USING STATISTICAL MODEL AND MACHINE LEARNING FOR CRYPTOCURRENCY PRICE PREDICTION

LE QUOC KHANH[1], LE GIA KIET[2], AND NGUYEN THI THUY[3]
[1]Faculty of Information Systems, University of Information Technology, (e-mail: 21520283@gm.uit.edu.vn)
[2]Faculty of Information Systems, University of Information Technology, (e-mail: 21522255@gm.uit.edu.vn)
[3]Faculty of Information Systems, University of Information Technology, (e-mail: 25122662@gm.uit.edu.vn)

ABSTRACT Cryptocurrency price prediction is a challenging yet crucial task in the dynamic realm of financial markets. In this article, we explore the efficacy of various statistical models and machine learning algorithms to forecast cryptocurrency prices. Leveraging a diverse toolkit including Linear Regression, ARIMAX (AutoRegressive Integrated Moving Average with Exogenous Variables), Recurrent Neural Networks (RNN), Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM) networks, Vector Autoregression (VAR), XGBoost, and LightGBM, we aim to capture the complex patterns inherent in cryptocurrency price movements. Technical indicators serve as the primary features, offering insights into market sentiment and trends. Through rigorous evaluation and comparison of these models, we seek to discern their strengths and weaknesses in accurately predicting cryptocurrency prices, contributing to the advancement of predictive analytics in the volatile domain of digital assets.

INDEX TERMS *Cryptocurrency price, Forecasting, Technical Indicators, Linear regression, ARIMAX, RNN, GRU, LSTM, VAR, XGBoost, LightGBM*

I. INTRODUCTION

Cryptocurrency has become a significant and rapidly growing market, with thousands of different digital currencies available for trading. With the rise of cryptocurrency, there has been an increasing demand for accurate and reliable price prediction models. This paper aims to explore the use of statistical models and machine learning algorithms to predict the price of three popular cryptocurrencies: Bitcoin, Ethereum, and Dogecoin.

The price of cryptocurrencies is influenced by a variety of factors, including market demand, investor sentiment, and global economic conditions. These factors make predicting the price of cryptocurrencies a challenging task. However, with the use of statistical models and machine learning algorithms, it is possible to identify patterns and trends in the data that can be used to make more accurate predictions.

This paper scrutinizes a comprehensive suite of models, encompassing traditional statistical methods and advanced machine learning algorithms. Specifically, we explore the application of Linear Regression, ARIMAX (AutoRegressive Integrated Moving Average with Exogenous Variables), Recurrent Neural Networks (RNN), Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM) networks, Vector Autoregression (VAR), XGBoost, and LightGBM. We will evaluate the performance of these models using historical price data and assess their ability to accurately predict future price movements.

Technical indicators provide valuable insights into market sentiment and trends, aiding in predicting price movements. Integrating these indicators into machine learning models offers a promising strategy to leverage the abundant cryptocurrency market data. This study employs various stock market evaluation indicators, such as moving averages and momentum indicators, to derive features for machine learning. The goal is to enhance predictive accuracy by enabling algorithms to identify meaningful patterns.

In summary, by using statistical models and machine learning algorithms to predict the price of cryptocurrencies, investors and traders can make more informed decisions and potentially increase their returns. Additionally, these models and algorithms can be used by cryptocurrency exchanges and financial institutions to manage risk and improve their trading strategies.

II. RELATED WORKS

There have been multiple studies conducted on the use of statistical models and machine learning algorithms for cryptocurrency price prediction. In a article by Gouxuan Son (2024) [1], two models that we concerned in three distinct models employed, Xboost and LightGBM, for predicting

Bitcoin prices was conducted. Another paper by Ziyang Yuan (2023) [2] used KNN, XGBoost and LightGBM to predict the price of Gold and Bitcoin Price. Especially, the investigation found that LightGBM is more effective and space-saving. Haydier, Albarwari and Ali compared between VAR and ARIMAX Time Series Models in Forecasting [3]. The results showed that the VAR model is better than the ARIMAX model for their observed data depending on the MSE criterion.

In another article, [4] the authors compared the performance of LSTM and GRU models in predicting Bitcoin prices. Additionally, they approved that the GRU model was able to capture long-term dependencies in the Bitcoin price data, while the LSTM model struggled to do so. Meanwhile, [5] compared and proved that LSTM is also better than RNN.

Recent research emphasizes the pivotal role of feature selection in developing effective and interpretable models for cryptocurrency price prediction. Huang, Huang, and Ni (2019) [6] showcased this significance by integrating high-dimensional technical indicators to predict bitcoin returns. Similarly, Mudassir et al. (2020) [7] employed a machine learning approach using such features for time-series forecasting of Bitcoin prices. These studies underscore the increasing acknowledgment of technical indicators as valuable features in enhancing predictive accuracy within cryptocurrency markets.

Based on insights gleaned from numerous prior literature studies, this research aims to predict cryptocurrency prices utilizing a diverse array of predictive models, including Linear Regression, ARIMAX, RNN, GRU, LSTM, VAR, XGBoost, and LightGBM.

III. MATERIALS

A. DATASET

The data is collected from finance.yahoo.com, downloading the daily data of Bitcoin, Ethereum and Dogecoin from 2018-Mar-01 to 2024-Mar-01, including close price, open price, high price, low price, Adjust close and the volume of trading coins with Currency in USD.

Including attributes

- Date: Represents the date of the trading day.
- Open: Refers to the opening price of Bitcoin/Ethereum/Dogecoin on that particular day.
- High: Indicates the highest price reached by Bitcoin/Ethereum/Dogecoin during that day.
- Low: Represents the lowest price reached by Bitcoin/Ethereum/Dogecoin during that day.
- Close: Refers to the closing price of Bitcoin/ETH/Dogecoin on that day.
- Adj Close: Represents the adjusted closing price, which accounts for factors like dividends and stock splits.
- Volume: Refers to the trading volume of Bitcoin on that day, i.e., the total number of Bitcoin/Ethereum/Dogecoin units traded.

As the goal is to forecast the price, only data relating to column “Close" (USD) will be analyzed

B. DESCRIPTIVE STATISTICS

	DOGE	BTC	ETH
Count	2193	2193	2193
Mean	0.067	22727.117	1298.755
Median	0.053952	19191.63086	1213.599976
Mode	0.002653	6741.75	none
Min	0.002	3236.762	84.308
25%	0.003	8368.83	228.73
50%	0.054	19191.631	1213.6
75%	0.084	35510.289	1924.566
Max	0.685	67566.828	4812.087
Std	0.09	16431.978	1145.552
Variance	0.008166821	270009888.9	1312289.819
Kurtosis	6.38367187	-0.652902729	-0.181911084
Skewness	2.186196643	0.702529234	0.81204465
Range	0.683	64330.066	4727.779

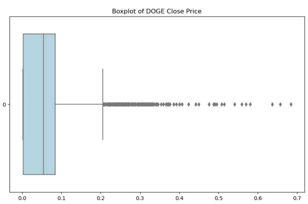

FIGURE 1. DOGE’s Box Plot

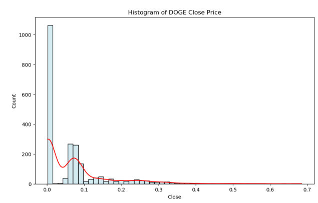

FIGURE 2. DOGE’s Histogram

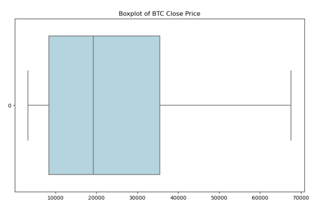

FIGURE 3. BTC’s Box Plot

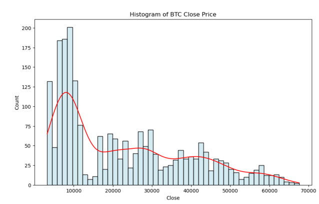

FIGURE 4. BTC’s Histogram

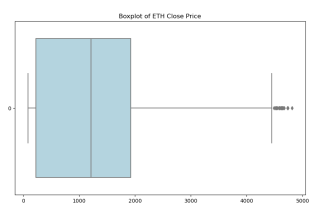

FIGURE 5. ETH’s Box Plot

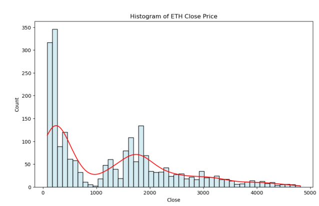

FIGURE 6. ETH’s Histogram

- Across all three cryptocurrencies, there are remarkable differences in mean values, indicating diverse price levels. Moreover, the considerable range between minimum and maximum values highlights the wide fluctuations in prices, portraying substantial volatility within the market.
- The high standard deviation, positive kurtosis, and skewness values suggest non-normal distributions with

fat tails and right skewness. This indicates frequent occurrence of outliers and a tendency for prices to be skewed towards higher values, which means the occurrence of high prices isn't significant.

- These measures contributes to a summary that it is potential for high profits but also heightened risks when investing in these three cryptocurrencies.

IV. METHODOLOGY

V. RESULT

A. EVALUATION METHODS

VI. CONCLUSION

A. SUMMARY

B. FUTURE CONSIDERATIONS

ACKNOWLEDGMENT

REFERENCES

[1] Sun, G. (2024). Cryptocurrency price prediction based on Xgboost, LightGBM and BNN. *Applied and Computational Engineering*, 49(1), 273–279. https://doi.org/10.54254/2755-2721/49/20241414

[2] Yuan, Z. (2023). Gold and Bitcoin Price Prediction based on KNN, XGBoost and LightGBM Model. *Highlights in Science, Engineering and Technology*, 39, 720–725. https://doi.org/10.54097/hset.v39i.6635

[3] Haydier, E., Albarwari, N., & Ali, T. H. (2023, December 1). The Comparison Between VAR and ARIMAX Time Series Models in Forecasting. *Iraqi Journal of Statistical Sciences*. https://doi.org/10.33899/iqjoss.2023.181260

[4] Dutta, A., Kumar, S. S., & Basu, M. (2020). A gated Recurrent unit approach to bitcoin price prediction. *Journal of Risk and Financial Management*, 13(2), 23. https://doi.org/10.3390/jrfm13020023

[5] Gunarto, D. M., Saadah, S., & Utama, D. Q. (2023). Predicting cryptocurrency price using RNN and LSTM method. *Jurnal Sistem Informasi Dan Komputer/Jurnal Sisfokom*, 12(1), 1–8. https://doi.org/10.32736/sisfokom.v12i1.1554

[6] Huang, J.-Z., Huang, W., & Ni, J. (2019). Predicting bitcoin returns using high-dimensional technical indicators. *The Journal of Finance and Data Science*, 5(3), 140–155. https://doi.org/10.1016/j.jfds.2018.10.001

[7] Mudassir, M., Bennbaia, S., Ünal, D., & Hammoudeh, M. (2020). Time-series forecasting of Bitcoin prices using high-dimensional features: a machine learning approach. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-020-05129-6

[8] Metsileng, L. D., Moroke, N. D., & Tsoku, J. T. (2018). Modelling the BRICS exchange rates using the Vector Autoregressive (VAR) model. *Journal of Economics and Behavioral Studies*, 10(5(J)), 220–229. https://doi.org/10.22610/jebs.v10i5(j).2511