Bitcoin price forecasting using ARIMA model

Tuan Phu Phan Tha Thanh Le Trung Minh Nguyen Nam Van Hai Phan
Thuan Do Thanh Hoang Khoa Dang Chau
FPT University - HCMC Campus

March 9, 2024

Abstract

Bitcoin, the first decentralized cryptocurrency, utilizes a transaction log with incentives for honest participation. As a virtual currency with the potential to disrupt traditional payment and monetary systems, Bitcoin is of significant interest to economists. The Bitcoin trading market experiences considerable volatility. This paper explores the use of the ARIMA model to forecast future Bitcoin prices. The ARIMA model is a statistical tool that leverages time series data to analyze trends and predict future values within a series.

1 Introduction

Bitcoin, introduced in 2008 as an alternative to traditional currency systems, has evolved into a widely traded speculative asset. Despite its potential to revolutionize financial systems, Bitcoin's primary use is as an investment vehicle, similar to tech stocks in the past. Its growing popularity and significant market capitalization underscore its disruptive potential and the need for further analysis. [3]. We noticed that the price of Bitcoin is very volatile and it is hard to predict the price of Bitcoin. In our work, we are using ARIMA model to forecast the price of Bitcoin. ARIMA stands for Auto Regressive Integrated Moving Average, and it is a class of stochastic processes used to analyze time series [1].

The remainder of the paper is organized as follows. In Section 2, we provide a brief overview of the mathematics used in the ARIMA model. In Section 3, we describe the dataset used in our analysis. In Section 4, we outline our methodology for forecasting Bitcoin prices using the ARIMA model. In Section 5, we present our numerical results. In Section 6, we present our conclusions and discuss the implications of our findings.

2 Background

2.1 Auto regression

An autoregressive (AR) model is a type of statistical model used for understanding and predicting future values in a time series based on its own past values. It is a representation of a type of random process; as such, it is used to describe certain time-varying processes in nature, economics, etc. The concept is similar to regression analysis, where the value of the dependent variable is assumed to be a linear combination of past values (lags) of itself.

Autoregressive models are foundational to time series forecasting. These models are widely used in various fields such as economics, finance, natural sciences, and engineering. The basic premise of an AR model is that the current observation is a sum of past observations with some stochastic noise. The order of an AR model, often denoted by 'p', indicates the number of lagged observations in the model.

For instance, an AR model of order 1, AR(1), would predict the current value of the series based on the immediately preceding value. An AR(2) model would use the two preceding values, and so on. [5] The general form of an AR(p) model is given by:

$$x_t = c + \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + ... + \varphi_p x_{t-p} + \varepsilon_t$$

where x_t is the time series, c is a constant, $\varphi_1, \varphi_2, ..., \varphi_p$ are the parameters of the model, and ε_t is white noise.

2.2 Intergrated

Intergrated(I) is the number of differences we consider between the current value and past value in order to make the time series stationary. Stationarity means that the statistical properties of the series, such as mean and variance, remain constant over time. To form a stationary series, the simplest method is to take the difference. Some financial series also convert to logarithms or yields. The order of difference to form a stationary series is also called the order of integration. The d-order difference process of the series I(d) is performed as follows:

$$\mathrm{I}(d) = \Delta^d(x_t) = \underbrace{\Delta(\Delta(\ldots\Delta(x_t)))}_{\text{d times}}$$

Normally the chain stops after the co-integration process I(0) or I(1). Very few series we have to take to the second difference. In some cases we will need to transform logarithms or square roots to create a stationary series.

2.3 Moving average

As an alternative to the autoregressive representation in which the x_t on the left-hand side of the equation are assumed to be combined linearly, the moving average model of order q, abbreviated as MA(q), assumes the white noise wt on the right-hand side of the defining equation are combined linearly to form the observed data. [5] The moving average model of order q,or MA(q) model, is defined by:

$$x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + ... + \theta_q w_{t-q}$$

Where $w_t \sim wn(0, \sigma_w^2)$, and $\theta_1, \theta_2, ..., \theta_q$ are the parameters of the model.

3 Dataset

For getting data of Bitcoin we decide to craw it from website Coinmarketcap which provides information and data such as prices, trade volumes, market capitalization on cryptocurrencies. Firstly, we use api from this platform to build our library for getting real-time pricing information of Bitcoin crypto, so we can obtain the latest data for creating model thanks for commonly used library like requests.

Then we extract the data (in JSON format), change the Dataframe format, which finally can use for analyzation before construct ARIMA model. For analysis and prediction, we utilized a dataset consisting of 163 price points for the average monthly Bitcoin price from August 2010 to February 2024. Figure 1 shows the visualization of the series.

To ensure the accuracy and reliability of our model, we split the dataset into two parts: 90% for the training set and 10% for the testing set. This division allows us to train the model on a significant portion of the data and evaluate its performance on a separate, unseen portion.

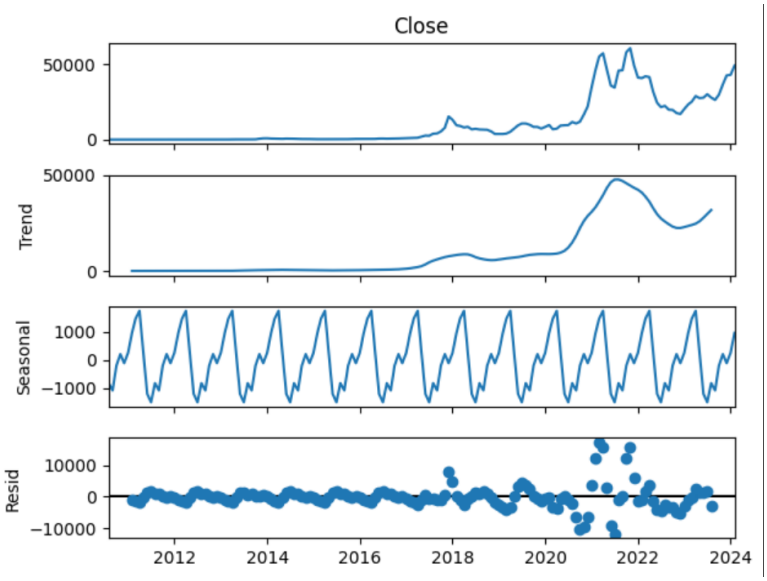

Figure 1: Bitcoin Close price visualization

4 Methodology

4.1 ARIMA (Model identification)

A class of models is constructed based on specific hypotheses. In this section, we will select the general ARIMA formula to model price data. This choice was made after carefully examining key characteristics of the Bitcoin price series. If y_t denotes the Bitcoin price at time t, the proposed general ARIMA formula is as follows:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + ... + \theta_q \epsilon_{t-q} + \varepsilon_t \quad (1)$$

where y_t **is the price at time t**, ϕs are the parameters of the model AR, θs are the parameters of the model AM, and ε_t is the error term. [4]

4.2 Check stationarity

One of the necessary conditions for regression in time series models is that the series must be stationary. Here, we utilize the Augmented Dickey-Fuller (ADF) test through the statsmodels package in Python. If the p-value is less than 0.05, the series exhibits stationarity. We show the ADF result of the series before and after intergrated by figure 2 and figure 3 respectively.

4.3 Choosing parameters ARIMA(p,d,q)

Initially, it may be necessary to transform the original price data to ensure the series is stable, or stationary (with constant mean and variance). In this step, differentiation is applied to the data to make the data stationary. Typically, we would apply first or second-order differences. That's how to choose parameter d.
Subsequently, we must ascertain the suitable degrees for the AR and MA components of the models. To determine the optimal degrees, we will utilize an autocorrelation plot (ACF) and a partial autocorrelation plot (PACF). [4]

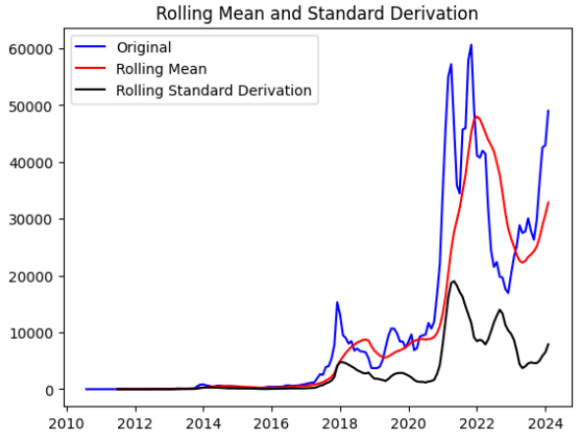

Figure 2: ADF result of the original series. p-value = 0.836223

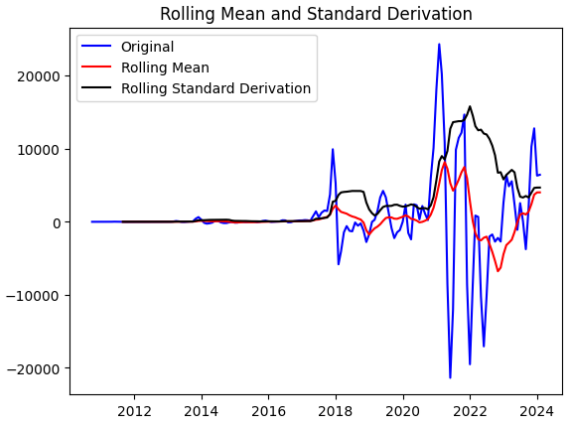

Figure 3: ADF result of the series after intergrated (d=2). p-value = 0.005389

Determine the degree of model AR(p): The PACF plot will be utilized to determine the degree p of the AR process, based on the peaks observed at consecutive lag levels. If the highest lag exhibits a peak outside the 5% confidence interval, the degree value for AR will be determined by this lag.
Determine the degree of model MA(q): The ACF plot will be used to determine the degree q of the MA process. The process is similar to determining AR, relying on the largest lag outside the 5% confidence interval. We visualize the PACF and ACF plot of the time series before and after intergrated in Figure 4.

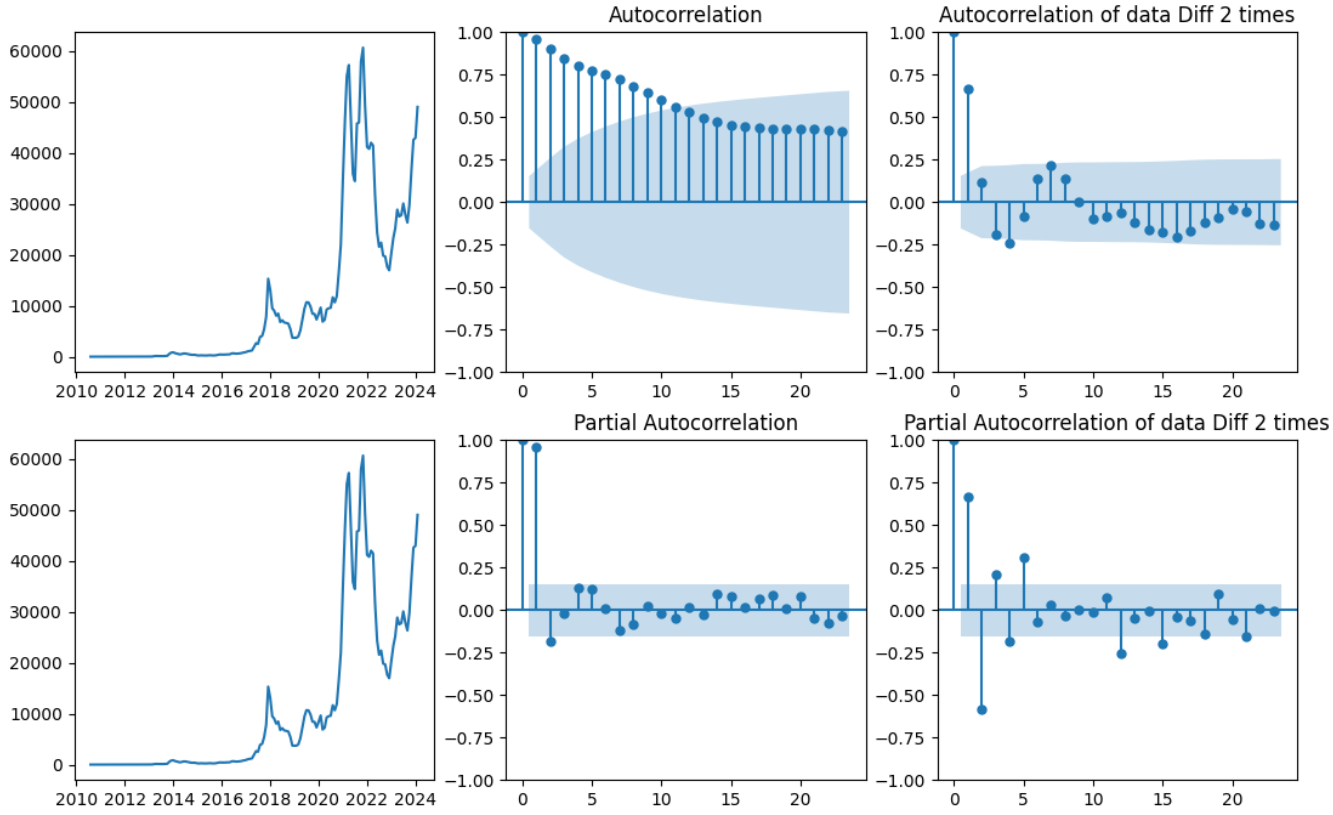

Figure 4: PACF and ACF visualization

4.4 Build ARIMA model

From the steps of determining p, d, q, we will obtain an ARIMA(p, d, q) process and perform regression on the training dataset. After deriving the prediction model, we will apply it to the test dataset and cross-check to verify if the predicted values match the actual values. Table 1 presents the top 5 ARIMA(p, d, q) models with the smallest Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

ARIMA(p, d, q)	RMSE	MAE
(5, 2, 5)	9954.92	8036.04
(4, 2, 5)	10806.41	8345.89
(5, 1, 2)	11123.29	8451.14
(0, 2, 1)	11445.91	8643.19
(1, 2, 5)	11470.02	8612.56

Table 1: RMSE and MAE values for different models

After processing, we choose ARIMA(5, 2, 5) as the model to predict the time series.

4.5 Evaluation

When using the ARIMA model, we often refer to two typical error parameters to assess its accuracy: MAE and RMSE. MAE measures the average difference between actual and predicted values within a week. It is calculated by summing the absolute differences between actual and predicted values for all observations in a week and then dividing by the number of observations. A value closer to 0 indicates more accurate forecasting.

RMSE is a commonly used metric to measure the accuracy of a regression model. It represents the square root of the average of the squared differences between the predicted values and the actual values. We show the MAE and RMSE formular in 2 and 3 respectively.

Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2)$$

Root Mean Square Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

Where:

- y_i is observed value
- $\widehat{y_i}$ is predicted value
- n is the number of observations

5 Numerical results

We have posted the dataset and Python code here.

5.1 Case Studies and forecasting

After resampling the time series dataset by month, we proceed to normalize the data for training ARIMA. We difference the data twice and conduct the Augmented Dickey-Fuller (ADF) test. With a p-value of 0.005, the second-differenced time series exhibits stationarity.

In addition to the parameter selection method utilizing plots of ACF and PACF, we employ loops to consider all parameter combinations and select the optimal model. While this approach may consume more time, it ensures optimal performance in terms of error function and accuracy when determining the model parameters. The time series prediction on the test set is depicted in Figure 5.

We forecast the future closing price of Bitcoin using the ARIMA model for the next 12 months in the figure 6.

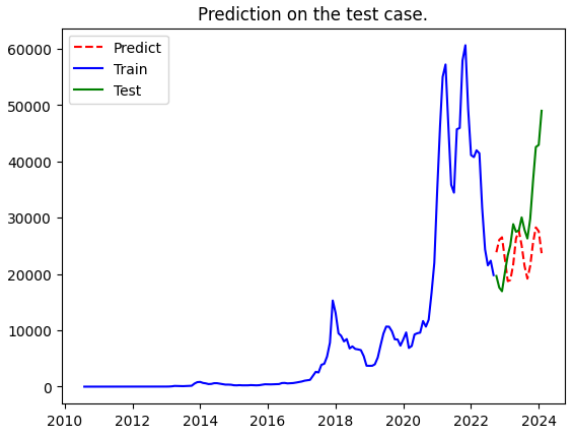

Figure 5: Prediction on the test case

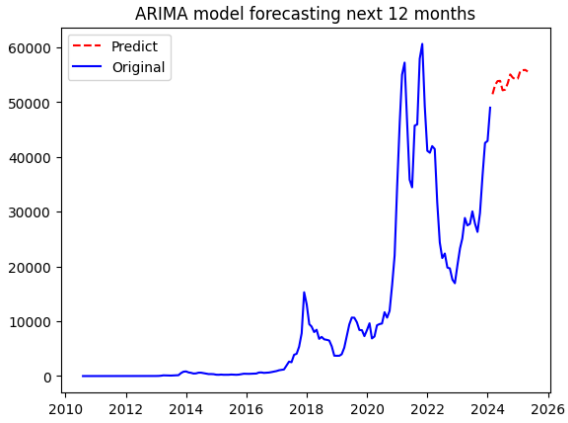

Figure 6: Prediction for the next 12 months

5.2 Discussion

During the process of analyzing the Bitcoin time series and making predictions using the ARIMA model, we have observed certain limitations. These include the model's tendency to provide short-term forecasts, typically ranging from 10 to 15 future data points. Furthermore, when dealing with seasonal Bitcoin closing price datasets, ARIMA often falls short in delivering optimal results. Research suggests that the SARIMA model exhibits greater efficacy in forecasting seasonal time series compared to ARIMA [2].

6 Conclusion

This study investigates the applicability of the Autoregressive Integrated Moving Average (ARIMA) model for evaluating price patterns and forecasting future trends in the Bitcoin market. By leveraging time series data, the ARIMA model offers valuable insights into Bitcoin market behavior. This research contributes to the growing body of knowledge on quantitative tools for navigating the complexities of cryptocurrency markets. As the cryptocurrency ecosystem continues to evolve, the adoption of adaptable models like ARIMA will be crucial for market participants and economic researchers alike.

References

[1] Javier Contreras et al. "ARIMA models to predict next-day electricity prices". In: *IEEE transactions on power systems* 18.3 (2003), pp. 1014–1020.

[2] Ashutosh Kumar Dubey et al. "Study and analysis of SARIMA and LSTM in forecasting time series data". In: *Sustainable Energy Technologies and Assessments* 47 (2021), p. 101474. ISSN: 2213-1388. DOI: https://doi.org/10.1016/j.seta.2021.101474.

[3] Satoshi Nakamoto. "Bitcoin: A peer-to-peer electronic cash system". In: *Decentralized business review* (2008).

[4] Th Ánh Hng Phm et al. "ng dng mô hınh chui thi gian d báo bnh truyn nhim ti tnh Hng Yên= Applying Time Series Models for Predicting The Rate of Influenza Flu in Hung Yen Province". In: (2022).

[5] Robert H Shumway et al. "ARIMA models". In: *Time series analysis and its applications: with R examples* (2017), pp. 75–163.