

Research on prediction model of optimal trading strategy- Taking bitcoin and gold as an example

Yuze Zhao^{1,#}, Lanqing Zhang^{2,#}, Siyang Liu^{2,#}

¹ School of Shipping Economics and Management, Dalian Maritime University, Dalian, China

² Faculty of Information Science and Technology, Dalian Maritime University, Dalian, China

These authors contributed equally to this work.

Abstract. In order to determine the best trading strategy between gold and bitcoin, this paper develops a prediction model. Firstly, the algorithm selects the optimal algorithm from SVM algorithm, logistic regression algorithm, KNN algorithm, random forest algorithm and decision tree algorithm. Secondly, sensitivity analysis is used to verify the feasibility and correctness of the model. Finally, the best trading strategy of gold and bitcoin is determined according to qualitative analysis and quantitative calculation.

Keywords: The best trading strategy; Gold and bitcoin; Sensitivity analysis.

1. Introduction

Today, the world financial market is more active than ever before, and a large number of transactions are carried out in this virtual market every day and every hour. Market traders always like to buy, hold and sell stable assets to maximize their final return. Trading strategies have many common practices. Machine learning technology is a popular method to predict the stock market recently [1].

Taking gold and bitcoin as examples, this paper develops a transaction prediction model to provide guidance for traders. Firstly, the algorithm selects the optimal algorithm from SVM algorithm, logistic regression algorithm, KNN algorithm, random forest algorithm and decision tree algorithm. Secondly, sensitivity analysis is used to verify the feasibility and correctness of the model. Finally, the best trading strategy of gold and bitcoin is determined according to qualitative analysis and quantitative calculation.

See Table 1 for the definition of symbols in this paper.

Table 1. The definition of symbols

Symbol	Description	Unit
	The missing sample data ($i=2$)	
x_i	The missing samples of the first two days and the last two days of the sample ($i=0,1,2,3,4$)	-
y_i	The missing data in row i of the table of gold	\$
sum	Sum of money	\$
inv	Investible amount	\$
gua	Guaranteed amount	\$
gd	Gold stock	-
bt	Bitcoin Stock	-
gro	Predicting the growth probability	-
grog	Predicting the probability of gold growth	-
grob	Predicting the probability of bitcoin growth	-
byg	Amount used to purchase gold	\$
byb	Amount used to purchase bitcoins	\$
ac	accuracy	-
δ	Relative error	-
ε	The extent to which the total return on investments after changing commissions deviates from the previous total return	-

2. The best daily trading strategy

The model is built and validated based on the following data sources: daily gold prices from the London Bullion Market Association; daily bitcoin prices from Nasdaq; Lbma-gold.csv; Bchain-mkpru.csv [2,3].

2.1 Processes and Methods

Processing data by Newton interpolation method (LBMA-GOLD) [4]. See Table 2 for data processing results.

Table 2. The missing data derived by Newton interpolation

	$x_0=1$	$x_1=2$	$x_2=3$	$x_3=4$	$x_4=5$
y_{76}	1133.65	1131.35	1130.7	1134.6	1145.9
y_{79}	1134.6	1145.9	1148.1	1151	1164.25
y_{328}	1264.55	1264.55	1270	1279.4	1291
y_{331}	1279.4	1291	1303	1312.05	1314.9
y_{581}	1259.75	1258.15	1261	1268	1279
y_{584}	1268	1279	1281.5	1282.9	1290.45
y_{834}	1479	1482.1	1496.8	1511.5	1514.75
y_{837}	1511.5	1514.75	1517.9	1527.1	1548.75
y_{1088}	1877.1	1875	1872.1	1874.3	1887.6
y_{1091}	1874.3	1887.6	1918.1	1943.2	1940.35

The data were randomly divided into training set and evaluation set in the ratio of 8:2.

2.2 Processes and Methods

SVM algorithm, logistic regression algorithm, KNN algorithm, random forest algorithm and decision tree algorithm are used for multiple training on the training set respectively [5-9]. Through simple cross validation, the observation data are randomly selected repeatedly for model training and validation.

Determine the ratio of the investable amount (*inv*) and the guaranteed amount (*gua*) based on the accuracy of the algorithm (*ac*).

$$inv = sum \cdot ac, gua = sum \cdot (1 - ac) = sum - inv \quad (1)$$

If the failure of prediction leads to loss, the guarantee amount will be redistributed. For example, the accuracy rate is 90%. Assuming that there is \$1000, the investment amount will be \$900, and the principal guaranteed capital of \$100 will be retained. Unless otherwise specified, the distribution amount specified later is the investable amount.

Based on the characteristics that the problem has certain requirements for the prediction of all three states (growth, flat and decline), the optimal algorithm is selected based on the accuracy value after multiple iterations.

The algorithm predicts $gro > 60\%$ as growth, $40\% < gro < 60\%$ as flat, and $gro < 40\%$ as down.

For fund allocation, the algorithm predicts the growth probability of bitcoin and gold on that day, which is discussed in the following scenarios.

(1) If both are determined to be growing, the funds are allocated based on the ratio of growth probability $\frac{grog}{grob}$.

$$byg = inv \cdot \frac{grog}{grog + grob} \quad (2)$$

$$byb = inv \cdot \frac{grob}{grog + grob} \quad (3)$$

(2) If it is determined that one grows the other falls or remains flat, then the amount to be invested is determined by the probability of growth $grog$, and the falling one is not purchased.

$$byg = grog \cdot inv \quad (4)$$

$$byb = grob \cdot inv \quad (5)$$

(3) If both sides do not intend to grow, they will not buy on the same day. About when to sell. If it is predicted that there will be a decline tomorrow, and the price of gold or bitcoin on that day is higher than yesterday, then we will allocate the sales volume according to the probability of decline $(1-gro)$.

Sell volume are as:

$$ac \cdot gd \cdot (1 - grog) \text{ or } ac \cdot bt \cdot (1 - grob) \quad (6)$$

Because gold can only be bought and sold on market trading days, none gold trading days remain on the sidelines, do not do any gold trading, and only analyze bitcoin.

Subsequently, the model is used as the basis for printing confusion matrix, calculating accuracy and recall, drawing ROC curve and calculating AUC value.

Finally, the model is used to predict the daily stop loss of gold and bitcoin from September 11, 2016 to September 10, 2021.

For the fund allocation after predicting the stop loss scenario, we will allocate the fund according to the growth probability of gold and bitcoin through the weighting method according to the accuracy of the algorithm, and predict the data of tomorrow with the data of the day and before according to the problem.

Therefore, if the forecast is growth, we will choose to buy some today; If the forecast drops, we will choose to sell the part today; If the forecast is flat, we will choose to wait and see.

2.3 Validate model performance

Make confusion matrix (Fig. 1).

The accuracy rate is specific to the predicted outcome and indicates how many of the samples with positive predictions are actually positive samples. The calculation formula is:

$$P = \frac{TP}{TP + FP} \quad (7)$$

After substituting the data, the result is calculated as 0.823.

The recall rate is sample specific and indicates how many positive cases in the sample were correctly predicted. The calculation formula is:

$$P = \frac{TP}{TP + FN} \quad (8)$$

After substituting the data, the result is calculated as 0.841.

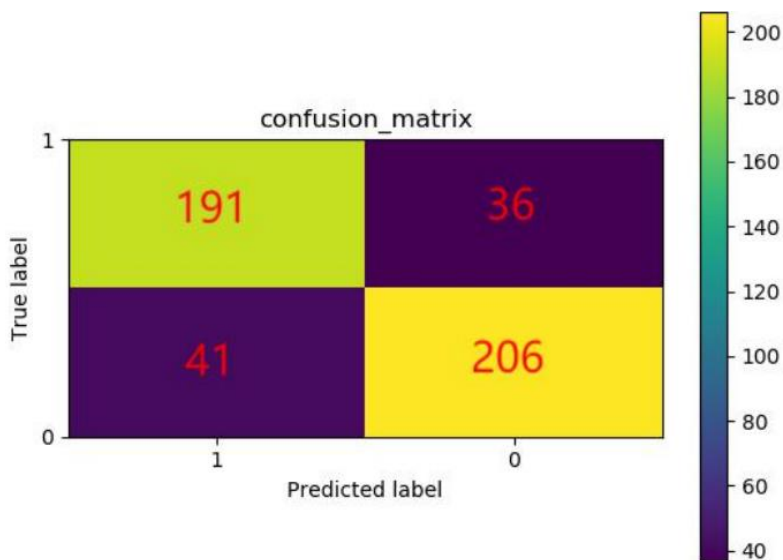


Fig. 1 Confusion matrix of the model on the test set

The accuracy rate indicates how many correct predictions were made out of all the predicted results. The calculation formula is:

$$P = \frac{TP + TN}{TP + FP + TN + FN} \quad (9)$$

After substituting the data, the result is calculated as 0.861.

The precision rate indicates how many of the samples predicted to grow are actually growing, and the recall rate indicates how many of the growths in the sample are correctly predicted. Since the requirement of the question is to predict growth as correctly as possible, it is clearly biased to assess the model's merit by accuracy or recall alone. Introduce the accuracy metric to select the optimal model for training. In our program, the model accuracy fluctuates between 0.83 and 0.86, and is able to correctly predict about 85% of the data.

The final prediction is that the initial \$1,000 investment on September 10, 2021 will be worth \$74,821.33.

3. Validate model performance

After multiple training on the training set by the SVM algorithm, logistic regression algorithm, KNN algorithm, and random forest algorithm, they are validated on the validation set, and the optimal model trained by each algorithm is taken to make their ROC curves and AUC curves.

ROC curves are often used to evaluate models because they have good properties. When the distribution of positive and negative samples in the test set changes, the ROC curve can remain unchanged. Class imbalance often occurs in real data sets, where there are many more negative samples than positive samples (or vice versa), and the distribution of positive and negative samples in the test data may change over time. The ROC curves of the above algorithms are shown in the following Fig. 2.

AUC is defined as the area under the ROC curve and takes a value generally between 0.5 and 1. The AUC value is used as an evaluation criterion because often the ROC curve does not clearly indicate which classifier works better, but as a value, the classifier corresponding to a larger AUC works better (Fig. 3).

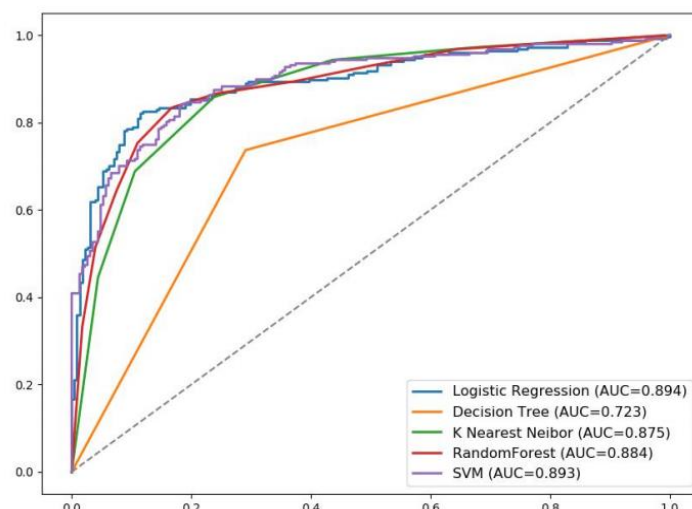


Fig. 2 ROC curves

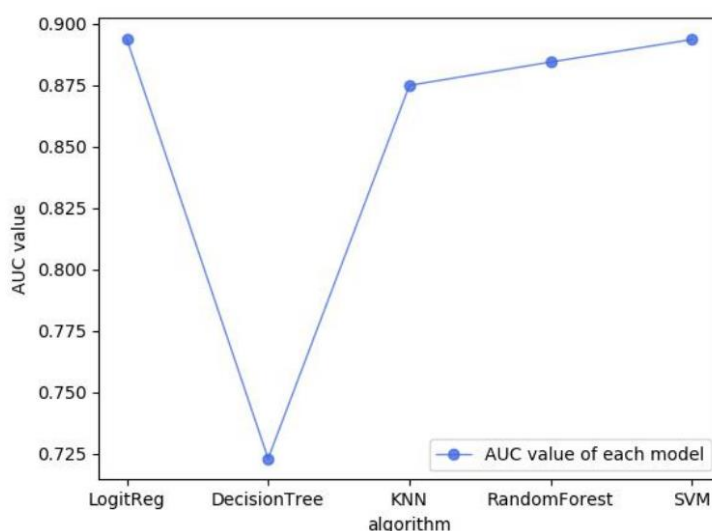


Fig. 3 AUC curves

Finally, five models are tested, and SVM algorithm is finally determined to be used as the model to predict the final result, with an accuracy of 85%.

4. The evidence of the best strategy

Test the sensitivity of the model to transaction costs [10-12]. In our model, we consider the transaction cost to be the commission we pay for buying gold and bitcoin, and we set the commission parameter to a constant in the model.

Therefore, modify the commission parameter in the model was built at the beginning. In the model above, set the commission parameters for gold and bitcoin to $agold=1\%$ and $abitcoin=2\%$ respectively.

So when modifying the parameters of gold and bitcoin, modified them in a ratio of 1:2 steps respectively. That is, the commissions for gold and bitcoin respectively increased by 0.25% and 0.5% twice, and then decreased by 0.25% and 0.5% twice for gold and bitcoin, the total change range for gold and bitcoin was $1\% \pm 0.5\%$ and $2\% \pm 1\%$ respectively, which is a total of four sets of data tested.

Then the comparison is made with the original data respectively. The data from the tests and the original data are shown in the following Table 3. In particular, group 3 is the original data, and groups 1, 2, 4, and 5 are the test data after changing the commission.

Table 3. Impact of gold and bitcoin commission rate changes on total investment returns

Groups of data	1	2	3	4	5
Gold commission rate	0.5%	0.75%	1%	1.25%	1.5%
Bitcoin commission rate	1%	1.5%	2%	2.5%	3%
Total investment returns	82302.25	79054.38	74821.33	71937.28	67518.17

The change in total revenue after the change in commission is shown in Fig. 4.

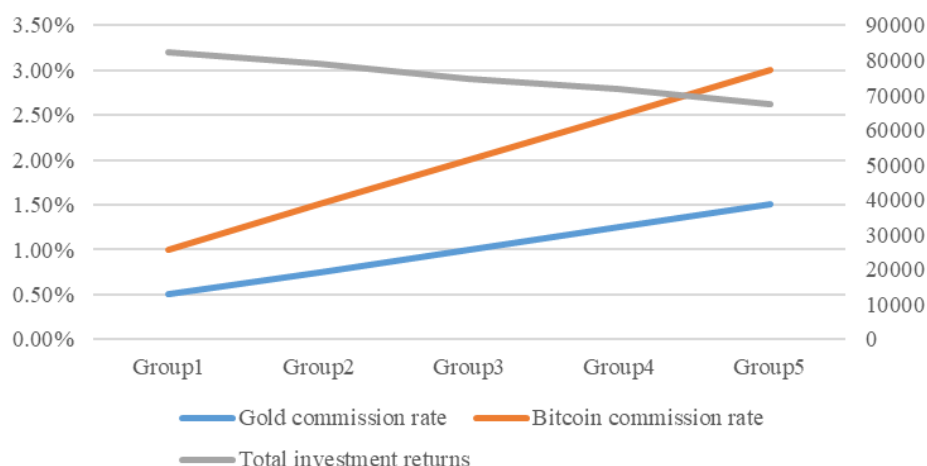


Fig. 4 Total returns with gold and bitcoin commission rate chart

From the Fig. 4, the total return on investment does not change much and the trend is relatively smooth, which indicates that our model is reasonably designed with the original commission parameters. This indicates that our model is reasonably designed with the original commission parameters and that the total return shows a de-creasing trend as the commission continues to increase.

In order to describe more precisely the change in the total return on investment after a change in commission, start with the analogy of the concept of relative error among physics experiments.

In physics, call the relative error the value obtained by multiplying the ratio of the absolute error caused by the measurement and the true value of the measured by 100%, which is generally expressed as a percentage of.

$$\delta = \frac{|measure\ value - actual\ value|}{actual\ value} \times 100\% \quad (10)$$

The magnitude of the relative error often reflects the extent to which the measurement deviates from the actual value.

Analogously to our problem, we can also introduce a quantity ε in the same way to better describe the extent to which the total return on investment after changing the commission deviates from the total return on investment when the original commission is given.

$$\varepsilon = \frac{|Returns\ after\ changing\ commission - Original\ total\ returns|}{Original\ total\ returns} \times 100\% \quad (11)$$

Separately use $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5$ to express the ε_1 values of the 1st, 2nd, 4th, and 5th test data. The calculated results are shown below: $\varepsilon_1=9.9\%, \varepsilon_2=5.7\%, \varepsilon_4, =3.9\%, \varepsilon_5=9.8\%$.

From the above data, can be seen that the greater the difference between the transaction cost and the original transaction cost, the greater the degree of deviation of the final total return of the investment from the original return.

If the absolute value is removed from Eq. (11), the ε is negative. As the commission continues to grow, the total return on investment decreases. It can be concluded that if the changed commission is

higher than the original commission, the rate of return will be lower, and the more cautious you need to be when buying gold and bitcoin. If the changed commission is lower than the original commission, the total return will be higher, and you can choose to sell your gold and bitcoin.

5. Conclusion

After a series of training and verification, the SVM algorithm with the best performance is finally selected as our model. At the same time, the decision-making indicators choose a more cautious investment mode, which can steadily increase capital and have a certain risk resistance. In the sensitivity test, the model also has good sensitivity to transaction costs, which can ensure that correct decisions can still be made in the case of market shock and macro mobilization.

References

- [1] Chenkai Zhao. 2019. Comparing the Effectiveness of Multiple Quantitative Trading Strategies. In Proceedings of the 11th International Conference on Computer Modeling and Simulation (ICCMS 2019). Association for Computing Machinery, New York, NY, USA, 47–51.
- [2] <https://www.lbma.org.uk/>
- [3] <https://www.nasdaq.com/>
- [4] Ying Yang, zefei Yan, Ying Hua. Application of interpolation method and fitting method in infectious diseases [J] Science and technology information, 2021,19 (22): 183-186.
- [5] Tao Xi, Weixiong Xu, Zongshuai Gao, Lijing Wang. Input voltage control model of construction elevator based on Rf-Ga-Svm algorithm [J] Journal of Tianjin University of technology, 2022,41 (02): 60-66.
- [6] Fuxin Wang, Xiaoman Jiao, Yantong Shi. Application of binary logistic regression algorithm in cosmetic efficacy evaluation [J] Brand and standardization, 2022 (01): 60-62.
- [7] Jifang Jin, Chengyuan Liu, Xiaohong Fan, Xiaoyi Duan, Jiayu Liu. Comparative study on energy analysis attack based on template and KNN algorithm [J] Telecommunications Science, 2022,38 (04): 121-129.
- [8] Xin Zhang, lincui Zhong, Jun Wu, Yanjing Hu, Xiaomin Song, Jingchun Song. Risk factors analysis of thrombotic events in trauma patients based on random forest algorithm [J/OL] PLA medical journal: 1-11 [May 15, 2022].
- [9] Shuhui Xu, Xin Zhang. Application progress of decision tree algorithm in nursing field [J] Nursing and rehabilitation, 2022,21 (04): 92-94 + 98.
- [10] Rongrong Chen, Zheng Ding. Ecological sensitivity analysis of Quanzhou City Based on GIS [J] Environment and development, 2022,34 (04): 100-109 + 116.
- [11] Defu Zhang, Xiongda Wang. Simplified mechanism of heavy oil substitutes based on sensitivity analysis [J] Internal combustion engines and accessories, 2022 (07): 39-41.
- [12] Zelin Ding, Yupeng Gao, Hongyang Zhang, Jing Wang. Research on hydraulic structure model materials based on sensitivity analysis and BP neural network [J] Hydropower energy science, 2022,40 (03): 158-161.