Shape, rectangle

Description automatically generated

**FINAL PROJECT**

**BUSINESS ANALYTICS :**

**PREDICT BITCOIN PRICE**

Instructor : Assoc. Prof. Nguyễn Đình Thuân

Nguyễn Thị Viết Hương

Class ID : IS403.N23.HTCL

Students group: Group 07

|  |  |
| --- | --- |
| Full name | Student ID |
| Phạm Nguyễn Cao Triết | 19521050 |
| Võ Xuân Thịnh | 19522290 |
| Lê Thị Ngọc Huyền | 20521421 |
| Nguyễn Duy | 20521234 |

**June 20th 2023, Ho Chi Minh city**

**UNIVERSITY OF INFORMATION TECHNOLOGY**

**INFORMATION SYSTEM FACULTY**

**TEACHER'S COMMENTS**

………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………

**TABLE OF CONTENTS**

[I. INTRODUCTION 4](#_Toc138158705)

[1.1 What is Bitcoin ? 4](#_Toc138158706)

[1.2 Introduction of research topic 4](#_Toc138158707)

[1.3 Forecasting methods 5](#_Toc138158708)

[II. RELATED WORK 5](#_Toc138158709)

[III. DATA AND METHODOLOGY 7](#_Toc138158710)

[3.1 Dataset 7](#_Toc138158711)

[3.2 Linear Regression (LR) 8](#_Toc138158712)

[3.3 Non-Linear Regression (NLR) 9](#_Toc138158713)

[3.4 Autoregression Integrated Moving Average (ARIMA) 9](#_Toc138158714)

[3.5 Seasonal Autoregressive Integrated Moving Average (SARIMA) 11](#_Toc138158715)

[3.6 Long Short-Term Memory (LSTM) 12](#_Toc138158716)

[IV. RESULT 13](#_Toc138158717)

[4.1 Linear Regression (LR) 13](#_Toc138158718)

[4.2 Non-Linear Regression (NLR) 14](#_Toc138158719)

[4.3 Autoregression Integrated Moving Average (ARIMA) 14](#_Toc138158720)

[4.4 Seasonal Autoregressive Integrated Moving Average (SARIMA) 16](#_Toc138158721)

[4.5 Long Short-Term Memory (LSTM) 17](#_Toc138158722)

[V. DISCUSSION 19](#_Toc138158723)

[5.1 (7-2-1 split) RMSE and MAPE based models comparisons. 19](#_Toc138158724)

[5.2 (6-2-2 split) RMSE and MAPE based models comparisons. 19](#_Toc138158725)

[5.3 (8-1-1 split) RMSE and MAPE based models comparisons. 19](#_Toc138158726)

[VI. CONCLUSION 20](#_Toc138158727)

[VII. GROUP WORK DISCUSSION 20](#_Toc138158728)

[VIII. REFERENCE 21](#_Toc138158729)

# INTRODUCTION

## 1.1 What is Bitcoin ?

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.

One of the key features of Bitcoin is its limited supply. There will only ever be 21 million bitcoins in existence, and this scarcity is designed to create value and prevent inflation. Bitcoins can be acquired through various means, such as purchasing them on cryptocurrency exchanges, receiving them as payment for goods or services, or mining, which involves using powerful computers to solve complex mathematical problems to validate transactions.

Bitcoin transactions are pseudonymous, meaning that while transaction details are recorded on the blockchain, the identities of the participants are not necessarily revealed. However, it's important to note that Bitcoin transactions are not completely anonymous, as it is possible to analyze the blockchain and potentially identify individuals based on transaction patterns.

Bitcoin has gained popularity and recognition as a form of digital currency, and it has been used for various purposes, including online purchases, investment, and remittances. Its decentralized nature and the potential for financial independence from traditional banking systems are some of the factors that have contributed to its appeal.

## 1.2 Introduction of research topic

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. Therefore, in this report, our team will employ a regression model combined with machine learning algorithms to forecast the price of Bitcoin.

## 1.3 Forecasting methods

Our team utilized multiple models to forecast the end-of-day Bitcoin prices, including Autoregression Integrated Moving Average (ARIMA), Long Short-Term Memory (LSTM), Linear Regression (LR), Non-Linear Regression (NLR), and Seasonal Autoregressive Integrated Moving Average (SARIMA). Additionally, we combined LSTM with ARIMA to enhance the accuracy of the predictions. Finally, we compared the accuracy of each model to determine the most suitable model for the task.

To evaluate the performance of each model, the team assessed the accuracy of the predicted Bitcoin prices by comparing them with the actual prices at the end of the day. we considered metrics such as mean absolute error (MAE), root mean squared error (RMSE), or mean percentage error (MPE) to quantify the differences between the predicted and actual prices. The model that demonstrated the lowest error and highest accuracy in predicting Bitcoin prices was identified as the most appropriate model for the task.

# 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.

# DATA AND METHODOLOGY

## 3.1 Dataset

The data was collected from the Investing.com website. The Bitcoin price data was collected from June 1, 2021 to June 1, 2023 including the following attributes: Date, Price, Open, High, Low, Volume, Change.

Link dataset : https://www.investing.com/crypto/bitcoin/historical-data

|  |  |
| --- | --- |
| Attributes | Description |
| Date | The date of the transaction |
| Price | The closing price of the previous period |
| Open | The opening price of the day |
| High | The highest price reached during the day |
| Low | The lowest price reached during the day |
| Vol | The total trading volume for the day |
| Change | The percentage change or growth rate |

A screenshot of a computer

Description automatically generated with low confidence

*Figure 1 : Data sources*

A picture containing line, plot, screenshot

Description automatically generated

*Figure 2 : Data graph depicting Bitcoin price overview*

## 3.2 Linear Regression (LR)

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]

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.

To find the best-fit line for each independent variable, multiple linear regression calculates three things: [6]

* The regression coefficients that lead to the smallest overall model error.
* The t statistic of the overall model.
* The associated p value (how likely it is that the t statistic would have occurred by chance if the null hypothesis of no relationship between the independent and dependent variables was true).
* It then calculates the t statistic and p value for each regression coefficient in the model.

## 3.3 Non-Linear Regression (NLR)

Nonlinear regression is a mathematical model that fits an equation to certain data using a generated line. As is the case with a linear regression that uses a straight-line equation (such as Ӷ= c + m x), nonlinear regression shows association using a curve, making it nonlinear in the parameter. The nonlinear regression model when at least one parameter is a nonlinear function :

A black text on a white background

Description automatically generated with medium confidence

Formula: A simple nonlinear regression model is expressed as follows:

A picture containing font, white, calligraphy, typography

Description automatically generated

Where:

* x X is a vector of P predictors
* x **β** is a vector of k parameters x
* F (-) is the known regression function x
* is the error term

## 3.4 Autoregression Integrated Moving Average (ARIMA)

ARIMA model (Autoregressive Integrated Moving Average) performs well in time series forecasting and is most effective when the data exhibits time dependence. ARIMA model combines autoregressive (AR), moving average (MA), and differencing (I) components.

To decide whether to build an ARIMA model, the time series data needs to be stationary. Stationarity refers to a time series where the mean, variance, and covariance do not change over time, and there is no trend present. The Dickey-Fuller test and Augmented Dickey-Fuller test are commonly used to test for stationarity in the time series data. These tests help determine whether the time series data satisfies the stationarity assumption required for ARIMA modeling.

A picture containing text, font, line, screenshot

Description automatically generated

*Figure 3: Principles of Station*

The Dickey-Fuller test:

∆𝑌𝑡 = 𝛿𝑌𝑡−1 + 𝑈𝑡

∆𝑌𝑡 = 𝛽0 + 𝛿𝑌𝑡−1 + 𝑈𝑡

∆𝑌𝑡 = 𝛽0 + 𝛽1𝑡 + 𝛿𝑌𝑡−1 + 𝑈𝑡

Check out Augmented Dickey Fuller:

∆𝑌𝑡 = 𝛼 + 𝛽𝑡 + 𝛾𝑌𝑡−1 + 𝛿1𝑌𝑡−1 + ⋯ + + 𝛿𝑝−1𝑌𝑡−𝑝+1 + 𝜀𝑡

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

𝐴𝑅(𝑝) = 𝑌𝑡 = 𝜀𝑡 + 𝑐 + 𝛼1𝑌𝑡−1 + 𝛼2𝑌𝑡− 2 + ⋯ + 𝛼𝑝𝑌𝑡−𝑝

Difference I is the difference between present and past values.

First difference : ∆ Y = 𝑌𝑡 − 𝑌𝑡−1

Second difference: ∆ Y = (𝑌𝑡 − 𝑌𝑡−1) − (𝑌𝑡−1 − 𝑌𝑡−2) = 𝑌𝑡 − 2𝑌𝑡−1 + 𝑌𝑡−2

Difference of order d: ∆ 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 + 𝛽0 + 𝛽1𝑌𝑡−1 + 𝛽2𝑌𝑡− 2 + ⋯ + 𝛽𝑞𝑌𝑡−q

## 3.5 Seasonal Autoregressive Integrated Moving Average (SARIMA)

SARIMA stands for Seasonal-ARIMA and it includes seasonality contribution to the forecast. The importance of seasonality is quite evident and ARIMA fails to encapsulate that information implicitly. The Autoregressive (AR), Integrated (I), and Moving Average (MA) parts of the model remain as that of ARIMA. The addition of Seasonality adds robustness to the SARIMA model

A picture containing text, font, white

Description automatically generated

where m is the number of observations per year. We use the uppercase notation for the seasonal parts of the model, and lowercase notation for the non-seasonal parts of the model. Similar to ARIMA, the P, D, Q values for seasonal parts of the model can be deduced from the ACF and PACF plots of the data.

Together, the notation for an SARIMA model is specified as:

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

Where the specifically chosen hyperparameters for a model are specified, for example:

SARIMA(3,1,0)(1,1,0)12

Importantly, the m parameter influences the P, D, and Q parameters. For example, an m of 12 for monthly data suggests a yearly seasonal cycle. A P=1 would make use of the first seasonally offset observation in the model, e.g . t-(m\*1) or t-12. A P=2, would use the last two seasonally offset observations t-(m \* 1), t-(m \* 2). Similarly, a D of 1 would calculate a first order seasonal difference and a Q=1 would use a first order errors in the model (e.g. . moving average). A seasonal ARIMA model uses differencing at a lag equal to the number of seasons (s) to remove additive seasonal effects. As with lag 1 differencing to remove a trend, the lag s differencing introduces a moving average term. The seasonal ARIMA model includes autoregressive and moving average terms at lag s.

## 3.6 Long Short-Term Memory (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]

A diagram of a flowchart

Description automatically generated with low confidence

*Figure 4: Model of LSTM architectural*

**Forget gate:**

**Input gate:**

**Temporary cell state:**

**Current cell state:**

**Output gate:**

# RESULT

## 4.1 Linear Regression (LR)

- Result of the best model of LR for each train-test-val split A graph with blue and orange lines

Description automatically generated with low confidence

*Figure 5: Result of LR model* (7-2-1 split)

A picture containing text, plot, line, diagram

Description automatically generated

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

A picture containing text, plot, line, diagram

Description automatically generated

*Figure7: Result of LR model* (8-1-1 split)

## 4.3 Autoregression Integrated Moving Average (ARIMA)

- Result of the best model of ARIMA for each train-test-val split

A picture containing text, plot, line, diagram

Description automatically generated

*Figure 8 : Result of ARIMA model* (7-2-1 split)

A picture containing text, plot, line, diagram

Description automatically generated

*Figure 9 : Result of ARIMA model* (6-2-2 split)

A picture containing text, plot, line, diagram

Description automatically generated

*Figure 10 : Result of ARIMA model* (8-1-1 split)

## 4.3 Seasonal Autoregressive Integrated Moving Average (SARIMA)

- Result of the best model of SARIMA for each train-test-val split

A picture containing text, plot, line, diagram

Description automatically generated

*Figure 11 : Result of SARIMA model* (7-2-1 split)

A picture containing text, plot, line, diagram

Description automatically generated

*Figure 12 : Result of SARIMA model* (6-2-2 split)

A picture containing text, plot, line, diagram

Description automatically generated

*Figure 13 : Result of SARIMA model* (8-1-1 split)

## 4.4 Long Short-Term Memory (LSTM)

- Result of the best model of LSTM for each train-test-val split

A picture containing text, line, diagram, plot

Description automatically generated

*Figure14 : Result of LSTM model* (7-2-1 split)

A picture containing text, plot, line, diagram

Description automatically generated

*Figure 15 : Result of LSTM model* (6-2-2 split)

A picture containing text, line, plot, diagram

Description automatically generated

*Figure 16 : Result of LSTM model* (6-2-2 split)

# DISCUSSION

## 5.1 (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% |

## 5.2 (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% |

## 5.3 (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.

# 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.

# GROUP WORK DISCUSSION

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| WORK | C.TRIET | X.THINH | N.HUYEN | N.DUY |
| Problem Statement | X |  |  |  |
| Data collection | X |  |  |  |
| Visualizing data | X |  |  |  |
| Data Analysis | X |  |  |  |
| Data preprocessing | X |  |  |  |
| Models research | X |  |  |  |
| Linear Regression model | X |  |  |  |
| Arima Model | X |  |  |  |
| Sarima Model | X |  |  |  |
| LSTM Model | X |  |  |  |
| Compiling codes | X |  |  |  |
| Visualizing results | X |  |  |  |
| Analyzing results | X |  |  |  |
| Writing report | X |  |  |  |

# REFERENCE

[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, Canda: Mc GrawHill, pp. 777-784.

[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, 92-107.

[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.