Best Feature Selection using Correlation Analysis for Prediction of Bitcoin Transaction Count

Se-Hyun Ji
Computer Information and Science
Korea University
Sejong, Korea
sxzer@korea.ac.kr

Ui-Jun Baek
Computer Information and Science
Korea University
Sejong, Korea
pb1069@korea.ac.kr

Mu-Gon Shin
Computer Information and Science
Korea University
Sejong, Korea
tm0309@korea.ac.kr

Young-Hoon Goo
Computer Information and Science
Korea University
Sejong, Korea
gyh0808@korea.ac.kr

Jun-Sang Park
LG Electornics
A&B Center
Seoul, Korea
junsang.park@lge.com

Myung-Sup Kim
Computer Information and Science
Korea University
Sejong, Korea
tmskim@korea.ac.kr

***Abstract*—Cryptocurrency made on the basis of block-chain technology Bitcoin is drawing the attention of individuals, corporations, governments and financial institutions today. As the number of Bitcoin transactions increases over the past years, the scale of the Bitcoin market has been increasing day by day. Predicting the number of transactions contained in a Bitcoin block is important in a Bitcoin network. The aim of this paper is to propose a learning feature selection method for designing a machine learning model that predicts the number of transactions contained in the Bitcoin block by applying the machine learning algorithm. Selecting the appropriate feature to design a machine learning model is crucial things to the performance of the model. We apply correlation analysis to select the appropriate learning feature of the transaction count prediction model in the Bitcoin block and verify the validity of the proposed method through experiments.**

Keywords—Cryptocurrency, Bitcoin, Machine Learning, Feature Selection, Coefficient Analysis

I. INTRODUCTION

Bitcoin developed by Satoshi Nakamoto is a cryptocurrency made based on block-chain technology [1]. Currently, Bitcoin attracts the attention of government, business, and financial institutions. In the last few years, the number of transactions in Bitcoin has increased tremendously. According to coinmarketcap.com, as of May 2019, Bitcoin's market capitalization amounts to about $100 billion. As the number of Bitcoin transactions increases, the Bitcoin network is rapidly developing, but there is also a problem. For example, Bitcoin transaction processing costs have increased, but transaction acknowledgment times have been delayed. For this reason, predicting the number of transactions contained in the Bitcoin block is important in coping with the growth and problems of Bitcoin networks. In order to predict the number of transactions contained in the Bitcoin block as in the [2], various researches is underway to predict the number of transactions in the Bitcoin block by applying various machine learning algorithms.

The aim of this paper is to propose the learning feature selection method for designing the machine learning model that predicts the number of transactions contained in the Bitcoin block by applying the machine learning algorithm. In this research, we collect 84 kinds of Bitcoin block and transaction statistical feature to predict the number of transactions contained in the Bitcoin block. The performance of the machine learning model depends on the learning feature. Therefore, it is important to find a training feature appropriate for predicting the number of transactions in the Bitcoin block among the 84 kinds of Bitcoin block and transaction statistical feature. Using the feature of all kinds as a learning feature to model's learning feature cannot guarantee the performance of the model. To learn feature using the number of all cases is inefficient. There is a need for an efficient method for predicting the number of transactions contained in a Bitcoin block using a machine learning algorithm. So, we propose the machine learning feature selection method applying correlation analysis in an efficient way.

In this paper, following the introduction and describe to the various machine learning algorithm used for Bitcoin network analysis in related work. In the methodology, we propose two correlation analysis method algorithms selecting learning features for predicting the number of transactions in the Bitcoin block. The proposed method verifies validity through experiments.

II. RELATED WORK

Research on predicting the number of transactions contained in the Bitcoin block is lacking. Therefore, this Section describes the research of analysis of Bitcoin data by applying a machine learning algorithm.

A. Artificial Neural Network

Artificial neural network (ANN) is machine learning algorithms represented by computer systems that abstract biological neurons [3]. [4] collected the Bitcoin transaction data to predict the price of Bitcoin and used transaction data for the ANN model's feature. As a result, they reported the price direction accuracy of 55% using ANN. [5] were also predicted the price of Bitcoin using ANN, and they tried to investment based on the ANN model's prediction. Their method attained higher performances than the primitive investment methods. In addition, they found important implications on the relationship of ANN model performance with the input data.

B. Multilayer Perceptron

Multilayer Perceptron (MLP) is a class of feed-forward artificial neural networks [6]. [7] researched the relationship between the features of Bitcoin block and the next day

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea Government(MSIT) (No.2018-0-00539-001, Development of Blockchain Transaction Monitoring and Analysis Technology) and Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2018R1D1A1B07045742)

change in the price of Bitcoin by applying MLP and used a trading strategy through the MLP model got a result in making 85% profit in return. [8] used the additional feature with Bitcoin block data like currency feature, and they used correlation analysis to their feature for predicting the price of Bitcoin. [9] researched Bitcoin transaction fees estimation using MLP and to help Bitcoin users save expenditures in their funds in their transaction fees.

C. Recurrent Neural Network

Recurrent neural network (RNN) is a class of artificial neural networks where connections between nodes from a directed graph along a temporal sequence [10]. RNN is an algorithm specialized for the prediction of time series data. [11] predicted the exchange rate of BTCEUR by applying RNN and their results showed the possibility of applying RNN to time series data. [12] used the sentiment on Twitter as input data for RNN for predicting the price of Bitcoin, and They got the overall price prediction accuracy using RNN is found to be 77.62%. [13] used consists of fourteen input variables related to the Bitcoin prices as input data for RNN, and they attempted to predict the market trend of Bitcoin using RNN.

D. Long Short-Term Memory

Long short-term memory is an artificial recurrent neural network (RNN) architecture. In general, LSTM shows better performance than RNN. [14] used the price of Bitcoin as input data for LSTM for predicting the price of Bitcoin and compared it to the ARIMA model. They compared the performance of the ARIMA model with that of the LSTM model and found that the LSTM model showed better performance. [15] predicted the exchange rate of the Bitcoin using ANN and suggested that it would perform better if LSTM is applied. [16] proposed a new forecasting framework with the LSTM model to forecasting the daily price of Bitcoin, and they got excellent forecasting accuracy of the proposed model.

III. METHODOLOGY

A. Correlation Analysis

Correlation analysis is a method of analyzing the linear relationship between two variables. The two variables can be independent or correlated, and the strength of the relationship between two variables is called the correlation [17]. We tried two correlation analysis.

The first correlation analysis uses the Spearman correlation coefficient. The Spearman correlation coefficient is a correlation coefficient when ranking is used instead of a data value. The data are sorted in order from smallest to largest, and the correlation coefficient is obtained by using rankings. Spearman correlation coefficients reveal whether there is a correlation between two variables. The Spearman correlation coefficient has a value between -1 and +1. If the rank of two variables is completely matched, it is +1, If the rank of the two variables is completely opposite, it is -1 [18].

The second correlation analysis uses the Pearson correlation coefficient. The Pearson correlation coefficient is a measure of the linear correlation between two variables. The Pearson correlation coefficient has a value between +1 and -1 due to the Kosi-Schwartz inequality, where +1 is a perfect positive linear correlation, 0 is a linear correlation, and -1 is a perfect linear relationship It means [19]. Table 1 shows the detailed analysis of the Pearson correlation coefficient. The Pearson correlation coefficient is denoted by r.

Table 1. Interpretation of Pearson correlation coefficient

Range of r	Degree of Relationship
$-1.0 \le r \le -0.7$	A strong negative linear relationship
$-0.7 \le r \le -0.3$	A distinct negative linear relationship
$-0.3 \le r \le -0.1$	A weak negative linear relationship
$-0.1 \le r \le +0.1$	Not a linear relationship
$+0.1 \le r \le +0.3$	A weak positive linear relationship
$+0.3 \le r \le +0.7$	A distinct positive linear relationship
$+0.7 \le r \le +1.0$	A strong positive linear relationship

B. Data Collect

Table 2. Raw features of Bitcoin

Feature of Bitcoin	Explanation
nTx	The number of transactions contained in the block
Weight	The weight of the block
Size	The size of the block
vSize	The virtual size of the block
nVin	The number of inputs the transaction contains
nVout	The number of outputs the transaction contains
Value	The amount of value of the transactions
Fee	The fee of the transaction
Tx.Size	The size of each transaction
Tx.vSize	The virtual size of each transaction
Vin.Value	The transaction input's value
Vout.value	The transaction output's value

Table 3. Bitcoin block's raw features of statistical processed

Data Unit	Raw Feature	1st Statistical Process	2nd Statistical Process	Number of Features
Block	**nTx**			1
	Weight			1
	Size			1
	Vsize			1
Transaction	**nVin**		Sum Max Min Avg Stdv	5
	nVout			5
	Value			5
	Fee			5
	Tx_vSize			5
	Tx_Size			5
	Vout_value	Sum Max Min Avg Stdv		25
	Vin_value			25

Since genesis Bitcoin blocks contain an only simple pattern of data, we have collected block data in 200,000 height blocks starting from 100,000 height blocks in the Bitcoin network. We collected 12 features from the raw data of the Bitcoin block and 12 features are shown in Table 2. In addition, 84 features were collected by statistical processing from the feature of Table 2, and the collected features are shown in Table 3. We distinguished Bitcoin features by block and transaction unit. The summation, maximum, minimum, average, standard variation of transaction units can be obtained. The collection criteria are based on block.

C. *Experimental Features*

To select experimental data, we performed two correlation analyzes from between 84 collected Bitcoin statistical features and the number of transactions contained in a block. The number of transactions contained in the Bitcoin block is the number of transactions contained in the next generated block of the block containing the statistical data. The selected features are listed in ascending order of the coefficients of Figure 1 and Figure 2.

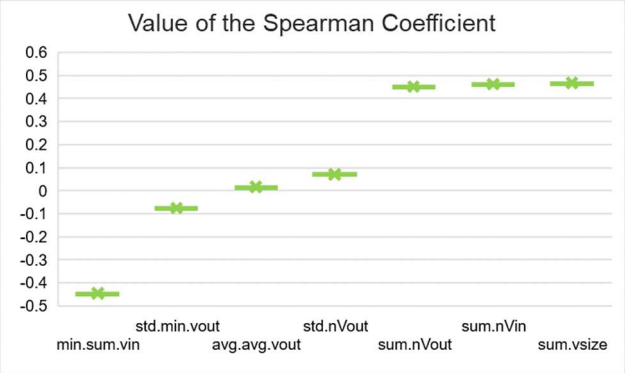

Figure 1. Selected features from the Spearman correlation analysis

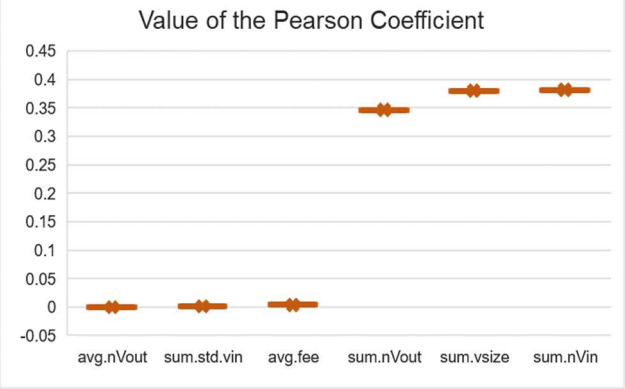

Figure 2. Selected features from the Pearson correlation analysis

In the case of applying the Spearman correlation coefficient analysis, three features with no correlation with four features with a correlation coefficient of -1 or 1 with the number of transactions contained in the Bitcoin block were selected for the performance comparison of the machine learning model. For the two correlation analyzes, sum.nVout, sum.nVin, sum.vSize resulted in a high correlation coefficient. In the case of applying the Pearson correlation coefficient analysis, three features with no linear relationship with three features with a positive linear relationship with the number of transactions contained in the Bitcoin block were selected for the performance comparison of the machine learning model.

The selected experimental features are composed of 80% of the train set, 10% of the validation set, 10% of the test set in order to evaluate the performance of the model. Table 4 shows the experimental data configuration.

Table 4. Configuration of experimental machine learning data set

Height of Bitcoin Block	Configuration of Data Set
100,000 ~ 180,000	Train Set
180,001 ~ 190,000	Validation Set
190,001 ~ 200,000	Test Set

D. *Machine Learning Model*

We chose two machine learning models. One is the ANN model and the other is the LSTM model. We did not use complex models because we focused on comparing the performance of machine learning models according to learning features. The structure of the ANN model is shown in figure 3. Input has an associated weight(w), which is assigned on the basis of input's relative importance. The neuron applies a function f to the weighted sum of its inputs. The output Y from the neuron is computed as shown in figure 3. The function f is non-linear and is called the activation function. The purpose of the activation function is to introduce non-linearity into the output of the neuron [20]. In this research, the input of the ANN model is the feature of the experimental data and the output of the ANN model is the number of transactions contained in the next generated block.

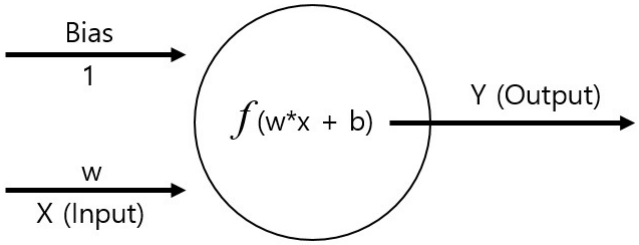

Figure 3. Structure of ANN Model

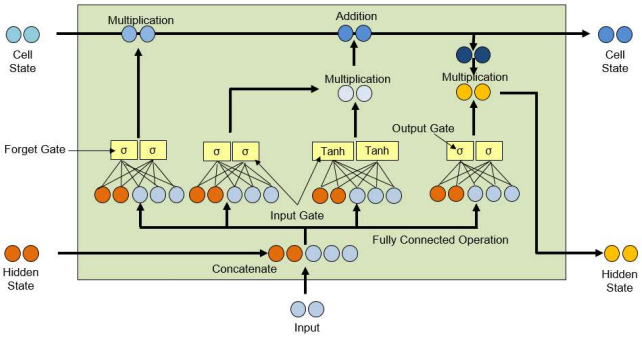

Figure 4. Structure of LSTM Model

The structure of the LSTM model is shown in figure 4. The LSTM cell composed of the input gate, the forget gate, the output gate, cell state, and hidden units. The input gate composed of the sigmoid function and the hyperbolic function. The input gate can store current information and decide whether to store new information. The forget gate and the output gate compose of The sigmoid function. The forget

gate can forget the past information and decide whether to store the previous information. The output gate controls the output value of the updated LSTM cell. The LSTM model can update or delete information to the cell state using these three gates. The hidden unit is a neural network in the LSTM cell. The hidden unit determines the shape of the output. In this research, the input of the LSTM model is the feature of the experimental data and the number of transactions contained in the Bitcoin block and output of the LSTM model is the number of transactions contained in the next generated block.

IV. EXPERIMENTS

The experiment is conducted using the machine learning feature selected through the correlation analysis. The machine learning model goes through the training phase, validation phase and testing phase. The training phase is the step of learning the same data several times. Since the same data is repeatedly learned, the machine learning model suitable for the training data is completed. The validation phase is the step of verifying the performance of the machine learning model using data not used for the training phase. The testing phase is the evaluation of the performance of the performance of the completed machine learning model through the training phase and the validation phase. The performance of the machine learning model is evaluated by the mean square error(MSE) and the mean absolute error(MAE) of the actual value and the predicted value of the machine learning model. The closer the MSE and the MAE are to zero, the better the performance of the machine learning model. We applied the Min-Max scaler to all experimental data to optimize the learning of the machine learning model. Therefore, the values of all learning features are converted to values between 0 and 1. The results of the performance of the ANN model are shown in Table 5, 6.

Table 5. Performance of ANN model with the Pearson correlation analysis

Feature	Learning Phase	MSE	MAE
avg.nVout	Validation	0.0201	0.0992
	Test	0.0227	0.1077
sum.std.vin	Validation	0.0200	0.0991
	Test	0.0226	0.1077
avg.fee	Validation	0.0199	0.0992
	Test	0.0225	0.1079
sum.nVout	Validation	0.0177	0.0887
	Test	0.0201	0.0956
sum.vsize	Validation	0.0164	0.0879
	Test	0.0188	0.0944
sum.nVin	Validation	0.0163	0.0879
	Test	0.0187	0.0943

The MSE and MAE of the ANN model which learned the selected feature by applying the Pearson correlation analysis showed that the larger the value of the Person correlation coefficient, the smaller it was. The MSE and MAE values of the ANN model applying by selecting sum.nVin, which is the feature with the largest Pearson correlation coefficient, as the learning feature is the smallest. Therefore, it is concluded that feature with a large number of the Pearson correlation coefficient values derived from applying the Pearson correlation analysis is feature appropriate for learning of the ANN model, rather than a feature with the Pearson correlation coefficient value close to zero.

Table 6. Performance of ANN model with the Spearman correlation analysis

Feature	Learning Phase	MSE	MAE
min.sum.vin	Validation	0.0199	0.0986
	Test	0.0225	0.1071
std.min.vout	Validation	0.0200	0.0991
	Test	0.0226	0.1077
avg.avg.vout	Validation	0.0200	0.0990
	Test	0.0226	0.1075
std.nVout	Validation	0.0201	0.0992
	Test	0.0227	0.1077
sum.nVout	Validation	0.0177	0.0887
	Test	0.0201	0.0956
sum.nVin	Validation	0.0163	0.0879
	Test	0.0187	0.0943
sum.vsize	Validation	0.0164	0.0879
	Test	0.0188	0.0944

The MSE and MAE of the ANN model which learned the selected feature by applying the Spearman correlation analysis showed that generally the larger the value of the Spearman correlation coefficient, the smaller it was. However, sum.vsize, which is the feature having the largest Sperman correlation coefficient value, the mse and the mae were little higher than sum.nVin. The learning features similar to the Pearson correlation analysis were selected, but the relationship between the Spearman correlation coefficient and the mse and mae of the ANN model is relatively irregular. As a result of applying two correlation analysis methods, the Pearson correlation analysis is more suitable as a method of selecting learning feature of the ANN model which predicts the number of transactions contained in the Bitcoin block and sum.nVin is the best feature that can be used for the ANN model to predict the number of transactions contained in the Bitcoin block. Finally, the ANN model with a sum.nVin learning feature was designed, and the prediction of ANN model test data is shown in figure 5.

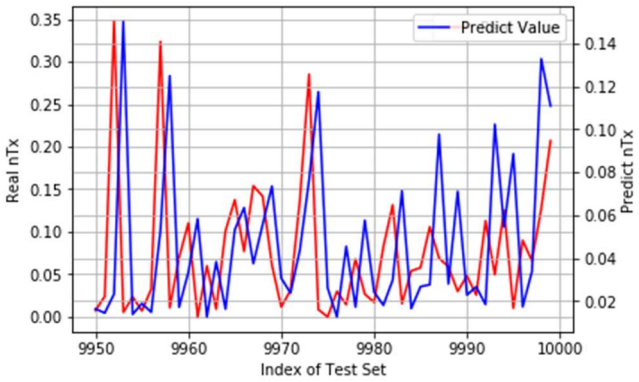

Figure 5. Prediction of the number of transactions contained in the Bitcoin block using ANN Model

In figure 5, the red line indicates the number of actual transactions in the test set and the blue line indicates the number of predicted transactions.

Unlike the ANN model, the LSTM model has many types of hyper-parameters that can be arbitrarily set. The hyper-parameters of the LSTM model are simply constructed as shown in Table 7 to maximize the influence of the learning features on the LSTM model.

Table 7 Hyper-parameters of LSTM model

Hyper-Parameter	Method / Value
Sequence Length	3
The number of hidden Units	1
Loss Function	MSE / MAE
Optimizer	Adam

LSTM The results of the performance of the LSTM model are shown in Table 8, 9.

Table 8. Performance of LSTM model with the Pearson correlation analysis

Feature	Learning Phase	MSE	MAE
avg.nVout	Validation	0.0219	0.0829
	Test	0.0251	0.0900
sum.std.vin	Validation	0.0204	0.0872
	Test	0.0235	0.0899
avg.fee	Validation	0.0219	0.0872
	Test	0.0251	0.0950
sum.nVout	Validation	0.0130	0.0827
	Test	0.0154	0.0891
sum.vsize	Validation	0.0130	0.0823
	Test	0.0154	0.0891
sum.nVin	Validation	0.0130	0.0823
	Test	0.0153	0.0889

Table 9. Performance of LSTM model with the Spearman correlation analysis

Feature	Learning Phase	MSE	MAE
min.sum.vin	Validation	0.0174	0.0941
	Test	0.0202	0.1028
std.min.vout	Validation	0.0133	0.0828
	Test	0.0156	0.0899
avg.avg.vout	Validation	0.0134	0.0828
	Test	0.0157	0.0899
std.nVout	Validation	0.0139	0.0830
	Test	0.0163	0.0901
sum.nVout	Validation	0.0130	0.0827
	Test	0.0154	0.0891
sum.nVin	Validation	0.0130	0.0823
	Test	0.0153	0.0889
sum.vsize	Validation	0.0130	0.0823
	Test	0.0154	0.0891

Overall, the performance of the LSTM model outperforms that of the ANN model. Especially, when sum.nVin was selected as a feature, the LSTM model shows about 0.003 lower MSE score and 0.005 lower MAE score than the ANN model. The MSE of the LSTM model obtained by applying the Pearson correlation analysis was smaller as the value of the Pearson correlation coefficient was larger. However, the MAE is irregularly relationship to the Pearson correlation coefficient as compared with the MSE. Spearman analysis showed a similar tendency to the experiment of the ANN model. As a result, sum.nVin is a learning feature that shows the best performance in ANN and LSTM models and Pearson correlation analysis is more accurate in selecting features of machine learning than Spearman correlation analysis. Finally, the LSTM model with a sum.nVin learning feature was designed, and the prediction of LSTM model test data is shown in figure 6. The attributes in figure 6 match figure 5.

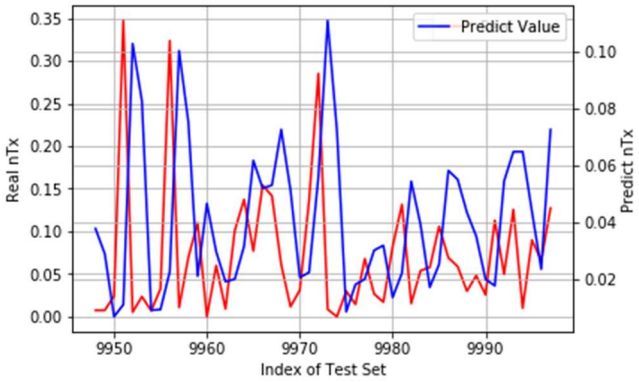

Figure 6. Prediction of the number of transactions contained in the Bitcoin block using LSTM model

V. CONCLUSION

In this paper, we propose a learning feature selection method of a machine learning model to predict the number of transactions contained in the Bitcoin block, which is an important factor in the Bitcoin network. We statistically processed Bitcoin block data to collect 84 Bitcoin statistical features and performed two correlation analyzes. The validity of the proposed method is verified through experiments of the ANN model and the LSTM model. Future research will investigate various correlation analysis methods other than the two correlation analyzes, and then apply correlation analysis to experimental data to find learning data appropriate for machine learning. Finally, we will design a sophisticated machine learning model that predicts the number of transactions contained in the Bitcoin block.

REFERENCES

[1] *NAKAMOTO, Satoshi, et al. Bitcoin: A peer-to-peer electronic cash system. 2008.*

[2] Gabriel Bianconi, Mahesh Agrawal, Predicting Bitcoin Transactions with Network Analysis, snap.stanford.edu, last modified Sep 10, 2018, accessed May 20, 2019, https://snap.stanford.edu/class/cs224w-2017/projects/cs224w-65-final.pdf.

[3] SCHALKOFF, Robert J. Artificial neural networks. New York: McGraw-Hill, 1997.

[4] GREAVES, Alex; AU, Benjamin. Using the bitcoin transaction graph to predict the price of bitcoin. No Data, 2015.

[5] NAKANO, Masafumi; TAKAHASHI, Akihiko; TAKAHASHI, Soichiro. Bitcoin technical trading with an artificial neural network. Physica A: Statistical Mechanics and its Applications, 2018, 510: 587-609.

[6] "Multilayer perceptron" Wikipedia, last modified May 05, 2019, accessed Apr 30, 2019, https://en.wikipedia.org/wiki/Multilayer_perceptron.

[7] SIN, Edwin; WANG, Lipo. Bitcoin price prediction using ensembles of neural networks. In: 2017 13th International Conference on Natural

Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD). IEEE, 2017. p. 666-671.

[8] JANG, Huisu; LEE, Jaewook. An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information. IEEE Access, 2018, 6: 5427-5437.

[9] AL-SHEHABI, Abdullah. Bitcoin Transaction Fee Estimation Using Mempool State and Linear Perceptron Machine Learning Algorithm. 2018.

[10] "Recurrent neural network" Wikipedia, last modified May 03, 2019, accessed May 15, 2019, https://en.wikipedia.org/wiki/Recurrent_neural_network.

[11] KODAMA, Osamu; PICHL, Lukáš; KAIZOJI, Taisei. Regime change and trend prediction for Bitcoin time series data. In: CBU International Conference Proceedings. 2017. p. 384-388.

[12] PANT, Dibakar Raj, et al. Recurrent Neural Network Based Bitcoin Price Prediction by Twitter Sentiment Analysis. In: 2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS). IEEE, 2018. p. 128-132.

[13] SEO, Yunbeom; HWANG, Changha. Predicting Bitcoin Market Trend with Deep Learning Models. Quantitative Bio-Science, 2018, 37.1: 65-71.

[14] KARAKOYUN, E. S.; CIBIKDIKEN, A. O. Comparison of ARIMA Time Series Model and LSTM Deep Learning Algorithm for Bitcoin Price Forecasting. In: The 13th Multidisciplinary Academic Conference in Prague 2018 (The 13th MAC 2018). 2018. p. 171-180.

[15] PICHL, Lukáš; KAIZOJI, Taisei. Volatility analysis of bitcoin. Quantitative Finance and Economics, 2017, 1: 474-485.

[16] WU, Chih-Hung, et al. A New Forecasting Framework for Bitcoin Price with LSTM. In: 2018 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, 2018. p. 168-175.

[17] "Correlation and dependence " Wikipedia, last modifed May 8, 2019, accessed May 15, 2019, https://en.wikipedia.org/wiki/Correlation_and_dependence.

[18] "Spearman's rank correlation coefficient" Wikipedia, last modified April 2, accessed May 15, 2019, https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient.

[19] "Pearson correlation coefficinet" Wikipedia, last modified April 30, 2019, accessed May 15 , 2019, https://en.wikipedia.org/wiki/Pearson_correlation_coefficient#cite_note-1.

[20] "Artificial Neuron Models", accessed May 5, 2019, https://www.willamette.edu/~gorr/classes/cs449/ann-overview.html