Predictive Analysis of Cryptocurrency Residual Log-returns Based on Lightgbm

Zenan Ji[1, a], Yihong Zhang[2]

[1]School of Data Science, Zhejiang University of Finance & Economics, Hangzhou, 310018, China
[2]School of Science, North China Institute of Science & Technology, Langfang, 065201, China
[a]jizenan0606@163.com.

Abstract: *Cryptocurrencies have taken an increasingly important place in the international financial markets. However, due to the complexity of the factors affecting cryptocurrencies and the high volatility of cryptocurrency prices, predicting the future trend of cryptocurrencies has been a difficult topic. In this paper, we establish a prediction model based on LightGBM, aiming to find the optimal machine learning algorithm to predict different cryptocurrencies. Firstly, we preprocessed the data to ensure data integrity and availability. Secondly, we trained the training set using the LighGBM algorithm. Then, to find the optimal LightGBM algorithm for different cryptocurrencies, we optimized the algorithm using Grid Search. Finally, we evaluated the results by Root Mean Squared Error (RMSE). It shows that the Root Mean Squared Error of the prediction results for all cryptocurrencies is below 8%, which means LightGBM can be a good choice for monitoring cryptocurrencies.*

Keywords: *Cryptocurrency; Residual log-returns; LightGBM; Grid Search*

1. Introduction

In the digital age, cryptocurrencies have become one of the most important financial assets. As of 2021, there are even thousands kind of cryptocurrencies in the financial market. Cryptocurrency relies on its decentralized, de-trusted and collectively maintained distributed ledger technology to grow rapidly in the financial market. It brings a technological breakthrough in human social life, and is also affecting the transaction behavior and market rules of the financial industry. [1] As an unconventional currency, the emergence of cryptocurrencies is mysterious and dangerous. Cryptocurrencies are known to be highly volatile, which may lead to huge gains, but also may bring financial risks that are difficult to prevent. [2]

Unlike ordinary financial products, there are various factors that affect the price of cryptocurrencies, making price prediction a complex and technically challenging task. The trading of cryptocurrencies is similar to the trading of stocks, and although a part of machine learning models are maturely applied in the field of stock price prediction research, the research on cryptocurrency price prediction is still in its early stage.[3]

Previously, S. Saadah and A. A. Ahmad Whafa[4] predicted bitcoin,Ethereum and XRP using KNN, SVM and LSTM. They found that the prediction accuracy of LSTM is around 80%, which means that LSTM is suitable for cryptocurrency worthy algorithms. Gullapalli S [5] designed and implemented TDNNs and RNNs using the NeuroSolutions artificial neural network (ANN) development environment to build predictive models for Bitcoin. Sovbetov Y [6] studied and empirically demonstrated the factors that influence the price of cryptocurrencies. Some scholars like Valencia F, Gómez-Espinosa A, Valdés-Aguirre B [7] and A. Inamdar, A. Bhagtani, S. Bhatt, P. M. Shetty [8] have also gotten good results using sentiment analysis to predict the value of cryptocurrencies.

In this paper, we constructed a predictive model based on LightGBM and performed an empirical analysis using data from 14 different cryptocurrencies.

2. Data processing

Data for cryptocurrencies are obtained from the contest of ‘G-Research Crypto Forecasting’ on the official Kaggle website. This dataset includes all transaction data for 14 different Cryptocurrencies (such as Bitcoin, Eos. IO and Ethereum) from January 1, 2018 to September 21, 2021 on a minute to minute basis. It has a total of 24236806 data records consisting of Total number of trades, opening price, highest price, lowest price, closing price, transaction volume, VWAP (the average price of the asset over the time interval, weighted by volume) for each time-step and the target of residual log-returns for the asset over a 15 minute horizon.

However, it was found that 3% of the transaction data had missing target values. Due to the small number of missing data and the fact that our training and prediction are not based on the time series model, we chose to delete the transaction records with missing values.

Then, we separate the data of different cryptocurrencies from the dataset for subsequent training and prediction of the LGBM model. Figure 15 shows the correlation between different cryptocurrencies, in which we can discover that there is some correlation between many cryptocurrencies.

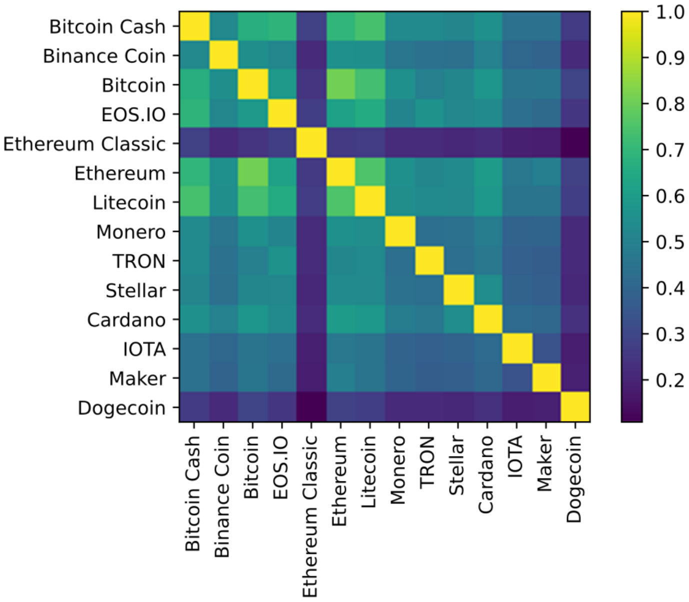

Figure 1: Close price change of Bitcon

3. Methodology

3.1. LightGBM

The LightGBM algorithm is an improved algorithm for gradient boosting decision trees. The most important thing for tree models is to identify the optimal splitting points of features. LightGBM uses histogram algorithm to find the optimal splitting points, which greatly improves the training efficiency.

The histogram algorithm works by converting continuous feature values into a histogram, which requires bucketing the feature values at first, i.e., converting them into N integers, and then obtaining a histogram of width N. In the process of algorithm implementation, the values obtained after bucketing are used as indexes for data traversal, and during the traversal process, the histogram then accumulates the required data and finally identifies the optimal separation points.

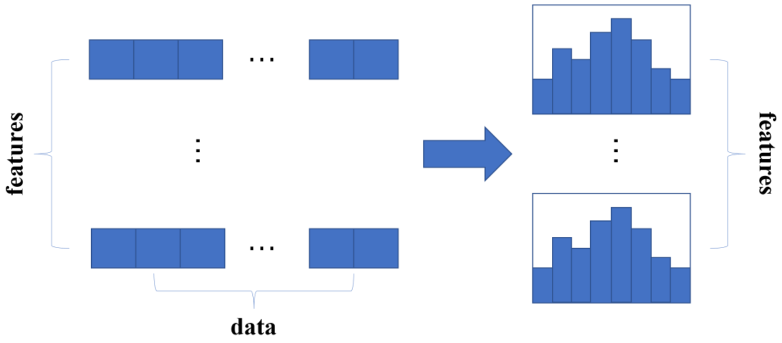

Figure 2: Histogram algorithm

3.2. The model based on LightGBM

In order to predict the residual log-returns of cryptocurrencies, the following model is built in this paper based on the LightGBM algorithm.

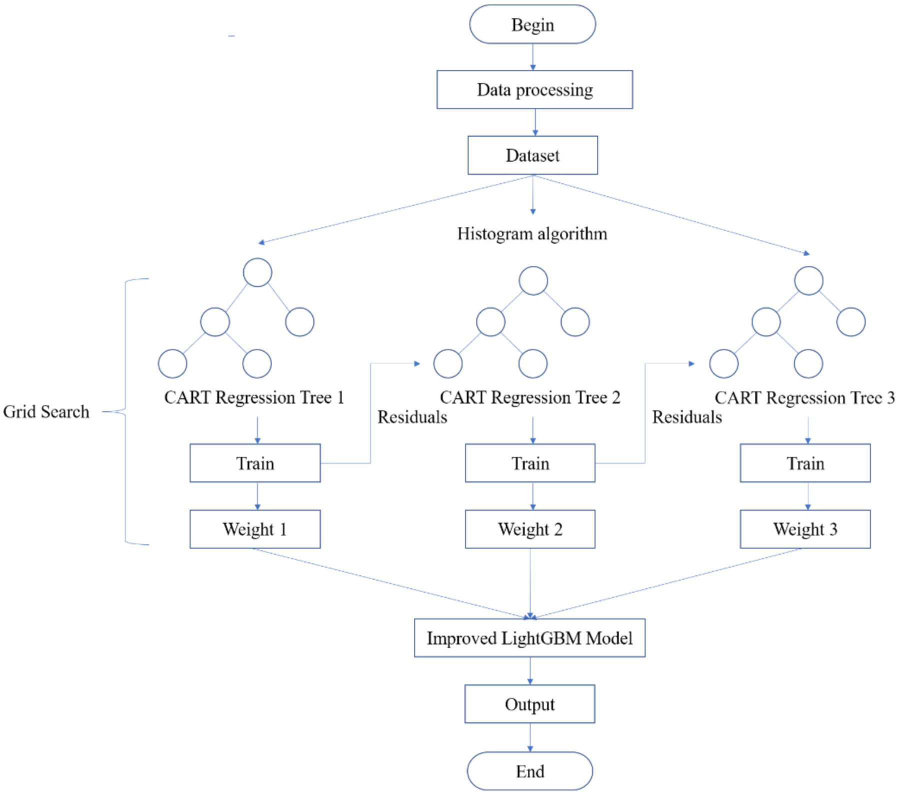

Figure 3: The model based on LightGBM

Firstly, process the data and input the processed dataset of Cryptocurrencies transactions. **Secondly**, use the histogram algorithm to find the optimal splitting points of features, and also use Leaf-wise leaf growth strategy with depth limitation to generate CART regression tree. **Thirdly**, calculate the residuals of the first CART regression tree, and the residuals from the previous round are used as training samples for the next CART regression tree to continuously fit the residuals. **Then**, the CART regression trees generated in each round of training are weighted and summed to obtain the final model. **Finally**, use the improved LightGBM Model to forecast the cryptocurrencies and get the final forecast results.

4. Model improvement based on Grid Search

4.1. Grid Search

In machine learning, each model algorithm usually has a number of different parameters, each with a different meaning. If the parameters are not set properly, the algorithm may be over-fitted or under-fitted. Therefore, in order to improve the model performance, we need to tune the parameters of the model. There are many methods to adjust the parameters, in this paper we choose Gird Search to find the optimal parameters.

The main idea of the grid search algorithm is to combine all the possible values of the parameters to obtain a ‘grid’ of all the parameter combinations. Then, all the parameter combinations are traversed for model training, and the model is evaluated by cross-validation. The parameter combination with the best model effect is the optimal parameter.

4.2. Parameters of the LightGBM algorithm

The LightGBM algorithm is an improved algorithm for gradient boosting decision trees, which is still essentially a tree model. Therefore, it involves some parameters of the decision tree algorithm in addition to the parameters of the algorithm itself. Table 1 lists the common parameters of the LightGBM algorithm and their meanings.

Table 1: Common parameters and meanings of LightGBM model

parameter	meaning
learning_rate	learning rate
boosting	boosting algorithm type
n_estimators	number of basic leaner
max_depth	maximum depth of the tree
num_leaves	<2^(max_depth)
min_data_in_deaf	the minimum number of records a leaf may have
feature_fraction	randomly selected ratio of parameters
bagging_fraction	proportion of the selected sample
reg_lambda	L2 regularization factor

Since the LightGBM algorithm involves many parameters and some of them affect each other, the choice of Grid Search algorithm to select the optimal combination of parameters is more accurately.

5. Results

In this paper, the data of different cryptocurrencies are divided according to a 4:1 ratio,which means the training set accounts for 80% of the total data set and the test set accounts for 20% of the total data set. Then, we use LightGBM to train the training set and finally find the optimal LightGBM model for different cryptocurrencies by Grid Search. After constructing the LightGBM models corresponding to different cryptocurrencies, we input the test set data into the LightGBM models and obtained the prediction results.

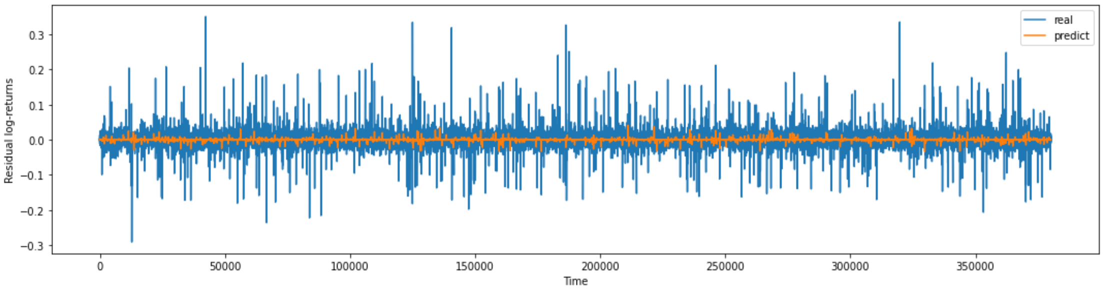

Figure 4: LightGBM model prediction results of Bitcoin

To further observe the predictive effectiveness of the LightGBM model, we evaluated the RMSE (Root Mean Squared Error) of the model on the test set:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2} \tag{1}$$

Table 2: RMSE of different cryptocurrencies on the test set

cryptocurrency	RMSE
Bitcoin Cash	0.0529
Binance Coin	0.0557
Bitcoin	0.0346
EOS.IO	0.0520
Ethereum Classic	0.0671
Ethereum	0.0400
Litecoin	0.0480
Monero	0.0640
TRON	0.0529
Stellar	0.0557
Cardano	0.0539
IOTA	0.0714
Maker	0.0616
Dogecoin	0.0656

As can be seen from the above table, the RMSE of the LightGBM model stays between 0.0346 and 0.0714, which indicates that the prediction accuracy of the LightGBM model is pretty high.

6. Conclusion

In this paper, we use the LightGBM algorithm to build a model for predicting the residual log-returns of virtual currencies. lightGBM uses a histogram algorithm to accelerate the training of the model and reduce memory usage. After training with 14 different cryptocurrencies, we can see that the predictions are generally very good even though there are some unavoidable errors. In particular, the RMSE of the algorithm is all controlled within the 8% range.

To conclude all, LightGBM can be a good choice for monitoring cryptocurrencies in a digital age where cryptocurrencies are becoming more and more popular. In future work, we should try more machine learning algorithms for prediction and analysis of cryptocurrencies, and we can combine machine learning algorithms with finance to better maintain the financial market.

References

[1] Guo Yan,Wang Lirong,Han Yan. Blockchain technology in financial market:scenario application and value outlook [J]. Technology Economics, 2017, 36(07): 110-116.
[2] Li Kangzhen, Wang Hao. Research on the use of digital cryptocurrency crimes and its detection and prevention revelation [J]. Journal of Hunan Police Academy, 2018, 30(03): 57-66.
[3] Zeng Fancheng. Research on cryptocurrency price prediction based on deep learning method and quantile regression [D]. Jiangxi University of Finance and Economics, 2021.
[4] S. Saadah and A. A. Ahmad Whafa, "Monitoring Financial Stability Based on Prediction of Cryptocurrencies Price Using Intelligent Algorithm," 2020 International Conference on Data Science and Its Applications (ICoDSA), 2020, pp. 1-10, doi: 10.1109/ICoDSA50139.2020.9212968.
[5] Gullapalli S. Learning to predict cryptocurrency price using artificial neural network models of time series [J]. 2018.
[6] Sovbetov Y. Factors influencing cryptocurrency prices: Evidence from bitcoin, ethereum, dash, litcoin, and monero [J]. Journal of Economics and Financial Analysis, 2018, 2(2): 1-27.
[7] Valencia F, Gómez-Espinosa A, Valdés-Aguirre B. Price movement prediction of cryptocurrencies using sentiment analysis and machine learning [J]. Entropy, 2019, 21(6): 589.
[8] A. Inamdar, A. Bhagtani, S. Bhatt and P. M. Shetty, "Predicting Cryptocurrency Value using Sentiment Analysis," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019, pp. 932-934, doi: 10.1109/ICCS45141.2019.9065838.