Investment Decision Model Based on Random Forest and Genetic Algorithm

Longlong Li[1], Ziyan Xiang[2], Hongrui Shen[3], Yutong Wu[4], Yongqi Ji[5,] Shengwei Wang[1,*]

[1]Department of Computer Science and Engineering, Northwest Normal University, Lanzhou, Gansu, 730070

[2]Department of Mathematics and Statistics, Northwest Normal University, Lanzhou, Gansu, 730070

[3]College of Economics, Northwest Normal University, Lanzhou, Gansu, 730070

[4]College of Chemistry and Chemical Engineering, Northwest Normal University, Lanzhou, Gansu, 730070

[5]College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou, Gansu, 730070

*Corresponding author: wangshengwei0308@163.com

Keywords: Random Forest, Target planning, Genetic Algorithm.

Abstract: This paper will propose a reasonable, scientific, and practical decision-making plan for the quantitative investment of gold and bitcoin by establishing a model. Firstly, the random forest model was used to predict the market trend of gold and bitcoin in the next few days. The results showed that the fitting degree of the forecast model exceeded 99%. Subsequently, the objective function is established based on whether the gold trading stops the next day and the constraint conditions are determined. The Genetic Algorithm is used to solve the problem, which is necessary to prove that the proposed scheme is optimal. The RMSE index is calculated to measure the accuracy of the Random Forest Prediction.

1. Introduction

It is important to have an accurate, fast, reasonable, and effective forecasting method in today's fast-moving markets. After comparing grey forecasts, time series methods, and other forecasting methods based on past data, it was decided to use machine learning methods. The random forest model was chosen as the most accurate among the machine learning methods. The random forest is insensitive to multivariate covariance. The results are more robust to missing and unbalanced data and can predict up to several thousand explanatory variables very well.

Genetic algorithms originate from computer simulations of biological systems and are a stochastic global search and optimization method that simulates the replication, crossover, and mutation in natural selection and heredity. Starting from an initial population, the population evolves to an increasingly better region of the search space through random selection, crossover, and mutation operations to produce a group of individuals that are better suited to the environment [7,10].

2. Model Establishment and Solution

2.1. Forecasting and target planning models

(1). Prediction based on Random Forest algorithm

In the random forest model, new samples are judged separately by the decision tree after entering each of the decision trees in the random forest. The random forest improves prediction accuracy without a significant increase in computing power. The random forest is insensitive to multicollinearity, and the results are more robust to missing and unbalanced data, and it predicts the effects of up to several thousand explanatory variables well [5].

The random forest generates a new training sample set by repeatedly drawing k samples (k is generally the same as N) at random from the original training sample set N by the self-help method

resampling technique and then generates n classification trees to form a Random Forest based on the self-help sample set [12].

The training and test sets are divided according to the characteristics and objectives of the data. To divid the pre-processed data, the first 70% of the data is used as the training set, and the least 30% of the data is used as the test set. The gridSearchCv function is also called here for moderation to produce optimal results.

The results of the model predictions compared to the actual values are partially shown in Figure 1.

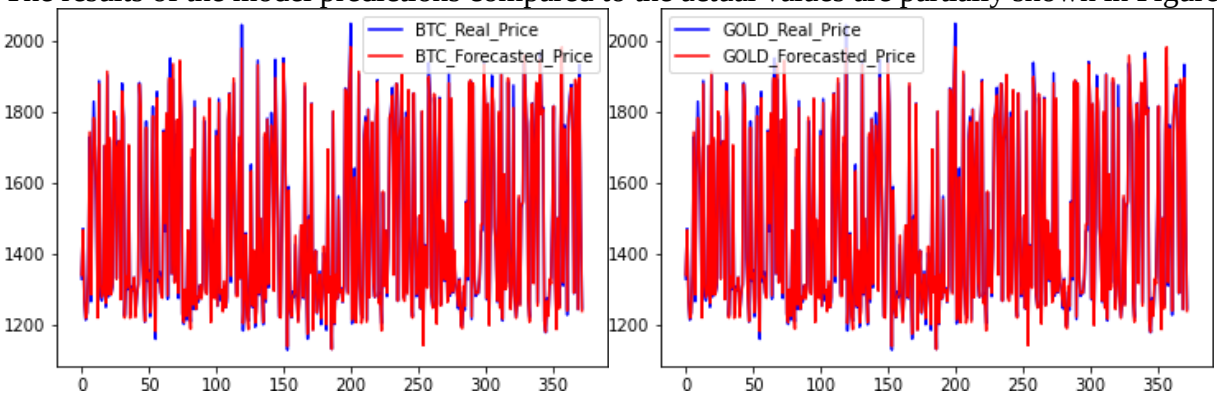

Figure 1 Comparison of Real Prices and Forecast Prices for Bitcoin (above) and Gold (below)

(2). Single-objective planning based on Genetic Algorithm

The following model is set up to maximize the assets after the next transaction. As no funds can be borrowed from outside parties during the transaction, the control constraints are that the value of cash, gold, and bitcoin cannot be negative. Again, as the gold and bitcoin assets will not be negative in the market during the transaction, they will be doomed to lose if there are more fees and will not be considered if the fees are greater than the value of the assets. This condition also reduces time and space complexity for subsequent calculations [4,6].

(1) Day i can trade gold:

The total value of assets at the end of trading on the day i is $W_i = C_i + P_{G_i}G_i + P_{B_i}B_i$, the market price $P'_{G_{i+1}}$ and $P'_{B_{i+1}}$ on day i+1 are derived from the forecast and the trading strategy on the day i represented by the two unknown quantities ΔG_i and ΔB_i. The trading strategy is determined by the volume of gold and bitcoin traded.

The objective function and constraints are as follows:

$$max\, W'_{i+1} = C_i - \alpha\%$$

$$max\, W'_{i+1} = C_i - \alpha\% P_{G_i}|\,\Delta G_i| \;-\; \beta\%\, P_{B_i}|\,\Delta B_i\,| \;-\; P_{G_i}\,\Delta G_i - P_{B_i}\Delta B_i \;+\; (G_i + \Delta G_i)\,P_{G_{i+1}} + (B_i + \Delta B_i)\,P_{B_{i+1}} \quad (1)$$

s.t.

$$\begin{cases} C_i - \alpha\% P_{G_i}|\,\Delta G_{i+1}| \;-\; \beta\%\, P_{B_i}|\,\Delta B_{i+1}\,| \;-\; P_{G_i}\,\Delta G_{i+1} - P_{B_i}\Delta B_{i+1} \geq 0 \\ G_i + \Delta G_i \geq 0 \\ B_i + \Delta B_i \geq 0 \\ P_{G_{i+1}}/P_{G_i} \geq 1 + \alpha\% \\ P_{B_{i+1}}/P_{B_i} \geq 1 + \beta\% \end{cases} \quad (2)$$

The cash, gold, and bitcoin holding in after today's trading were:

$$C_{i+1} = C_i - \alpha\% P_{G_i}|\,\Delta G_i| \;-\; \beta\%\, P_{B_i}|\,\Delta B_i\,| \;-\; P_{G_i}\,\Delta G_i - P_{B_i}\Delta B_i \quad (3)$$

$$Bi{+}1 = Bi + \Delta Bi{+}1 \quad (4)$$

$$Gi{+}1 = Gi + \Delta Gi{+}1 \quad (5)$$

(2) No gold trading on the day i:

The situation is the same as in the previous process, but gold is not available for trading and closed. The value of gold is the same as it was on the previous day, but it cannot be traded. Only Bitcoin can be traded. Finding the maximum expected total assets while requiring that the value of the transaction is not negative, i.e., satisfying the condition that cash and bitcoin are not negative after the transaction.

$$max\, W_{i+1} = C_i - \beta\%\, P_{B_i} |\, \Delta B_i \,| - P_{B_i} \Delta B_i + G_i\, P_{G_{i+1}} + (B_i + \Delta B_i)\, P_{B_i+1} \tag{6}$$

$$\text{s.t.} \begin{cases} C_i - \beta\%\, P_{B_i} |\, \Delta B_{i+1} \,| - P_{B_i} \Delta B_{i+1} \geq 0 \\ B_i + \Delta B_i \geq 0 \\ P_{B_{i+1}} / P_{B_i} \geq 1 + \beta\% \end{cases} \tag{7}$$

Meet the condition that cash and bitcoin are not negative after the transaction.

The cash, gold, and bitcoin holding in after today's trading were:

$$C_i - \beta\%\, P_{B_i} |\, \Delta B_i \,| - P_{B_i} \Delta B_i \geq 0 \tag{8}$$

The actual total assets on the second day were:

$$W_{i+1} = C_{i+1} + P_{B_i} B_{i+1} + P_{G_i} G_{i+1} \tag{9}$$

A genetic algorithm is a population-based operation that takes all individuals and uses only the basic genetic operators: the selection operator, the crossover operator, and the variation operator. Selection, Crossover, and Mutation are the three main operators of a genetic algorithm, and they form the genetic operations that give the genetic algorithm features not found in other methods.

This paper solves the model to obtain charts of total assets, bitcoin holdings, and gold holdings over a 5-year cycle.

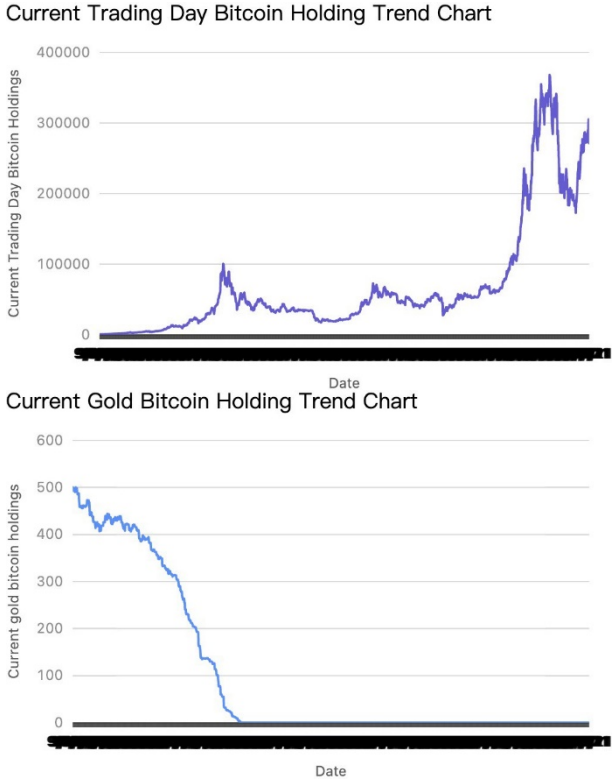

Figure 2 Trend of Bitcoin and Gold Holdings

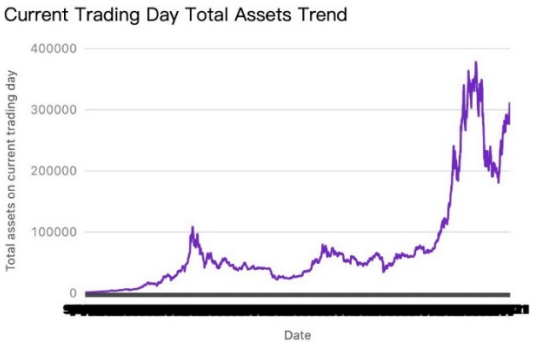

Figure 3 Trend of Total Assets Holdings

From Figures 2 and Figure 3, we see a gradual increase in the value of bitcoin investments relative to gold. Although there are ups and downs during this period, in general, the investment value of Bitcoin is gradually greater than that of gold, and the total personal assets are also on an upward trend. The large ups and downs in the curve indicate that the trading model is flawed and does not consider the robustness of gold trading and the preservation of value.

(3). A revision of the Genetic Algorithm based single-objective planning model

In the daily trading process, first understand whether the gold market will be open in the next few days. If both bitcoin and gold can trade freely, the first part will be used, and if there is a break in the gold market in the next few days, it will be reconsidered. For example, although today's decision is the best, the gold cannot be traded for the next two days, which greatly limits capital activity and cannot be invested in time. Therefore, a function *f* should be explored, which is a function of the gold valuation on the next day, the valuation on the next gold market opening, and the number of days during which the gold cannot be listed. Using this function will improve the formula in process one. To enhance the solid role that gold plays in investment, we change from a simple short-term decision to a long-term decision under certain conditions from a gold break.

(1) Gold may be traded on day *i* and not on Days *i*+1 to *i* + *t*:

After the market price is announced on Day *i*, it is not allowed to trade from Day *i*+2. It is necessary to decide the investment strategy on Day i+1 carefully. The value of gold traded is multiplied by the risk assessment function *f*. Estimated total assets equal to the sum of the cash held, the value of gold held, and the cash value of bitcoin held after the market price formula.

$$max\, W_{i+1} = C_i - \alpha\% P_{G_i} |\, f\Delta G_i| \; - \; \beta\%\, P_{B_i} |\, \Delta B_i \,| \; - \; P_{G_i} f\Delta G_i - P_{B_i}\Delta B_i \; + \; (G_i + f\Delta G_i\,)\, P_{G_{i+1}} + \; (\, B_i + \; \Delta B_i)\, P_{B_{i+1}} \tag{10}$$

$$\text{s.t.} \begin{cases} G_i \; + f\Delta G_i \geq 0 \\ B_i \; + \Delta B_i \geq 0 \\ P_{G_{i+1}}/P_{G_i} \geq 1 + \alpha\% \\ P_{B_{i+1}}/P_{B_i} \geq 1 + \beta\% \end{cases} \tag{11}$$

At the same time, the conditions must be met that neither good nor bitcoin can be negative after the transaction of today.

Cash, gold, and bitcoin held after Day *i*+1 transaction is modeled in the previous content.

(2) Gold cannot be traded on day *i* and gold cannot be traded on Days *i*+1 to *i* + *t*.

Meanwhile, the situation is the same as in the previous process, but gold cannot be traded. Cash and bitcoin that meet the condition are not negative after the transaction.

Next, we will construct the function *f* [11].

$$f = \lambda \frac{P_{G_{i+t}} - P_{G_{i+1}}}{t} \left(x_1 \frac{W'_{i+1} - P_{B_i} B_{i+1} - C_{i+1}}{G_i\, P_{G_{i+1}}} + x_2 \frac{W_{i+2} - P_{B_i+1} B_{i+2} - C_{i+2}}{G_{i+2}\, P'_{G_{i+1}}} + \cdots + \varepsilon\right) \tag{12}$$

$$\text{s.t. x1 + x2 + } \cdots \text{ + xn = 1} \tag{13}$$

Where λ, x_1, x_2, ..., x_n are parameters to be determined, and ε is any small positive real number.

The establishment of this function relates to the amount of investment during the golden week. The more days there is, the lower the activity of the fund, which also means the greater the risk. Based on the previous data, we calculated the data by the undetermined coefficient method.

In the last calculation, the risk function *f* will control the gold holdings, consider the investment periodically, and improve the loopholes that cannot be considered for the long-term transaction before the model.

As shown in Figures 4 and 5, using the improved model to solve the problem, a steady increase in total assets is seen, which indicates that the model correction is effective.

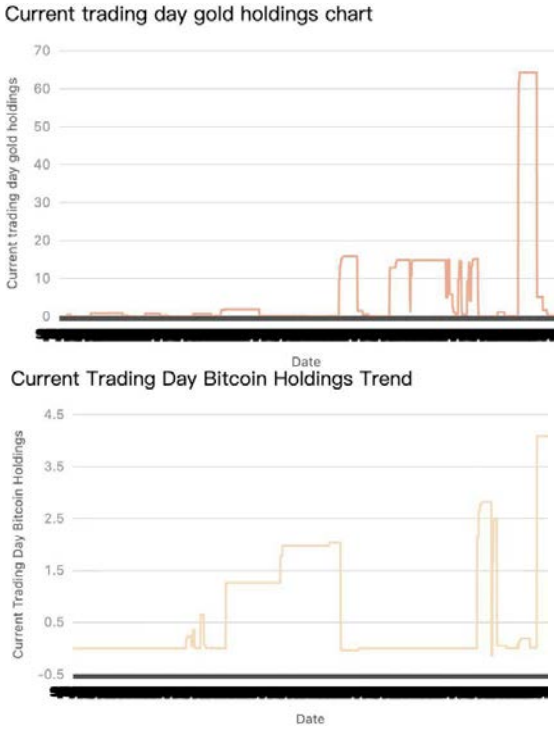

Figure 4 Trend of Bitcoin (below) and Gold (above) Holdings (Revision)

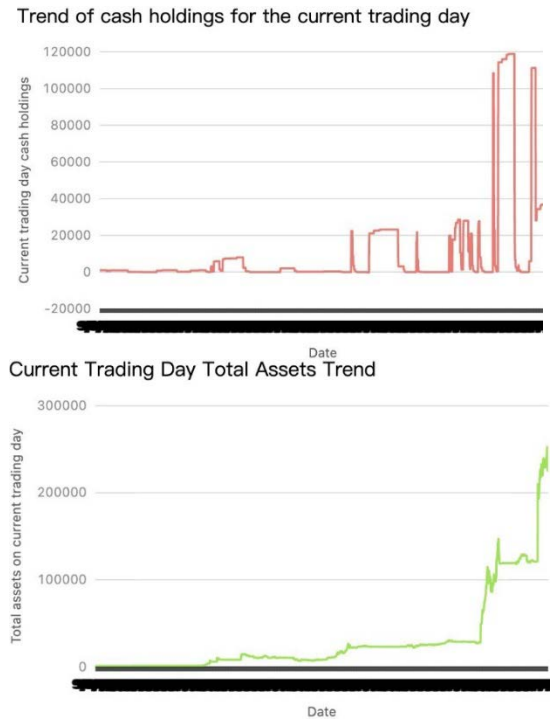

Figure 5 Trend of cash (above) and total assets (below) holdings (Revision)

3. Analysis and Evaluation of Models

3.1. Establish root mean square error model

$$RSME(S,\overline{S}) = \sqrt{\sum_{t=1}^{n}(S_t - \overline{S}_t)^2/n} \tag{14}$$

The root means square error measures the deviation between the observed value and the actual value. It is the square root of the ratio between the square of the deviation between the predicted value and the actual value and the number of observations *N*. In practical measurements, the number of observations *N* is always limited, and the truth value can only be replaced by the most reliable (best) value. The standard error is very sensitive to very large or very small errors in a set of measurements, so the standard error can well reflect the precision. The data was used to calculate each evaluation score, as shown in Table 1.

Table 1. Scoring Metrics for Predictive Models

	MSE	*RMSE*	*MAE*	R^2
GOLD	250.320	15.82	11.13	0.99
BTC	116975.16	1081.52	497.63	0.99

For one thing, the error increases with increasing MSE, MAE, and RMSE. According to the properties and value of these indicators above, it is concluded that our model is generally precise. For another thing, the coefficient of determination of gold and Bitcoin is very close to one, representing that the model fits well.

3.2. Establish an investment risk model

Order transaction costs:

$$f(x_i) = \begin{cases} Mx_i p_i, x_i > 0 \\ 0, x_i = 0 \end{cases} \tag{15}$$

Then the net income is:

$$R = \sum_{i=0}^{2} M(1 + r_i)x_i - M \tag{16}$$

The overall risk is:

$$Q = \max Mxiqi,\ i = 0, 1, 2 \cdots \tag{17}$$

The constraint conditions are:

$$\sum_{i=0}^{2}(f(x_i) + Mx_i) = M \tag{18}$$

3.3. Solving investment risk models

Fixed income minimizes *Q*

$$\min Q = \max xiqi,\ i = 0, 1, 2 \cdots \tag{19}$$

$$\sum_{i=0}^{2}(r_i - p_i)x_i = R \tag{20}$$

$$\sum_{i=0}^{2}(1 + p_i)x_i = 1 \tag{21}$$

Make $\rho_i = (r_i - p_i)/(1 + p_i)$, ρ_i denotes the net rate of return on the *i*-th investment, then ρ_i must be greater than ρ_0, Otherwise, if $\rho_1 < \rho_0$, then no investment is made in S_i because the net rate of return on investment in the project is less than that of a bank deposit and the risk of loss is greater than that of a bank deposit. If ρ_i is ordered from smallest to largest, and set to largest, then it is easy to see that there must be $0.03 \leq R \leq \rho_k$ feasible solutions to the model.

When R=0.03, no gold or bitcoin is purchased, Q=0; when $R = \rho_k$, all funds are used to purchase S_i; when $0.03 \leq R \leq \rho_k$, there is the following conclusion.

3.4. Conclusion

if $0.03 \leq R \leq \rho_k$, (x_0, x_1, x_2) is the optimal solution of the model, then $x_1q_1 = x_2q_2$. Moreover, for the model where fixed returns minimize risk, this conclusion can also be stated: the overall risk is minimal when the risk losses of each investment are equal, i.e., $x_1q_1 = x_2q_2$, provided that the total of the first five investments is constant.

References

[1] Xinyue Liu." Empirical Study of Quantitative Investing Model".Proceedings of 2015 2nd International Conference on Education, Management and Information Technology(ICEMIT 2015). Ed. Atlantis Press, 2015, 289-293.

[2] Blitz David,and Vliet Pim van."The Conservative Formula: Quantitative Investing Made Easy." The Journal of Portfolio Management 44.7(2018): doi:10.3905/jpm.2018.44.7.024.

[3] Venter Pierre J.,and Maré Eben."Univariate and Multivariate GARCH Models Applied to Bitcoin Futures Option Pricing." Journal of Risk and Financial Management 14.6(2021): doi:10.3390/JRFM14060261.

[4] Ahin Telli, Hongzhuan Chen."Multifractal behavior in return and volatility series of Bitcoin and gold in comparison." Chaos, Solitons, and Fractals: the interdisciplinary Journal of Nonlinear Science, and Nonequilibrium and Complex Phenomena 139. (2020): doi: 10.1016/j.chaos.2020.109994.

[5] Mingqin Chen, et al."A Quantitative Investment Model Based on Random Forest and Sentiment Analysis." Journal of Physics: Conference Series 1575.1(2020): doi:10.1088/1742-6596/1575/1/012083.

[6] Christian Pierdzioch, and Marian Risse."A machine learning analysis of the rationality of aggregate stock market forecasts." International Journal of Finance Economics 23.4(2018): doi:10.1002/ijfe.1641.

[7] Qi Yue, et al. "Evaluation and analysis of the application effect of genetic algorithm in specially constructed portfolio selection model". Proceedings of the 19th China Management Science Annual Conference. Ed., 2017, 294-299.

[8] Huang Yonghao, and Chen Xi. "Portfolio Optimization with Proportional Charges Based on Sensitivity Analysis." Control and Decision-making 29.07 (2014): 1181- 1186.DOI: 10.13195/j.kzyjc.2013.0597.

[9] Liu Xiaojuan. "Portfolio Investment with Lower Bound of Capital and Its Sensitivity Analysis." Journal of Xinxiang University (Natural Science Edition) 29.03(2012):193- 194+199. DOI:

[10] Zhang Shijie. "The Application of Machine Learning under Big Data in Stock Market Forecast." Journal of Guiyang University (Social Science Edition) 16.04 (2021): 43-48. DOI: 10.16856/J. Cnki.52-1141/C.2021.04.007.

[11] Lin Dai. "Bitcoin Investment Risk Analysis." Accounting Newsletter. 05 (2015): 3-4. DOI: 10.16144/J. Cnki. ISSN 1002-8072.2015.05.001.

[12] Fang Kuangnan, et al. "Summary of research on stochastic forest method." Statistics and Information Forum 26.03(2011):32-38.