Predicting cryptocurrency investment suitability using machine learning techniques

Xiaoke Song

Department of Mathematics, University of California at Berkeley, Berkeley, 94720, United States

20020618sxk@berkeley.edu

Abstract. The study aims to predict the close prices of four different cryptocurrencies (Bitcoin, Ethereum, Dogecoin, and Cardano) using machine learning techniques and determine which of these cryptocurrencies is suitable for investment. To achieve this goal, we used two popular gradient boosting algorithms: Extreme Gradient Boosting (XGBoost) and Light Gradient-Boosting Machine (LightGBM). Prediction accuracy of the trained model is evaluated by Mean Absolute Error (MAE) generated by the methodology of Cross-Validation. Our results show that both XGBoost and LightGBM can effectively predict the close prices of the four cryptocurrencies, with LightGBM achieving slightly better performance in terms of prediction accuracy. Based on our analysis, we were able to identify which cryptocurrencies were suitable for investing and provide recommendations for potential investors. Overall, our study highlights the potential of machine learning techniques in predicting cryptocurrency close prices and identifying suitable investment opportunities.

Keywords: cryptocurrencies, XGBoost, LightGBM, prediction.

1. Introduction

Cryptocurrencies have gained significant attention in recent years as an alternative form of digital currency [1]. Among the many different cryptocurrencies available, Bitcoin (BTC), Ethereum (ETH), Dogecoin (DOGE), and Cardano (ADA) are some of the most well-known and widely traded [2]. Figure 1 shows how the volume of four selected cryptocurrencies in the market changed from January 1, 2021 to March 31, 2022, also reflecting the high-demand for the selected cryptocurrency in the market. As the popularity of cryptocurrencies continues to grow, there is increasing interest in predicting their close prices and identifying which ones are suitable for investment [3].

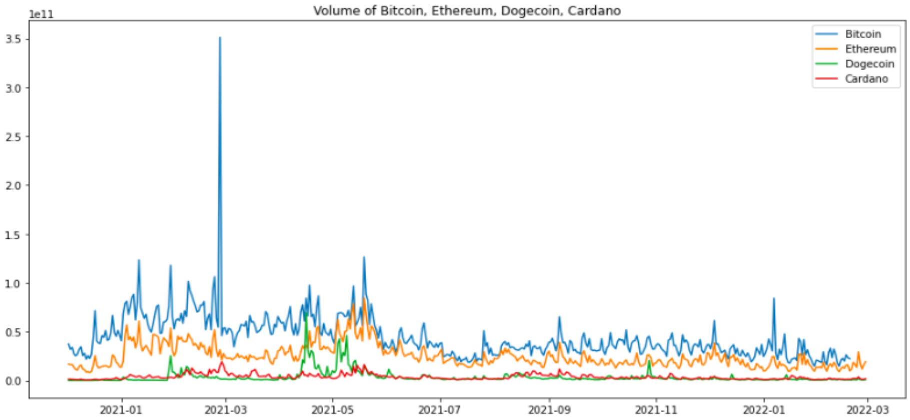

Figure 1. Volume of Bitcoin, Ethereum, Dogecoin, and Cardano.

In this research paper, we aim to predict the close prices of four different cryptocurrencies (BTC, ETH, DOGE, and ADA) using machine learning techniques and determine which of these cryptocurrencies is suitable for investment. Specifically, we will utilize two popular gradient boosting algorithms, XGBoost [4] and LightGBM [5], to make our predictions. Gradient boosting algorithms have been widely used in a variety of applications, including predicting stock prices [6] and forecasting energy demand [7], and have shown promising results in these domains. To the best of our knowledge, there have been few studies that have focused on using machine learning techniques to predict cryptocurrency close prices and identify suitable investment opportunities [8]. Therefore, this study aims to fill this gap by providing a comprehensive analysis of the predictability of close prices for the four selected cryptocurrencies and identifying which ones are suitable for investment.

The whole paper is structured as follows. The first section provides an introduction to the study. The second section discusses the fluctuations in the close price of the selected cryptocurrencies and the importance of predicting them. The third section outlines the methodology of the study, beginning with data processing. The fourth section presents and discusses the experimental results. The last section of the paper, section five, will present the conclusion.

2. Exploratory data analysis

2.1. Data description

In this study, we collected daily close price data for the four selected cryptocurrencies from January 1, 2021 to March 31, 2022. The data was obtained from Kaggle [9-12], a widely used website that provides real-time market data for various cryptocurrencies. The daily close price data is important to analyze the fluctuations in the prices of the four selected cryptocurrencies and to develop prediction models for each cryptocurrency. It is important to consider the fact that the original data for each selected cryptocurrencies may have different time periods, as can be seen in Figure 2. This is because the four cryptocurrencies have different origin times and may have been introduced to the market at different times. To control for this variable and ensure that the number of training and testing data is almost the same for each selected cryptocurrency, we extracted the last year’s worth of data for each cryptocurrency. This would allow them to have a comparable amount of data for each cryptocurrency and would help to eliminate any bias that may be introduced by using data from different time periods.

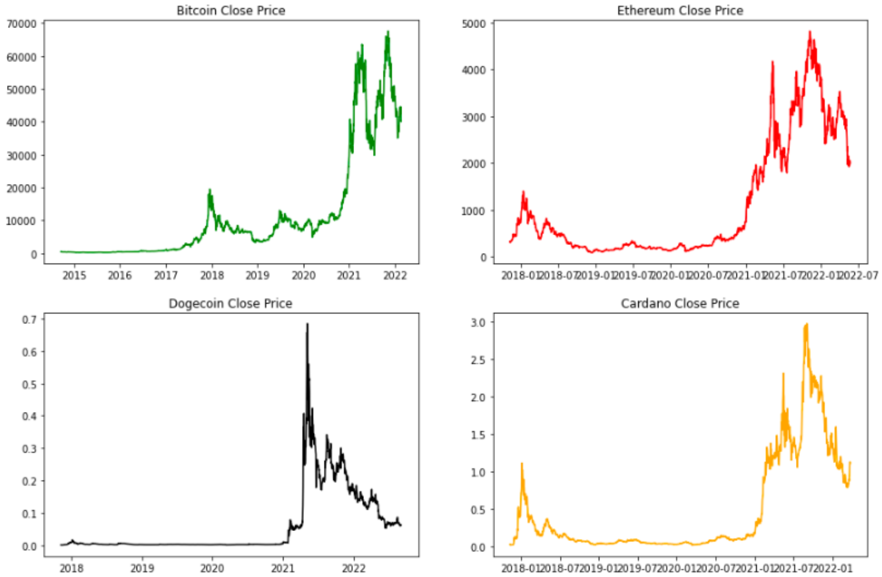

Figure 2. Close prices variation for original data.

2.2. Exploratory data analysis

From both Figure 2 and Figure 3, the close price of BTC, ETH, DOGE, and ADA has varied significantly over the last several years. Bitcoin is the first and most well-known cryptocurrency [13]. Its price has been highly volatile, with significant fluctuations over the last several years. In late 2017, the price of BTC reached an all-time high of nearly $20,000, before falling significantly in 2018 [14]. Since then, the price of BTC has recovered somewhat, but it has continued to experience significant fluctuations. Ethereum is a decentralized platform that runs smart contracts: applications that run exactly as programmed without any possibility of downtime, censorship, fraud or third-party interference [15]. The close price of ETH has also been highly volatile over the past few years. In late 2017, the price of ETH reached nearly $1,400, before falling significantly in 2018 [14]. Dogecoin is a cryptocurrency that was created as a joke in 2013, but has since gained a significant following [16]. The close peak of price of DOGE occurred in late 2021, $0.70. Cardano is a decentralized public blockchain and cryptocurrency project that is focused on providing a secure and scalable platform for the development of decentralized applications [17]. In late 2021, the price of ADA reached an all-time high of nearly $1.40.

It is clear from the analysis of both figures that the close price of the four selected cryptocurrencies is highly volatile. Thus, due to the fluctuations in their close prices in recent years, predicting the close price of these cryptocurrencies can be important for making informed investment, identifying market trends, and understanding the factors driving the price of a particular cryptocurrency.

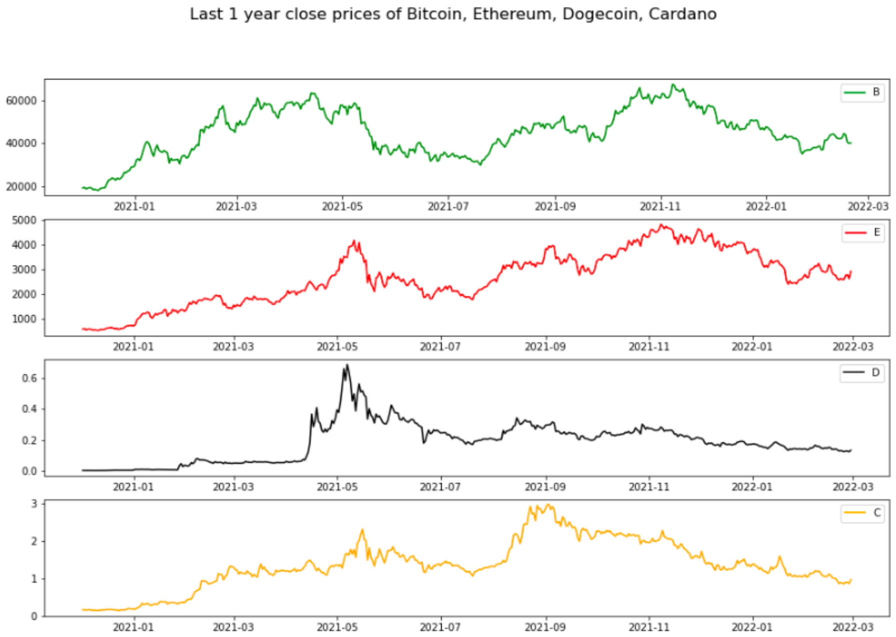

Figure 3. Close prices variation over the last year.

3. Methodology

To achieve our research goal of predicting the close prices of four different cryptocurrencies (BTC, ETH, DOGE, and ADA) and identifying which ones are suitable for investment, we followed the methodology described below.

3.1. Data processing

The data processing part of the study involves several steps to prepare the data for analysis and model training. One of these steps is to replace the “none” and “null” values in the close price data for the selected cryptocurrencies. The prediction models that are mentioned, XGBoost, and LightGBM, are known to be sensitive to missing or incomplete data. Therefore, it is important to ensure that the data used to train and test these models is as clean and complete as possible.

To replace the “none” and “null” values, we used techniques called imputation and interpolation. Imputation involves replacing missing or incomplete data with a substitute value, such as the mean or median of the data. Interpolation involves estimating the missing or incomplete data based on the surrounding data points. By replacing the “none” and “null” values, we are able to ensure that the data is complete and accurate, which can help to improve the performance of the prediction models.

3.2. Prediction model

To predict the close prices of the four cryptocurrencies, we used two popular gradient boosting algorithms: XGBoost [4] and LightGBM [5].

XGBoost (Extreme Gradient Boosting) and LightGBM (Light Gradient Boosting Machine) are both gradient boosting algorithms that can be used for predictive models and are particularly effective on tabular data. Both algorithms use decision trees as weak learners and build an ensemble model by adding decision trees sequentially, with each tree correcting the mistakes of the previous tree.

However, there are several key differences between XGBoost and LightGBM. LightGBM is generally faster than XGBoost, especially on large datasets. This is because LightGBM uses a more efficient tree building algorithm as can be seen from Figure 4 and 5 [18], which reduces the time it takes to fit the model [5]. Moreover, LightGBM is more memory-efficient than XGBoost, which means it can handle larger datasets without running out of memory. This is because LightGBM uses a technique called gradient-based one-side sampling (GOSS), which reduces the size of the data that needs to be loaded into memory during training [5]. Overall, LightGBM is generally considered to be a more efficient and scalable alternative to XGBoost, and is often the preferred choice for large-scale machine learning tasks. On the other hand, both algorithms can be effective in our study, and the best choice for a particular task will depend on the specific needs of the task at hand.

In this study, both XGBoost and LightGBM are both efficient to predict the close prices of the selected cryptocurrencies. We have compared the performance of these algorithms on the same dataset and evaluated the performance of the algorithms using various evaluation metrics, such as mean absolute error (MAE).

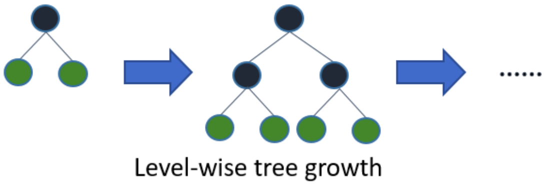

Figure 4. Algorithm behind XGBoost.

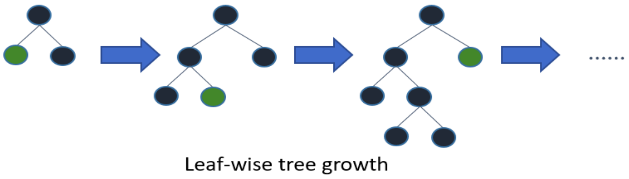

Figure 5. Algorithm behind LightGBM.

3.3. *Evaluation metrics*

We implemented cross-validation techniques [19] to ensure the robustness and generalizability of our models. Specifically, we employed a k-fold cross-validation approach, dividing the training set into k smaller subsets and iteratively training and evaluating the models on each subset. This allowed us to account for any potential biases or variations in the data and to obtain more accurate estimates of model performance. In our study, to evaluate the performance of the trained models, we used 5-fold cross-validation. In this approach, the data is divided into five equal-sized folds, and the model is trained and evaluated on each fold in turn. The final evaluation score is obtained by averaging the scores across all folds.

4. Results and discussion

4.1. *Experimental results*

We gathered a dataset containing historical close price data for each of the four cryptocurrencies. This data was then split into a training set and a testing set, with the training set used to train the XGBoost and LightGBM models and the testing set used to evaluate their performance. Figure 6 and Figure 7 respectively show the comparisons for Bitcoin between the original close prices and test predicted close prices with XGBoost and LightGBM separately. Furthermore, depending on these two figures, the LightGBM model is a more accurate prediction model than XGBoost. We used the trained models to make predictions on the close prices of the four cryptocurrencies in the testing set and compared these predictions to the actual close prices to evaluate the accuracy and reliability of the models. Figures 8 and 9 show how we compare the predictions to the actual close prices in scatter plots. The purpose of this study was to compare the performance of two machine learning algorithms, XGBoost and LightGBM, in predicting the close prices of four different cryptocurrencies over the next ten days. The cryptocurrencies included Bitcoin (BTC), Ethereum (ETH), Dogecoin (DOGE), and Cardano (ADA). Figures 10, 11, 12, 13 respectively show the predictions of close prices of BTC, ETH, DOGE, and ADA with two different prediction models. The blue line represents the prediction result generated by XGBoost, and the red line stands for the outcome predicted using LGBM.

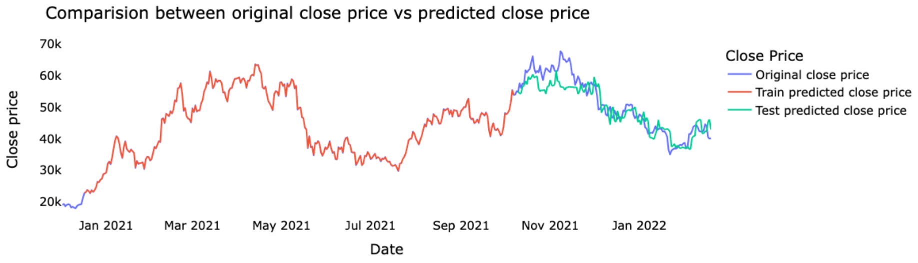

Figure 6. Actual and forecasting prices of BTC with XGBoost.

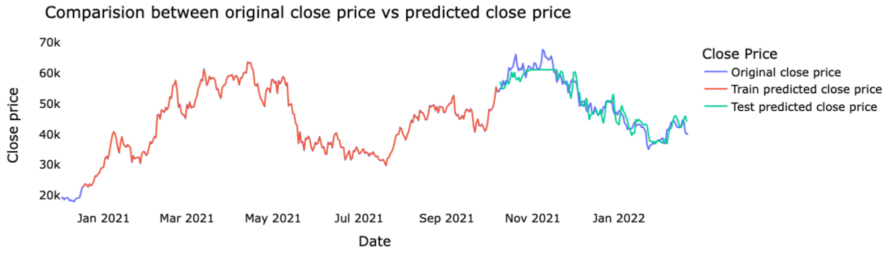

Figure 7. Actual and forecasting prices of BTC with LightGBM.

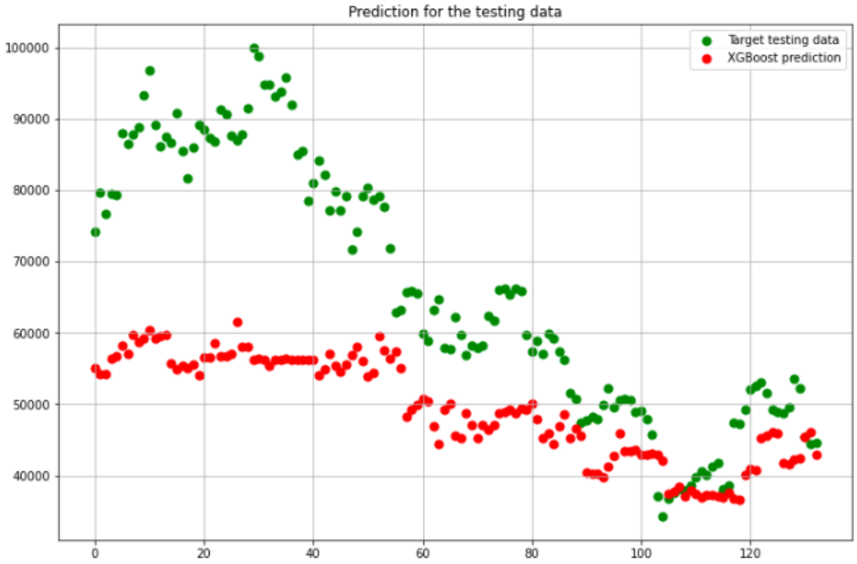

Figure 8. Prediction vs. actual value of BTC with XGBoost.

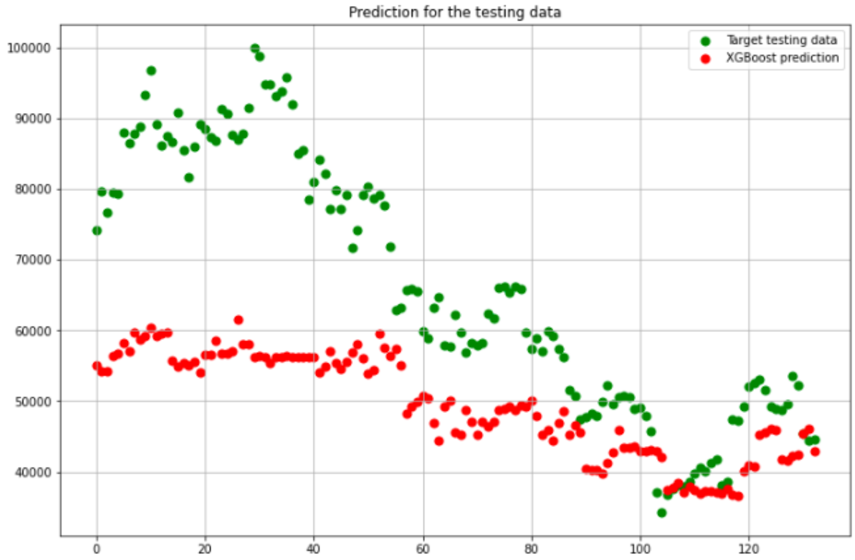

Figure 9. Prediction vs. actual value of BTC with LightGBM.

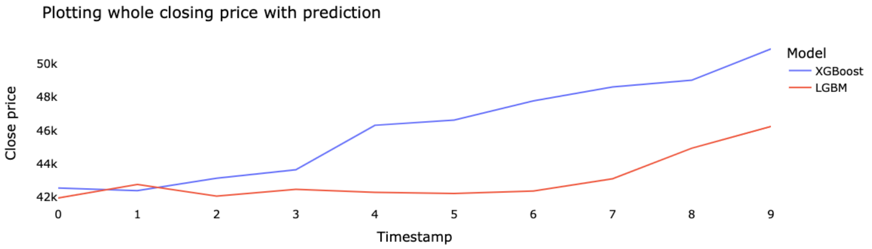

Figure 10. Prediction of BTC.

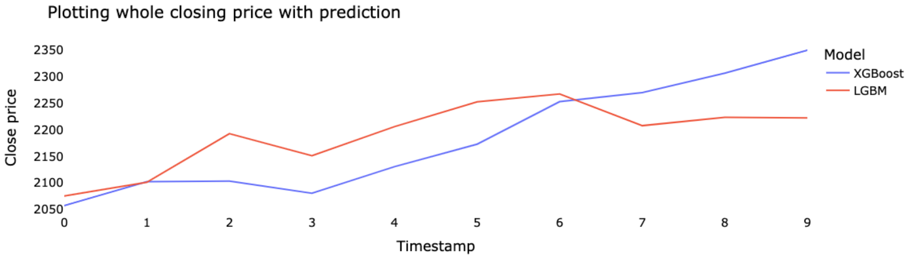

Figure 11. Prediction of ETH.

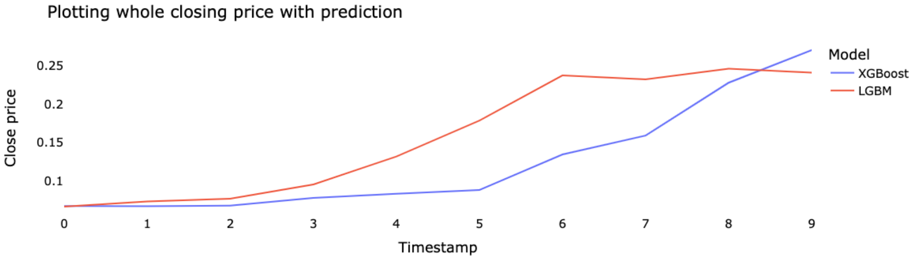

Figure 12. Prediction of DOGE.

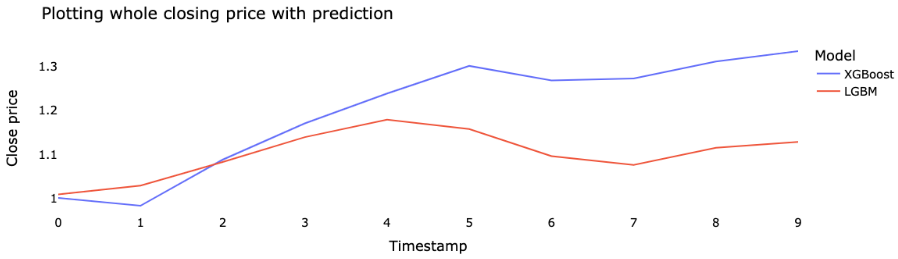

Figure 13. Prediction of ADA.

4.2. Discussion

In general, the results of this experiment suggest that both XGBoost and LIghtGBM can be effective in predicting the close prices of cryptocurrencies, but the performance of the models may vary depending on the specific cryptocurrency being considered. The results of the experiment showed that the close prices of all four cryptocurrencies were increasing. However the close prices of BTC and DOGE showed a consistent upward trend, while the close prices of ETH and ADA fluctuated significantly. When using XGBoost to make predictions, the model was able to accurately predict the close prices of BTC and DOGE with a high degree of accuracy. However, the model struggled to accurately predict the close prices of ETH and ADA, with the predictions for these cryptocurrencies showing a larger degree of variability. On the other hand, the LightGBM model performed well in predicting the close prices of all four cryptocurrencies, with the predictions for ETH and ADA showing a higher degree of accuracy compared to the XGBoost model.

5. Conclusion

This study has shown that the XGBoost and LightGBM machine learning algorithms are able to accurately predict the close prices of Bitcoin (BTC), Ethereum (ETH), Dogecoin (DOGE), and Cardano (ADA) to some extent. In general, the close prices of all four cryptocurrencies showed an increase over the short-term period studied. The close prices of BTC and DOGE were consistently increasing, suggesting that these cryptocurrencies have the potential to generate high returns and may be suitable for investment. On the other hand, the close prices of ETH and ADA showed significant fluctuations, indicating that there may be some risk involved in investing in these cryptocurrencies. In order to make informed investment decisions, it is important to carefully consider the potential risks and rewards of investing in cryptocurrencies. This study has provided insights into the trends and patterns in the close prices of BTC, ETH, DOGE, and ADA, which can be useful in making informed investment decisions. However, it is important to note that cryptocurrency prices are highly volatile and can be affected by a wide range of factors. As such, it is essential to conduct thorough research and carefully consider the potential risks and rewards before making any investment decisions.

References

[1] Böhme, R., Christin, N., Edelman, B., Moore, T., & Rik, V. (2015). Bitcoin: Economics, technology, and governance. Journal of Economic Perspectives, 29(2), 213-238.

[2] Coinmarketcap. (2021). Cryptocurrency market capitalizations. Retrieved from https://coinmarketcap.com/

[3] Lee, S., Kim, J., & Han, I. (2020). A study on the predictability of cryptocurrency prices using machine learning techniques. Expert Systems with Applications, 147, 113643.

[4] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 785-794).

[5] Ke, Q., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T.-Y. (2017). LightGBM:

A highly efficient gradient boosting decision tree. In Advances in neural information processing systems (pp. 3149-3157).

[6] Huang, J., Li, J., & Li, H. (2018). Stock price prediction with technical indicators using SVM. Neural Computing and Applications, 30(8), 2407-2414.

[7] Xie, Y., Chen, Y., & Gao, L. (2020). Forecasting hourly electricity demand using gradient boosting decision tree. Energy, 214, 117934.

[8] Kim, D., Kwon, Y., & Han, I. (2018). Predicting cryptocurrency prices using machine learning. In International Conference on Advanced Data Mining and Applications (pp. 365-377). Springer, Cham.

[9] Bitcoin Price Dataset. (2022, February 19). Kaggle. https://www.kaggle.com/datasets/meetnagadia/bitcoin-stock-data-sept-17-2014-august-24-2021

[10] Ethereum Crypto Price. (2022, May 24). Kaggle.

[11] https://www.kaggle.com/datasets/ranugadisansagamage/ethereum-crypto-price

[12] Dogecoin Historical Data. (2022, September 4). Kaggle. https://www.kaggle.com/datasets/dhruvildave/dogecoin-historical-data

[13] Cardano Data. (2022, March 25). Kaggle. https://www.kaggle.com/datasets/varpit94/cardano-data

[14] Nakamoto, S. (2008). Bitcoin: A Peer-to-Peer Electronic Cash System. Retrieved from https://bitcoin.org/bitcoin.pdf

[15] CoinDesk. (2021). Bitcoin Price Index. Retrieved from https://www.coindesk.com/price/bitcoin

[16] Buterin, V. (2014). A next-generation smart contract and decentralized application platform. Ethereum White Paper. Retrieved from https://ethereum.org/greeter

[17] Palmer, B. (2021). The story of Dogecoin, the joke cryptocurrency that became too successful to kill. The Verge. Retrieved from https://www.theverge.com/2021/5/9/22420981/dogecoin-history-explainer-what-is

[18] Cardano Foundation. (2021). About Cardano. Retrieved from https://www.cardano.org/about/

[19] Saha, S. (2022, November 14). XGBoost vs LightGBM: How Are They Different. neptune.ai. https://neptune.ai/blog/xgboost-vs-lightgbm

[20] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In International joint conference on artificial intelligence (pp. 1137-1145).