FORECASTING CRYPTOCURRENCIES PRICES USING TIME SERIES APPROACH.

LE NGOC YEN KHOA[1], DANG LUU HA[2], DOAN THI MY LINH[3], TRAN MINH QUANG[4], AND NGUYEN HUU PHUNG[5]

[1]Faculty of Information Systems, University of Information Technology, (e-mail: 21522224@gm.uit.edu.vn)
[2]Faculty of Information Systems, University of Information Technology, (e-mail: 21520798@gm.uit.edu.vn)
[3]Faculty of Information Systems, University of Information Technology, (e-mail: 21522285@gm.uit.edu.vn)
[4]Faculty of Information Systems, University of Information Technology, (e-mail: 21522519@gm.uit.edu.vn)
[5]Faculty of Information Systems, University of Information Technology, (e-mail: 21522451@gm.uit.edu.vn)

ABSTRACT This study utilizes a variety of forecasting models to analyze and predict the exchange rates of Binance Coin (BNB), Bitcoin (BTC), and Ethereum (ETH) against USD from 2019 to 2024. By employing models such as TimesNet, Random Forest, CNN-LSTM, Bagging Model, and VAR to forecast fluctuations, along with linear regression, our research aims to provide valuable insights for investors, financial analysts, and policymakers in navigating the volatile cryptocurrency market.

INDEX TERMS **Keyword - Time series, statistical method, cryptocurrency exchange rates, machine learning, deep learning**

I. INTRODUCTION

In today's era, the cryptocurrency market has become an integral part of the global financial system, and monitoring and forecasting the exchange rates of cryptocurrencies against USD is crucial. In this scope, we focus on the three most popular types of cryptocurrencies: Binance Coin (BNB), Bitcoin (BTC), and Ethereum (ETH).

This study will employ a range of forecasting models to understand and predict fluctuations in the exchange rates of these cryptocurrencies against USD. Specifically, we will use linear regression models to analyze long-term trends, and then utilize neural networks such as TimesNet and CNN-LSTM to explore complex relationships between factors and forecast short-term fluctuations.

Additionally, we will also apply machine learning models such as Random Forest, Bagging Model, and VAR (Vector Autoregression) to ensure flexibility and accuracy in forecasting. The combination of these models will provide a comprehensive view of trends and fluctuations in cryptocurrency exchange rates against USD, aiding investors and financial experts in making informed and effective investment decisions.

II. RELATED WORKS

There are some research works done on this topic:

Yecheng Yao et al. [1] proposed a deep learning way to predict the price of cryptocurrencies. The paper emphasized on using Long Short-Term Memory (LSTM) and Recurrent Neural Network (RNN) for predictive analysis of cryptocurrency price. The factors that are considered are market cap, volume, circulating supply, and maximum supply. The authors also gave idea about some factors like political environment and human regulations.

Muhammad Ali Nasir et al. [2] proposed a model where a cryptocurrency returns, and volumes are forecasted using search engines. A weekly dataset from 2013 to 2017 is used and captured a dependence structure by using empirical approaches like VAR framework, a copulas approach etc.

The main focus is given on prediction of values of bitcoins and Ethereum. The model contains a linear model to take input tweets and Google trends data. This model measures the overall interest regarding the cryptocurrency price rise or fall in terms of volume. The authors also support the idea of using linear regression model along with lagged variables to get the better results.

Aggarwal A. et al. [3] discussed about various parameters affecting bitcoin price prediction based on Root Mean Square Error (RMSE) using various deep learning models like Convolutional Neural Network (CNN), Long ShortTerm Memory (LSTM) and Gated Recurrent Unit (GRU).

Alessandretti, L., ElBahrawy, A., Aiello, L.M. and Baronchelli, A. [4] have tested the performance of three models in predicting daily cryptocurrency price for 1681 cryptocurrencies. Two of the models are based on the gradient boosting decision treeljbljkbs and one is based on Long ShortTerm memory. In all the cases, investment portfolios are built based on the predictions and performance is compared based on the return on investment.

Mittal R. et al. [5] have predicted the cryptocurrency prices based on their open, low and high cost. Some Machine Leaning Algorithms are used to predict the cryptocurrency daily price changes. The Dataset used in this paper is relatively smaller. Multivariate Linear Regression has been used to predict the Highest and lowest prices of cryptocurrency. R is used as a platform to predict the cryptocurrency prices based on the dependent features. Various statistical measures like F Score and p value are used to test the model's accuracy and probability respectively.

Suhwanji. et al. [6] have developed and compared various deep learning-based Bitcoin price prediction models using Bitcoin blockchain information. More specifically, they tested the state-of-the-art deep learning models such as deep neural networks (DNN), long short-term memory (LSTM) models, convolutional neural networks (CNN), deep residual networks (ResNet), and their combinations. For regression problems, LSTM slightly outperformed the other models, whereas for classification problems, DNN slightly outperformed the other models Although CNN and ResNet are known to be very effective in many applications, including sequence data analysis, their performance was not particularly good for Bitcoin price.

Phumudzo Lloyd Seabe. et al. [7] concluded three types of deep learning techniques—LSTM, GRU, and Bi-LSTM—were used to predict the prices of three major cryptocurrencies, as measured by their market capitalization: Bitcoin, Ethereum, and Litecoin. The results of the study showed that the Bi-LSTM model provided the most accurate predictions for all three currencies, followed by the GRU model. This suggests that the combination of forward and backward flows in bi-directional models improves the performance of time-series prediction.

Nguyen Dinh Thuan. et al. [8] concluded predicting the price of cryptocurrencies is a difficult job, requiring extensive and in-depth research on the cryptocurrency market. In addition, it also needs the support of machine learning and statistical modeling. in forecasting the next day's price of this currency. The combination of the model or the hybrid model leads to the predictive model with more effective quality as higher results, low error rate in detail.

III. MATERIALS

A. DATASET

The dataset used in this study was sourced from https://www.investing.com/, a reputable financial platform renowned for its comprehensive and up-to-date market information. This dataset covers the values of the three most popular cryptocurrencies in the world: Binance Coin, Ethereum, and Bitcoin against the US dollar from 01/03/2019 to 01/03/2024, providing a robust temporal scope for our analysis.

Each entry in the dataset includes key financial indicators:

Date: This indicates the date when the data was recorded.

Open: This represents the opening price of the respective cryptocurrency pair for that day, which is the price of the cryptocurrency when the market opens for trading.

High: This denotes the highest price that the respective cryptocurrency pair reached on that day.

Low: This signifies the lowest price that the respective cryptocurrency pair reached on that day.

Close: This indicates the closing price of the respective cryptocurrency pair for that day, which is the price of the cryptocurrency when the market closes for trading.

Adj Close: This is the adjusted closing price used to calculate adjustments such as stock splits, new stock issuances, stock dividends, etc. For some cryptocurrencies, the adjusted closing price and the closing price may be very similar or identical.

Volume: This indicates the total trading volume of the respective cryptocurrency pair on that day.

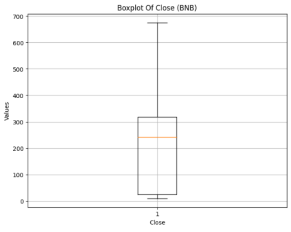
Since the objective is to forecast close prices, only data relating to the "Close" column (USD) will be processed.

By utilizing the data available on this platform, we ensure a reliable foundation for our analysis.

B. DESCRIPTIVE STATISTICS

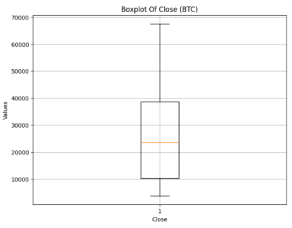
Bang 1. BNB, BTC, ETH's Descriptive Statistics

	BNB	BTC	ETH
Count	1828	1828	1828
Mean	212.5066	26000.4734	1489.504
Std	171.7517	16088.1356	1160.549
Min	9.3861	3761.5571	110.6058
25%	25.4486	10279.4622	259.3103
50%	242.5276	23651.3789	1574.2838
75%	318.8924	38693.0576	2158.4348
Max	675.6841	67566.8281	4812.0874

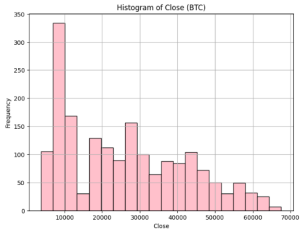


Hình 1. Binance Coin(BNB) Close Value's boxplot

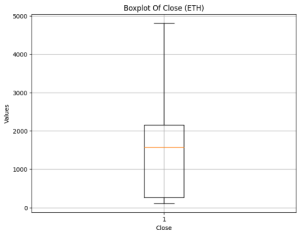
Hình 2. Binance Coin(BNB) Close Value's histogram



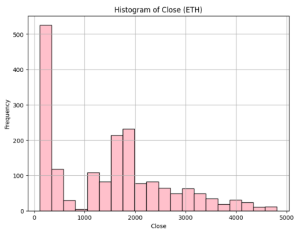
Hình 3. Bitcoin(BTC) Close Value's boxplot



Hình 4. Bitcoin(BTC) Close Value's histogram



Hình 5. Ethereum(ETH) Close Value's boxplot



Hình 6. Ethereum(ETH) Close Value's histogram

C. TOOL

During the research process, we found that there are many popular Python tools and libraries used for data analysis, supporting deep learning, and data visualization. After researching and selecting, we decided to use some main tools: numpy, pandas, sklearn, matplotlib.pyplot. The use of these tools has helped us define data, understand the meaning of data more deeply. At the same time, the ability to visualize data has greatly supported the process of understanding details as well as describing datasets clearly, providing broader perspectives in data exploration.

D. DATA SPLIT RATIO

In our analysis of time series data, we split the dataset into training and testing sets using different proportions: 70% for training and 30% for testing, 80% for training and 20% for testing, and 90% for training and 10% for testing. These ratios help us examine how the model's performance is affected by the distribution of data in each set.

The commonly used 7:3 ratio allocates 70% for training and 30% for testing, striking a balance between providing enough training data and ensuring distinct sets for fine-tuning and evaluation. Another option is the 8:2 ratio, which favors an 80% training set, beneficial for more complex models requiring a larger training dataset. In some cases, a cautious approach like the 9:1 ratio may be preferred, especially with a large dataset and a simpler model. This ratio ensures sufficient training data while providing a substantial testing set for performance evaluation.

E. MODEL EVALUATION

RMSE, or Root Mean Square Error, represents the square root of the average squared error in the predicted yi values. Essentially, it gauges the disparity between predicted and actual values. Lower RMSE values indicate superior predictive models.

Mean Absolute Percentage Error (MAPE) is a metric utilized to evaluate the accuracy of a forecasting or prediction model. It calculates the average percentage deviation between predicted and actual values, providing a measure of how well the model performs in terms of percentage error.

MSE, or Mean Squared Error, calculates the average of the squared differences between predicted and actual values in a model. It provides a measure of the overall accuracy by quantifying the magnitude of errors.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}|\frac{y_i - \hat{y}_i}{y_i}|$$

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$$

Where:

y_i is the observer value,

$\hat{y}_i$ is the predicted value,

n is the number of observers

IV. METHODOLOGY

A. LINEAR REGRESSION

Linear Regression is a widely used statistical machine learning method that models the linear relationship between a dependent variable (y) and one or more independent variables (X). In other words, it predicts the value of y based on the values of X. A multiple linear regression model has the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

Where:

- Y is the dependent variable (Target Variable).
- $X_1, X_2, \ldots, X_k$ are the independent (explanatory) variables.
- β_0 is the intercept term.
- $\beta_1, ..., \beta_k$ are the regression coefficients for the independent variables.
- ε is the error term.

B. AUTOREGRESSIVE INTEGRATED MOVING AVERAGE (ARIMA)

Theory: We know that most time series exhibit correlation between past values and the current value, with stronger correlation as the observations are closer in time. The ARIMA model aims to capture this correlation by introducing lag variables to create a forecasting model that fits the values of the time series well.

The ARIMA model stands for Autoregressive Integrated Moving Average. It represents a linear regression equation

of the input variables (also known as dependent variables in statistics) with two main components:

- **Autoregression (AR):** This component is represented by the AR term. It includes a set of lagged values of the current variable. The lag order represents how far back in time the series is shifted. The length of the lag in the AR process depends on the lag parameter p. Specifically, the AR process of the series Y is represented as follows:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \epsilon_t$$

- **Moving Average (MA):** The moving average process involves shifting or altering the average value of the series over time. Since our series is assumed to be stationary, the moving average process essentially represents a white noise series. The moving average process seeks a linear relationship between random errors (ϵ_t). This series must be a white noise series satisfying the following properties:

$$E(\epsilon_t) = 0$$
$$Var(\epsilon_t) = \sigma^2$$
$$Cov(\epsilon_t, \epsilon_s) = 0 \text{ for } t \neq s$$

The moving average process is represented in terms of white noise as follows:

$$Y_t = \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \ldots - \theta_q \epsilon_{t-q}$$

Integrated: The integrated process involves differencing or taking the first order of differencing. The common requirement for time series algorithms is that the series must exhibit stationarity. Most series tend to increase or decrease over time. Therefore, the correlation between them is not due to an actual relationship but rather a common correlation over time. By transforming the series into a stationary series, time-dependent factors are removed, making the series easier to forecast. To create a stationary series, the simplest method is to take the first order of differencing. The degree of differencing required to create a stationary series is called the order of integration. The differencing process for a series is as follows:

$$\text{First order difference: } Y'_t = Y_t - Y_{t-1}$$
$$\text{Differencing order } d: Y_t^{(d)} = Y_t^{(d-1)} - Y_{t-1}^{(d-1)}$$

The ARIMA(p, d, q) regression equation can be expressed as:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \ldots - \theta_q \epsilon_{t-q} + \epsilon_t$$

Therefore, in general terms, ARIMA is a combination of two processes: autoregressive and moving average. Past data is used to forecast future data. Before training the model, it's necessary to transform the series into a stationary series by taking the first order of differencing or using other methods like logarithm transformation or taking the difference of logs. Additionally, the model must adhere to strict conditions regarding the absence of autocorrelation and the residuals being white noise. This is the basic theory of econometrics. In the field of machine learning, the main concern is to select a model with the smallest forecasting error. Next, we will use the *vnquant* package, a package developed by our team to support the community in extracting stock data more conveniently.

C. VAR - VECTOR AUTOREGRESSION

Vector Autoregression (VAR) is a statistical model commonly used in econometrics, particularly when multiple time series influence each other. This means that the relationship between the time series is bidirectional. It estimates each equation of each time series variable based on the lagged values of the variable (p) and all other variables. In essence, it captures the interdependencies among multiple time series variables.

An AR(p) model equation typically takes the following form:

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t$$

where: - y_t is the variable of interest at time t. - c is a constant term. - ϕ_i are the coefficients to be estimated for each lag i. - ϵ_t is the error term assumed to be white noise.

The best prediction for variable y is a linear function of the variables x.

VAR models are widely used in various fields such as economics, finance, and environmental science to analyze multivariate time series data and make predictions based on the interdependencies among the variables.

References

[1] Yecheng Yao, Jungho Yi, Shengjun Zhai, Yuwen Lin, Taekseung Kim Guihongxuan Zhang, Leonard Yoonjae Lee, "Predictive Analysis of Cryptocurrency Price Using Deep Learning", International Journal of Engineering and Technology, Volume 7, Issue 3.27, pp. 258-264, 2018 [https://www.sciencepubco.com/index.php/ijet/article/view/17889]

[2] Muhammad Ali Nasir, Toan Luu Duc Huynh, Sang Phu Nguyen and Duy Duong, "Forecasting cryptocurrency returns and volume using search engines", Financial Innovation 5, Article number 2, 2019 [https://jfin-swufe.springeropen.com/articles/10.1186/s40854-018-0119-8]

[3] Aggarwal A., Gupta I., Garg N., and Goel A., "Deep Learning Approach to Determine the Impact of Socio-Economic Factors on Bitcoin Price Prediction", Twelfth International Conference on Contemporary Computing (IC3), 2019 [https://ieeexplore.ieee.org/document/8844928]

[4] Alessandretti, L.,ElBahrawy, A.,Aiello, L.M. and Baronchelli,A., 2018. Anticipating cryptocurrency prices using machine learning. Complexity, 2018. [https://www.hindawi.com/journals/complexity/2018/8983590/]

[5] Mittal R., Arora S. and Bhatia M. P. S., "Automated cryptocurrencies prices prediction using machine learning", International Journal on Soft Computing, Volume 8, Issue 4, pp. 1758-1761, 2018

[6] Suhwan Ji, Jongmin KimORCID and Hyeonseung Im, "A Comparative Study of Bitcoin Price Prediction Using Deep Learning", Department of Computer Science, Kangwon National University, Chuncheon-si, Gangwon-do 24341, Korea, 2019 [https://www.mdpi.com/2227-7390/7/10/898]

[7] Phumudzo Lloyd Seabe, Claude Rodrigue Bambe Moutsinga, andEdson Pindza, "Forecasting Cryptocurrency Prices Using LSTM, GRU, and Bi-Directional LSTM: A Deep Learning Approach", 2023 [https://www.mdpi.com/2504-3110/7/2/203]

[8] Nguyen Dinh Thuan, Nguyen Minh Nhut, Hoang Tung, Vu Minh Sang, "PREDICTING THE CLOSING PRICE OF CRYPTOCURRENCY USING HYBRID ARIMA, REGRESSION AND MACHINE LEARNING", Ky yeu Hoi nghi KHCN Quoc gia lan thu XIV ve Nghien cuu co ban va ung dung Cong nghe thong tin (FAIR), TP. HCM, ngay 23-24/12/2021