BTC, ETH and Dogecoin Price Prediction Based on OLS, Random Forest and XGBoost

Ruhan Hou *

Faculty of Engineering, University of Bristol, United Kingdom

*Corresponding author: ni18080@alumni.bristol.ac.uk

Abstract. In recent years, the digital world is fast speeding developed from decentralised concept to blockchain, then to cryptocurrency. Especially, cryptocurrency is a popular trending in recent decades that attracts different experts from various field. Its high volatility has been attracted plenty of investors while also brings the difficulty for realizing the price forecasting. On this basis, this study uses public cryptocurrency dataset and three analytical models to predict the direction of cryptocurrency's price. To be specific, three underlying assets covering large proportion in cryptocurrency are selected, i.e., Bitcoin, Ethereum and Dogecoin. According to the analysis, the prediction results of different models and approaches will be presented. At the end of study, it gains that the optional model with appropriate hyperparameters based on the judgement of metrics values, which offers relevant suggestions for future works. These results shed light on guiding further exploration of cryptocurrency price prediction in terms the state-of-art machine learning scenarios.

Keywords: Cryptocurrency; machine learning; financial market prediction; random forest; XGBoost.

1. Introduction

From cash to bank card, and from bank card to online payment, only a few decades of time, they way of payment has developed dramatically fast. However, even contactless payment is convenient, it is still controlled and regulated, chasing every activity that card holders paid. On this basis, Friedrich Hayek published 'The Denationalization of Money' in 1976 [1], advocated the establishment of competitively issued private moneys, pointing out the proposal: abolish central banking, allow private money to be issued, and compete freely. This competitive process would discover the best currency. The creation of a virtual payment method, which functions like currency in that it is non-traceable, anonymous, instantaneous, free to use, and can communicate with computer networks, is an original concept that has long been mentioned in the evolution of the fast-moving world [2]. In a 2008 publication, Satoshi Nakamoto, who is accused of creating the original version of the Bitcoin programme, presented the idea of cryptocurrencies to the world [3]. Since Satoshi unleashed a cryptocurrency, it immediately spread throughout the quick-paced world, enabling users to purchase it online and then use it to make purchases anonymously and freely. Blockchain is Satoshi's invention, and it underpins Bitcoin and all subsequent cryptocurrencies.

A blockchain is a method of storing data that makes it challenging or impossible to alter, hack, or defraud the system [4]. It is a ledger of transactions, similar to what a bank keeps, but copies of it are spread across computers around the globe, automatically updating with every transaction (every new transaction will be logged to every participant's ledger), and always up-to-date and 100 percent validated [5]. The idea's strength is that it is decentralised, or distributed ledger technology, which refers to a decentralised database that is managed by several participants as opposed to being centrally controlled by a single entity. However, maintaining this distributed ledger requires a lot of labour (DLT, where transactions are recorded with immutable cryptographic signatures called hashes) [4].

This is an incredibly potent idea since it not only addressed the issue of money on the internet, but it also addressed the issue of trust. In addition, similar to all other financial market, cryptocurrencies' investments are treated like investing in stock market, seizing the moment to buy low and sell high with different ones. Nowadays, there are several cryptocurrencies, e.g., Ethereum (ETH), Litecoin, Ripple, and even the peculiar Dogecoin. People have wagered a total of $500 billion on various wagers. Since it is hard to analyse all cryptocurrencies, the three most representative cryptocurrencies are selected and analysed using machine learning algorithms here, which are Bitcoin, Ethereum and Dogecoin.

Based on the research of Rajat Kumar, Deepti Mishra, et al. Recently, they built a Fbprophet model as key analysis model to predict the price of Bitcoin, which increases the accuracy of the prediction compared to LSTM (Long short-term memory) and ARIMA (Autoregressive integrated moving average) [6]. Their model relatively eliminates shortages in the LSTM and ARIMA models when analysing cryptocurrency assets by considering seasonal fluctuations in the data. It is a valuable rouse to be added into this study as one of the features to analyse. To forecast cryptocurrency price direction, the analysis process not only considers specific features but also uses machine learning models with optional parameters. Syed and Perry highlighted using random forest and traditional logit econometric models to achieve high accuracy for predicting Bitcoin price direction [7]. They gained prediction accuracy greater than 85% for 10 to 20 forecast horizons. Compared to the research of Rajat Kumar, Deepti Mishra, etc. [6], Syed and Perry mentioned that inflation is not an important predictor for Bitcoin, but technical indicators are important features for prediction since they.

Since the explosion and development of decentralization technologies, lots of innovative concepts are created, from digital currency to Web 3.0, this technology is driving the concept of decentralization. In this case, in order to understand this development deeply, it starts with analysing the cryptocurrency assets valuation, and trying to forecast its trending based on machine learning in this study. In this study, two main contributions are introduced, cryptocurrency analysis and forecast, and machine learning models comparison. Based on the recent research, three cryptocurrencies are analysed and predicted by using three machine learning models. The process includes data collection, model's parameters adjusting, results comparison, limitation, and some summaries with future work.

2. Data & Method

2.1 Date Preparation

The analysis is started from importing data, which comes in the form of historic pricing data for Bitcoin (BTC-USD), Ethereum (EHT-USD), and Dogecoin (DOGE-USD) from 2018-01-01 to 2022-11-16. The dataset is directly fetched from the yahoo finance by using yfinance (an open-source library can be used to access the financial data available on Yahoo Finance [7]) and loading it into a pandas data frame. It contains 6 features for each trading day as shown in Table 1 [8].

Table 1. The cryptocurrency's features and relevant meaning.

Features	Meaning
Date	The value's date
Open	The cryptocurrency's opening price at the start of the trading day.
High	The cryptocurrency's peak price on a trading day.
Low	The cryptocurrency's lowest point of the price on a trading day.
Close	The cryptocurrency's closing price at the end of a trading day.
Adj Close	The worth of a cryptocurrency after dividends have been distributed is valued by its amended closing price.
Volume	The total volume of cryptocurrency exchanged over a specific time period.

To check the quality of dataset, the closing price of three cryptocurrencies are visualised in Fig. 1. It displays that the trending of cryptocurrency dramatically increases from 2021. It is essential to check the quality of dataset, thus, visualise the correlation between each feature (Fig. 2). To gain a proper data frame, transform the data into a processable shape to train within the models. Next step is selecting specific features (Open, High, Low, Close, Adj Close, Volume and Prediction) and scaling the data to a standard value range. After checking the quality of dataset, the next step is splitting training and test set. The target is 'Adj Close' and selected features contain 'Open', 'High', 'Low'

('Close' is dropped out since it is high correlated to the target). Once all the preparation and splitting are settled down, it can start to build models with proper hypermeters.

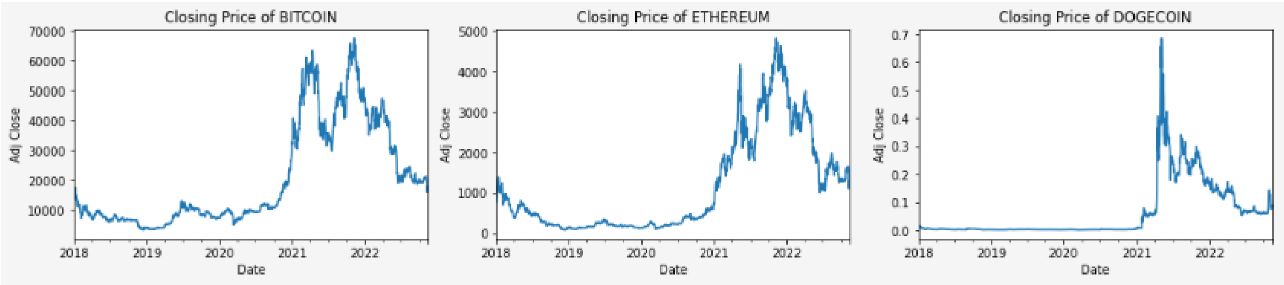

Fig. 1 The closing price of Bitcoin, Ethereum, and dogecoin

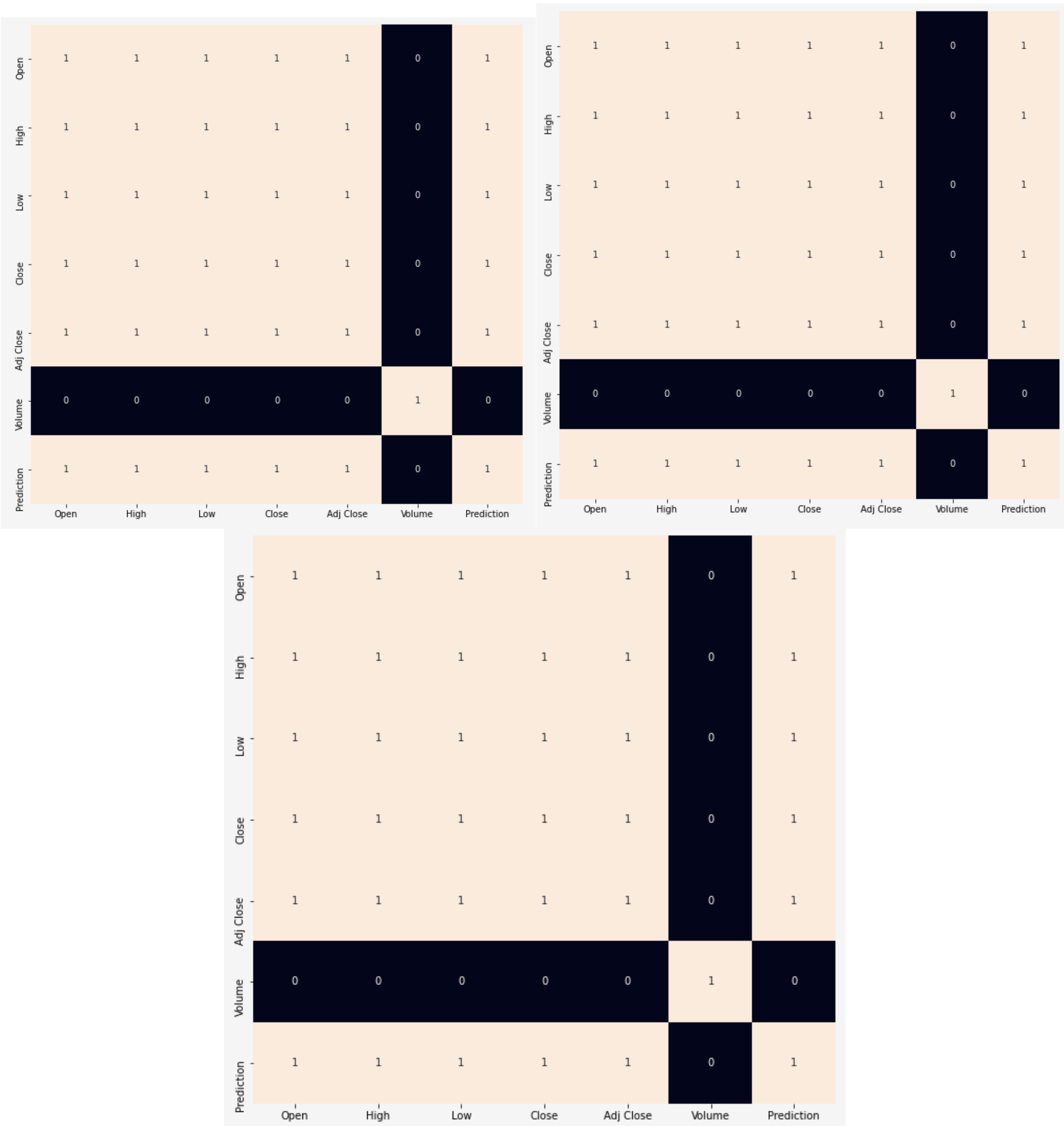

Fig. 2. The correlation of features (top-left: Bitcoin, top-right: Ethereum, bottom: Dogecoin)

2.2 Machine Learning Models

In this study, there are three machine learning algorithms used to analyse the cryptocurrency price direction, namely OLS, random forest, and XGBoost. The idea starts with relatively simple model to much more complex model and adjust related parameters to evaluate the sample. Then, one compares different models with different parameters to find out which features are the most relevant and affective. Ordinal least squares regression is a well-liked and popular technique for finding the coefficients of linear regression equations that depict the relationship between one or more independent quantitative variables and a dependent variable [9]. The technique of random forest is quite simple since it offers an interpretable decision-making process [10]. For a single tree, all nodes, except terminal nodes, are called decision nodes which typically have two child nodes and contain a conditional expression involving a single feature of the input data. The input data starts from the root node and is evaluated through each layer of nodes with specific features until it reaches the end of tree, the decision tree can produce an output class (classification) or a target value (regression). Then, each tree is modified the training procedure by random forest. XGBoost (Extreme Gradient Boosting) is stronger version of tree model, integrating many weak classifiers together to constitute a strong classifier [11]. It is one of boosting algorithms, consisting of a set of classification and CART regression trees. The goal of this technique is to train a new function that will fit the residuals from the previous prediction. After training, it needs to predict the score of a sample once it has k trees. Each tree will fit into a matching leaf node, and each leaf node will correlate to a score, based on the features of this sample. In order to get the projected value of this sample, it must sum up the associated scores of each tree. Simply, this model is summing the prediction of multiple trees together. It is a highly efficient algorithm that can solve problems beyond billions of examples [12].

2.3 Metrics

Regression-based machine learning models are used by data scientists to predict the values of continuous attributes. Like all supervised machine learning problems, a regression model is trained with a set of features (X) and learns the influence of a target variable (y). In regression analysis, the aim is a continuous variable, such as the price of a house. A number of measures must be taken into account while assessing the success of the regression model, much like with other machine learning tasks. The metrics will vary depending on the training model's data and application. Given the possibility of bias, the magnitude of the inaccuracy, and the potential effects of using the model in practise, it is usually necessary to assess overall performance using a variety of indicators, with benchmark comparisons frequently being the most helpful.

R-squared (R^2) is a statistical measure that indicates the percentage of the model's goodness of fit. It is calculated as the ratio of the regression sum of squares to the total sum of squares, which shows the proportion of the dependent variable's total variation that can be explained by the regression model. It is also used as a benchmark in investing [13]. The range of the results is [0,1], and the closer it is to 1, the higher the proportion of the sum of squares from regression to all other squares will be. Additionally, the closer the regression line is to each observation point, the better the regression's ability to match the data. The formula for R-squared can be presented as follows:

$$\mathrm{R}^2 = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}} \tag{1}$$

Mean Absolute Error (MAE) is the amount of error in the measurements, which is the difference between the expected value and predicted value [14]. The difference will be positive when using the MAE since the score is measured as the average of the absolute error values. The range of the outcome is $[0, +\infty)$, when the predicted value is exactly in the line with the real value, it is equal to 0, which is the ideal model, the larger the error, the larger the value. The MAE value can be calculated as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i| \tag{2}$$

Here. n is the number of errors, Σ is the summation symbol, $|\hat{y}_i - y_i|$ is the absolute errors.

The Mean Absolute Percent Error (MAPE) measures the size of the error in percentage terms, which is equivalent to a weighted version of MAE [15]. Since the variable's units are scaled to percentage units, which is the most typical measurement used to anticipate error, it is simpler to understand. Additionally, the benefit of MAPE is that it offers a reference point for quick comparison, specifically the 0 value. The MAPE result is as near to 0 as feasible when the predicted y is precisely the same as the observed y, as the minimal MAPE value is 0. However, MAPE has a shortcoming is when the actual value y is very small, even if a little error, MAPE 's calculations are huge. The MAPE value can be wrote as follows:

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{3}$$

In order to evaluate the success of regression machine learning models, an unusual metric known as MDAPE (Median Absolute Percentage Error), which is the median of all absolute percentage errors discovered between forecasts and their corresponding actual values, is utilised [16]. The resulting value is also easier to understand since it is returned as a percentage unit. The formula of MDAPE can be expressed as follows:

$$\mathrm{MDAPE} = median\left(\frac{|\mathrm{actual} - \mathrm{prediction}|}{\mathrm{actual}}\right) \tag{4}$$

Compared to the MAPE, MDAPE is much less sensitive to outliers than MAPE because MDPAE returns the median value of all the errors, while MAPE returns the mean.

Accuracy is an index used to evaluate a classification model. Simply, the model predicts the correct amount as a percentage of the total. In this study, it is calculated using the percentage of difference between 100 and MAPE's mean.

$$\mathrm{Accuracy} = 100 - mean(\mathrm{MAPE}) \tag{5}$$

3. Results & Discussion

3.1 BTC

By the end of 2017, Bitcoin had been so widely used that its value had caught up to that of the major banks. And with just a few simple steps, owners may use Bitcoin to pay for things that are often purchased with unlawful services [17]. Given their popularity among individuals looking to hide, virtual currencies can be difficult for law authorities to police. the privacy provided by currency, but only cash that was not issued by governments. Instead, it was created using code that was driven by internet users, thus, does not qualify as legal tender. It is probably the biggest innovation that have seen since the internet. And money is counterfeit, unless issued by the federal government. The fundamental promise of Blockchain is decentralization. The first blockchain test was using bitcoin. a monetary system with no rules, no laws, and no one in charge.

Table 2. The metrics' results of Bitcoin

Bitcoin	R^2	MAE	MAPE	MDAPE	Accuracy
OLS	0.997858	370.5114	1.1589%	0.7909%	98.8400%
Random Forest	0.992622	742.7505	2.6597%	1.7062%	97.3400%
XGBoost	0.993763	0.002055	1.8026%	1.2904%	98.2000%

Seen from the Table 2, the mean absolute error is 0.002055 using XGBoost, which is relatively small, which means XGBoost model perform quite well than other two models. However, the mean absolute percentage error is 1.8026%, which means that the mean of our predictions deviates from the actual values by 1.8026%, and it is larger than OLS model's MAPE (1.1589%). Then, the accuracy of OLS model has the highest result 98.84% and Random Forest model gains 97.34%, which is quite less than others. Using OLS model to predict, it can gain 99.2091% accurate in predicting since the median absolute percentage error is 0.7909%.

3.2 ETH

Programmer Vitalik Buterin, who was inspired by Bitcoin, originally put up the idea of Ethereum in 2013 or 2014, describing it as "next generation cryptocurrency and decentralised application platform." A blockchain platform with its own cryptocurrency, named Ether (ETH) or Ethereum, was launched in 2015 after starting to develop in 2014 through an initial coin offering (ICO) crowdfunding campaign. After Bitcoin, it is the most widely used cryptocurrency [18]. Using the same machine learning models with appropriate hyperparameters, the results of related metrics' results are gained (Table 3). The MAE all quite small compared to them of Bitcoin. Similarly, OLS model presents the highest accuracy 98.49% than other Random Forest and XGBoost models even XGBoost model given the lowest MAE value 0.006526.

Table 3. The metrics' results of Ethereum.

Ethereum	R^2	MAE	MAPE	MDAPE	Accuracy
OLS	0.99731	33.61963	1.5120%	1.1411%	98.4900%
Random Forest	0.995369	48.3457	2.2610%	1.7074%	97.7400%
XGBoost	0.950591	0.006526	5.2584%	3.4000%	94.7400%

3.3 Dogecoin

In 2013, Adobe software developer Jackson Palmer and IBM software engineer Billy Markus created the Shiba Inu-inspired cryptocurrency known as Dogecoin as a joke based on an online meme [19]. The market valuation of the company is now about $500 million. Unexpectedly, XGBoost gains the worse accuracy 64.3503% compared to OLS and Random Forest models (as shown in Table 4). The R-squared value is negative since R-squared refers to the coefficients a linear relationship, then it is reasonable to be negative when using XGBoost model for predicting. Although XGBoost has the smallest MAE, it cannot be considered to use for predicting Dogecoin price direction since its accuracy is too low to be used.

Table 4. The metrics' results of Dogecoin

Dogecoin	R^2	MAE	MAPE	MDAPE	Accuracy
OLS	0.997858	370.5114	1.1589%	0.7909%	98.8400%
Random Forest	0.992622	742.7505	2.6597%	1.7062%	97.3400%
XGBoost	-0.38271	0.041756	35.6497%	35.2859%	64.3503%

3.4 Comparison & Suggestions

After using the model of OLS for predicting cryptocurrencies' price direction, the coefficients of three cryptocurrencies are gained as listed in Table 5. It shows that open price has negative relation with the predicted price, and other two features have positive relation with the predicted value.

Table 5. The coefficient of OLS model for Bitcoin, Ethereum, Dogecoin

OLS	Coefficient		
	Open	**High**	**Low**
Bitcoin	-0.5447961	0.90731258	0.63334671
Ethereum	-0.5875953	0.89070471	0.69442179
Dogecoin	-0.5342767	0.72913922	0.81379449

For Random Forest model, there are 7 parameters that are settled down before fitting with training data (summarized in Table. 6). Each parameter value is calculated by using best parameter function from initialising the RandomizedSearchCV object (it provides a statistical distribution or list of hyper parameters) with estimators is RandomForestRegressor of cross validation is 5 with n_jobs is -1 (it is parallel processing).

Table 6. The hyperparameters of Random Forest model

Parameter	Value	Parameter	Value
max_depth	14	min_samples_split	2
min_samples_leaf	3	max_features	auto
n_estimators	20	bootstrap	TRUE
random_state	2		

Initially, the XGBoost regressor has only one parameter ‘n_estimator=1000’ at default, which gained the relatively unexpected value (shown in Table 7). Since their accuracies are all unvaluable, hyperparameter tunning is necessary at this point. Technically, XGBoost can perform better than the previous models, however it displays lower accuracy than other two models. In this case, in order to optimise the XGBoost model, there are several parameters (summarized in the Table 8) added into the hyperparameter tunning list. Based on this, the performance of the model increased about 98% accuracy for predicting Bitcoin, 95% accuracy for predicting Ethereum, and about 1% accuracy for Dogecoin, which are displayed in Table 9.

Table 7. The results of metrics for XGBoost model with n_estimator=1000

	R^2	MAE	MAPE	MDAPE	Accuracy
Bitcoin	-7.99	32088.35	99.999568%	99.999529%	0.0004321%
Ethereum	-5.89	2312.42	99.991035%	99.991148%	0.0089650 %
Dogecoin	-0.43	0.04241	36.464147%	35.518848%	63.5358531%

Table 8. The optimised hyperparameter of XGBoost model

Parameters	Values	Parameters	Values
n_estimator	[10, 20, 50, 100, 200, 300, 400, 500, 600]	**seed**	0
learning_rate	0.1	**subsample**	0.8
n_estimators	500	**colsample_bytree**	0.8
max_depth	5	**gamma**	0
min_child_weight	1	**reg_alpha**	0
reg_lambda	1		

Table 9. The XGBoost model with optimised hyperparameters

XGBoost	R^2	MAE	MAPE	MDAPE	Accuracy
Bitcoin	0.99376294	0.002055	1.8026%	1.2904%	98.2000%
Ethereum	0.95059136	0.006526	5.2584%	4.0000%	94.7400%
Dogecoin	-0.3827148	0.0417558	35.6497%	35.2859%	64.3503%

After adding new parameters, the performance of model is optimised, but it is not the optimal model since the accuracy relatively low compared to the two previous models. Additionally, tunning these parameters with different values to optimise the model. The results seem to provide different proper model for different cryptocurrencies. In general, OLS model offers the highest accuracy for predicting all three cryptocurrencies, which generates more usable results for cryptocurrencies price prediction. Although XGBoost are potential to yield a valuable result, it gains relevantly different performances based on different training data and hyperparameters, which is also applies to Random Forest model as well.

4. Conclusion

In summary this paper investigates the three cryptocurrencies price direction (i.e., Bitcoin, Ethereum, and Dogecoin) based on three different analytical regression models (OLS, Random Forest, and XGBoost). The whole analytical process starts from data preparation to creating features to model training, and then fitting the models to make predictions. According to the analysis, three models can capture hidden dynamics and make predictions. Although hyperparameter tunning is used in this study, it is not enough to predict complex cryptocurrency market since there are various affective features that have been considered in the step of selecting related features. In the future, the potential work may explore other datasets with various relevant features for predicting more accurate values.

References

[1] Hayek F A. Denationalization of Money (1976). The Money Changers. Routledge, 2015: 157-160.

[2] Trust No One: The Hunt for the Crypto King. Retrieved from: https://www.netflix.com/gb/title/81349029

[3] Hayes A. Who Is Satoshi Nakamoto. Investopedia, Investopedia, 28 Sept. 2022, Retreived from: https://www.investopedia.com/terms/s/satoshi-nakamoto.asp.

[4] Di Pierro M. What is the blockchain?. Computing in Science & Engineering, 2017, 19(5): 92-95..

[5] Rathore R K, Mishra D, Mehra P S, et al. Real-world model for bitcoin price prediction. Information Processing & Management, 2022, 59(4): 102968.

[6] Basher S A, Sadorsky P. Forecasting Bitcoin price direction with random forests: How important are interest rates, inflation, and market volatility? Machine Learning with Applications, 2022: 100355.

[7] Crypto Real Time Prices & Latest News. Yahoo! Finance, Yahoo!, Retrieved from: https://finance.yahoo.com/crypto/.

[8] Yfinance. PyPI, retrieved from: https://pypi.org/project/yfinance/.

[9] Norris E. The Linear Regression of Time and Price. Investopedia, Investopedia, 6 Jan. 2022, Retrieved from: https://www.investopedia.com/articles/trading/09/linear-regression-time-price.asp.

[10] James C. Machine Learning. COMS30035, Retrieved from: https://uob-coms30035.github.io/.

[11] Chen T, He T, Benesty M, et al. Xgboost: extreme gradient boosting. R package version 0.4-2, 2015, 1(4): 1-4.

[12] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.

[13] Figueiredo Filho D B, Júnior J A S, Rocha E C. What is R2 all about?. Leviathan (São Paulo), 2011 (3): 60-68.

[14] Wang, Weijie, and Yanmin Lu. “Analysis of the Mean Absolute Error (Mae) and the Root Mean Square ...” Researchgate, Mar. 2018, https://www.researchgate.net/publication/324265077_Analysis_of_the_Mean_Absolute_Error_MAE_and_the_Root_Mean_Square_Error_RMSE_in_Assessing_Rounding_Model.

[15] Oracle Help Center. Working with Planning. 1 Nov. 2022, Retrieved from: https://docs.oracle.com/en/cloud/saas/planning-budgeting-cloud/pfusu/insights_metrics_MAPE.html.

[16] Aladag C H, Egrioglu E, Kadilar C. Improvement in forecasting accuracy using the hybrid model of ARFIMA and feed forward neural network. American Journal of Intelligent Systems, 2012, 2(2): 12-17.

[17] Fauzi M A, Paiman N, Othman Z. Bitcoin and cryptocurrency: Challenges, opportunities and future works. The Journal of Asian Finance, Economics and Business, 2020, 7(8): 695-704.

[18] Tikhomirov S. Ethereum: state of knowledge and research perspectives. International Symposium on Foundations and Practice of Security. Springer, Cham, 2017: 206-221.

[19] Young I. Dogecoin: A brief overview & survey[J]. Available at SSRN 3306060, 2018.