# PREDICT BITCOIN PRICE

1st Pham Nguyen Cao Triet
IS403.N23.HTCL
*University of Information Technology*
Ho Chi Minh city
19521050@gm.uit.edu.vn

***Abstract*** - Crypto has been a long-lasting and potential investment field attracting much investment in this field every year. In particular, favorite cryptos such as Bitcoin (BTC) have attracted many investments in recent years. The volatility of crypto is very unpredictable that makes investor have difficulty in investing this field. The goal of this paper is determine what accuracy the direction of crypto price in USD can be forecasted. This study will use models like Linear regression linear regression , ARIMA , SARIMA, LSTM forecasting the close price of BTC The comparison results will be based on two evaluation parameters: RMSE and MAPE.

***Keywords*** - Bitcoin prediction, machine learning, linear regression, ARIMA, SARIMA, LSTM.

## I. INTRODUCTION

Bitcoin is a digital currency and a decentralized form of money that was created in 2009 by an unknown person or group of people using the pseudonym Satoshi Nakamoto. It was the first cryptocurrency to be introduced, and it operates on a technology called blockchain. Bitcoin is not controlled by any central authority, such as a government or financial institution. Instead, it relies on a peer-to-peer network of computers to verify transactions and maintain the integrity of the system. This network is known as the blockchain, which is a public ledger that records all Bitcoin transactions. In recent years, along with the development of technology, the value of BTC has been increasing and has played a significant role in the investment community. Particularly, some countries, companies, and organizations have started accepting payments in the form of the virtual currency BTC, with Tesla being a notable example. These factors have attracted many researchers to engage in forecasting the price of Bitcoin. Regression models and machine learning algorithms are commonly used in cryptocurrency price prediction problems.

## II. RELATED WORK

Financial market forecasting is a widely researched area with conflicting evidence on market predictability and efficiency. Regression analysis is a well-established method for examining signals that can explain asset returns and predict profitability. Linear regression, a straightforward mathematical technique, is widely used for making predictions and easily adaptable to software and computation. It enables businesses to accurately analyze raw data, forecast future values and trends, and leverage relevant existing data.The ARIMA model is best suited for linear connections between present data and historical data, according to Robert et al. (1979) [2]. Additionally, Brockwell et al. (2001) hypothesized that the ARIMA model would provide more accurate predictions if the data were broken down by month [3]. Three key components make up the ARIMA model: the autoregressive component (AR), the stationary time series component (I), and the moving average component (MA) [4]. When forecasting time series, Gujarati (2006) and R. Carter Hill et al. (2011) recommend using the ARIMA model.[5] [6]Gradient boosted trees model is very advantageous especially in the context of price prediction for a number of reasons as follows. Firstly, it is not required to normalize the data in this case as it is sensitive to arithmetic range of data and features. Secondly, it is a very scalable machine learning model due to its construction process and finally, it is also a rule-based learning method . A number of works dealing with prediction and forecasting of sales as well as cryptocurrency prices in the literature have successfully employed gradient boosted trees model [7,8 ] Papers by Sean et al. utilizing LSTM [10] They suggest a method for determining the price of Bitcoin that combines Recurrent Neural Network, Long Short Term Memory, and Ruchi.Based on the historical pattern, Mittal et al. [11] offer an automated machine learning technique for predicting cryptocurrency prices (daily trend). Using LSTM, Chih-Hung et al. [12] developed a new framework for forecasting the price of bitcoin. They offered two different LSTM models (standard LSTM and LSTM with AR(2) model) with 208 records of data and compared their results to MSE, RMSE, MAE, and MAPE. A common stock market prediction model was created by Fei Qian et al.[13] based on LSTM under

various market-impacting factors, and for this study, they chose three stocks with comparable tendencies.

# III. METHODS

## A. *Data collection*

We using dataset Bitcoin from 01/06/2021 to 01/06/2023 (during 2 years).

## B. *Linear Regression*

Regression analysis is a tool for building mathematical and statistical models that characterize relationships between a dependent variable and one or more independent, or explanatory, variables, all of which are numerical. This statistical technique is used to find an equation that best predicts the y variable as a linear function of the x variables. A multiple linear regression model has the form: [5]

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

Where:

- **Y** is the dependent variable.
- **X_1,…X_k** are the independent (explanatory) variables.
- **β_0** is the intercept term.
- **β_1,… β_k** are the regression coefficients for the independen variables.
- **ε** is the error term.

## C. *ARIMA*

ARIMA stands for Auto Regressive Integrated Moving Average. It is a well-known forecasting model in financial and data science time series application [18]. The auto-regressive moving average (ARMA) models are used in stationary cryto data only, this model contains three combination models which are AR (p), MA (q) and I (d).

Auto Regressive (AR) is a model that predicts current values based on past values.

$$AR(p) = Yt = \varepsilon t + c + \alpha 1 Yt{-}1 + \alpha 2 Yt{-}2 + \cdots + \alpha p Yt{-}p$$

Difference I is the difference between present and past values.

First difference :

$$\Delta Y = Yt - Yt{-}1$$

Second difference:

$$\Delta Y = (Yt - Yt{-}1) - (Yt{-}1 - Yt{-}2)$$

$$= Yt - 2Yt{-}1 + Yt{-}2$$

Difference of order

$$\text{d: } \Delta Y = Yt - 2Yt{-}1 + Yt{-}d$$

A Moving Average (MA) is a linear model of the current value against past errors so that the process of changing the mean of a time series can be performed.

$$MA(q) = Yt = Ut + \beta 0 + \beta 1 Yt{-}1 \cdots + \beta q Yt{-}q$$

## D. *SARIMA*

This is where SARIMA (Seasonal Autoregressive Integrated Moving Average) models come in. As an extension of the ARIMA method, the SARIMA model not only captures regular difference, autoregressive, and moving average components as the ARIMA model does but also handles seasonal behaviour of the time series. It can be used to predict CTC prices, the spread of diseases as well as sales of companies. The main advantage of SARIMA over ARIMA is that it can be used to process seasonal time series to make long term predictions more accurate.

In general, the ARIMA Seasonal model is notated as follows:

SARIMA(p,d,q)(P, D, Q)s

$$\text{ARIMA} \quad \underbrace{(p,d,q)}_{\text{Non-seasonal part of the model}} \quad \underbrace{(P,D,Q)_m}_{\text{Seasonal part of the model}}$$

## E. *LSTM*

Long Short-Term Memory (LSTM) is an improvement from the RNNs, which able to solve the Gradient Problem. The LSTM models essentially extend the RNN's memory to enable them to keep and learn long-term dependencies of inputs. The figure below will show the LSTM Architecture. [16]

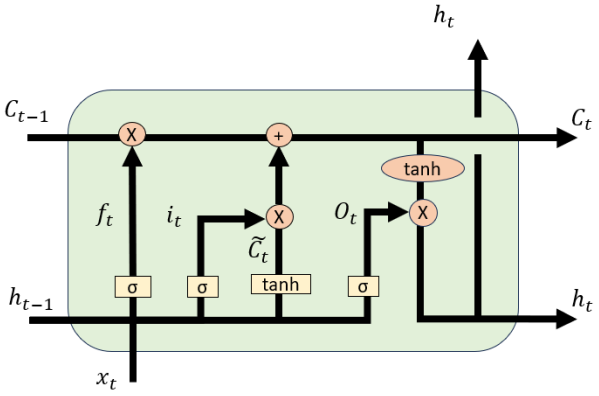


*Figure 1: Model of LSTM architectural*

- **Forget gate:**

$$f_t = \sigma(W_f . [h_{t-1}, x_t] + b_f)$$

- **Input gate:**

$$i_t = \sigma(W_i . [h_{t-1}, x_t] + b_i)$$

- **Temporary cell state:**

$$\widetilde{C}_t = tanh(W_c . [h_{t-1}, x_t] + b_c)$$

- **Current cell state:**

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C_t}$$

- **Output gate:**

$$o_t = \sigma(W_i \,.\, [h_{t-1}, x_t] + b_o)$$

## IV. EVALUATION METHODOLOGY

In this research, predictive models are evaluated according to two criteria: MAPEand RMSE

- Mean Absolute Percentage Error – MAPE

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

- Root Mean Squared Error – RMSE

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

## V. ANALYSIS

### A. Visualization

In this study, we use data price of Bitcoin during 2 years from 06/01/2021 to 06/01/2023

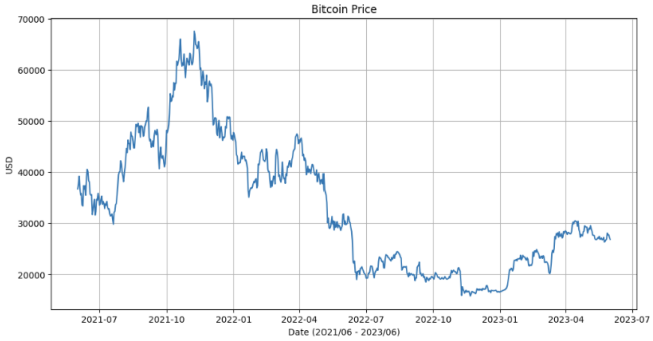


*Figure 2 : Data depicting Bitcoin price overview*

### B. Splitting data

At the same time, we split the data sets into 70% training data - 20% testing data - 10% validate data, 80% training data - 10% testing data - 10% validate data, 60% training data - 20% testing data - 20% validate data

## VI. RESULT

### A. (7-2-1 split) RMSE and MAPE based models comparisons.

| Model | Test RMSE | Valid RMSE | Test MAPE | Valid MAPE |
|---|---|---|---|---|
| LR | 5451.54 | 16286.77 | 22.35% | 57.83% |
| ARIMA | 3184.44 | 8811.25 | 14.22% | 31.06% |
| SARIMA | 3581.26 | 8745.58 | 15.47% | 30.84% |
| LSTM | 1813.04 | 1407.65 | 8.22% | 4.36% |

### B. (6-2-2 split) RMSE and MAPE based models comparisons.

| Model | Test RMSE | Valid RMSE | Test MAPE | Valid MAPE |
|---|---|---|---|---|
| LR | 8054.65 | 7295.32 | 42.91% | 22.65% |
| ARIMA | 5887.13 | 3411.44 | 30.81% | 11.93% |
| SARIMA | 6950.68 | 2925.45 | 36.29% | 10.55% |
| LSTM | 1395.54 | 1719.62 | 6.40% | 4.93% |

### C. (8-1-1 split) RMSE and MAPE based models comparisons.

| Model | Test RMSE | Valid RMSE | Test MAPE | Valid MAPE |
|---|---|---|---|---|
| LR | 8363.76 | 17938.78 | 33.35% | 63.73% |
| ARIMA | 6048.22 | 11298.76 | 24.29% | 39.98% |
| SARIMA | 6391.66 | 11604.60 | 25.59% | 41.10% |
| LSTM | 1777.51 | 1047.41 | 5.28% | 3.21% |

With the Bitcoin Price dataset being splitted into 3 patterns of train-test-val (**6-2-2, 7-2-1 and 8-1-1)** we have the best predictive models respectively : LSTM models.

### D. Visualize predict price

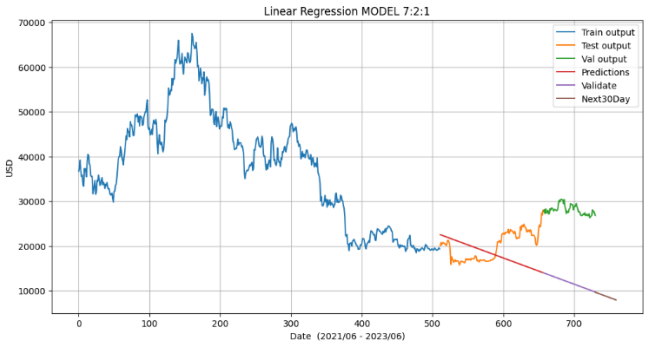


*Figure 3: Result of LR model* (7-2-1 split)

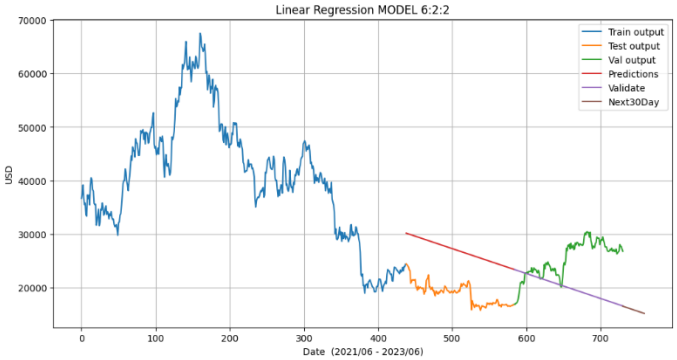


*Figure 4: Result of LR model* (6-2-2 split)

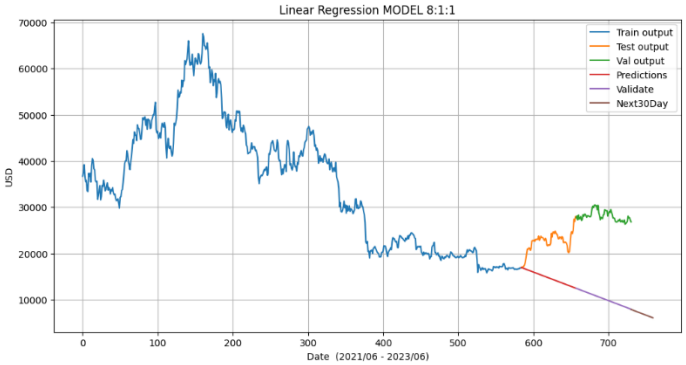


*Figure 5: Result of LR model* (8-1-1 split)

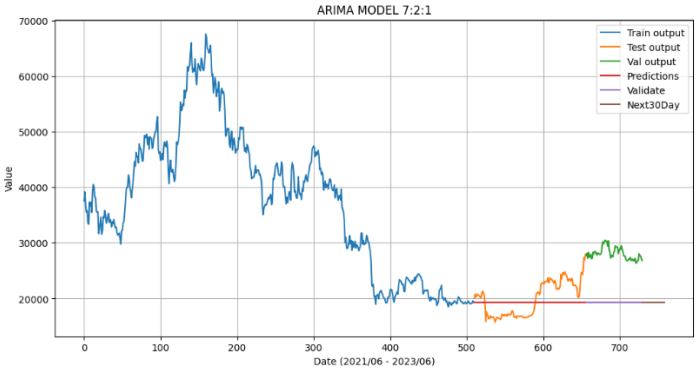


*Figure 6 : Result of ARIMA model* (7-2-1 split)


*Figure 7 : Result of ARIMA model* (6-2-2 split)

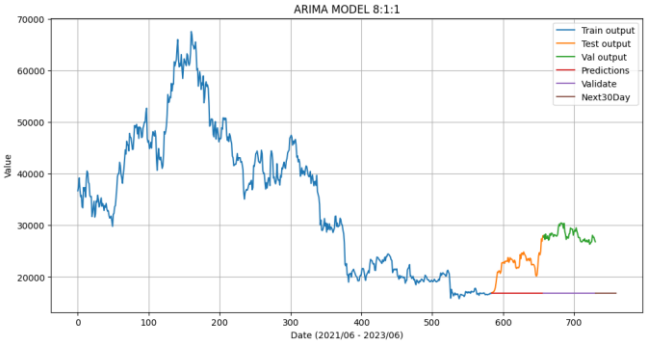


*Figure 8 : Result of ARIMA model* (8-1-1 split)

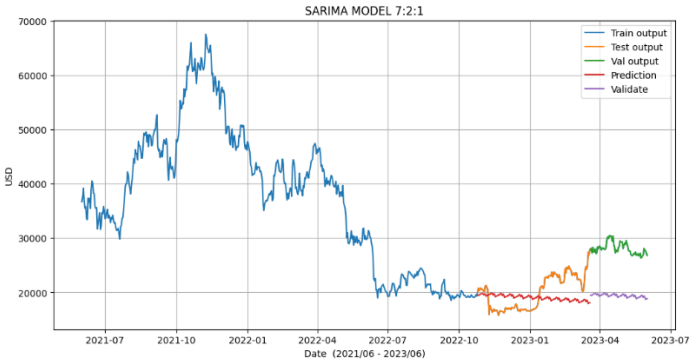


*Figure 9 : Result of SARIMA model* (7-2-1 split)

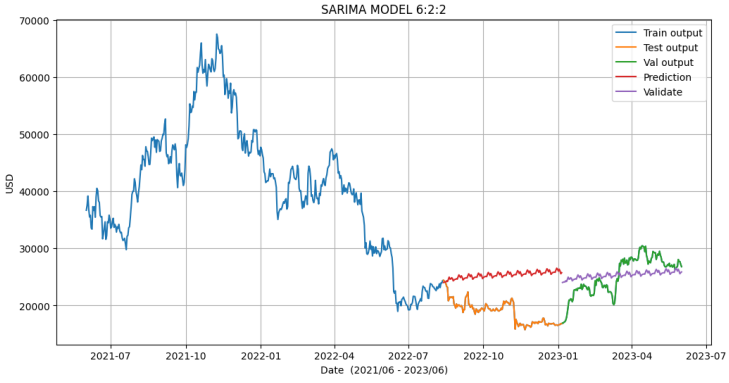


*Figure 10 : Result of SARIMA model* (6-2-2 split)

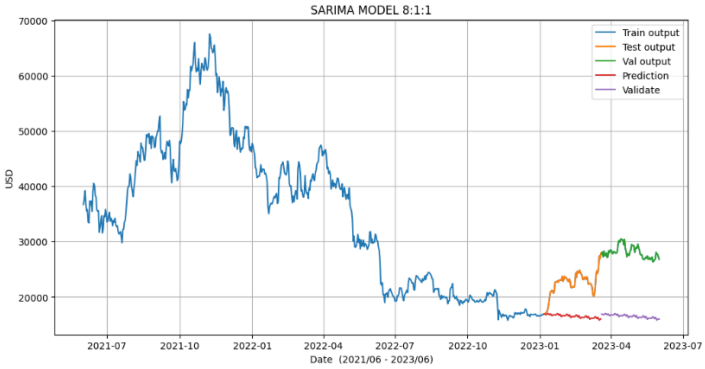


*Figure 11 : Result of SARIMA model* (8-1-1 split)

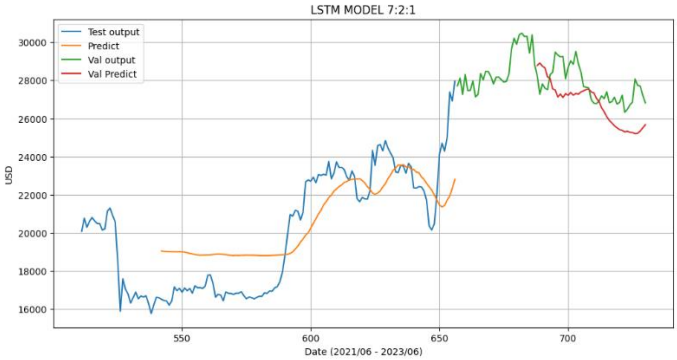


*Figure12 : Result of LSTM model* (7-2-1 split)

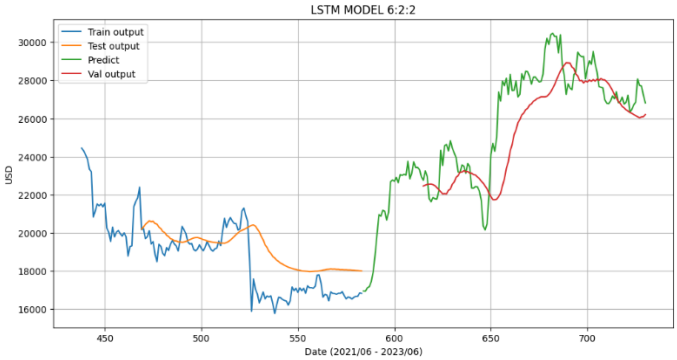


*Figure 13 : Result of LSTM model* (6-2-2 split)

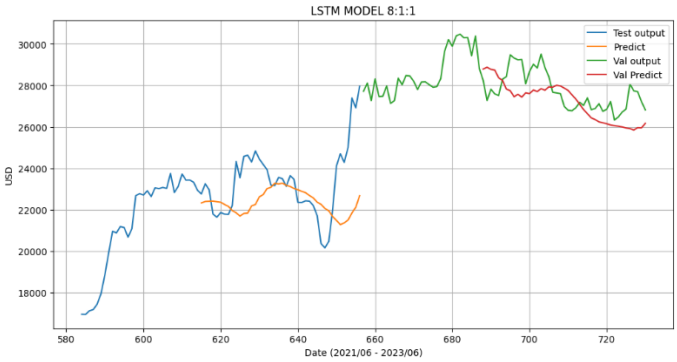


*Figure 14 : Result of LSTM model* (6-2-2 split)

## VII. CONCLUSION

As the most popular cryptocurrency, bitcoin has drawn interest from economists, financiers, and even computer scientists. Its significant price volatility and changes make forecasting difficult but also appealing. In this research workwe divided the training, testing and validate data sets by 70% - 20% - 10%, 60% - 20%- 20% and 80% - 10% - 10% and statistics model and machine learning to predict the close price of Bitcoint. The ARIMA model learned more from training data, linear learning such as LR has high errors. Besides, ARIMA model has learns good, but ARIMA still regression like LR, so the the predicted values has show a straight line. From that, it is difficult for predicting. And the excellent learning and small error prediction is LSTM model (Long Short Term Memory). Finally, with this study, we hope help the investor in invest the cryptocurrency.

## REFERENCES

[1] What is Bitcoint?https://coin98.net/bitcoin-btc-la-gi

[2] P. J. Brockwell and R. A. Davis (2001), Introduction to Time Series and Foreasting, 2rd ed., New York: Springer Link, pp. 180196.

[3] G. Box andJenkin (1970), Time Series Analysis, Forecasting and Control, 4 ed., San Francisco: Holden-Day, 1970, pp. 234-239.

[4] D. N. Gujaati and D. C. Porter (2009), Basic Econometrics, 5 ed., vol. 5

[5] R. Hill, W. E. Griffiths and G. C. Lim (2011), Principles of Econometrics, 4 ed., New Jersey: John Wiley & Sons, Inc., pp. 512- 517.

[6] Sun, X., Liu, M., & Sima, Z. (2018). A novel cryptocurrency price trend forecasting model based on LightGBM. Finance Research Letters.

[8] Guo, T., Bifet, A., & Antulov-Fantulin, N. (2018). Bitcoin Volatility Forecasting with a Glimpse into Buy and Sell Orders. 2018 IEEE International Conference on Data Mining (ICDM)

[9] Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in regression analysis: the problem revisited. The Review of Economic and Statistics

[10] S. McNally, J. Roche, and S. Caton, “Predicting the price of Bitcoin using Machine Learning," in Parallel, Distributed and Network-based Processing (PDP), 2018 26th Euromicro International Conference on, 2018, pp. 339- 343. [11] R. Mittal, S. Arora, and M. P. S. Bhatia, "AUTOMATED CRYPTOCURRENCIES PRICES PREDICTION USING MACHINE LEARNING," 2018. [12] C.-H. Wu, C.-C. Lu, Y.-F. Ma, and R.-S. Lu, “A New Forecasting Framework for Bitcoin Price with LSTM," in 2018 IEEE International Conference on Data Mining Workshops (ICDMW), 2018, pp. 168-175.

[13] F. Qian and X. Chen, “Stock Prediction Based on LSTM under Different Stability," in 2019 IEEE 4th International Conference on Cloud Computing and

[14] S. Hochreiter and J. Schmidhuber, “Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735 1780, 1997.

[15] Yi Dnhui, Wang Yan. Applied time series analysis. 5h ed.. Beijing, China: China Renmin University Press; 2019.