Bitcoin Price Prediction with Random Forest Technique using Python

التنبؤ بأسعار البتكوين بطريقة الغابة العشوائية باستخدام برنامج Python

Nawal Chicha[1]
University of Khemis Miliana – Algeria
n.chicha@univ-dbkm.dz

Received: 12/12/2021 *Accepted: 05/10/2022* *Published: 11/11/2022*

Abstract:

This study applies a Random Forest technique to analyze the Bitcoin fluctuations from January 2019 to December 2020. A Random Forest model, which attributes the right currency margin in forecasting, was designed using Python. The prices fluctuations were examined by various measures such as the root mean square error (RMSE), the mean absolute error (MAE), the mean absolute percentage error (MAPE), and R^2. The results show that this technique can forecast trends in Bitcoin and cryptocurrencies.

Keywords: Bitcoin, Random Forest, Volatility, Prediction, Cryptocurrencies

ملخص

تعتمد هذه الدراسة على طريقة الغابة العشوائية لتحليل تقلبات أسعار البتكوين خلال الفترة الممتدة من جانفي *2019* إلى ديسمبر *2020*. تم بناء نموذج الغابة العشوائية، والذي يمنح هامشا مناسبا في التنبؤ بتقلبات الأسعار، باستخدام برنامج *Python*. وتم فحص تقلبات الأسعار باستخدام مقاييس مختلفة هي خطأ مربع متوسط الجذر، ومتوسط الخطأ المطلق، ومتوسط نسبة الخطأ المطلق، ومعامل التحديد. تظهر النتائج أنه يمكن استخدام هذه التقنية للتنبؤ بالاتجاهات في أسعار البتكوين والعملات المشفرة الأخرى.

الكلمات المفتاحية: البتكوين ، الغابة العشوائية ، التقلبات ، التنبؤ، العملة المشفرة.

1. INTRODUCTION

Cryptocurrencies, like stock markets, are frequently used as a form of trading, which allows speculating on the price of cryptocurrencies, generating buy and sell orders for assets. The cryptocurrency boom has prompted many investors to make concerted efforts to forecast cryptocurrency prices to increase their profits. One of the main disadvantages of the cryptocurrency market is the high volatility of its prices, caused by the risk it entails. That is why it is essential to analyze the prices fluctuation. This fluctuation, present in cryptocurrencies, is related to the price fluctuation at different times. As a result, forecasting rises and falls in cryptocurrency prices are being investigated.

[1] - *Corresponding author: Nawal Chicha: n.chicha@univ-dbkm.dz*

Forecasting techniques are used in this study to predict the values of cryptocurrencies to reduce the market's high volatility[1]. Specifically, the Random Forest technique will study the most capitalized digital currency on the market: Bitcoin.

Bitcoin is the world's first decentralized cryptographic financial network, where its value has gone from zero to billions of dollars, and behind it, there is a story in relation to cryptocurrencies, innovation, technological advances, and new forms of financial assets.

Bitcoin became operational in January of the following year, the first decentralized cryptocurrency. The creation of Bitcoin created a new form of the financial system never seen before, one that is decentralized and has its system for the establishment and transaction of currency backed by users. In contrast to the discretionary and circumstantial decisions of central banks, the proposed system in Bitcoin establishes a system in which the creation of currency is completely transparent.

Bitcoin uses blockchain technology to support a decentralized infrastructure and transparent monetary policy[2]. The Blockchain is a technology that allows the safe transfer of data from one part to another without the involvement of third parties. By the end of the third quarter of 2019, the number of people with active Bitcoin wallets ranged between 13 and 40 million, according to Statista. Furthermore, with 38% and 27%, respectively, most users are concentrated in Asia and Europe.

Due to its importance, it is necessary to predict bitcoin prices to anticipate market movements. To achieve our goal, we will use data from January 1, 2019, to December 1, 2020, of Bitcoin values, mainly: the opening, maximum, minimum, and closing prices. A Random Forest technique is used to create predictive models of the closing price based on selected characteristics.

The model will be evaluated out in a validation set using the root mean square error (RMSE), the mean absolute error (MAE), the mean absolute percentage error (MAPE), and the coefficient of determination (R^2). The significance of this study can be found in the fact that it is crucial to understand a topic of great interest, which is generating incentives worldwide due to its way of operating, due to its risks, limitations, and restrictions.

In this context, the study is organized as follows: the second section introduces the Bitcoin prices fluctuations and the Random Forest. The third section presents the methodology. The fourth section analyzes and discusses the study results, then the conclusion.

2. Prediction Bitcoin Prices Using Random Forest: A Literature Review

2.1. Bitcoin fluctuation: An overview

Cryptocurrencies, since their inception, were conceived as a new alternative within innovative payment mechanisms. Cryptocurrencies can be divided into three broad categories: Bitcoin, Altcoin, and Tokens. Bitcoin was officially the first cryptocurrency, using a blockchain with the same name. The cryptocurrencies created later are known as altcoin or alternative currencies because they emerged as an alternative to Bitcoin.

The last type of cryptocurrency is known as tokens, which do not have their own Blockchain but also function as payment methods and can even represent shares of a company or become a certain amount of cryptocurrencies. The common characteristics of most cryptocurrencies can be summarized in the following table:

Table 1. Main characteristics of cryptocurrencies

Characteristics	**Definition**
Decentralization	They are not related to any government or financial body
Cryptograph y	It uses cryptographic techniques to ensure the security of its transactions, making it impossible for them to be forged or duplicated
Transparency	All transactions are public and freely accessible thanks to Blockchain technology
Volatility	It is common for the price to fluctuate, which gives a feeling of insecurity
Regulation	Depending on the country, for example, in European countries, there is some regulation, while in Japan, they are fully regulated
Irreversibilit y	Third parties do not intervene (P2P)
Privacy	It is not necessary to provide personal data, but in practice, to avoid money laundering

Source: *Author elaboration based on a meta-analysis review*

The main difference between the types of cryptocurrencies is their integration with the Blockchain. However, they all work as digital assets, and it is possible to obtain historical prices, so their integration with the program is compatible. According to Biczok (2018), more than a thousand varieties of alternative currencies exist, and only 18 of them have managed to exceed a volume of one billion dollars. Bitcoin is the most valuable cryptocurrency, and it is distinguished by its use of Blockchain technology to[3]:

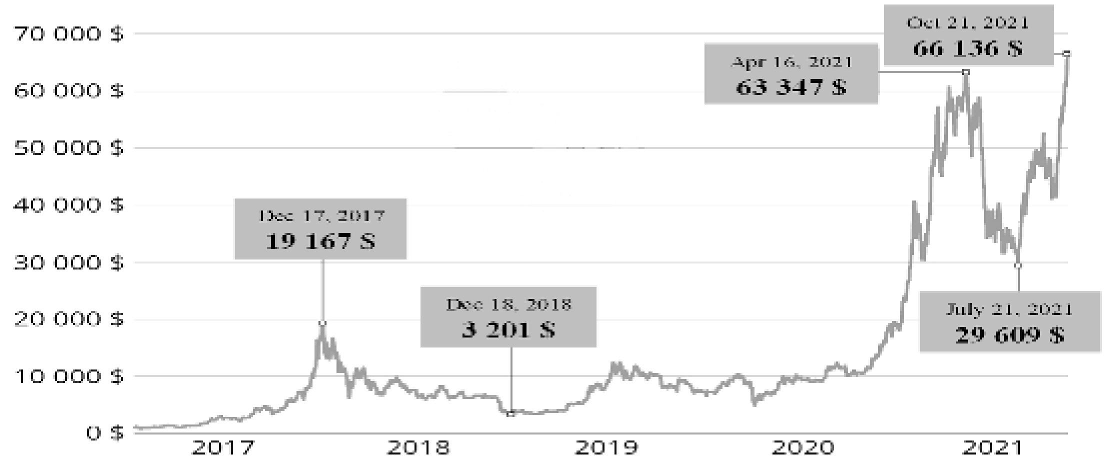
- Achieve decentralization concerning any governmental authority;
- Ensure the anonymity of each transaction, as it is impossible to track any transaction.
- All transactions are recorded in a publicly accessible ledger;
- Money is transferred more quickly than through banks or national entities;
- Each transaction is irreversible, as it cannot be corrected or canceled;

Bitcoin, known by its acronym BTC within the internet network, is a virtual currency that can be used as a means of payment in shops, stores, warehouses, or other places in the same way as physical money. The operation of the Bitcoin network as a whole is a series of complex processes[4], including portfolio registration and key generation, transaction verification, and the corresponding updating of balances, block creation, or mining.

Bitcoin has its origin in 2009, where Satoshi Nakamoto created the first virtual, independent, and decentralized currency. This currency is very volatile in terms of its value due to speculation[5]. In the new business model through the internet, Bitcoin is considered one of the fastest-growing payment mechanisms; the first Bitcoin block gave a reward of 50 BTC in 2009; this value is equivalent to more than 275,000 US dollars in the Blockchain market.

As shown in the following figure, after reaching an all-time high of over 63,000$ in April 2021, Bitcoin had fallen below 30,000$ by the end of July due to a bleak market outlook and market restrictions.

Fig.1. The evolution Bitcoin price (from 2017 to October 21, 2021)



Source: https://www.coindesk.com/price/bitcoin/

Therefore, the concretization of the transfer of knowledge about the new payment mechanisms and currencies to society via Blockchains is regarded as critical in the formalization of the cryptocurrency market, where knowledge about this new one must inevitably be incorporated market to ensure a continuous flow of shared knowledge.

The data obtained regarding Bitcoin purchase and sale movements show a significant increase compared to the behavior of the 2015-2016 cut, in a market valued at 185,000 million dollars and with a 50% growth expectation for 2019. According to experts, the final consolidation of this currency in the financial market will occur in 2020.

2.2. Random Forest Technique Application

Because most cryptocurrencies are decentralized, their prices are not affected by monetary policies, inflation rates, or interest rates, but rather by the perception of users based on information available in the news, internet sites, and media, as well as other non-fundamental factors[6].

Similarly, the supply and demand of cryptocurrencies play a fundamental role in the valuation of these assets since, unlike traditional currencies, some cryptocurrencies have a limited supply of them. An analysis of the instrument in question is required to obtain information for correct decision-making and due to the constant changes in the prices of cryptocurrencies.

It is common to think of a single regression model to predict an outcome when considering statistics. However, more recent data science research suggests that "assembly methods," in which multiple models are created from the same data set and intelligently combined, perform better than predictions based on a single model[7].

Random Forest is a technique of Machine Learning algorithm that defines a prediction model based on the decision trees[8]. The Random Forest technique used to forecast Bitcoin prices comprises multiple decision trees[9].

Random forests have been used to predict the direction of a trend in stock prices. In 2016, Khaidem et al. used random forests to predict trends in the prices of Apple and General Electric shares listed on the NASDAQ stock exchange; and the share prices of Samsung Electronics.

A Random Forest is an ensemble tool of decision trees combined with Bagging[10]. It causes each tree to be trained with different data samples for the same problem. When using Bagging, what is happening is that different trees see different portions of the data. No tree sees all the training data. In this way, when combining their results, some errors are compensated for others, and we have a prediction that generalizes better[11].

Overfitting is a critical issue that can wreak havoc on results, but the Random Forest algorithm will not overfit the model if there are enough trees in the forest. Another benefit is that the Random Forest classifier can handle missing values. Finally, the Random Forest classifier can be modeled for categorical values. When selecting the attributes, there are different measures on how to diversify each node correctly, such as the Gini impurity or the Shannon entropy. The following equation gives the Gini impurity:

$$G(N) = \sum_{i=1}^{c} p(i) * (1 - p(i))$$

While C is the number of classes, and p(i) is the probability of misclassifying a randomly chosen item.

It should be noticed here that a pure database is one in which all of our data relates to a single class. Its value ranges between 0 and 1, with 0 indicating that all elements belong to a single class and one indicating that the elements are randomly distributed across several classes. A Gini index of 0.5 indicates that all elements are distributed equally in some classes.

The Random Forest is based on another technique known as Bagging, which consists of performing a Bootstrap on the experiment's data sampling on the input variables and using a fixed number of variables.

As a result, the model considers a set of random characteristics in each division rather than the entire totality present in the model, and then the dimensionality is reduced (Hinton, 2006). The Random Forest model takes into account the following:

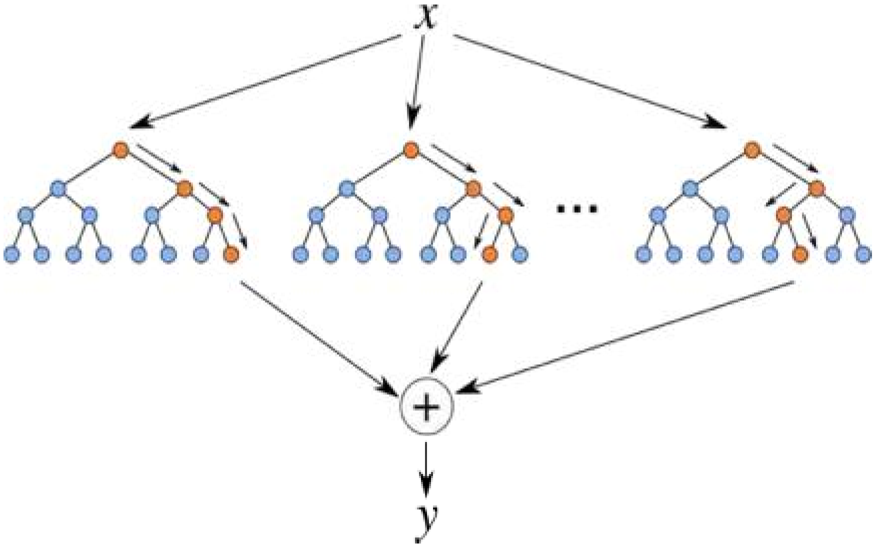
N numbers of cases in the training set are randomly sampled with replacement. This sample constitutes the training set to build tree i.

Assuming that the number of input variables is M, the number of variables selected for each node is m ($m < M$), which is kept constant for the generation of the block. The block is then related by partitioning the node and using the best division of those m attributes. In each division, it is recommended to obtain m ≈ √(M) predictor variables and minimum size of nodes 1, for classification problems and m ≈ M and minimum size of nodes 5, for three regression problems. The number of predictors considered in each division is approximately equal to the square root of the total number of predictors. The best way to find the optimal value is to evaluate the Out Of Bag-MSE for different values of m. In general, if the variables selected at each node are highly correlated, small values of m lead to good results.

By this, m observations are used at each node for training and M − m for testing.

Globally, given a training set D of size n, m samples with replacement are generated that will serve as new training sets $\mathrm{D_i}$ $(\mathrm{i} = 1, \dots, \mathrm{m})$ of size n'. Afterward, each decision tree is trained with a set $\mathrm{D_i}$. Moreover, having a new observation, it enters the nodes of each tree to make a prediction, as can be seen in figure 2.

Fig.2. Structure of a random forest



Source: *https://www.paradigmadigital.com/techbiz/machine-learning-dummies/*

When obtaining the predictions of all the forest trees, these are combined to get the final prediction. The effectiveness of random forest models for the validation data will usually be carried out by calculating:

- RMSE (root mean square error):

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

- MAE (mean absolute error):

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

It should be noticed that the RMSE and MAE have a value range of $[0, \infty]$. In both cases, the lower the value, the less efficient the prediction model.

When significant errors are not desired, RMSE is usually most useful.

- MAPE (mean absolute percentage error):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right|$$

- R^2 (R squared or coefficient of determination):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\bar{y}_i - y_i)^2}$$

R^2 indicates the quality of the prediction model to replicate the results.

It is frequently used to explain how well or poorly the selected independent variables explain the variability in their dependent variables.

3. Methodology

The process used to structure the prediction model consists of the following parts:

3.1. Data and Method

To develop the model, it was necessary to extract historical prices of the cryptocurrencies of interest: Bitcoin. Because the prices of these change depending on the market in which they are acquired, it was necessary to include the option of obtaining historical data from different markets.

In this part, we must decide what data is required to achieve the paper's purpose. The relevant information must be chosen, along with why that information was chosen. Understanding what each variable contributes to us and verifying their accuracy is critical for obtaining the correct data.

Once these steps are completed, it is very important to clean the data, using the necessary amount and excluding all empty or erroneous information that may appear in the database. We analyzed how the changes in data can contribute and what positive impact they may have on our next application of the Random Forest technique.

As stated in the objectives of this study, the prediction of Bitcoin is analyzed to evaluate the Random Forest technique and deal with financial uncertainty. Bitcoin prices will be predicted using a validation set and then compared to real prices. Python has been used to obtain the desired data from Bitcoin and create a model and analyze its results.

Subsequently, it was necessary to select a cryptocurrency exchange platform from which the historical prices would be extracted and the operations carried out. Cryptocurrency markets are similar to traditional financial markets[12], but they are mainly focused on cryptocurrency transactions: buying and selling cryptocurrencies through local currencies, such as dollars, euros, or other cryptocurrencies.

Some of the most important markets include: "Binance," "Coinbase," "Bittrex," and "Bitfinex." The difference between them and between all cryptocurrency markets in general lies mainly in the offer prices of cryptocurrencies, their trading volumes, and their transaction fees.

Table 2. Comparison between Cryptocurrencies platforms (as of November 6, 2020)

Platform	**Criteria**				
	Exchange ($)	**Volume**	**Liquidity***	**Taker fee**	**Maker fee**
Binance	19,559,285,062	541	855	0.1%	0.1%
Coinbase	1,109,238,524	413	99	0.5%	0.5%
Bitrex	525,677,046	413	297	0.1%	0.2%
Bitfinex	907,860,038	441	185	0.16%	0.26%

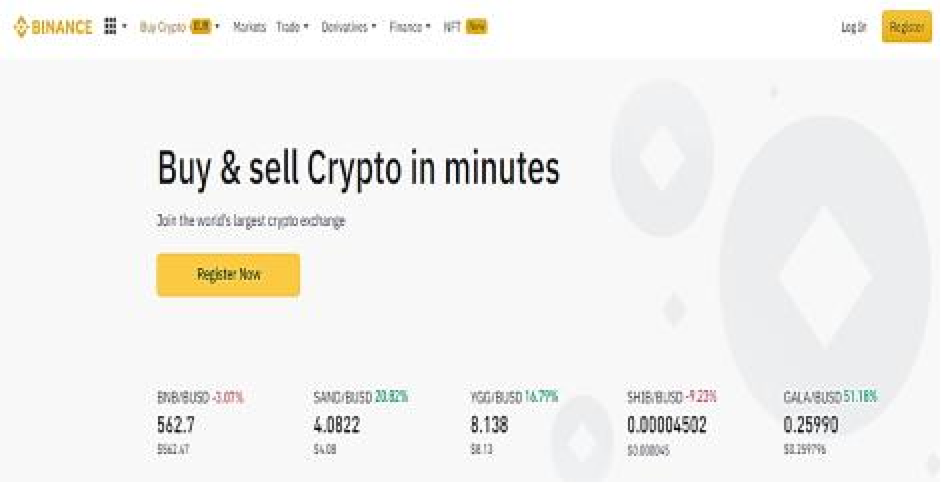
* Liquidity is measured as a number between 0 and 1000

***Source**: Author elaboration based on data retrieved from Platforms*

Table 2 shows the volume, liquidity, number of currencies available in the market, and transaction fees for the markets as mentioned above. Due to its high liquidity, the great variety of cryptocurrencies available, and low transaction costs, it was decided to extract the information from Binance.

Therefore, all of the data was explicitly obtained through the Binance trading platform, one of the most popular among investors due to its high performance, security, and stability. This platform has over 100 cryptocurrencies available for use.

Fig.2. Data source



***Source**: https://www.binance.com/en*

The Bitcoin data was collected in CSV format (to be used on Python). It should be mentioned that the value of Bitcoin is collected in the US dollar. This type of presentation was chosen because of its widespread use in exchanges and its high utility.

The data was collected from January 1, 2019, to December 1, 2020.

The data for this study consists of:

- Timestamp: start of the time interval;
- Open: the value of the opening price of Bitcoin in the set interval;
- High: the value of the maximum price of Bitcoin in the set interval;
- Low: the value of the minimum price of Bitcoin in the set interval;
- Close: the value of the closing price of Bitcoin in the set interval;
- Volume: the amount of an asset invested in the set interval;
- Close time: end of the set interval;
- Y: closing price of n + 1. A prediction will be made with the characteristics of n, which will then be compared with the actual results.

After that, we must choose methods that allow predicting Bitcoin price. Many techniques can be applied, but the one that can best respond to the needs based on the data provided must be selected. Various techniques can be used and then compare the results of each one, which is very helpful.

We choose to apply Random Forest for this study by using Python libraries. There are several libraries in Python that allow us to extract information about the historical prices of Bitcoin**.**

3.2. Variables of the study and Correlation Analysis

After data collection, the variables were identified. These variables make use of the closing and opening values and the highs and lows. The variables of this study consist of:

a)- Dependent variable: the dependent variable is the variable to predict. It refers to the closing price of the n + 1. With the independent variables of n, we study the price of the next interval. This variable is collected in a Database and coded as y.

b)- Independent variables: The variables used to build the prediction model are:

- The opening;
- Closing;
- Minimum and maximum prices of n.

The correlation analysis shows inferior correlation results, for example, between the dependent variable and volume (0.125) and between the dependent variable and the number of operations (0.281).

That is why we will not consider these characteristics for constructing the model. Because of this, we have deleted these variables from the model. As a result, the essential variables will be investigated in order to determine which independent variables have the most significant influence on Y.

In this step, each model that has been generated is explored, and its effectiveness checked. It is important to check the correct operation of the evaluated models and then select those that a priori give rise to better results.

Before building the model, a relationship was carried out between the variables. The results are presented in the following table. It is worth noting that the ignore variable has been removed because it provided no useful information to the model.

The results show that the correlation between the open, high, low, and close variables is positive and significant. As the two last variables are not helpful to predict the Bitcoin price, we only choose Open, High, Low and Close as features in the modeling phase.

Table 3. Correlation Results

Independent variables	Correlation value
Open	0.979
High	0.967
Low	0.971
Close	0.883
Volume	0.125
Close time	0.281

***Source**: Author elaboration based on Python outputs*

The data is divided into two subsets to solve a regression problem with the Random Forest technique: the training data and the test data. For this, it is necessary to note that the training and validation sets formed by the four independent variables are formed as expected.

Then, two sets are created orderly: (i) the initial 75% of the total data set corresponds to the training set and; (2) the remaining 25% to the most recent test set. The train set is the set of data used to train the model and with which its parameters will be determined. In the case of random forests, the parameters are the variables and thresholds used to separate the data set at each node.

The validation set is used to select the model and compare hyperparameters. Hyperparameters are external settings that are not learned during training but are defined to improve performance.

In the case of random forests, some hyperparameters include the number of trees in the forest, the maximum and minimum depth of each tree, and the minimum number of observations required at each node.

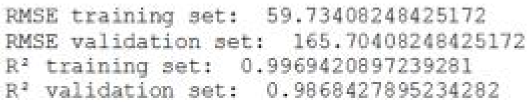
Our data is collected in 23 months, so almost six last months will be used to validate the test set. Note that the training set begins on January 1, 2019, and ends on May 12, 2020. Therefore, the validation set begins on May 12, 2020, and ends on December 1, 2020.

4. RESULTS AND DISCUSSION

4.1. Model building

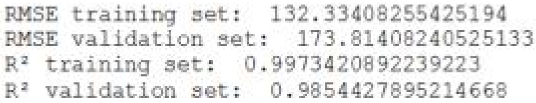
Before creating a prediction model, RMSE will be calculated to compare its value and the value of R^2 in both sets. The errors and parameters will not be shown together until the results are completed, but to explain the procedure, some results and the changes made in each model need to be clarified.

The first model is created with default parameters except removing the randomness each time it is executed. Once the model is built, it is trained. Using Python Libraries, RMSE and R^2 values were obtained:



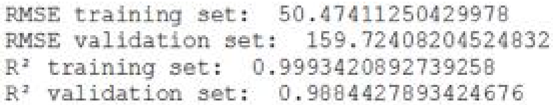
```
RMSE training set:  59.73408248425172
RMSE validation set:  165.70408248425172
R² training set:  0.9969420897239281
R² validation set:  0.9868427895234282
```

To obtain better results, a new model is proposed, increasing the number of samples at the end node, reducing the tree's depth.



```
RMSE training set:  132.33408255425194
RMSE validation set:  173.81408240525133
R² training set:  0.9973420892239223
R² validation set:  0.9854427895214668
```

Using these parameters does not improve the error in the validation set, and using less depth, a priori, does not benefit the model. So we will build a new model with the same trees as the previous one and a depth that we consider to be decent in the first instance, and we get the following results:

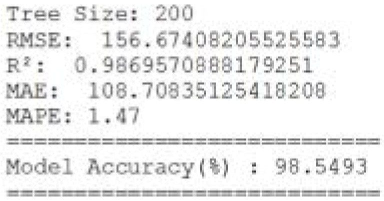


```
RMSE training set:  50.47411250429978
RMSE validation set:  159.72408204524832
R² training set:  0.9993420892739258
R² validation set:  0.9884427893424676
```

Because the RMSE is reduced and the R^2 increases, we assume that making these changes benefits the model. To evaluate the prediction model's effectiveness in terms of the parameter number of trees, a Python function was created that returns the errors in the model while varying the number of trees.

Consider all the default values except for the number of trees, which will vary from 10 to 10 in each model up to 500 trees; RMSE, MAE, MAPE, and R^2 will be calculated. It shows a stable RMSE at 157, decreasing but entails greater computational expense.

However, of all the trees studied, the one with a lower RMSE and a higher R^2 has 200 estimators. It is presented below:



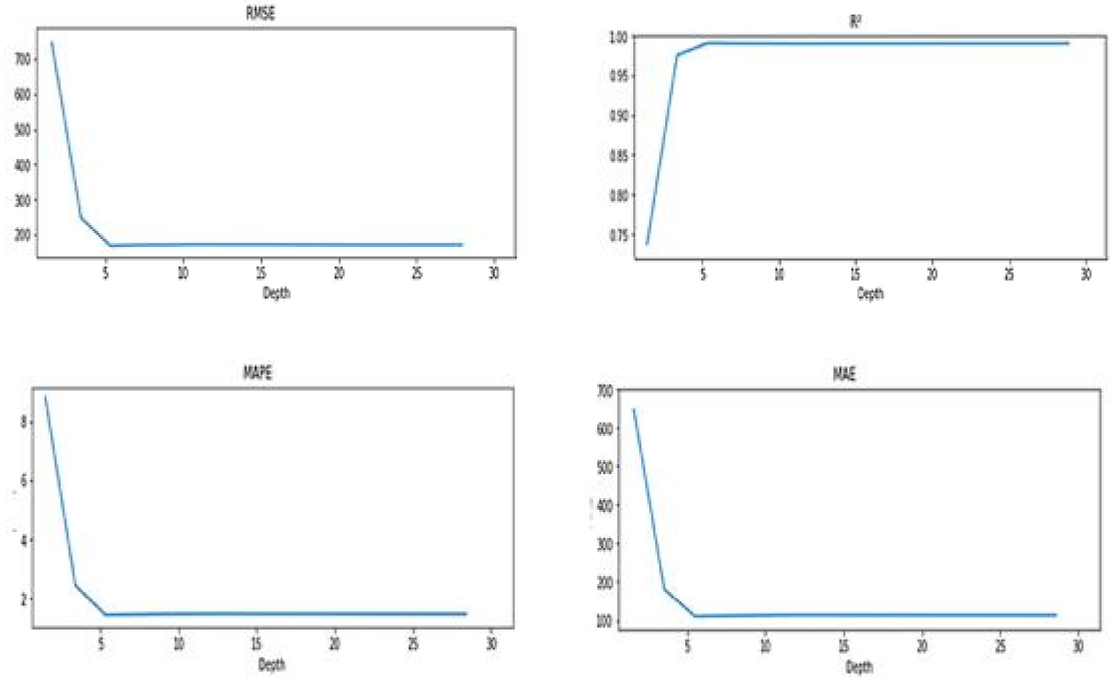
```
Tree Size: 200
RMSE:  156.67408205525583
R²:  0.9869570888179251
MAE:  108.70835125418208
MAPE: 1.47
============================
Model Accuracy(%) : 98.5493
============================
```

Therefore, it is suggested to choose 200 trees.

Something very similar to what was proposed with trees has been proposed to test the effectiveness of the prediction model in terms of the maximum depth parameter. The results of the previous 200 trees will be compiled to maximize the parameters of depth and number of trees.

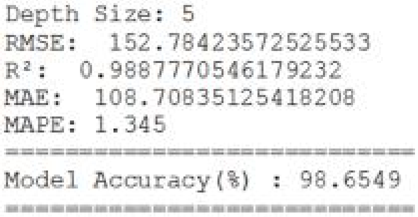
We will use the same function that returns the validation set's errors RMSE, MAE, MAPE, and R^2. The model begins to fully stabilize at the fifth depth, as shown in the following figure:

Fig.3. Depth Function (RMSE, R^2, MAE, MAPE)



***Source**: Python outputs*

However, of all the possibilities investigated, the one with the lowest RMSE and the highest R^2 has depth 5. It is shown below:



```
Depth Size: 5
RMSE:  152.78423572525533
R²:  0.9887770546179232
MAE:  108.70835125418208
MAPE: 1.345
============================
Model Accuracy(%) : 98.6549
============================
```

It should be noted that the developed model corresponds to the Bitcoin US Dollar exchange rate. The obtained errors and R2are shown as a function of depth and under the parameters.

4.2. Model selection and Discussion

In this section, the result of the chosen model will be discussed with the corresponding model. That is the chosen model in which explanatory variables have been used at time t, and the model with explanatory variables at time t and t-1 will be compared. Using the same parameters in the models, with four and eight explanatory variables, it will be possible to determine whether or not there have been improvements by simply changing the number of variables.

It is assumed that the last node must contain at least 25 samples. These 25 observations are required to perform a new tree node division. Increasing this figure reduces the depth of the tree, which is detrimental in the proposed model (errors increase).

The forest has 40 trees, 30 more than the default model. It is expected that the more trees there are, the more accurate the results will be, though this is not always the case because the improvement does not have to be significant.

To accomplish this, it was tested with four characteristics and their square root, i.e., between 4 and 2. The modified final model achieves the highest R^2 value, followed by the last three models. However, the outcome is very similar among those discussed. The same thing happens with the RMSE, MAE, and MAPE errors.

It is worth noting that the average price of Bitcoin (in US dollar) in the selected time period is 7500 $, and the RMSE is around 151. The best results in this model have been obtained by varying the number of trees and the depth, though there are some hyperparameters for which no manual search has been performed.

The first model is selected for having RMSE, MAE, MAPE, and R^2, very similar to the model with the best results. It is essential to mention that the non-selection of some hyperparameters such as the maximum or the minimum of observations in a node can accompany the overfitting above problem. For this reason, this decision has been made and not to choose the modified model.

It seems trivial to interpret that the maximum and minimum of the interval n are very likely to affect the interval n+1 if there are no very significant changes (Dash & Dash, 2016). Then, analytically, the variables that matter most in the model are shown in the following table

Table 3. Table title (this is an example of table 1)

Independent variables	Importance
Open	0.118823
High	0.210046
Low	0.221035
Close	0.473445

***Source**: Author elaboration based on Python outputs*

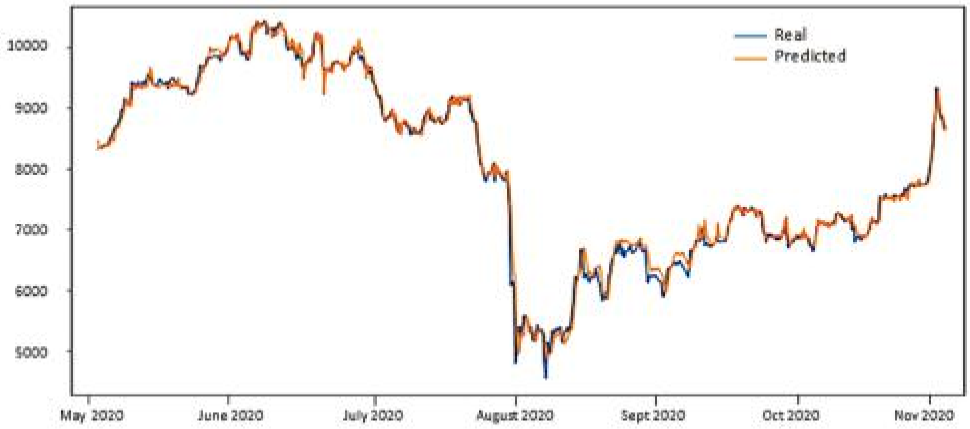
The most important variable for predicting the closing price is the "close" variable with 47.3%. The other variables, mainly: "Open," "High," and "Low," have a percentage ranging between 10% and 23%, which can be considered important for the model.

The maximum price of the "Low" is also particularly important, as is the minimum price of the "high" variable, both touching 22.1% and 21%, respectively.

The "Open" price variable seems to be less significant, although it contributes around 12% of the importance of the model.

Therefore, a predicted and real Bitcoin price can be presented graphically.

Fig.4. real and predicted Bitcoin price using the Random Forest technique



***Source**: Python outputs*

All this indicates that, as previously mentioned, having more prices is very beneficial to getting accurate predictions. The importance of the variables depends entirely on the model, but the models generally have in common that the closing price always influences. The results indicate that the closing price is the feature that has managed to make a more accurate prediction in the validation set.

5. CONCLUSION

Random Forest has undoubtedly caught the attention of many investors in recent years. The fields of application of this technique seem innumerable, and the results obtained in this work show that they can be helpful in asset trading. The results in this work show that even the method not as sophisticated as random forest could help predict trends in cryptocurrencies and other financial assets.

This study analyzes the Bitcoin price to subsequently make predictions of the closing price using the Random Forest technique. Beyond the training and validation data, the realization and implementation of a prediction system are proposed in order to invest effectively and reflect the balance of gains and losses after a specific time.

Similarly, it is suggested that this data be processed using other techniques and tested for effectiveness before selecting the algorithm that produces the best results. Precisely, it was initially thought that neural networks, specifically the LSTM extension, would be used. On the other hand, expanding the database with data prior to those chosen indicates that it can be very beneficial in relation to the previously stated conclusions.

Similarly, it seems logical to believe that if the results were good over the chosen period, it might be worthwhile to extend this technique to make short-term predictions. It could, for example, be proposed to make predictions over the next three months, updating the data obtained up to the present date. In addition, it is suggested that more variables be used and that the best results be obtained.

5. Bibliography List :

Journal article

Al-Yahyaee, K.H., Mensi, W., & Yoon, S.M. (2018). Efficiency, multifractality, and the long-memory property of the Bitcoin market: A comparative analysis with stock, currency, and gold markets. Financ. Res. Lett. 27, 228–234

Biau, G. (2012). Analysis of a random forests model. The Journal of Machine Learning Research, 13(1), 1063–1095

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5–32

Breiman, L. (1996). Bagging predictors. Mach. Learn. 24, 123–140

Chen, T.H., Chen, M.Y., & Du, G.T. (2021). The determinants of bitcoin's price: Utilization of GARCH and machine learning approaches. Comput. Econ. 57, 267–280

Dahir, A.M., Mahat, F., Noordin, B.A.A., & Razak, N.H. (2019). Dynamic connectedness between Bitcoin and equity market information across BRICS countries: Evidence from TVP-VAR connectedness approach. Int. J. Manag. Financ. 16, 357–371.

Dash, R., & Dash, P. K. (2016). A hybrid stock trading framework integrating technical analysis with machine learning techniques. The Journal of Finance and Data Science, 2(1), 42–57.

Easley, D., O'Hara, M., Basu, & S. (2019). From mining to markets: The evolution of bitcoin transaction fees. J. Financ. Econ. 134, 91–109

Hinton, G.E., & Salakhutdinov, R.R. (2006). Reducing the dimensionality of data with neural networks. Science 313, 504–507.

Jang, H., & Lee, J. (2017). An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information. IEEE Access, 6, 5427–5437.

Makarov, I., & Schoar, A. (2020). Trading and arbitrage in cryptocurrency markets. Journal of Financial Economics, 135(2), 293–319.

Khaidem, L., Saha, S., & Dey, S. R. (2016). Predicting the direction of stock market prices using random forest. arXiv preprint arXiv:1605.00003.

Mitchell, T. M. (1999). Machine learning and data mining. Communications of the ACM, 42(11), 30–36.

Internet websites

Nakamoto Satoshi (2008). Bitcoin: A peer-to-peer electronic cash system, Available at: http://bitcoin.org/bitcoin.pdf (consulted on 28/11/2021)

Biczok, D. (2018). The future of bitcoin and the blockchain technology, Available at: http://investas.lu/CMS/images/PDFs/Biczok_Master_Thesis.pdf (consulted on 28/11/2021)

7. Citations:

[1] *Chen, T.H., Chen, M.Y., Du, G.T. (2021). The determinants of bitcoin's price: Utilization of GARCH and machine learning approaches. Comput. Econ. 57, 267–280*

[2] *Nakamoto Satoshi (2008). Bitcoin: A peer-to-peer electronic cash system, Available at: http://bitcoin.org/bitcoin.pdf (consulted on 28/11/2021)*

[3] *Biczok, D. (2018). The future of bitcoin and the blockchain technology, Available at: http://investas.lu/CMS/images/PDFs/Biczok_Master_Thesis.pdf (consulted on 28/11/2021)*

[4] *Dahir, A.M., Mahat, F., Noordin, B.A.A., & Razak, N.H. (2019). Dynamic connectedness between Bitcoin and equity market information across BRICS countries: Evidence from TVP-VAR connectedness approach. Int. J. Manag. Financ. 16, 357–371.*

[5] *Easley, D., O'Hara, M., Basu, S. (2019). From mining to markets: The evolution of bitcoin transaction fees. J. Financ. Econ. 134, 91–109*

[6] *Makarov, I., & Schoar, A. (2020). Trading and arbitrage in cryptocurrency markets. Journal of Financial Economics, 135(2), 293–319*

[7] *Jang, H., Lee, J. (2017). An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information. IEEE Access, 6, 5427–5437*

[8] *Mitchell, T. M. (1999). Machine learning and data mining. Communications of the ACM, 42(11), 30–36*

[9] *Breiman, L. (2001). Random forests. Machine learning, 45(1), 5–32*

[10] *Breiman, L. (1996). Bagging predictors. Mach. Learn. 24, 123–140*

[11] *Biau, G. (2012). Analysis of a random forests model. The Journal of Machine Learning Research, 13(1), 1063–1095*

[12] *Al-Yahyaee, K.H., Mensi, W., Yoon, S.M. (2018). Efficiency, multifractality, and the long-memory property of the Bitcoin market: A comparative analysis with stock, currency, and gold markets. Financ. Res. Lett. 27, 228–234*