

Development and application of investment prediction model based on gold and bitcoin

Xuanwu Wang^{1,#}, Sirun Zheng^{2,#}

¹ The Division of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai, China

² School of Business, Macau University of science and technology, Macau, China

These authors contributed equally to this work.

Abstract. How to predict the change trend of asset prices in the future and decide different operation modes in advance to obtain the maximum benefits is the concern of investors. Taking gold and bitcoin as examples, this paper develops an appropriate mathematical model that uses only the past daily price stream to help traders determine whether to buy, hold or sell assets in their portfolio every day. At the same time, the robustness of the model is analyzed by robustness. The study found that holding US \$1000 on September 11, 2016 will eventually maximize profits on September 10, 2021.

Keywords: The maximum benefits; Gold and bitcoin; Mathematical model; The robustness of the model.

1. Introduction

Nowadays, more and more people participate in the trading and trading of funds, stocks, bonds and assets. The two most important trading assets are gold and bitcoin [1]. Generally speaking, successful market traders often buy and sell volatile assets with the goal of maximizing total return. The ultimate way to obtain the maximum benefit is to determine different operation modes in advance by predicting the change trend of asset prices in the future.

The task of this paper is to develop an appropriate mathematical model that uses only the past daily price flow to help traders determine whether to buy, hold or sell assets in their portfolio every day. On September 11, 2016, he held us \$1000 and finally maximized profits on September 10, 2021 [2]. See Figure 1 for the definition of symbols in this paper.

Nomenclature	
Symbols	Definition
BT	Bitcoin price today
GT	Gold price today
BN	Bitcoin price the next day
GN	Gold price the next day
BP	Bitcoin price predicted
GP	Gold price predicted
BH	Bitcoin held value today
GH	Gold held value today
BTVT	Bitcoin transaction volume today
GTVT	Gold transaction volume today
TA	Total assets
IRB	The interest rate of the bitcoin
IRG	The interest rate of the gold

Figure 1. The definition of symbols

2. Model establishment

A decision tree is a tree structure in which each internal node represents a test on an attribute, each branch represents a test output, and each leaf node represents a category. Common decision tree algorithms are C4.5, ID3 and CART (Figure 2).

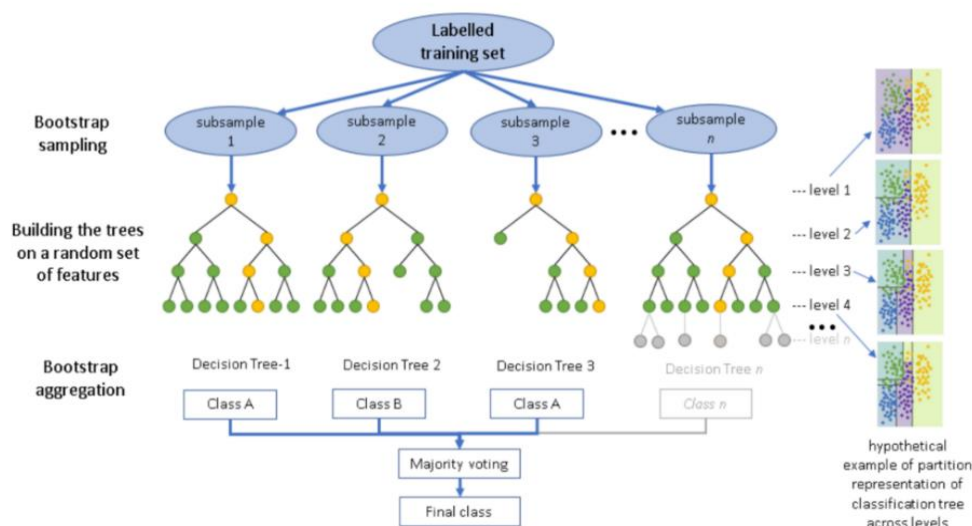


Figure 2. The flow chart of the process of the random forest application

Due to the randomness of random forest selection of training samples and features, and the characteristics of classifying test samples many times [5]. Applying this algorithm for classification can better solve the problem of weak generalization ability of a single decision tree, and make up for decision-making to a certain extent. Trees are prone to overfitting defects.

2.1 Model Running Process (Figure 3)

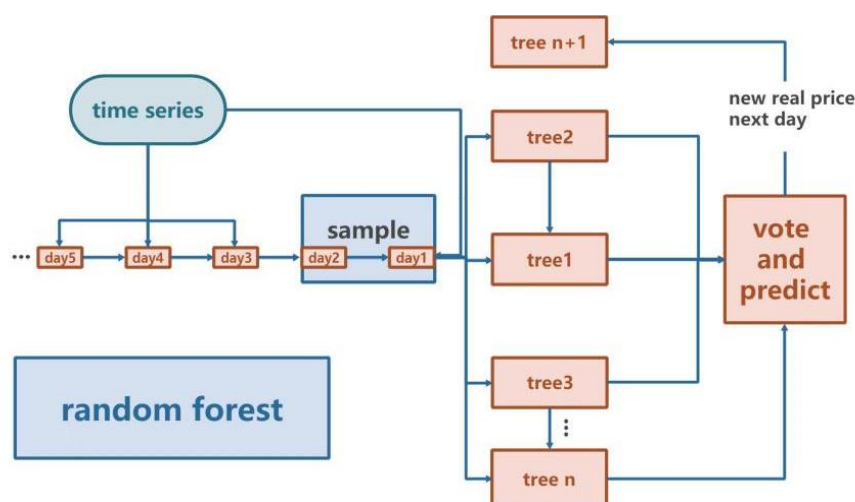


Figure 3. The flow chart of the running process of the random forest [6]

1. Feed data into this random forest model every two days starting on day three. Set random_state to the number 5. Through the method of bootstrap, the data of the first two days are taken as samples in sequence according to the time series.
2. Based on the stock price, we build a CART decision tree for it.
3. Repeat the above two steps. Each iteration will generate a decision tree.
4. Once a new stock price enters the random forest every day, after the decision of each tree, the best predicted price is obtained by voting as the basis for subsequent asset allocation.

2.2 Data integration and implementation of the model

Because the trading hours of gold and bitcoin are different, gold will be closed on holidays, so we use python's pandas merge to connect the prices of the two assets through time (Figure 4,5). The sliding window transformation of time series data is used to convert time series data into regression data. Simply put, it is to change a single series of data into X->Y regression data (Table 1,2).

Table 1. The comparition of gold

	0	1	2	3	4	...	1821	1823
Y	609.67	610.92	608.82	610.38	609.11	...	52677.40	46809.17
ShifX_0	621.65	609.67	610.92	608.82	610.38	...	51769.06	52677.40

Table 2. The comparition of bitcoin

	0	1	2	3	4	...	1260	1261
Y	1323.65	1321.75	1310.80	1308.35	1314.85	...	1786.00	1788.25
ShifX_0	1324.60	1323.65	1321.75	1310.80	1308.35	...	1802.15	1786.00

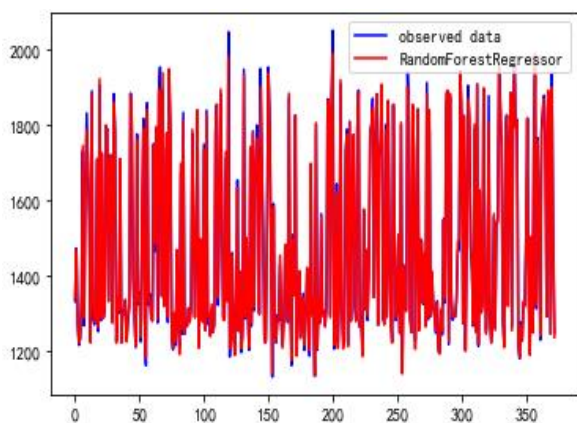


Figure 4. The comparision of gold

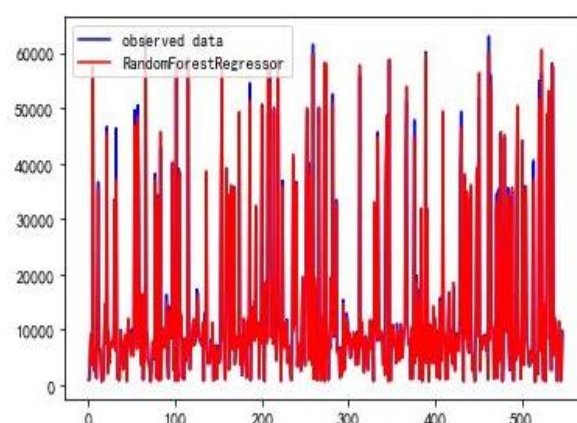


Figure 5. The comparision of gold

2.3 Data integration and implementation of the model

Genetic algorithm (GA) is a computational model that simulates the biological evolution process of natural selection and genetic mechanism in Darwin's biological evolution theory (Figure 6,7). It is a method to find the optimal solution by simulating the natural evolution process [3,4].

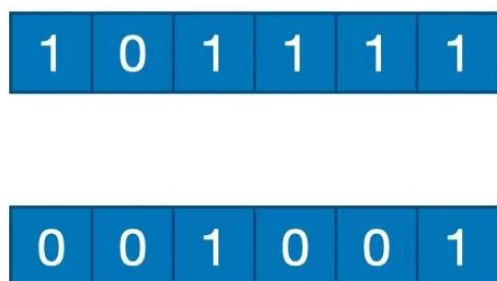


Figure 6. Principles of genetic

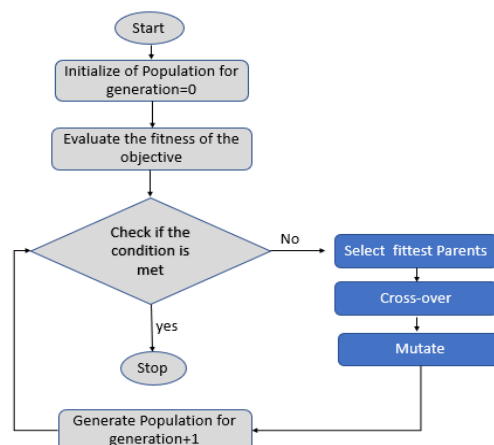


Figure 7. The flow chart of the process

Currently, there is no general-purpose traditional optimization method that can handle various types of constraints. Suppose, if we predict that both products are falling, and the decline exceeds the

value of the fee, we will choose to sell both products. If one of the two products rises and the rise exceeds the handling fee, then we will choose to trade.

Then we will go through the four parts of the genetic algorithm. Python's genetic algorithm package handles our constraints. We utilize the constraint method of penalty method. Among them, $f(BTVT)$, $f(GBTVT)$ become our objective function.

$$f(GTVT) = TA = \frac{GN}{GT}(GTVT + GH_- + \frac{BN}{BT}(f(BTVT) + BG) \quad (1)$$

Where, $\frac{BN}{BT} - 1 > 0.02$, $\frac{GN}{GT} > GC = 0.01$, $GT \cdot VT \in [-GH, GH]$, $GT \cdot VT / 0.99 = BT \cdot VT / 0.98$, $(GP_1, GP_2, GP_3, GP_4 \dots GP_n) \in GP$, $(BP_1, BP_2, BP_3, BP_4 \dots BP_n) \in BP$.

Under certain conditions, the solution of the evaluation function is equivalent to the solution in the original problem. The construction of fitness evaluation function usually adopts two ways of addition and multiplication. The additive approach simply adds an additional penalty term to the original objective function. That is, the evaluation function is:

$$eval(x) = f(x) + p(x) \quad (2)$$

Here x represents the chromosome. $f(x)$ (>0) is the objective function. $P(x)$ is the penalty term, and then the crossover operation.

By searching through data and assays, we determined the mutation value to be 0.001. After the method is determined, we initialize the population. Then we select a large enough population. Determine the number and length of chromosomes used. We determine the number of generations. Put the predicted prices of two assets each day into the model to get a daily customized trading strategy. Through these constraints, we found a more reasonable genetic parameter (the largest genetic subband is located at 500, the population size is set at 100, and the mutation probability is 0.001). Although the more genetic subbands and the population size, the better, but due to computer limitations, we chose a relatively suitable number.

On September 10, 2021, the initial investment value of \$1,000 on September 11, 2020 became \$276,239.9.

3. Some evidences for the best strategies are offered by our model

3.1 Data integration and implementation of the model

In the process of selecting our mathematical model, we introduced several statistical formulas to measure the accuracy of the model for ease of understanding.

$$SSE = \sum_{i=1}^n w_i (y_i - \hat{y}_i)^2 \quad (3)$$

The closer SSE (Sum of Squares due to Error) is to 0, the better model selection and fitting, and the more successful data prediction.

MSE (Mean-Square Error):

$$MSE = SSE/n = \frac{1}{n} \sum_{i=1}^n w_i (y_i - \hat{y}_i)^2 \quad (4)$$

RMSE (Root Mean Square):

$$RMSE = \sqrt{MSE} = \sqrt{SSE/n} = \sqrt{\frac{1}{n} \sum_{i=1}^n w_i (y_i - \hat{y}_i)^2} \quad (5)$$

R-Squared:

$$R - square = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad (6)$$

3.2 The model catching for the most profit according to the problem

Random forest can use Out of Bag Error (OOB error) to measure the performance of random forest. In an ideal case, about 36.8 % of the total training data forms the OOB sample. This can be shown as follows.

If there are N rows in the training data set. Then, the probability of not picking a row in a random draw is:

$$\frac{N-1}{N} \quad (7)$$

Using sampling-with-replacement the probability of not picking N rows in random draws is:

$$\left(\frac{N-1}{N}\right)^N \quad (8)$$

Which in the limit of large N becomes equal to:

$$\lim_{n \rightarrow \infty} \left(1 - \frac{1}{N}\right)^N = e^{-1} = 0.368 \quad (9)$$

Therefore, about 36.8 % of total training data are available as OOB sample for each DT and hence it can be used for evaluating or validating the random forest model.

The effectiveness of OOB error in measuring the performance of random forests has been demonstrated. Of course, the training set and the test set can also be divided, which can measure the difference in performance between different random forests, because the test set is guaranteed to be the same. Compared with time series analysis models or autoregressive models, the model we use is more reasonable and accurate. Therefore, we can use out-of-bag samples to predict the future price very early, and generally there will be no large deviation.

Therefore, it is a continuous learning process, rather than a training set that exists all the time like a time series. In other words, traders can invest earlier based on forecasts and get more returns.

4. Robustness and Sensitivity analysis [7-9]

Cart algorithm in random forest model has pruning loss function [10]:

$$C_\alpha(T_t) = C(T_t) + \alpha|T_t| \quad (10)$$

α is the regularization parameter (same as the regularization of linear regression), $C(T_t)$ is the prediction error of the training data, and $|T_t|$ is the number of leaf nodes in the subtree T . When $\alpha = 0$, that is, there is no regularization, the original generated CART tree is the optimal subtree. For a fixed α , there must exist a unique subtree that minimizes the loss function $C_\alpha(T_t)$.

Table 3. Sensitivity calculation

Agold/abitcoin	Final value	Sensitivity
1%/2%	276239	0
3%/2%	304562	10.25%
1%/4%	230980	16.38%
4%/2%	322828	16.87%
1%/1%	265234	3.98%
2%/2%	295752	7.06%
1%/3%	256211	7.25%

Use the control variable method, use 1% as the gradient to change, and take the original gold and Bitcoin transaction fees $\alpha_{gold} = 1\%$ and $\alpha_{bitcoin} = 2\%$ The final value of \$276,239 is used as the standard value. By changing the values of α_{gold} and $\alpha_{bitcoin}$, the final value under different

transaction costs is calculated through the asset return forecast distribution model, and the difference between it and the standard value of \$276,239 is divided by the standard value to obtain the sensitivity under the transaction cost (Table 3).

Based on the sensitivity values obtained above, produced a specific range map. According to this map, we can see that in the relatively small changes in transaction costs, the final rate of return of our quantitative investment strategy does not appear relatively large. changes, has a good anti-risk ability.



Figure 8. Sensitivity analysis

As can be seen from the above Figure 8, in the relatively small changes in transaction costs, the final rate of return of our quantitative investment strategy has not changed significantly, and it has a good ability to resist risks.

Based on the analysis of the above decision indicators, robustness and sensitivity, our quantitative trading decision model has good economic benefits under the premise of anti-interference and anti-risk, and it is recommended to be adopted by traders.

References

- [1] Ye Wuyi, Sun Liping, Miao Boqi Research on dynamic cointegration of gold and bitcoin -- Based on semi parametric Midas quantile regression model [J] Systems Science and mathematics, 2020, 40 (07): 1270-1285
- [2] <https://www.comap.com/>
- [3] https://blog.csdn.net/weixin_26632369/article/details/108173515
- [4] Wen Yu and Renhou Li. Study Based on Genetic Algorithms of Constrained Optimization, Computer Science, vol.29, no.6, pp.98-101, 2002.
- [5] Cootes T.F., Ionita M.C., Lindner C., Sauer P. (2012) Robust and Accurate Shape Model Fitting Using Random Forest Regression Voting. In: Fitzgibbon A., Lazebnik S., Perona P., Sato Y., Schmid C. (eds) Computer Vision – ECCV 2012. ECCV 2012. Lecture Notes in Computer Science, vol 7578. Springer, Berlin, Heidelberg.
- [6] Manish Kumar and M.Thenmozhi. Forecasting Stock Index Movement: A Comparision of Support Vector Machines and Random Forest. Indian Institute of Capital Markets 9th Capital Market Conference Paper, 2006.
- [7] Sun Jinhua, Wang Sanyin, Yu Cuiping Research on the difference of nonlinear relationship between accounting conservatism and enterprise technological innovation [J] Chinese certified public accountant, 2022 (03): 65-69
- [8] Guan Xuehua, Chen Zhihe, ye Zhiheng Dynamic simulation and sensitivity analysis of water resources system in Zhongshan City [J] Water resources protection, 2022, 38 (02): 103-111
- [9] Xi Xinqiang, Gong Xintong, Zhang shangyun, Yang Nan, sun Shucun Structure and robustness of "pollinator plant herbivore" interaction network in alpine meadow in Northwest Sichuan [J] Chinese Science: Life Science, 2022, 52 (03): 449-458
- [10] Han Yuhao, Cao Lijun, Wang Youqi, Zhang Jingjing Analysis of pest prediction model based on decision tree cart algorithm [J] Modern agriculture, 2022 (01): 45-47