Baltic J. Modern Computing, Vol. 13 (2025), No. 1, 1-11
https://doi.org/10.22364/bjmc.2025.13.1.01

Evaluation of the Model for Bitcoin Price Prediction Using Machine Learning Algorithms and Blockchain Technology

Mimoza MIJOSKA, Blagoj RISTEVSKI

University St. Kliment Ohridski - Bitola, Faculty of Information and Communication Technologies, ul. Partizanska nn, 7000 Bitola, North Macedonia

mijoska.mimoza@uklo.edu.mk, blagoj.ristevski@uklo.edu.mk

ORCID 0000-0002-4248-2760, ORCID 0000-0002-8356-1203

Abstract: Blockchain technology can be used to analyze and process data through the effective integration of financial resources. Likewise, machine learning is one of the most notable technologies in recent years. Both technologies are data-driven, and therefore there is a rapidly growing interest in integrating them for more secure and efficient data sharing and analysis. This paper shows how these two technologies, blockchain technology and machine learning, can be combined to predict bitcoin volatility. To analyze and predict the volatility of bitcoin, real-time series bitcoin data was used, and the random forest algorithm was utilized. To evaluate the model, the following statistical errors were analyzed: mean absolute error, root mean square error, mean absolute percentage error, median absolute percentage error and symmetric mean absolute percentage error in cases using the different split ratios of the training and test sets. The obtained results have shown that the prediction model is well-designed.

Keywords: blockchain technology, machine learning, random forests, bitcoin volatility, statistical errors

1. Introduction

Businesses are often streamlined and enhanced by the emergence and applicability of new technological advancements. Blockchain is one of those technologies which is bringing a paradigm shift in our various old and traditional business models.

Blockchain technology was introduced in 2008 with the publication of Satoshi Nakamoto's paper - "Bitcoin: a peer-to-peer electronic cash system" (Nakamoto, 2008). Blockchain technology was first used in the cryptocurrency Bitcoin. The first Bitcoin transactions took place in January 2009. Apart from their use in the economic domain, bitcoin and blockchain technology solve an important problem in informatics and computer technology that has been an obstacle to building a functional digital monetary system for years. With this technology, the problem of double use is solved, i.e. the risk that the cryptocurrency can be used two or more times is eliminated. Virtual currency developers must prevent users from being able to spend their funds more than once. The interest of enterprises, industries and governments around the world in blockchain technology is high, as the application of this technology is much larger than the domain of cryptocurrencies.

In 2014, a consortium called R3 was founded to start research and development of blockchain technology. In March 2017, this group counted about 75 companies, and 200 companies in March 2018, to reach 400 companies in March 2022 (Lukić, 2016) (Kramer, 2019). The formation of such a strong corporation with a lot of research and implementation of blockchain technology, especially in the financial sector, indicates that a new era in the development of banking is coming.

This paper describes the calculation of Bitcoin's realized volatility and discussion of the obtained results. The remainder of the paper is organized as follows. Section 2 highlights the principles of blockchain technology. In the next section, machine learning algorithms are described with particular emphasis on the random forest algorithm used in the research. Section 4 describes the calculation of Bitcoin's realized volatility. The discussion of the obtained results of this research is presented in the fifth section. In the last section are given concluding remarks and directions for further works.

2. Blockchain technology

"Block-chain" is a coined word made up of the words "block" and "chain". Blockchain is a distributed replicated database organized in the form of a single linked list - a chain, where nodes are blocks of transaction data. To connect the blocks, cryptographic algorithms, namely a hash function, are used in such a way that it is impossible to change the content of one block without changing the content of all the blocks that follow it. This is a very important feature of blockchain technology, as it ensures the immutability of the data entered into it. Blockchain technology enables digital transactions without intermediaries.

In 2008, several powerful financial institutions and insurance companies in the United States were on the verge of bankruptcy. These circumstances led to the immediate intervention of the federal government, to avoid a domestic and possibly global financial collapse (Senbet and Wang, 2012).

These events illustrate the dangers of living in a digital, connected world that depends on intermediaries to generate transactions and makes people vulnerable to digital exploitation and crime. It is an academic challenge to create a digital infrastructure for disbursal, without intermediaries, that has no corrupt or error-prone central authority and is secure and trustworthy. In a blockchain, ledgers are distributed across the entire network and there is no need for an intermediary to be in the middle of a transaction. The technology maintains multiple copies of data, similar to a peer-to-peer file-sharing system. Each node gets a copy of the entire database (Ashurst and Stefano, 2021).

Blockchain technology is a type of distributed ledger. Bitcoin blockchain technology uses Proof-of-Work Mining (PoW), which is the oldest publicly proven method used to achieve distributed consensus (Zhang et al., 2020).

The concept of the actual software architecture of blockchain technology is explained by breaking the concept of blockchain into two separate components - block and chain. A block can be thought of as a data container. In the case of the Bitcoin blockchain, each block contains data (such as Bitcoin transactions), block headers, block identifiers and Merkle trees (Vujičić et al., 2018). A block is a set of data that is collected and processed to fit into it through the mining process. Each block is identified through a cryptographic hash and a time stamp. When a new block is formed, it will contain the hash of the previous block, enabling blocks to form a chronologically ordered chain from the first block ever generated in the entire blockchain (also called a "genesis block") to the newly formed block. This process is repeated over and over to develop and maintain the network (Mijoska and Ristevski, 2021).

Blockchain is a technology that is constantly evolving. The most common types of blockchains are public blockchain (Gu, 2018), private blockchain and hybrid blockchain (Samuel, 1967).

3. Machine learning algorithms

Machine learning is a scientific field that allows computers to learn without being explicitly programmed (Géron, 2017).

Well-known machine learning algorithms are the random forests, k-nearest neighbors (k-NN) algorithm, artificial neural networks, support vector machines, the Naïve Bayesian classifier, etc. Machine learning algorithms come in many forms and can be classified according to the amount and type of supervision they receive during training. There are three main categories: supervised learning (Kotsiantis et al., 2006), unsupervised learning (Géron, 2017), and reinforcement learning (Géron, 2017).

In supervised learning, the desired output for the model is already known. It is presented with only an input example and has to learn to produce the predicted output (Liu and Xiaoguang, 2021).

3.1. Random forests

The Random Forest is a supervised learning algorithm used for both regression and classification. It is among the most popular machine learning algorithms due to its high flexibility and easy implementation. Consisting of multiple decision trees, just as a forest has many trees, each tree represents one vote in most decisions. Coincidence in this algorithm is used to improve its accuracy and reduce overload, which can be a huge question for such a sophisticated algorithm. These algorithms make a decision based on a random selection of data samples and receive predictions from each tree. After that, they choose the best sustainable solution through votes. The purpose of this method is to reduce the variance of the final model. It’s certainly one of the most sophisticated algorithms as it builds on the functionality of decision trees. Assuming your dataset has “*m*” features, the random forest will randomly choose *k* features where $k<m$. Now, the algorithm will compute the root node among the k features by selecting a node that has the highest information gain (Vadapalli, 2021).

After that, the algorithm splits the node into child nodes and repeats this process *n* times. Now you have a forest with *n* trees. Finally, you’ll perform bootstrapping, i.e, combine the results of all the decision trees present in your forest. Technically, it is an ensemble algorithm. The algorithm generates the individual decision trees through an attribute selection indication. Every tree relies on an independent random sample. In a classification problem, every tree votes and the most popular class is the final result. On the other hand, in a regression problem, you'll compute the average of all the tree outputs and that would be your final result (Vadapalli, 2021).

4. Realized volatility

Realized volatility is defined as an estimate of the variation in returns for an investment product over a defined period, by analyzing its historical returns. An evaluation of the degree of uncertainty and/or possible financial loss/gain from an investment in a business can be calculated using volatility/variability in the entity's share prices. The most common method of estimating variability in statistics is by calculating the standard deviation, i.e. variation in values from the mean. The realized volatility or actual volatility in the market is caused by two components - a continuous volatility component and a jump component, which influence the stock prices. Continuous volatility in a stock market is affected by intra-day trading volumes. For example, a single high-volume trade transaction can introduce a significant variation in the price of an instrument (Chauvet et al., 2010).

This paper evaluates the model that predicts the price of bitcoin. High variance intra-day data is used by analysts to estimate hourly/daily/weekly or monthly frequency levels. The resulting data can be used to estimate the volatile movement of sales. Analysts use high-variance daily data to estimate hourly/daily/weekly or monthly frequency levels. The data can then be used to estimate volatile sales movement. During the analysis, data whose frequency is 1 hour from the Gemini platform were taken (see WEB, a) and then using that data, the achieved volatility is calculated with a daily frequency. The Gemini exchange tracks and creates files for daily, hourly and minute data on the prices of the time series for the physical market for pairs, US dollar (USD) and the most popular cryptocurrencies such as bitcoin, etherium, lightcoin and others. Each file can be downloaded in .csv format. There are OHLC (Open/High/Low/Close) pricing data in each file that is updated daily. For this paper, granular hourly data are taken back to the 2015 year, for the market price of the pair of bitcoin/dollar.

Achieved volatility is measured by calculating the standard deviation from the average price of the asset over a given period. Since volatility is non-linear, the realized variance is first calculated by converting values taken from the stock market into logarithmic values and measuring the standard deviation of the log-normal returns. The achieved variance is calculated by calculating the sum of the squares of the standard deviation. The achieved volatility is calculated as the square root of the achieved variance (Yokuma and Armstrong, 1995).

To calculate the achieved volatility of bitcoin, an application was created in the programming language R (Mijoska et al., 2022).

A date sequence is then added using the seq() function, which can generate the general or regular sequences from the given inputs, defining the start and end time points with a frequency of 1 hour. In the time series "08.10.2015 13:00:00" is taken as the starting date, and "12.01.2022 12:00:00" is taken as the end date. The price of bitcoin was taken at the close of the calculations. To enhance the accuracy of the results, the logarithmic values of the bitcoin price are calculated.

For the needs of the research, a free data set is downloaded from the website (see WEB, b) in .csv format for the last 3 years for 9 features of bitcoin. The following characteristics were used: miner revenue divided by the number of transactions, miner revenue as a percentage of transaction volume, the total estimated USD value of blockchain transactions, total USD value of block rewards and transaction fees that are paid to miners, the total number of confirmed transactions per day, the average number of transactions per block in the past 24 hours, the total value of all outgoing transactions per day, the total USD value of trading volume on major Bitcoin exchanges, and the total value in USD on all transaction fees paid to miners.

To prepare the data for machine learning it is necessary to pre-process data. Normalization is a crucial step in data pre-processing for any machine learning application and model fitting. The algorithm will be more affected by the high-end values if the data is not transformed. This means that they will probably be more accurate in predicting high values than low values. The min-max normalization method was chosen for the data used in this research (Chauvet et al., 2010).

A machine learning algorithm is used to predict the realized bitcoin volatility. The random forest algorithm is chosen, which is included in the ensemble's learning methods. Ensemble learning is a type of supervised learning technique where the basic concept is to generate several training models and then simply combine their output rules or their Hx hypothesis, construct a strong model that works very well, does not overload and also balances bias and variance Bias-Variance Tradeoff. The idea is that instead of creating a single complicated and complex model that could have a large variance that leads to overload or be too simple and have a large bias that leads to insufficient fit, many training models can be generated in the training set, which eventually combine. One such technique is the random forest, which is a common joining technique used to improve the predictive outcome of decision trees by averaging them to reduce tree variance. In this algorithm, only a random subset of m predictors is used whenever we split into a training set and a test set. The number of randomly selected variables to create each tree is the main setting parameter in random forests. Turning off some of the predictors makes sense, as the result would be that each tree uses different predictors. This implies that 2 trees generated on the same training data will have randomly different variables selected in each division so that the trees will be unrelated and independent of each other. The final result of the ensemble model is determined by counting the majority of votes from all decision trees. This concept is known as bagging. Since each decision tree takes a different set of training data as input, deviations in the original training data set do not affect the final result obtained by aggregating the decisions from each tree. Therefore, bagging as a concept reduces the variance without changing the bias of the complete ensemble (Jacobucci, 2018).

The Bitcoin volatility prediction algorithm uses the forecastML package in the R programming language. When using machine learning algorithms, the model is first generated using training data, and then the test data values are predicted. After the data preprocessing is completed, the research continues by dividing the data into a training and a test set. The data used to predict the volatility of bitcoin contain 1095 observations, starting from 14.01.2019 to 12.01.2022. In the initial analysis, the first 995 observations are taken as the training set and the remaining 100 observations are used as the test set.

The forecasting method uses three different forecasting horizons in the initial analysis. These different horizons are used to be able to predict in the short and long term, to combine the predictions in the final forecast and thus minimize the error (Aristeidou, 2020). The function randomForest(), which is used for classification and regression and also can be used for assessing proximities among data entries, is then defined with its arguments. The first step in the prediction process is to create some validation windows to perform nested cross-validation. Next, we train our model and present the predictions, residuals, and some error metrics. It is then predicted on the test set using the validation windows and the actual versus predicted values are displayed. In the beginning, the size of each forecast horizon is defined (Kumar, 2019). Horizon is an argument of the create_lagged_df() function that creates the training model and prediction dataset. This function creates a list of datasets with lagged, grouped, static and dynamic features to train a forecasting model for specific forecast horizons or to predict the future with a trained model. A horizon represents a numeric vector of one or more forecast horizons, measured in rows of data. If dates are given, a horizon of value 1 would equal 1*frequency in calendar time (Chamorro-Courtland, 2021).

5. Discussion of the obtained results

In this paper, an analysis of the results obtained with tests was made in the case when the training and testing sets are divided on a precisely determined date of the time series, and a second case when the data to be taken in the training and testing set are randomly selected with the function sample.split() from the caTools library, when predicting bitcoin market price volatility. This function is used to partition a dataset into training and testing sets for model building. Analyzes of different situations are made, in which the mean absolute error (MAE) (Dewi and Rung-Ching, 2019), root mean square error (RMSE) (Kreinovich, 2014), mean absolute percentage error (MAPE) (Willmott and Matsuura, 2005), median absolute percentage error (MDAPE) (Chai and Roland, 2014) and symmetric mean absolute percentage error (sMAPE) (Khair, 2017) is measured to properly evaluate the accuracy of the prediction model.

In this paper, an analysis of the results obtained using a different division of the training and testing sets was made, namely for the following ratios 90 %: 10 %, 80 %: 20% and 66.6 %: 33.3%.

Standard errors were examined first in the training set using validation windows. From the obtained results it can be concluded that the mean absolute error (MAE) and root mean square error (RMSE) values do not change significantly by changing the ratio of training and testing set divisions. The smallest value for MAE = 0.017 and for RMSE = 0.021 when training and test datasets are randomly selected from the given dataset.

Next, the standard errors are analyzed in the testing set, using validation windows in predicting bitcoin volatility using a different split of the training and test sets and a different choice of split method. It can be concluded that the mean absolute error (MAE) and root mean square error (RMSE) get the smallest values when the training and test sets split ratio is 90 %: 10% with a fixed split of the time series data MAE = 0.009, and RMSE = 0.011.

In the next step, the standard errors are analyzed in the training set, without using validation windows when predicting bitcoin volatility. From the obtained results, it can be concluded that the value of mean absolute error (MAE) and root mean square error (RMSE) are almost identical when using a different division of the training and testing set, for the ratio 90 %:10 %, 80 %: 20 % and 66.6 %: 33.3 %, regardless of the way of partitioning the training and testing sets with a value of 0.006 and 0.010, respectively.

Figure 1 compares the mean absolute error (MAE) and root mean square error (RMSE) using a different split of the training and test sets and a different choice of split method, in the test set without using validation windows in predicting bitcoin volatility.

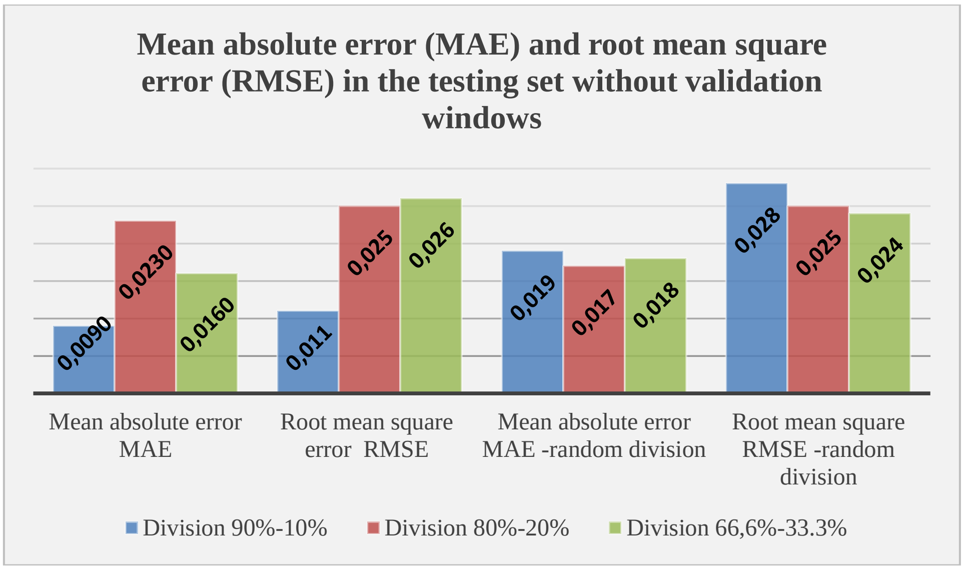

Figure 1. Mean absolute error (MAE) and root mean square error (RMSE) in the test set without validation windows in situations using a different split of the training and test sets.

From the chart in Figure 1, it can be concluded that the mean absolute error (MAE) and root mean square error (RMSE) have the smallest values when the training and test set split ratio is 90 %: 10 % with a fixed split of the time series data, MAE = 0.009 and RMSE = 0.011.

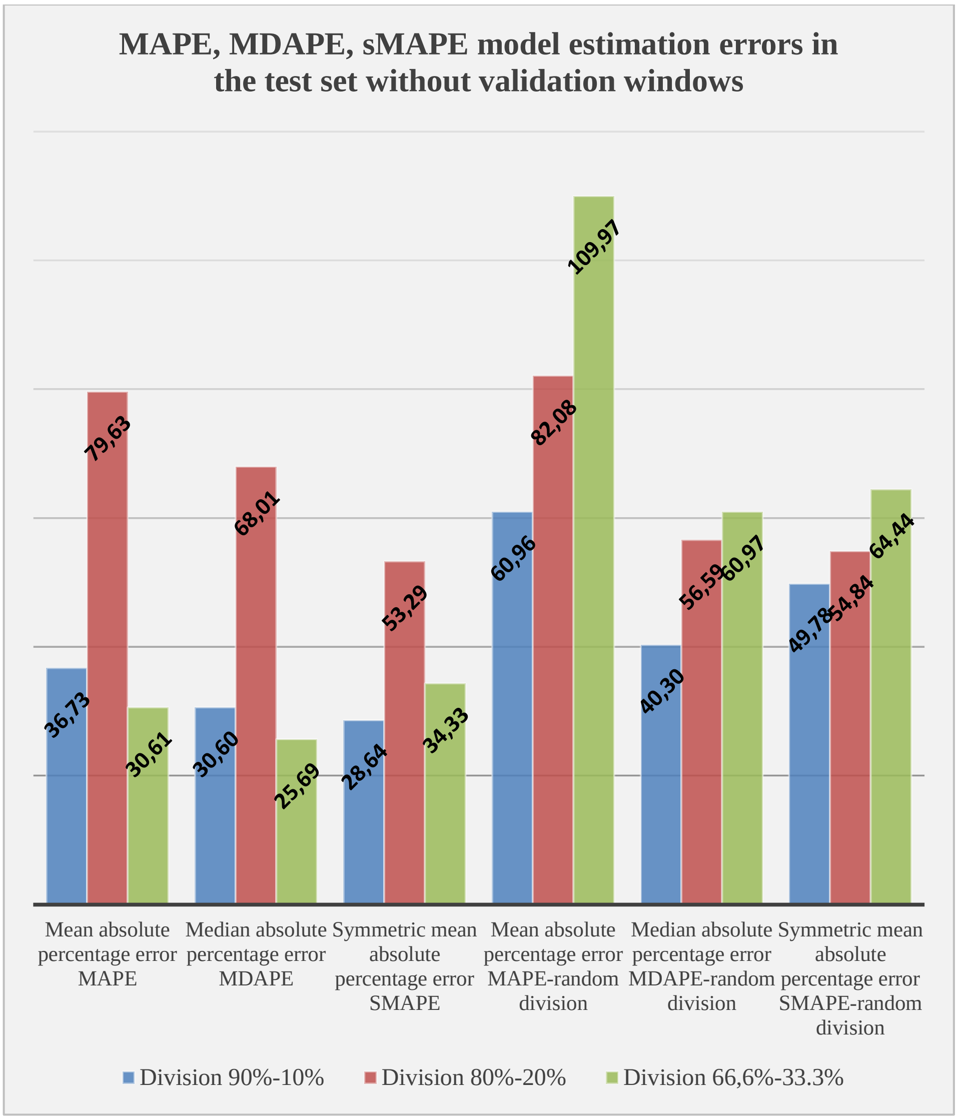

Figure 2. Mean absolute percentage error (MAPE), median absolute percentage error (MDAPE) and symmetric mean absolute percentage error (sMAPE) in situations using different partitioning of the training and test sets

From the chart shown in Figure 2, it can be concluded that the absolute percentage error (MAPE) and the median absolute percentage error (MDAPE) have significantly lower values when a fixed division of the training and testing sets is used and a selected ratio of 66.6 %: 33.3 % and thus it is concluded that this prediction model is better.

The performance metrics results are enhanced and comparable, compared with the results obtained by using two models Gated Recurrent Unit (GRU) and Bidirectional Long Short-Term Memory (BiLSTM) (Aljadani, 2022).

6. Conclusion

Blockchain technology has the potential to revolutionize the underlying payment system technology and credit information systems in banks, significantly upgrading and transforming them. The traditional system is incomplete because there is no way to prevent double-spending of money. To solve this, there is a peer-to-peer network that uses a "proof-of-work" algorithm to keep a public history of transactions. For an attacker to be able to change nodes is computationally almost impossible if honest nodes control a majority of the network.

This paper used a random forest machine learning algorithm to predict time series of realized fluctuations in the stock market price of Bitcoin and investigated whether blockchain information could be used to predict the volatility and price of Bitcoin. Many people in the world use Bitcoin as an investment because of its high volatility and in this way, they can get huge profits and losses in a short time.

In this paper, the volatility of the market price of bitcoin is modeled as a basis for measuring the risk factor in financial services using blockchain technology. Predicting the change in the value of bitcoin improves the operation of financial services, reduces the risk factor when investing, working on stock exchanges, saving, etc.

This model can also be useful for detecting anomalies and fraudulent activities in financial operations. When the actual price behavior of a cryptocurrency changes significantly from the modeled behavior, it can indicate the effect of external factors such as major global events as well as fraudulent activities. Further research could examine whether there are any macroeconomic or financial variables and indices that affect bitcoin volatility. In this paper, a specific machine learning algorithm, random forests, is chosen to predict the time series of realized volatility of Bitcoin. The same procedure can be done using another machine learning algorithm such as neural networks, support vector machines, logistic regression, lasso, k-nearest neighbor regression, etc. Additionally, one can examine which of these algorithms predicts with greater accuracy. Different types of variability can be examined as dependent variables of the model, or different types of methodology in which the prediction will not be a time series, i.e. regression, but classification where the prediction is made using an increasing or decreasing categorical variable.

In this paper, an analysis of the results obtained with tests is made in the case when the training and testing sets are divided on a precisely determined date of the time series, and a second case when the data to be taken in the training and testing set is randomly selected with the function sample.split() from the caTools package, when predicting bitcoin market price volatility. Different criteria such as forecast error measurements, the speed of calculation, interpretability and others have been used to assess the quality of forecasting. Forecast error measures or forecast accuracy are the most important in solving practical problems (Shcherbakov et al., 2013). With the analysis made in this paper and the resulting values of the standard errors, it can be said that this model can be successfully used to improve financial services, such as predicting the volatility of

Bitcoin, which is becoming more and more popular every day for people who want to invest in this cryptocurrency. Predicting the change in the value of bitcoin improves the operation of financial services, and reduces the risk factor when investing, working on stock exchanges, saving, etc.

This model can also be valuable for identifying anomalies and detecting fraudulent activities in financial operations. When the actual price behavior of a cryptocurrency changes significantly from the modeled behavior, it can indicate the effect of external factors such as major global events as well as fraudulent activities. Further research could explore whether there are any macroeconomic or financial variables and indices that influence bitcoin volatility.

References

Aljadani, A. (2022). DLCP2F: a DL-based cryptocurrency price prediction framework. *Discover Artificial Intelligence, 2(1)*, 20.

Aristeidou, C. (2020). Study of the volatility of bitcoin cryptocurrency using machine learning methods: an implementation in R, available at http://nemertes.library.upatras.gr.

Ashurst, S., and Stefano, T. (2021). Blockchain applied: practical technology and use cases of enterprise blockchain for the real world. *Productivity Press*.

Chai, T., and Roland R., D. (2014). Root mean square error (RMSE) or mean absolute error (MAE). *Geoscientific Model Development Discussions 7, no. 1,*, pp. 1525-1534.

Chamorro-Courtland, C. (2021). The Future of Clearing and Settlement in Australia: Part II-Distributed Ledger Technology. *Company & Securities Law Journal 38, no. 7.*

Chauvet, M., Senyuz, Z., Yoldas, E. (2010). What Does Realized Volatility Tell Us About Macroeconomic Fluctuations? *Unpublished working paper*.

Dewi, C., and Rung-Ching, C. (2019). Random forest and support vector machine on features selection for regression analysis. *Int. J. Innov. Comput. Inf. Control 15, no. 6*, pp. 2027-2037.

Géron, A. (2022). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. *O'Reilly Media, Inc*.

Gu, J. B. (2018). Consortium blockchain-based malware detection in mobile devices. *IEEE Access 6*, pp. 12118-12128.

Jacobucci, R. (2018). Decision tree stability and its effect on interpretation, available at https://osf.io/m5p2v/

Khair, U. H. (2017). Forecasting error calculation with mean absolute deviation and mean absolute percentage error. *Journal of Physics: Conference Series, vol. 930, no. 1, IOP Publish*, p. 012002.

Kotsiantis, S., Zaharakis, I., Pintelas, P. (2006). Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review, 26(3)*, pp.159–190.

Kramer, M. (2019). An overview of blockchain technology based on a study of public awareness. *Global Journal of Business Research 13, no. 1*, pp. 83-91.

Kreinovich, V. H. (2014). How to estimate forecasting quality: A system-motivated derivation of symmetric mean absolute percentage error (SMAPE) and other similar characteristics, available at https://scholarworks.utep.edu.

Kumar, M. S. (2019). Credit card fraud detection using random forest algorithm. *In 3rd International Conference on Computing and Communications Technologies (ICCCT)*, pp. 149-153.

Liu, X. and Xiaoguang, D. (2021). TanhExp: A smooth activation function with high convergence speed for lightweight neural networks. *IET Computer Vision 15, no. 2*, pp. 136-150.

Lukić, V. (2016). Potentials and limits of private digital currencies, available at http://www. ekof. bg. ac. rs/wp-content/uploads/2016/03/Seminar-katedre-2017-Potencijali-i-ograni% C4% 8Denja-privatnih-digitalnih-valuta-PDF. pdf.

Mijoska, M. and Ristevski, B. (2021). Possibilities for applying blockchain technology–a survey. *Informatica 45, no. 3*.

Mijoska, M., Ristevski, B., Savoska, S., Trajkovik, V. (2022). Predicting Bitcoin Volatility Using Machine Learning Algorithms and Blockchain Technology, available at https://repository.ukim.mk.

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. *Decentralized Business Review: 21260*.

Samuel, A. L. (1967). Some studies in machine learning using the game of checkers. II—Recent progress. *IBM Journal of research and development 11.6*, pp. 601-617.

Senbet, L. W. and Wang, T. (2012). Corporate financial distress and bankruptcy: A survey. *Foundations and Trends® in Finance 5, no. 4*, pp. 243-335.

Shcherbakov, M. V., Brebels, A., Shcherbakova, N., Tyukov, A., Janovsky, T., Kamaev, V. (2013). A survey of forecast error measures. *World applied sciences journal 24, no. 24*, pp. 171-176.

Vadapalli, P. (2021). Random Forest Classifier: Overview, How Does it Work, Pros & Cons, available at https://www.upgrad.com/blog/random-forest-classifier.

Vujičić, D., Jagodić, D., Ranđić, S. (2018). Blockchain technology, bitcoin, and Ethereum: A brief overview. *17th international symposium infoteh-jahorina (infoteh)*, pp. 1-6.

WEB(a). *Gemini Exchange Data*, available at https://www.cryptodatadownload.com/data/gemini.

WEB(b). *Blockchain Charts,* available at https://www.blockchain.com/charts.

Willmott, C. J. and Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research 30, no. 1,*, pp. 79-82.

Yokuma, J. T., and Armstrong, J. (1995). Beyond accuracy: Comparison of criteria used to select forecasting methods. *International Journal of Forecasting 11, no. 4*, pp. 591-597.

Zhang, L., Xie, Y., Zheng, Y., Xue, W., Zheng, X. (2020). The challenges and countermeasures of blockchain in finance and economics. *Systems Research and Behavioral Science 37, no. 4,*, pp. 691-698.

Received January 10, 2024, revised September 16, 2924, accepted November 11, 2024