Research on the Investment Strategy of Quantitative Trading Based on Random Forest and Genetic Algorithm

Yuwei Chen*, Yueyang Zhang, Yufeng Lan, Yujia Liu, Yaoxiang Lin

School of Economics and Finance, South China University of Technology, Guangzhou, Guangdong, 510000, China

*Corresponding author. Email: chenyuwei0414@163.com

Abstract. Gold and Bitcoin are two common investments in the market. Investors profit from the capital market through specific trading strategies. Based on previous price data, this paper established a prediction-decision model, which provides investors with an optimal trading strategy and effectively improves the return on investment. To begin with, we use the past daily price of gold and bitcoin to calculate the Bollinger band, DMA, and MACD characteristics. We used the Random Forest algorithm to predict the next day's price and established a loop to predict all price data. Finally, we got the prediction data with the goodness of fit (R^2) as high as 0.98. We intuitively see that the prediction effect is pretty good from the predictive value-real value-line chart. Then, we established the transaction decision-making objective model, used the Genetic Algorithm to select the daily transaction amount, obtained the optimal transaction strategy, and then established a loop to obtain all the optimal strategies. The final total assets were 243,424.76, and the annual return rate reached 6085.62%.

Keywords: Quantitative Investments, Random Forest, Genetic Algorithm.

1. Introduction

Traditional investment - gold and emerging investment - bitcoin are both very mainstream investments in the market, yet their trading strategies are highly different. Investors tend to derive significant returns from investing in both.[1] How to preserve and increase the value of assets has become a problem faced by many people.[2] With this demand, quantitative investment, as an emerging form of financial technology, is becoming an investment approach that investors are paying more attention. Starting from 2017,[3] more investment institutions tend to transform their inherent investment trading methods into quantitative investments. Unlike traditional investment methods, quantitative investment mainly relies on data and models to find investment targets and strategies, using scientific and systematic methods to analyze big data thoroughly and achieve solid profits.

Quantitative investment is ineffective under the traditional efficient market hypothesis, which believes that investors cannot get excess returns by analyzing market information. However, in recent decades, researchers' analysis of the capital market has found much empirical evidence that quantitative investment can achieve excess returns by analysis. [4]

Traditional stock market trend prediction mainly uses linear regression and time series statistical methods, such as B. S. Bini, and Tessy Mathew (2016) used the Clustering and Regression Techniques method to predict the stock trend.[5]

Recently, the SVM has been one of the most popular algorithms employed in Machine Learning. In addition, Support Vector Regressor (SVR) is also widely used in market trends prediction because of its characteristics of noise tolerance and tuning of hyperparameters, which greatly improves the accuracy. ANNs are used for traditional stock price trend prediction but proved to be inefficient. A recurrent Neural Network (RNN) has become a popular choice in market forecasting. The RRN has been widely used to predict the stock price and has given good accuracy, but this accuracy is different according to the duration used for prediction. [6] KNN is another Machine Learning algorithm, which is also used to predict the stock market price, but with poor accuracy.[7] Bin Li(2017) used Machine Learning algorithms such as Support Vector Machine, Neural Network, and Adaboost to predict the trend of stock price fluctuation by 19 technical indicators and found that these algorithms had higher prediction accuracy. The portfolio constructed according to the prediction also achieved better

investment performance.[8] Krauss et al. (2017) integrated Deep Neural Network (DNN), Gradient Boosting Decision Tree (GBDT), and Random Forest strategy and used the earnings of all stocks in the past to predict the rise and fall of S&P500.[9] Nair, Mohandas, and Sakthive used similar methods to establish the relationship between the change of common technical indicators and the changing trend of the index and constructed investment strategies by identifying the rise and fall, which achieved good results.

2. Financial product price prediction model based on random forest

Random Forest refers to a classifier that uses multiple trees to train and predict samples. It belongs to an integrated algorithm and has a good application effect in classification, regression, and clustering. The Stock Trend Prediction is a regression prediction that uses the Classification and Regression Tree (CART) as the basic unit to construct the forest, and each CART is trained in turn. The training samples of each CART are obtained from the original training set by Bootstrap.

CART is easily over-fitting, but Random Forest uses Bootstrap and Aggregate (also known as Bagging) to solve the over-fitting problem. Meanwhile, the variance ability of the model is also enhanced by the presence of randomness.

We choose three indicators as the characteristics of the Random Forest by the prices of the past period ($T_m \sim T_n$), including the Different between Moving Average (DMA), Bollinger Bands (BOLL), and Moving Average Convergence / Divergence (MACD).

Comparing the best segmentation points of all characteristics that need to be collected for multiple eigenvalues, and ayre selecting the best characteristics segmentation points. All decision trees were integrated using average or voting. Assuming that there is a total of m samples and n characteristics data sets, t decision trees need to be constructed at most, and the number of characteristics of each decision tree is k, then the implementation process of Random Forest algorithm is as follows:

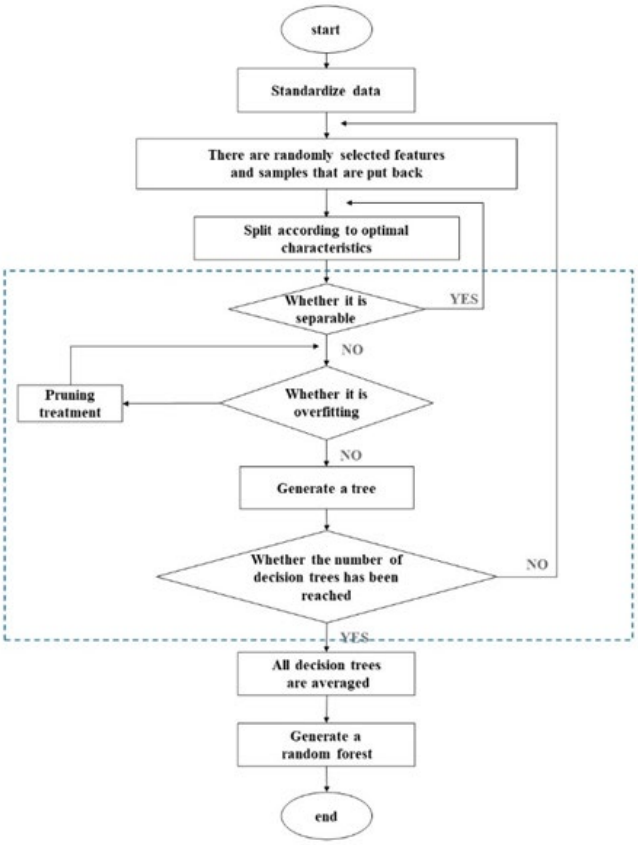

Figure 1 The Overview of Random Forest

BOLL was created by Mr. John Bollinger, which uses statistical principles to calculate the standard deviation (std) of stock prices and their trust intervals, to determine the range of stock price volatility and future trends, using bands to show the safe high and low-price levels of stock prices, thus also known as Bollinger Bands. Its upper and lower limit range are changing with the stock price rolling. The stock price fluctuates in the upper and lower limits of the range, the width of this band area, with the size of the stock price fluctuations and changes. The stock price increases and decreases in magnitude, and the band area becomes wider, up, and down in narrow consolidation, then the band area becomes narrower.

The calculation method of indicators is as follows:

Step 1. Calculate MA

$$MA = \frac{sum\ of\ the\ closing\ prices\ in\ N\ days}{N} \tag{1}$$

Step 2. Calculate the standard deviation (std)

$$s^2 = \frac{(x_1 - M)^2 + (x_2 - M)^2 + \cdots + (x_n - M)^2}{n} \tag{2}$$

Step 3. Calculate $Mid - track$, $Upper - track$, and $Lower - track$:

$$Mid - track = MA \tag{3}$$

$$Upper\ track = Mid - track + 2std \tag{4}$$

$$Lower\ track = Mid - track - 2std \tag{5}$$

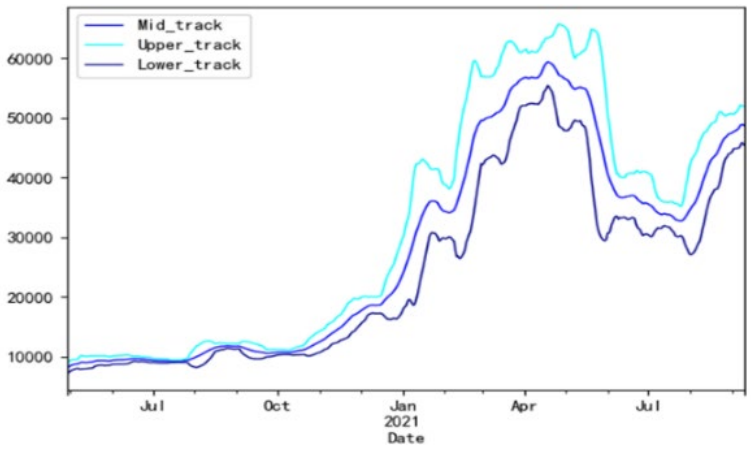

Figure 2 Bitcoin's partial BOLL lines

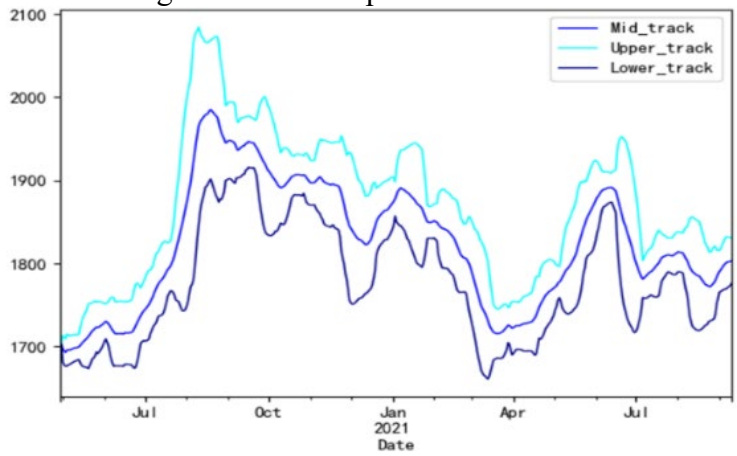

Figure 3 Gold's partial BOLL lines

DMA uses the average of two different periods to determine the amount of current buying and selling energy and future price trends. A positive number indicates a long market run, and a negative number indicates a short market run.

The calculation method is as follows:

Step 1. Calculate MA:

$$MA = \frac{sum\ of\ the\ closing\ prices\ in\ N\ days}{N} \tag{6}$$

Step 2. Calculate DMA:

$$DMA = MA(10) - MA(50) \tag{7}$$

$MACD$ is an improved moving average indicator proposed by Gerald Appel in 1979. The main disadvantage of the simple moving average (MA) is that it has a certain lag and indicates that every value within this period will have the same impact on the future trend. $MACD$ uses smoothed moving averages ($EMAs$), which greatly alleviate the problems of simple moving averages with lags and short backtest periods that lead to frequent trading signals. The $MACD$ indicator is the main technical analysis indicator of the stock market, which predicts the short-medium-term trend of the stock price through the study of the relationship between EMA, DIF and DEA. The study of the moving average connected by DIF and DEA, and the study of the bar chart (BAR) drawn by subtracting the DEMvalue from DIF.

The calculation method is as follows:

Step 1. Calculate the $EMA_t(m)$:

$$EMA_t(m) = EMA_{t-1}(m) \times \frac{m-1}{m+1} + P_t \times \frac{2}{m+1} \tag{8}$$

Step 2. Calculate the $DIFF_t$:

$$DIFF_t = EMA_t(m) - EMA_t(n) \tag{9}$$

Step 3. Calculate the DEA_t:

$$DEA_t = DEA_{t-1} \times \frac{p-1}{p+1} + DIFF_t \times \frac{2}{p+1} \tag{10}$$

Step 4. Calculate $MACD_t$:

$$MACD_t = 2 \times (DIFF_t - DEA_t) \tag{11}$$

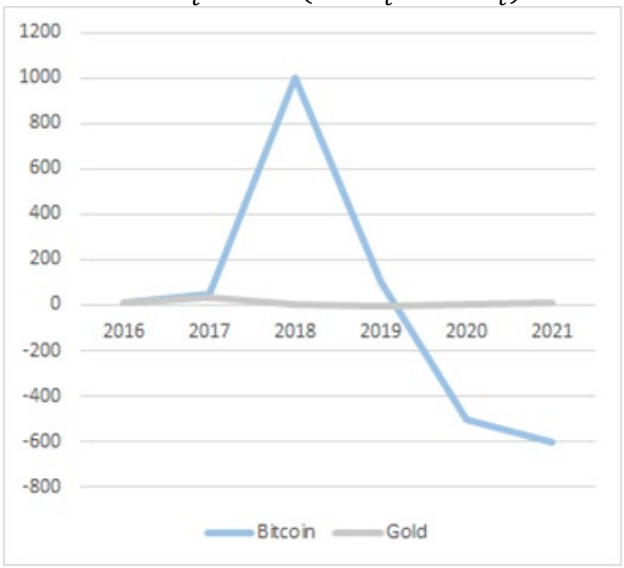

Figure 4 Moving Average Convergence / Divergence

Incorporating characteristics into the model for prediction:

Step 1. Traverses each characteristic in the input sample where the training dataset is located and recursively divides each region into two subregions. Calculate the sum of squares of residuals of n characteristics and their corresponding cut points, then a pair of (j, s) were found, which satisfies that the sum of squares of residuals of left and right subtrees is minimized, respectively. On this basis, the sum of the two is minimized again. The mathematical expression follows:

$$\min_{j,s}[\min_{c_1}\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2 + \min_{c_2}\sum_{x_i\in R_2(j,s)}(y_i-c_2)^2] \quad (12)$$

Where R_1 and R_2 represent the two subsets divided (the regression tree is a binary tree); c_1 and c_2 represent the mean values of R_1 and R_2 samples, respectively; j represents the sample characteristics of $BOLL$, $MACD$ and DMA; s represents the dividing point. y_i represents the actual value of the sample target variable.

Step 2. The selected (j,s) is used to divide the region and determine the corresponding output value. The sample mean value formula is:

$$\hat{c}_m = \frac{1}{N_m}\sum_{x_i\in \mathrm{Rm}(j,s)} y_i, x\in R_m, m=1,2 \quad (13)$$

$$R_1(j,s)=\{x \mid x^{(j)}\le s\} \quad (14)$$

$$R_2(j,s)=\{x \mid x^{(j)}\le s\} \quad (15)$$

Step 3. Continues to call steps 1 and 2 on two subregions until the partition cannot continue.

Step 4. Divides the input sample into M regions, namely: R_1,R_2, …, R_m to generate decision trees. The formula is as follows:

$$f(x)=\sum_{M}^{m=1}\hat{c}_m I(x\in R_m) \quad (16)$$

Where c represents the average value of the corresponding region; $\boldsymbol{i}$ represents whether the condition is met, which is 1, otherwise 0.

Step 5. Use a playback sampling to obtain a dataset with m samples (possibly duplicate samples) from the original dataset after m times of sampling. From n characteristics, the principle of no-replacement sampling is adopted, and k characteristics are removed as input characteristics. Repeat the above process t times on the new dataset to construct t decision tree.

Step 6. Generates a Random Forest by averaging the generated t decision tree.

After random forest prediction, our prediction results are as follows:

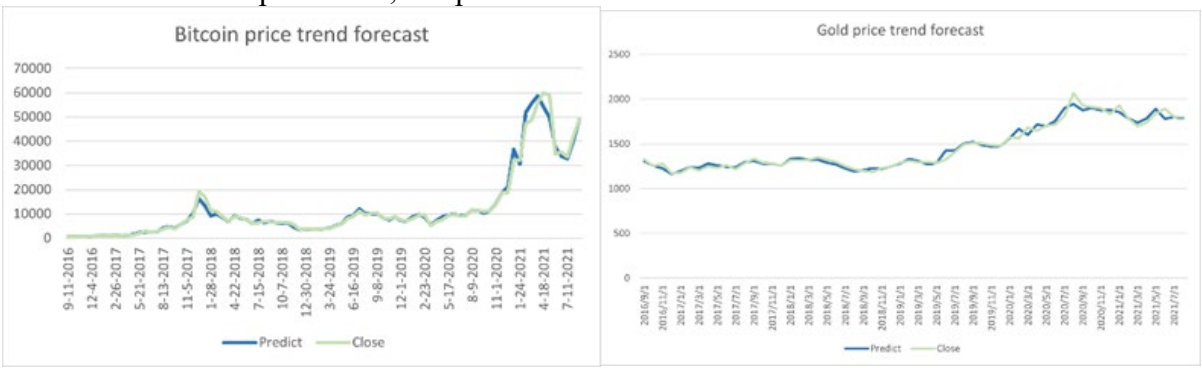

Figure 5 Comparison of solved results with actual values

In summary, the fit of our prediction results obtained using Random Forest is pretty good, and all the fitted indicators are excellent, especially the goodness-of-fit R^2 reaches 0.98.

3. Investment Strategy Based on Genetic Algorithm

We established constrained equations to determine the Trading Strategy.

Use W_n to represent the total value on the trading day T_n. $w_{T_n}^g$, $w_{T_n}^b$ and M_{T_n} denote the asset values of gold, bitcoin, and cash on the trading day, respectively T_n:

$$W_n = w_{T_n}^g + w_{T_n}^b + M_{T_n} \quad (17)$$

We need to maximize W_{n+1} at T_{n+1} by trading. Thus, the objective function model is established as follows:

$$MAX \quad W_{n+1} = \beta_g\left(w_{T_n}^g + c_{T_n}^g\right) + \beta_b\left(w_{T_n}^b + c_{T_n}^b\right) + M_{T_{n+1}} \tag{18}$$

$$M_{T_{n+1}} = M_{T_n} - \left(c_{T_n}^g + c_{T_n}^b\right) - \left(0.01 \times \left|c_{T_n}^g\right| + 0.02 \times \left|c_{T_n}^b\right|\right) \tag{19}$$

$$\beta_g = \frac{p_{T_{n+1}}^g}{p_{T_n}^g}, \beta_b = \frac{p_{T_{n+1}}^b}{p_{T_n}^b} \tag{20}$$

It is represented the ratio of gold to bitcoin T_{n+1} and T_n price respectively, and the yield of gold and bitcoin on T_{n+1} were calculated as $\beta_g - 1, \beta_b - 1$. $c_{T_n}^g$ and $c_{T_n}^b$ represent the transaction value of gold and bitcoin on the trading day, respectively, on the day T_n. Considering the market trading mechanism, we establish constraint equations:

$$\begin{cases} -w_{T_n}^g - w_{T_n}^b < c_{T_n}^g + c_{T_n}^b < M_{T_n} \\ -w_{T_n}^g < c_{T_n}^g < M_{T_n} - c_{T_n}^b \\ -w_{T_n}^b < c_{T_n}^b < M_{T_n} - c_{T_n}^g \end{cases} \tag{21}$$

$\left(c_{T_n}^g + c_{T_n}^b\right)$ is the total transaction value. Its value is the amount of gold and bitcoin traded. $0.01 \times \left|c_{T_n}^g\right|, 0.02 \times \left|c_{T_n}^b\right|$ are the transaction costs of buying and selling. The cash of the current day (M_{T_n}) is deducted from today's total trading volume, and the cash for the next day can be obtained.

When gold is on a non-trading day, the functions are as follows:

$$MAX \quad W_{n+1} = w_{T_n}^g + \beta_b\left(w_{T_n}^b + c_{T_n}^b\right) + M_{T_{n+1}} \tag{22}$$

$$M_{T_{n+1}} = M_{T_n} - c_{T_n}^b - 0.02 \times \left|c_{T_n}^b\right| \tag{23}$$

Where the constraint function is as follows:

$$-w_{T_n}^b < c_{T_n}^b < M_{T_n} \tag{24}$$

J. Holland proposed the genetic Algorithm (GA) in 1975, inspired by biological evolution. GA is a highly parallel, random, and adaptive optimization algorithm based on the survival of the fittest. It represents the solution of the problem as the survival process of the 'chromosome.' The continuous evolution of the generation of the 'chromosome' group, including replication, crossover, and mutation, finally converges to the individual of the 'most suitable environment' to find the optimal solution.[10]

Using the model, we obtain the following graph of the trend of the total asset, and the Final total assets were 243,424.76, with annual yields of 60.86%.

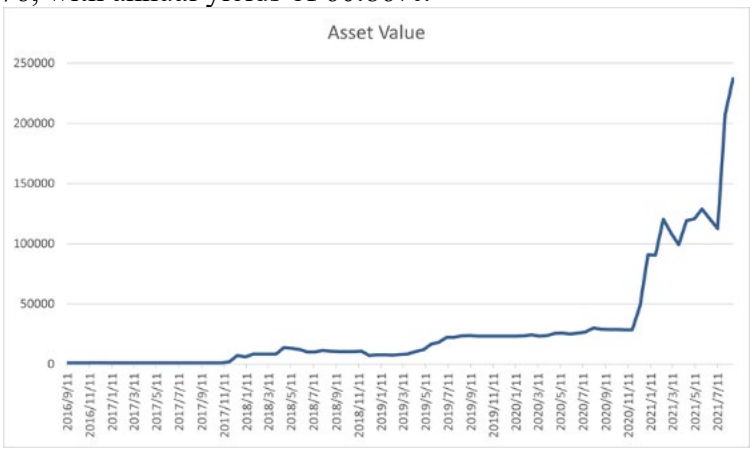

Figure 6 Total assets curve before adding risk indicators

4. Conclusion

This paper uses the past daily price of gold and bitcoin to calculate the Bollinger band, DMA, and MACD characteristics. We used the Random Forest algorithm to predict the next day's price and established a loop to predict all price data. We intuitively see that the prediction effect is pretty good from the predictive value-real value-line chart. Then, we established the transaction decision-making objective model, used the Genetic Algorithm to select the daily transaction amount, obtained the optimal transaction strategy, and then established a loop to obtain all the optimal strategies. The final total assets were 243,424.76, and the annual return rate reached 6085.62%. We used a random forest algorithm to predict. Compared with the traditional prediction algorithm, the prediction accuracy is greatly improved. However, when the total assets are small, they may be limited by the investment unit and cannot be traded accurately.

References

[1] BROCK W, LAKONISHOK J, LEBARON B. Simple technical trading rules and the stochastic properties of stock returns [J]. The Journal of Finance, 1992, 47(1): 1731-1764.

[2] LAI M M, LAU S H. The profitability of the simple moving averages and trading range breakout in the Asian stock markets [J]. Journal of Asian Economics, 2006, 17(1): 144-170.

[3] MARSHALL B R, NGUYEN N H. VISALTANACHOTI N. Time series momentum and moving average trading rules [J]. Quantitative Finance, 2017,17(3): 405-421.

[4] RAD H, LOW R K Y, FAFF R. The profitability of pairs trading strategies: Distance, cointegration and copula methods [J]. Quantitative Finance, 2016, 16(10): 1541-1558.

[5] B.S. Bini, Tessy Mathew. Clustering and Regression Techniques for Stock Prediction[J]. Procedia Technology,2016,24(C):

[6] Ding Y, Cheng L, Pedrycz W, et al. Global Nonlinear Kernel Prediction for Large Data Set With a Particle Swarm-Optimized Interval Support Vector Regression[J]. IEEE Transactions on Neural Networks & Learning Systems, 2017, 26(10):2521-2534.

[7] Wen M, Li P, Zhang L, et al. Stock Market Trend Prediction Using High-order Information of Time Series[J]. IEEE Access, 2019:28299-28308.

[8] Bin Li, Yan lin, Wenxuan Tang. ML-TEA: A set of quantitative investment algorithms based on machine learning and technical analysis[J]. Systems Engineering Theory and Practice. 2017, 37 (5)： 1089-1100.

[9] Krauss, C, X. A. Do,and N. Huck. Deep Neural Networks, Gradient -boosted Trees, Random Forests: Statistical Arbitrage on the S&P 500[J]. European Journal of Operational Research, 2017, 259（2）: 689-702

[10] He Juan, Luo Guangyi, Yang Sha, Zeng Lijuan, Kang Shuai. Research on Quantum Entanglement Genetic Algorithm and Its Optimization Performance [J]. Information Technology and Informatization, 2021(12):27-30.'