows the characteristic of tailing except at x = 100, which is to big to be the parameter q. Thus, we have to choose our q to be 0.

From the PACF plot of figure 2, similarly , there is no particular spikes at the start and the data also commonly shows the characteristic of tailing. Therefore, we choose the parameter p to be 0 as well.

However, from the previous ACF plot, we observe that there is one obvious spike around x = 100 and another one around x = 230. Therefore, we suspect it has seasonality and the period m could possibly be 100.

- **Determination of the seasonal parameters:**

The F-test on seasonal dummies shows we have no evidence to reject m = 100. Then, we further differ the data with the lag 100. From the ACF plot of figure 3, we can see an exceptionally large spike at x = 100, showing the characteristic of tailing off to zero, indicating the parameter Q should be 1. From the PACF plot of figure 3, we can see an exceptionally large spike at x = 100 and another one at x = 200. Also, in the latter one, the characteristic of tailing off to zero is observed, indicating the parameter P should be 0.

Now that we have obtained our model SARIMA$(0,1,0)(2,1,1)_{100}$, we proceed to its validation.

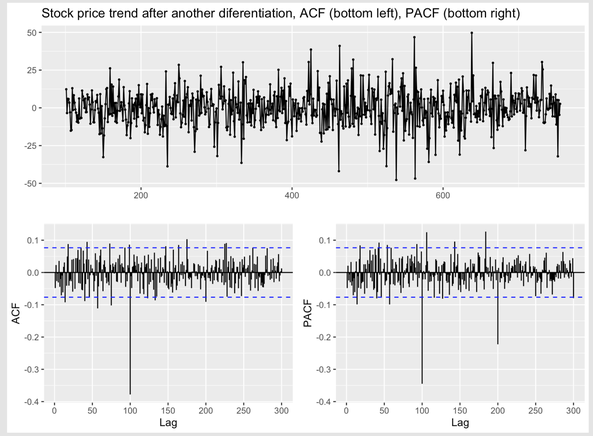

Fig 3. Trend, ACF, PCF of differentiated with lag 100 data

Validation

- **Normality**

According to Fig. 4, the normality of the residuals is also acceptable by the normal Q-Q plot.

- **Autocorrelation**

According to Fig. 4, the residual vs. fitted values plot and the residual vs. previous residual plot shows no strong correlation between the errors. We check the independence of our model's residuals and the squares of them by applying the Ljung-Box test with lag from 1 to 6, which are shown in Table 1. Since the p-values of residuals are all comparatively large, we could not reject the null hypothesis that they are white noise series. However, the p-value of squares of the residuals are all significantly smaller than 0.05. Then, our model may have conditional heteroskedasticity problem and we further turn to the Garch model to solve it, which is stilled being worked on.

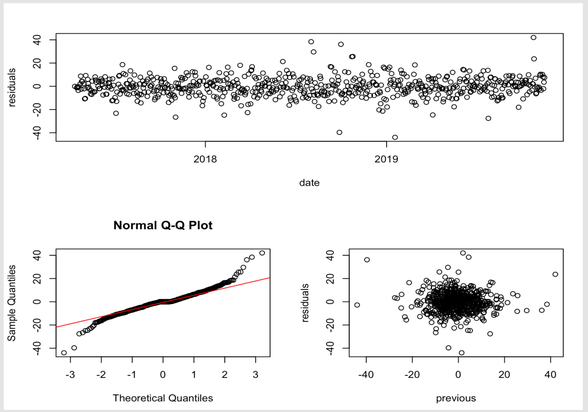

Fig 4. Residual vs. Fitted Value, Q-Q Plot, Residuals vs. Previous Residuals

- **Accuracy on validation set**

Taking the stock price from Nov.19 to Nov.26 (altogether six data points) as validation set, we find out a 4.28 % error between the actual values and our prediction, which is acceptable. Although our estimation does not float in that great extent as the actual values does, the trend suggested by our model is in accordance with the general behavior of the actual values, which is surprisingly good.

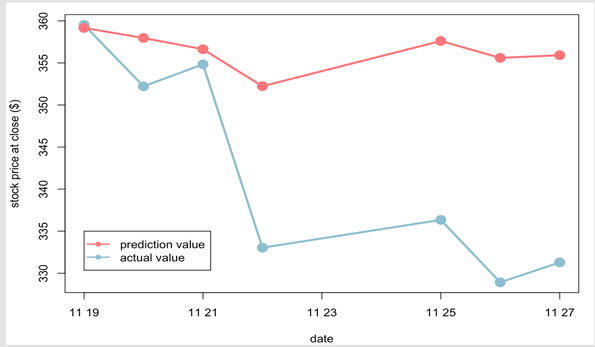

Fig 5. Prediction Value vs. Actual Value

Lag	P-value of Residual	P-value of Square of Residual
1	0.4149	7.869e-05
2	0.5876	0.0002713
3	0.7319	0.000174
4	0.8629	0.0004924
5	0.9152	0.0008573
6	0.9157	0.001724

Table 1. Result of Ljung-Box Test

Results

Since we do not yet have the actual stock prices for the next week, the prediction is for the following five days, which is shown in Table 2.

Date	Predicted Value
11.28	332.0648
11.29	334.3883
12.2	335.6379
12.3	332.8252
12.4	335.8570

Table 2. Prediction Value of the following five days

Conclusion

By first taking the first difference, we successfully obtain a relatively stationary time series. By plotting the ACF and PACF plot of the differentiated series, we have detected seasonality and determined that both p and q, m, Q, P and D. Overall, having checked the residuals and the prediction quality, we claim that SARIMA$(0,1,0)(2,1,1)_{100}$ is a valid model with reasonably good prediction power. We also give the prediction values for the following five days, which could serve as reference.

- **Possible Issue while handling massive data**

The prediction may be influenced by data points from a long time ago, which don't meet current trend and lead to bad prediction.

The introduction of seasonality will make the algorithm extremely computationally expensive, massive data may require a massive amount of computation power.

The effect of conditional heteroskedasticity problem will be enlarged.