JSDS: JOURNAL OF STATISTICS AND DATA SCIENCE
VOLUME 1, No 2, October 2022
e-ISSN: 2828-9986
https://ejournal.unib.ac.id/index.php/jsds/index

Price Prediction Using ARIMA Model of Monthly Closing Price of Bitcoin

Apriliyanus Rakhmadi Pratama[1*]
[1] IAIN Sultan Amai Gorontalo, Gorontalo, Indonesia

* Corresponding Author: Apriliyanus.pratama@iaingorontalo.ac.id

Article Info

Article History:
Received: July 12, 2022
Revised: October 29, 2022
Accepted: October 29, 2022
Available Online: October 31, 2022

Key Words:
ARIMA
Bitcoin
Forecasting
Univariate

Abstract

The rising of bitcoin's user as a digital currency and investments causing an instability and an uncertainty in price movement and increasing the risk of trading, therefore in this study we try to forecast the future value of bitcoin price using ARIMA Models. 2 candidate models are selected by the lowest value of AIC and using the performance indicators ME, RSME, MAE, MPE, and MAPE conclude ARIMA (1,1,0) are the best ARIMA model, then the next 5 months future price forecasted using the best model. While ARIMA (1,1,0) is the best model, the model failed to follow price movement as shown in the forecasted price.

1. INTRODUCTION

Bitcoin is a phenomenal digital currency that shaking economic foundation since its nature are decentralized and powered by its users with no central authority or middleman. Bitcoin's transactions are recorded and monitored by its users using cryptograph network technology or called blockchain. Created in 2008 and used for the first time in 2009 as it launched as open software [1]. Nowadays, bitcoin has gathered many interests, not just people, communities, or companies but countries [2]. Like a double edge sword, bitcoin's price rising and falling as its user increasing follows the market flow of supply and demand causing a instability and uncertainty. Despite its volatility, bitcoin is still in early phase and one promising digital currency in the future. Therefore, in this study, we try to apply time series analysis method to forecast the future price of bitcoin and trying to follow its price movement.

2. METHOD

2.1 Nonstationary

Nonstationary is a condition where time series data had no zero mean, constant variance over time, and constant autocorrelation structure over time. Performing stationary test on time series data formally conducted by unit roots test's Dickey-Fuller [3] with uses the null hypothesis H0: data had unit root / time series data nonstationary against alternative hypothesis H1: data had no unit root / time series data is stationary. Ideally, reject H0 as p-value less than significance level alpha = 0.05 and conclude time series data is stationary.

Handling nonstationary data is carried out through differencing process $(1-B)^d$ with d as differencing order while transforming time series data helps to reduce non constant variances, most commonly transformation method is Box-Cox Transformation [4]. Box-Cox transformation depend on the parameter λ and are defined as

$$w_t = \begin{cases} \log(y_t) & if\ \lambda = 0 \\ \dfrac{(y_t^\lambda - 1)}{\lambda} & otherwise \end{cases}$$

where y_t is original observations and w_t is transformed observations. Detail estimation value of parameter λ shown in Table 1.

2.2 ARIMA Models

Autoregressive Integrated Moving Average (ARIMA) is combination of Autoregressive (AR) model, Moving Average (MA) model, and Integrated (I) as order of differencing process [3]. ARIMA model is suitable to handle a nonstationary time series data that has gone through differencing process [5], generally denoted as ARIMA (p, d, q) while p, d, q represents order of AR, I, and MA respectively.

ARIMA (p,d,q) model have general notation as follows:

$$\varphi(B)Y_t = \phi(B)(1-B)^d Y_t = \theta_0 + \phi\theta(B)a_t \tag{1}$$

where $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ is a stationary autoregressive operator, $\varphi(B)Y_t = \phi(B)(1-B)^d$ is a nonstationary generalized autoregressive operator with d unit roots, $\theta(B)$ is a invertible moving average operator, and $\theta_0 \sim N(0, \sigma_t^2)$.

2.3 Model Identification

General approach model identification [6] Forecasting process of time series data as follows:

1. Plot the data and identify any unusual observations.
2. If necessary, transform the data (using a Box-Cox transformation) to stabilize the variance.
3. If the data are non-stationary, take first differences of the data until the data are stationary.
4. Examine the ACF/PACF: Is an ARIMA (p, d, 0) or ARIMA(0, d, q) model appropriate?
5. Try your chosen model(s) and use the AIC to search for a better model.
6. Check the residuals from your chosen model by plotting the ACF of the residuals and doing a portmanteau test of the residuals. If they do not look like white noise, try a modified model.
7. Once the residuals look like white noise, calculate forecasts.

Identification process of ACF and PACF could follow general pattern criteria as shown in Table 2

Table 1. λ Estimations on Box-Cox Transformation

Estimation λ	Transformation
-1.0	$\frac{1}{Y_t}$
-0.5	$\frac{1}{\sqrt{Y_t}}$
0.0	ln Y_t
0.5	$\sqrt{Y_t}$
1.0	Y_t

Table 2. General Pattern Criteria ACF and PACF

Model	Plot ACF	Plot PACF
AR(*p*)	Dies down	Cut off lag *p*
MA(*q*)	Cut off lag *q*	Dies down
ARMA(*p*,*q*)	Dies down after lag (*q*-*p*)	Dies down after lag (*p*-*q*)

2.4 Parameter Estimation

Maximum Likelihood Estimation method is used as a parameter estimation as the method used all the information in the time series data. Given set time series observations $Y_1, Y_2, \ldots, Y_n$, likelihood function L is defined to be function of the ϕ, θ, μ, and σ_t^2 given the observations.

Without assumption $E(Y_t) = \mu = 0$, equation 1 can written as:

$$\theta_0 = \phi(B)(1-B)^d Y_t - \phi\theta(B)a_t \tag{2}$$

Then, likelihood function L can be written as:

$$L(\phi,\theta,\mu,\sigma_t^2|\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma_t^2)^{\frac{n}{2}}} exp^{-\frac{\sum\theta_0^2}{2\sigma_t^2}} \tag{3}$$

Again, log likelihood function can be written as:

$$\ell(\phi,\theta,\mu,\sigma_t^2|\boldsymbol{\theta}) = \log\Big(L(\phi,\theta,\mu,\sigma_t^2|\boldsymbol{\theta})\Big) = \log\left(\frac{1}{(2\pi\sigma_t^2)^{\frac{n}{2}}} exp^{-\frac{\sum\theta_0^2}{2\sigma_t^2}}\right)$$

$$\ell(\phi,\theta,\mu,\sigma_t^2|\boldsymbol{\theta}) = -exp^{\frac{\sum\theta_0^2}{2\sigma_t^2}} - \frac{n}{2}\log(2\pi\sigma_t^2) \tag{4}$$

2.5 Diagnostic Checking

2.5.1 Autocorrelation Test of The Residuals

The residuals are ideally independently distributed and exhibit no serial correlation. Ljung-Box Test [3] is common method that used to test the residuals independency with null hypothesis H0: The Residuals are independently distributed against alternative hypothesis H1: The residuals are not independently distributed and exhibit seral correlation. Statistics test is expected to fail to reject H0 within significant level and conclude that the residuals are independently distributed.

The test statistics for the Ljung-Box Test as follows:

$$Q = \frac{n(n+2)\sum_{k=1}^{h} p_k^2}{n-k} \tag{5}$$

where, n is number of sample size, p_k is sample autocorrelation at lag k ($k = 1, \dots, h$).

2.5.2 Normality of The Residuals

The residuals also need to be normally distributed. One of many methods are by creating a Histogram. A histogram counts the number of observations between some ranges. To not violate the normality assumption, the histogram should be centered around zero and should show a bell-shaped curve. A high frequency at the extremes of the histogram could indicate that the residuals are not normally distributed.

2.6 Model Selection

Given a collection of models, Akaike's Information Criterion (AIC) estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection. AIC estimates the relative amount of information lost by a given model, the less information a model loses, the higher the quality of that model. In estimating the amount of information lost by a model, AIC deals with the trade-off between the goodness of fit of the model and the simplicity of the model.

General notation of AIC written as[7]:

$$AIC = -2\log(L) + 2(p+q+k+1) \tag{6}$$

where L is the likelihood of the data, and k is the number of predictors in the model, $k = 1$ if $c \neq 0$ and $k = 0$ if $c = 0$.

2.7 Performance Indicators

Measuring the performance of the models carried out by applying few methods. Such as, Mean Error (ME), Root Mean Squared Error (RSME), Mean Absolute Error (MAE), Mean Percentage Error (MPE), and Mean Absolute Percentage Error (MAPE). Ideally, the best accuracy shown as the value of closer to zero. The equations are as follows:

$$ME = \frac{\sum_{t=1}^{n} x_t - f_t}{n} \tag{7}$$

$$RSME = \sqrt{\sum_{t=1}^{n} \frac{(x_t - f_t)^2}{n}} \tag{8}$$

$$MAE = \frac{\sum_{t=1}^{n} |x_t - f_t|}{n} \tag{9}$$

$$MPE = \frac{\sum_{t=1}^{n} \frac{x_t - f_t}{x_t}}{n} \times 100\% \tag{10}$$

$$MAPE = \frac{\sum_{t=1}^{n} \left|\frac{x_t - f_t}{x_t}\right|}{n} \times 100\% \tag{11}$$

where x_t is observed value at time t and f_t is forecasted value at time t, and n is number observations.

3. RESULT AND DISCUSSION

On this study, we used bitcoin price data on monthly basis and was obtained from the website https://coinmarketcap.com, start from 01 January 2014 until 31 May 2022. At the end of October 2021, the bitcoin price closed at a peak price of $61,318.96 while in the last 2 years, the lowest bitcoin price was recorded at the end of March 2020 at $6,438.64.

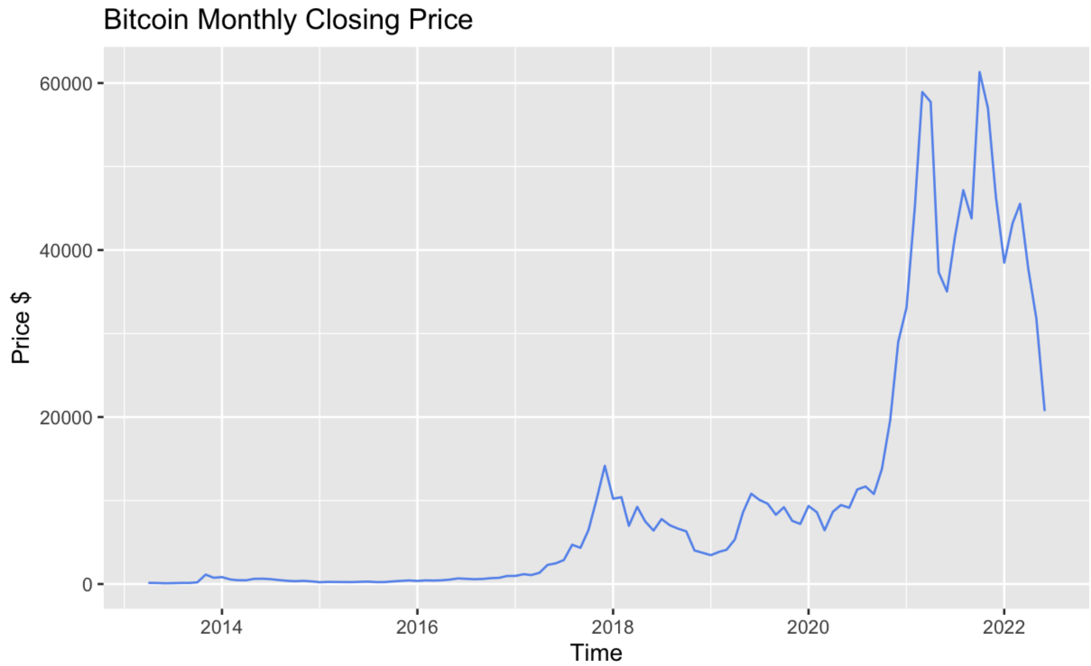

Figure 1. Bitcoin Monthly Price Period January 2021 – May 2022

3.1. Data Preparation

Figure 1 clearly indicates that the data have upward trend and non-constant variance depicting a non-stationary time series. To stabilize the variance, we performed box-cox transformation. The train-test split evaluation [8] methodology is applied to our data. Data Training is used to find the most suitable models, while data testing is used as a comparison to forecast value from the best model. Data Testing starts from January 2022 – 31 May 2022, and data Training starts from January – December 2021.

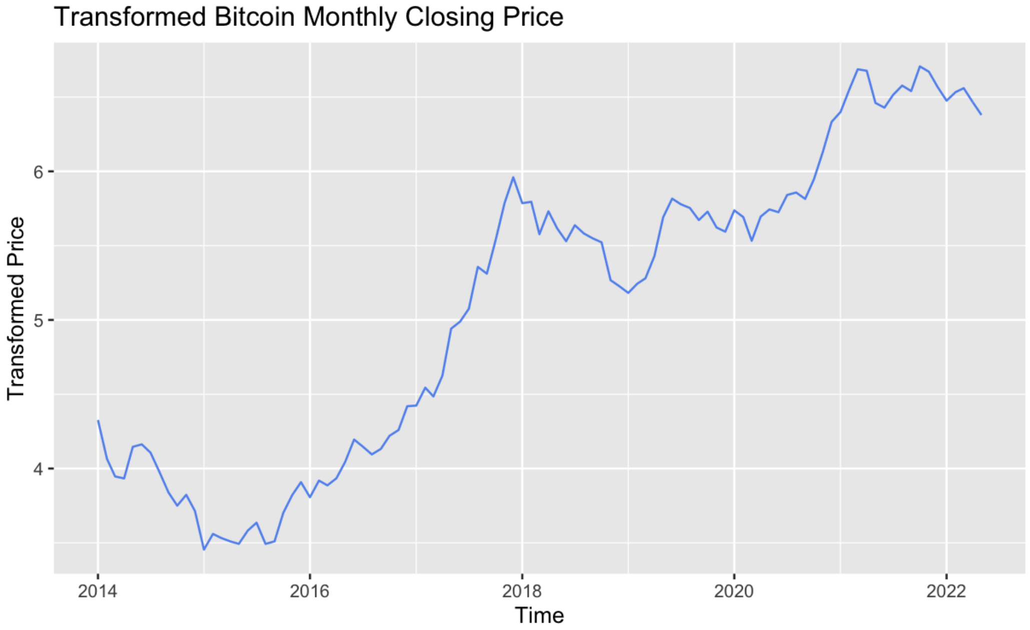

Figure 2. Transformed Bitcoin Monthly Price

3.2. Data Stationarity

Figure 2 clearly shows transformed data had upward trend, meaning data has non-constant mean. Supported by the result of the Augmented Dickey-Fuller (ADF) test conducted on the Training data, which obtained p-value 0.8684. Since p-value > alpha 0.05 indicates that we failed to reject null hypothesis, we could say that our training data is not stationary. Differencing process was applied one time to achieve constant mean time series, as shown in Figure 3. ADF re-test was carried out to obtained p-value = 0.02411. Since p-value lower than alpha 0.05, we could finally be safe to reject null hypothesis.

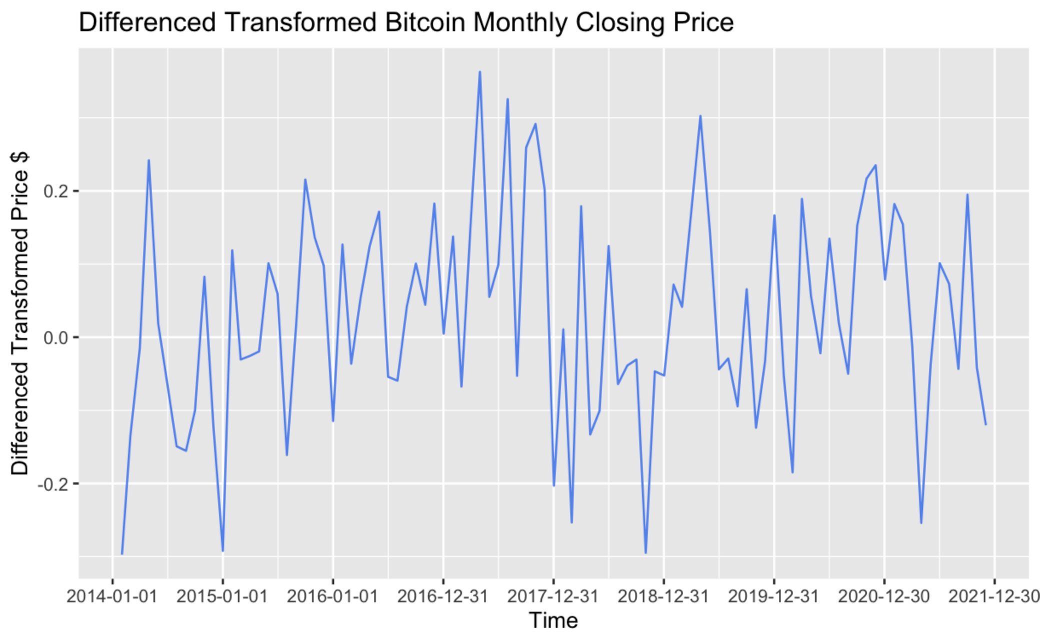

Figure 3. Plot Differenced Data Training

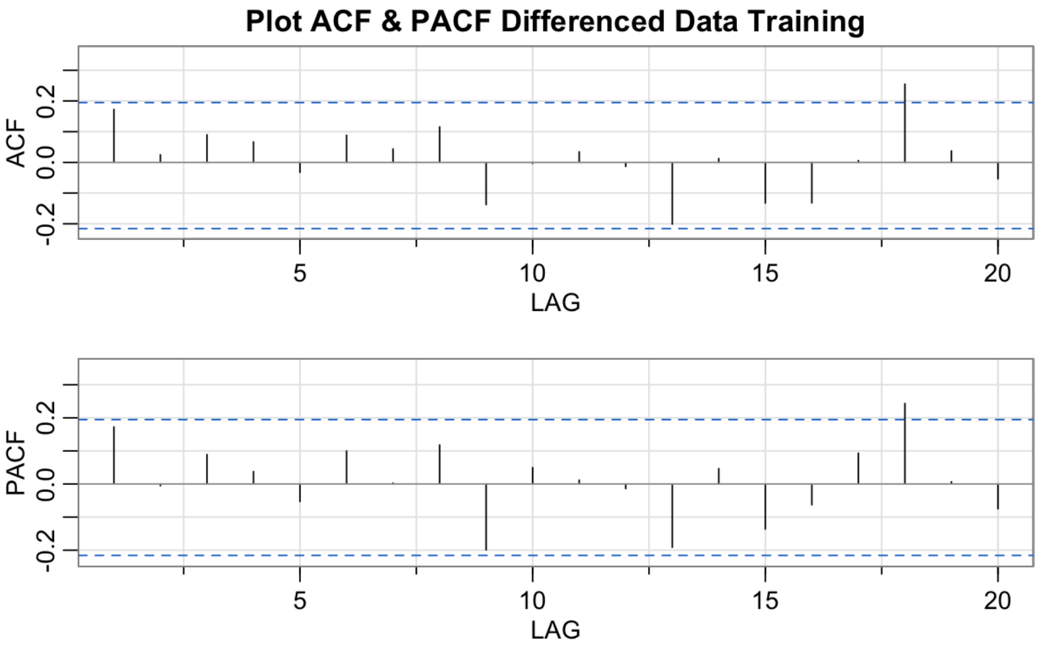

Figure 4. Plot ACF & PACF on Differenced Data Training

3.3. ARIMA model identification

Figure 4 shows Both ACF and PACF plot did not show any pattern resemble geometric decay or sharp drop. Ljung-Box then applied to check whether our differenced data is white-noise as null hypothesis or not white noise as alternative hypothesis. The test results p-value = 0.008118 which is larger than alpha 0.05, that conclude our differenced data is not a white noise.

To identify the model candidates, "auto.arima()" was used, with arguments: stepwise = False, approximation = False, seasonal = False, allowdrift = False, and AIC value is used to measure the best candidates. The result shown in Table 3. The white-noise test was carried out on model candidate's residue using the "Box.test()" syntax. The result p-value were shown in Table 4 indicates that all model candidate's residue were white-noise. Figure 5 display one of residual plot from model candidates ARIMA (1,1,4), the residual plot from the model shows that the variation of residuals stays much the same across the historical data, therefore the residual can be treated as constant, while ACF plot show there is no significant correlations in the residuals series. The histogram suggests that the residual may be normal with a little heavy on the right side. Consequently, forecast from the model candidates will be quite good.

Table 3. Model Candidates with The Lowest AIC Value

Model Candidates	AIC
ARIMA (1,1,0)	-97.28916
ARIMA (1,1,4)	-97.23681
ARIMA (0,1,1)	-97.03295
ARIMA (1,1,1)	-96.11443
ARIMA (4,1,1)	-95.74804
ARIMA (2,1,0)	-95.37009

Table 4. p-value from White-Noise Test on Residue

Model Candidates	p-value
ARIMA (1,1,4)	0.7476709
ARIMA (4,1,1)	0.5848354
ARIMA (1,1,0)	0.5143566
ARIMA (0,1,1)	0.5070393
ARIMA (1,1,1)	0.4954557
ARIMA (2,1,0)	0.420635

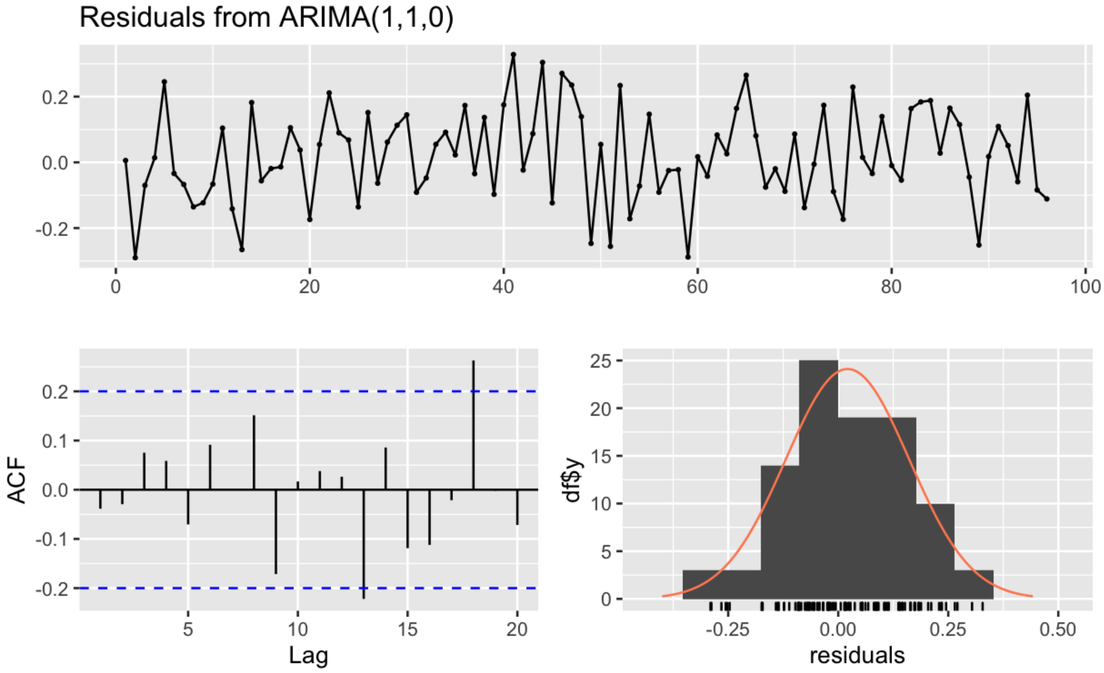

Figure 5. ARIMA (1,1,4) Residual Check

3.4. Forecasting

Top 2 model candidates from Table 3 are used to make prediction for the next 5 month. Figure 6 visualize the forecast against data Test on blue line, we can see that the test value has uptrend in first 3 months and continue to fall in the next 2 months with overall data Test in a downtrend. Both candidates did also have downtrend but with different approach, ARIMA (1,1,0) has slow gradually decreasing, while ARIMA (1,1,4) has a decreasing zig-zag pattern. For better evaluating forecast accuracy, Table 7 show the result of accuracy test between model candidates against data Test. ARIMA (1,1,0) clearly has the best accuracy since all its accuracy test result closest to 0 compared to other candidates as the results are: ME = -4168.446, RMSE = 6332.401, MAE = 5032.100, MPE = -12.31296, MAPE = 14.20949.

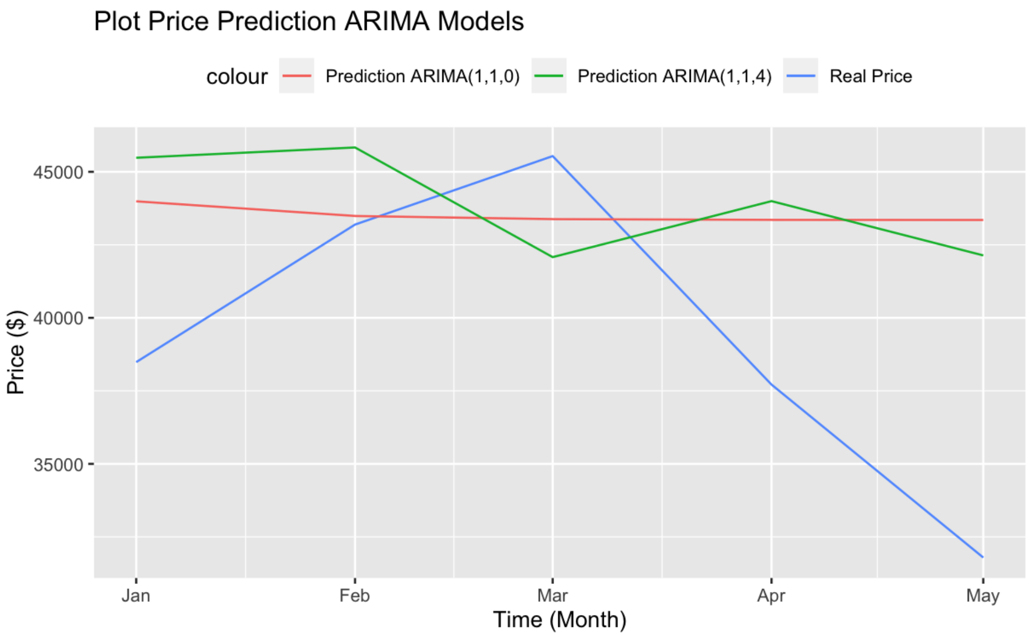

Figure 6. Forecasting of Model ARIMA (1,1,0) and ARIMA (1,1,4)

Table 7. Result of Forecast Accuracy Test

Models	ME	RMSE	MAE	MPE	MAPE
ARIMA (1,1,0)	-4168.446*	6332.401*	5032.100*	-12.31296*	14.20949*
ARIMA (1,1,4)	-4560.867	6548.114	5944.953	-13.17859	16.21795

4. CONCLUSION

Model ARIMA (1,1,0) clearly win the competition between candidate models, with result of accuracy tests closest to 0. While the ARIMA (1,1,0) has the same downtrend with the data Test, in Figure 6 we can see clearly that how bad the model's performance to follow price movement. If we track the problem, time series data has non-constant variance and non-constant mean in latest series compared to early series. Recall Figure 1, it is noticeable prior 2017 the price movements are not as dynamic as from 2017 onwards, therefore this data maybe not accurate representation of how Bitcoin currently behaves.

As a conclusion, ARIMA (1,1,0) is the best model by the result of accuracy test compared to the other model, but clearly cannot be used as main tool to predict the bitcoin price due to its poor performance to predict price movement. Further study is needed to overcome the characteristic of Bitcoin Price time series data.

REFERENCES

[1] A. M. Antonopoulos, *Mastering Bitcoin*, vol. 50, no. 4. 2016. [Online]. Available: https://www.bitcoinbook.info/

[2] O. Lopez and E. Livni, "In Global First, El Salvador Adopts Bitcoin as Currency," *The New York Times*, 2021. https://www.nytimes.com/2021/09/07/world/americas/el-salvador-bitcoin.html (accessed Dec. 09, 2021).

[3] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th ed. Canada: John Wiley & Sons, Inc., Hoboken, New Jersey, 2016.

[4] G. E. Box and D. R. Cox, "An analysis of transformations," *Journal ofthe Royal Statistical Society, Series B*, vol. 26, pp. 211–252, 1964, doi: 10.1080/01621459.1982.10477788.

[5] W. W. S. Wei, *Time Series Analysis: Univariate and Multivariate Methods*, 2nd ed. Greg Tobin, 2006.

[6] R. J. Hyndman and G. Athanasopoulos, "Forecasting: Principles and Practice (2nd ed)," *OTexts: Melbourne*, 2018. https://otexts.com/fpp2/ (accessed Jul. 06, 2022).

[7] H. Akaike, "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974, doi: 10.1109/TAC.1974.1100705.

[8] R. Medar, V. S. Rajpurohit, and B. Rashmi, "Impact of Training and Testing Data Splits on Accuracy of Time Series Forecasting in Machine Learning," *2017 International Conference on Computing, Communication, Control and Automation, ICCUBEA 2017*, 2018, doi: 10.1109/ICCUBEA.2017.8463779.