Prediction on Bitcoin Price Trends based on Machine Learning Algorithms

Yiyun Zhou*

Department of Mathematics, London School of Economics and Political Science, London, United Kingdom

*Corresponding author: y.zhou133@lse.ac.uk

Abstract. With the rapid growth of technology, cryptocurrency like Bitcoin is attracting more and more attention. Its high volatility in prices creates many difficulties for predicting and there has been much work on this. This paper aims to provide a comparison of various machine learning models like linear regression, SVM, random forest, and neural networks for predicting the directions for Bitcoin close prices. The dataset used is from Jan 2012 to March 2021 and all four prices are used for predictions: Close, Open, High, and Low. Two different methods are used to fit the different types of machine learning algorithms: for regressors, close price predictions are first done and then construct in the predicted direction; for classifiers, direction predictions are done directly. Accuracy is used to do the comparison, which is the percentage of correct direction predictions made via the algorithm. It is shown that LSTM, a neural network algorithm generates the highest accuracy of about 58% and the random forest classifier has the lowest accuracy of about 55.47%.

Keywords: Linear Regression; SVM; Random Forest; RNN; LSTM.

1. Introduction

Bitcoin was created in 2009. Since then, it has attracted much attention from media from all over the world. One reason is for its difficult mining process and its fundamental properties of it: anonymity and security. Despite those, bitcoin is also known for its great fluctuation in price. With high volatility comes great potential for excessive profits while trading Bitcoin. Both the potential for profits and the high level of difficulty for prediction have attracted lots of investors and researchers devoted into the topic of predicting Bitcoin prices. How can computers program themselves (from experience and some basic structure) is a topic that machine learning, at the nexus of computer science and statistics, seeks to answer [1].

The use of machine learning algorithms to forecast bitcoin values is not a novel concept; several studies have been conducted in this area. The author [2] mentioned that they achieved an accuracy of 97.5 percent for the linear regression model and 95.8 percent for the decision tree model when they used order book data to perform price prediction. Authors [3] discussed utilizing trading data from 1-minute intervals to forecast Bitcoin values. Four models are used, which include both regression models and neural networks. The paper showed that GRU gives the best result of R^2 at 99.2% and concluded that more varied parameters should be considered, i.e., reactions to prices from social media. Authors [4] talked about using network-based features to predict Bitcoin prices. The models used are linear regression, SVM, logistics regression, and neural networks. The highest accuracy is achieved by the neural network, at about 55.1%. The paper concluded the two most informative features are the net flow through accounts and transactions made by new addresses.

The object of this study is to produce a comparison of some most-known machine learning algorithms on the Bitcoin price direction prediction. More specifically, is to assess how many ups and downs there were between the actual market data and the forecasted market data. In later portions of this study, this will be covered.

2. Data and Features

2.1 Bitcoin

Bitcoin is a decentralized virtual currency that acts as a means of payment in a specific community. Cryptocurrencies are the general name for decentralized virtual money systems that frequently rely on the exchange of encrypted communications [5]. The essential ideas on which Bitcoin is built are the anonymity and security that this offers [5]. The issuers are not under the governance of the government. With those properties, there also exist some risks relating to Bitcoin. The use of Bitcoin for criminal activity is a concern, however, there are currently laws against the three criminal activities listed below: money laundering, Bitcoin-specific crime, and criminality that uses Bitcoin as a facilitator [6]. Another risk associated with trading Bitcoin is the market risk which is created by the fluctuations in the exchange rate of bitcoin to indicated currency. Figure 1 shows the exchange rate for Bitcoin in terms of USD for the period between September 2021 to September 2022. The biggest difference is over 40.0K USD within this time period.

Figure 1. Exchange rate between Bitcoin and USD from September 2021 to August 2022

The most often cited factors influencing the price of bitcoin include economic, transactional, technological, and interest-related. Large speculation can be used to explain short-term fluctuation, but long-term fundamentals may drive total variation, which is also consistent with previous studies [7]. Although bitcoin is viewed as a completely speculative asset, there is nevertheless a correlation between its price and basic elements like the amount of money in circulation and its use in trade [8]. Additionally, Bitcoin appears to be susceptible to cyberattacks that might destabilize the Bitcoin system [9].

2.2 Bitcoin Database

All data used in this paper are obtained from Kaggle, which can be accessed via https://www.kaggle.com/datasets/mczielinski/bitcoin-historical-data. The following variables are updated every minute in this database: Open, High, Low, Close, Volume in BTC and the indicated currency, and Weighted Bitcoin Price., ranging from Jan 2012 to March 2021. In this paper, the author is aiming to predict close prices, which can be useful for making daily trading strategies. During the time of Jan 2012 and March 2021, the lowest close price was 4.33 USD (2012-02-19) and the highest price was 60458.89 USD (2021-03-14). The mean price for this period is 4605.64 USD with a standard deviation of 8207.37.

The detailed trend is shown below in Figure 2. After 2021, the close price of Bitcoin experienced significant rise from 10000 USD to 60000 USD.

Figure 2. Bitcoin close price change from Jan 2012 to March 2021

2.3 Feature Selection

The correlation between all variables is shown below in Figure 3. It is shown that Open, High, Low, and Close prices have positive correlations, which is close to +1. Therefore, this paper will use Open, High, Low, and Close price to predict close price. Features and descriptions are shown in Table 1.

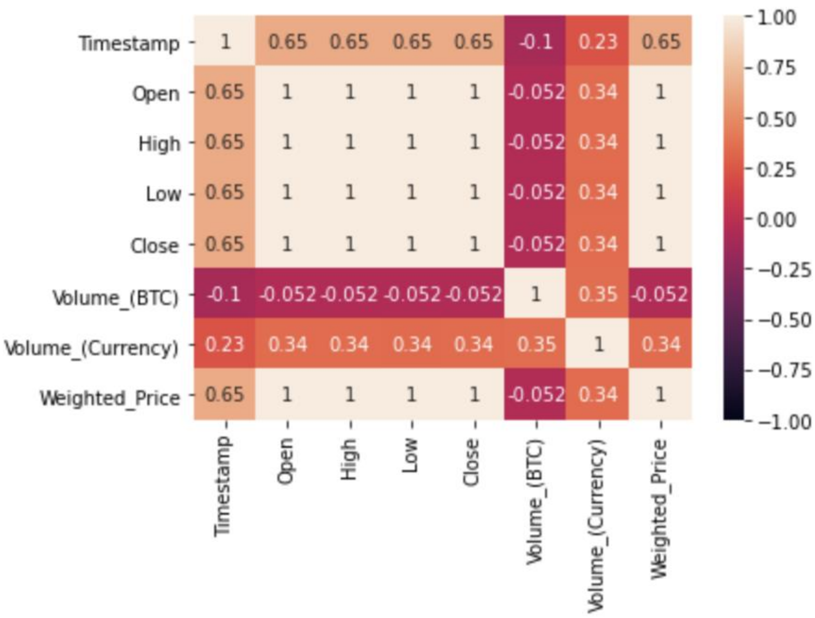

Figure 3. Correlation map between all variables in the dataset

Table 1. Selected features and the descriptions

Feature	Description
Open	The first price given within the time range
High	The highest price had within the time range
Low	The lowest price had within the time range
Close	The last traded price within the time range

2.4 Training set and test set

In this paper, there are two types of training and test set ratio used. The details are shown in Table 2.

Table 2. Training sets and testing sets

Model	Training Data			Test Data		
	Starting date	End date	Number of sets	Starting date	End date	Number of sets
Linear Regression	2011/12/31	2019/5/25	2700	2019/5/26	2021/3/31	676
SVM						
Random Forest						
Simple RNN	2011/12/31	2016/8/16	1688	2016/8/17	2021/3/31	1688
LSTM						

3. Methodology

3.1 Data Pre-processing

The Bitcoin Database used in this paper consists of 4857377 entries. To remove all the Nan values in the database. The minute-by-minute data is converted using the 'Groupby' function into a daily database of 3376 rows, and all subsequent work is based on this new database.

3.2 Python Packages

Every model used in this study is run using Python. Scikit-learn, the most well-known machine learning package, has techniques for grouping, classification, and regression. Three algorithms from this library are employed in this paper: random forest, support vector machine, and linear regression. TensorFlow, a library with an emphasis on machine learning and artificial intelligence, is another often used library. This study uses two algorithms from this library: long short-term memory and a simple recurrent neural network.

3.3 Problem Formulation

The problem defined in this paper is to predict close prices and thus predict trends. A new variable 'log_diff' is defined by taking the logarithm difference between day *t* and day (*t-1*). The log difference trend is shown below as Figure 4 The biggest difference is -0.51 in downward change and +0.29 in upward change. The overall change is not volatile, as over 99% of the changes are within ±0.2.

A new variable called 'Direction_up' is also defined via the following process: direction=1 if log difference is positive and direction=0 if log difference is negative. For regressors, prediction of the close price is done, and then construct 'Direction_up' to count the total number of predicted ups and downs. For classifiers, prediction of trend is done directly.

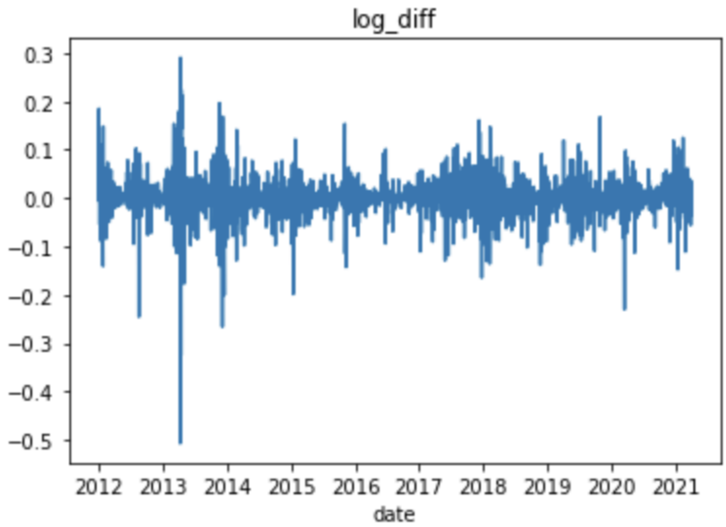

Figure 4. 'log_diff' change from Jan 2012 to March 2021

3.4 Predictive models

This part will mainly talk about all machine learning techniques and predictive models used and discussed in this paper. The models used involve both regressors and classifiers.

3.4.1 Linear regression

The linear regression method is a linear function that takes features as inputs and calculates a weighted sum for all the input features. The equation is shown below as Equation 1:

$$\hat{y} = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n \tag{1}$$

Where:
$\hat{y}$ *is the predicted value*
n *is the number of features*
x_i *is the* i^{th} *feature value*
a_i *is the* i^{th} *model parameter*

To train a linear regression model, the value of a_is is found such that root means square error (RMSE) is minimized. In this paper, multiple linear regression is used to predict the close price for Bitcoin.

3.4.2 Support Vector Machine (SVM)

Support Vector Machine (SVM) method or classifier is a machine learning algorithm that separates a data set by finding a separating hyperplane. It is regarded as the most adaptable method for creating unambiguous and precise limits [10,11]. It is capable of doing both linear and non-linear classifications and their performances are often compared regarding the same problem. When dealing with complicated datasets that have several training sets but few features, linear SVMs perform better in terms of speed and execution time than non-linear ones [12]. Although non-linear SVMs are losing explanatory power, they seem to perform stably across various problems, hence non-linear SVM becomes a more preferred choice when comparing with linear SVM [13]. In this paper, both linear and polynomial kernels are used to find the optimal algorithm.

3.4.3 Random Forest

Decision trees must first be discussed before talking about the random forest. A decision tree is a type of method that predicts the response of observation by finding the mean response of the region that the observation belongs to. Since all splitting rules can be shown in a tree, thus this type of method is called a decision tree. Decision trees themselves have high variance, hence a further method of bagging is conducted to reduce variance. Bagging is short for Bootstrap aggregation; it resamples the training set and fit classifiers for each new set. The overall forecast for a classifier problem is determined by a majority vote of all classifiers. Random Forest is an improved method based on a bagging decision tree. It allows the features to also be randomly chosen so that the variance is further reduced.

It is possible to employ Random Forest as both a regressor and a classifier. In this paper, a random forest classifier is used to predict the Bitcoin close price trends directly.

3.4.4 Simple Recurrent Neural Network (Simple RNN)

A recurrent Neural Network (RNN) is a neural network model that can analyze time series data and predict future time data points. It has an issue with long-term dependent data that might result in a vanishing gradient [3], but it can accept a sequence of data inputs and produce an equivalent length of the sequence of data, which is useful for forecasting prices in the future. In this paper, Simple RNN is used to predict the close prices for Bitcoin. The simple RNN parameters used in this paper are shown in Table 3.

Table 3. Parameters in simple RNN

Parameter	Value
time step	10
activation	Rectified Linear Unit
epoch	100
loss	Mean Squared Error

3.4.5 Long-Short-Term-Memory (LSTM)

The idea of LSTM is invented to solve the vanishing gradient problem in RNN [14]. LSTM could learn to recognize important inputs and store them in the long-term and learn to extract them whenever it is appropriate. A detailed graph of a basic LSTM cell is presented in Figure 5 [3]. Three gates are presented in a cell: the forget gate, the input gate, and the output gate. They each involve one logistic function to decide whether the outputs produced at each step should proceed to the corresponding routes or not. The LSTM parameters used in this paper are shown in Table 4.

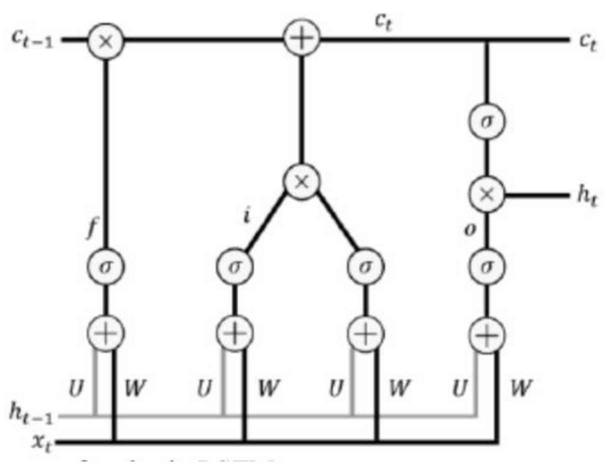

Figure 5. Structure of nodes in LSTM

Table 4. Parameters in LSTM

Parameter	Value
time step	10
activation	Rectified Linear Unit
epoch	100
batch size	32
loss	Mean Squared Error

4. Results Analysis

This section starts by presenting the result table consisting of root mean square error for both training and testing set (Train RMSE, Test RMSE), accuracy for predicting directions (Accuracy), and the training set to test set ratio (Train: Test). The results are shown in Table 5.

Table 5. Accuracy and error results for algorithms

Model	Train RMSE	Test RMSE	Accuracy	Train: Test
SVM	0.67	0.66	56.07%	8:2
Random Forest Classifier	0.64	0.67	55.47%	8:2
Linear Regression	159.84	720.46	59.47%	8:2
Simple RNN	0.05	2518.92	56.04%	5:5
LSTM	0.02	572.59	58.12%	5:5

Table 5 shows that linear regression has the highest accuracy in predicting directions for Bitcoin Close prices, which is 59.47% with test RMSE at about 720.46. The second highest accuracy is 58.12% which is achieved by LSTM with test RMSE at about 572.59. Both linear regression and LSTM are regressors, as they are used to predict the close prices first, then calculate the accuracy of the predictions.

Table 6 shows the accuracy with cross-validation for SVM and Random Forest Classifier. Cross-validation is a method that aims at finding a generalizing accuracy by splitting the training set into new test and training sets. It is done by using cross_val_score from scikit-learn python package. The two accuracy results are both lower than the results get from Table X1, which indicates a potential overfitting issue with my models.

Table 6. Accuracy with cross-validation table

Model	Accuracy with cross-validation
SVM	55.48%
Random Forest Classifier	49.48%

Figure 6 and Figure 7 illustrate a further comparison of predicting close prices under two neural network algorithms: Simple RNN and LSTM. The two predictions are plotted as line graphs and compared with the real bitcoin close prices. It is clear to see that LSTM has a relatively higher accuracy while predicting compared with Simple RNN. This conclusion can also be shown by the lower RMSE that LSTM has in Table 5.

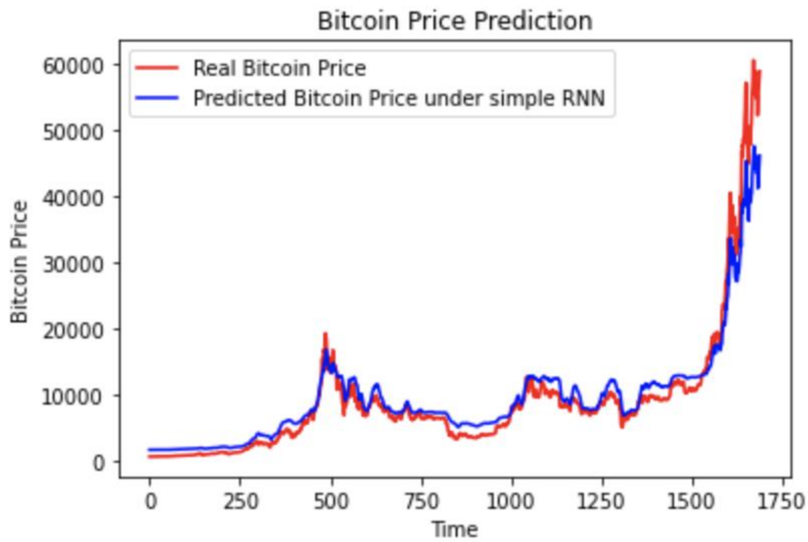

Figure 6. Real Bitcoin price and prediction under simple RNN

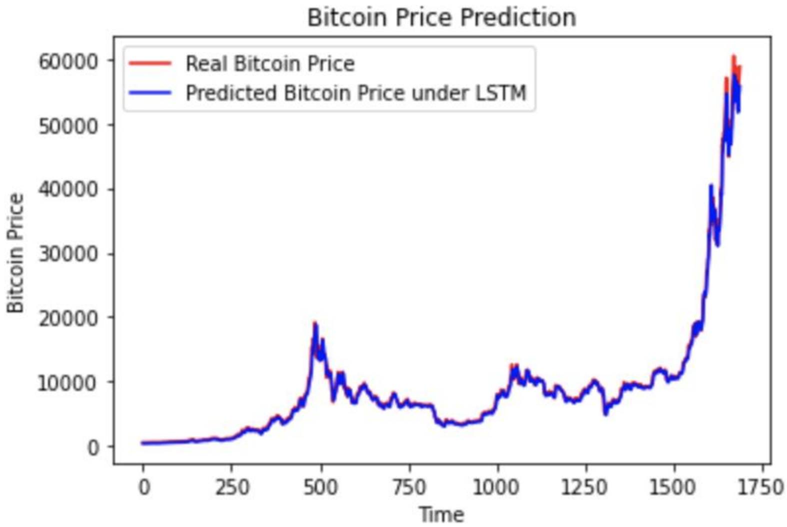

Figure 7. Real Bitcoin price and prediction under LSTM

5. Discussion

There has been a lot of research done relating to applying machine learning models in the Bitcoin market. Lots of models are evaluated under different datasets and circumstances. The results obtained in this paper don't necessarily align with some of the existing literature. [15] reveals that Random Forest has the highest forecasting performance among models like SVM and Artificial Neural Network, which is the opposite of the results in this paper.

The choice of model for predicting bitcoin prices has always been a popular topic. The most common ones, other than machine learning models, are the time series ones. Auto-Regressive Integrated Moving Average (ARIMA) model is one of them. Previous literature has shown that LSTM deep learning model has a lower RMSE and MAE compared with ARIMA Model [16,17]. This outperformance is due to the 'iterative' optimization algorithm that LSTM has [17]. For example, how many epochs there are shows the number of times that the LSTM model is trained with the same data set.

Other than model selection, feature engineering and construction is also a commonly discussed topic. Firstly, a more varied and technical feature selection and engineerings, such as Simple 14 days moving average and moving average convergence/divergence (MACD) [15] or some network-based features like what [4] used. With more technical features, higher accuracy is expected. The movements of Bitcoin close prices have chaotic and nonlinear characteristics in conditions of increasing economic uncertainties [15]. Hence some macroeconomic indicators may also be useful while predicting Bitcoin prices.

6. Conclusion

The extreme volatility of the price of bitcoin and the possibility of making excessive gains from trading it encourages researchers to develop methods for making the most precise price predictions. A comparison of the abilities of five different models to forecast the closing price of Bitcoin is conducted in this paper. The paper discusses five different algorithms for machine learning, which are linear regression, SVM, random forest, Simple RNN, and LSTM, and their performances in predicting the directions for Bitcoin close prices. All results are shown in the results analysis part and linear regression has the highest accuracy, 59.47% while LSTM has more 'well-rounded' results with the lowest test RMSE, which is 572.59, and a second high accuracy,58.12% among all regressors. As a deep learning architecture, LSTM has an advantage over traditional machine learning algorithms, like regression, on prediction problems.

In the future, the research could put more focus on deep learning as results in this paper indicate that a better and more accurate result could be potentially generated via deep learning.

References

[1] Mitchell, T. M. (2006). The discipline of machine learning (Vol. 9). Pittsburgh: Carnegie Mellon University, School of Computer Science, Machine Learning Department.

[2] Rathan, K., Sai, S. V., & Manikanta, T. S. (2019). Crypto-Currency price prediction using Decision Tree and Regression techniques. IEEE Xplore. https://doi.org/10.1109/ICOEI.2019.8862585.

[3] Phaladisailoed, T., & Numnonda, T. (2018). Machine Learning Models Comparison for Bitcoin Price Prediction. 2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE). https://doi.org/10.1109/iciteed.2018.8534911.

[4] Greaves, A., & Au, B. (2015). Using the bitcoin transaction graph to predict the price of bitcoin. No data, 8, 416 - 443.

[5] Segendorf, B. (2014). What is bitcoin. Sveri gesRiksbankEconomicReview, 2014, 2 - 71.

[6] Böhme, Rainer, Nicolas Christin, Benjamin Edelman, and Tyler Moore. 2015. "Bitcoin: Economics, Technology, and Governance." Journal of Economic Perspectives, 29 (2): 213-38. DOI: 10.1257/jep.29.2.213

[7] Bouoiyour, J., Selmi, R., Aviral Kumar Tiwari, & Olaolu Richard Olayeni. (2016). Volume 36, Issue 2 What drives Bitcoin price? http://www.accessecon.com/Pubs/EB/2016/Volume36/EB-16-V36-I2-P82.pdf.

[8] Kristoufek, L. (2015). What Are the Main Drivers of the Bitcoin Price? Evidence from Wavelet Coherence Analysis. PLOS ONE, 10(4), e0123923. https://doi.org/10.1371/journal.pone.0123923.

[9] BOUOIYOUR, J., SELMI, R., & TIWARI, A. K. (2015). Is Bitcoin Business Income or Speculative Bubble? Unconditional vs. Conditional Frequency Domain Analysis. Annals of Financial Economics, 10(01), 1550002. https://doi.org/10.1142/s2010495215500025.

[10] Lin, K., Basudhar, A., & Missoum, S. (2012). Parallel construction of explicit boundaries using support vector machines. Engineering Computations, 30 (1), 132–148. https://doi.org/10.1108/02644401311286099.

[11] Basudhar, A., & Missoum, S. (2010). An improved adaptive sampling scheme for the construction of explicit boundaries. Structural and Multidisciplinary Optimization, 42 (4), 517–529. https://doi.org/10.1007/s00158-010-0511-0.

[12] Hitam, N. A., & Ismail, A. R. (2018). Comparative Performance of Machine Learning Algorithms for Cryptocurrency Forecasting. Indonesian Journal of Electrical Engineering and Computer Science, 11 (3), 1121. https://doi.org/10.11591/ijeecs.v11.i3.pp1121 - 1128.

[13] Huerta, R., Corbacho, F., & Elkan, C. (2013). Nonlinear support vector machines can systematically identify stocks with high and low future returns. Algorithmic Finance, 2 (1), 45 – 58. https://doi.org/10.3233/af-13016.

[14] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9 (8), 1735 – 1780. https://doi.org/10.1162/neco.1997.9.8.1735.

[15] Pabuçcu, H., Ongan, S., & Ongan, A. (2020). Forecasting the movements of Bitcoin prices: an application of machine learning algorithms. Quantitative Finance and Economics, 4 (4), 679 – 692. https://doi.org/10.3934/qfe.2020031.

[16] Karakoyun, E. S., & Cibikdiken, A. O. (2018, May). Comparison of arima time series model and lstm deep learning algorithm for bitcoin price forecasting. In The 13th multidisciplinary academic conference in Prague (Vol. 2018, pp. 171-180).

[17] Siami-Namini, S., Tavakoli, N., & Siami Namin, A. (2018). A Comparison of ARIMA and LSTM in Forecasting Time Series. 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). https://doi.org/10.1109/icmla.2018.00227.