Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/49/20241414

Cryptocurrency price prediction based on Xgboost, LightGBM and BNN

Guoxuan Sun

Business School, Hohai University, Nanjing 211100, China

2163910123@hhu.edu.cn

Abstract. The valuation and prediction of cryptocurrency prices have become increasingly important in the financial market. Therefore, this study aims to focus on the selection and evaluation of machine learning models for cryptocurrency valuation. Thus, two types of machine learning models, gradient boosting trees (Xgboost and LightGBM) and neural networks, are compared to determine their effectiveness in generating features for cryptocurrency valuation. Additionally, correlation tests are conducted to identify the most suitable input variables for the models. The results demonstrate that the generated features have a significant impact on the accuracy of machine learning predictions for cryptocurrency prices. It highlights the potential of machine learning models in accurately predicting and evaluating the value of cryptocurrencies. Overall, the findings of this study contribute to the understanding of the role of machine learning in cryptocurrency valuation and provide valuable insights for investors and researchers. By leveraging machine learning techniques, investors can make informed decisions and develop effective investment strategies in the cryptocurrency market. This study contributes to cryptocurrency valuation research. Leveraging machine learning enables informed decisions and effective investment strategies. Furthermore, the findings inform the development of advanced machine learning models and algorithms for cryptocurrency valuation.

Keywords: Cryptocurrency valuation, prediction, machine learning, gradient boosting trees, neural networks.

1. Introduction

Cryptocurrency is a digital currency that uses a public transaction ledger, known as the blockchain, to record transactions. Since the concept of cryptocurrency was introduced in the 1980s, an increasing number of people have been involved in the research and creation of cryptocurrencies [1]. In 2008, the first decentralized cryptocurrency based on blockchain technology, Bitcoin, was introduced. It provided a development standard for many subsequent digital currencies. After the introduction of Bitcoin, an increasing number of cryptocurrencies entered the market, such as Ethereum, Ripple, Tether, Cardano, Stellar, Litecoin, and Zcash [2]. On January 1st, 2023, there are 22,163 cryptocurrencies in the global cryptocurrency market, with a total market capitalization of approximately $798.688 billion. These blockchain-based cryptocurrencies offer secure, transparent, traceable, and immutable transactions, which is why individual investors, large institutions, and companies are heavily investing in them [3].

The price of cryptocurrencies not only affects the commodity trading market but also the investment market [4]. Therefore, the price of cryptocurrencies has always been a subject of curiosity for researchers

DOI: 10.54254/2755-2721/49/20241414

worldwide. The price of cryptocurrencies is unstable and depends on various factors such as transaction costs, mining difficulty, market trends, popularity, prices of alternative coins, stock markets, emotions, and certain legal factors [5]. Thus, it is crucial to have a model that can accurately predict the cryptocurrency market with the same level as the stock market. Additionally, having real-time knowledge of price changes can bring higher profits and lower investment risks for investors [6].

In recent years, the application of machine learning techniques to predict and evaluate the value of cryptocurrency assets has gained significant attention in the academic community. Catania et al. delved into the challenges of forecasting cryptocurrencies, particularly addressing model and parameter instability, utilizing methods such as support vector machines and random forests [7]. Kumar and Shah introduced a novel approach by leveraging Long Short-Term Memory networks (LSTM) for Bitcoin price forecasting, comparing its efficacy with other machine learning methodologies [8]. While the primary focus of Bouri et al. was on Bitcoin's potential as a hedge against global uncertainty, their research also touched upon the role of machine learning in assessing the value of cryptocurrency assets [9]. Oyedele et al. investigate the performance evaluation of genetic algorithm tuned Deep Learning and boosted tree-based techniques in predicting cryptocurrency closing prices, with CNN model showing the least mean average percentage error and highest explained variance score compared to other models [6]. Parekh et al. propose a hybrid and robust framework, DL-Gues, for cryptocurrency price prediction, considering the interdependency of cryptocurrencies and market sentiments. They validate the framework using price history and tweets of various cryptocurrencies, including Dash and Bitcoin-Cash. Collectively, these studies underscore the growing importance and potential of machine learning in the realm of cryptocurrency valuation and prediction [10].

In the trading market of cryptocurrencies, the trading information of cryptocurrencies is similar to that of stocks [11]. In the stock exchange market, one often uses the basic information of stock trading to generate some characteristics to aid the investment decisions. This paper assumes that cryptocurrencies have some of the trading nature of stocks, thus generating some feature-aided machine learning predictions. The results show that some of the generated features have a good effect on the prediction of machine learning. This study tries to find out which model is more effective for the valuation of cryptocurrencies by trying two types of models: gradient lifting tree and neural network. At the same time, this paper attempts two kinds of gradient lift trees, Xgboost and LightGBM, and evaluates the prediction results and model performance.

2. Data and method

This paper utilizes the Global Cryptocurrency Database, a comprehensive and curated dataset available on the Kaggle platform. This dataset encompasses over 7,500 cryptocurrencies, each paired with the US dollar (USD). It provides extensive information including the Coin Name, Trading Symbol, Date of Price (accurately time-stamped for precise tracking), Opening Price, Highest Price, Lowest Price, Closing Price, and Adjusted Closing Price. Leveraging this dataset, the paper employs certain stock market evaluation indicators to generate data features that aid in machine learning. These indicators include the High-Low Range, ATR, 5-day simple moving average, 20-day index moving average, RSI, MACD, among others. Finally, the input features are screened based on variable correlation, which will be elaborated upon subsequently.

The principle of Xgboost is rooted in the Gradient Boosting Decision Tree (GBDT) algorithm. Xgboost further enhances and refines GBDT to achieve optimization. The fundamental concept behind Xgboost lies in iteratively boosting the model's predictive capability through gradient ascent. In each iteration, Xgboost calculates the residuals between the current model's predicted values and the actual values. It then trains a new decision tree to capture and reduce these residuals. By repeatedly adding new decision trees to the model, Xgboost effectively minimizes the residual errors. Through this iterative process, Xgboost progressively enhances the model's predictive power. LightGBM is an advanced machine learning algorithm that builds upon the Gradient Boosting Decision Tree (GBDT) framework. It enhances the efficiency and scalability of GBDT through the integration of two novel techniques. The first technique, known as Gradient-based One-Side Sampling (GOSS), selectively excludes data

DOI: 10.54254/2755-2721/49/20241414

instances with smaller gradients, effectively reducing the dataset size. This approach accelerates the training process without compromising accuracy. The second technique, Exclusive Feature Bundling (EFB), reduces the dimensionality of the feature space by grouping mutually exclusive features together. By bundling these features, the computational complexity is significantly reduced, resulting in improved efficiency. In summary, LightGBM leverages the power of GOSS, EFB, and gradient boosting to enhance the efficiency and scalability of GBDT. BNN Bayesian Neural Networks (BNNs) are artificial neural networks trained using Bayesian statistical methods. Unlike traditional neural networks with fixed parameters, it introduce randomness by treating the parameters as random variables. The training process of BNN involves estimating the posterior distribution of the parameters through Bayesian inference, which entails setting the prior distribution based on domain knowledge or experience and updating it with observed data. In the prediction process, BNNs sample from the posterior distribution to obtain prediction results, enabling the quantification of uncertainty and providing interpretability. Overall, BNNs leverage Bayesian inference to train and predict, treating parameters as random variables, and offer advantages in adapting to diverse data scales and mitigating overfitting.

Parameter scanning is a crucial step in optimizing model performance. Through systematic exploration of the parameter space, the optimal parameter combination can be identified, leading to improved accuracy and generalization. Given the limited parameter space, this study employs grid search and cross-validation to evaluate the performance of each parameter combination across diverse data subsets. Common parameters include learning rate, maximum tree depth, subsampling ratio, column sampling ratio, among others. The trained models are evaluated using various performance metrics, including Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Squared Error (MSE). These metrics are employed to assess the effectiveness of the proposed models.

3. Results and discussion

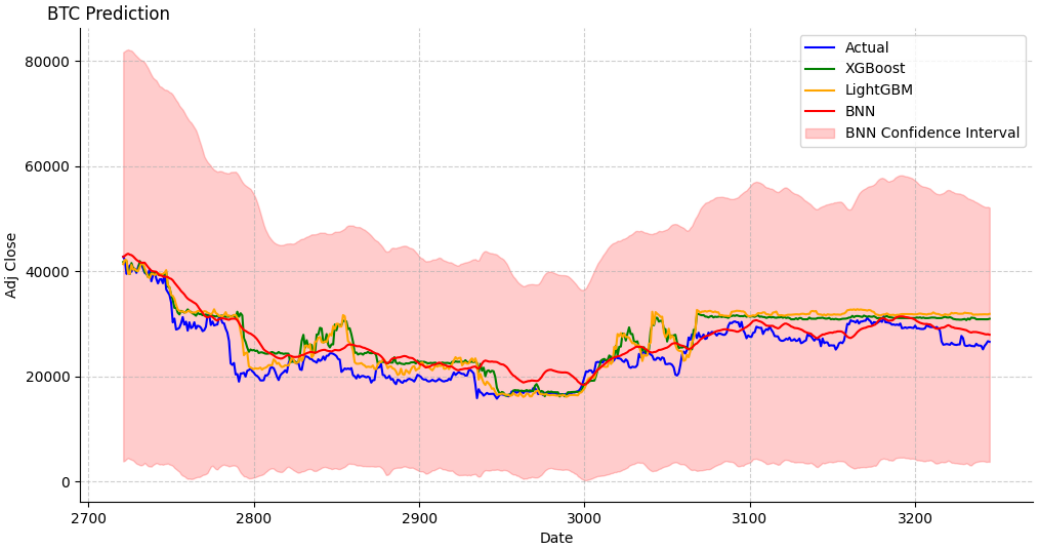
Feature engineering serves as a critical step in constructing high-performance machine learning models. To forecast the daily adjusted market price of the cryptocurrency, this paper utilizes statistics from the cryptocurrency exchange market as the primary data source. Given the availability of open market data, it is imperative to investigate the utilization of such data and its derived information for predictive purposes. The subsequent sections will elucidate the process of constructing appropriate data through feature construction, extraction, transformation, selection, as well as the creation of lagging data sets, and the partitioning of training and test sets. The dataset from the open market is limited in its features and not suitable for machine learning prediction. Therefore, this study focuses on feature construction. Initially, time series features are generated based on the data. Additionally, considering the similarities between the cryptocurrency trading market and the secondary trading market of stocks, this study incorporates evaluation indicators from the stock market to generate additional features.

In the feature extraction phase, the time data is primarily segmented. Subsequently, in the feature transformation phase, the mean value of the data is standardized to enhance the model's generalization capability and mitigate the bias impact among features. It is worth noting that tree models such as Xgboost and LightGBM are insensitive to the scale and range of features, thus eliminating the necessity of standardizing their respective features. Feature selection plays a pivotal role in constructing models that are both high-performing and interpretable, while also ensuring efficiency. This study employs the embedded selection method, utilizing both filtering and wrapping selection techniques to screen the features. Through multiple screening iterations, 20 out of the initial 31 features are selected as input for further analysis in this paper. This paper aims to predict the adjusted price of the cryptocurrency on the eighth day using data from the first seven days. To achieve this, a lag data set is constructed. Subsequently, the data set is divided into a training set and a test set. The training data set comprises 75% of the total data set, while the test data set comprises 25% of the total data set.

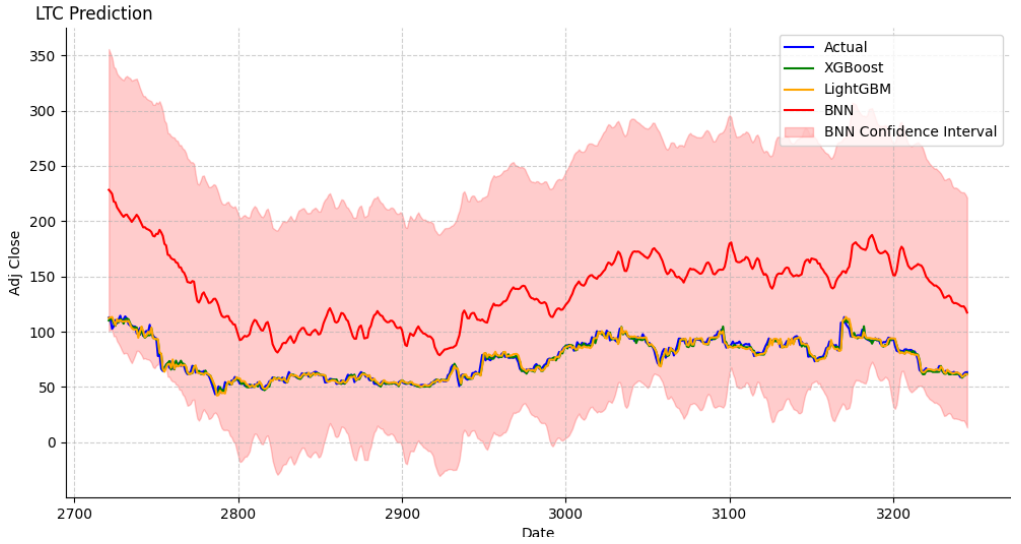
In the quest to predict cryptocurrency asset valuation, three distinct models were employed: Xgboost, LightGBM, and Bayesian Neural Network (BNN). Each model was trained on the same dataset and evaluated based on their predictive accuracy on a test set. The performance metrics used for evaluation

DOI: 10.54254/2755-2721/49/20241414

were Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Squared Error (MSE).

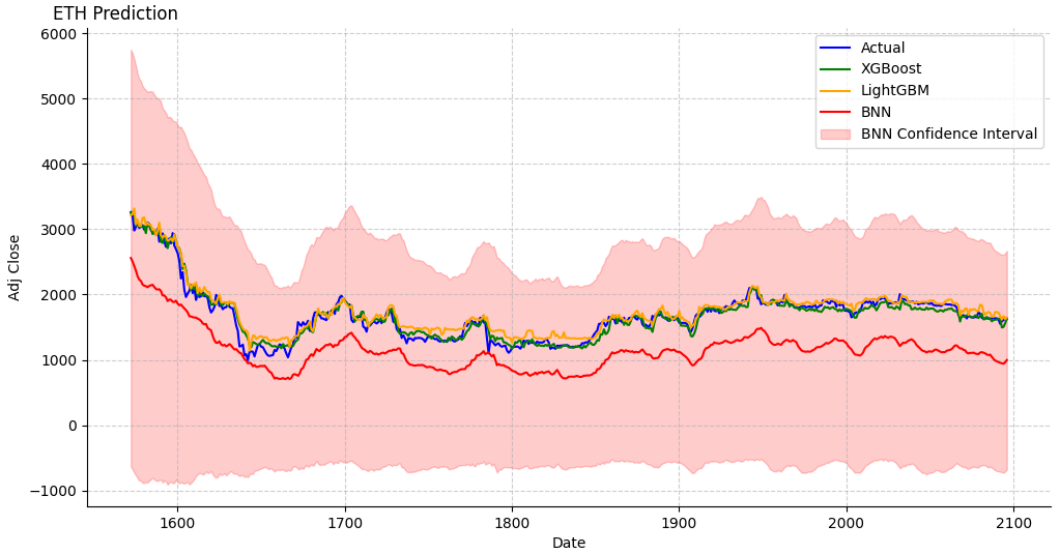


(a)Price prediction of BTC using Xgboost,LightGBM and BNN.



(b)Price prediction of LTC using Xgboost,LightGBM and BNN.

DOI: 10.54254/2755-2721/49/20241414



(c)Price prediction of ETC using Xgboost,LightGBM and BNN.

Figure 1. Prediction results of three types of cryptos (Photo/Picture credit: Original).

The results given in Fig. 1 and Table 1 underscore the effectiveness of gradient boosting algorithms, particularly Xgboost, in predicting cryptocurrency asset valuations. The superior performance of Xgboost can be attributed to its robustness to overfitting, its ability to handle missing data, and its capacity to model non-linear relationships effectively. LightGBM's performance, while commendable, was overshadowed by Xgboost. This might be due to the specific hyperparameters chosen or the nature of the dataset. However, given its efficiency and speed, LightGBM remains a viable option for large datasets or real-time predictions. The underperformance of the BNN model is noteworthy. While BNNs offer the advantage of quantifying uncertainty, their training can be intricate, requiring careful hyperparameter tuning and potentially more extensive training data. The high MAPE suggests that the model might have been overconfident in its incorrect predictions. However, the ability to provide confidence intervals is a unique feature that can be invaluable, especially when making decisions based on predictions. In conclusion, while Xgboost emerged as the most accurate model for this dataset, the choice of model should be based on the specific requirements of the application. If speed and efficiency are paramount, LightGBM might be more suitable. If quantifying uncertainty is crucial, despite its current performance, further tuning and experimentation with BNNs might yield better results.

Table 1. Test results of Xgboost & LightGBM & BNN

Model	Currency	MAPE	MAE	RMSE	MSE
	BTC	8.81547946%	2435.14416324	3053.99011659	9326855.63221627
Xgboost	ETH	4.50295675%	71.16473528	95.94090535	9204.65731880
	LTC	3.50816841%	3.75292358	5.87991289	34.57337555
	BTC	9.47079039%	2732.40891799	3422.30555321	11712175.29954738
LightGBM	ETH	6.19807429%	91.03123158	119.67813827	14322.85678015
	LTC	3.83224775%	4.24136556	6.84120757	46.8021210
	BTC	21.39565090%	5268.22780124	6065.81215840	36794077.14105232
BNN	ETH	12.49766489%	220.19256104	267.34408372	71472.85910145
	LTC	73.27843508%	72.44863413	76.80087365	5898.37419386

4. Limitations and prospects

Although the predictive models used in cryptocurrency asset valuation have shown promising results, it is essential to recognize their limitations. A comprehensive understanding of these limitations is crucial

DOI: 10.54254/2755-2721/49/20241414

for interpreting the results accurately and making informed decisions based on the predictions. The performance of the models is highly dependent on the quality and representativeness of the data. In this study, the models were trained using historical cryptocurrency data, which may not fully capture the intricate dynamics and evolving nature of the cryptocurrency market. Factors such as regulatory changes, market sentiment, and technological advancements are challenging to quantify and integrate into the models, potentially constraining their predictive accuracy. Model Assumptions Each model is constructed based on specific assumptions and constraints. For example, Xgboost and LightGBM assume an additive relationship between input features and the target variable, represented by decision trees. Although these assumptions are generally applicable, they may not hold true for all cryptocurrency assets or market conditions. Likewise, the BNN model assumes a specific prior distribution for the underlying data, which may not always be accurate. The performance of the models can vary across different datasets and time periods. Factors such as the time range, feature selection, and target variable in the training data can influence the predictive ability of the models. Therefore, it is important to exercise caution when applying these models to new and unseen data, as their performance may differ.

Despite their limitations, the results obtained from predictive models offer valuable insights into the valuation of cryptocurrency assets. Based on these findings, there are several avenues for future research and development aimed at improving the predictive accuracy and applicability of these models. Incorporating Additional Features Expanding the feature set in the models has the potential to enhance their predictive power. Incorporating factors like social media sentiment, news sentiment, and macroeconomic indicators can offer a more comprehensive understanding of market dynamics. Furthermore, incorporating domain-specific features, such as blockchain transaction data or network metrics, can provide valuable insights into the underlying fundamentals of specific cryptocurrencies. Ensemble Methods Ensemble methods, such as stacking or boosting, have the potential to enhance overall predictive performance by combining the predictions of multiple models. By leveraging the strengths of different models and mitigating their weaknesses, ensemble methods can yield more robust and accurate predictions. Adaptive Models Developing adaptive models that can effectively respond to changing market conditions and evolving data patterns is essential in the dynamic cryptocurrency market. These models, equipped with the ability to automatically adjust parameters or update training data in real-time, have the potential to enhance predictive accuracy and responsiveness to market changes.

5. Conclusion

Machine learning models have emerged as a promising tool for accurately predicting and evaluating the value of cryptocurrencies, akin to their application in stock trading. Numerous studies have delved into the utilization of diverse machine learning models, such as gradient boosting trees and neural networks, to generate features that facilitate cryptocurrency valuation. These models have exhibited commendable prediction results and performance. However, it is crucial to acknowledge that the cryptocurrency market is influenced by a multitude of factors, including transaction costs, market trends, and legal considerations, which can constrain the accuracy of predictions. Notwithstanding these limitations, the application of machine learning techniques in cryptocurrency valuation and prediction has garnered significant attention within the academic community. Real-time knowledge of price fluctuations can yield higher profits and mitigate investment risks for investors. The ability to precisely forecast cryptocurrency prices can provide valuable insights for individual investors, large institutions, and companies. Moreover, comprehending the factors that impact cryptocurrency prices can contribute to a better understanding of the overall market dynamics and potentially inform investment strategies. In conclusion, machine learning models offer a promising approach to predict and evaluate the value of cryptocurrencies. Further research and advancements in this field can lead to more precise predictions and a deeper comprehension of the cryptocurrency market.

References

[1] Rice M 2019 Digital commons vol 14.

DOI: 10.54254/2755-2721/49/20241414

[2] Tanwar S, Patel N P, Patel S N, Patel J R, Sharma G and Davidson I E 2021 Deep Learning-Based Cryptocurrency Price Prediction Scheme With Inter-Dependent Relations IEEE Access vol 9 pp 138633-138646.
[3] Patel M M, Tanwar S, Gupta R and Kumar N 2020 A Deep Learning-based Cryptocurrency Price Prediction Scheme for Financial Institutions Journal of information security and applications vol 55 pp 102583.
[4] Deepika P and Kaur E R 2017 International Journal of Trend in Research and Development vol 4(4) pp 4-6.
[5] Sovbetov Y 2018 Journal of Economics and Financial Analysis vol 2(2) pp 1-27.
[6] Oyedele A A, Ajayi A O, Oyedele L O, Bello S A and Jimoh K O 2023 Performance evaluation of deep learning and boosted trees for cryptocurrency closing price prediction Expert Systems with Applications vol 213 p 119233.
[7] Catania L, Grassi S and Ravazzolo F 2019 Forecasting cryptocurrencies under model and parameter instability International Journal of Forecasting vol 35(2) pp 485-501.
[8] Kumar D and Rath S K 2020 Predicting the Trends of Price for Ethereum Using Deep Learning Techniques Artificial Intelligence and Evolutionary Computations in Engineering Systems pp 103-114.
[9] Fang L, Bouri E, Gupta R and Roubaud D 2019 Does global economic uncertainty matter for the volatility and hedging effectiveness of Bitcoin? International Review of Financial Analysis vol 61 pp 29-36.
[10] Parekh R, Patel N P, Thakkar N, Gupta R, Tanwar S, Sharma G and Sharma R 2022 DL-GuesS: Deep Learning and Sentiment Analysis-Based Cryptocurrency Price Prediction IEEE Access vol 10 pp 35398-35409.
[11] Liang J, Li L, Chen W and Zeng D 2019 Towards an Understanding of Cryptocurrency: A Comparative Analysis of Cryptocurrency, Foreign Exchange, and Stock IEEE International Conference on Intelligence and Security Informatics (ISI) pp 137-139.