Using machine learning for cryptocurrency trading

Jifeng Sun
Shenzhen Environmental Science and New Energy Technology Engineering Laboratory Tsinghua-Berkeley Shenzhen Institute Tsinghua University
ShenZhen, China

Yi Zhou
Graduate School at Shenzhen Tsinghua University
ShenZhen, China

Jianwu Lin
Graduate School at Shenzhen Tsinghua University
Shenzhen,China
lin.jianwu@sz.tsinghua.edu.cn

***Abstract*—In this study, we use random forest to predict several cryptocurrencies' prices by using part of factors in Alpha101 [1] to represent features from the history of cryptocurrencies' market data on Binance and Bitfinex. The result shows our strategy with some factors from Alpha101 is effective in cryptocurrency trading.**

***Index Terms*—cryptocurrency, machine learning ,trading strategy**

I. Introduction

Cryptocurrency is a kind of digital currencies, which is a medium of exchange to ensure the security of transactions based on cryptography. Bitcoin, the most popular cryptocurrency, became the first decentralized currency in 2009. Since then, several similar cryptocurrencies have been created. Although it has only been nine years since its birth, in recent years, especially since 2017, the cryptocurrency market has been growing rapidly. The unit price of Bitcoin at the beginning of 2017 was less than $1,000. However, by the end of the year, its unit price had exceeded $13,000 and the market value had increased by more than 12 times. As of the end of December 2017, the number of cryptocurrencies with a market value of more than $1 billion has reached more than thirty. Although the cryptocurrency market suffered a severe decline in 2018, there are still thirteen cryptocurrencies with a market value of more than $1 billion. Cryptocurrency has become an important part of the global financial market that can no longer be ignored.

Cryptocurrency, as an emerging alternative asset, can be invested through a variety of strategies.

The rest of this paper is organized as follows: Section II discusses the common trading strategies for cryptocurrencies. We want to study the feasibility of to leverage traditional asset transaction analysis in predicting highly volatile cryptocurrency prices.

Section III reviews previous research in the field of cryptocurrencies, including the use of machine learning and deep learning techniques for bitcoin price forecasting and other types of time series forecasting in financial markets.

Section IV proposes an approach to the predict the cryptocurrency price. The Section V describes the implementation of the experiment, data sets, and performance metrics. Section VI gives the results in tabular and graphical format.

Finally, Section VII provides directions for future research.

II. Trading Strategy

Fundamental strategy. Investors can appraise cryptocurrencies based on fundamentals and technology, and select highquality cryptocurrencies with high potential for growth. Different from analyzing traditional investment stocks, in which the fundamental analysis of stocks focuses on profit, income, growth rate, market share, and corporate strategy, the operating companies of many cryptocurrencies only have white papers and teams. The fundamentals of the cryptocurrency must be related to the blockchain. The fact that the activity of the technical community is an important factor supporting the price of the cryptocurrency will continue to affect the price of the currency and can serve as the reference for investments.

Multifactor strategy. The multi-factor strategy is a widely used stock picking strategy. The basic idea is to find some factors that are most relevant to the rate of asset's return, and build a stock portfolio based on the factors, which be expected to make portfolio outperform in the future. The key to build a multifactor model is to find the correlation between factors and yields. In cryptocurrency trading, by analyzing market data and even social media data, investors can Multifactor strategy to predict the future trend of cryptocurrency, and trade.

Trend tracking.It is also known as the momentum trading strategy,which means buying stocks or other securities that have had high returns over the past several months, and selling those that have had poor returns over the same period.It is one of the most common quantitative investment strategies. The application of momentum investment strategy covers multiple financial markets, not only in the stock market, but also in futures, foreign exchange and other markets.

Arbitrage.The arbitrage strategy is a strategy of using the temporary inconsistency between the price and the rate of return of certain financial products to obtain income in the financial market. When this price change produces a risk-free return, it is called a risk-free arbitrage strategy. During actual investment. We can easy to find that different currencies are traded at different prices in different transactions. By buying currencies on low-priced platforms, and selling on high-priced platforms, we can carry out cross-currency arbitrage, or arbitrage between different platforms, or arbitrage on digital currency futures contracts.

978-1-5386-8500-6/19/$31.00 ©2019 IEEE

III. Related Work

Currently, research on cryptocurrency price forecasts is at an early stage. There are several related articles that study cryptocurrency transactions from different methods and perspectives.Some research that predict the price of cryptocurrencies through machine learning techniques and some people have analyzed the factors affecting the popularity of cryptocurrency. Some people have built different models to describe the trend of cryptocurrency.

A. Analysis of the factors

Stefan Hubrich [2] finds momentum, value and carry factors are constructed based on the market and issuing data of 11 cryptocurrencies. The results show that all three factors were effective. Among all, momentum factors performed best. After factor combination, higher risk adjusted return can be obtained.

Shehhi et al. [3] finds more than half of the participants believe that the currency name and logo affect the choice of using and mining a cryptocurrency. The ease of mining, community, anonymity, privacy is also one reason for people to choose which cryptocurrencies to use and to mine. Phillips RC [4] uses wavelet coherence to study co-movement between a cryptocurrency price and its related factors, for a number of examples. Ross C. Phillips shows that the medium-term positive correlation between factors extracted from the Internet and prices is significantly enhanced when the price series is foamy.

B. predicting the price trend using machine learning and deep learning

Isaac Madan [5]tells us the 10-minute data maybe more effective than the 10-second data performing on the model, which is reflected in the Accuracy - it reflects the trend more than the 10-second data, while the random forest performed better than the generalized linear models. The result is shwon at TABLE I. Phillips RC [6] builds a trading strategy based on historical social network data and epidemic modeling. The resulting trading strategy is superior to the buying and holding strategy. This study not only proves the wider use of epidemic-detecting hidden Markov models in identifying bubble-like behavior in time series, but also proves that social media data can play an important role in forecasting cryptocurrency movements.

Yecheng Yao [7] indicates that how market open is may play a key role in influencing all other parameters. In addition, the size of the data set may affect future predictions, as the results of models trained with large data sets perform better.

Anton Misnik and Sergey Krutolevich [8] find that increasing the market data input can improve the accuracy of the prediction. And the Long short Memory network was slightly more accurate than the multi-layer Perceptron(MLP) network in this paper. The Long short Memory network(LSTM) which is a long-term and short-term memory network suitable for the processing and prediction of important events with relatively long intervals and long delays in time series. MLP, is a forward-structured artificial neural network, which can be used to fit complex functions or solve classification problems.

TABLE I
COMPARISONS BETWEEN TIME INTERVALS

STATISTIC	10 SECOND GLM	10 MINUTE GLM	10 MINUTE RANDOM FOREST
Sensitivity(TPR)	0.5429	0.534	0.540
Precision(TNR)	0.574	0.551	0.581
Precision(PPV)	0.574	0.551	0.581
Accuracy(ACC)	0.085	0.539	0.574

IV. Methods

A. Data Collection

There are many platforms for cryptocurrency transactions without a uniform data specification for each platform, which brings some trouble to data collection. This article lists some common data sources, distinguished by whether they are free or not.

TABLE II
Dataset Links

name	link	free or not
Bincentive	https://github.com/Bincentive/TradingData	free
tushare	https://tushare.pro	free
bitcoincharts	http://api.bitcoincharts.com	free
kaiko	https://www.kaiko.com/pages/historical-data	charge

For convenience, we choose the data provided by the Bincentive, which is a company who creates an intelligent mirror trading swap ecosystem based on artificial intelligence (AI), big data, social trading, personal hedge funds, and blockchain technology.

We selected the data of these two exchanges: Binance and Bitfinex. The data frequency includes 1 minute, 5 minutes, 30 minutes, one hour,and one day. The data range is from August 2017 to December 2018, including ETC-USDT, IOTA-USDT, LTC-USDT, NEO-USDT, TRX-USDT, XLM-USDT, XRP-USDT, ETH-USDT, ADA-USDT, BCH-USDT, BNB-USDT, BTC-USDT.In order to ensure sufficient data volume for training, we selected 5-mintue frequency data to do the following experiments.

Due to the lack of data sources and special data requirements, fetch data,and build a self-built database is a task sometimes must to do. Fortunately, many exchanges provide detailed API documentation to help user access to their data.

B. Data Preprocessing

In data mining, we call the raw data obtained as “dirty data”, that is, the data will always have some noise, missing values, and inconsistent data dimensions. The existence of such ”dirty data” will directly affect our later modeling effects. Therefore, we must first preprocess the data before modeling, process the irregular data into rule data, complete the missing values, and delete the outliers. Before dealing the data, we first calculate the correct rate using the following formula:

$$\alpha = (x.high/x.open) * (x.close/x.low) * (x.high/x.low)$$

$$\beta = \begin{cases} 1, & \text{if } \alpha \geq 1 \\ 0, & \text{if } \alpha < 1 \end{cases}$$
$$\kappa = \frac{\sum_{i=1}^{l} \beta_i}{l} * 100\% \quad (1)$$

We define κ as the correct rate of the data, α and β is an intermediate variable. l is the length of the dataset. x is an open-high-low-close data.

We use the above formula to judge the correctness of the data format. This means that the market high should be higher than the market open, the market close should be higher than the market low, and the market high should be higher than the market low.

We can see from Fig. 1. that the data format in 124 documents are mostly correct, and there are only a few errors in the data in the three files. For these three files, we choose to drop them.

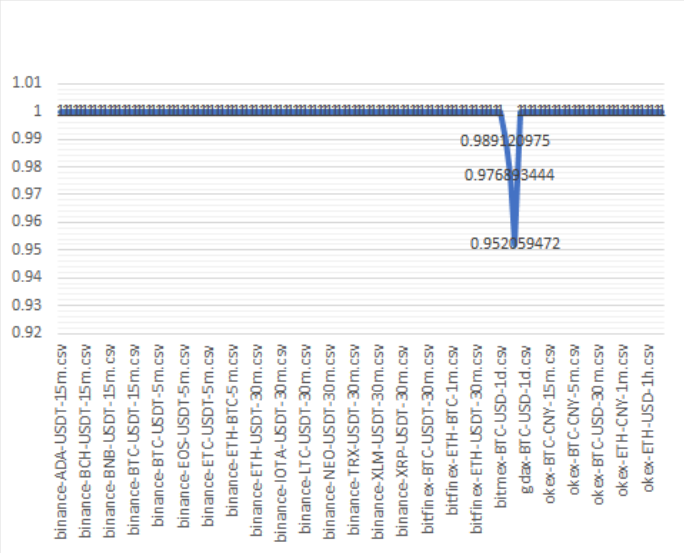

Fig. 1. Data Integrity

Missing value processing

Due to the unstable data interface provided by the exchanges, the collected data maybe has missing values. In our experiments, we choose to ignore the short missing range data and use polynomial interpolation to deal with the long missing range data, such as Lagrange interpolation and Newton interpolation.

Outlier processing

Outlier is defined as a measured value of a set of that deviates from the mean by more than two standard deviations. The measured value that deviates from the average value by more than three standard deviations is called an abnormal value of the height abnormality. In our calculation of the alpha101 factor, the data anomaly may be amplified. When processing data, the abnormal value of the height anomaly should be eliminated. We discard the factors that produce more nan and INF, and use 0 to replace the outliers generated.

standardization of data

The standardization of data is used to unify the dimensions between different factors. The difference between data with different dimensions may be very large. For example, the price of bitcoin is generally distributed between $3000 and $20000, and the volume may be distributed between 0 and 100. So when these two data are modeled together, the big gap between them will have a huge impact on the results. Therefore, in order to eliminate the difference in the dimensions between different factors, the data needs to be standardized.

There are three main methods for data standardization:

Minimum-maximum normalization, Zero-mean normalization and decimal scaling normalization.

The principle of the minimum-maximum normalization is to use the linear transformation method to project the numerical value of the original data to among 0 to 1. The conversion formula is as follows:

$$x^* = \frac{x - min}{max - min} \quad (2)$$

x is the data sample before normalization. x^* is the data sample after normalization. *max* in the formula represents the maximum value in the sample and *min* represents the minimum value in the sample. In the following formulas, the same symbol represents the same meaning

The advantage of this approach is that it eliminates the dimension of the data and maintains the relationship between the various parts of the data. The disadvantage of this method is that if a certain value in the data set is large, the value will be close to 0 after normalization.

Zero-mean normalization is also called standard deviation normalization. The processed data has a mean of 0 and a standard deviation of 1. The conversion formula is:

$$x^* = \frac{x - \mu}{\sigma} \quad (3)$$

μ is the mean of the raw data. σ is the standard deviation of the raw data, and 0-mean normalization is the most commonly used data normalization method.

Decimal scaling normalization is to map the raw value to [-1,1] by moving the decimal point of the raw value.

The conversion formula is:

$$x^* = \frac{x}{10^k} \quad (4)$$

In this paper, we use Zero-mean normalization to deal with each column of data.

The z-score standardization is applicable to the case where the maximum and minimum values of the data are unknown, and there is no case where the values in the data set deviate from the mean value, resulting in a large difference between the values after normalization. It's appropriate to use this method here.

After the above normalization, the dimensions of each part of the data are unified, the outliers are also eliminated, and the data missing problem is solved.

Then, we divided the data into training set and test set according to the ratio of 9:1.

According to the trading commissions of each exchange and the slippage that may occur during the real trading, we label The difference between the current close price and the price after n bar greater than 0.5% as class 1 and the data decline rate greater than 0.5% as class -1. The rest is labeled as class 0, n is a parameter that need to be optimized. In this paper, we test some integer in the range of 0 to 300.

We also collect commissions for common exchanges and supported cryptocurrency types and pair trading symbols, which is shown in the table below.

TABLE III
THE EXCHANGES

name	support currencies and pair trading symbols	commission
Ggtrade	more than 100 symbols	free
OKEX	91 currencies,274 symbols	0.03%
Kucoin	294 symbols	0.10%
Binance	99 currencies,235 symbols	0.10%
Bittrex	199 currencies,271 symbols	0.25%
Bitfinex	31 currencies,84 symbols	0.2%-0.1%
Huobi	48 currencies,88 symbols	0.2%

C. Feature Selection

At the end of 2015, World Quant published the paper ***"101 Formulaic Alpha"*** [1], which gives 101 real-world alphas. They emphasizes that 80% were still in use at that time. Some of the factors came from the websim platform, this part of alpha has certain logic, and the expression is usually simple; the other part is mined by genetic programming.

In this paper, we choose 16 factors in Alpha101, which using Open-High-Low-Volumes data and raw OHLCV data to do feature engineering. The rest factors in Alpha101 that require cross sectional indices are not considered. However, maybe some cross sectional indices can well describe cryptocurrencies will be found in the future not to long. In the traditional multi-factor strategy, the validity test of factors is an important step. The method is to select the target portfolio according to the single factor value sequence, and repeat several cycles of the method at a time, and then calculate the return of the asset portfolio. According to the situation of the target combination selected by each factor, the factors that can produce continuous returns are screened out as the effective factors for multi-factor stock selection. However, The validity test of factors is not carried out in this paper. This is because that the classification algorithm adopted in this paper is the random forest. Every tree created does not have to be used all factors, but randomly selected factors, and finally integrate all the trees to get the classification results. Thanks to the structure of the random forest, we can get all the importance weight of factors, which also help to study the factors affecting price changes.

D. Classification algorithm selection

According to the result of [5], we find that the random forest model performed very well on the bitcoin data, and using the random forest algorithm can help us to obtain the importance of each factor and build the trading strategy in the future. Therefore, we chose to use this model as the classification algorithm, applied to the data set.

E. Model performance

In the experiment, we set the cycle of model training according to the periodicity of the investors' actual trades.The model was retrained daily according to the date of the dataset.The final results are weighted according to the amount of data.

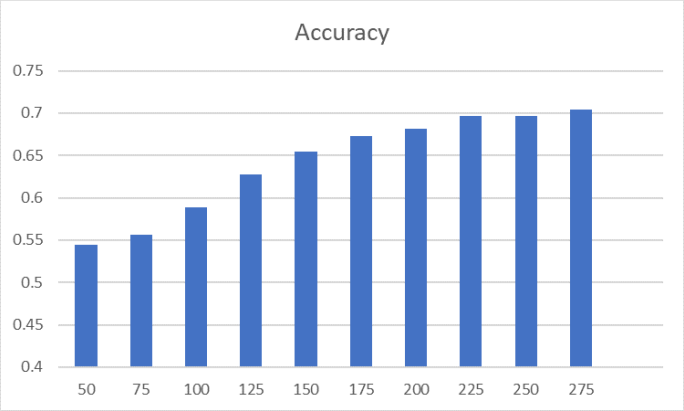

Fig. 2. Accuracy Variation

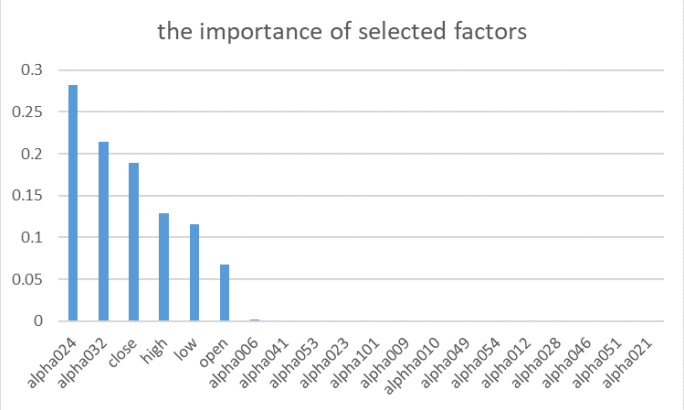

Fig. 3. factor importance

F. Result analysis

According to the results all the above,we can draw conclusions:

- The histogram in Fig.2. shows the average accuracy in all data set of different time intervals when setting the training data's label which introduced in *Data Preprocessing*. That means the model is more accurate at predicting the future farther ahead. And we can find that the accuracy

first increases and becomes stable as the time interval increases.

- The histogram in Fig.3. the alpha24 and alpha32 show important degree of predictability in the experiment. The original market data such as high, close, open also play important role in the model. It shows that factors like alpha101 are useful for discovering future price variation in cryptocurrencies.

V. Analysis of Backtesting Results

We select the most famous cryptocurrencies——bitcoin and Ethereum 's historical market data at five-minute frequency for the backtesting of the strategy. the strategy use the model which we introduce in the part IV above. In the backtesting part of the strategy, we retrain the model daily in the backtesting period and make different types of trading orders according to the forecast results at the same time.

The strategy is:

Algorithm 1 The Strategy (model is updated everyday)

Model's input: Bitcoin or Ethereum's five-minute frequency market bar data,position:
Model's body: RandomForest
Model's output: Prediction result

```
if prediction==1 then
    if position==0 then
        long entry
    else if position <0 then
        short exit and long entry
    end if
else if prediction==-1 then
    if position==0 then
        short entry
    else if position >0 then
        long exit and short entry
    end if
else
    keep taking
end if
```

The following two graph represent the backtesting performance of the strategy from Feb.6, 2018 to Aug.5, 2018, with a commission and slippage rate at one in ten thousand. The first part of the graph is the curve of total capital. And the second part shows the market close price during that time and The green and red triangles represent the entry and exit price for each trade order. The third part shows the capital of per transaction.The fourth part of the chart means the Max Drawdown and the last part means position.

From these two figures, we can see that strategy can capture the trend change of Bitcoin,and Ethereum, but when the bitcoin price has a relatively large price reversal, there will be a large drawdown.

Therefore, in the future, we need to improve the details of the strategy, such as setting a reasonable stop loss price, and need to have a more detailed classification for price fluctuations. More importantly, we need to control the transaction exposure.

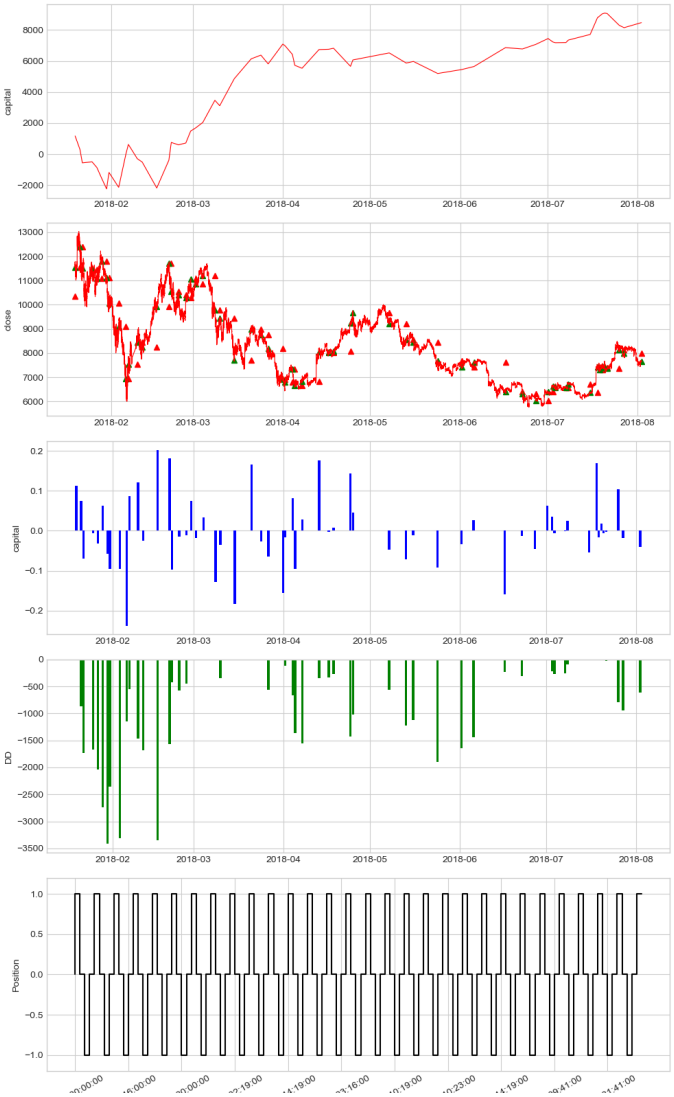

Fig. 4. performance of strategy on binance-BTC

VI. Conclusion & Future Work

In this paper we use random forests to build prediction model on different cryptocurrencies market OHLCV data from binance and bintfinex. We use some factors of alpha101 to do feature engineering. Based on the performance of all random forest models on the dataset, we can see that some factors play important role in predicting price movements and model can make more accurate predictions over longer time intervals. The backtesting result on bitcoin and Ethereum also support this conclusion. In the future work, We will try with a more parameter tuning method to get higher accuracy models and add more factors to increase the predictive power of the model.

In the future we will try to use more kinds of statistical model algorithms to compare their performance on the dataset.

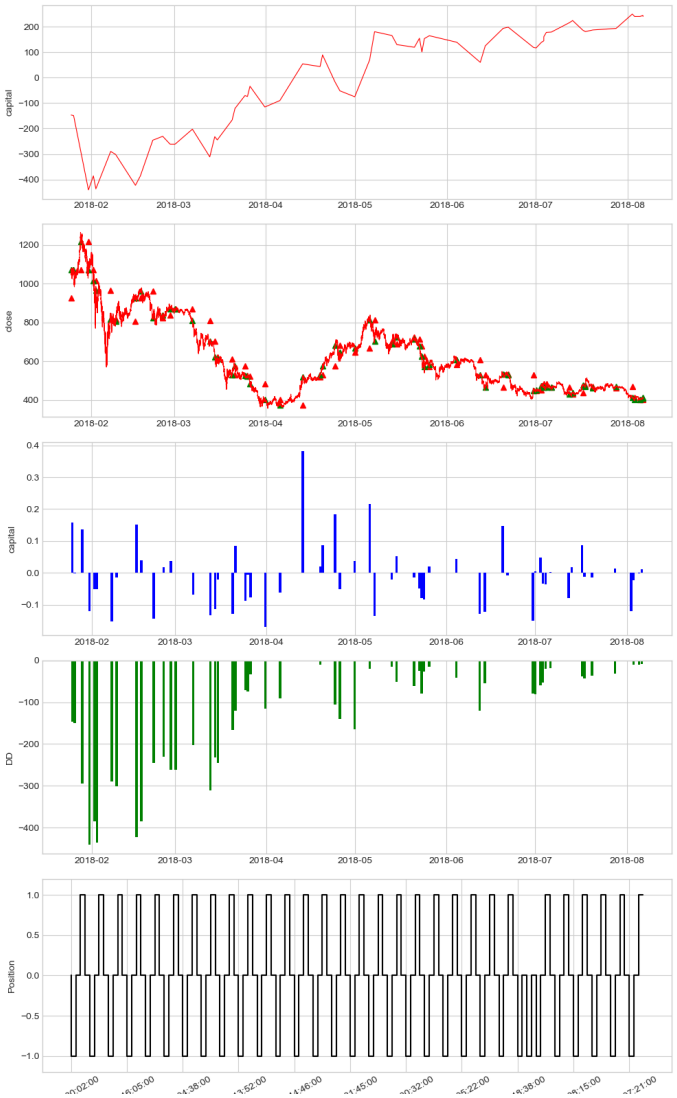

Fig. 5. performance of strategy on bitfinex-ETH

We will also try deep learning algorithms to predict the data. At the same time, we are deeply aware that using only market data to build the model is not enough. We will pay more attention to factors such as the cryptocurrency community activity, the discussion of social media, the status of the cryptocurrency itself and other factors.

REFERENCES

[1] Kakushadze, Z. (2016). 101 Formulaic Alphas. Wilmott, 2016(84), 72-81. doi:10.1002/wilm.10525

[2] Stefan Hubrich, "'Know When to Hodl 'Em, Know When to Fodl 'Em': An Investigation of Factor Based Investing in the Cryptocurrency Space" , SSRN Electronic Journal· January 20172139/ssrn.3055498

[3] Al Shehhi A, Oudah M, Aung Z, "Investigating factors behind choosing a cryptocurrency", IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), (2014), pp.1443-1447.

[4] Phillips RC, Gorse D. Cryptocurrency price drivers: Wavelet coherence analysis revisited. PloS one. 2018 Apr 18;13(4):e0195200.

[5] Madan, Isaac, Shaurya Saluja, and Aojia Zhao. "Automated Bitcoin Trading via Machine Learning Algorithms." (n.d.): 1-6. cs229.stanford.edu. Web

[6] Phillips RC, Gorse D, "Predicting cryptocurrency price bubbles using social media data and epidemic modeling", IEEE Symposium Series on Computational Intelligence (SSCI), (2017), pp.1-7.

[7] Yecheng Yao, Jungho Yi, Shengjun Zhai, Yuwen Lin, Taekseung Kim, Guihongxuan Zhang, Leonard Yoonjae Lee, "Predictive Analysis of Cryptocurrency Price Using Deep Learning ",International Journal of Engineering & Technology, 7 (3.27) (2018) 258-264

[8] Anton Misnik, Sergey Krutolevich, Siarhei Prakapenka, Max Vasilyeu "Neural Network Approximation Precision Change Analysis on Cryptocurrency Price Prediction ",FTI 2018. 96-101.