Journal of Informatics
Electrical and Electronics Engineering, 2023,
Vol. 04, Iss. 03, S. No. 064, pp. 1-16
ISSN (Online): 2582-7006

Bitcoin and Cryptocurrency Exchange Market Prediction and Analysis Using Big Data and Machine Learning Algorithms

Branković Adnan[1], Jukić Samed[2]

[1]International Burch University, Faculty of Engineering, Natural and Medical Sciences, Department of Information Technologies, Francuske revolucije bb, Ilidža 71210, Bosnia and Herzegovina
[2]International Burch University, Faculty of Engineering, Natural and Medical Sciences, Department of Information Technologies, Francuske revolucije bb, Ilidža 71210, Bosnia and Herzegovina
[1]adnan.brankovic@stu.ibu.edu.ba, [2]samed.jukic@ibu.edu.ba

How to cite this paper: B. Adnan and J. Samed, "Bitcoin and Cryptocurrency Exchange Market Prediction and Analysis Using Big Data and Machine Learning Algorithms," *Journal of Informatics Electrical and Electronics Engineering* (JIEEE), Vol. 04, Iss. 03, S No. 064, pp. 1–16, 2023.

https://doi.org/10.54060/jieee.2023.64

Received: 01/04/2023
Accepted: 01/06/2023
Online First: 08/08/2023
Published: 25/11/2023

Abstract

Due to economic uncertainty and the financial crisis of 2008, a desire for an unregulated currency arose, leading to the invention of Bitcoin. Using a pseudonym called Satoshi Nakamoto, Bitcoin was created in 2009, anonymously or by a group of unknown individuals. Since Bitcoin has been the most valuable cryptocurrency in recent years, its prices have fluctuated dramatically, making it difficult to predict their prices. Investors, businesses, risk managers, and market analysts can all benefit from being able to predict Bitcoin prices. By using the Bitcoin transaction data obtained from the Bitstamp website in this study, several different Machine Learning models are employed to determine the most accurate model for predicting Bitcoin prices. These models are based on 1-minute interval exchange rates in USD from January 1, 2012, to January 8, 2022. Analysis was performed primarily with Python, but it was also used and Hadoop, a distributed data storage and processing framework that uses the map-reduce programming model to allow efficient parallel processing of Big Data. Based on the results of our research, comprising three experiments, autoregressive-integrated moving average (ARIMA) makes the most accurate prediction of Bitcoin prices, with a 95.98% success rate.

Keywords

Bitcoin, Cryptocurrency, Machine Learning, Big Data, ARIMA, LSTM

1. Introduction

The use of currencies has been around since ancient times when they were used as a medium of exchange for goods and services that are valuable which were not meant to be exchanged for themselves, but for another good or service. For thousands of years, the concept has taken the form of a physical object with limited supply, either natural (precious metals) or artificial (tokens issued by monopolists). Today, as a result, fiat currencies are now free to float and are only backed by the faith and credit of the governments that issue them. It is important to note that both fiat and gold-based currencies have some weaknesses, including inflation, government-bound value (depending on government stability), limited privacy (government can easily trace currency transactions), high fees and limits (withdrawals and spending limits per day and high international transfer fees), but Bitcoin attempts to overcome these weaknesses. With Bitcoin, however, there is no need for a single recordkeeper, so it solves both the issue of controlling the creation of digital currency and preventing its duplication at the same time. Validation can be difficult, but those who do are rewarded by being able to create new Bitcoins under a controlled environment. To access the Bitcoin network, users must download a Bitcoin software program and participate in the Bitcoin network, which allows transactions to be updated and verified by all participants as well as engage in operations, so as a cryptographic currency, Bitcoin is controlled through the use of cryptography. As opposed to a standard fiat currency, the value of Bitcoin is independent of any group, company, government, regardless of the amount of money in circulation. A public ledger called blockchain records payments verified and recorded by users on the computer network by mining, a process of contributing computing power. Transaction fees and new Bitcoins are received in exchange for this service. The process of creating Bitcoins does require real resources (computer hardware and energy), but since Bitcoin mining is free, the term "mining" may suggest that Bitcoin is not a fiduciary. While Bitcoin's market value is determined by how many are created each day (regardless of the network size), once created, they have no value other than as a medium of exchange.

Bitcoin users can buy and sell cryptocurrencies using their cryptocurrency exchange account, which can be used to create various types of orders related to the cryptocurrency market, including buy, sell, and speculation orders, just as they would on other traditional trading platforms [1-5].

Bitcoin is accepted as a payment option by more and more companies. By the end of 2022, there will be more than 15.200 companies including the most famous ones such as Microsoft, AT&T, Burger King, Twitch, and many other well-known companies. In October 2022, Bitcoin balances totaled USD 2,35 billion, compared with USD 2,238 billion in U.S. currency circulating, according to the latest data. Cryptocurrencies must catch up with Visa's capabilities (over 24,000 transactions per second (TPS)) if they are to gain mass adoption, since the Bitcoin network processes only seven transactions per second (TPS), whereas Ethereum processes twenty transactions per second (TPS), due to the limited number of transactions in each block, which is estimated that Bitcoin blocks (one or a few transactions per block) are generated once every ten minutes. At the moment, the smallest amount of Bitcoin a user can send or receive in a transaction is 0.93 USD (0.0000546 BTC), while the largest recorded Bitcoin transaction was 1.1 billion (161,500 BTC) on April 10, 2020.

There were only two nations in the world accepting Bitcoin as legal tender until May 2022, and those are El Salvador and the Central African Republic, while other nations had different cryptocurrency regulations. World Bank estimates 71% of the Central African Republic's 5.4 million inhabitants live below the international poverty line, despite its rich diamond, gold, and other mineral resources because years of political instability and violence have gripped the country. The government may have been told that this will bootstrap payments, but it is unclear how because internet coverage in the Central African Republic is just 11%. El Salvador faced protests after it introduced its Bitcoin Law, and the International Monetary Fund also criticized the country, because of concerns about its potential impact on financial stability and consumer protection [5-10].

As cryptocurrency prices fluctuate and are very unstable, making forecasting difficult, we compared multiple machine learning models for forecasting cryptocurrency market movements in order to fill this gap in the field. Cryptocurrencies are similar to stocks in that they do not have the same risk factors as stock investments, such as negotiations and fluctuations in stock prices.

Bitcoin price prediction is important for several reasons. The following are some of the key reasons:

Investment decisions: By utilizing Bitcoin price predictions, investors will be able to make better decisions about when to buy, hold, or sell Bitcoins. They will also be able to determine when to enter or exit the market depending upon the price path.

Risk management: Investors can also manage risk with Bitcoin price predictions by knowing how much volatility is expected in the market, so they can adjust their risk management strategies if prices are expected to be volatile.

Business planning: A sense of where the price of Bitcoin is headed can be crucial for businesses in the cryptocurrency industry in planning and decision-making, and when these businesses know where the price of Bitcoin is heading, they can make strategic decisions regarding their operations, investments, and expansion plans.

Market analysis: A cryptocurrency price prediction is also important for market analysts who wish to understand market trends and movements, and analysts can make predictions about the future of the cryptocurrency market and identify investment and growth opportunities by studying price predictions and analyzing market data.

The skill of predicting Bitcoin prices can be helpful to investors, businesses, risk managers, and market analysts [11-17].

2. Bitcoin Price Over Time

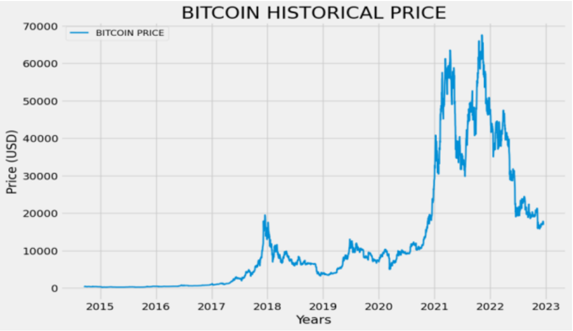

Figure 1. Bitcoin historical price

A graphic representation of Bitcoin price movement from the beginning of 2015 to the end of 2022 can be seen in Figure 1. After the registration of bitcoin.org on August 18th, 2008, the Bitcoin project was recorded on SourceForge.net, an open-source projects community resource, on October 31st, after following a link to Satoshi Nakamoto's paper named “Bitcoin: A Peer-to-Peer Electronic Cash System”. As a result of Satoshi Nakamoto's mining of the first Bitcoin block (the genesis block) in January 2009, a Bitcoin network was born, and on January 12th, the first Bitcoin transaction, known as block #170, was made by Satoshi Nakamoto and Hal Finney, an American software developer. The Bitcoinwiki website indicates that during its history, Bitcoin has gained more legitimacy among lawmakers and legacy financial institutions by 2017 [16-19]:

- Japan passing a law accepting Bitcoins as a legal payment method,
- Russia announcing it will legalize cryptocurrencies, such as Bitcoin, and
- Skandiabanken, Norway's largest online bank, has integrated Bitcoin accounts.

In response to Google's and Facebook's efforts to protect investors from fraud, Twitter announced it would ban cryptocurrency promotion during March 2018. Bitcoin's price burst into action once again in 2020 when the economy shut down due to COVID-19 when it started the year at around 7,000 USD. As a result of the pandemic shutdown and government policies that followed, investor fears regarding the global economy grew, accelerating Bitcoin's growth. As of November, Bitcoin was trading at 19,000 USD Bitcoins price reached just under 29,000 USD in December 2020, an increase of 416% from the beginning of the 2020 year. The price of Bitcoin surpassed 40,000 USD in January 2021 in less than a month 2021, breaking the price record set in 2020. Coinbase, which is a cryptocurrency exchange, went public on April 7, 2021, pushing Bitcoin prices to new all-time highs of over 60,000 USD. Institutional interest pushed the price further upward until Bitcoin reached 63,558 USD by April 12, 2021. In mid-December 2021, Bitcoin dropped to 46,164 USD after reaching an all-time high of 68,789 USD on November 10, 2021, before closing at 64,995.17 USD. The price began fluctuating more as investors grew concerned about inflation and the emergence of a variant of COVID-19 called Omicron. During the period between January and May 2022, Bitcoin's price dropped gradually, reaching 47,445 USD by March 2022 before dropping further to 28,305 USD on May 11th. The price of bitcoin closed below 30,000 USD for the first time since July 2021, and crypto prices plummeted on June 13th, so Bitcoin dropped below 23,000 USD for the first time since December 2020 [21-38].

3. Related Work

Machine learning models were used to predict Bitcoin prices by Sean McNally, Jason Roche, and Simon Caton (2018) [39]. They used a dataset obtained from CoinDesk's official website which ranges data from 19 August 2013. until 19 July 2016. and includes Open, High, Low, and Close (OHLC) data. The researchers in their work concluded that LSTM achieved by far the highest accuracy, while RNN achieved the lowest RMSE and higher accuracy when compared to ARIMA Bitcoin price prediction. Also, LSTM outperformed RNN, but not significantly, needed significantly more time to train the model, while ARIMA predictions showed the best performance in terms of sensitivity, specificity, and precision results.

A research paper by Thearasak Phaladisailoed and Thanisa Numnonda (2018) [20] attempted to identify which machine learning algorithm is most efficient for predicting Bitcoin prices with the highest level of accuracy. They used 1-hour interval exchange rate in USD from January 1, 2012, to January 8, 2018, obtained from the official Kaggle website and experimented with several regression models with the sci-kit-learn library. The research indicates that deep learning models perform better than Theil-Sen regression and Huber regression. GRU gives the best results of MSE at 0.00002 and R2 at 0.992 or 99.2%. Huber regression uses a much shorter calculation time than LSTM and GRU.

A study by Wei Chen, Huilin Xu, Lifen Jia, and Ying Gao (2020) [21] attempted to predict the Bitcoin exchange rates using economic and technological determinants using Machine Learning models in their analysis. All the models were designed based on datasets obtained from the Bitcoincharts website, where data is collected via APIs and websites for the Bitcoin exchange rate which is considered as the output target.

By using economic and technological determinants, the authors have concluded that LSTM could be more effective as an early prediction tool than Artificial Neural Networks (ANN) and Random Forests (RF), both of which employed the previous exchange rate. Also, the information that is derived from economic and technological determinants is of greater value for predicting Bitcoin exchange rates than information obtained from previous exchange rates.

4. Methodology

4.1. Data Collection

By analyzing Bitcoin transaction data from the Bitstamp website, machine learning models in this study model 1-minute interval trading rates in USD between January 1, 2012, and January 8, 2022. The dataset is in CSV file. It is important to note that some values in our data set were missing or meaningless, as they were collected from APIs and websites. Therefore, the mean was calculated whenever necessary in order to replace the missing data; if the value was missing, we removed the corresponding date from the data set.

4.2. Feature Selection

Machine learning models are able to perform predictions more easily if useful patterns are extracted from data, as part of the feature selection process. Features of the dataset from our study are as follows:

- Close (latest trade);
- Open (opening trade);
- High (highest trade during the day);
- Low (lowest trade during the day);
- Weighted price (Bitcoin price);
- Volume_(BTC) (total trade volume of day in BTC);
- Volume_(Currency) (total trade volume of day in USD);
- Timestamp (data recorded time).

We created models using only Close, Open, High, and Low features when predicting Weighted prices, and the next step was to divide the data into a training set and a test set, as the ratio is 70:30. As part of data science, data splitting is essential, particularly for the creation of models based on the data, and this technique ensures that data models and processes based on data models, such as machine learning, are accurate.

4.3. Data preparation

In order to determine whether a time series is stationary, one of the most common statistical tests is the augmented Dickey-Fuller test (ADF Test). Assuming the presence of unit roots, the p-value obtained should be less than the significance level (0.05) to reject the null hypothesis. By observing the unit root in the time series, we can infer that it is stationary and that it requires a certain number of different operations to become stationary if the time series contains a unit root. The following techniques can often be used to convert a nonstationary time series to one that is stationary:

The Box-Cox transformation

As many time series models assume that the data will remain stationary, the Box-Cox transformation can be useful for transforming a non-stationary time series into a stationary one. The Box-Cox transformation is defined as:

$Y = (X^{lambda} - 1) / lambda$

where X is the original time series and Y is the transformed time series. The parameter lambda is chosen such that the transformation results in a stationary time series. As a result, the lambda value that yields the most stationary time series is chosen as the optimal transformation. In order to model a time series, the optimal transformation must be applied to the original time series.

Seasonal differentiation

As many time series models assume that data is stationary, seasonal differentiation converts a non-stationary interval into a stationary one, making it useful for time series modeling. After decomposing a time series into seasonal, trend, and residual components, the seasonal component is subtracted to obtain a stationary time series. Because seasonal differentiation can be an effective method for making a time series stationary, it is important to carefully consider whether it is the right approach.

STL seasonal decomposition

The decomposition of time series into seasonal, trend and residual components is known as the seasonal decomposition of time series (STL). This method is useful in identifying and modeling underlying patterns in time series, as well as transforming non-stationary time series into stationary ones. By removing the trend and seasonal components, the residual component can also be used to transform the original time series into a stationary one by analyzing the underlying patterns in the data.

The seasonal trend and residual components can then be used to reconstruct the original time series or to analyze the underlying patterns in the data.

In order to avoid bias caused by variable measures at different scales, it is typically necessary to apply feature-wise normalization just prior to the model fitting process, such as MinMax Scaling before model fitting occurs. During normalization, numeric columns within a dataset are scaled so they retain all their information and ranges without losing any information. When preparing data for machine learning, normalization is often applied. Some Machine Learning models can benefit from MinMax Scaling since the backpropagation is more stable and easier compared to using original unscaled data. As MinMaxScaler subtracts the minimum value from a feature, it divides it by the range, preserving the original distribution shape of the maximum and minimum.

4.4. Modeling

We used the Keras library to build ARIMA and LSTM machine learning models due to Bitcoin's continuously fluctuating price. We also compared ensemble supervised machine learning models according to the coefficient of determination and execution time, using the Pysan library in Python. The predictive performance of the models was compared using the root mean square error (RMSE), the mean absolute error (MAE), and the mean absolute percentage error (MAPE).

4.4.1. Autoregressive-Integrated Moving Average (ARIMA)

As a statistical model based on past data, the autoregressive integrated moving average (ARIMA) can help predict future values. This is a time series model that includes both autoregressive and moving average terms, as well as the difference between present and past values (the "integrated" component). An ARIMA model combines autoregressive (AR) and moving average (MA) terms to forecast the future based on past data. Using the AR component of an ARIMA model, the model is able to account for past values of the time series, while using the MA component to account for random shocks or noises in the data. In other words, the integration component allows the time series to remain stationary over time by keeping the mean and variance constant.

The ARIMA model is typically denoted by the notation ARIMA (p, d, q), in which p indicates the order of the autoregressive component, d indicates the order of the integration component, and q indicates the order of the moving average component. A number of methods can be used to estimate these parameters, such as the Box-Jenkins method or maximum likelihood estimation, and the equations are expressed as:

$$y'_t = c + \phi_1 y'_{t-1} + \ldots + \phi_p y'_{t-p} + \theta_1 \epsilon_{t-1} + \ldots + \phi_q \epsilon_{t-q} + \epsilon_t$$

It is common for time series techniques to assume that the data are stationary. There is no change in the mean, variance, or autocorrelation structure of a stationary process over time, which makes it possible to define it mathematically in precise terms. It is important to note that the series we are dealing with is flat in appearance, without a trend, constant variance and autocorrelation over time, and is free of periodic fluctuations.

4.4.2. Long Short-Term Memory (LSTM)

In contrast to other RNNs, Long Short-Term Memory (abbreviated “LSTM”) have feedback connections that enable them to handle random sequences of input, which allows them to capture long-term dependencies in data. RNNs perform poorly in handling long-term dependencies due to their inability to retain a "memory" of past inputs. The LSTM, however, is able to maintain a "memory" of past inputs with extended persistence. Each episode of training causes the network's connection weights and biases to change, similar to how synaptic strength changes in the brain store long-term memories; each time step, the activation patterns in the network change, similar to how short-term memories are stored by changes in electrical firing patterns in the brain. In LSTM architecture, RNNs can maintain a short-term memory for thousands of timesteps, thereby providing "long short-term memory”. Because LSTMs can recall past information and use it to predict the future, they are well-suited for time series forecasting.

A time series forecast can be made by using an LSTM in the following way:

- Divide time series data into training and testing sets;
- Preprocess the data, such as by normalizing the values or applying min-max scaling;
- Define the model architecture (number of layers, units in each layer, and the input and output shapes);
- Train the model on the training data using an appropriate loss function and optimization algorithm;
- Evaluate the model on the testing data;
- Use the trained model to make predictions on new data.

4.4.3. Ensemble Supervised Machine Learning Models

As a general “meta-approach” (an acronym for “most effective tactics available”) to machine learning, ensemble learning combines predictions from multiple models to improve predictive performance. In bagging, various samples of the same dataset are fitted with many decision trees and the predictions are averaged. In stacking, a number of different models are fitted to the same data, and another model is used to determine how to combine the predictions. As part of boosting, ensemble members are added sequentially to correct predictions made by previous models, resulting in a weighted average. Recent years have seen an increase in applications of ensemble learning due to the increasing computational power that allows training large ensembles within a reasonable timeframe. The LightGBM (Light Gradient Boosting Machine) framework relies on decision trees to optimize the model and reduce its memory consumption by using gradient boosting. By splitting the tree leaf-wise rather than growing it level-wise, LightGBM speeds up the boosting process by selecting the leaf with the largest delta loss to grow, and since the leaf is fixed, the leaf-wise algorithm has less loss than level-wise boosting.

A decision tree method called extra trees (short for extremely randomized trees) is used by AutoML when training with decision trees. As with random forests, the extra trees algorithm creates many decision trees, but each tree is sampled randomly, without replacement. This generates a dataset for each tree containing unique samples, and each tree receives a specific number of features from the total set of features, which are also randomly selected. Combining decision trees with a technique known as Bootstrap and Aggregation, commonly called bagging, Random Forests are capable of performing regression and classification tasks. For each model, random sampling and feature sampling is performed from the dataset to form sample datasets. This part is called Bootstrap. Random Forest uses multiple decision trees as base learning models.

The gradient-boosting regression tree derives from an ensemble method, which is based on a decision tree. Prediction results are calculated from seed to leaf, starting from tree roots, branching based on conditions, and ending at the goal leaf. When the hierarchy of a decision tree is too deep, it can result in overfitting test data.

There are many practical approaches to supervised learning, including the Decision Tree, which can be used for classifying and predicting, with the latter method being more practical in its application. In a tree-structured classifier, there are three types of nodes: the Root Node represents the entire sample, the Branch Node represents the decision rules, and the Leaf Node represents the outcome.

4.5. Proposed Method

Analysis was performed primarily with Python, but it was also used and Hadoop, a distributed data storage and processing framework that uses the map-reduce programming model to allow efficient parallel processing of Big Data. As a result of this parallel processing, large datasets can be processed quickly and efficiently, and the project was developed in a localized Linux virtual machine, which served as a userfriendly development environment for various integrated development environments (IDEs). Cleansed datasets were saved as CSVs and then pushed to Hadoop Distributed File System (HDFS) where map-reduce operations were used to group data and train a machine learning model. Several Python packages are used for analyzing and processing databases within the Python programming language, including the most famous: NumPy, SciPy, Pandas, Matplotlib, Plotly, Seaborn, Ggplot, etc.

We conducted three experiments using the cleaned dataset after splitting it into training and test segments. An ARIMA machine learning algorithm was used in the first experiment, and the results of the gained method were presented as tables and graphs after an in-depth review of the data, along with a detailed explanation of the model evaluation process and efficiency indicators. The second experiment implemented the LSTM model of the machine learning algorithm, determined the time period for when the model achieved higher efficiency, and explained the efficiency indicators of the model. As part of the LSTM model, Bitcoin's value for the next 30 days was forecasted by the model. During the last, third experiment, using the Pysan package the five best models of ensemble supervised machine learning models were analyzed and compared based on model execution time and coefficient of determination. The Proposed Method flowchart is shown in Figure 2.

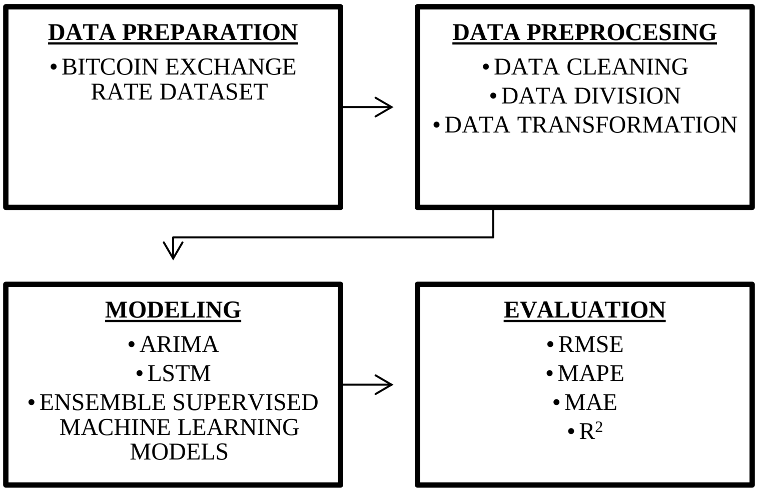

Figure 1. Proposed Method

5. Experimental Results

5.1. Experiment I

Autoregressive-Integrated Moving Average (ARIMA)

TIME SERIES TRANSFORMATON MODEL	p-value
Box-Cox Transformations (DF test)	0,998863
Seasonal differentiation (DF test)	0,444282
STL-decomposition (DF test)	0,000024

Figure 2. Time series decomposing

Upon decomposing the time series, it is apparent from the results shown in Figure 3, that in Box-Cox Transformations and Seasonal differentiation, results indicate that our sample did not provide sufficient evidence to reject the null hypothesis. So, the time series is in fact non-stationary. As a result of STL decomposition, we reject the null hypothesis, concluding that time series are stationary, and we will use this method to form our ARIMA model.

Based on the obtained results in Figure 4, we can conclude that all of the residuals are zero-mean, and a large positive residual correlate to an unexpected rise in price for the years 2013, 2014, 2017, and 2018, while a large negative residual corresponds to an unexpected decline in price for years 2015, 2019, and 2020. Based on the ACF plot of the residuals from the ARIMA model, all autocorrelations are within the threshold limits, indicating that the residuals behave like white noise, which means that the model has explained the variance well in the dependent variable, and nothing is left to extract in terms of information. Moreover, the autocorrelation (ACF) does not show any significant lag, so by examining the parameters presented, we conclude our model is suitable for prediction.

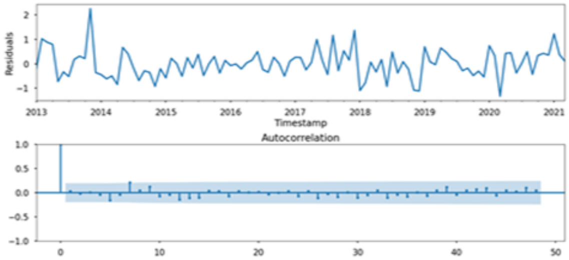

Figure 3. ACF plot

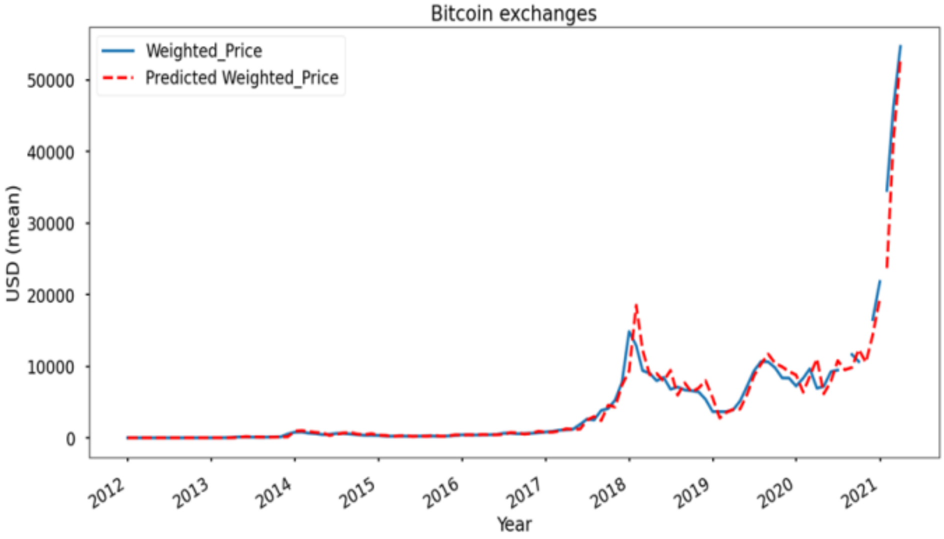

Figure 5. ARIMA Bitcoin price prediction

In Figure 5, we can clearly see that the ARIMA model accurately predicts Bitcoin prices and that they are very similar to the actual prices. An indicator of forecast error, the Mean Absolute Percentage Error (MAPE) represents the average percentage difference between the forecast and the actual value, divided by the actual value as a percentage. We estimate that our model predicts the next 15 observations with about 95.98% accuracy based on its MAPE of 4,02%. In terms of mean absolute error, the actual forecast does worse out of the sample than a naive forecast did in the sample since our MASE is 1,4520. If we assume that the out-of-sample data will be quite similar to the in-sample data (which depends on the problem at hand), then MASE>1 suggests discarding the actual forecast in favor of a naive forecast, as we can only determine how well a naive forecast performs in the sample, not out of sample. In the case of Bitcoin (or any other financial asset), there are a number of factors that can influence its price, and past performance does not necessarily predict future results. It can be challenging to create a prediction model that accounts for all of these factors. The degree of accuracy of any Bitcoin price prediction based on ARIMA or any other algorithm will depend on how well the model is built and how well the data is collected.

5.2. Experiment II

Long Short-Term Memory (LSTM)

The price of Bitcoin fluctuated drastically from 200 USD in 2014 to 22,000 USD in 2019 and 3,000 USD in 2020. Based on our LSTM model, we are predicting Bitcoin's Close Price, so we will simply take a look at a period of one year to avoid this type of fluctuation in data while just considering the Close Price and Date in the formation of the LSTM model.

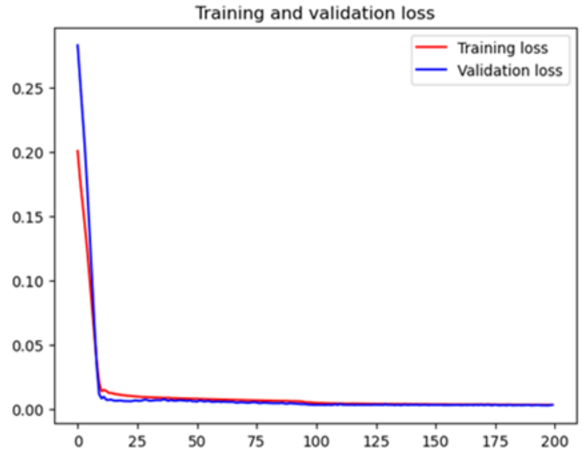

Figure 6. Training and validation loss graph

As a result of training and validation losses, we are able to gain a better understanding of how learning performance changes over time and can diagnose any problems that can lead to underfitting or overfitting models. Additionally, they will inform us about the epoch at which the trained model weights should be used for inferencing. A good fit on the 200 epoch is shown by the training and validation losses plot shown in Figure 6, so our model is ready to be used.

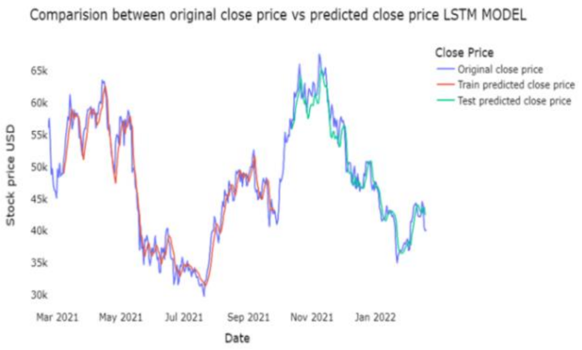

Figure 4. LSTM model price prediction

	METRIC	SCORE
TRAIN DATA	RMSE	2107.1636
	MSE	4440138.690
	MAE	1662.1254
	EXPLAINED VARIANCE	0.9495
	R2 SCORE	0.9489
TEST DATA	RMSE	2082.4879
	MSE	4336755.987
	MAE	1614.2076
	EXPLAINED VARIANCE	0.9505
	R2 SCORE	0.9469

Figure 5. LSTM model metrics

Based on the results shown in Figure 8, in a dataset, RMSE (root mean squared error) measures the difference between predicted and true values. The approach is often used for evaluating the performance of LSTM models on regression problems. An improved fit to data is indicated by a lower RMSE. Often used to solve regression problems, MSE stands for Mean Squared Error, and it is calculated by averaging the squared differences between the predicted and actual target values. It is more accurate for a model if the RMSE is closer to 0, but RMSE is calculated according to the target we are predicting, so there is no general rule for determining what is considered a good metric value because it cannot evaluate metric value outside of the context of the dataset working in. Since Bitcoin prices, and in general, the prices of cryptocurrencies, tend to fluctuate greatly, an RMSE of 2,000 is most likely considered good for a Bitcoin price prediction model, which tends to be over 65,000 USD, and from our research, this is considered an acceptable value for a good model. The same indicators are MSE and MAE which also show that it is a good prediction model.

An explained variance score measures the amount of variance the regression model explains in the target variable. This score ranges from 0 to 1, with a higher score indicating that the model is more accurate. It is calculated by comparing the sum of square residuals with the total sum of squares for the target variable. R2 measure is known as the coefficient of determination and is the ratio between the squares of residuals and the squares of the target variable. It is based on the number of squares of the residuals of the regression model divided by the squares of the target variable. A higher score indicates a better fit, and it ranges between 0 to 1.

Our predictive model has an accuracy rate of around 95%, based on the indicators R2 and explained variance score. When we consider that it aims to predict prices that undergo significant fluctuations in a short period of time, as well as those that are affected by external market and economic factors, this is more than good. According to important metrics, our model has a good predictive ability, which can be clearly seen on the chart shown in Figure 7, so we predicted the movement of Bitcoin prices over the next 30 days. Based on our model, a slight increase in Bitcoin value is predicted after the sudden drop in Bitcoin value, specifically within the next 15 days, as shown in Figure 9.

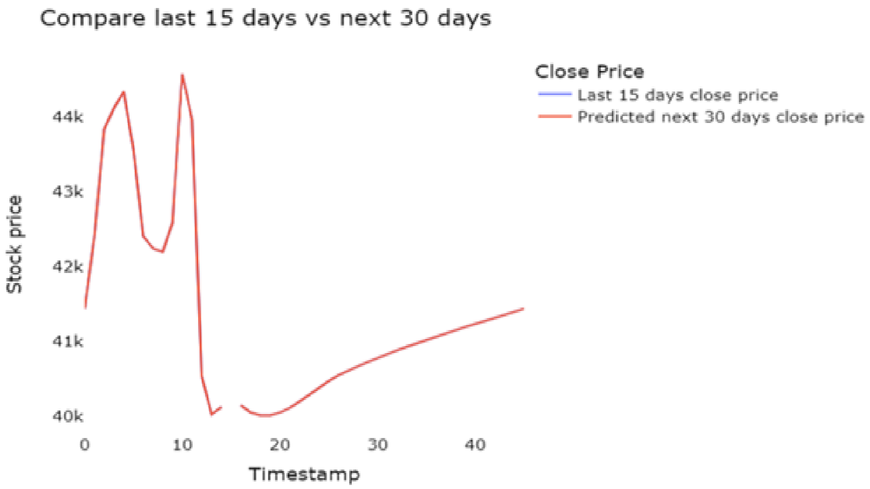

Figure 9. Next 30 days LSTM prediction

5.3. Experiment III

Ensemble Supervised Machine Learning Models

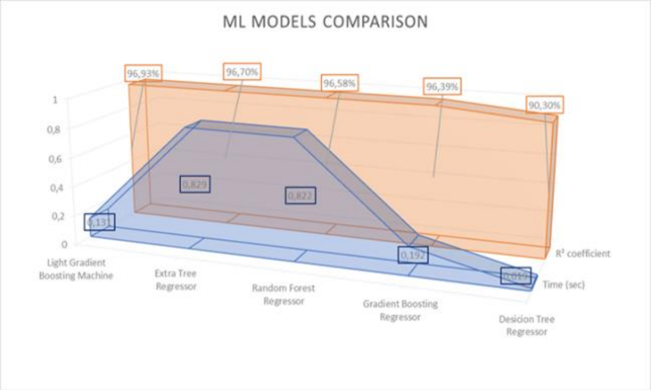

Figure 6. Ensemble supervised machine learning models

Figure 7. Ensemble supervised ML models performance

MODEL	EXECUTION TIME	R² coefficient
Light Gradient Boosting Machine	0,131s	96,93%
Extra Tree Regressor	0,829s	96,70%
Random Forest Regressor	0,822s	96,58%
Gradient Boosting Regressor	0,192s	96,39%
Desicion Tree Regressor	0,019s	90,30%

A model with good accuracy and fast execution time is generally considered to be a high-performing model. Based on our research, as shown in Figure 11, and also graphically shown on the chart in Figure 10, the Light Gradient Boosting Machine method has the highest coefficient of determination, whereas Decision Tree Regressor has the fastest execution time, but the worst coefficient of determination. It is important to note, however, that there is often a trade-off between accuracy and execution time, and finding the right balance will depend on the specific requirements of the research application. Since Bitcoin's value can be affected by a number of factors, including market demand, regulatory changes, and global economic conditions, it is difficult to accurately predict its price, as it is subject to a number of factors that can affect its value. Aside from being highly volatile, the Bitcoin price can fluctuate significantly within a short period of time. Consequently, it is impossible to make reliable predictions of Bitcoin or any other cryptocurrency prices with a high degree of certainty, and investors should always be aware of the potential risks and uncertainties involved.

6. Discussion

The RNN and LSTM deep learning models, according to Sean McNally, Jason Roche, and Simon Caton (2018), are significantly more efficient at predicting Bitcoin prices than ARIMA models, which are better at recognizing long-term relationships. The LSTM prediction achieved the highest accuracy and the RNN prediction achived lowest RMSE, while the ARIMA prediction did poorly. As a result of our research, which was carried out within three experiments, we concluded that ARIMA showed a higher accuracy in predicting Bitcoin price when applied to longer time frames compared to the LSTM. As a result of using LSTM algorithm, we achieve significant high levels of accuracy when predicting Bitcoin price over a shorter period of time, in our case it is about one year, where 95% accuracy was achieved. Considering Bitcoin's fluctuating price, price prediction in a short period of time isn't a huge advantage and help to investors, so it isn't much of a help to them. The ARIMA algorithm has a greater advantage because when it comes to a longer period of prediction, it has a higher prediction accuracy, so investors are able to make timely decisions about making or withdrawing their investments. The results of this study demonstrate that machine learning models can correctly predict short-term and long-term Bitcoin market movements. Furthermore, we predicted the future price of Bitcoin with LSTM algorithm in a short time interval (30 days) successfully, which can be very useful for future investors interested in short-term investments.

7. Conclusion

As the most valuable cryptocurrency, Bitcoin has been highly fluctuating in recent years, making it difficult to predict its prices, and has experienced some major fluctuations in price since its launch in 2009. However, despite these fluctuations, Bitcoin has gained widespread acceptance and adoption as a digital asset and a store of value. A growing interest in Bitcoin and other cryptocurrencies from institutional investors, governments, and mainstream financial institutions has led to more stable prices and a broader range of applications. The number of cryptocurrency users in the world is estimated at 295 million, according to Debthammer, and Forbes estimates that there are over 20,000 cryptocurrency projects around the globe. Bitcoin's future depends on a number of factors, including continued adoption, regulatory developments, and competition from other cryptocurrencies. However, others caution that the market is still highly speculative and subject to significant risks, although some experts predict the value of Bitcoin will continue to increase as it becomes a widely accepted alternative to traditional currencies. In this study, we intend to identify the most decent and most efficient model for predicting the price of bitcoin, based on a variety of machine learning algorithms. According to the results obtained it is evident that machine learning models, such as ARIMA and LSTMs, are effective in predicting Bitcoin, with the ARIMA providing a better capability for recognizing longer-term relationships among Bitcoin prices. Due to the high variance of this task, achieving impressive validation results can be difficult, so in consequence, it remains a challenging undertaking, where it is important to remain mindful of the fine line between overfitting and a underfittngm model. There was a marginally better performance from the LSTM than the ARIMA, but not significantly. However, the LSTM has a smaller learning curve. According to both models, accuracy is similar, and ARIMA and LSTM predict the next value with an accuracy of about 95%. Regarding ensemble supervised machine learning models the Light Gradient Boosting Machine method has the highest coefficient of determination (96,93%), whereas the Decision Tree Regressor has the fastest execution time (0,019 sec), but the worst coefficient of determination (90,30%). In terms of accuracy and execution time, it is important to remember that there is often a trade-off between the two, and determining the right balance depends on the research application itself. Research on upcoming advanced methods can further enhance this work's analysis of forecasting prices, enabling a more comprehensive picture to be obtained. The paper's primary objective is to adjust the volatility of prices and forecast Bitcoin prices accurately.

8. Future Work

As part of our research paperwork, we utilized advanced forecasting methods, including the application of an ARIMA (Auto-Regressive Integrated Moving Average) model to capture the temporal patterns in Bitcoin prices. We also employed machine learning techniques such as long short-term memory (LSTM) networks to capture complex patterns and dependencies in the data. Additionally, we utilized ensemble methods, such as bagging, boosting, or stacking, to combine multiple forecasting models and enhance prediction accuracy.

In the future, conducting further research and incorporating additional variables may be necessary to enhance Bitcoin price prediction. One possible approach is to implement sentiment analysis, where social media, news articles, and other textual data are analyzed to extract sentiment related to Bitcoin. This analysis can provide valuable insights into market sentiment, which has the potential to impact Bitcoin prices. Additionally, fundamental analysis can be considered, which involves incorporating factors such as network usage, transaction volume, adoption rate, regulatory changes, and macroeconomic variables that can influence the value of Bitcoin.

To improve the analysis, it is possible to add new variables that would enhance accuracy. These variables can include relevant market indicators such as trading volume, liquidity, volatility indices, or the performance of other cryptocurrencies. Additionally, incorporating technical indicators like Moving Averages can be beneficial. Moving Averages calculate different types of averages to capture trends and momentum in Bitcoin prices. Another useful indicator is the Relative Strength Index (RSI), which measures the speed and change of price movements, providing insights into potential overbought or oversold conditions and indicating potential price reversals. It is also worth considering macroeconomic indicators such as inflation rates, interest rates, GDP growth, or unemployment rates, as these factors can indirectly influence cryptocurrency prices.

Improving the volatility adjustment of Bitcoin prices is important for accurately capturing the inherent volatility of the cryptocurrency. A number of potential avenues should be considered, including Stochastic Volatility Models, such as the Heston model or the SABR model, explicitly capture the dynamics of volatility as a separate process. These models can provide more accurate volatility adjustments for Bitcoin prices. Also, consider employing regime switching models, such as Markov switching models or Hidden Markov models, to capture different volatility regimes in Bitcoin prices. These models can adapt to changes in market conditions and adjust volatility estimates accordingly. There is also a possibility of use option pricing models, such as the Black-Scholes model or its extensions, to estimate implied volatility from Bitcoin options. Implied volatility reflects market participant expectations of future volatility and can provide additional insights into volatility adjustments.

References

[1]. B. B. Mandelbort, “When can price be arbitraged efficiently? A limit to the validity of the random walk and martingale properties, “Review of Economic Statistics, vol.53, no.3, pp.225–236, 1971.

[2]. E. F. Fama & K. R. French, “Permanent and temporary components of stock prices,” Journal of Political Economy, vol.96, no.2, pp.246–273, 1988.

[3]. A. W. Lo & A. C. Mackinlay, “Stock market prices do not follow random walks: Evidence from a simple specification test,” Review of Financial Studies, vol.1, no.1, pp. 41–66, 1988.

[4]. W. Brock, J. Lakonishok & B. L. Baron, “Simple technical trading rules and the stochastic properties of stock returns,” Journal of Finance, vol.47, no.5, pp.1731–1764, 1992.

[5]. S. Nadarajah & J. Chu, “On the inefficiency of Bitcoin,” Economics Letters, vol.150, no. C, pp.6–9, 2017.

[6]. P. Ciaian, M. Rajcaniova & d’Artis Kancs, “The economics of Bit Coin price formation,” Applied Economics, vol.48, no.19, pp.1799-1815, 2016

[7]. A. S. Hayes, “Cryptocurrency value formation: an empirical study leading to a cost of production model for valuing Bitcoin,” Telemat Information, vol.34, no.7, pp.1308-1321, 2016.

[8]. M. B. Taylor, Bitcoin and the age of bespoke silicon, In Proceedings of the International Conference on Compilers, Architectures and Synthesis for Embedded Systems, CASES’13. Piscataway, NJ, USA: IEEE Press, pp.1-10, 2013.

[9]. I. Magaki, M. Khazraee, L. V. Gutierrez et al., “ASIC clouds: specializing the datacenter. In Proceedings of the 43rd International Symposium on Computer Architecture, ISCA’16, pp.178-190, 2016.

[10]. O’ Dwyer KJ, Malone D. Bitcoin mining and its energy footprint,” Irish Signals & Systems Conference and China–Ireland International Conference on Information and Communications Technologies IET; pp.280-285, 2014.

[11]. H. Vranken, “Sustainability of Bitcoin and blockchains,” Current Opinion in Environmental Sustainability, Applied Economics, vol.28, pp.1–9, 2017.

[12]. P. Franco, “Understanding Bitcoin: Cryptography,” Engineering and economics, John Wiley & Sons, ISBN: 978-1-119-01916-9, pp.1-288, 2014.

[13]. S. Ranjan, I. Singh, S. Dua, et al., “Sentiment analysis of stock blog network communities for prediction of stock price trends,” Indian J Finance, vol.12, no.12, 2018.

[14]. F. Valencia, A. G. Espinosa, B. V. Aguire, “Price movement prediction of cryptocurrencies using sentiment analysis and machine learning,” Entropy vol.21, no.6, pp.1–12, 2019.

[15]. V. S. Pagolu, K. N. Reddy, G. Panda, et al., “Sentiment analysis of twitter data for predicting stock market movements,” Entropy, pp.1345-1350, 2016.

[16]. J. Abraham, D. Higdon, J. Nelson, et al., “Cryptocurrency price prediction using tweet volumes and sentiment analysis,” International Conference on Information and Communications Technologies, vol.1, no.3, pp.1-21, 2018.

[17]. S. Bird, E. Klein & E. Loper, Natural language processing with python: analyzing text with the natural language toolkit. O’Reilly Media USA, ISBN: 978-0-596-51649-9, pp.1-463, 2009.

[18]. S. Galeshchuk, O. Vasylchyshyn, A. Krysovatyy, “Bitcoin response to twitter sentiments. ICTERI workshops, pp.1-9, 2018.

[19]. B. Gunter, N. Koteyko, D. Atanasova, “Sentiment analysis: a market-relevant and reliable measure of public feeling,” Int J Mark, vol.56, no.2, 2014.

[20]. T. Phaladisailoed & T. Numnonda. Machine Learning Models Comparison for Bitcoin Price Prediction,” 10th International Conference on Information Technology and Electrical Engineering (ICITEE), pp.506-511, 2018.

[21]. W. Chen, H. Xu, L. Jia, et al., “Machine learning model for Bitcoin exchange rate prediction using economic and technology determinants,” International Journal of Forecasting, vol.37, no.1, pp.28-43, 2021.

[22]. S. A. Fattah, “Time Series Modeling for U.S. Natural Gas Forecasting, “Proceedings of International Petroleum Technology Conference IPTC 10592, 2015.

[23]. R. Yu, G. Xue, V. T. Kilari, et al., “Coin Express: A Fast Payment Routing Mechanism in Blockchain-Based Payment Channel Networks,” 27th International Conference on Computer Communication and Networks (ICCCN) pp.1-9, 2018.

[24]. S. A. Rakhshan, M. S. Nejad, M. Zaj, et al., "Global analysis and prediction scenario of infectious outbreaks by recurrent dynamic model and machine learning models: A case study on COVID-19," Computers in Biology and Medicine, vol.158, 2023.

[25]. L. D. Agati, Z. Benomar, F. Longo, et al., IoT/Cloud-Powered Crowdsourced Mobility Services for Green Smart Cities," IEEE 20th International Symposium on Network Computing and Applications (NCA), pp.1-8, 2021.

[26]. R. K. Rathore, D. Mishra, P. Singh Mehra, et al., "Real-world model for bitcoin price prediction," Information Processing & Management, vol.59, no.4, 2022.

[27]. A. Dutta, S. Kumar, M. Basu, "A Gated Recurrent Unit Approach to Bitcoin Price Prediction," Journal of Risk and Financial Management, vol.13, no.2, 2020.

[28]. J. J. Jaber, R. S. Alkhawaldeh, S. M. Alkhawaldeh, et al., "Chapter 133 Predicting Bitcoin Prices Using ANFIS and Haar Model," Springer Science and Business Media LLC, vol.1056, 2023.

[29]. A. Ekstrom & H. Sorenason, "Statistical analyses," The R Primer, pp.1-86, 2011.

[30]. K. Kamiya, K. Shimizu, A. Igarashi, et al., "Intraindividual comparison of changes in corneal biomechanical parameters after femtosecond lenticule extraction and small-incision lenticule extraction," Journal of Cataract and Refractive Surgery, vol.40, no.6, pp.963-970, 2014.

[31]. A. Aljadani, "DLCP2F: a DL-based cryptocurrency price prediction framework," Discover Artificial Intelligence, vol.2, no.1, 2022.

[32]. L. Raju, G. Sowmya, S. Srividhya, et al., "Advanced Home Automation Using Raspberry Pi and Machine Learning," 7th International Conference on Electrical Energy Systems (ICEES), pp.600-605, 2021.

[33]. S. A. Reddy, S. Akashdeep, R. Harshvardhan, et al., "Stacking Deep learning and Machine learning models for short-term energy consumption forecasting," Advanced Engineering Informatics, vol.52, 2022.

[34]. B. Panda, K. A. Haque, "Extended data dependency approach," Proceedings of the ACM symposium on Applied computing, pp.446-452, 2002.

[35]. N. R. Shetty, L. M. Patnaik, N. H. Prasad, "Emerging Research in Computing, Information, Communication and Applications," Springer Science and Business Media LLC, vol.2, 2022.

[36]. A. Abdelsamea, A. A. El. Moursy, E. E. Hemayed, et al., "Virtual machine consolidation enhancement using hybrid regression algorithms," Egyptian Informatics Journal, vol.18, no.3, pp.161-170, 2017.

[37]. M. J. Crawley, "Time Series Analysis," The R Book, ISBN 978-0-470-97392-9 2007.

[38]. E. Akyildirim, A. Goncu, A. Sensoy. "Prediction of cryptocurrency returns using machine learning," Annals of Operations Research, vol.297, pp.3-36, 2020.

[39]. S. M. Nally, J. Roche, S. Caton. "Predicting the Price of Bitcoin Using Machine Learning," 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP), pp.339-343, 2018.

Corresponding author: Adnan Branković
Institution: International Burch University, Faculty of Engineering, Natural and Medical Sciences, Department of Information Technologies
E-mail: adnan.brankovic@stu.ibu.edu.ba