Forecasting Bitcoin Closing Price by Four Machine Learning Algorithms

Jingjing Zhang*

School of Economics and Management, Dalian University of Technology, Dalian, China

* Corresponding Author Email: 20211011321@mail.dlut.edu.cn

Abstract. Bitcoin has increased in popularity as a speculative asset. Since 2013, eventually becoming the most recognizable cryptocurrency. But it's worth noting that the price of Bitcoin has a very high degree of volatility and diversity, which means the ability to estimate prices accurately is crucial for making wise financial decisions. Although recent research has implemented machine learning to predict Bitcoin prices with greater precision, such as Long short-term memory (LSTM), few have focused on traditional machine learning methods. In this article, the author chose a data set including nearly eight years of daily bitcoin price data for closing price prediction. Four different machine learning algorithms were used simultaneously: the Linear Regression (LR), the Decision Tree (DT) and the Random Forest (RF). An artificial neural network, the Multilayer Perceptron (MLP) was also used in this study. The author altered parameter values using the cross-validation method before creating the models in order to get more precise predictions. Finally, Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and R-squared are used as indicators to assess the outcomes from each model. The study's findings demonstrated that all three metrics of Linear Regression outperformed the performance of the other three models. Perhaps future research could focus more on traditional machine learning algorithms instead of going after complex models.

Keywords: Bitcoin, machine learning, linear regression.

1. Introduction

Satoshi Nakamoto created Bitcoin in 2008 as the world's first digital currency based on the technology of blockchain to overcome the fundamental flaws in confidence in the transaction-based architecture [1]. As the first distributed virtual currency, Bitcoin has witnessed record high prices in recent years. When Bitcoin first appeared, one US dollar could be exchanged for 1,300 Bitcoins. However, the price of Bitcoin grew exponentially in the second half of 2017, reaching a high of $19,000 in November. In February 2021, it surpassed the $58,000 per coin threshold and the market capitalization surpassed $1 trillion. In recent years, the price of bitcoin has risen to all-time highs, and the capital markets' interest in and acceptance of it have grown significantly. These developments have been accompanied by a constant inflow of cash into the bitcoin market. Bitcoin has a lot of benefits over other financial assets that make it appropriate for statistical arbitrage and portfolio construction [2]. Similarly, in terms of the type of transactions and data recording, Bitcoin is quite comparable to the traditional stock market, so it is possible to anticipate the price of Bitcoin using the same techniques as the stock market [3]. By introducing nonlinear characteristics into the forecasting system to deal with non-stationary financial time series, algorithms based on machine learning have been successfully applied to stock market forecasting. The findings have indicated that the methodology is put into use to forecast more accurately [4]. However, because of the erratic price swings of bitcoin and the constantly shifting structure of the market, traders as well as investors are exposed to dangers that could result in significant monetary losses, making it challenging for scholars to innovative methods for price prediction that increase accuracy [5].

In this study, Bitcoin closing price prediction is performed with LR, DT, RF and MLP from machine learning methods by using daily Bitcoin closing prices between 2014-2022. In the stage of training, the cross-validation methodology is utilized to create a model with great performance regardless of the data collection [6]. RMSE, MAE and R-squared are three statistical metrics used to evaluate the effectiveness of the model behaved. For the Bitcoin data set, it can be seen that the LR

model performs better than the other three models regarding price prediction precision. This might encourage researchers to rethink conventional machine learning algorithms.

2. Literature Review

To predict the future value of bitcoin, numerous machine learning algorithms have been created over time. Jing [7] applied the ARDL model, LSTM model, and SPCA-CNN-LSTM to the weekly data of bitcoin price between 2015 and 2020 to predict the bitcoin price. The experimental findings show that the SPCA-CNN-LSTM model predicts more accurately than the other two models and can obtain a high degree of fit for the bitcoin price, while the LSTM model makes the worst predictions. According to McNally et al., LSTM is more accurate in forecasting the Bitcoin price than the ARIMA model [8]. Seckin et al. predicted the price of the cryptocurrency using linear regression (LR) and support vector machines (SVM), both of which are machine learning techniques [6]. The suggested SVM model for Bitcoin data set outperforms the LR model in terms of price prediction performance. Li et al. predicted price fluctuations through LR and KNN algorithms [9]. The degree of certainty of fitness for Bitcoin in the LR prediction is 98.43%, whereas it is 95.06% in the KNN method. Niamkova et al. looked at how COVID-19's impact on various countries affected the overall price of Bitcoin [10]. They twice ran the LSTM model on the data set and utilized the RF model to determine and rank the relevance of the feature, finally discovered that the COVID-19 data increased the LSTM model's ability to predict Bitcoin values. The author found that in recent years, LSTM and neural network algorithms have gradually gained popularity among researchers due to their good predictive performance. However, in some situations, more sophisticated algorithms don't perform as well as conventional machine learning techniques like LR and RF. Ranjan et al. also found that statistical methods such as logistic regression predicted daily prices more accurately than complex machine learning algorithms such as XGBoost predicted prices at 5-minute intervals [11]. As related algorithms continue to improve, the author believe suggest researchers should not overlook simple machine learning algorithms that perhaps perform better in certain prediction scenarios.

3. Research Methodology

This section of the article discusses the dataset's source and composition, some data preprocessing, the four machine learning methods employed, and finally the metrics used to evaluate the prediction outcomes.

3.1. Data Collection and Description

The data set was obtained from Kaggel, a platform where programmers and researchers may produce and exchange code, host databases, and hold machine learning competitions.

Table 1 presents the basic statistical information of the data set from September 17, 2014, to March 25, 2022. As can be seen, the price of Bitcoin has a very high value, rising as high as about $70,000. However, the lowest price is just about $200, highlighting the price of bitcoin's extreme volatility.

Table 1. Data Description

	Open	High	Low	Close
count	2747	2747	2747	2747
mean	11668.600	11981.035	11325.597	11682.892
sted	16323.684	16759.569	15825.585	16330.192
min	176.890	211.731	171.510	178.103
25%	609.122	611.895	606.309	609.234
50%	6371.850	6500.870	6285.630	6376.710
75%	10728.271	10992.469	10412.890	10755.395
max	67549.734	68789.625	66382.063	67566.828
range	67372.837	68577.894	66210.553	67388.725
var	1.399	1.399	1.397	1.397
dis	10119.149	10380.574	9806.581	10146.161

As shown in Fig. 1, the price of Bitcoin has risen very rapidly over this eight-year period, with sharp fluctuations over short periods of time. When looking at the log value of the Bitcoin price, some repeating patterns in the price swings appear to persist even if the Bitcoin price appears to follow a random walk.

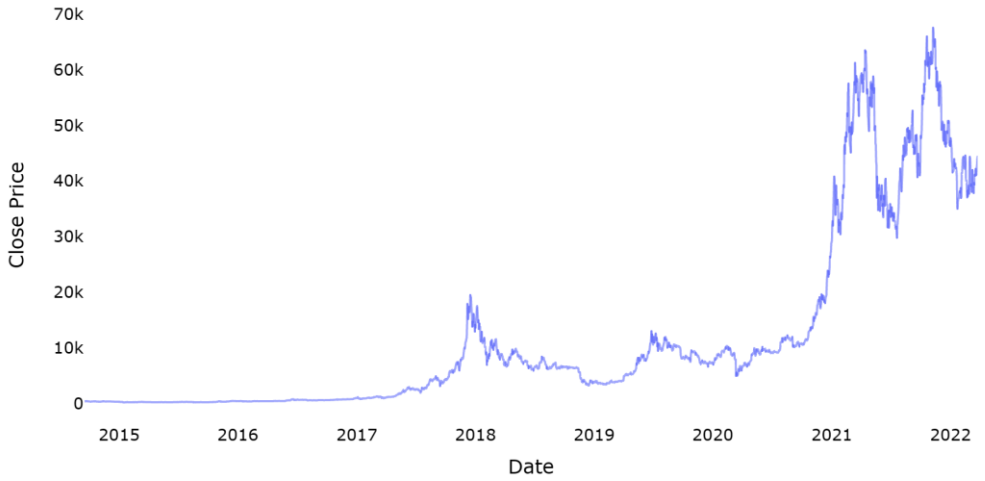

Fig. 1 Bitcoin's Closing Price Movement Trend

3.2. Data Pre-processing

Pre-processing has been done on the data before it is fed into the various models for analysis and comparison. In order not to affect the correlation of the variables in the dataset, the date columns were eliminated prior to modeling. Then, this paper scaled our data and changed the numerical value of our input parameters using the StandardScaler module. The vectors of attributes are brought to the same scale by sizing the features, which shields the model from the negative effects of severe data fluctuations and enhances learning.

The author then explored the correlation between the variables. The heat map in Fig. 2 demonstrates a good link between the different factors, with correlation coefficients of 0.7 or higher for all of them. In addition, as shown in Fig. 3, the dataset's joint distribution plots display the interrelationship between the variables two by two. It is seen a strong linear relationship between the variables.

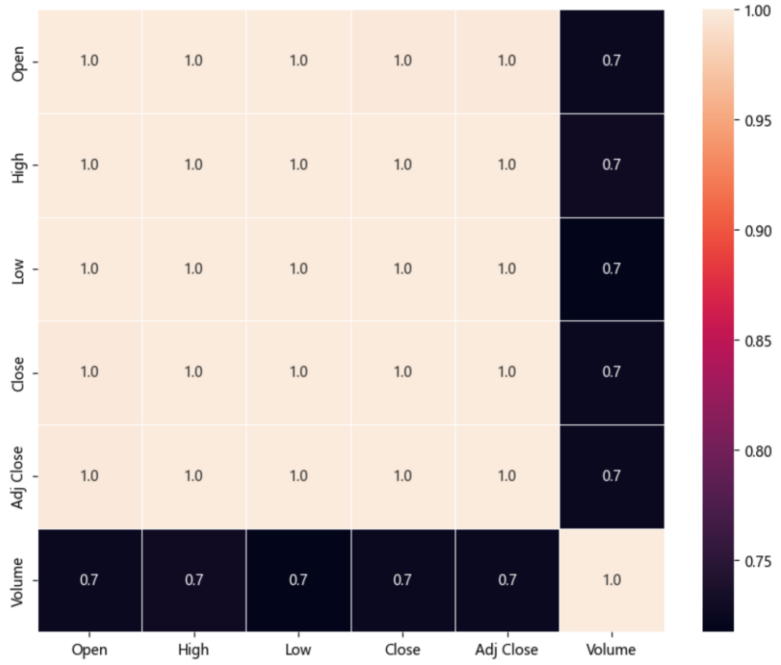

Fig. 2 Heat map

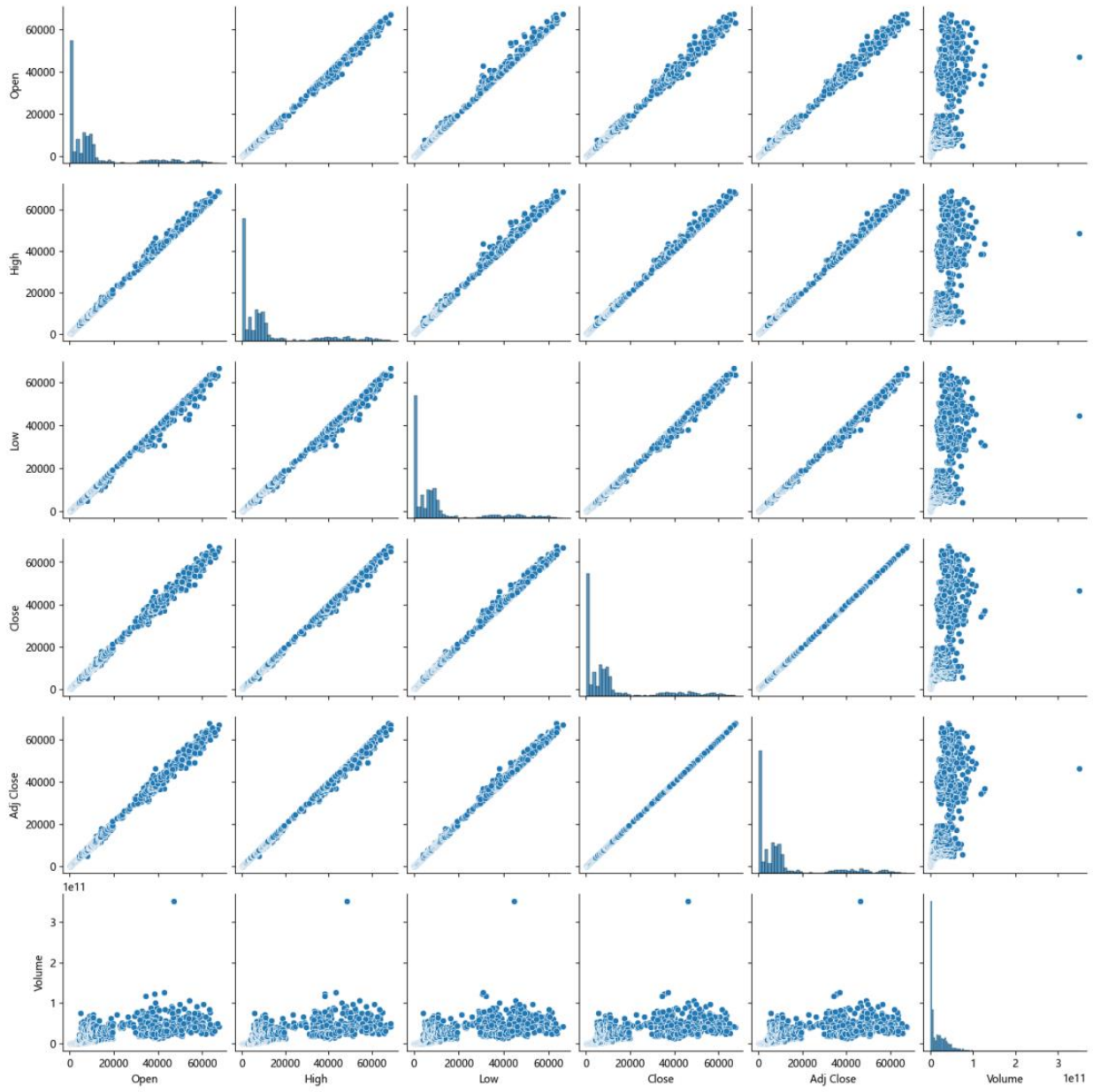

Fig. 3 Joint distribution plots

The machine-learning prediction model was designed with the expectation that it would perform well when given fresh, unproven data. Therefore, this paper divided the bitcoin pricing data into two halves in order to model new, unforeseen data. First, a bigger subset of the data—representing 75% of the original data—was utilized as a training set. Second, a more condensed group, with the other twenty-five percent acting as a test set. The best model should then be chosen based on how effectively it performed on the test set after being built using the training set as a foundation for a predictive model.

3.3. Linear Regression (LR)

The following equation is used in the linear regression approach to model the connection between a dependent variable and one independent variable:

$$y = b_0 + \overrightarrow{b_1} \cdot \overrightarrow{x_1} \tag{1}$$

where b_0 is the intercept and b_1 is the vector of the slope coefficients, and y and x_1 are the dependent and independent variables, respectively. In this study, the elements of the independent variable x_1 vector correspond to the previous closing price and y denotes the expected closing price.

Through this method, the data are approximated and the relationship between the dependent and independent variables are best captured by the curve that most closely resembles the data. The algorithm chooses the line that minimizes the squared differences sum $\sum_i (y_i - \overline{y_i})^2$, that is, the separation between the actual points and those that the line of greatest fit crosses, after calculating and storing the amount $(y - \bar{y})^2$ for each trend line that can be plotted with our data.

3.4. Decision Tree(DT)

Due to the wide range of categorization mathematical methods and the allure of accessibility, decision tree learning is regarded as an information mining tool. Decision tree learners are the essential tools of many real-world artificial intelligence applications owing to their quickness and consistency in delivering comprehensible responses that are frequently unexpectedly correct [12]. The fundamental purpose of a decision tree is to classify known case structures into a tree structure, then use that structure to summarize specific laws in the cases; the resulting decision tree can then utilize that structure to generate predictions outside of the sample. When compared to other data mining techniques, the evolution of decision trees is the most representative method that has been employed in previous research for general classification [13]. It has been confirmed that employing decision tree algorithms to illustrate problem relationships makes the information more understandable than using more standard, general statistical techniques [14].

3.5. Random Forest (RF)

Random forest is a collaborative method that creates a variety of individual regression systems by combining the idea of decision trees with the bootstrapping and aggregating process [15]. Using the Random Forest technique, a significant amount of conditionally autonomous trees are created in order to address the issue of overfitting. This entails integrating the predictor results from different tree-based models scattered across the model and minimizing variance [16]. Both classification and regression frequently employ such methods [17].

3.6. Multilayer Perceptron (MLP)

A multilayer perceptron (MLP) is an artificial neural network with a feed-forward algorithm that produces a set of outputs from a collection of inputs. It consists of a minimum of three layers of neurons. All neurons, excluding the input neurons, have a nonlinear activation function. It was first created by Frank Rosenblatt [18] and is better suited for use in describing intricate interactions between different predictor variables [19].

4. Results

Three indicators were used to assess the models' performance, including RMSE, MAE and R-squared. Their formulas are shown below:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \hat{x_i})^2}{n}} \tag{2}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}|x_i - \hat{x_i}|}{n} \tag{3}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{x_i} - x_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{4}$$

Where n denotes the sample size, x_i the predicted value, $\hat{x_i}$ the real value, and $\bar{x}$ the mean value. Additionally, decreasing RMSE and MAE values and an R2 value that is nearer 1 show that the model is more accurate at predicting outcomes. Our modeling indicates that the prediction performance is robust, and the outcomes are displayed in Figs 4–7.

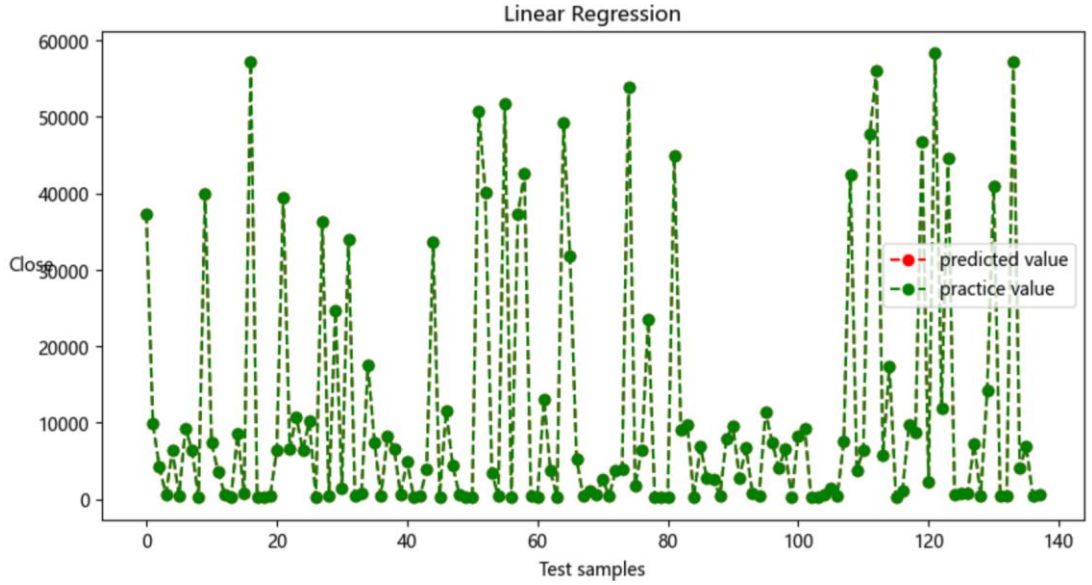

Fig. 4 Predictive Results of Linear Regression

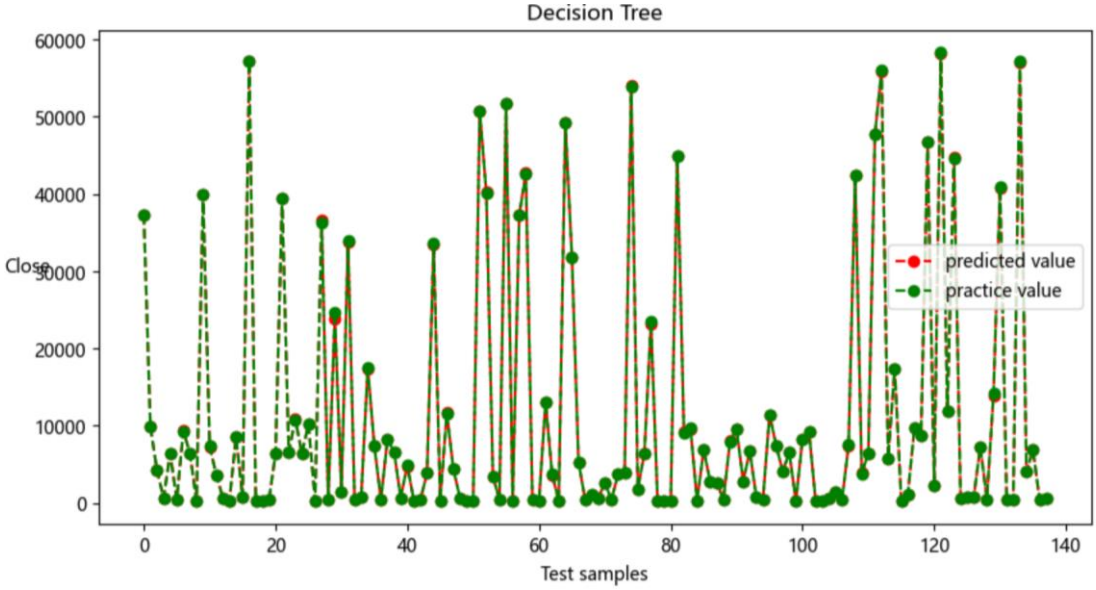

Fig. 5 Predictive Results of Decision Tree

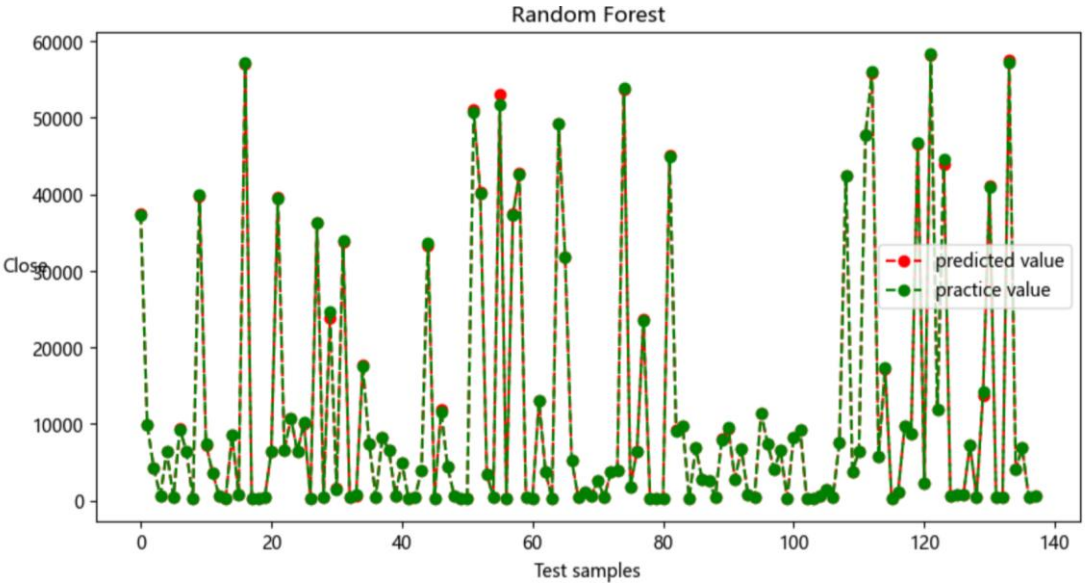

Fig. 6 Predictive Results of Random Forest

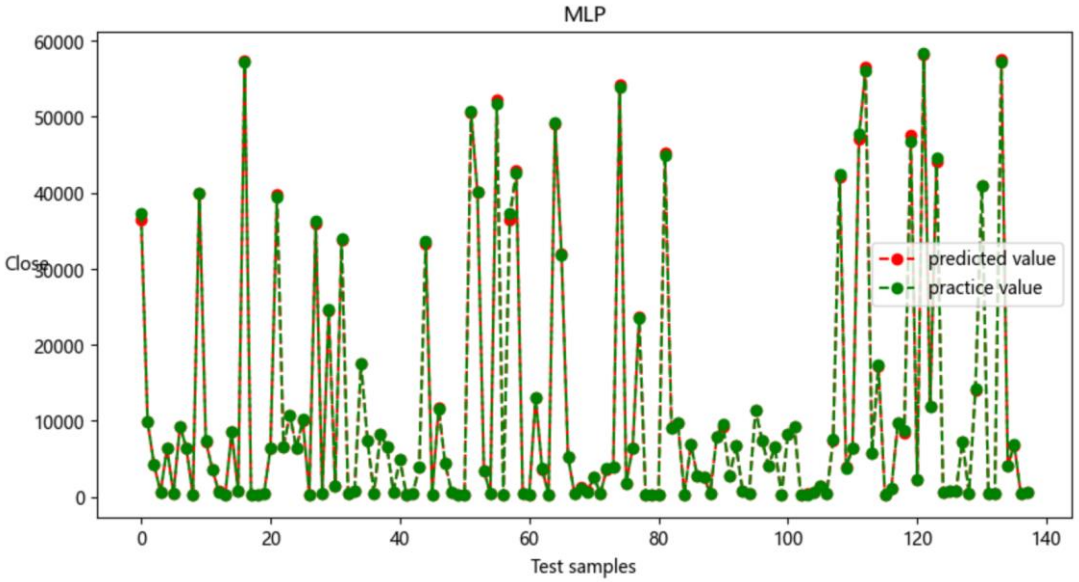

Fig. 7 Predictive Results of Multilayer Perceptron

Table 2 presents the metrics for the four algorithms.

Table 2. Comparing of the Results

Algorithms	Main tuning Parameters	RMSE		MAE		R2 score	
		Train	Test	Train	Test	Train	Test
LR	n_jobs: 1 positive: False	$1.047e^{-11}$	$1.044e^{-11}$	$7.415e^{-12}$	$7.415e^{-12}$	1.000000	1.000000
RF	max_depth:9 min_samples_split: 2 n_estimators: 3	89.238	190.507	32.368	32.368	0.999970	0.999863
DT	max_depth: 9 min_samples_split: 2	20.404	113.051	13.620	13.620	0.999998	0.999952
MLP	hidden_layer_sizes: (80, 90, 80) max_iter: 600	171.627	197.088	80.110	80.110	0.999890	0.999853

5. Conclusion

From the results of the study, The values of R-squared for all four models are very close to 1. However, it can be seen that the Linear Regression has a far better performance than the other three models in all three indicators. The algorithm with the second most accurate prediction results is the Decision Tree, which has a value of about 113 for RMSE and 14 for MAE. In contrast to the results of many studies, the Multilayer Perceptron is the worst predictor, with an RMSE close to 200. Despite the many limitations and shortcomings of this paper, the results of the study show that more traditional or simpler algorithms do not necessarily perform worse than complex algorithms in some cases, which may remind researchers to refocus their attention on traditional algorithms. Perhaps the combination of traditional and complex algorithms can show better accuracy and fitness for some prediction problems.

References

[1] Chen, Z., Li, C., & Sun, W. Bitcoin price prediction using machine learning: an approach to sample dimension engineering. Journal of Computational and Applied Mathematics, 2019, 365: 112395.

[2] Mei Y. Research on Statistical Arbitrage Strategies of High Frequency Data in Bitcoin Market. Guilin University of Electronic Technology, 2021, Thesis for master's degree.

[3] Han, R. Z. Bitcoin Price Prediction Based on Machine Learning. Shandong Normal University, 2022, Thesis for master's degree.

[4] Yuan, W., Chin, K. S., Hua, M., Dong, G., & Wang, C. Shape classification of wear particles by image boundary analysis using machine learning algorithms. Mechanical Systems and Signal Processing, 2016, 72-73: 346-358.

[5] Tripathi, B., & Sharma, R. K. Modeling bitcoin prices using signal processing methods, bayesian optimization, and deep neural networks. Computational Economics, 2022.

[6] Seckin, K., Aytac, A., Zehra, S., & Rifat, H. Prediction of Bitcoin prices with machine learning methods using time series data., Signal Processing and Communications Applications Conference, 2018: 1-4.

[7] Pengfei, J. Comparative Research of Bitcoin Price Prediction Based on Multiple Models. Shanxi University, 2021, Thesis for master's degree.

[8] Sean, M., Jason, R., & Simon, C. Predicting the Price of Bitcoin Using Machine Learning, Parallel, Distributed and Network-Based Processing, 2018: 339-343.

[9] Jingjing, L., Xinge, R., Xianyi, L., & Sihai, G. Gold and Bitcoin Optimal Portfolio Research and Analysis Based on Machine-Learning Methods, Sustainability, 2022, (New Rochelle, N.Y.): 14-21

[10] Palina, N., & Rafael, M. Improved Bitcoin Price Prediction based on COVID-19 data, arXiv preprint arXiv, 2023: 2301.10840

[11] Sumit, R., Parthajit, K., & Malvika, S. Bitcoin Price Prediction: A Machine Learning Sample Dimension Approach, COMPUTATIONAL ECONOMICS, 2023, 61(4): 1617-1636.

[12] Witten, I. H., & Frank, E. Data mining: practical machine learning tools and techniques. Acm Sigmod Record, 2010.

[13] Questier, F., Put, R., Coomans, D., Walczak, B., & Heyden, Y. V. The use of cart and multivariate regression trees for supervised and unsupervised feature selection. Chemometrics & Intelligent Laboratory Systems, 2005, 76(1): 45-54.

[14] Alos, A. D. Z. Decision tree matrix algorithm for detecting contextual faults in unmanned aerial vehicles. Journal of intelligent & fuzzy systems: Applications in Engineering and Technology, 2020, 38.

[15] LeoBreiman. Bagging predictors. Machine learning, 1996.

[16] Francisco, O., José, M., María Jesús, S., & Pablo, S. A random forest-based model for crypto asset forecasts in futures markets with out-of-sample prediction, Research in International Business and Finance, 2022, 64

[17] Breiman. Random forests. MACH LEARN, 2001,45(1): 5-32.

[18] Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological Review, 1958, 65: 386-408.

[19] Abdou, H. A., Mitra, S., Fry, J., & Elamer, A. A. Would two-stage scoring models alleviate bank exposure to bad debt?. Expert Systems with Applications, 2019, 128: 1-13.