Prediction of Ethereum Prices Using Linear Regression and Long Short-term Memory

Wenni Zhang *

Department of Economic, University of York, York, UK

*Corresponding author: 101119@yzpc.edu.cn

Abstract. Contemporarily, the popularity and use of cryptocurrencies has risen along with their prices and Ethereum is the second most popular and largest cryptocurrency after Bitcoin. Cryptocurrencies are based on the blockchain, which is a decentralized technology that has the power to change any banking system. They have become an attractive investment for both traders and individuals looking to invest. The price of Ethereum fluctuates and is affected by various factors, e.g., the crypto trading exchange as well as supply and demand. Ethereum is so valuable because it can be used as cash and one also pay Ethereum in full or in part to someone in exchange. Besides, it is easily guaranteed by the blockchain. Unlike stocks, the price of Ethereum is much more variable because it is traded 24 hours a day and there are no closing times. On this basis, this paper compares the results of two different models, namely linear regression and Long Short-Term Memory networks (LSTM). The dataset comprised in the closing prices of the last 372 days for Ethereum. The performance of the obtained models is critically evaluated using statistical indicators Root Mean Squared Error (RMSE) and the study have drawn our conclusions based on the RMSE result. The paper demonstrates a technique for using time series data in both models and determining each model's RMSE. These results shed light on guiding further exploration of prediction Ethereum prices and trends.

Keywords: Ethereum; LSTM; Linear Regression; Cryptocurrency.

1. Introduction

Cryptocurrency is a comprehensive product of modern cryptography and computer science that emerged from the exploration of decentralised transaction mechanisms. As a brand new financial asset, cryptocurrency has been in high demand in the market over the past two years. In 2009, Satoshi Nakamoto released the first cryptocurrency based on blockchain technology, Bitcoin. Some users exchanged 10,000 Bitcoins for a pizza to make the first transaction. With the development of cryptocurrencies, hundreds of well-known institutions have started accepting cryptocurrencies including Bitcoin and Ethereum [2]. Under the impact of the global COVID-19 in 2020, the price of cryptocurrency has skyrocketed. The price of Bitcoin has exceeded $50,000 and the price of Ethereum is over $3,000. Many scholars have studied the factors that influence the price of Bitcoin [3].

Bitcoin and Ethereum are the two most popular cryptocurrencies. In 2013, the concept of Ethereum (Ethereum) was proposed by Vitalik Buterin,which is a public chain platform based on blockchain technology and has smart contracts. Chain technology allows for a wider range of trials and is not controlled by a central authority. Compared to traditional currencies, Ethereum has more advantages, such as security, investment, anonymity and convenience. Meanwhile, the annual growth rate of cryptocurrency prices can be up to 400%, which is a high-yield investment [4, 5]. Cryptocurrency prices can be considered a time series forecasting problem, but the price of it is subject to cyclical ups and downs. Therefore, the cryptocurrency trading needs a standardised approach to accurately predict the trends of price fluctuations. Econometric models are the most commonly used tools in the study of traditional financial markets in the Bitcoin market.Models commonly used in field research include vector autoregression (VAR), ordinary least squares (OLS) and quantile regression (QR).

Throughout the previous few years, machine learning techniques have been implemented to forecast asset and cryptocurrency values and returns.Machine learning techniques have been effectively used to anticipate the stock market by introducing non-linear features into forecasting models to handle non-stationary financial time series, and results demonstrate that the strategy is more

successful for predicting. These machine learning methods include more well-known models including SVM, RF, as well as artificial neural networks (ANN), which have gained popularity recently. More complex deep recurrent neural networks RNNs and long learning methods such as Long Short-Term Memory networks (LSTMs) can achieve higher predictive accuracy [6].

Using long short-term memories (LSTM) and Multi-layer perceptron (MLP), Deepak Kumar conducted research on the price movements of Ethereum.They concluded that LSTM was slightly, but not very significantly, superior to MLP [7]. Monish and Mridul compared three different models for predicting the Ethereum price, namely recurrent neural networks (RNNs), Long Short-Term Memory (LSTMs) and bidirectional Long Short-Term Memory (Bi-LSTMs). They used only the closing price as the parameter and conclude that the bidirectional LSTM is the best model [8]. Rakshit mention the implementation of linear regression and Long Short-Term Memory (LSTM), both of which make predictions using the Indian stock market index Nifty 50. They come to the conclusion that the LSTM model can produce more precise findings and is best suited for time series data [9]. Singh et al. presented three types of machine learning (ML) for predicting stock prices, including linear regression, LSTM and decision tree. The conclusion was that the best approach with the highest accuracy is LSTM [10]. Ebenesh and Anitha investigated two classification algorithms, linear regression (LR) and long short-term memory, to predict three equities (AAPL, MSFT, and AMZN) (LSTM). Compared to the LSTM model, the LR model performs better than the LSTM model in evaluating the predictive parameters of the stock index [11]. Researches also predict cryptocurrency prices through sentiment analysis using linear regression (LR) and Long Short-Term Memory (LSTM). According to the results of this study, which only compares the Open, Close, High and Low characteristics, Long Short-Term Memory has a higher accuracy than linear regression [12].

Exploring the price of Ethereum can provide people with a reasonable and objective understanding of it and help them assess the price of it more reasonably before investing. It serves as a reference for the development of cryptocurrencies and has practical significance for the research and promotion of national cryptocurrencies. The rest part of the paper is organzied as follows. Section 2 provides the data and method of this study. Section 3 presents the results of forecasting models and discussions. Section 4 reports some limitations and prospects. Section 5 provides some concluding remarks.

2. Data & Method

2.1 Data Collection & Cleaning

The factors include the price of WTI crude oil, the price of gold, the price of the bitcoin, T-bond, the S&P 500 Index (S&P500) and the Nasdaq Composite (NASDAQ). At the same time, the study also collected information on several key global exchange rates, including the euro/dollar (EUR/USD) and the UK GBP/USD. The information on global economic indicators and global exchange rates was obtained from investing.com as shown in Fig. 1. There are cases where some values are missing or the data are not meaningful. Therefore, missing data that can be replaced are completed by interpolation; otherwise, the data corresponding to the missing values are removed from the dataset.

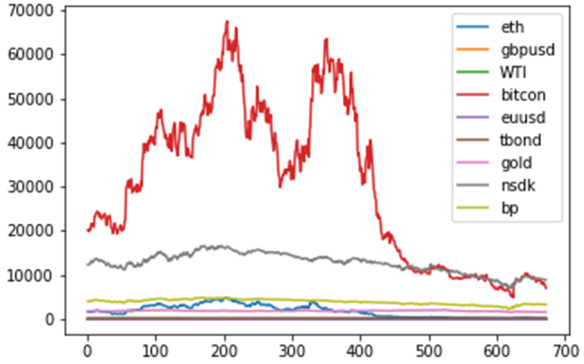

Fig. 1 Display of the collected data.

Next, this study makes a graph of the correlation matrix for each variable. If the correlation matrix is shown in yellow, it means that both factors have a positively correlation; if it is shown in blue, it means that negative correlation exists between the variables. For the shades of the two colors, the darker the color, the greater the correlation between the variables. The diagram of the correlation matrix for each variable studied in this experiment is illustrated in Fig. 2. Here, 0 is Ethereum, 1 is GBP/USD, 2 is WTI oil, 3 is bitcoin price, 4 is EUR/USD, 5 is T-bond, 6 is the price of gold, 7 is NASDAQ and 8 is S&P500.

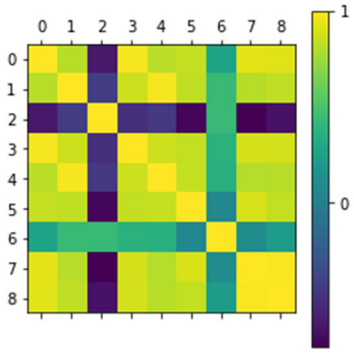

Fig. 2 Correlation matrix for each variable.

2.2 Normalisation of Data

When utilizing the sigmoid (default) or tanh activation functions, specifically, LSTMs are sensitive to the scaling of the input data.The data must be rescaled to a range of 0 to 1 (also known as normalisation). In this experiment, the MinMaxScaler preprocessing class from the scikit-learn library is used to normalise the data set. The normaliztion equation is given as follows:

$$x_{\text{normalization}} = \frac{x - \text{Min}}{\text{Max} - \text{Min}} \tag{1}$$

The study separates the processed data set into a test set and a training set, using the data from January 1, 2021, to August 1, 2022 as the test set and the remaining data from January 1, 2020, to February 29, 2022 as the training set.

2.3 Models & Metrics

2.3.1 Models

Linear regression is a method of statistical analysis which uses the least squares function of the linear regression equation to analyze the quantitative relationship between at least two variables. It is used extensively. Its expression has the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{2}$$

where e is a normal distribution with a mean of 0. There is only one independent variable and one dependent variable in a regression analysis, and their connection can roughly be approximated by a straight line. A linear regression analysis with one variable is the formal name for this particular regression investigation. Multiple linear regression analysis is used when the dependent variable and the independent variables are linearly related, and the regression study includes two or more independent variables.

The 1980s and 1990s focused on recurrent neural networks, while deep learning algorithms first appeared in the early 2000s. Recurrent neural networks are a type of neural network where all nodes are connected in a chain and sequence data is input recursively in the direction of the sequence evolution.The Long Short-Term Memory networks (LSTM) and the Bidirectional Recurrent Neural Networks (Bi-RNN) are typical recurrent neural networks among them. Recurrent neural networks have the following advantages when learning nonlinear properties of sequences: memory, parameter sharing, and Turing completeness. RNN are utilized in numerous time series forecasting tasks as well as natural language processing (NLP), including speech recognition, language modeling, machine translation, etc. A time-cyclic neural network called Long Short-Term Memory (LSTM) was created

expressly to address the issue of general RNNs' long-term dependence (recurrent neural networks). A concatenated form of recurring neural network modules exists in all RNNs. The recurrent structural module only has a very basic structure in conventional RNNs, like a tanh layer.

Figure 3 exhibits the structure of the LSTM network. Three gates are added to the hidden layer: forget gate, input gate and output gate The core of the LSTM is the horizontal line through the recurrent unit, this part controls the updating of the unit's state. There is a simple linear interaction as shown in Eq. (1).

$$C_n = f_n * C_{n-1} + i_n * \tilde{C}_n \tag{3}$$

A forget gate is defined to selectively delete some information. h -1 is the output of the previous layer. The input of the previous cell, they all give a value between 0 and 1 by the sigmoid function:

$$f_n = \text{sigmoid}\left(W_f[h_{n-1}, x_n] + b_f\right) \tag{4}$$

The sigmoid function is expressed as

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{5}$$

The input gate defines the new information to be added to the recurrent unit selected by following

$$i_n = \text{sigmoid}(W_n[h_{n-1}, x_n] + b_n) \tag{6}$$

$$\tilde{C}_n = \tanh(W_C[h_{t-1}, x_t] + b_C) \tag{7}$$

Optionally, one adds new information to the loop unit. is the output of the sigmoid function, the tanh function is the output vector of . The output gate defines the information to be output. The tanh function processes the state of the cell and outputs a value between -1 and 1. Multiplying this value by the output of the sigmoid function, one obtains the output of the cell.

$$O_n = \text{sigmoid}(W_o[h_{n-1}, x_n] + b_o) \tag{8}$$

$$\tilde{C}_n = \tanh(W_C[h_{t-1}, x_t] + b_C) \tag{9}$$

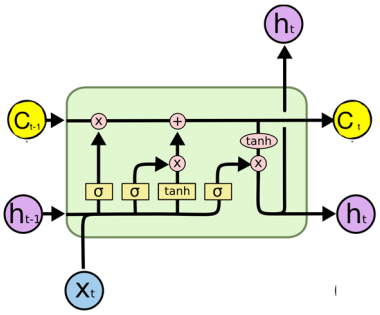

Fig. 3 LSTM Architecture.

2.3.2 Metrics

In general regression analysis, the degree of discrepancy between the true value and the predicted value is typically used as the benchmark for evaluating the quality of the findings. To test the predictive power of the two forecasting models for the price of Ethereum, choose the Root Mean Square Error (RMSE) as the evaluation standard to measure the performance of the different models.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{10}$$

3. Results & Discussion

3.1 Linear Regression

The study define the module method LinearRegression(), use fit to train the training data. A model is already trained and used to immediately predict the data from the test set when all training steps have been finished as shown in Fig. 4. In order to determine the evaluation indicators for the previously established model, the study uses data from the test dataset, and output the evaluation indicators RMSE is 0.09

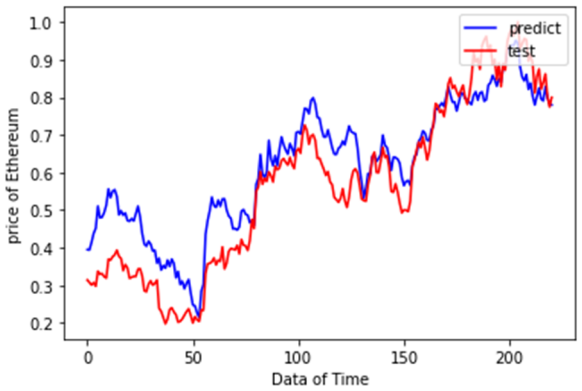

Fig 4. Final Resultant graph of LR and the Root Mean Absolute Error rate (0.09%)

3.2 LSTM

The study conducts pre-processing of inventory data. The data must be modified prior to employing an LSTM model. It entails changing the data set into a supervised learning problem and normalizing the variables (both the input and output values) in order to provide the data at the previous time (t-1) and forecast the data at the present time (t). This paper defines a function called series_to_supervised() that generates multivariate time series as a dataset for supervised learning. A Keras LSTM layer is fed a 3-dimensional (L, S, F) array of numbers, where L is the length of sequences, S is the size of time step and F is the number of sequence features.

The experiment builds the model quickly using the Keras Deep Learning Framework. A sequential model is created, the first layer of LSTM with 11 units of nodes and return sequence as True is added, dropout is set to 0.5, a dense layer is added to increase the dimension to 1 (regression problem), relu is used as the activation function (Although sigmoid is also utilized, its experimental performance falls short of that of relu), and Mean Absolute Error is selected as the loss function (MAE). Each batch size is 8, the model utilizes 50 epochs, and Adam is the optimization algorithm, and the attribute of a time step(t-1) is the input variable. Set the parameter validation_data in the function fit() to record the loss on the training set and the test set . The LSTM model in this experiment mainly uses the RMSE as the evaluation measure. Mainly compare the error loss of the training set train_loss and the error loss of the prediction set pre_loss as presented in Fig 5. Obviously, the lower the loss value, the better the fitting performance of the model. Depicted in Fig 6, the blue line in the figure is the prediction of the model fed with the training dataset and the red line is the true value.The result shows that most of the blue and red lines largely overlap, which shows that the LSTM model is accurate enough, and the RMSE is 0.044.

The study conducts the comparison of prediction results and real values. The difference between the prediction and the true value, allows us to visually see how well the model fits. According to the results listed in Table. 1, the LSTM model has the better predictive performance based on indicator RMSE, followed by the LR model.

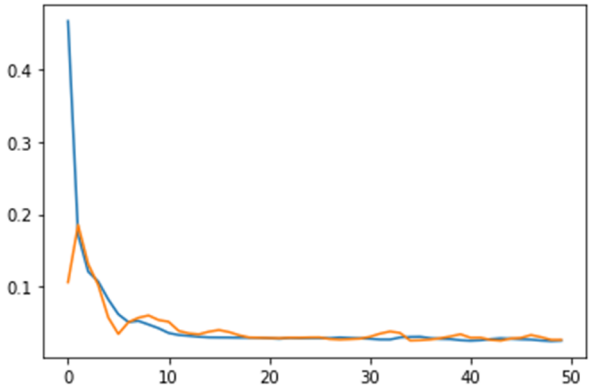

Fig 5. the loss on the training set and the test set.

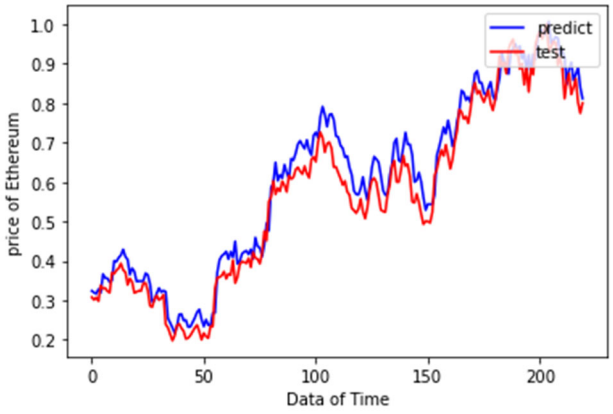

Fig 6. Final Resultant graph of LSTM and the Root Mean Absolute Error rate (0.04%)

Table 1. RMAE of LR and LSTM

Parameters	LR	LSTM
RMAE	0.091	0.044

4. Limitations & Prospects

This paper mainly presents innovative machine learning methods for predicting the Ethereum price. From an economic point of view, the research on the factors influencing the Ethereum price is not yet comprehensive and is limited to the analysis of the main 8 influencing factors. In fact, the Ethereum price level is influenced by many factors, e.g., political factors, do not find relevant variables for quantitative analysis. Different factor variables may have different effects, and it is difficult for the model to capture this difference. Therefore, the prediction of Ethereum price is not yet comprehensive enough. In the future, one will have a more comprehensive discussion on the factors that influence the Ethereum price. This paper offers only a limited explanation for the high volatility of Ethereum prices and the analysis of prices is also not discussed in different time periods, e.g. separate models are built before and after major policy announcements or the COVID-19 outbreak. Therefore, it can be considered to analyze and discuss the Ethereum price in different time periods and create models for each phase to enable a more effective prediction of the Ethereum price.

5. Conclusion

In summary, this study investiagte the feasibility of ETH price predcition based on linear regression and Long Short-Term Memory (LSTM). According to the metric RMSE, The Long Short-Term Memory (LSTM) model is the better one at forecasting the Ethereum price and provides a more accurate result, as one can clearly see the differences between the two models. The model can be further enhanced by considering impacts of parameters such as hash rate and by changing various hyperparameters. There are still many possibilities and things to explore, and one day better models will come out. Overall, these results offer a guideline for personal investment in Ethereum and provide the state with a basis for formulating relevant policies.

References

[1] Nakamoto S. Bitcoin: A peer-to-peer electronic cash system. Decentralized Business Review, 2008: 21260.

[2] Astuti Intan Dwi, Suryazi Rajab, Desky Setiyouji. Cryptocurrency Blockchain Technology in the Digital Revolution Era. Aptisi Transactions on Technopreneurship (ATT), 2022, 4.1: 9-15.

[3] Fröhlich Michael, et al. Blockchain and Cryptocurrency in Human Computer Interaction: A Systematic Literature Review and Research Agenda. arXiv preprint arXiv:2204.10857, 2022.

[4] Hougan Matt, and David Lawant. Cryptoassets: The Guide to Bitcoin, Blockchain, and Cryptocurrency for Investment Professionals. CFA Institute Research Foundation, 2021.

[5] Dhanda Namrata. Cryptocurrency and Blockchain: The Future of a Global Banking System. Regulatory Aspects of Artificial Intelligence on Blockchain. IGI Global, 2022, 181-204.

[6] Zhang Jiefe. Research on Bitcoin Price Prediction and Influencing Factors based on Machine Learning. Master Dissertation, 2021.

[7] Kumar D, Rath S K. Predicting the trends of price for ethereum using deep learning techniques. Artificial Intelligence and Evolutionary Computations in Engineering Systems. Springer, Singapore, 2020, 103-114

[8] Monish S, Mridul Mohta, Shanta Rangaswamy. ETHEREUM PRICE PREDICTION USING MACHINE LEARNING TECHNIQUES–A COMPARATIVE STUDY. IJEAST, 2022.

[9] Shah Rakshit, et al. Linear Regression vs LSTM for Time Series Data. 2022 IEEE World Conference on Applied Intelligence and Computing (AIC). IEEE, 2022.

[10] Singh Sarthak, Shaurya Rehan, and Vimal Kumar. Stock Price Prediction Using Linear Regression, LSTM and Decision Tree. No. 7805. EasyChair, 2022.

[11] Ebenesh C, Anitha K. A Novel Approach to Minimize the Mean Square Error in Predicting Stock Price Index using Linear Regression in Comparison with LSTM Model. 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS). IEEE, 2022.

[12] Uday K C H, Deekshith K, and Alekhya V. Bitcoin price prediction based on linear regression and lstm. South Asian Journal of Engineering and Technology, 2022, 12.3: 87-95.