Bitcoin Return Prediction based on OLS, Random Forest, LightGBM, and LSTM

Gaohao Zhu[1, *]

[1]Department of Economics, University of Southern California, Los Angeles, USA

*Corresponding author: robinzhu@usc.edu

Abstract. Since Bitcoin was proposed in 2008, it has become a very valuable asset and an important part of many investors' portfolios. It's important to both understand Bitcoin mechanics and predict its valuation with the help of the state-of-art machine learning tools. The study develops four different models, including Ordinary Least Squares (OLS) regression model, Random Forest, Light Gradient Boosting Machine (LightGBM), and Long Short-Term Memory (LSTM), to predict the return of Bitcoin and compare the performance of these models. According to the analysis, the daily changes in the high, low, close price of Bitcoin, and close price of Tesla stock, and gold price between yesterday and today are all strongly correlated to the Bitcoin return on tomorrow. The statistical approach, or OLS modeling, has the simplest algorithm whereas the highest accuracy rate. The LightGBM model and LSTM model have lower accuracy rates in order, but still exceed the 50% (random benchmark). The Random Forest model, as another type of decision tree algorithm, has similar prediction results with the LightGBM model but a lower accuracy rate that fails to reach the benchmark. Based on the analysis, multiple factors affect the Bitcoin return, and these results provide an insight for investors to the cryptocurrency market and the macroeconomic environment. It validates the effectiveness of several machine learning algorithms in Bitcoin return forecasting and supports future developments in related fields.

Keywords: Bitcoin Prediction; Cryptocurrency; Machine Learning; LSTM; Forecasting Performance.

1. Introduction

The concept of Bitcoin was invented by Satoshi Nakamoto in 2008 [1]. Initially, Bitcoin was designed as an electronic cash system on a peer-to-peer basis without a third-party financial institution. Moreover, it was powered by the blockchain technology, a distributed database in which all transactions of cryptocurrency (such as Bitcoin) can be recorded. At the very beginning, many people were new to the fields of blockchain and cryptocurrency. Therefore, few people used them and the market value of Bitcoin wasn't very high. As a result, it was difficult to purchase any material good with this type of cryptocurrency in most cases. Contemporarily, as the market has more confidence in blockchain and cryptocurrency, Bitcoin has become very popular as a choice of asset management. It is traded by many investors and has the highest conversion rate with the US dollar among all cryptocurrencies.

Therefore, the valuation of Bitcoin has been a heated topic that many researchers are focusing on. Recent studies have employed different statistical approaches and machine learning algorithms to predict the Bitcoin prices. Some researchers use linear regression methods to predict the closing prices and get high accuracy scores [2-5]. Some scholars use random forest, a type of decision tree, to make the prediction [6-10]. While some researchers try to use the boosting machine (e.g., Light Gradient Boosting Machine (LightGBM)) to improve the prediction accuracy [11], other researchers start to apply neural networks including Long Short-Term Memory (LSTM) to the price prediction [5- 7, 12-21].

According to previous studies that involve a comparison between various machine learning algorithms, the LSTM model usually has the best performance in Bitcoin price prediction, defeating all other models in terms of accuracy scores. Those studies often use daily data of Bitcoin prices, and import Time Series modeling. In other words, the prices are recorded day by day, and they are selected into the model following the sequence of time, without being randomly splitting into unordered pieces. In fact, the daily prices of Bitcoin don't fluctuate from that of the previous day much. In most cases, it's easy for machine learning models to follow the trend/pattern in the market, and predict the daily prices very well. However, such predictions won't give investors a strong indication of either buying or selling Bitcoin. Since the price will mostly fluctuate in a moderate range, the predicted price is not that sufficient in pointing out the direction of the market movement due to the small daily changes in price.

The purpose of this study is to predict the return ratio of Bitcoin using different machine learning algorithms, including linear regression, random forest, lightGBM, and LSTM. The return can be either a profit (when the next day's price is higher) or a loss (when the next day's price is lower). Hence, it can give investors a stronger signal of buying Bitcoin or not, which is identified as a two-classification problem. In addition, the study is aimed to find out which machine learning model above has the highest accuracy score in such prediction on return, through a comparative analysis. With daily data of BTC-USD from 2016-07-01 to 2022-07-01, all 4 models will be trained and tested, and their accuracy scores will be determined by a confusion matrix of binary classification. The model with the highest accuracy score is the one with the best performance in predicting Bitcoin return.

The rest of the paper is organized as follows. The Section 2 will explain the research methodology, upon data, models, processing, and accuracy metrics. The Section 3 will show the results of correlation and accuracy, give an explanation of the obtained results, and state some limitations of this study. Eventually, the Section 4 will summarize this study.

2. Methodology

The prediction on return is done by four models separately, and each model generates an accuracy score for comparison. If the next day's price is higher than that day's (i.e., the return is a profit), the direction of market movement is up and it is displayed as 1. Otherwise, the return is a loss and displayed as 0. The models compute the binary classification problem individually, predicting the return as either a profit or a loss. The predicted values will be compared with the actual values of the market movement, through a confusion matrix that gives the accuracy score.

2.1 Data and basic models

The data includes daily prices and volumes of Bitcoin, Tesla stock, CBOE Market Volatility Index (VIX), and they are all downloaded from Yahoo Finance. Additionally, the daily prices of gold from a Kaggle dataset are included as well. Data within the timeframe from 2016-07-01 to 2022-07-01 is incorporated into one dataset for study. To conduct a prediction on the return, the study scales the data into log forms instead of using the data's absolute values. For the daily open, high, low, close prices and volumes of Bitcoin, a common logarithm (logarithm with a base of 10) is calculated for each day. Then, a log difference between today and yesterday is generated by using Day(n)'s log value to minus Day(n-1)'s log value. The same process is repeated for eight times, which turns into eight independent variables: log difference of open price of Bitcoin, log difference of high price of Bitcoin, log difference of low price of Bitcoin, log difference of close price of Bitcoin, log difference of volume of Bitcoin, log difference of Tesla stock price, log difference of VIX, and log difference of gold price. For those days when the market is closed, the missing value of Tesla stock price, VIX, and gold price is filled with the average value of prices on the nearest trading day before and after.

The return on Bitcoin is calculated by the log difference of closing prices between today and tomorrow. After obtaining this value by using Day(n+1)'s log to minus Day(n)'s log, it's turned into the binary form of 0 and 1, depending on whether the return is negative or positive. The two-classified return is the dependent variable that the study investigates.

2.2 Trainning models

In total, four models are constructed based on four machine learning algorithms: Ordinary Least Squares (OLS), Random Forest, Light Gradient Boosting Machine (LightGBM), and Long Short-

Term Memory (LSTM). They are given the same input of independent variables in time series, which are the predictors, and they are expected to output the dependent variables in a binary form.

2.2.1 OLS

The mathematical descriptions of the linear model can be given as follows:

$$Y = f(X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8) \tag{1}$$

Here, Y is a binary classification that represents for Profit or Loss of Bitcoin Return on tomorrow; X_1 represents for the Log Difference between yesterday's and today's Open Prices of Bitcoin; X_2 represents for the Log Difference between yesterday's and today's High Prices of Bitcoin; X_3 represents for the Log Difference between yesterday's and today's Low Prices of Bitcoin; X_4 represents for the Log Difference between yesterday's and today's Close Prices of Bitcoin; X_5 represents for the Log Difference between yesterday's and today's Volumes of Bitcoin; X_6 represents for the Log Difference between yesterday's and today's Close Prices of Tesla Stock; X_7 represents for the Log Difference between yesterday's and today's Close Prices of VIX; X_8 represents for the Log Difference between yesterday's and today's Close Prices of Gold; Y is a function of the eight independent variables above.

The OLS model is based on linear regression. After inputting the dataset, the model will generate datapoints based on all the variables. Then, it draws a best-fit line that minimizes the residuals (distance between the datapoints and the best-fit line). The best-fit line gives an equation for prediction. The predicted y-variable has a range from 0 to 1, and the benchmark is set at 0.5. In this case, if the predicted value is greater than 0.5, it more approaches to 1, and thus it's classified as 1. Otherwise, the predicted value is classified as 0. Hence, the OLS model classifies the return into two groups based on the line that does the prediction accordingly.

2.2.2 Random Forest

Random forest is a type of decision tree, and it's widely used in classification problems. In a decision tree, each node will split into different classes, until each node contains samples from a single class. The sample that is being tested will trace down the tree nodes and be assigned into a class. Bagging helps improve the performance of a decision tree and consists of the random forest. In this study, the number of estimators ranges from 100 to 400, and the random state is always 101.

2.2.3 LightGBM

The Light Gradient Boosting Machine is a new Gradient Boosting Decision Tree algorithm. The difference between LightGBM and traditional decision trees is that LightGBM will continuously reduce the loss of a leaf by taking gradients. By calculating the loss of a loss, the LightGBM is expected to improve the performance of prediction over time, until it overfits the data. For the parameters the study uses, the max_depth is from 2 to 10; the number of estimators is from 100 to 400; the learning rate is from 0.1 to 0.5; the min_child_samples are from 20 to 50; the random state is always 1 (the default value).

2.2.4 LSTM

The Long Short-Term Memory neural network is a state-of-art algorithm of deep learning used in time series modeling. It processes partial data and gives some output that can be carried to the next layer for another processing along with additional input. It also includes a dropout layer that randomly excludes some data from training, in order to reduce overfitting. In general, the LSTM model can deal with enormous datasets, and requires much data to improve accuracy. The study tests parameters of different values repeatedly: the window length is from 3 to 10; the number of neurons is from 20 to 200, the number of epochs is from 20 to 200; the dropout size is from 1/10 to 1/5; the learning rate is from 0.01 to 0.2. Some parameters are held constant, i.e., the random state is 42 and the batch size is 32.

2.3 Processing & Metrics

All the models are given the same input, which is a compiled dataset including the date, all the predictors, and the dependent variable. With time series modeling, the four algorithms will process the data and predict on the return of Bitcoin. To examine the accuracy of the prediction, a confusion matrix of binary classification is applied. Within the confusion matrix, two columns are the two situations: return is positive (i.e., profit) and return is negative (i.e., loss). The two rows are: predicted return is positive and predicted return is negative. After plugging the prediction results into the confusion matrix, the accuracy score will be automatically calculated, based on how well the predicted values fit the actual ones.

3. Results & Discussion

All data is combined into one set for analysis after feature engineering. By doing a linear regression, all the eight independent variables have a statistically significant relationship with the Bitcoin return. Among them, the log difference of high price, low price, volume of Bitcoin, and gold price have a positive correlation with the Bitcoin return. The remaining predictors, including the log difference of open price, close price of Bitcoin, close price of Tesla stock, and VIX index, have more or less negative correlation with the Bitcoin return. In Figure 1, the Bitcoin historical close prices are shown.

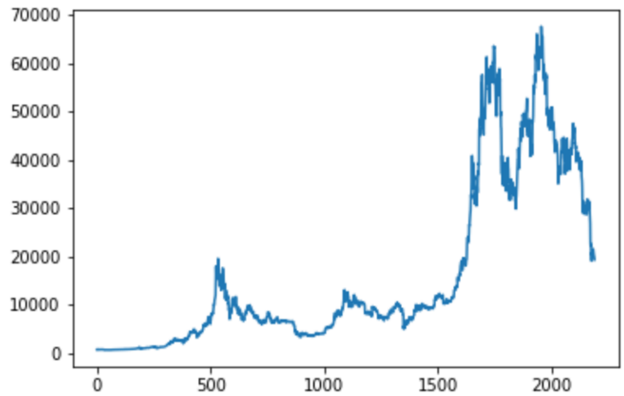

Fig. 1 Bitcoin Close Prices (2016-07-01 to 2022-07-01)

After training and testing each model, a confusion matrix is implemented to evaluate the accuracy of the four models. In the circumstance of doing a binary classification of Bitcoin return, the statistical OLS model turns out to be the most accurate, followed by the LightGBM model, with LSTM model ranked at the third. Random Forest model shows the lowest accuracy rate, not even beating the possibility of a random coin flip.

3.1 Correlation analysis

The linear regression model shows the correlation between independent variables and the dependent variable, as depicted in Figure 2. Although the log difference of Bitcoin volume has a positive correlation with the Bitcoin return, the relationship is very weak, as the regression coefficient is only 0.0204. Similarly, the log difference of Bitcoin open price and VIX index have a negative correlation that is very weak, with regression coefficients being -0.0956 and -0.0684 in order.

The log difference of high price and low price of Bitcoin have a positive correlation with the return. The regression coefficients are 0.3184 and 0.4468 in order, which indicates a moderately strong relationship with the response variable. Besides, the log difference of close price of Tesla stock has a strong negative correlation with the Bitcoin return, with a regression coefficient of -0.7315. In other words. when Tesla, one big holder of Bitcoin, has its stock's value increasing, the Bitcoin return is expected to decrease. This will be explained later in the section.

Among all the predictors, the log difference of Bitcoin close price and gold price have the strongest relationship with the Bitcoin return. As for the log difference of Bitcoin close price, it has a negative correlation with the return, with a regression coefficient of -1.7720. Log difference of gold price, however, is positively correlated with the Bitcoin return, having a regression coefficient of 2.6552. As gold depreciates, people feel pessimistic for the market, so the Bitcoin return declines as well.

In between the independent variables, some correlations have also been found. A positive correlation with a coefficient of 0.47 exists between log difference of high price and low price of Bitcoin, which is self-explanatory as they are data from the same day and the Bitcoin price won't fluctuate much in one day. The two prices are also correlated with the open and close prices, with a moderate strength. However, it suggests that the log difference of open price of Bitcoin isn't really correlated to that of the close price, since the correlation coefficient is only -0.028.

Another important result is that the log difference of Tesla stock close price is negatively correlated to the VIX index. The correlation coefficient is -0.34, showing that as the Tesla stock appreciates, the market becomes less volatile. Consequently, more people invest in the stock market rather than Bitcoin, dragging down the return of Bitcoin. This helps explain why the log difference of Tesla's stock is negatively correlated to the Bitcoin return.

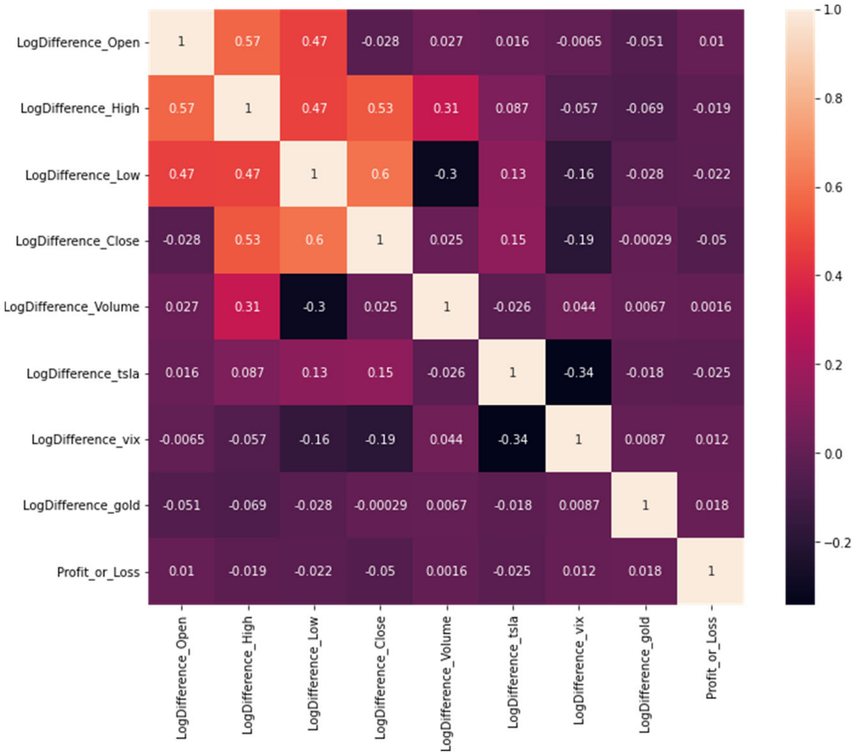

Fig. 2 Correlation Map.

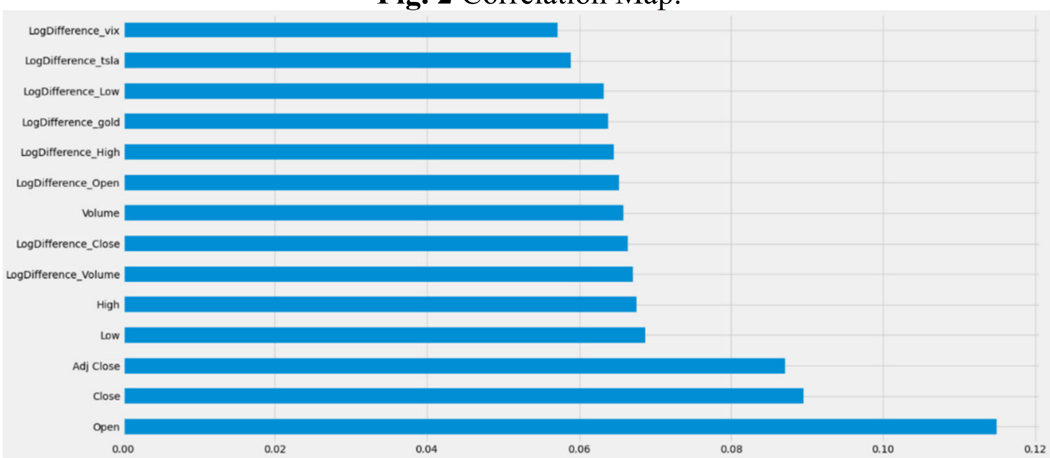

Fig. 3 Feature Importance

In addition, according to the analysis on the feature importance (displayed in Figure 3), the open and close prices have the most importance related to the response variable on return. The two have a feature importance score exceeding 0.8, which is greater than what other 6 variables have.

3.2 Comparions of the accuracy

The first five years' data (2016-07-01 to 2021-07-01) are separate as the training set, and the remaining year's data (2021-07-01 to 2022-07-01) is the test set. The models are not given the data of Bitcoin return in the test set, but are asked to do the predictions of the return by using the data of predictors in the test set. Applying the confusion matrix to the prediction results gives the accuracy rates of each model.

Based on this standard of comparison, the simplest algorithm (i.e., Ordinary Least Squares model) turns out to have the best performance. With the linear regression algorithm, the OLS model reaches an accuracy rate of 55.0685%. It is not very accurate but it exceeds the possibility of a random coin flip successfully. As followed, the LightGBM model is accurate for 50.6849% of the time. The one with a lower accuracy rate is the LSTM model, which has a rate of 50.1393% and barely exceeds the benchmark of a random coin flip. These three models above at least show some effectiveness in predicting Bitcoin return, with over 50% accuracy rate. The Random Forest model has the worst performance by comparison, and it only has an accuracy rate of 49.8630%, not even reaching the 50% standard line.

3.3 Explanation

The comparative study points out some uncommon results of using different statistical and machine learning models to do Bitcoin forecasting. Firstly, all four models have accuracy rates that are not very high, around 50% and not over 60%. It's partly likely caused by the unconventional way in which the data is preprocessed. Many previous studies did prediction on the Bitcoin price, so they used absolute values of the Bitcoin prices. In contrast, this study aims to predict the return of Bitcoin and scales the data in order to do a binary classification. Hence, the difference in how the daily data of Bitcoin is preprocessed can affect the prediction results and thereby the prediction accuracy. Moreover, additional features have been added to analyses. Specifically, the log difference of the VIX index has a very weak correlation with the Bitcoin return but is still considered for predicting the return. Therefore, additional features can add difficulty to the prediction and lower the accuracy rate to some extent. In general, an accuracy rate over 50% can still indicate the usefulness of the model, as it exceeds the benchmark of prediction.

Secondly, the LSTM model is not the one with the best performance in predicting Bitcoin return, which may overturn the claim that the LSTM model has met the state of art. Shown in Figure 4, the error of prediction by LSTM is decreasing over training. However, for dealing with time series data on a daily basis and doing classification of only one day, there is possibility that the LSTM model cannot completely fulfill its strength. Sometimes, the simplest algorithm, similar to the linear regression model, can complete tasks very well and even outdo more complicated models.

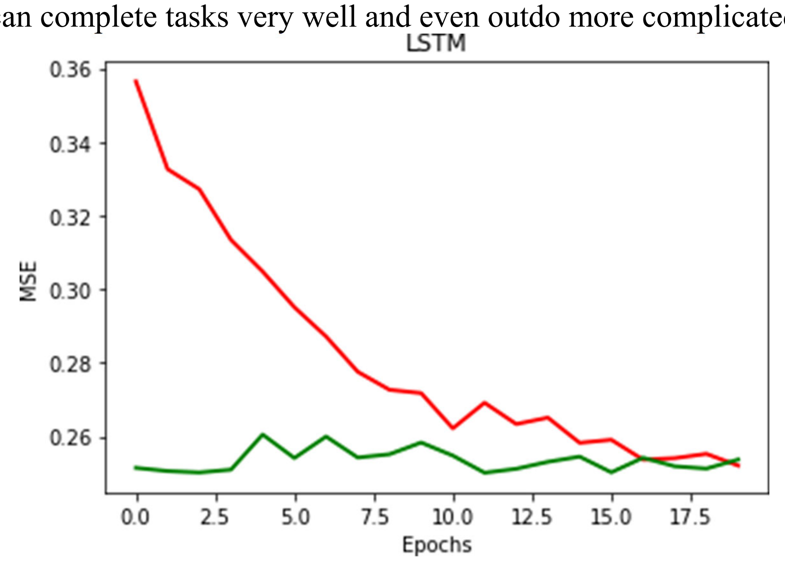

Fig. 4 LSTM Training Error

Although the LightGBM model has an accuracy rate slightly higher than that of the Random Forest model, their performances are very similar. On the one hand, it suggests that the LightGBM model is an advanced type of decision tree algorithm and does boosting to improve its performance. On the other hand, it implies that decision tree algorithms may have similar results of prediction, especially in the classification problems.

3.4 Limitations

There are some shortcomings and drawbacks of this study due to limited time and expertise. One limitation is that the research calculates the log difference between yesterday's value and today's value, but not extends the calculation to earlier days. In the real market, the Bitcoin prices from several days ago may still have an impact on today's price. The study doesn't include the log difference between several days ago and today into consideration, since it can greatly enlarge the dataset that requires more computational power. In addition, such impacts will decrease over time, so they are relatively less worth researching than the direct influence from yesterday's price. Another limitation is that the models predict the Bitcoin return on only one day. The response variable is the return on tomorrow, instead of the return on a series of days following tomorrow. If predicting on the return of a longer time range, some models (e.g., the Long Short-Term Memory Neural Network) may significantly improve their performances and become more useful in real-life investments in Bitcoin. Such limitations exist in this study, and future researchers who are interested can make the adjustments and improvements as suggested above.

4. Summary

In conclusion, this study constructs four models to predict on the return of Bitcoin and compares their performances. Overall, this study discovers multiple factors that have an impact on Bitcoin return. These factors include the daily changes in the high, low, close price of Bitcoin, and close price of Tesla stock, and gold price between yesterday and today. The OLS modeling, yet simple, has the highest accuracy rate. The LightGBM model and LSTM model have lower accuracy rates in order, but are still somehow effective. The Random Forest model, one type of decision tree algorithm, has similar prediction results with the LightGBM model but fails to solve the classification problem well. Nevertheless, the study has limitations on exploring the impacts of Bitcoin's past prices, as it includes influence from yesterday but not earlier days. Moreover, long-term return prediction, i.e., for several following days rather than tomorrow, ought to be presented. These results offer a guideline for investing in the cryptocurrency market and the macroeconomic environment. Besides, it demonstrates the potential for implementation of various machine learning algorithms in Bitcoin return forecasting. In the future, it can support more developments in related fields.

References

[1] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. Decentralized Business Review, page 21260, 2008.

[2] Karunya Rathan, Somarouthu Venkat Sai, and Tubati Sai Manikanta. Crypto-currency price prediction using decision tree and regression techniques. In 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), pages 190–194, 2019.

[3] Mohammad Ali and Swakkhar Shatabda. A data selection methodology to train linear regression model to predict bitcoin price. In 2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT), pages 330–335, 2020.

[4] Karasu S, Altan A, Saraç Z, et al. Prediction of Bitcoin prices with machine learning methods using time series data. 2018 26th signal processing and communications applications conference (SIU). IEEE, 2018: 1-4.

[5] Nicola Uras, Lodovica Marchesi, Michele Marchesi, and Roberto Tonelli. Forecasting bitcoin closing price series using linear regression and neural networks models. PeerJ Computer Science, 6: e279, 2020.

[6] Zheshi Chen, Chunhong Li, and Wenjun Sun. Bitcoin price prediction using machine learning: An approach to sample dimension engineering. Journal of Computational and Applied Mathematics, 365:112395, 2020.

[7] Isaac Madan, Shaurya Saluja, and Aojia Zhao. Automated bitcoin trading via machine learning algorithms. Retrieved from: http://cs229. stanford.edu/proj2014/Isaac% 20Madan, 20, 2015.

[8] Erdinc Akyildirim, Ahmet Goncu, and Ahmet Sensoy. Prediction of cryptocurrency returns using machine learning. Annals of Operations Research, 297(1):3–36, 2021.

[9] Ahmed Ibrahim, Rasha Kashef, and Liam Corrigan. Predicting market movement direction for bitcoin: A comparison of time series modeling methods. Computers & Electrical Engineering, 89:106905, 2021.

[10] Siddhi Velankar, Sakshi Valecha, and Shreya Maji. Bitcoin price prediction using machine learning. In 2018 20th International Conference on Advanced Communication Technology (ICACT), pages 144–147. IEEE, 2018.

[11] Xiaolei Sun, Mingxi Liu, and Zeqian Sima. A novel cryptocurrency price trend fore- casting model based on lightgbm. Finance Research Letters, 32:101084, 2020.

[12] Chih-Hung Wu, Chih-Chiang Lu, Yu-Feng Ma, and Ruei-Shan Lu. A new forecasting framework for bitcoin price with lstm. In 2018 IEEE International Conference on Data Mining Workshops (ICDMW), pages 168–175, 2018.

[13] Ebru Seyma Karakoyun and AO Cibikdiken. Comparison of arima time series model and lstm deep learning algorithm for bitcoin price forecasting. In The 13th multidisciplinary academic conference in Prague, volume 2018, pages 171–180, 2018.

[14] Hari Krishnan Andi. An accurate bitcoin price prediction using logistic regression with lstm machine learning model. Journal of Soft Computing Paradigm, 3(3):205– 217, 2021.

[15] Yan Li and Wei Dai. Bitcoin price forecasting method based on cnn-lstm hybrid neural network model. The journal of engineering, 2020(13): 344–347, 2020.

[16] Yiqing Hua. Bitcoin price prediction using arima and lstm. In E3S Web of Conferences, volume 218, page 01050. EDP Sciences, 2020.

[17] Sean McNally, Jason Roche, and Simon Caton. Predicting the price of bitcoin using machine learning. In 2018 26th euromicro international conference on parallel, distributed and network-based processing (PDP), pages 339–343. IEEE, 2018.

[18] Iman Abu Hashish, Fabio Forni, Gianluca Andreotti, Tullio Facchinetti, and Shiva Darjani. A hybrid model for bitcoin prices prediction using hidden markov models and optimized lstm networks. In 2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), pages 721–728. IEEE, 2019.

[19] Wei Chen, Huilin Xu, Lifen Jia, and Ying Gao. Machine learning model for bitcoin exchange rate prediction using economic and technology determinants. International Journal of Forecasting, 37(1):28–43, 2021.

[20] Takuya Shintate and Luk´aˇs Pichl. Trend prediction classification for high frequency bitcoin time series with deep learning. Journal of Risk and Financial Management, 12(1):17, 2019.

[21] Suhwan Ji, Jongmin Kim, and Hyeonseung Im. A comparative study of bitcoin price prediction using deep learning. Mathematics, 7(10):898, 2019.