SOCIOECONOMIC ANALYTICS

Predicting Bitcoin Cryptocurrency Price Behavior based on ARIMA and NNAR modelling

Patrícia Virgínia de Santana Lima [1] *
David Venâncio da Cruz [2]
Albaro Ramon Paiva Sanz [3]

[1 2 3] Departamento de Estatística, State University of Paraíba, Campina Grande, Brazil
Emails: [1] patricia.lima@aluno.uepb.edu.br; [2] davidvenacio4@gmail.com; [3] albaro.sanz@servidor.uepb.edu.br
* Corresponding author

How to cite this paper: Lima, P. V. de Santana, da Cruz, D. V., Sanz, A. R. P. (2024). Predicting Bitcoin Cryptocurrency Price Behavior based on ARIMA and NNAR modelling. *Socioeconomic Analytics,* 2(1), 121-129. https://doi.org/10.51359/2965-4661.2024.265073

RESEARCH ARTICLE

Socioeconomic Analytics
https://periodicos.ufpe.br/revistas/SECAN/
ISSN Online: 2965-4661

Submitted on: 11.11.2024.
Accepted on: 15.12.2024.
Published on: 27.12.2024.

Abstract

In this work, we develop specific models to predict the behavior of the Bitcoin cryptocurrency using a public database (Yahoo! Finance) to track price trends. The methodologies used are ARIMA and NNAR, and the validation of the models is carried out based on the daily closing values of assets. Both models fail to differ significantly. However, the adjusted model NNAR (2.2) fits slightly better with the original data series, presenting an MPE (Mean Percentage Error) of -0.102. Some prospects on the seasonality of data are discussed.

Keywords

prediction, Time Series Analysis, Bitcoin, forecast. ARIMA, NNAR.

1. Introduction

The Bitcoin cryptocurrency that emerged in 2008, created by an anonymous named Satoshi Nakamoto (Ulrich, 2004), with the publication of the following work “Bitcoin: A Peer-to-Peer Electronic Cash System” with the main objective of decentralizing the market generating independence from government interversion in transactions. In the face of a vast growth of investors, the search for tools to analyze this asset is increasing. As a result, many investors make decisions through indicators such as: Fibonacci, candlestick pattern and forecast models, as shown in this work (NAKAMOTO, 2009).

Since its creation, several works have been developed within data science in an attempt to predict the behavior of cryptocurrencies. In 2018, Velankar, Valecha and Maji proposed methods for Bitcoin prediction based on data from CoinMarktCap, after filtering the data and applying Bayesian Regression and Generalized Linear Models. Regarding recurrent neural networks, MCNally (2016) makes a comparison between RNN (Optimized Bayesian Neural Network) and LSTM (Long Short Term Memory) in which LSTM outperforms RNN in relation to fit measures, although without significant difference.

In this specific area of cryptocurrency forecasting, we still have, Kim *et al*. (2016) who proposed prediction models based on the comments and responses published in online communities related to cryptocurrencies, specifically on the Bitcoin, Ethereum and Ripple forums, with the participants comments influencing the fluctuations of cryptocurrency prices and trading volume. This is done through an algorithm called VADER, the comments of the participants are filtered, a statistical analysis is carried out by the sentiment they transmit (HUTTO, GILBERTO, 2014), excluding comments considered extreme positive or negative, the result of this work was positive for the idea worked, that the opinion of users can be used to predict fluctuations in this market, however, the authors recommend more qualitative criteria for forecasting.

The most recent work using neural networks for the prediction of Ether is Duarte and Lima (2019), who compare two neural networks Multilayer perceptrons (MLP) and Bidirectional Long Short Term Memory Units (bi-LSTM), and resulted in bi-LSTM with a slightly better result than MLP.

With the time series models especially ARIMA, Azari 2009, proposed models with several parameters for a series with a period of 3 years and daily series, which concluded that for long series the models could not “capture” all the oscillations, because the cryptocurrency market is still relatively new, that is, “emerging”, which makes speculations sensitive, news, political, economic and military crises. Which makes it a high volatility market. According to NuBank, between 2009 and 2011, bitcoin gained an impressive 30,000 percent in value. So the lower the lag, the better the price prediction.

According to Silva (2016), the biggest problem with bitcoin is volatility. As a way to account for price fluctuations, it has a profound impact on investment strategies, that is, the series usually present extreme observations and asymmetric and volatile reactions compared to the past.

The use of this market with a view to the possibility of investment and ultimately financial return, however, it must be taken into account that these types of operations are of great risk, warn Radityo, Munajat and Budi (2017), that due to the high volatility, forecasting tools are necessary to assist possible investors in decision-making in the exchange of cryptocurrencies.

The present work aims to apply the ARIMA and NNAR models on the cryptocurrency data, verifying the residuals generated by the adjusted models in order to obtain Bitcoin price predictions, analyzing the errors and analyzing the errors of the market.

2. Material and methods

The analyzed database was made available by Yahoo Finance containing 8430 rows and 6 columns, with daily values with the following attributes: Lowest value reached, highest value reached, opening (initial value), closing (last value reached), adjusted and trading volume of the Bitcoin cryptocurrency, between 01/01/2021 and 11/05/2024.

2.1. Time Series

A Time Series can be defined as a sequence of data observed at regular time intervals, i.e., a random variable ordered in time. Thus, we have Y_t where $(Y_t)_a^z = \{Y_a, \dots, Y_z\}$, where Y_t is a random variable observed once regular (BOX; JENKINS, 2015).

2.2. Integrated autoregressive moving average model

According to Cruz (2010), the ARIMA model is a generalization of the autoregressive moving average (ARMA) model. The ARIMA representation (p, d, q) refers respectively to the autoregression, integration and moving average orders: p is the autoregressive operator, d number of differences, q is the number of terms of the moving average. Soon after the adjusted ARIMA model, it is verified that the residuals of the model are independent, called white noise.

The model given by equation 2 is called the Autoregressive model of order AR(p).

$$\boldsymbol{Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_p Z_{t-p} + a_t} \tag{2}$$

in which the terms $\boldsymbol{Z_{t-1}, Z_{t-2} + \cdots + Z_{t-p}}$ are independent of $\boldsymbol{a_t}, \boldsymbol{\phi_1}$ is the parameter (weight) that describes how $\boldsymbol{Z_t}$ relates to the value for $\boldsymbol{Z_{t-1}}$ (data series) and $\boldsymbol{a_t}$ is the random error.

The methodology implemented by Box and Jenkins and Reinsel (2015) for the construction of ARIMA models is composed of three steps: Identification; Estimation and Verification. Therefore, step 1 is more relevant because it determines the values of P, a, and b. This determination is made by means of the estimated autocorrelations and partial autocorrelations, which are expected to adequately represent the true theoretical quantities that are unknown.

2.3 Neural Network Autoregression (NNAR).

The Autoregressive Neural Network (NNAR), introduced in 2018 by Hyndman and Athanasopoulos, is a type of time series prediction model based on artificial neural networks (ANNs). One of the main advantages of NNAR models is their ability to capture nonlinear relationships between data, which makes them more flexible and powerful than linear models such as ARIMA [Maleki *et al*., 2018]. This model was created for a feedforward neural network with a hidden layer, denoted by NNAR (p,k), where p indicates the number of neurons in the hidden layers and k indicates the lagged input.

The Neural Network Autoregression model is defined with a multi-layer network involving a linear combination as the activation function, the output points are inputs to the next. Also according to Thoplan, (2014), the linear function is given by the following expression:

$$\boldsymbol{z_j = b_j + \sum_{i=1} w_{i,j} x_i} \tag{3}$$

The entrances to the hidden neuron $\boldsymbol{j}$ are linearly combinations, in which the weights $\boldsymbol{w}$ and the parameters $\boldsymbol{b}$ are defined from the data. It should be noted that in the hidden layers, the nonlinear function is used as a sigmoid, the activation function, defined by the following equation (4):

$$\boldsymbol{f(z) = \frac{1}{1+e^{-z}}} \tag{4}$$

2.4. Accuracy

To measure the performance of the models, the following were used: The MAE (Mean Absolute Error) is calculated from the average of the absolute errors, the modulus of each error is used, avoiding underestimation, so the value is less affected by the extreme points (outliers). Each error is interpreted as the difference between $\boldsymbol{y}$ and $\boldsymbol{\hat{y}}$, so we have:

$$\boldsymbol{MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|} \tag{5}$$

And the MAPE (The average absolute percentage error) is the average of the absolute errors during a given period multiplied by 100% so that the results obtained are in the form of percentages (Junita, *et al.*, 2024), defined as:

$$\boldsymbol{MAPE = \frac{1}{n}\sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{y_i} \times 100} \tag{6}$$

2.5. Model residues

To validate the adjusted model, the residuals must meet the following conditions:

1. Randomness - Graph of the residuals *versus* the order of data collection; and

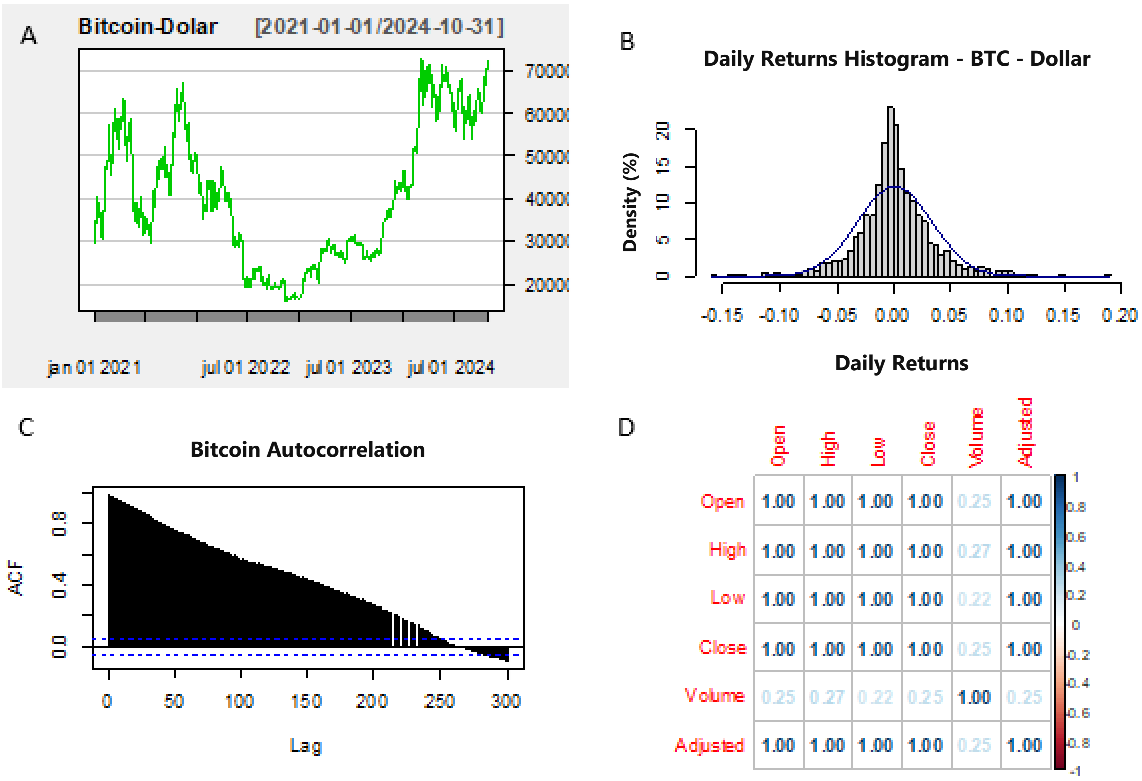
2. Autocorrelation function. According to Morettin and Toloi (2018), to verify whether a series is white noise, the graph of the autocorrelation function is constructed with its respective confidence interval. If the residual correlations are within the range, the process is stated to be white noise.

3. Normality - For the analysis of the residuals, the assumption of normality must be observed, and it is possible to observe it through the histogram and normality test.

3. Results and discussions

This section presents the results obtained, through an analysis of the time series, between January 2020 and October 2024, the graphs presented below, it is possible to verify that:

Figure 1: Graph (A) Time series between January 2020 and October 2024; Chart (B) Daily returns (oscillation) of Bitcoin; Chart (C) Autocorrelation of prices; Chart (D) Price and volume correlation matrix.



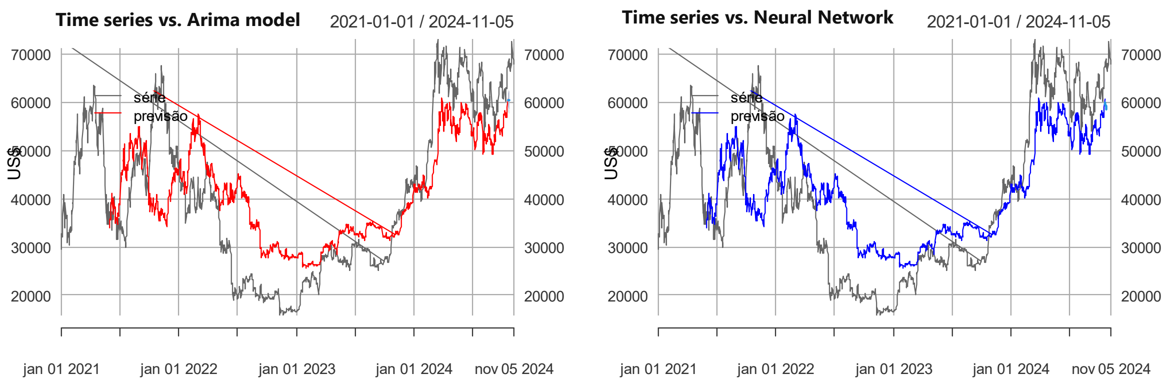
It can be seen in Chart (A) that the series referring to the closing values of Bitcoin between July 2022 and March 2023, the asset suffered a depreciation in price reaching the lowest mark in the period of 15787.28 dollars, currently the asset is in an uptrend. Graph B, it is shown that the price fluctuations (returns) of the Bitcoin cryptocurrency in their daily periodicity, after being quantified on logarithmic bases, have a distribution that resembles the normal distribution. For the autocorrelation graph (C) Values close to 1 or -1 indicate a strong influence of the respective past record on the current record, while a value close to 0 indicates a weak influence.

For the time series analyzed, that is, in the behavior of the prices of the Bitcoin cryptocurrency, it is possible to verify that the prices measured daily demonstrate positive autocorrelation approximately the 270th day before. From this, prices indicate a negative autocorrelation. Regarding the correlations between prices (Chart D), only the volume of currencies traded in a given time does not have a significant correlation with the variation in the price of the respective asset.

3.1. The adjusted models

For the purpose of modeling the behavior of the daily closing price of Bitcoin, two models were used: An Arima-type model (AutoRegressive Integrated Moving Average model) that combines differentiation methods and the autoregression and moving average models. The second model was the Neural Network Auto Regressive (NNAR) is the Artificial Neural Network (ANN) that Hyndman and Athanasopoulos (2018), this model has the advantage of being able to efficiently predict when there is a nonlinear relationship between the months and so we don't have to worry about making transformations, like Box-Cox in taking the seasonal difference. The models fitted below (Figure 2).

Figure 2: Time series of the asset's closing prices and the ARIMA (1,1,0) and NNAR (2,2) models.



The model used for the series is Arima (1,1,0), a model composed of an autoregressive term and only one difference needed to make the series stationary, the second NNAR model (2,2) which considers 2 lag's in the input layer, 2 seasonal lag.

3.2. Metrics

Table 1 shows the metrics for the validation of the adjusted models: mean absolute error (MAE) and mean absolute percentage error (MAPE) for the models:

Table 1: Values obtained during the validation of the prediction models.

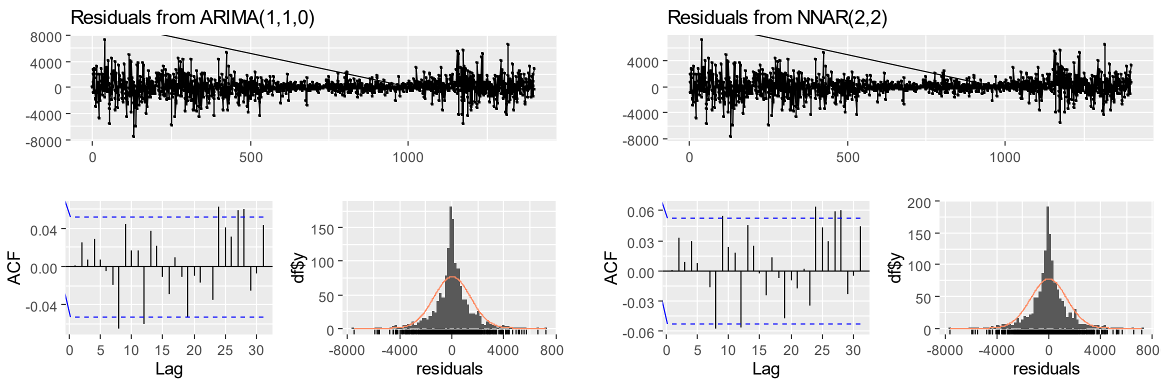
Models	Variance	MAE	MAPE
Arima (1,1,0)	1947146	916,6333	0,9921052
NNAR (2,2)	1927664	919,3747	0,9950723

The models present similar variances, while the MAE Arima model (1,1,0) presents better performance, that is, it obtains a lower average among the absolute errors. Regarding the MAPE, the second prediction model performs better when the assumed value is low, for this there are four classification levels: Highly accurate (<10%); accurate (10% - 20%); regular (20% -50%) and inaccurate (>50%), respectively (Junita, T. P., & Kartikasari, M. D., 2024). Therefore, both models are highly accurate, with MAPE around 0.99%, with a small discrepancy between the observed and predicted values.

3.3. Residual analysis

The next step was to verify the quality of the fit of the models, for this we will analyze the residuals of the models, looking at the errors generated in the estimation process, presented below in Figure 3.

Figure 3: Residual analysis for asset series adjusted models.



It is observed that the residuals obtained in both models are characterized as white noise, the sequences of errors (residue) oscillate around zero, that is, the average of the residuals is zero without bias, the variance of the residues produced by the models is constant, the dispersion of the residues does not change over time, which means that the residues are independent and identically distributed, assuming a covariance equal to zero.

The autocorrelation functions for both models show that the autocorrelations for the prediction errors practically do not exceed the significance thresholds for lags 1-30, indicating that there is not much evidence of non-zero autocorrelations in the lags 1-30. Therefore, we can conclude that the residuals are not self-correlated.

In addition, prediction errors have constant variance over time and are typically distributed with zero mean, seen in the Residuals histograms for both models. Therefore, the ARIMA (1,1,0) and NNAR (2,2) models are valid.

4. Conclusion

Based on the results obtained, they show that the ARIMA and NNAR models, even presenting satisfactory indexes for validation. The NNAR Neural network (2,2), even presenting more complex structures, could not have greater precision than the ARIMA model (1,1,0), so to estimate it is recommended to use this simpler model, since it can represent the dynamics of the system with the same precision.

It was observed, analyzing the series and the literature, that it is necessary to obtain models that capture seasonality, as the price behavior in relation to this asset presents periods of high and low cycles, influenced by external and internal factors, known in the financial market as "crypto winter and summer". Therefore, for future studies, models of multiplicative methods are recommended, since there are seasonal variations that change proportionally to the level of the series.

Acknowledgements

Our thanks go first to the event "Wanda 2024" and its organizers, and the State University of Paraíba (*Universidade Estadual da Paraíba*), for the availability granted to students.

Conflict of Interest Declaration

The authors have no conflicts of interest to declare. All co-authors agree with the content of the manuscript and there is no financial interest in reporting.

References

1. Azari, A. (2019). Bitcoin price prediction: An ARIMA approach. *arXiv preprint arXiv:1904.05315*.
2. Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.
3. Cruz, D. V., Cunha Filho, M., & Falcao, A. P. S. T. (2014). Controle Estatístico de processos aplicado ao monitoramento do pH e turbidez das águas no abastecimento de Campina Grande-PB. *Rev. Bras. Biom*, *32*(4), 459-477.
4. Duarte, F. L. C., & de Almeida Lima, N. C. (2019). Redes Neurais Aplicadas Na Predição do Preço do Ether: MLP vs bi-LSTM. *Anais do Encontro de Computação do Oeste Potiguar ECOP/UFERSA (ISSN 2526-7574)*, *1*(3).
5. Finance, Y. (2020). Yahoo Finance. *Retrieved from finance. yahoo. com: https://finance. yahoo. com/recent-quotes* , (acesso em 02 de novembro de 2024).
6. Junita, T. P., & Kartikasari, M. D. (2024). Application of the neural network autoregressive (nnar) method for forecasting the value of oil and gas exports in Indonesia. *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, *18*(1), 0341-0348.
7. Kim, Y. B., Kim, J. G., Kim, W., Im, J. H., Kim, T. H., Kang, S. J., & Kim, C. H. (2016). Predicting fluctuations in cryptocurrency transactions based on user comments and replies. *PloS one*, *11*(8), e0161197.

8. Hyndman, R. J., & Athanasopoulos, G. *Forecasting: Principles and Practice, 2nd edn. OTexts, Melbourne (2018)*.
9. Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media* (Vol. 8, No. 1, pp. 216-225).
10. Maleki, A., Nasseri, S., Aminabad, M. S., & Hadi, M. (2018). Comparison of ARIMA and NNAR models for forecasting water treatment plant's influent characteristics. *KSCE Journal of Civil Engineering*, *22*, 3233-3245.
11. McNally, S. (2016). *Predicting the price of Bitcoin using Machine Learning* (Doctoral dissertation, Dublin, National College of Ireland).
12. Morettin, P. A., & Toloi, C. M. (2018). *Análise de séries temporais: modelos lineares univariados*. Editora Blucher.
13. Nakamoto, S., & Bitcoin, A. (2008). A peer-to-peer electronic cash system. *Bitcoin.–URL: https://bitcoin. org/bitcoin. pdf*, *4*(2), 15.
14. NUBANK. Menu. Disponível em:<https://nubank.com.br/> Acesso em 15 nov. 2024.
15. Radityo, A., Munajat, Q., & Budi, I. (2017, October). Prediction of Bitcoin exchange rate to American dollar using artificial neural network methods. In *2017 international conference on advanced computer science and information systems (ICACSIS)* (pp. 433-438). IEEE.
16. Shintate, T., & Pichl, L. (2019). Trend prediction classification for high frequency bitcoin time series with deep learning. *Journal of Risk and Financial Management*, *12*(1), 17.
17. Thoplan, R. (2014). Simple v/s sophisticated methods of forecasting for mauritius monthly tourist arrival data. *International Journal of Statistics and Applications*, *4*(5), 217-223.
18. Ulrich, F. (1892). Bitcoin-a moeda na era digital. *Journal, volume*, *2*, 239.
19. Velankar, S., Valecha, S., & Maji, S. (2018). Bitcoin price prediction using machine learning. In *2018 20th International Conference on Advanced Communication Technology (ICACT)* (pp. 144-147). IEEE.