Stock Investment and Trading Strategy Model Based on Autoregressive Integrated Moving Average

Litao Liu
Institute for Disaster Management and Reconstruction, Sichuan University
Pittsburgh Institute, Sichuan University
Chengdu, China
1569589244@qq.com

2022 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS) | 978-1-6654-7053-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/TOCS56154.2022.10015924

***Abstract*—Gold and Bitcoin have long been popular among investors as traditional and newer assets that can fight inflation and preserve value, respectively, which has led to an increasing number of investors interested in investing in Bitcoin and gold. This paper focuses on a trading decision model based on time series analysis combined with innovative bull and bear market forecasts. By comparing the forecasting effectiveness of Autoregressive Integrated Moving Average (ARIMA) and XGBoost neural network, we decided to choose the more effective ARIMA model for stock price forecasting. ARIMA models combined with bull and bear market forecasting models for conservative stock traders analyze the risk and return of trading and propose trading decisions. In addition to this, we verified the optimality of the developed model and analyzed the sensitivity of the model. Adjustments were made to the gold, bitcoin and cash positions on 23 different dates through the trading decision methodology we provided. Ultimately, the model we built is expected to deliver around 260 times profitability over the dates of the data used.**

Keywords*—*ARIMA, XGBoost, Regression-prediction, Trading Strategy

I. Introduction

Gold has long been favored by investors as a traditional, inflation-proof asset that can fight inflation and preserve its value. For several years, Bitcoin has sometimes been referred to by economics experts as a new type of gold, a safe-haven asset that can replace gold in the fight against inflation. This has led to an increasing number of investors interested in investing in bitcoin and gold [1, 2].

Assets with volatile values tend to attract investors with their high returns, and these investors always expect to achieve more rewards with each decision. However, the instability of market prices predisposes the decision to a certain level of risk. Therefore, to gain more benefits and avoid asset losses, it is necessary for investors to forecast market prices. Investors can explore the pattern of price changes based on market price data, make reasonable forecasts and decide their investment strategy [3]. To get an optimal trading strategy, the present model is built essentially as a portfolio model that contains three main parts: forecasting, planning and optimal strategy test.

In this paper, we plan to use time series analysis to regress market value forecasts for gold and bitcoin to explore the application of Autoregressive Integrated Moving Average (ARIMA) model time series models in the field of financial investment. That is, a forecast of the market value of the assets of the day on a certain day based on the data of the previous days, and the adjustment of the asset holdings based on the forecast value to maximize the return.

II. Method

A. Auto-regressive Integrated Moving Average (ARIMA) Model

As we mentioned above, the ARMA(p,q) model is a combination of the AR(p) model and the MA(q) model, and the ARIMA model differs the data on the basis of ARMA. It is worth noting that the ARMA model can only deal with smooth time series. By differencing the original data in a certain number of times, the non-stationary time series have the opportunity to be converted into a stationary time series, and thus the data can be analyzed. Thus, the ARIMA model allows for the analysis of non-stationary time series. The AR(p) model is shown below.

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_p y_{t-p} + \varepsilon_t \quad (1)$$

where is a white noise series with variance . The essence of the AR model is to treat the 1st to pth order lagged terms as independent variables for regression. In the following, the qth order moving average model MA(q) will be shown [4].

$$y_t = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \cdots + \beta_q \varepsilon_{t-q} \quad (2)$$

In order to be able to handle non-stationary time series, we need to introduce differencing to convert ARMA to ARIMA model. Combining the difference terms, we can obtain the formula for the general ARIMA model.

$$y_t^{'} = a_0 + \sum_{i=1}^{p} a_i y_{t-i}^{'} + \varepsilon_t + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i} \quad (3)$$

where $y_t^{'} = \Delta^d y_t = (1-L)^d y_t$.

Therefore, $\mathrm{ARIMA}(p,d,q)$ can be expressed as

$$(1-\sum_{i=1}^{p} a_i L^i)(1-L)^d y_t = a_0 + (1+\sum_{i=1}^{q} \beta_i L^i)\varepsilon_t \quad (4)$$

B. Mathematization of Trading Decision-Planning Models

According to our decision planning model, it is clear that the Overall Evaluation Index (OEI) plays an important role in the first round of determination. The definition of OEI is the ratio of benefit index to risk index, and the representation to calculate the composite asset evaluation index for gold and bitcoin are as follows.

$$OEI = \frac{BI}{RI} \quad (5)$$

where BI and RI are asset valuation indices for the last 30 trading days.

BI is defined as an asset Benefit Index (The ratio of the difference between the forecast price and the current price to the current price, and when the benefit index < 0, the BI is set to 0) and RI is defined as a Risk Index (The product of the variance of the last 30 trading days closing value and the rate of decline). They can be represented by the following equations respectively.

$$BI = \frac{(P' - P)}{P} \quad (6)$$

The calculation of the risk index involves a new concept-Decline Rate (DR-The ratio of the number of days down in the last thirty days to the number of days traded in the last thirty days), and they are given by the following equations.

$$RI = \frac{1}{n}[(P_1 - \overline{P})^2 + (P_2 - \overline{P})^2 + \cdots + (P_{30} - \overline{P})^2] \times DR \quad (7)$$

Then it is a case of making a determination of the increase based on data from the approaching 30 trading days and making the appropriate decisions based on the guidance. The Bitcoin Allocation Ratio (BAR -Ratio of the Bitcoin Overall Evaluation Index to the sum of the Bitcoin Overall Evaluation Index and the Gold Overall Evaluation Index) and the Gold Allocation Ratio (GAR-Ratio of the Gold Overall Evaluation Index to the sum of the Bitcoin Overall Evaluation Index and the Gold Overall Evaluation Index) are given by the following equation,

$$BAR = \frac{OEI_{bitcoin}}{OEI_{bitcoin} + OEI_{gold}} \quad (8)$$

$$GAR = \frac{OEI_{gold}}{OEI_{bitcoin} + OEI_{gold}} \quad (9)$$

III. Results And Discussion

A. Data Regression Results based on ARIMA Model

1) ARIMA (1, 1, 7) model for gold trading data

The parameter pdq is variable and the figure 1 below shows some results of the ARIMA model obtained by simulation with different parameters. Finding the right parameters is difficult, fortunately, with the SPSS Expert Modeler it is possible to find a specific ARIMA model best suited to the time-series data, making the prediction results more confident.

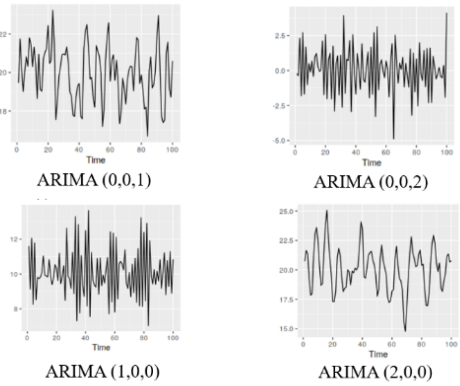

Fig. 1. The possible simulation results corresponding to ARIMA when p, d and q takes different parameters.

If the bias correlation function of the smooth series is truncated and the autocorrelation function is trailing, it can be concluded that the series is suitable for the AR model; if the bias correlation function of the smooth series is trailing and the autocorrelation function is truncated, it can be concluded that the series is suitable for the MA model; if the bias correlation function and autocorrelation function of the smooth series are both trailing, the series is suitable for the ARMA model.

We performed a time series regression analysis on the gold data. By using SPSS expert modeler, we get the best-fit model is ARIMA (1, 1, 7), i.e., the model contains three components, a first-order AR model, a seventh-order MA model, and a single difference operation to smooth the data.

After estimating the time series model, we need to perform a white noise test on the residuals. If the residuals are white noise, it means that the model we selected can completely identify the laws of the time series data, i.e., the model is acceptable; if the residuals are not white noise, it means that there is still some information not identified by the model, and we need to correct the model to identify this part of information.

We plan to analyze the ACF and PACF plots of the data residuals, as well as incorporate the Q-test proposed by Ljung and Box in 1978 to help us test whether the residuals are white noise.

From the model statistics table, we can see that the p-value obtained from the Q-test on the residuals is 0.062>0.05, i.e., we cannot reject the original hypothesis that the residuals are white noise series, so the ARIMA (1,1,7) model can identify the gold value data very well. Ultimately, the parameters of the model ARIMA (1,1,7) are obtained in the following table I and table II.

TABLE I. ARIMA (1,1,7) Model Results Statistics.

Number of predictive variables	Model fit statistics	Q Test (18)		
	Stable R-squared	Statistics	DF	Significance
0	.026	22.922	14	0.062

TABLE II. Optimal Time Series Model Parameters for Gold Value Data.

		Estimate	Standard Error	t	Significance
AR	Delay 1	0.649	0.139	4.667	0
MA	Delay 1	0.621	0.139	4.455	0
	Delay 4	0.123	0.024	5.076	0
	Delay 7	-0.093	0.023	-3.984	0

Finally, the regression value of daily gold value is obtained. Visualize the real data and the regression data to get a comparison chart, as shown in Figure 2.

Fig. 2. Comparison of actual and forecast gold prices.

2) ARIMA (0, 1, 2) model for Bitcoin value data

A similar approach was used to perform a time series analysis of Bitcoin-related data, again with the aim of obtaining daily regression values to inform investors' daily decisions. The expert modeler suggests that we use the ARIMA (0,1,2) model for Bitcoin.

It can be seen from the model statistics table III, the p-value obtained from the Q-test on the residuals is 0.450>0.05, i.e., we cannot reject the original hypothesis that the residuals are white noise series, so the ARIMA (0,1,2) model is able to identify the b-value data well. The following table IV gives the parameter settings for this model based on bitcoin data.

TABLE III. ARIMA (0,1,2) MODEL RESULTS STATISTICS.

Number of predictive variables	Model fit statistics	Q Test (18)		
	Stable R-squared	Statistics	DF	Significance
0	.238	17.057	17	0.450

TABLE IV. OPTIMAL TIME SERIES MODEL PARAMETERS FOR BITCOIN VALUE DATA.

		Estimate	Standard Error	t	Significance
Constants		0.002	0.001	2.476	0.013
MA	Delay 2	-0.066	0.022	-3.005	0.003

Finally, the regression value of daily gold value is obtained. Visualize the real data and the regression data to get a comparison chart, as shown in Figure 3.

Fig. 3. Comparison of actual and forecast bitcoin prices.

B. Trading Decision-Planning Model Results

Through computer-aided modeling, incorporating the above model development ideas into our program, in a time series analysis regression model for joint analysis, and recording the trade logs through the editor, we obtained the following table of trading strategies. According to our decision opinion, we get the final asset value of $1000 on September 11, 2016 will expand to $259,963.61 on September 10, 2021.

TABLE V. DISTRIBUTION OF ASSETS AFTER EACH TRANSACTION.

Transaction Date	Cash (USD)	Gold (troy ounce)	Bitcoin (pcs)
2016/10/21	0	0	1.55
2017/1/4	1712.46	0	0
2017/1/7	0	0	1.87
2019/5/27	4180.84	0	0
2017/5/28	0	0	1.86
2017/12/7	30214.63	0	0
2017/12/8	0	23.91	0
2018/1/29	31820.18	0	0
2018/1/30	0	0	3.06
2018/3/7	30362.04	0	0
2018/3/9	0	0	3.27
2018/4/27	0	21.66	0
2019/1/14	27714.79	0	0
2019/1/15	26946.35	0	0.21
2019/1/16	18534.01	0	2.48
2019/1/17	0	0	7.5
2019/6/27	95029.23	0	0
2019/7/1	0	67.68	0
2019/7/8	93808.16	0	0
2019/7/14	0	0	8.07
2020/11/9	122538.23	0	2
2020/11/11	0	0	7.84
2021/1/12	0	146.84	0

As shown in the table V above, we need to make 23 times adjustments over a five-year period. For example, based on the trading decision given in the table above, we need to make an asset adjustment on October 21, 2016 to bring ourselves to [C,G,B] in line with [0,0,1.55]. Then until January 4, 2017 in the asset adjustment, the adjusted assets need to meet [C,G,B] = [1712.46,0,0].

C. Optimal Strategy Testing

The model is essentially a combination of a forecasting model and a decision planning model. To prove whether the trading decision provided by the model is the best decision, the forecasting model and the decision model can be analyzed separately.

1) Selected Prediction Model Testing

In order to verify that the prediction results of the time series analysis ARIMA model are more reasonable, we used python computer-aided analysis software in combination with the XGBoost regression prediction model to perform regression prediction analysis on gold and bitcoin trading data again [5-7].

We plan to use the first 30 trading days of gold and bitcoin each as the training set and the rest of the data as the prediction set, using the following machine learning parameters, as shown in Table VI.

TABLE VI. SUMMARY OF XGBOOST REGRESSION PREDICTION LEARNING PARAMETERS.

Parameter Name	Parameter Value (Gold)	Parameter Value (Bitcoin)
Training time	0.436s	0.829s
Data Slicing	1	0.5
Data shuffle	No	No
Cross-validation	No	No
Base Learners	gbtree	gbtree
Number of base learners	100	100
Learning Rate	0.1	0.1
L1 canonical term	0	0
L2 canonical term	1	1
Sample Collection Sampling Rate	1	1
Tree feature sampling rate	1	1
Node feature sampling rate	1	1
Minimum weights of samples in leaf nodes	0	0
Maximum depth of the tree	10	10

The trained model was used to perform regression analysis on the trading data to get the daily predicted values. A comparison of the predicted and true XGBoost values for gold and bitcoin is shown below.

Tne MSE responds to the expected value of the squared difference between the parameter estimate and the true value of the parameter in parameter estimation, which effectively reflects the accuracy of the estimate relative to the true value. Using excel-assisted analysis, we obtain a comparison of the ARIMA model estimates and the XGBoost model estimates, as shown in Figure 4.

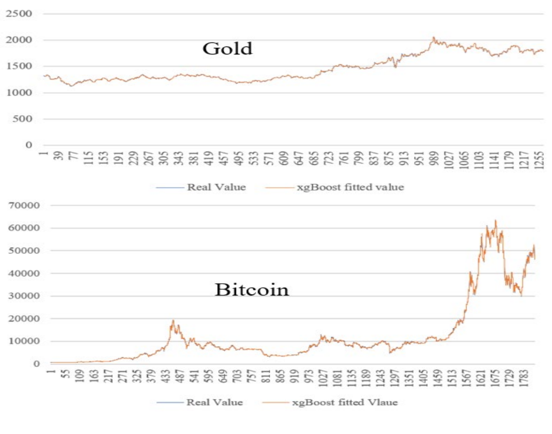

Fig. 4. Fitted vs. true values of gold (top) and bitcoin (bottom) in the XGBoost regression model.

TABLE VII. COMPARISON TABLE OF MSE VALUES CORRESPONDING TO DIFFERENT FORECASTING MODELS.

	XGBoost	ARIMA
Gold	30.49	13.79
Bitcoin	1751.81	756.26

From the data in the table VII, it can be seen that the ARIMA model is more suitable for the data analysis of this question than the XGBoost regression prediction model. Therefore, the selected prediction model is more favorable. If time allows, more models can be compared to discover the best model, but on the present basis, ARIMA belongs to the best prediction model.

2) Trading Model Testing

For the trading model, we judge an uptrend by adjusting the controllable parameter, i.e., if there are greater than or equal to i days up in the last 30 trading days. At this point i can be taken as any integer from 0 to 30. The relationship with the magnitude of the final value of i is given by the following figure 5.

As can be seen from the figure, the i=21 chosen by the model is the optimal parameter value. That is, the resulting decision is still optimal from the perspective of the decision model.

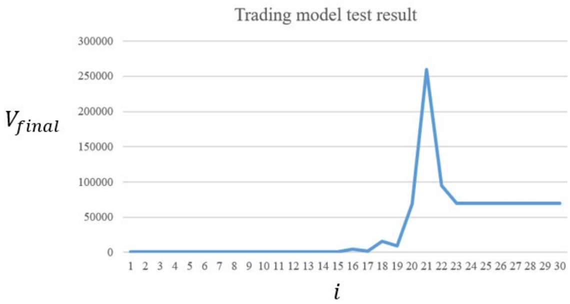

Fig. 5. Relationship between the value of i and the final value of the asset on 9/10/2021.

3) Impact of Transaction Costs on Strategy and Outcomes

To explore the impact of transaction costs on outcomes and decisions, we assign a sequence of transaction costs α_{gold} and α_{bitoin} for gold and bitcoin, i.e., from 0 to 20%, run the model again and record the output to obtain a three-dimensional plot of transaction costs versus the final value of the asset, as shown in Figure 6.

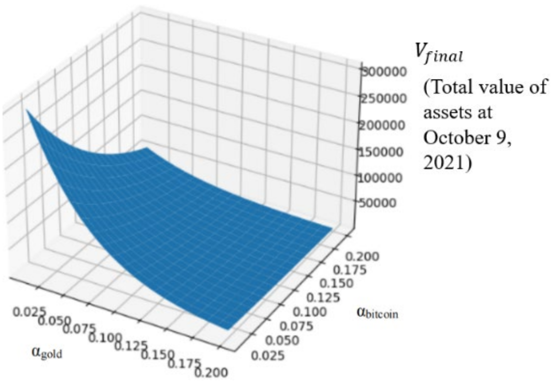

Fig. 6. Three-dimensional diagram of transaction costs versus final total value.

The image reveals that transaction costs are always detrimental to achieving the maximum return for the investor. For example, for a fixed value of α_{bitoin}, approximately smaller α_{gold} corresponds to a larger V_{final} and vice versa. We choose the case (α_{gold}=1%, α_{bitoin}=5%) and show the results of the decision in this case, as shown in Table VIII.

TABLE VIII. TRADING MODEL RESULTS (A_{GOLD}=1%, A_{BITOIN}=5%).

Transaction Date	Cash (USD)	Gold (troy ounce)	Bitcoin (pcs)
2016/10/21	0	0	1.5
2017/1/4	1609.22	0	0
2017/1/7	0	0	1.7
2019/5/27	3691.93	0	0
2017/5/28	0	0	1.6
2017/12/7	25072.77	0	0
2017/12/8	0	19.85	0
2018/1/29	26405.1	0	0
2018/1/30	0	0	2.46
2018/3/7	23676.15	0	0
2018/3/9	0	0	2.47
2018/4/27	0	15.87	0
2019/1/14	20308.92	0	0
2019/1/15	19745.82	0	0.15
2019/1/16	13581.4	0	1.76
2019/1/17	0	0	5.33
2019/6/27	65437.63	0	0
2019/7/1	0	46.6	0
2019/7/8	64596.8	0	0
2019/7/14	0	0	5.9
2020/11/9	79293.4	0	0
2020/11/11	0	0	4.92
2021/1/12	0	89.29	0

By comparing the optimal decision results in 5.3, we find that the change in transaction cost does not affect the transaction date, asset allocation, or transaction method, but only threatens the investor's ultimate interest. In this case, the investor's asset value on September 10, 2021 is $158,078.

IV. CONCLUSION

In summary, based on time series analysis and decision planning models, we have succeeded in not only giving guidance to traders on trading strategies, but also providing a more innovative decision planning model. The developed model fully takes into account the investment risks and benefits, determines the asset allocation through a certain process and thus performs asset reconciliation, and the obtained results are simple and easy to understand, which helps to guide investors in their decision making. This paper focuses on conservative traders as an example, and in fact, by modifying the model parameters, it is possible to achieve trading strategies that suit the personality of different types of traders. The sensitivity analysis, by proving the best decisions, allows us to ensure that the developed model has good stability.

The generalization of the machine learning model offers the possibility to optimize the model in this paper. Machine learning more fully exploits the historical data is what allows us to obtain more accurate predictions, with the expectation of improving the model accuracy to some extent.

REFERENCES

[1] Lv D, Huang Z, Li M, et al. Selection of the optimal trading model for stock investment in different industries[J]. PLoS ONE, 2019, 14(2).

[2] Systems Science and Mathematics, 2020, 40(7): 1270-1285 Dyhrberg A H. Hedging capabilities of bitcoin. is it the virtual gold?[J]. Finance Research Letters, 2016, 16: 139-144.

[3] Kannan K S, Sekar P S, Sathik M M, et al. Financial stock market forecast using data mining techniques[C]//Proceedings of the International Multiconference of Engineers and computer scientists. 2010, 1: 4.

[4] Enders W. Applied econometric time series[M]. John Wiley & Sons, 2008

[5] Wade C. Hands-On Gradient Boosting with XGBoost and scikit-learn: Perform accessible machine learning and extreme gradient boosting with Python[M]. Packt Publishing Ltd, 2020.

[6] Brownlee J. XGBoost with Python[J]. Machine Learning Mastery, 2019.

[7] Dong W, Huang Y, Lehane B, et al. XGBoost algorithm-based prediction of concrete electrical resistivity for structural health monitoring[J]. Automation in Construction, 2020, 114: 103155.