Finance Research Letters 32 (2020) 101084

Contents lists available at ScienceDirect

Finance Research Letters

journal homepage: www.elsevier.com/locate/frl

A novel cryptocurrency price trend forecasting model based on LightGBM

Sun Xiaolei[a,*], Liu Mingxi[a], Sima Zeqian[b]

[a] Institutes of Science and Development, Chinese Academy of Sciences, Beijing 100190, China
[b] ZhongQingyu Big Data Technology Co., Ltd., Beijing 100020, China

ARTICLE INFO

Keywords:
Cryptocurrency
Trend forecasting
LightGBM
Forecasting performance

ABSTRACT

Forecasting cryptocurrency prices is crucial for investors. In this paper, we adopt a novel Gradient Boosting Decision Tree (GBDT) algorithm, Light Gradient Boosting Machine (LightGBM), to forecast the price trend (falling, or not falling) of cryptocurrency market. In order to utilize market information, we combine the daily data of 42 kinds of primary cryptocurrencies with key economic indicators. Results show that the robustness of the LightGBM model is better than the other methods, and the comprehensive strength of the cryptocurrencies impacts the forecasting performance. This can effectively guide investors in constructing an appropriate cryptocurrency portfolio and mitigate risks.

1. Introduction

Since Nakamoto (2008) introduced Bitcoin, cryptocurrencies have gradually become the investment target of investors all over the world. Nowadays, the cryptocurrency market contains about 2000 kinds of coins for which the volume of transactions and circulation are huge but uneven. Considering that the investment risk of cryptocurrencies is greater than other products, forecasting the fluctuation tendency of cryptocurrency price is of great importance.

At present, some scholars devote themselves into market efficiency, price volatility of cryptocurrencies centering on the Bitcoin (Bouri et al., 2017; Balcilar et al., 2017; Urquhart, 2017; Ji et al., 2018; Jiang et al., 2018). Aiming to extend earlier research efforts, Ji et al. (2018) set out a comprehensive literature review on the correlations between the most prominent cryptocurrency (i.e., Bitcoin) and a set of financial assets. Although the cryptocurrency market itself is extremely complex and risky, it still represents an alternative investment instrument with the unique characteristic of high return and a low correlation with financial assets. In particular, Bitcoin serves as a hedge, a safe haven, and a diversifier for oil price movements in terms of diversification opportunities and downside risk reductions (Selmi et al., 2018).

Existing literature recognizes Bitcoin as an investment asset (Bouri et al., 2017; Ji et al., 2018; Selmi et al., 2018). Guesmi et al. (2018) highlight in detail how the Bitcoin market allows hedging the investment risk against all other financial assets. Generally speaking, cryptocurrencies offer diversification and hedging benefits for investors by considerably reducing portfolio risk. In addition, more and more scholars are noting what driving factors determine the price fluctuation of cryptocurrencies. For example, regarding Bitcoin, Balcilar et al. (2017) reveal that volume can predict the returns, and Demir et al. (2018) examine the forecasting power of the daily economic policy uncertainty index on the daily Bitcoin returns. Additionally, Corbet et al. (2018) examine the fundamental drivers of the price and use these variables to detect and datestamp bubbles. Instructively, these studies generate a

* Corresponding author.
E-mail address: xlsun@casipm.ac.cn (X. Sun).

https://doi.org/10.1016/j.frl.2018.12.032
Received 12 November 2018; Received in revised form 14 December 2018; Accepted 26 December 2018
Available online 27 December 2018

better understanding of how market information is embedded in the fluctuation of cryptocurrency prices, which also helps increase the forecasting performance.

With the booming of the cryptocurrency market, diverse cryptocurrencies attract huge amounts of capital flow, and information can be fully transmitted among different currencies. Also, the fluctuation trends of cryptocurrencies show a certain similarity (Corbet et al., 2018). Impacted by common macroeconomic drivers, the trend of one cryptocurrency is not only determined by its own dynamics but also closely correlated with other cryptocurrencies. Remarkably, even if investors only focus on one particular cryptocurrency when forecasting or designing a portfolio, it is difficult to avoid the systematic risk of the whole cryptocurrency market. In order to utilize the market information as much as possible, this paper seeks to construct a dataset containing major currencies to forecast the price trends of cryptocurrency markets.

With regard to forecasting cryptocurrency, time-series techniques, as well as machine learning algorithms, are the most common alternative models. Yet, it is still a challenge to predict the cryptocurrency price by using traditional econometric models, which only work for time series. In this paper, we focus on the forecasting performance of machine-learning techniques when predicting the price trend. For a certain cryptocurrency, when the closing price is less than that of the previous day, the original label is recorded as 0; otherwise, the original label is recorded as 1. In this case, trend forecasting of cryptocurrencies can be transferred into a typical dichotomy problem in machine learning. In this paper, three different kinds of data mining methods, namely LightGBM, SVM (Support Vector Machines) and RF (Random Forests), were utilized to implement the trend forecasting of cryptocurrencies. Specifically, SVM is a typical and effective method to solve classification problem, which has been widely used in several different application fields (Geng et al., 2016; Kumar and Gopal, 2009; Soualhi et al., 2015). As for LightGBM and RF, they are ensemble learning methods mainly based on the Decision Tree algorithm, and they are also frequently-used in classification tasks.

Focusing on trend forecasting, the rest of this paper is organized as follows. Section 2 describes the model specifications. Section 3 presents the data and the forecasting results. Finally, Section 4 presents our conclusions.

2. Model specifications

In this section, as a relatively new algorithm, the LightGBM algorithm is introduced in detail. LightGBM is a novel GBDT (Gradient Boosting Decision Tree) algorithm, proposed by Ke and colleagues in 2017, which has been used in many different kinds of data mining tasks, such as classification, regression and ordering (Ke et al., 2017). The LightGBM algorithm contains two novel techniques, which are the gradient-based one-side sampling and the exclusive feature bundling, respectively.

Given the supervised training set $X = \{(x_i, y_i)\}_{i=1}^{n}$, LightGBM aims to find an approximation $\hat{f}(x)$ to a certain function $f^*(x)$ that minimizes the expected value of a specific loss function $L(y, f(x))$ as follows:

$$\hat{f} = \arg\min_{f} E_{y,X} L(y, f(x)) \tag{1}$$

LightGBM integrates a number of T regression trees $\sum_{t=1}^{T} f_t(X)$ to approximate the final model, which is

$$f_T(X) = \sum_{t=1}^{T} f_t(X) \tag{2}$$

The regression trees could be expressed as $w_{q(x)}$, $q \in \{1, 2, \ldots, J\}$, where J denotes the number of leaves, $q$stands for the decision rules of the tree and w is a vector that denotes the sample weight of leaf nodes. Hence, LightGBM would be trained in an additive form at step t as follows:

$$\Gamma_t = \sum_{i=1}^{n} L(y_i, F_{t-1}(x_i) + f_t(x_i)) \tag{3}$$

In LightGBM, the objective function is rapidly approximated with Newton's method. After removing the constant term in (3) for simplicity, the formulation can be transformed as follows:

$$\Gamma_t \cong \sum_{i=1}^{n} (g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)) \tag{4}$$

where g_i and h_i denote the first- and second-order gradient statistics of the loss function. Let I_j denote the sample set of leaf j, and (4) could be transformed as follows:

$$\Gamma_t = \sum_{j=1}^{j} ((\sum_{i \in I_j} g_i) w_j + \frac{1}{2} \Big(\sum_{i \in I_j} h_i + \lambda) w_j^2 \Big)) \tag{5}$$

For a certain tree structure $q(x)$, the optimal leaf weight scores of each leaf node w_j^* and the extreme value of Γ_K could be solved as follows:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{6}$$

Table 1
The main parameters of LightGBM.

Parameters	Interpretation
num_leaves	This is the number of leaves per tree.
learning_rate	This controls the speed of iteration.
max_depth	This describes the maximum depth of the tree. It is capable of handling model overfitting.
min_data	This is the minimum number of the records a leaf may have. It is also used to deal with overfitting.
feature_fraction	This is the fraction of features selected randomly in each iteration for building trees.
bagging_fraction	This specifies the fraction of data to be used for each iteration and is generally used to speed up the training and avoid overfitting.

$$\Gamma_T^* = -\frac{1}{2}\sum_{j=1}^{J}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} \tag{7}$$

where Γ_T^* could be viewed as the scoring function that measures the quality of the tree structure q. Finally, the objective function after adding the split is:

$$G = \frac{1}{2}\left(\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right) \tag{8}$$

where I_L and I_R are the sample sets of the left and right branches, respectively. Unlike the traditional GBDT-based techniques, such as XGBoost and GBDT, Light GBM would grow the tree vertically whereas other algorithms grow trees horizontally, which makes LightGBM an effective method in processing large-scale data and features.

Generally, the forecasting accuracy would be significantly influenced by the hyper-parameters. Thus, before using LightGBM, we should first determine the number and the range of variation of its hyper-parameter. In this paper, Python 3.7 was used to perform all the computation. The main parameters of LightGBM are shown in Table 1.

3. Data and results

Based on trading volume, market ranking, circulation, and data availability, 42 kinds of primary cryptocurrencies were selected for our research to represent the market. The daily trading data from January 1, 2018, to June 30, 2018, were collected from https://www.investing.com/. Existing research generally considers that Bitcoin price volatility is correlated to some macroeconomic variables such as stock index, exchange rate, and oil price and so on. This paper selected the Dow Jones 30 index, the US S&P 500 index, the Hang Seng index, the U.S. dollar index futures, the Shanghai Stock Composite Index, the Shenzhen Component Index, the RMB to U.S. dollar exchange rate, the FTSE China A50 and the WTI crude oil futures as the economic variables (characteristics) that could affect the price fluctuation of cryptocurrency market.

After deleting holiday transaction data, the original data were naturally logarithmically processed, and the changed variables were taken as the input variables in forecasting models. After dealing with the closing prices of 42 selected cryptocurrencies, 4873 samples were ultimately obtained. Next, correlation analysis was implemented between the log values of cryptocurrencies prices and selected key economic indicators. Finally, 40 feature variables were obtained with the threshold of 0.04, as shown in Table 2.

As mentioned above, the labels mask the trends of cryptocurrency markets and the trends are divided into two kinds: falling (0) or not falling (1). In order to compare the models' forecasting accuracy under different forecasting periods, the correspondence between original features and labels are reconstructed by adding the lag time, which are 2 days, 2 weeks and 2 months, respectively. The training set, validation set, and test set were categorized according to the ratios of 50%, 30%, and 20%. Specifically, when the forecasting period is 2 days or 2 weeks, the training set is classified into two categories. In the first category, the test set is the true subset of the training set, whereas, in the second category, the test set does not belong to the training set.[1] The former is used to test the pros and cons of the model classification performance itself, while the latter is used to verify the forecasting accuracy. Tables 3 and 4 show the classification performance measured by the AUC indicator,[2] and the forecasting accuracy can be calculated thus:

$$\text{Accuracy} = \frac{\text{number of correct falling prediction} + \text{number of correct not} - \text{falling prediction}}{\text{total number of the sample}}$$

Obviously, there are many hyper-parameters in both three algorithms. For LightGBM, the most important hyper-parameters in the whole selection and optimization process are 'feature_fraction' and 'bagging_fraction', which largely determines the randomness of the model. In this paper, the Grid Search method, which is a widely used parameter optimization algorithm, was utilized to determine

[1] It is similar to the in-sample prediction and out-of-sample prediction in traditional time series prediction. Whether test samples are independent or not, the prediction accuracy of the model from different perspectives could be obtained.

[2] ROC curve (receiver operating characteristic curve) is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. AUC is the area under the ROC curve, which is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. The closer the value of AUC is to 1, the higher the prediction accuracy.

Table 2
Identified feature set.

Feature	Feature	Feature	Feature
The closing price of S&P 500 Index	The closing price of Dow Jones Index	The closing price of Dollar Index	The closing price of the Shanghai Composite Index
The opening price of S&P 500 Index	The opening price of Dow Jones Index	The opening price of Dollar Index	The opening price of the Shanghai Composite Index
The highest price of the S&P 500 Index	The highest price of Dow Jones Index	The highest price of Dollar Index	The highest price of the Shanghai Composite Index
The lowest price of the S&P 500 Index	The lowest price of Dow Jones Index	The lowest price of Dollar Index	The lowest price of the Shanghai Composite Index
The trading volume of S&P 500 Index	The trading volume of Dow Jones Index		Label of the Shanghai Composite Index
	Label of Dow Jones Index	Label of RMB to U.S. dollar exchange rate	
The closing price of the Hang Seng Index			The closing price of the Shenzhen Component Index
The opening price of the Hang Seng Index	The opening price of WTI crude oil futures	The closing price of the FTSE China A50 Index	The opening price of the Shenzhen Component Index
The highest price of the Hang Seng Index	The highest price of WTI crude oil futures	The opening price of the FTSE China A50 Index	The highest price of the Shenzhen Component Index
The lowest price of the Hang Seng Index	Label of WTI crude oil futures	The highest price of the FTSE China A50 Index	The lowest price of the Shenzhen Component Index
The trading volume of the Hang Seng Index		The lowest price of the FTSE China A50 Index	
Label of Hang Seng Index	Fluctuation of cryptocurrencies	Label of FTSE China A50 Index	

Table 3
Forecasting performance in the first category of training sets.

	AUC	AUC	AUC	Period (Accuracy)			1st day of the period (Accuracy)		
Forecasting period	2-month	2-week	2-day	2-month	2-week	2-day	2-month	2-week	2-day
LightGBM model	0.963	0.981	0.987	0.776	0.881	0.762	0.762	0.905	0.548
SVM model	0.830	0.889	0.953	0.853	0.893	0.952	0.762	0.952	0.929
RF model	0.966	0.976	0.989	0.971	0.981	0.988	0.952	0.952	0.976

Table 4
Forecasting performance in the second category of training sets.

	AUC	AUC	Period (Accuracy)		1st day of the period (Accuracy)	
Forecasting period	2-week	2-day	2-week	2-day	2-week	2-day
LightGBM model	0.50	0.97	0.607	0.476	0.952	0.93
SVM model	0.50	0.50	0.607	0.476	0.952	0.93
RF model	0.51	0.60	0.514	0.619	0.952	0.31

the best combination of the above two hyper-parameters in LightGBM. In particular, the range of 'feature_fraction' is set to (0.1, 1) with the step length of 0.1. The range of 'bagging_fraction' is also set to (0.1, 1) with the step length of 0.1. All the other hyper-parameters are set to the default values of the algorithm. Then, the optimal hyper-parameters would be used in the following training and testing processes. As for SVM, the Gaussian kernel function and Grid Search parameter optimization method were utilized to implement the prediction process. Traditionally, the regularization factor 'c' and the parameter 'g' of Gaussian kernel function are two main hyper-parameters need to be optimized. Similarly, the most important hyper-parameters in RF are the number of decision trees and the number of candidate features, which would also be optimized using the Grid Search method.

As shown in Table 3, when the test sets belong to the training sets, the forecasting performance of the LightGBM, SVM and RF models are better under the 2-week forecasting period than that under other periods. When the test sets are independent of the training sets, the forecasting performance of these three models is still better in the 2-week period, as shown in Table 4. The LightGBM and RF exhibit a better forecasting performance with their own advantages.

As shown in Table 4, the LightGBM model shows better results when using the second category of training sets. As the sample size increases, its advantages will become more and more obvious.

Taking the top 10 cryptocurrencies[3] as a subsample, as shown in Fig. 1, the forecasting performance is better than that of total 42 cryptocurrencies in the first category of training sets, while there is no significant difference between the top 10 and all 42

[3] Bitcoin, Ethereum, XRP, Bitcoin cash, EOS, Stellar lumens, Litecoin, Monero, Cardano, IOTA.

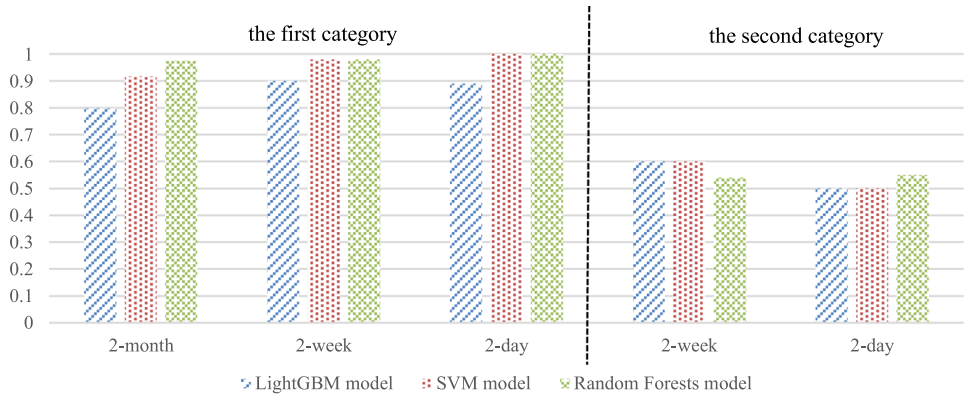

Fig. 1. The period performance with a subsample of top 10 cryptocurrencies.

cryptocurrencies in the second category. In short, the forecasting performance is correlated to the forecasting period and also to the competitiveness of the cryptocurrency. This also suggests that when investing in cryptocurrencies, we must fully consider the comprehensive strength and investment cycle of the cryptocurrencies, especially the medium term in a 2-week period.

Another major reason for the relatively high prediction accuracy is that price information for all the other cryptocurrencies is used in the forecasting models. That is, not only the fluctuation pattern of a certain kind of cryptocurrency but also the fluctuation patterns of many other cryptocurrencies are utilized simultaneously. Obviously, as a highly liquid market, the interactions and capital flows between different currencies are very important, so this could effectively improve the predictive ability of the models.

4. Conclusions

In order to utilize the market information as much as possible, this paper seeks to construct a dataset containing major cryptocurrencies for a period of January 1, 2018, to June 30, 2018. Along with the feature variables, a novel GBDT algorithm, LightGBM, is adopted to forecast the price trend (falling, or not falling) of cryptocurrency market. Some conclusions have been obtained, which could effectively guide investors in constructing an appropriate cryptocurrency portfolio.

We found that the adopted methods are more suitable for medium-term (2 weeks) prediction, and for a given cryptocurrency, the higher its comprehensive strength, the better the forecasting performance obtained. Comparatively, the LightGBM model outperforms SVM and RF in robustness, which would make it an effective forecasting model when managing a large number of data instances and a large number of features simultaneously.

In the future, we will select more potential influencing factors and develop novel forecasting models by introducing econometric models and deep learning algorithms. Also, the forecasting target will be expanded from fluctuation trend to price level with proposed forecasting models, which can be applied to the forecasting problem for large-sample and multi-feature datasets.

Acknowledgments

This work was supported by National Natural Science Foundation of China (nos. 71771206 and 71425002), President's Youth Foundation of the Institutes of Science and Development, Chinese Academy of Sciences (CAS) (no. Y7X111Q01), and funded by CPSF-CAS Joint Foundation for Excellent Postdoctoral Fellows (no. 2016LH0004).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.frl.2018.12.032.

References

Balcilar, M., Bouri, E., Gupta, R., Roubaud, D., 2017. Can volume predict Bitcoin returns and volatility? A quantiles-based approach. Econ. Model. 64, 74–81.
Bouri, E., Gupta, R., Tiwari, A.K., Roubaud, D., 2017. Does Bitcoin hedge global uncertainty? Evidence from wavelet-based quantile-in-quantile regressions. Finance Res. Lett. 23, 87–95.
Corbet, S., Lucey, B., Yarovaya, L., 2018. Datestamping the Bitcoin and Ethereum bubbles. Finance Res. Lett. 26, 81–88.
Demir, E., Gozgor, G., Lau, C.K.M., Vigne, S.A., 2018. Does economic policy uncertainty predict the Bitcoin returns? An empirical investigation. Finance Res. Lett. 26, 145–149.
Geng, Y., Chen, J., Fu, R., Bao, G., Pahlavan, K., 2016. Enlighten wearable physiological monitoring systems: on-body rf characteristics based human motion classification using a support vector machine. IEEE Trans. Mob. Comput. 15 (3), 656–671.
Guesmi, K., Saadi, S., Abid, I., Ftiti, Z., 2018. Portfolio diversification with virtual currency: evidence from bitcoin. Int. Rev. Financ. Anal in press.
Ji, Q., Bouri, E., Gupta, R., Roubaud, D., 2018. Network causality structures among Bitcoin and other financial assets: a directed acyclic graph approach. Q. Rev. Econ. Financ in press.
Jiang, Y., Nie, H., Ruan, W., 2018. Time-varying long-term memory in Bitcoin market. Finance Res. Lett. 25, 280–284.
Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. LightGBM: a highly efficient gradient boosting decision tree. Adv. Neur. Inf. Process.

Sys. 30, 3146–3154.
Kumar, M.A., Gopal, M., 2009. Least squares twin support vector machines for pattern classification. Expert Syst. Appl. 36 (4), 7535–7543.
Nakamoto, S., 2008. Bitcoin: a peer-to-peer electronic cash system. https://bitcoin.org/bitcoin.pdf.
Selmi, R., Mensi, W., Hammoudeh, S., Bouoiyour, J., 2018. Is Bitcoin a hedge, a safe haven or a diversifier for oil price movements? A comparison with gold. Energy Econ. 74, 787–801.
Soualhi, A., Medjaher, K., Zerhouni, N., 2015. Bearing health monitoring based on Hilbert–Huang transform, support vector machine, and regression. IEEE Trans. Instrum. Meas. 64 (1), 52–62.
Urquhart, A., 2017. Price clustering in Bitcoin. Econ. Lett. 159, 145–148.