UNIVERSITY OF INFORMATION TECHNOLOGY

FACULTY OF INFORMATION SYSTEMS

Predict Bitcoin price by Linear regression, Non linear regression, ARIMA, Neural networks auto regression, Boosting model and Sarimax

- PHẠM DUY HƯNG - 18520805
- BÙI QUANG GIÀU - 20521264
- VÕ ANH HÀO - 20521297

Predict Bitcoin price by Linear regression, Non-linear regression, ARIMA, Neural networks auto regression, Boosting model and Sarimax

1st Pham Duy Hung
STAT3013.N12.CTTT – EN
University of Information Technology
Ho Chi Minh city
18520805@gm.uit.edu.vn

2nd Bui Quang Giau
STAT3013.N12.CTTT – EN
University of Information Technology
Ho Chi Minh city
20521264@gm.uit.edu.vn

3rd Vo Anh Hao
STAT3013.N12.CTTT – EN
University of Information Technology
Ho Chi Minh city
20521297@gm.uit.edu.vn

Abstract

Keywords - Bitcoin price, prediction, Arima, linear regression, non-linear regression, boosting model, NNAR, Sarimax

I. INTRODUCTION

One of the most widely used cryptocurrencies is called bitcoin. In 2009, Satoshi Nakamoto released the initial version of Bitcoin, which continues to be the most valuable cryptocurrency today. Many of the altcoins that are currently available have their roots in bitcoin, which also serves as a significant turning point for online payment systems.

The current cryptocurrency has numerous components of money, including:

+ Used for payment and the exchange of goods and services and accepted by some people.

+ Some countries' currencies can be converted from cryptocurrencies.

+ The issuance of cryptocurrencies is also governed by laws to prevent the coin's value from declining due to an excessive number of coins being created.

You can think of Bitcoin as a digital money because BTC tokens do not exist in physical form. Transactions made using bitcoin are totally open and unrestricted. Money can be readily sent to anyone in the world. Instead of a central bank or government, this financial system is protected by thousands of computers located all over the world.

However, numerous national governments and significant corporations are starting to acknowledge Bitcoin. Although perhaps not many people have yet trusted this coin, it is currently the main driver behind transactions. Even a large number of exchanges already permit the public purchase and sale of bitcoin.

A significant area of financial research is empirical asset valuation. Due to their flexibility in selecting among a vast number of potential features and their capacity to investigate intricate, multidimensional correlations between features and target, models have found growing use in this sector. The new stream of crypto values has garnered little attention, despite a sizable body of research examining the valuation of stocks and bonds and producing a sizable number of indicators that may be able to anticipate the market. more vile In particular, a thorough analysis of the bitcoin market's short-term predictability is still lacking. Furthermore, the majority of studies simply examined technical features without considering the significance of the features included in the machine learning models. In light of this, we fill this research gap by comparing and contrasting various models to forecast the market movement of bitcoin, the most relevant cryptocurrency. With a market valuation of around $170 billion USD (as of September 2020), bitcoin controls about 58% of the cryptocurrency market. [1]

II. RELATED WORK

Financial market forecasting is a well-known and thoroughly researched area of financial study. About the predictability and efficiency of financial markets, there is conflicting evidence. Regression analysis on signals that have the potential to explain asset returns is a well-established method for examining profit-predicting signals. The mathematical method for making predictions provided by linear regression is rather straightforward and simple to understand. A well-established statistical method, linear regression is easily adaptable to software and computation. Businesses use it to accurately transform raw data and forecast values and trends for the near future using other relevant and existing data values.

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 [1]. 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 [7,8,9]

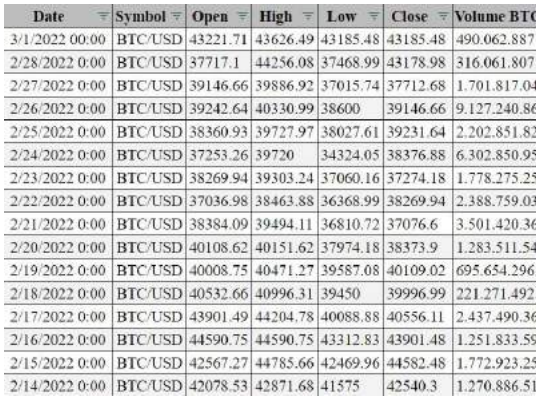
Papers by Sean et al. utilizing LSTM [10] They suggest a method for determining the price of Bitcoin that combines Recurrent Neural Network, Long Short Term Memory, and Ruchi.

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.

The LSTM forecasting algorithm is performed well.

III. DATA AND METHODOLOGY

A. Data sources



Date	Symbol	Open	High	Low	Close	Volume BTC
3/1/2022 00:00	BTC/USD	43221.71	43626.49	43185.48	43185.48	490.062.887
2/28/2022 0:00	BTC/USD	37717.1	44256.08	37468.99	43178.98	316.061.807
2/27/2022 0:00	BTC/USD	39146.66	39886.92	37015.74	37712.68	1.701.817.04
2/26/2022 0:00	BTC/USD	39242.64	40330.99	38600	39146.66	9.127.240.86
2/25/2022 0:00	BTC/USD	38360.93	39727.97	38027.61	39231.64	2.202.851.82
2/24/2022 0:00	BTC/USD	37253.26	39720	34324.05	38376.88	6.302.850.95
2/23/2022 0:00	BTC/USD	38269.94	39303.24	37060.16	37274.18	1.778.275.25
2/22/2022 0:00	BTC/USD	37036.98	38463.88	36368.99	38269.94	2.388.759.03
2/21/2022 0:00	BTC/USD	38384.09	39494.11	36810.72	37076.6	3.501.420.36
2/20/2022 0:00	BTC/USD	40108.62	40151.62	37974.18	38373.9	1.283.511.54
2/19/2022 0:00	BTC/USD	40008.75	40471.27	39587.08	40109.02	695.654.296
2/18/2022 0:00	BTC/USD	40532.66	40996.31	39450	39996.99	221.271.492
2/17/2022 0:00	BTC/USD	43901.49	44204.78	40088.88	40556.11	2.437.490.36
2/16/2022 0:00	BTC/USD	44590.75	44590.75	43312.83	43901.48	1.251.833.59
2/15/2022 0:00	BTC/USD	42567.27	44785.66	42469.96	44582.48	1.772.923.25
2/14/2022 0:00	BTC/USD	42078.53	42871.68	41575	42540.3	1.270.886.51

B. Multiple Linear Regression

- Regression analysis is a statistical technique to evaluate the relationship between variables that are related to each other, according to the study of the essay "A Study of Multiple Linear Regression Analysis" authored by Gulden Kaya Uyanik [10]. link between causes and effects. The fundamental goal of univariate regression is to create an equation for the linear relationship between the dependent and independent variables by analyzing the relationship between a dependent variable and an independent variable. Multilinear regression is the name given to regression models with one dependent variable and multiple independent variables.
- In multivariable regression analysis, an effort is made to take into account synchronous changes in the independent and dependent variables (Unver & Gamgam, 1999). The following was used to build the multivariate regression model:

$$y = \beta_0 + \beta_1 + \cdots + \beta_n x_n + \varepsilon$$

$$y = dependent\ variable$$

$$x_i = independent\ variable$$

$$\beta_i = parameter$$

$$\varepsilon = error$$

Examples of applications for multiple linear regressions include:

• Attempting to forecast a person's score given some knowledge attribute

• Makes an effort to forecast a student's overall test performance at a "A" level, based on results from a group of exams taken at age 18.

- Measures to evaluate a population's life expectancy with a variety of social and behavioral health characteristics (occupation, habits, eating habits, etc.).

This analysis, like basic linear regression and correlation, prevents us from drawing conclusions about the cause of an event, but it does allow us to look at the relationship between a group of explanatory variables and a certain dependent variable.

In terms of hypothesis testing, the null hypothesis is H0, which is the coefficient of 0 between the explanatory variable (x) and the dependent variable (y) in the case of simple linear regression. That is to say, no The dependent variable and the explanatory variable have a connection. The alternative hypothesis H1 states that there is a coefficient between x and y that is not zero. In other words, x and y do in fact have a relationship.

The null and empty hypotheses will be written as follows: H0: b1 = 0 H1: b10

C. Non-Linear Regression

Regression analysis that uses a non-linear relationship between a dependent variable and the independent variables is referred to as nonlinear regression. a complex nonlinear model that yields reliable results. Based on the specified data set, it analytically creates a curve that illustrates the relationship between variables. A best-fit curve can be produced by models with lots of flexibility. The mathematical function reflecting the variables (dependent and independent) in the non-linear relationship is the created and optimized curve in non-linear regression, which maps experimental data to a model. The nonlinear model's representational formula, which is acknowledged as a flexible type of regression analysis, is listed below.

$Y = f(X,\beta) + \epsilon$.

f is the regression function

ϵ is the error term

X are the vector parameters.

D. ARIMA

The ARIMA model is one of the top predictive models, according to Bishal DeyBidesh... (2022) [11].

The goal of this model is to enable the connection between time series data's past value and its present and potential future values. All participants in the foreign exchange market—analysts, exporters, importers, multinational corporations, speculators, and traders—believe that historical patterns can predict short-term market movements. future.

Model explanation: The differential autoregressive moving average model, also known as the ARIMA model [9], is a widely used model.

The prediction model was fit with fixed time series. It essentially combines the ARIMA model and the difference operation. Only linear relationships, not nonlinear ones, can be represented by this paradigm. The difference transforms the non-stationary time series into a fixed time series, after which the dependent variable only regresses on its lag value, current value, and delay of the random error term. The time series' historical and present values can be used by the model to forecast future values.

The structure of the ARIMA model (p, q) is as follows:

$$y(t+1) = a_0 + a_1 y(t) + \cdots + a_p y(t-p+1) + e(t+1) + b_1 e(t) + \cdots + b_q e(t-q+1)$$

E. Boosting model method

Our experiments show that boosting full decision trees usually yields better models than boosting weaker stumps. Unfortunately, our results also show that boosting to directly optimize log-loss, or applying Logistic Correction to models boosted with exponential loss, is only effective when boosting weak models such as stumps. Neither of these methods is hi is trained on the weighted train set. The error of hi determines the model weight αi and the future weight of each training example. There are two equivalent formulations. The first formulation, also used by Friedman, Hastie, and Tibshirani (2000) assumes yi ∈ {−1, 1} and hi ∈ {−1, 1}. The output of the boosted model is:

$$F(x) = \sum_{i=1}^{T} \alpha_i h_i(x)$$

Friedman et al. show that AdaBoost builds an additive logistic regression model for minimizing E(exp(−yF(x))). They show that E(exp(−yF(x))) is minimized by:

$$F(x) = \frac{1}{2} log \frac{P(y=1|x)}{P(y=-1|x)}$$

This suggests applying a logistic correction in order to get back the conditional probability:

$$P(y=1|x) = \frac{1}{1+exp(-2F(x))}$$

As we will see in the Empirical Results section, this logistic correction works well when boosting simple base learners such as decision stumps. However, if the base learners are powerful enough that the training data becomes fully separable, after correction the predictions will become only 0's and 1's (Rosset, Zhu, & Hastie 2004) and thus have poor calibration.

F. SARIMAX method

Two different types of orders must be provided in the SARIMAX models parameter. We refer to this order as a seasonal order in which we are required to submit four integers. The first one is comparable to the ARIMAX model (p, d, and q), and the other is to indicate the influence of seasonality. [5]
This is how the model can be represented mathematically.

$$\phi_p(L)\tilde{\phi}_P(L^s)\Delta^d\Delta_s^D y_t = A(t) + \theta_q(L)\tilde{\theta}_Q(L^s)\epsilon_t$$

Where :

- $\phi_p(L)$ is the non-seasonal autoregressive lag polynomial
- $\tilde{\phi}_P(L^s)$ is the seasonal autoregressive lag polynomial
- $\Delta^d\Delta_s^D y_t$ is the time series, differenced d times, and seasonally differenced D times.
- $A(t)$ is the trend polynomial (including the intercept)
- $\theta_q(L)$ is the non-seasonal moving average lag polynomial
- $\tilde{\theta}_Q(L^s)$ is the seasonal moving average lag polynomial

G. LSTM

Another kind of module offered for RNNs is LSTM (Long Short Term Memory). Hochreiter developed LSTM.

& Schmidhuber (1997)[14], which many scholars later expanded upon and made popular. The LSTM network (LSTM network) is made up of modules with recurrent consistency, just as the RNN.

The link between the hidden layers of RNN is different in LSTM, which is an upgraded form of RNN. In Figure 1, the RNN's explanation structure is displayed. The only structural difference between RNN and LSTM is the memory cell of the hidden layer. Additionally, the gradient issues are successfully resolved by the design of three specific gates. Figure 2 depicts the LSTM memory structure of the hidden layer. [12]

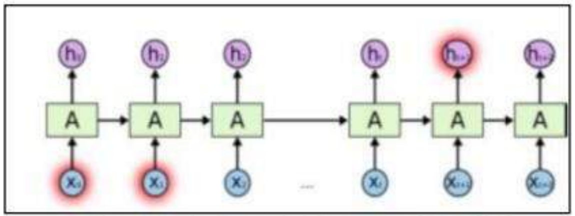


Figure 1. The Expanded Structure of RNN[21]

The RNN has flaws, which are illustrated in Figure 1. These flaws can be seen in the input X_o, X_1 which has a very wide range of information X_t,, X_{t+1}, so that when 1+1 requires information those which are relevant X_o, X_1 to RNN are unable to learn to link information because of old memory saved which will become increasingly useless over time because it is overwritten or replaced with new memory, a problem which was identified by Bengio,

Because LSTM may use memory cells and gate units to handle the memory at each input, it lacks the drawback of the RNN.

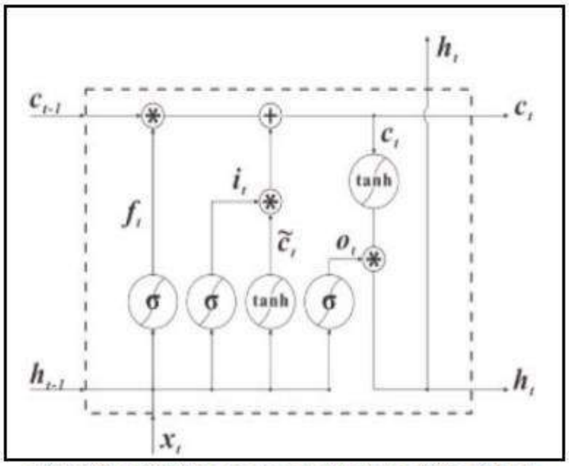


Figure 2. LSTM memory cell structure of the hidden

- i_i= the input gate, it's means information will be updated in the cell
- f_t= the forget gate, it's means information should be dropped from the cell.
- O_t= the output gate, it means how much information is output

- c_t= the candidate value for the states of the memory cell at time t.
- c_t = the state of the current memory cell at time t, which calculated by the combination of i_1 and c_t f_t and c_{t-1} through element-wise multiapplication
- h_t = it's mean output value filtered by output gate
-
- σ = is denoted denotes sigmoid function with the range 0 to 1, the function is used to put the value between -1 and 1.

IV. MODEL SETTING

1. *Original value*

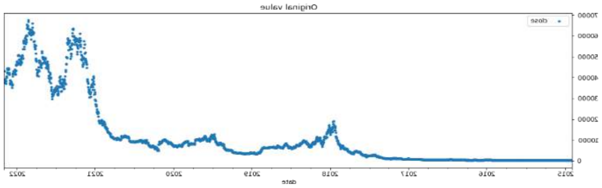


Figure 3. Price chart of BTC (2021-2022)

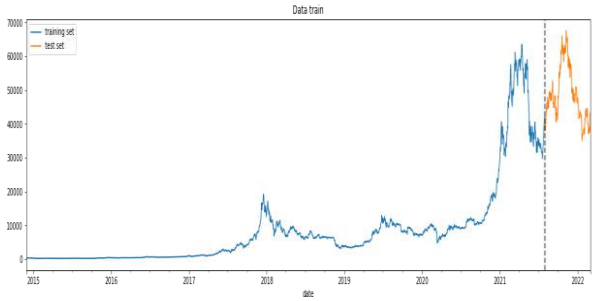


Figure 4. Train split value chart of BTC (2021-2022)

B. Linear Regression

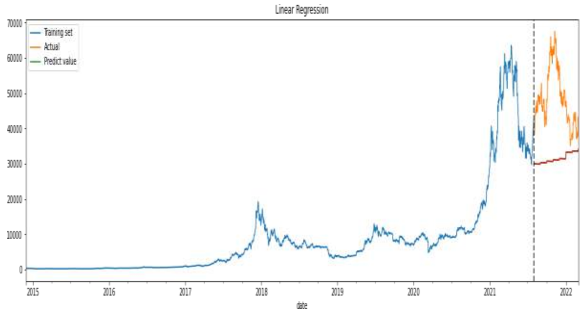
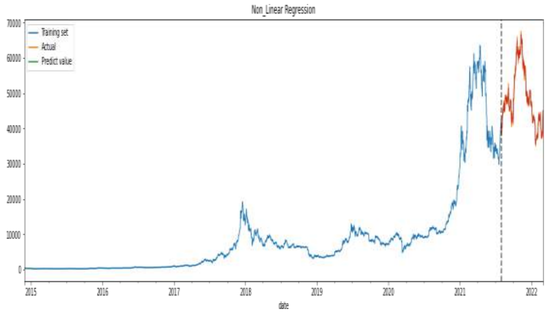


Figure 5. Compare the actual and predicted values in Linear Regression

C. Non Linear Regression



E. Boosting Model

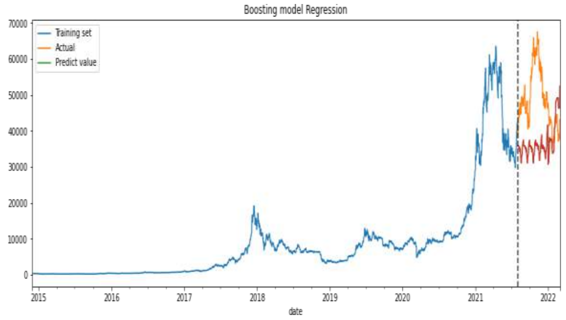


Figure 8. Compare the actual value and predicted value in BoostingModel

Figure 6. Compare the actual and predicted values in Non Linear Regression

D. ARIMA

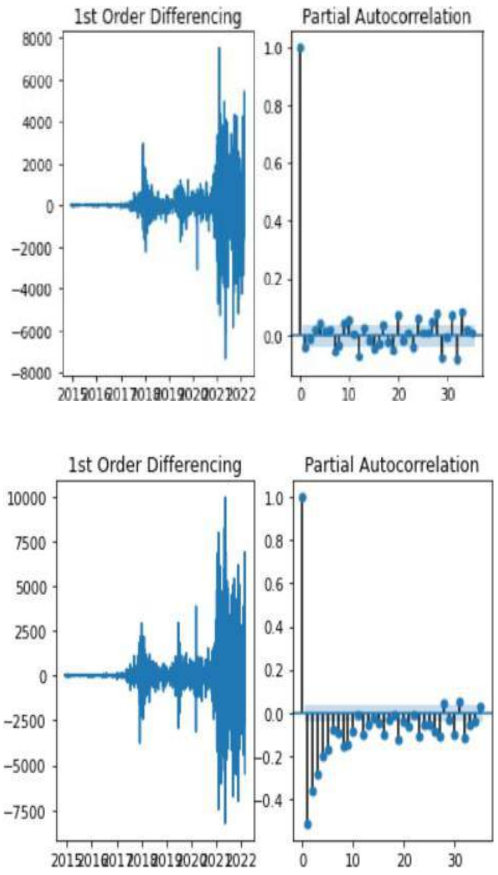


Figure 7. The actual and predicted values in ARIMA

F. SARIMAX

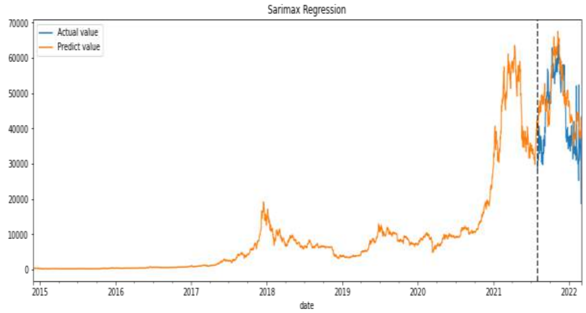


Figure 9. Compare the actual value and predicted value in SARIMAX

G. LSTM

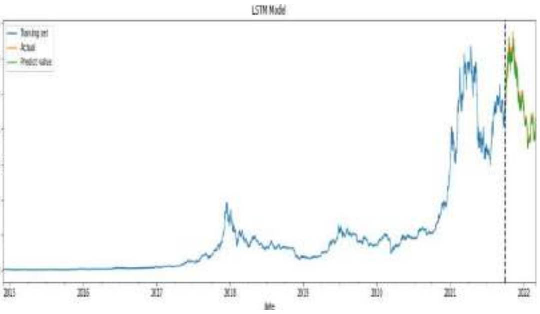


Figure 10. The actual value and predicted value in LSTM

REFERENCES

[1] R. B. MILLER and J. C. HICKMAN, "Time series analysis and forecasting", Transaction of society of actuaries, vol. 25, no.1,pp. 267-329, 1973.

[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.

[7] Yin, Haohua Sun, and Ravi Vatrapu. "A First Estimation of the Proportion of Cybercriminal Entities in the Bitcoin Ecosystem Using Supervised Machine Learning." 2017 IEEE International Conference on Big Data (Big Data), 2017.

[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.