2022 3rd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE) | 978-1-6654-5160-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICBAIE56435.2022.9985937

Crypto Profit Forecasting using hybrid Machine Learning method

Qi Zhong[1],
Tongji University,
Shanghai, China,
1850425@tongji.edu.cn,

Mingda Huo[1],
Jinan University,
Guangzhou, China,
md.huo@outlook.com,

***Abstract* - The resulting digital cryptocurrency has drawn investors' attention with its new architectural design and soaring prices. Over $40 billion worth of cryptocurrencies are traded every day. Quantitative investment is a new type of investment method. It belongs to the cross-integration of finance, mathematics, statistics and computer science. In our paper, we hybrid Catboost and XGboost for Crypto Forecasting. The dataset is provided by Kaggle platform. To compare our model's performance, we do compared experiments. From the compared experiments, our model owns the highest RMSPE: 0.0359. In contrast, the other methods like Gradient Boosting, XGBoost and CatBoost's accuracy are 0.0312, 0.0349 and 0.0355 respectively.**

Keywords: Index Terms— Crypto Forecasting, Catboost, XGboost, RMSPE

I. INTRODUCTION

The continuous development of Internet technology not only deeply affects people's way of life, but also promotes the vigorous development of the Internet financial industry. The resulting digital cryptocurrency has drawn investors' attention with its new architectural design and soaring prices. Over $40 billion worth of cryptocurrencies are traded every day. They are one of the most popular speculative and investment assets, but have proven to be quite volatile. Rapidly fluctuating prices have made millionaires of lucky few and led to huge losses for others. Researchers hope to predict the trend of Bitcoin in advance through quantitative methods.

Quantitative investment is a new type of investment method. It belongs to the cross-integration of finance, mathematics, statistics and computer science. It uses mathematical models to replace human subjective judgments and uses computer technology to find profit opportunities from a large amount of data. Then, the trading order is executed according to the pre-determined strategy. Because quantitative investment has the characteristics of strong stability, high efficiency and strong discipline, more and more investors are devoted to quantitative trading or combine it with subjective trading. Common quantitative investment strategies are multi-factor stock selection strategies, arbitrage strategies and CTA strategies. Gradually applying it to the field of quantitative investment, how to construct a more effective algorithm model, improve the alpha has become a core topic in the field of quantitative investment.

In our paper, we hybrid Catboost and XGboost for Crypto Forecasting. The dataset is provided by Kaggle platform. Related work is described in section II, and we introduce our methodology and experiment in section III and IV.

II. RELATED WORK

The efficient market hypothesis[1] was put forward in 1970 by Fama, who pointed out that in order for a market to be efficient, the price of the stock in the market can fully reflect the whole information of the market But in practice, not everyone is always maintain absolute rationality, and the market information is fully reflected in the price it is difficult to achieve, so the price is not completely random walk process, now the world is not the strength of the effective market, this is the article to be able to use a model to predict the direction of theoretical basis. Crypto can be seen as a kind of stock. Therefore, the method for predicting stock can be used in a similar way in predicting crypto.

Most financial time series are nonlinear, unstable and noisy. It is always a very challenging work to predict such time series in the field of financial analysis. With the discovery and development of many machine learning models, it has been found that machine learning model itself is an efficient nonlinear modeling method, which is very consistent with the corresponding characteristics of financial time series. Therefore, many scholars began to devote themselves to the research on the application of relevant algorithm models in financial asset price forecasting. At present, using machine learning model to predict financial time series has become a very common means.

Cao, Tay (2003) [2] used SVM model to predict SP500 index, and evaluated the prediction effect through the prediction accuracy index, indicating that SVM model is very effective in the analysis of stock index price prediction. Chang (2011) [3] predicted the stock price through two machine learning algorithms: hybrid neural network and decision tree. It is concluded that the hybrid algorithm is better than using neural network or decision tree alone. Selvin (2017) [4] used a variety of neural network algorithm models such as LSTM to predict the stock price of the Indian stock market, and found that several models can achieve good prediction results.

- Our Contribution

✓ We hybrid Catboost and XGboost model as to predict the crypto.

✓ According to the dataset, we mine factors which has high Information as the input of model.

✓ We do compared experiments to approve our model's performance.

III. METHODOLOGY

- XGBoost

In order to obtain a stable model, the method of ensemble learning is often used. According to the generation mechanism of weak learners, two kinds of methods can be obtained: one is sequence integration, such as AdaBoost, etc. the weak learners used in this method are generated in series in sequence. There is a strong dependency between the weak learners of the sequence integration method, which is reflected in the fact that higher weights will be given to mark the wrong results generated by the weak learners in the training process, which will be optimized in the subsequent training to improve the overall prediction effect. The other is parallel integration, such as random forest. This method uses a weak learner that generates parallel processing at the same time. The weak learners of the parallel integration method work independently of each other. By averaging the output results of each weak learner, the errors of a single learner can be significantly reduced.

Bousting algorithm is a sequence integration algorithm with a wide range of meanings. At present, the mainstream boosting algorithms include AdaBoost, GBDT and XGBoost. Xgboost algorithm [5] is an optimization algorithm based on GBDT, AdaBoost and other algorithms.

The predictive model of the XGBoost algorithm can be expressed as follows:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \quad (1)$$

The objective function of XGBoost is as follows, which is composed of two parts: its own loss function and the regularization penalty term.

$$\mathrm{Obj(t)} = \sum_{i=1}^{n} \mathrm{l}(y_i, \hat{y}_i^t) + \Omega(\mathrm{f_t}) + \mathrm{Contast} \quad (2)$$

$$\text{Here } \Omega(\mathrm{f_t}) = \gamma \mathrm{T_t} + \frac{1}{2}\lambda \parallel \mathrm{w} \parallel^2 \quad (3)$$

T_t represents the number of leaf nodes of regression tree, w represents the weight of leaf node, γ and is the punishment intensity coefficient.

The XGBoost algorithm process is a weak learner per iteration, so the objective function can be further represented as:

$$Obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) + C \quad (4)$$

$\hat{y}_i^{(t-1)}$ means keep functions added in previous round, $f_t(x_i)$ means new function. C is the Contast.

We need to use Taylor Expansion Approximation of Loss.

$$Obj^{(t)} \simeq \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t) + \text{constant} \quad (5)$$

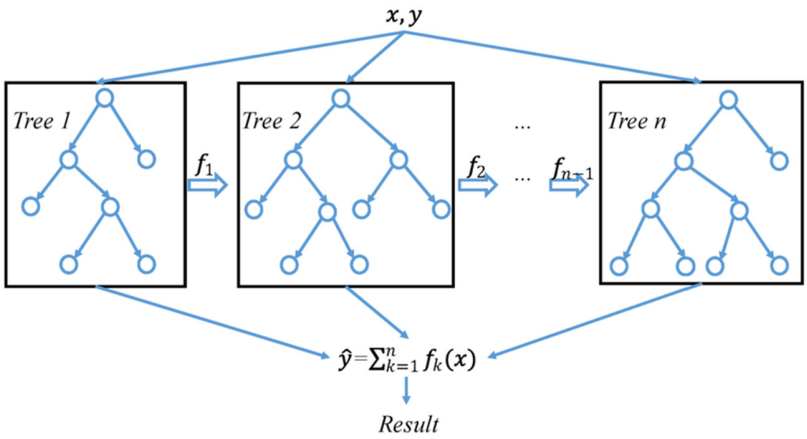

Figure 1: the structure of XGBoost

- CatBoost

CatBoost[6] is yandex's open source machine learning library in 2017. It is a kind of boosting algorithm. CatBoost, like XGBoost and LightGBM, is an improved method implemented under the framework of GBDT algorithm.[7] LightGBM effectively improves the computing efficiency of GBDT, and CatBoost performs better in algorithm accuracy than XGBoost and LightGBM.

XGBoost has certain advantages in distributed computing. However, XGBoost cannot handle category characteristics and can only accept numerical data. Therefore, before establishing a model, it is necessary to use various coding methods to convert category data into numerical data, and then establish a model. To solve this problem, CatBoost came into being. CatBoost can flexibly process categorical data. It can directly transfer the column index of categorical features. The model will automatically one-hot code it. If column index is not used, CatBoost will treat all data as numeric.

The difference between CatBoost and other machine learning algorithms lies in that the base learner of CatBoost algorithm is a symmetric decision tree, which is an

improvement of GBDT algorithm with fewer parameters, supporting category variables and high accuracy. In addition to dealing with categorical features efficiently and reasonably, CatBoost can also solve the problems of gradient deviation and prediction offset, so as to reduce the degree of over fitting and improve the generalization ability and accuracy of the learner. Of course, CatBoost also has some disadvantages: first, it requires a lot of memory and time to process category features; second, the setting of different random numbers may affect the prediction results of the model.

- Hybrid model

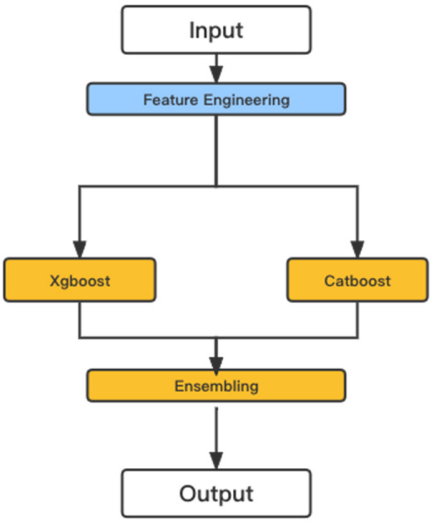

Figure 2: Hybrid Model

We combine LightGBM with CatBoost as a hybrid model, the model structure is shown in Figure 2.

IV. FACTOR MINING

- **Input feature**

Our dataset is provided by Kaggle. This dataset contains information on historic trades for several cryptoassets,

Table 1: input features

Asset_ID	An ID code for the crypto asset.
timestamp	A timestamp for the minute covered by the row.
Count	The number of trades that took place this minute.
Open	The USD price at the beginning of the minute.
High	The highest USD price during the minute.
Low	The lowest USD price during the minute.
Close	The USD price at the end of the minute.
Volume	The number of cryptoasset units traded during the minute.
VWAP	The volume weighted average price for the minute.
Target	15 minute residualized returns.

Figure 3 shows the timestamp of crypto asset.

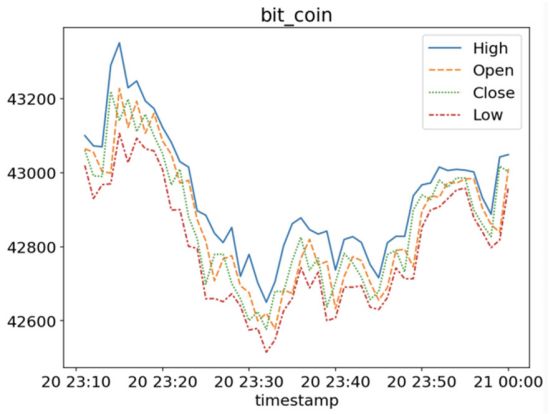

Figure 3: timestamp

- **Factor mining**

We choose eight factors as input of hybrid model.

- Upper_Shadow=high-max(close,open)
- lower_shadow=min(close,open)-low
- Close/Open
- Close-Open
- High-Low
- High/Low
- Mean=mean(Open, High, Low, Close)
- Median=median(Open, High, Low, Close)

V. EXPERIMENTS

- **Experimental setting**

Table 2: parameters of XGBoost

learning_rate	0.01
boosting_type	gbdt
Max_depth	8
Optimizer	Adam

Table 3: parameters of CatBoost

learning_rate	0.01
boosting_type	gbdt
Max_depth	8
Optimizer	Adam

- **Experimental metrics**

In our task, we use Pearson Correlation Coeffic[8] as the metrics.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{7}$$

Where:

$\text{cov}(X, Y)$ is the covariance of X and Y

σ_X is the standard deviation of X.

σ_Y is the standard deviation of Y.

- **Experimental result**

To compare our model's performance, we do compared experiments. Table 4 shows the compared experiment results. The less loss is, the better performance the model will owns.

Table 4: Compared Experiment Results

Models	RMSPE
Gradient Boosting	0.0312
Xgboost	0.0349
Catboost	0.0355
Proposed Model	**0.0359**

Where

$$\text{RMSPE} = \frac{1}{n}\sqrt{\sum_{i=1}^{n}\left(\frac{y_i - y_i'}{y_i}\right)^2} \qquad (8)$$

From the compared experiments, our model owns the highest RMSPE: 0.0359. In contrast, the other methods like Gradient Boosting, XGBoost and CatBoost's accuracy are 0.0312, 0.0349 and 0.0355 respectively.

VI. CONCLUSION

In our paper, we hybrid Catboost and XGboost for Crypto Forecasting. In section II, we describe some learning methods on financial data show our contribution. In section III, the methodology is introduced by us. The factor mining is presented in section IV and experiment's detail is allocated in section V. We will improve our model's performance in the future.

ACKNOWLEDGEMENT

Thanks to the Kaggle platform, we could advance our research in the long period and produce this paper documenting the work. And due to Qi Zhong and Mingda Huo's encouragement, the work could finish efficiently.

REFERENCES

[1] Malkiel B G. Efficient market hypothesis[M]//Finance. Palgrave Macmillan, London, 1989: 127-134.

[2] Cao LJ, Tay F E H. Support vector machine with adaptive parameters in financial time series forecasting[J]. IEEE Transactions on Neural Networks, 2003, 14(6): 1506-18.

[3] T. S. Chang. A comparative study of artivicial neural networks, and decision trees for digital game content stocks price prediction[J]. Expert Syst. 2011, 38: 14846-14851.

[4] Selvin S, Vinayakumar R, Gopalakrishnan E A, et al. Stock price prediction using LSTM, RNN and CNN-sliding window model[C]// 2017 International Conference on Advances in Computing, Communications and Informatic- s(ICACCI). IEEE, 2017.

[5] Tianqi c, Guestrin C. XGBoost: A Scalable Tree Boosting System[C]// 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.(S.1.]: ACM, 2016: 785-794.

[6] Dorogush A V, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support[J]. arXiv preprint arXiv:1810.11363, 2018.

[7] Liang W, Luo S, Zhao G, et al. Predicting hard rock pillar stability using GBDT, XGBoost, and LightGBM algorithms[J]. Mathematics, 2020, 8(5): 765.

[8] Bilgin M. Developing a cognitive flexibility scale: Validity and reliability studies[J]. Social Behavior and Personality: an international journal, 2009, 37(3): 343-353.