Cryptocurrency Price Forecast with Use of Time Series Model, Conventional Statistic Model and Machine Learning Approaches

Yiran Wang *
[*] Corresponding author: ywang085@e.ntu.edu.sg

Ms. School of Physics and Mathematical Sciences Nanyang Technological University Singapore

Abstract—Cryptocurrency is an emerging virtual currency gathering attention from the public, which is difficult to counterfeit or double-spend. One of the most significant discussions in cryptocurrency is its extremely high volatility which brings challenges on price prediction. Therefore, lots of research have been conducted to predict cryptocurrency by implementing different models. In this paper, various time series models, conventional linear models, and machine learning models are compared in terms of predictive performance on BTC-USD prices based on historical 5-year daily information and technical features derived. On this basis, various metrics are adopted including AIC, RMSE, MAE, and R-Square and are respectively evaluated and compared. Time series models, ARIMA and GARCH, have relatively poor predictive performance. For statistic models, linear regression, Ridge and Lasso are evaluated respectively, where both regularized models are not able to outperform the linear regression. Regarding to machine learning models, ensemble tree methods including random forest and LightGBM have relatively better performance than other types of models. Among all models tested for the same split, random forest has the lowest error and highest coefficient of determination, and its predictions are the most accurate. These results shed light on choosing from different models in cryptocurrency price prediction.

Keywords-Cryptocurrency; time series; conventional statistics; machine learning.

1 INTRODUCTION

Cryptocurrency is one of the topics that is currently on the spotlight. The idea of cryptocurrency came up in the late 1980s, while cryptocurrency market develops extremely fast in the past 15 years [1]. Typical currencies include fiat currency, private currency, virtual currency, alternative currency [2]. A cryptocurrency is a currency in digital or virtual form used in financial systems and its mechanism makes it nearly impossible to be counterfeited or double-spent [3]. Cryptocurrency can be converted via cryptographic procedures, which differentiates it from traditional currencies in the market [4], and the backing technology called Blockchain ensures it to be difficult to alter, hack, or fraud [5]. The most prominent cryptocurrency in the market is Bitcoin which was established in 2009, while thousands of cryptocurrencies emerge contemporarily [6]. The features of Bitcoin to mix monetary value with encryption technology provide it with broad attention from economics, computer science and cryptography fields [7].

Because of price volatility in the dynamic market, cryptocurrency price forecasting is deemed to be difficult. Previously, scholars revealed the volatility and shock transmission characteristics of

FFIT 2022, October 28-30, Shenzhen, People's Republic of China

DOI 10.4108/eai.28-10-2022.2328429

Bitcoin and national currencies, while economic policy uncertainty puts on little impact [8]. Methods that are widely used in cryptocurrency price prediction includes time series forecasting, machine learning, and deep learning methods (random forest, gradient boosting machine, and long short-term memory) [9, 10]. Time series models utilize the time-dependent feature of data to perform predictions, as compared with statistic regression methods that use features other than the objective variable as predictors [11]. Both time series models and conventional statistic models are commonly regarded as baseline and benchmark for relative simplicity. On the other hand, machine learning and deep learning methods are popular in prediction because of their outstanding performance in various environments [12].

Due to the emergence of artificial intelligence and its implementation on statistics and data mining, researchers have shown an increased interest in various algorithms in price prediction. The motivation of this study, as a result, is to compare commonly used methods in cryptocurrency price prediction. In this study, time series models, statistic models, and machine learning models are compared in terms of predictive performance (R-Square, MSE and AIC) on cryptocurrency price.

The rest part of the paper is organized as follows. The Section II will demonstrate the methodology in the research, and the Section III will discuss the performance of different models' predictions, potential limitation and future improvements. Eventually, a brief summary will be given in Section IV.

2 METHODOLOGY

The underlined cryptocurrency is chosen to be BTC-USD, which is a ticker symbol for Bitcoin and the United States dollar exchange rate. The price data of BTC-USD was downloaded from Yahoo Finance, which contains daily price information over the 5-year period (04/05/17 to 04/05/22), including open, high, low, close, adjusted close and volume. Technical features computation is then applied to the raw data to obtain KDJ, CCI, RSI, ATR and WR, which are five most important technical features in cryptocurrency price prediction.

In this study, time series models, statistic models, and machine learning models are implemented in the prediction. In both statistic models and machine learning models, features are selected as a combination of the technical predictors (KDJ, CCI, RSI, ATR and WR) and the price data, while the time series models only take close price as input. The evaluation and comparison of models is based on the coefficient of determination (R-Square), mean squared error (MSE), mean absolute error (MAE) and Akaike information criterion (AIC).

After preprocessing, the dataset is first split into training (first 80% of data) and test set (last 20% of data) prior to modeling, which is to ensure that the fitting process does not contain any external future information. Afterwards, models in interest are trained from the training set, with hyperparameters tuned by cross validation and parameters adjusted to obtain lowest training error satisfying model constraints. Finally, the fitted models are applied to the test data to make prediction, and the results are compared with the true prices in the test data.

3 RESULTS AND DISCUSSION

3.1 Exploratory Data Analysis

In order to visualize the general pattern of the equity, exploratory data analysis is first performed. Figure 1 and Figure 2 demonstrate close price and daily return over time. From 2020 October onwards, the price of BTC-USD has been increasing dramatically. In addition, both price and daily return data are fluctuating wildly with respect to time, which is consistent with the volatile characteristic of cryptocurrency. Figure 3 shows the correlation between predictors and close price. Accordingly, ATR has the highest correlation with price, and KDJ, CCI, RSI and WR have strong correlation with each other, while the correlation coefficients with price is low.

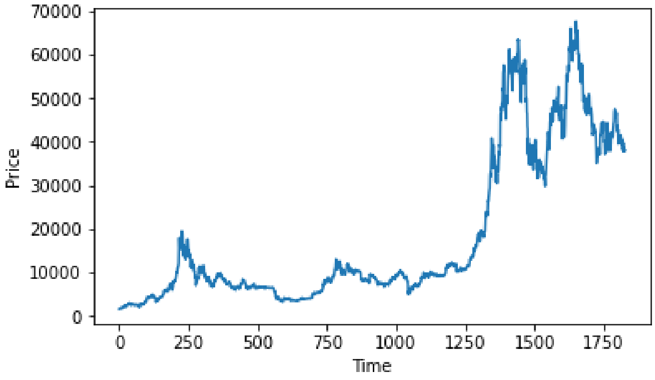

Figure 1. The price trends of BTC-USD.

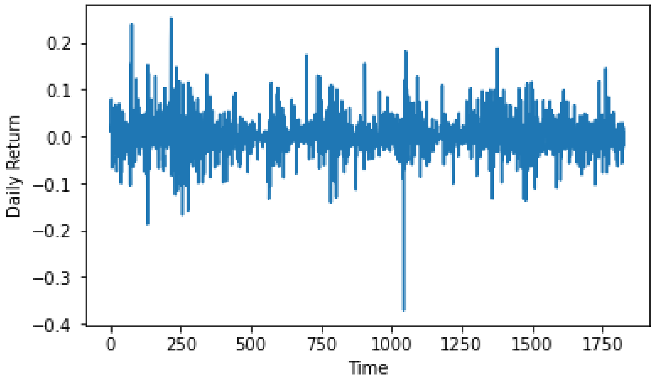

Figure 2. The return of BTC-USD.

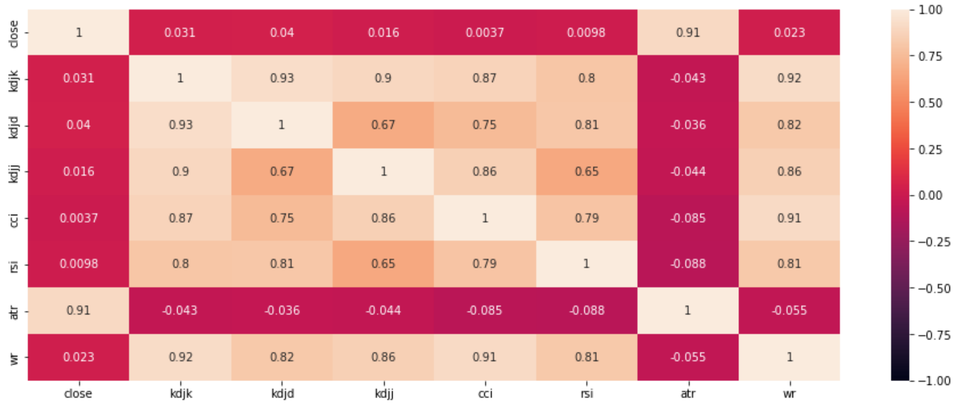

Figure 3. The correlation coefficients of features

3.2 Time Series Models

For time series models, Autoregressive Integrated Moving Average (ARIMA) and Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) models are tested. ARIMA is one of the simplest models in dealing with time series data. Therefore, it is suitable and convenient in analyzing price, and it integrates the combination of autocorrelation and moving average. However, ARIMA assumes constant variance. GARCH, on the other side, assumes varying variance with the term heteroskedasticity. Therefore, ARIMA and GARCH are separately evaluated in the research.

In time series analysis, close price, the objective variable, is the only variable in interest. Because the data input to ARIMA and GARCH needs to be stationary, prior to any fitting process, price data needs to be tested based on stationarity and possible differentiation might be needed. To perform this purpose, Autocorrelation Plot (ACF) and Partial Autocorrelation Plot (PACF) with and without first order differentiation are produced, as depicted in Figure 4. According to the plot, first order differentiation is necessary for the price data to become stationary. Augmented Dickey-Fuller test is then used to statistically conclude the stationarity of data, which gives a p-value of 1.9×10^{-12} on the data with first order differentiation. Hence, the hypothesis that the differentiated data is not stationary would be rejected.

Table 1 Time series model hyperparameters

	(1, 1, 1)	(1, 1, 2)	(2, 1, 1)	(2, 1, 2)
ARIMA AIC	23272.1	23273.2	23273.3	23244.7
	(1, 1)	(1, 2)	(2, 1)	(2, 2)
GARCH AIC	30596.6	30598.6	30596.6	30600.6

For ARIMA model, lag p was determined for AR based on the most significant lag in the PACF, lag q for MA based on ACF, and hyperparameter r determined by orders of differentiation. As illustrated in the first order differenced ACF and PACF in Figure 4 and Table 1, 1 or 2 for p and 1 or 2 for q might be tested, while r is chosen to be 1. In other words, one tests ARIMA(p, r, q) with (1, 1, 1), (1, 1, 2), (2, 1, 1), and (2, 1, 2). As given by the Akaike Information Criterion

(AIC), the most appropriate ARIMA model is ARIMA(2, 1, 2), which gives RMSE = 20924.3 in the prediction of the test set. The prediction graph is depicted in Figure 5.

For GARCH model, the same set of p and q is tested, and the best GARCH model among all is determined to be GARCH(1, 1). Owing to the mechanism of GARCH model, GARCH does not produce future realizations, but forecasts future volatility. Therefore, test RMAE, test RMSE and test R-Square are undefined for GARCH. In further research, Probability Integral Transform (PIT) might be used to evaluate the GARCH model. The Table 2 summarizes the metrics values of the two models.

Table 2 Time series model metrics

	ARIMA(2, 1, 2)	GARCH(1, 1)
Training AIC	23244.7	30596.6
Test MAE	18888.3410	Not Applicable
Test RMSE	20924.3375	Not Applicable
Test R-Square	-5.0837	Not Applicable

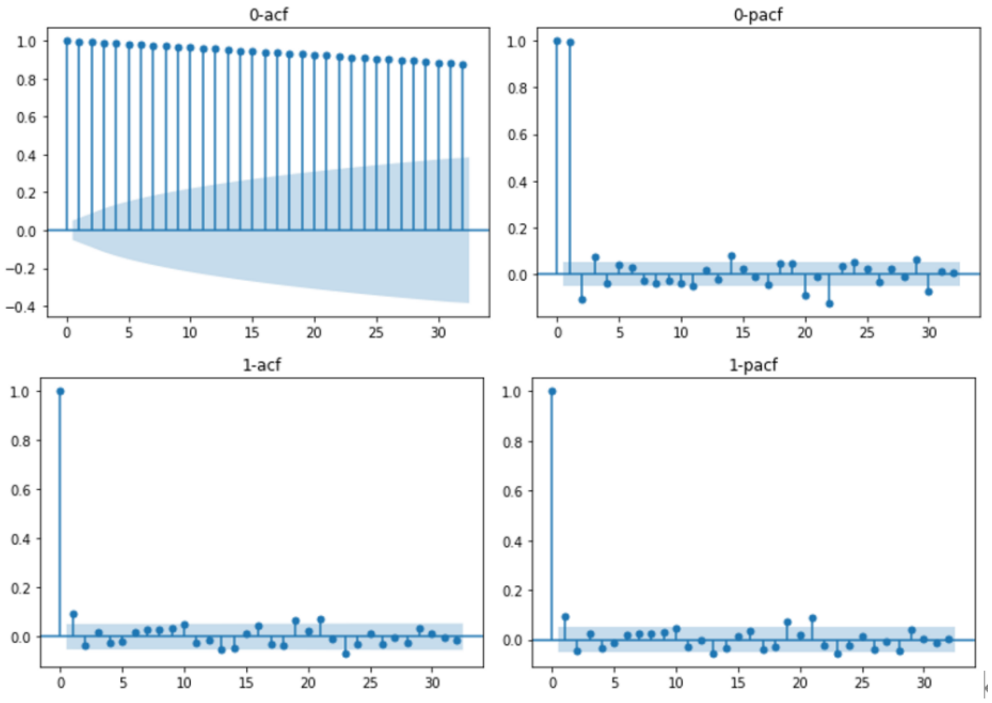

Figure 4. Time series ACF and PACF plotsE

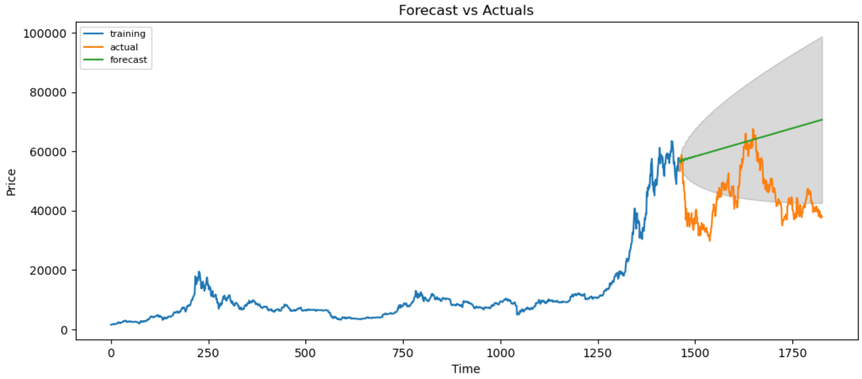

Figure 5. Time series (ARIMA) prediction

3.3 Conventional Linear Models

For conventional linear models, Ordinary Least Square (OLS), Ridge Regression and Lasso Regression are compared. OLS is the most basic algorithm for generalized linear regression, with the objective defined to minimize the test residual sum of square. However, on account of the fact that more predictors would always lead to better fitting in training set, overfitting might be witnessed in the circumstances of OLS. Hence, regularization could be applied. Lasso regression and Ridge regression use two different methods of regularization, which force the coefficient estimates to decrease, and therefore avoid overfitting and reduce variance in the model. Besides, as mentioned earlier in Section II, KDJ, CCI, RSI, ATR and WR are included as technical predictors, and Lasso regression also performs feature selection.

Linear regression, as the simplest statistic model without regularization, obtained training AIC = 11490.1, test MSE = 125.8, and R-Square score = -84.4. Ridge regression and Lasso regression add $l1$ and $l2$ regularization to avoid overfitting. Nevertheless, the results of Ridge and Lasso regression is no better than the linear regression in terms of test MSE, as indicated in the Table 3. Ridge regression has slightly better performance than Lasso, but both regularized models failed to outperform linear regression model.

Table 3 Linear model metrics

	Linear Regression	Ridge	Lasso
Training AIC	28653.2	Not Applicable	Not Applicable
Test MAE	13303.2277	13303.2026	13301.8992
Test RMSE	14696.8633	14696.8603	14696.9426
Test R-Square	-2.0013	-2.0013	-2.0014

3.4 Machine learning approaches

For machine learning models, different ensemble methods (random forest and Light GBM) are compared. Random forest is a tree ensemble method which is easy to tune and robust to parameter changes. LightGBM is another tree model that has more hyperparameters to tune and is generally deemed to have good performance. The predictors in machine learning models are the same as that in statistic models.

The predictive performance of the machine learning models is shown in Table 4. Among the two widely used machine learning models, random forest performs better with significantly lower test MAE, test RMSE and highest test R-Square, which is in line with previous studies [12].

Table 4 Machine learning model metrics

	Random Forest	LightGBM
Training AIC	Not Applicable	Not Applicable
Test MAE	11375.9496	13984.3090
Test RMSE	14576.3934	16091.5072
Test R-Square	-1.9523	-2.5980

3.5 Discussion

As indicated in the tables before that conclude prediction metrics, all models have relatively poor predictive performance on the test set. Possible reason for this phenomenon is that the price increases rapidly and is extremely volatile in the past 20 months according to price graph in Exploratory Data Analysis. Daily price data is also distorted as it loses much information as compared with tick-unit data. Therefore, previous price data might contain little information for the prediction in future. More resent and frequent data might be used instead to improve predictive performance and train-to-test ratio might be adjusted (4:1 in this study). Besides, more fundamental features such as NVT, MVRV, Mayer Multiple, and Mining Price-to-Earning Ratio might also be included, and feature selection might be conducted based on methods such as principal component analysis (PCA) or variable importance ranking.

Regarding to the data mining process, there is also room for improvements. In Ridge and Lasso regression, GridSearchCV algorithm in Python is used to perform cross-validation and tune penalizing hyperparameter, but since the function tests from a finite set of hyperparameter values, the best hyperparameter might be absent and some good model might be missed. In machine learning models, the training set might be divided into training and validation set in future investigation to further improve model performance.

Based on the analysis, cross comparison is performed to compare different types of models, which has a great reference value for researchers to choose in price mining. In addition, more cryptocurrency assets might be investigated in the future.

4 CONCLUSION

In summary, this paper investigates cryptocurrency price prediction of time series models, conventional linear models, and machine learning models based on the five-year historical daily price of BTC-USD with technical features as predictors. Various metrics including AIC, MAE, RMSE, and R-Square are computed for the models of the same split to compare the predictive performance. Among three types of models, machine learning models perform better in prediction, and the best machine learning model is random forest. Linear models with predictors as technical features have relatively higher error than machine learning models, with the best linear model being linear regression. Although regularization is meant to decrease bias and variance, regularized linear models cannot outperform linear regression. Time series models, in comparison, have even higher error.

In the future, in order to improve predictive performance, more recent and frequent data might be used, and the training-to-test ratio might be adjusted. Moreover, models with more better-established features might be tested to utilize all available information. For example, fundamental features might be included into the model, and feature selection might be conducted. Other the state-of-art models (e.g., XGBoost, LSTM) should also be evaluated, and the comparison might be applied to other cryptocurrency assets. In general, these results provide a solid guideline of different popular models in the price prediction performance of cryptocurrency, which can guide researchers to develop proper benchmark for investment.

REFERENCES

[1] L. Chaldaeva, and A. Danilin, "The history of blockchain technology and cryptocurrency market development", *Finance and Credit*, vol. 27, no. 4, pp. 740-762, 2021.

[2] P. Soni, Y. Tewari and D. Krishnan, "Machine Learning Approaches in Stock Price Prediction: A Systematic Review", *Journal of Physics: Conference Series*, vol. 2161, no. 1, p. 012065, 2022.

[3] I. Eyal, "Blockchain Technology: Transforming Libertarian Cryptocurrency Dreams to Finance and Banking Realities", *Computer*, vol. 50, no. 9, pp. 38-49, 2017.

[4] H. Jang and J. Lee, "An Empirical Study on Modeling and Prediction of Bitcoin Prices With Bayesian Neural Networks Based on Blockchain Information", *IEEE Access*, vol. 6, pp. 5427-5437, 2018.

[5] M. Saad, J. Choi, D. Nyang, J. Kim and A. Mohaisen, "Toward Characterizing Blockchain-Based Cryptocurrencies for Highly Accurate Predictions", *IEEE Systems Journal*, vol. 14, no. 1, pp. 321-332, 2020.

[6] N. Kyriazis, "A Survey on Empirical Findings about Spillovers in Cryptocurrency Markets", *Journal of Risk and Financial Management*, vol. 12, no. 4, p. 170, 2019.

[7] "Comparative Performance of Machine Learning Ensemble Algorithms for Forecasting Cryptocurrency Prices", *International Journal of Engineering*, vol. 34, no. 1, 2021.

[8] Y. Yang, Y. Wu, P. Wang and X. Jiali, "Stock Price Prediction Based on XGBoost and LightGBM", *E3S Web of Conferences*, vol. 275, p. 01040, 2021.

[9] A Primer on Money, Banking, and Gold (Peter L. Bernstein's Finance Classics). John Wiley & Sons, 2008.

[10] M. Wątorek, S. Drożdż, J. Kwapień, L. Minati, P. Oświęcimka and M. Stanuszek, "Multiscale characteristics of the emerging global cryptocurrency market", *Physics Reports*, vol. 901, pp. 1-82, 2021.

[11] R. Adhikari and R. Agrawal, An Introductory Study on Time Series Modeling and Forecasting. arXiv preprint arXiv:1302.6613, 2013.

[12] V. Dyntu and O. Dykyi, "CRYPTOCURRENCY IN THE SYSTEM OF MONEY LAUNDERING", *Baltic Journal of Economic Studies*, vol. 4, no. 5, p. 75, 2019.