Machine Learning-Based Prediction Of Bitcoin Price

Master of Quantitative Economics
University of California, Los Angeles
Student's Name: Zeng ZiYu
Faculty Sponsor: Randall R. Rojas
2023/12

Table of Contents

Machine learning-based prediction of Bitcoin price

Senior Capstone Project for Zeng ZiYu

ABSTRACT

In this paper we applied five selected Machine learning (including Elastic Net, RF, XGBoost, MLP and LSTM) methods to predict Bitcoin price. The sample data was covered from 1 February 2018 to 31 March 2023. Moreover, we utilized Fear and Greed Index as the sentiment analysis of Bitcoin investors which enhanced the performance of prediction. Comparing the results of RMSE, MAE and MAPE, the Elastic Net Model has performed better than other machine learning methods.

Keywords: Cryptocurrency, Bitcoin Price prediction, Predictive analytics, Machine learning

INTRODUCTION

About Bitcoin

In the wave of digitization sweeping across the globe and the continuous deepening of industrial informatization, the adoption of digital currency is a prevailing trend. However, traditional currency is inherently issued by banks with the endorsement of the state, carrying a natural attribute of centralized supervision. On the other hand, the internet inherently exhibits decentralized and anonymous characteristics. Thus, the question arises: what will be the future of digital currency.

In the early 20th century, Nobel laureate in economics, Hayek (2015), believed that currency should be allowed to compete freely, similar to goods and services, to enhance the efficient allocation of monetary resources. In his work "The Denationalization of Money," he advocated for the abolition of the central banking system. In 2008, Satoshi Nakamoto brought out a scheme of a digital currency which is not issued by any government or legal entity named Bitcoin. Bitcoin ia a cryptocurrency which rely on blockchain protocols and a distributed network of users to mint, store, and transfer. As of December 2023, there are more than 23,000 cryptocurrencies with a total market capitalization of over $1.57 Trillion (CoinMarketCap, 2023). Among all those cryptocurrencies, Bitcoin has been the most popular since its initiation. It conquered the cryptocurrency market with a 52.4% share and a market capitalization of $823B. After the boom and bust of cryptocurrency market in recent years, Bitcoin has been increasingly regarded as an investment asset traded in more than 16,000 markets around the world. There is an upward demand for developing accurate forecasting models for Bitcoin price, and researchers have long been motivated to better explain the behavior of Bitcoin price.

Despite its heaviest popularity, it is worth noting that Bitcoin price has increased from zero value at the time of its inception in 2009 to $1,242 at the end of 2013. At the end of 2014, its price has dropped to $309, but is increasing rapidly in 2017 to $19,783, in 2021 it was at its peak at around $69,000 (CoinMarketCap, 2023). Such market volatility with huge price movements is extremely unusual for traditional currencies, suggesting that there must be other determinants of price

formation, which are different from traditional financial assets. Because of its highly volatile nature, there is an upward need for good prediction on which to base investment decisions. However the true nature of Bitcoin remains a vexing problem. Some countries like Salvador legalized Bitcoin as legal tender, while countries like China prohibited the minting, storage, and circulation of Bitcoin. Some researchers supported that Bitcoin is likely to be a speculative bubble rather than a future currency or long-term investment (Bouoiyour and Selmi. 2015). In recent years, the majority of investors do not treat Bitcoin as a currency according to the criteria used by economists. Instead, they regard Bitcoin as a speculative investment similar to the stocks, and Bitcoin's value reflects the confidence of investors in cryptocurrency.

Bitcoin Characteristics

Forecasting the price of Bitcoin presents a complex and challenging task due to its inherent complex features. Firstly, in terms of supply, although the total production of Bitcoin is capped at 21 million, naturally leading to deflationary pressures, the actual impact on the market's supply is influenced by the sentiments of selling among speculators. In fact, speculators have shown strong confidence in Bitcoin in recent years, with investors inclined to hold onto Bitcoin, waiting for a price breakthrough before initiating transactions. This leads to a consistently restricted supply of Bitcoin in the market, thereby propelling upward pressure on Bitcoin price. Secondly, in terms of demand, because of Bitcoin's substantial price volatility and the flexibility in trading time and location, speculators can actively participate in trading through anonymous accounts on online exchanges. As a result, speculative activities have remained robust. Additionally, there are diverse approaches to participating in speculation, including mining, accumulating Bitcoin to diminish market supply and boost its value, or developing software to prevent hacking as an indirect investment in enhancing the Bitcoin ecosystem. Thirdly, Niederjohn (2015) conducted an empirical study on the impact of the financial market on Bitcoin. The results demonstrated that when stock exchange indices and exchange rates rise, the price of Bitcoin tend to slightly increase. In general, under favorable conditions in financial markets and macroeconomic situations, there is an upsurge in speculative demand for financial products, thereby exerting a positive influence on the price of Bitcoin. However, some researchers suggest that when traditional financial assets such

as stocks decline, Bitcoin demonstrates a substitution effect, resulting in increased purchasing and holding, ultimately driving up the price of Bitcoin. Finally, as a novel investment product, Bitcoin's appeal to investors significantly influences its price. Investors typically conduct thorough research on relevant information before making investment decisions; therefore, measures of investor attention and media hype such as Google Trends search volume index play an important role in analyzing sentiment in the highly speculative cryptocurrency market.

Given the intrinsic complexity of crypto market, methodological problems arise when one wants to predict Bitcoin price over time. It is challenging for four reasons: First, Bitcoin price is typically non-stationary, noisy and deterministically chaotic, with high volatility and irregular movements. In fact it is nearly comparable to a random walk process, and thus impossible to predict accurately. Prices often exhibit trends, seasonality, and structural breaks, requiring rigorous stationarity analysis and appropriate differencing or detrending techniques to ensure reliable modeling. Second, the distribution and statistical characteristics of Bitcoin price fluctuate at various time as the process is time-varying and influenced by variations of trading rules and investors. In the early years, Bitcoin was only used by geeks to examine Blockchain technology, while now it has become a popular asset for speculation. So applying early data contribute little to model performance due to different characteristics and complex properties. Third, it is challenging for researchers to disclose potential factors that trigger Bitcoin price movements (Dolatsara et al., 2022). Bitcoin price is heavily influenced by government policies and hacking attacks which are difficult to qualitative or predict. For example on August 2nd, 2016, Bitfinex was attacked by hackers and Bitcoin price collapse nearly 20% within a day. It is extremely difficult to predict such accident within models. Moreover, although more methods about feature selection are used, previous works have depended on the researchers' domain knowledge and lack a comprehensive consideration of feature dimensions. We address this problem by integrating the conclusions of recent empirical works by researchers. Fourth, since Bitcoin price data exhibits high nonlinearity and volatility, if we train the model too close to training data, there is a risk that the model would fail to generalize well to new, unseen data. Careful attention must be paid to avoid overfitting through appropriate model evaluation methods and robust regularization techniques.

Bitcoin Prediction Survey

Through investigating the literature, we can observe that the traditional approaches for forecasting are centered on statistical methods. Roy, Nanjiba, and Chakrabarty (2018), using annual bitcoin data from 2013 to 2017, applied autoregressive model (AR), moving-average model (MA) and autoregressive integrated moving-average model (ARIMA) to forecast the bitcoin price. They found that the ARIMA model was the best model to predict the bitcoin price. Sahar Erfanian (2021) applied generalized autoregressive conditional heteroskedasticity (GARCH). Wiedmer (2018) investigated the determinants of cryptocurrency prices utilizing a panel of 17 cross-sections. The study involved the application of unit-root and cointegration tests, with effects estimated using vector error correction models (ECM), dynamic ordinary least squares, and fully modified ordinary least squares. Causality flows were examined through weak exogeneity and Granger causality tests. The findings revealed that Metcalfe's law, community factors, and search engine queries exerted a significant impact on the cryptocurrency's price. Other methods include random walk (RW), hidden Markov model (HMM), autoregressive distributed lag model (ARDL) (Jing, 2021) have only achieved limited success and tend to be focused on a specific application area. These traditional statistical methods have struggled to identify complex intrinsic features, such as non-stationarity, high volatility, non-linearity, noise and unknown connections with other time series.

Therefore, researchers and investors have devoted considerable attention to abandoning strict assumptions and adopting a model-free, data-driven, flexible, and nonparametric approach. In other words, implementing machine learning models that encompass theoretical robustness and experimental effectiveness. Greaves and Au (2015) employed blockchain data for bitcoin price prediction, utilizing a Support Vector Machine (SVM), an Approximate Nearest Neighbor (ANN), linear regression, and logistic regression. An NN classifier with two hidden layers achieved the highest price accuracy at 55%, followed by logistic regression and SVM. Limited predictability was observed in this research using only blockchain data for training and prediction. The study concluded that incorporating features directly extracted from bitcoin exchanges, such as financial flow features, would likely enhance bitcoin price prediction accuracy. Mittal, Dhiman, Singh, and Prakash (2019) utilized ML techniques such as linear regression, polynomial regression, recurrent

NN (RNN), and long short-term memory (LSTM) models to identify the correlation between bitcoin prices and Twitter and Google search patterns. They concluded that when LSTM, RNN, and polynomial regression were applied to Google Trends and tweet volume, an improved accuracy in performance was demonstrated. Chen et al. (2020) compared statistical methods like linear discriminant analysis and logistic regression with more sophisticated ML approaches, including quadratic discriminant analysis, eXtreme Gradient Boosting (XGBoost), Random Forest (RF), LSTM, and SVM, for predicting daily bitcoin prices using high-dimensional features. ML models outperformed statistical models with an average prediction accuracy of 62.2% compared to 53.05%. The best performance was observed with the LSTM model, achieving an accuracy of 67.2%.

METHODOLOGY

Data and statistical analysis

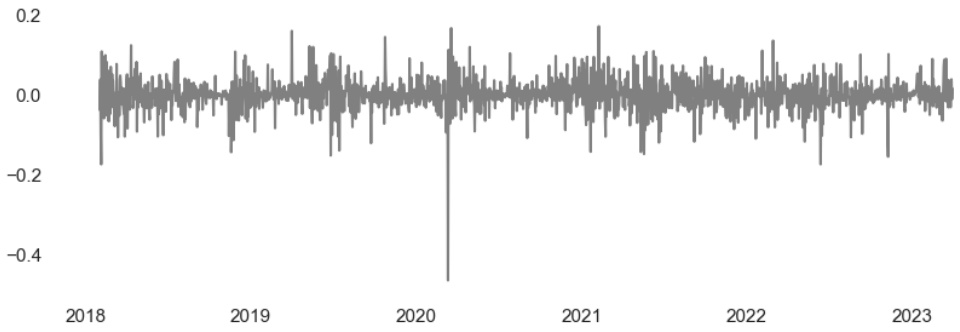

Fig. 1. The behavior of Bitcoin return time series.

The data was obtained from Kaggle.com, it consists of daily closing price (USD) and trading volume of Bitcoin, and Fear and Greed Index values for the overall crypto market. The return series of Bitcoin (calculated as a logarithmic difference in prices, i.e., $rt = (\ln p_t - \ln p_{t-1})$) is shown in Fig.1. Due to data availability, the sample data was covered from 1 February 2018 to 31 March 2023.

This dataset was chosen for three main reasons: First, as we mentioned in the introduction, the distribution and statistical characteristics of Bitcoin price has changed significantly compared with the early time, so we avoid using too much early data and hoping to capture current characteristics and complex properties. Second, we attempted to work with data at the hourly level and found that dealing with overly detailed information introduced too much noise, hindering the identification of the true data structure. Therefore, we opted for this dataset, which is structured without any missing values and is organized on a daily basis. Third, the speculation of Bitcoin is highly emotional, and people tend to overreact during periods of drastic price fluctuations, leading to irrational investment decisions. So we introduce Fear and Greed Index values in our models. This index is calculated through a comprehensive synthesis of social media analysis such as Reddit and Twitter, weekly public polling platforms, and Google Trend index. It is capable of reflecting the daily emotions and sentiments of speculators towards Bitcoin.

Table 1. Data and descriptive analysis.

Data	Mean	S.D	Min	Max	Skewness	Kurtosis	JB stat	JB stat p-value
Log Price	9.603501	0.818311	8.082329	11.120872	0.182770	-1.159851	116.153141	0
Log Volume	23.728518	0.821393	21.796106	26.583961	-0.606576	-0.440268	130.816989	0
Fear and Greed	42.363395	22.083345	5.000000	95.000000	0.586593	-0.619275	138.223049	0

In Table 1 we show a summary of the descriptive statistics of the main features used in modeling. Noticed that Fear & Greed Index (a value of 0 means "Extreme Fear" while a value of 100 represents "Extreme Greed") has a mean of 42, meaning that investors as a whole are not optimistic, and trading strategies tend to be conservative. Based on the obtained results, the kurtosis and skewness of data do not follow the corresponding characteristics in the normal distribution, and the p-value of Jarque-Bera's statistical test (JB stat p-value) is less than 0.05; therefore, the log close price and log volume of Bitcoin do not follow the normal distribution.

In order to collect appropriate data in this research, we have gathered the features that present the most impact in predicting Bitcoin price by reviewing the features and data used in similar studies. We wish to predict Bitcoin price on any day, given the history on L earlier days to achieve best forecasting performance. After tunning we fix L = 5 meaning we use the previous five trading days' data to forecast today's Bitcoin price. Besides, we noticed that the market sentiment tend to be higher in the weekend, so we add day of the week into our selected features. We also applied log transform to Bitcoin price and trading volume to enhance the accuracy and reliability of our Bitcoin price prediction models. At this stage of the modeling process, we separated 80% of our entire data as train sets and 20% as test sets. Data from 2018-02-01 to 2022-03-19 were considered the training set for model training and validation, and from 2022-03-20 to 2023-03-31, the test set for model testing. In Fig2, log price of Bitcoin has been illustrated, the training data is marked in blue, and the test data is marked in orange.

Fig. 2. Log Bitcoin price illustrated as training and testing data.

ElasticNet

In this framework, initially, we applied auto-regression test and seasonal decomposition to Bitcoin price, adhering to the structure proposed by Chen et al. (2020) to establish the ARIMA model as the baseline method. However, the absence of seasonality in Bitcoin prices resulted in unsatisfactory performance of the ARIMA model. Consequently, we tune order-5 autoregression (AR(5)) and other features with Elastic Net (a combination of Lasso and Ridge regularization) on the training data and assessed its performance on the test data.

The Elastic Net model is a linear regression model that combines both L1 (Lasso) and L2 (Ridge) regularization penalties in its optimization objective. The primary advantage of the Elastic Net model lies in its ability to encourage sparse feature selection like Lasso and regularization to correlated autoregression features like Ridge.

During the parameter tunning phase, the L1 ratio of the penalty parameter was adjusted and fixed at 0.9, resulting in the best prediction performance. As illustrated in Fig. 3, the final results of

each machine learning method's performance are presented. Noted that for visualization we only display first 3 month prediction on the test set. The Elastic Net model emerges as the superior performer, adept at tracking the market on high volatility days and avoiding substantial errors. Also, according to Table 2, the performance of the Elastic Net model during the proposed modeling process has been the best. One of the interesting points that can be mentioned about the results is that other machine learning methods like XGBoost and MLP may outperform Elastic Net model on training set based on train set MAE, but they suffer from overfitting problem, resulting in substantial error on test set.

Random Forest

Random Forest (RF) is an ensemble learning algorithm designed to enhance predictive accuracy and handle high-dimensional data by combining the predictions of multiple decision trees. During the tree construction phase, the algorithm selects the best feature and threshold at each decision tree node to split the data based on a chosen criterion. In the ensemble phase, a specified number of decision trees are built using bootstrap samples from the training data. Each tree is independently grown, incorporating random decisions at each split. In regression tasks, each tree predicts a continuous value, and the final prediction is obtained by averaging the predictions of all the trees.

In the parameter tunning phase, we set the number of trees at 1200, and use sqrt(n) of the features at each decision tree node. In Fig. 3, it is evident that the RF predictions display higher volatility compared to the Elastic Net model. However, the forecast results still fluctuate around the actual log price. According to Table 2, the performance of RF is comparable to that of XGBoost and MLP models, but unlike XGBoost and MLP, RF does not seem to suffer from overfitting problem.

XGBoost

Extreme Gradient Boosting (XGBoost) is a machine learning algorithm which minimizes the objective function through an iterative process. Unlike other boosting algorithms like AdaBoost and Gradient Boosting Machines (GBM), XGBoost inherently performs regularization. That is exactly one of the reasons we used this model. During training, XGBoost systematically selects the

best tree structure by considering all possible tree topologies iteratively and assigning scores to each leaf. The optimization employs the exact greedy algorithm to minimize the overall loss function.

In the parameter tunning phase, we chose deep tree structures and slow learning rate. In the end the model didn't fully converge so there is still room for improvement. In Fig 3, it is evident that the performance of the XGBoost model during fluctuation stages surpasses that of rapid-changing stages. There appears to be a certain lag in the prediction, with the model anticipating a smaller change than the actual market movement. According to Table 2, the performance of XGBoost is comparable to RF and MLP models, suggesting that XGBoost predictions may be more precise when the price change is relatively small.

MLP

Multilayer Perceptron (MLP) is a machine learning algorithm with multi-layer neural network architectures. MLP consists of multiple layers of interconnected nodes or neurons, each layer contributing to the extraction and transformation of features from the input data. MLP is particularly effective in capturing intricate relationships within data, making it well-suited for tasks that involve intricate patterns or non-linear dependencies. Its adaptability and capability to model complex functions contribute to our use in Bitcoin price prediction.

According to Fig 3 and Table 2, MLP prediction is similar to XGBoost prediction except for its overfitting problem. One notable drawback of MLPs is their inherent limitation in capturing sequential dependencies and temporal patterns effectively. MLPs process input data in a feedforward manner without considering the sequential nature of time series data. In Bitcoin markets, where price movements often follow intricate and non-linear patterns over time, MLPs may struggle to grasp the temporal dependencies crucial for accurate predictions.

LSTM

Long short-term memory model (LSTM) is a specialized type of recurrent neural network (RNN) designed to address the challenges of exploding and vanishing gradients in sequential data. LSTM models utilize memory cells to retain information over extended sequences, enabling them to capture long-term dependencies and temporal relationships within time series data more

effectively than traditional MLP models. Therefore, LSTM models are particularly effective for time series prediction tasks.

We applied six hypermeters for the LSTM model, including Layers size, Hidden unit size, learning rate, epoch size, Batch size, and Dropout rate. Unfortunately, the LSTM model failed to converge during the tunning process and the predictive performance was subpar. In theory and based on related literature, LSTM is expected to outperform MLP in price prediction task. Unfortunately, I failed to report that suitable hyperparameters and architecture could not be identified for the LSTM model. For the sake of research integrity, I have retained it in the study. However, the negative R^2 value indicates that its performance falls below that of a straightforward mean prediction, failing to meet the expected standards.

Table 2. The results of Machine Learning models.

Method	RMSE	R^2	MAPE	Train set MAE	Test set MAE
Elastic Net	0.032512	0.986361	0.215849	0.026083	0.021737
Random Forest	0.071290	0.934422	0.489661	0.053556	0.049163
XGBoost	0.065602	0.944468	0.483270	0.010610	0.048511
MLP	0.066513	0.942916	0.504581	0.009594	0.050593
LSTM	0.627536	-4.081394	5.765063	1.475317	0.572549

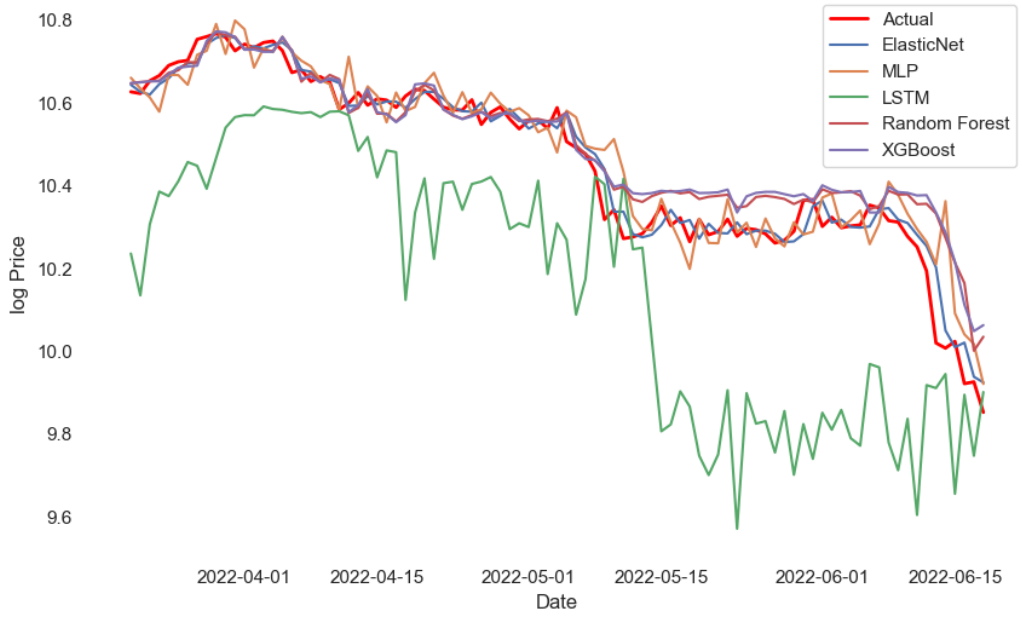

Fig. 3. The forecasting results of the Machine Learning models on test data

CONCLUSION

Bitcoin prices are inherently non-stationary, noisy, and deterministically chaotic, characterized by high volatility and irregular movements. Recognizing these complexities, numerous researchers have dedicated efforts to finding solutions for such Out of Specification (OOS) data. First we analyze the unique characteristics of Bitcoin and select features based on empirical studies. Then we applied machine learning models for log of Bitcoin price, employing techniques such as feature selection, time series cross-validation, sentiment analysis and hyperparameters tuning. Based on the obtained results, we can conclude that the proposed solutions contribute to an enhanced performance of the selected model on test data. Moreover, we observed that some features, such as Fear and Greed Index and day of the week, play an influential role in predicting the daily price of Bitcoin. By comparing the RMSE, MAE and MAPE of Elastic Net, Random Forest, XGBoost, MLP and LSTM models, we concluded that Elastic Net model stands out as the most fitting choice for Bitcoin price prediction, with RMSE of 0.033 and MAE of 0.22. By incorporating additional influential features, adjusting the model structure, refining preprocessing techniques and tuning optimal hyperparameters, we believe there is potential to develop more effective methods for predicting Bitcoin prices in the future.

Hopefully the study results have potential implications for investors, hedge fund managers, and researchers. We can achieve a better forecasting approach for Bitcoin daily price by implementing the machine learning models, especially in addressing data characteristics such as excess volatility, fat tails, and non-Gaussian distributions prevalent in cryptocurrencies.

REFERENCES

Bouoiyour, J., Selmi, R., & Tiwari, A. (2014). Is Bitcoin business income or speculative bubble? Unconditional vs. conditional frequency domain analysis. https://mpra.ub.uni-muenchen.de/id/eprint/59595

Chen, Z., Li, C., & Sun, W. (2020). Bitcoin price prediction using machine learning: An approach to sample dimension engineering. Journal of Computational and Applied Mathematics, 365, 112395. https://doi.org/10.1016/j.cam.2019.112395

CoinMarketCap 2023. accessed 11 December 2023. https://coinmarketcap.com/all/views/all/

Dolatsara, H. A., Kibis, E., Caglar, M., Simsek, S., Dag, A., Dolatsara, G. A., & Delen, D. (2022). An interpretable decision-support systems for daily cryptocurrency trading. Expert Systems with Applications, 203, 117409. https://doi.org/10.1016/j.eswa.2022.117409

Greaves, A., & Au, B. (2015). Using the bitcoin transaction graph to predict the price of bitcoin. Available via DIALOG. https://pdfs.semanticscholar.org/a0ce/864663c100582805ffa88918910da89add47.pdf. Accessed 8 Dec 2015.

Harrison, Ashley and Niederjohn, M. Scott and Clark, Jeff R. (2015). Is Bitcoin the Money of the Future? Social Education, Vol. 79, No. 2, March/April 2015. https://ssrn.com/abstract=2765457

Hayek , F. A. (1976). Denationalisation of Money. London : The Institute of Economic Affairs.

Mittal, A., Dhiman, V., Singh, A., & Prakash, C. (2019, August). Short-term bitcoin price fluctuation prediction using social media and web search data. In 2019 twelfth international conference on contemporary computing (IC3) (pp. 1-6). IEEE. DOI: 10.1109/IC3.2019.8844899

Pengfei Jing.(2021). Comparative Research of Bitcoin Price Prediction Based on Multiple Models. Shanxi University. https://link.cnki.net/doi/10.27284/d.cnki.gsxiu.2021.000592doi:10.27284/d.cnki.gsxiu.2021.000592.

Roy, S., Nanjiba, S., & Chakrabarty, A. (2018). Bitcoin price forecasting using time series analysis. In 2018 21st International Conference of Computer and Information Technology (ICCIT), IEEE, pp. 1–5.

Sahar Erfanian. (2021). Comparative Analysis of Machine Learning Algorithms for Cryptocurrency Price Prediction. SiChuan University. https://link.cnki.net/doi/10.27342/d.cnki.gscdu.2021.000756doi:10.27342/d.cnki.gscdu.2021.000756

Wiedmer, J. (2018). The price of cryptocurrencies: an empirical analysis (MSc thesis). Univerität Bern, Bern, Switzerland.