Research of Quantative Trading Strategy Based on LSTM

Junhao Su [a], Kaiwen Dong [b] and Ruiyang Ge [c, *]

Sichuan University, Chengdu 610064, China

[a] xue2329596062@163.com, [b] kevind0805@outlook.com, [c] ruiyangge_edu@163.com

*Correspondence Author

Abstract. Making strategies to maximize returns has always been the biggest problem almost every investor faces in the process of investment. In this paper, we will establish a model to analyze the internal characteristics of gold and bitcoin price data for further price prediction, which is of great help in formulation of trading strategy for the future. In Section 2 we hope to analyze the price trend of gold and bitcoin. At the beginning we establish a model based on ARIMA and get a relatively good prediction result. However, considering the errors may occur when the backward prediction unit gets long in this model, we try to use LSTM algorithm to better extract the nonlinear characteristics of the two time series of price. Through LSTM algorithm we get a much better result, and thus we choose LSTM as our final prediction model after comprehensive comparison. In Section 3, we use the predicted results to make trading strategies. We first divide investors into reckless and prudent types and establish models under different assumptions respectively. Then by applying greedy algorithm and establishing risk quantification model we approach the global optimal solution respectively. In the end, after testing the model and confirming its excellent performance, we summarize the advantages and shortcomings about this model comprehensively. Moreover, much more conclusions are drawn to analyze the possible model update in future.

Keywords: Price Prediction; ARIMA; LSTM; Risk quantification; Greedy Algorithm.

1. Introduction

With rapid development of finance and technology, more and more people enter the investment market, hoping to seek for the maximum profit and income by investment. Particularly, in trading market, gold and bitcoin are unique and the most popular assets.

Throughout its long and rich history, gold has been used as natural money and a store of value. It is a useful portfolio stabilizer and a source of liquidity in times of market turmoil. The empirical literature provides evidence of the traditional role of gold as a hedge against equities in normal times and as a safe haven against equities during stress periods [1]. During the period of time from 2011 to 2021, the price of gold fluctuated between $1100 per oz and $2000 per oz. Although it rose most of the past time, long-term trend wasn't found [2].

The key intention was to create a transaction system free from intervention by any central or monetary authority, be based on a mathematical algorithm instead of 'third-party trust', payments can be done electronically in a protected, verifiable and incontrovertible way [3].

Bitcoin is a cryptocurrency (or digital currency) formulated on the concept of 'peer-to-peer' network. In 2008, Nakamoto stated his new idea of e-money in Bitcoin: A Peer-to-Peer Electronic Cash System, and bitcoin announced its launch [4]. This is a virtual currency for payment with the intention of creating a transaction system free from intervention by any central or monetary authority, which is based on a mathematical algorithm instead of 'third-party trust' and payments can be done electronically in a protected, verifiable and incontrovertible way. At the same time, though, it also contributes to the characteristic that its value is determined by the market rather than itself: if the market consider bitcoin valuable, then bitcoin is valuable. Different from gold, the price of bitcoin rose from less than $100 to more than $50,000 just in the decade of 2011 to 2021. Bitcoin peaked at $69000 in November 2021 but it has fallen sharply by $23000 [2, 5]. As a result, although Bitcoin is a virtual asset that most people are optimistic about, high price volatility is inevitable, which means Bitcoin is quite risky instead of a safe haven asset.

Recently, attention has been shifted from gold in favor of a newly emerged asset, bitcoin, which is often presented as having similar properties to gold, specifically its hedging and safe haven characteristics. However, the fact is believed that most of the participants in current world's gold and bitcoin market are relatively young and inexperienced individual investors who can't make rational decisions to a large extent. Therefore, much care should be taken to predict the daily prices of gold and bitcoin, respectively, so that traders can determine their commissions for each purchase and sale, with a goal to maximize their total return as well as lower the risk.

2. Price Prediction of Gold and Bitcoin

2.1 Data Cleaning

First, we performed missing value tests on two sets of data and notice there are some gaps in the data of gold daily prices, so we choose the previous day's gold price to fill in the gaps. Then, in order to eliminate the effects of dimensionality, we standardized the data by Z-SCORE Rule:

$$y_i = \frac{x_i - \overline{x}}{s} \tag{1}$$

So strategy based on those data is much more meaningful and valuable. Before develop our model, we calculated some statistics and preliminarily recognized the characteristics of the data.

Table 1. Descriptive Statistics

Value	Sample size	Max	Min	AVG	Median
gold	1265	2067.15	1125.7	1464.141	1328.85
bitcoin	1826	63554.44	594.08	12206.07	7924.46
Value	STD	Var	Kurtosis	Skewness	CV
gold	249.432	62216.27	-1.016	0.651	0.17
bitcoin	14043.89	1.97E+08	3.095	2.013	1.151

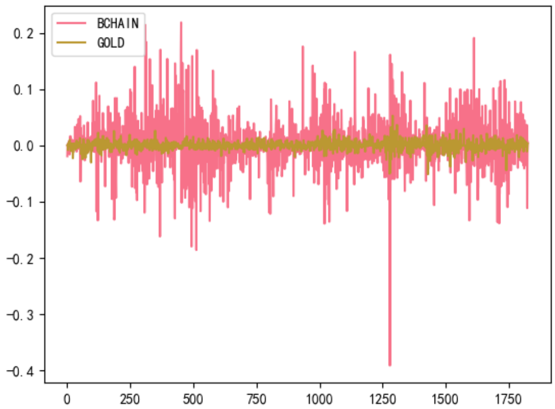

Fig. 1 Daily price fluctuation chart

From figure 1, bitcoin's price has been more volatile than gold's, which means trading strategies on bitcoin could have a bigger impact on the outcome. So in the process of building the model, we need to consider them in a targeted way.

2.2 2.2 Time Series Model-ARIMA

A time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time, thus it is a sequence of discrete-time data [6]. In order to cater for the characteristics of gold and bitcoin, we use autoregressive integrated moving average model (ARIMA) to forecast the prices of gold and bitcoin.

ARIMA (p, d, q) can be expressed in the following model:

$$\left(1-\sum_{i=1}^{p}\phi_i L^i\right)(1-L)X_t=\left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t \tag{2}$$

in which, {Xt} is a time series, L means Lag operator, E (εt) = 0, Var (εt) = σε2, Cov (εs , εt) = 0 (s ≠ t), Cov (Xs , εt) = 0 (∀s < t).

ARIMA model requires the sequence to meet the stationarity. We check the ADF test results and analysis whether it can significantly reject the hypothesis of sequence instability according to the analysis t-value ($p<0.05$).

What's more, ARIMA model requires the model to have pure randomness, that is, the residual of the model is white noise [7]. We confirm the model test table and test the white noise of the model according to the p-value of Q statistic ($p>0.05$ refer to white noise), or analysis the model based on the information criterion AIC and BIC. When the difference is order 1 and 2, the p-value is less than the given significance level a, which rejects the null hypothesis and can be considered as a stationary time series. Here are optimal difference sequence diagram (difference of order 1):

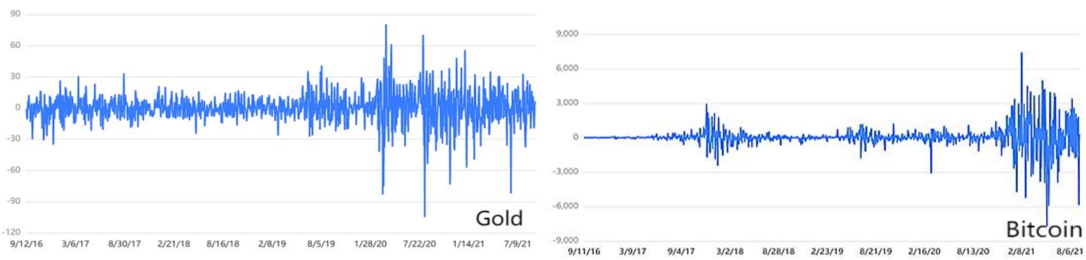

Fig. 2 Optimal differential sequence diagram of gold and Bitcoin

$R^2=\frac{\mathrm{ESS}}{\mathrm{TSS}}$ represents the fitting degree of time series, and the closer it is to 1, the better the effect is. R^2 for this model is 0.997, which can be seen from the model parameter table that the model basically meets the requirement.

Table 2. The Parameter Table of Bitcoin

	coefficient	STD	t	p	0.025	0.975
constant	0	0	1.223	0.221	0	0
ar. L1. D2. Value_z-score	-1.057	0.033	-31.765	0	-1.123	-0.992
ar. L2. D2. Value_z-score	-0.072	0.024	-3.047	0.002	-0.119	-0.026
ar. M1. D2. Value_z-score	-0.019	0.024	-0.807	0.42	-0.065	0.027
ar. M2. D2. Value_z-score	-0.981	0.024	-41.523	0	-1.027	-0.935

Table 3. The Parameter Table of gold

	coefficient	STD	t	p	0.025	0.975
constant	0	0	0.667	0.505	0	0
ar.L1.D2.USD (PM)_z-score	-0.827	0.091	-9.056	0	-1.006	-0.648
ar.L2.D2.USD (PM)_z-score	-0.003	0.029	-0.097	0.923	-0.059	0.054
ar.M1.D2.USD (PM)_z-score	-0.153	0.087	-1.764	0.078	-0.324	0.017
ar.M2.D2.USD (PM)_z-score	-0.847	0.087	-9.736	0	-1.017	-0.676

ARIMA (0,1,2) can be obtained based on AIC statistics to find the optimal parameters. Based on this model, a time series diagram containing true values, fitted values and predicted values is depicted as below.

Fig. 3 Price forecast sequence diagram of Gold and Bitcoin

Since the price data of gold and bitcoin are highly non-stationary, which does not meet the application conditions of ARIMA, the difference order is used to convert the non-stationary time series into stationary time series. When the difference order is 2, the converted time series can be judged to be stationary time series through the p value. Then through the p-value of Q statistics to gold and bitcoin price data after the conversion of the time series for white noise series, that is, the transformed sequence is suitable for ARIMA [8], through R2 and time series diagram we know that the application of ARIMA can be a good prediction of gold bitcoin price.

2.3 2.3 Recurrent Neural Networks-LSTM

The computing systems inspired by biological neural networks to perform different tasks with a huge amount of data involved is called artificial neural networks or ANN. Different algorithms are used to understand the relationships in a given set of data to produce the best results from the changing inputs.

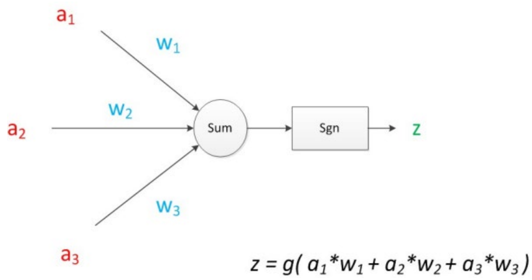

Fig. 4 Neural Network diagram

In Model 2, considering that the prices of gold and bitcoin show highly nonlinear characteristics, we first thought of applying Recurrent Neural Networks to predict prices. A recurrent neural network (RNN) is a type of artificial neural network which uses sequential data or time series data [9]. These deep learning algorithms are commonly used for ordinal or temporal problems recurrent neural networks utilize training data to learn. They are distinguished by their 'memory' as they take information from prior inputs to influence the current input and output.

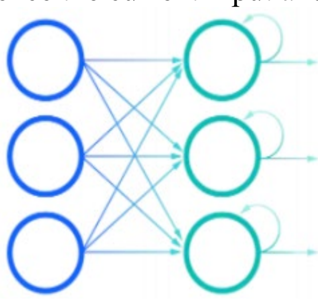

Fig. 5 Recurrent Neural Network diagram

When applying it to find the trends, we, in order to improve the reliability of the data, upgrade our algorithm to Long Short-term Memory (LSTM), which is not only suitable for extracting temporal features from time series, but also has the ability to learn long-term time series dependencies, and can solve the problem of gradient disappearance existing in Recurrent Neural Networks [10].

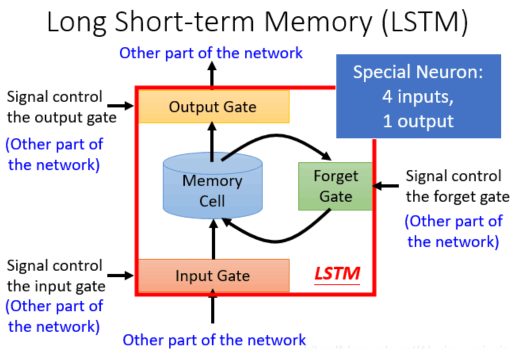

Fig. 6 The structure of LSTM

The data was converted by sliding window, the ratio of training set to test set was 9:1, and the number of training rounds was set to 400. The predicted results are as follows:

Table 4. Prediction effect of LSTM (raw data)

	MSE	RMSE	MAE	R^2
Gold	838.6093	28.95875	25.83932	0.9865
Bitcoin	2015210.4	1419.581	1253.082	0.9898

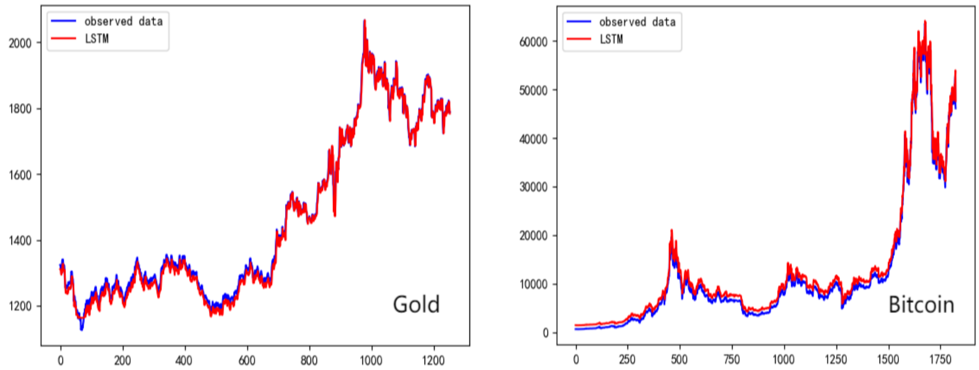

Fig. 7 Price forecast sequence diagram of Gold and Bitcoin based on LSTM

2.4 2.4 Model Contrast

First of all, the advantage of ARIMA lies in that only endogenous variables can be used to establish endogenous variables without recourse to other Exogenous variables, and the parameters that need to be considered are much less than those of LSTM, which provides great convenience for us to establish models.

However, the limitation is that ARIMA can only extract linear features in the data, while the prices of gold and bitcoin both have strong nonlinear characteristics, so once our prediction time is too long, ARIMA may have large errors.

We considered using only the 2016 to 2020 gold Bitcoin price for ARIMA's prediction, and tested it with the last year's data and found that the prediction was extremely poor.

Since LSTM adds a threshold structure that can control units on the basis of RNN, it can not only process time-series data for a long time, but also forget those unimportant characteristic information. It can solve the problem of gradient disappearance well, so as to achieve a balance between time series and nonlinearity.

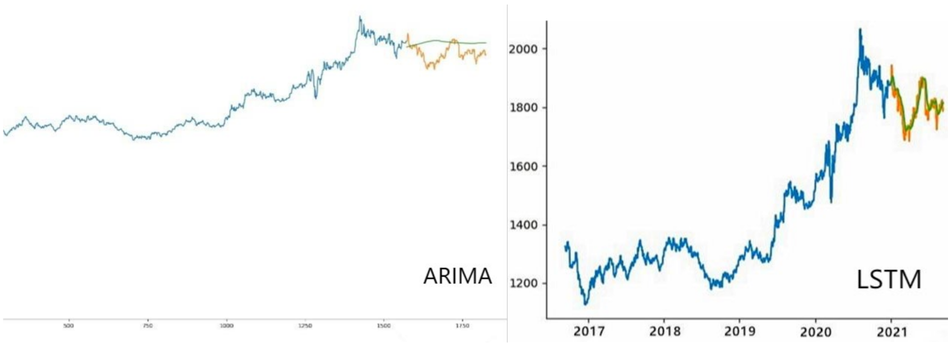

Fig. 8 Comparison between prediction trend and actual trend

3. Provide Stategy Based on Predicted Results

According to the models we developed before, we can predict the possible daily prices of gold and bitcoin depended only on price data up to that day. As we see, Model 2 fits the characteristics of the trading market better, so we choose it, to formulate the trading strategy from September 11, 2016 to September 10, 2021 to get the maximum profit. We may start with $500 in gold and Bitcoin, respectively, based on the sub-additive principle of economics that increases the average risk of a portfolio must reduce.

Normally, if gold and bitcoin go up the next day, we should buy them and sell them when they go down. First, we consider about the activist investors, which means they will trade assets immediately based on the outcome of the price forecast without caution of risk. Then we may approach the global optimal solution from the local optimal solution by greedy algorithm.

$$max\, z = \frac{n_1}{m_2}(z_1 + x_1) + \frac{n_2}{m_2}(z_2 + x_2) - \frac{x_1}{100} - \frac{x_2}{50} \tag{3}$$

$$\begin{cases} z_1 - z_2 < x_1 + x_2 < z_1 + z_2 \\ \dfrac{x_1}{100} \le \dfrac{n_1}{n_2} - 1 \\ \dfrac{x_2}{50} \le \dfrac{m_2}{n_2} - 1 \end{cases} \tag{4}$$

As we can see, the trader ends up with $2.7 million. However, it is rare for investors to take such risks in real transactions, and they usually take the risk into account, which called the steady investors. Therefore, we must incorporate risk indicators into the model. We obtained the following quantitative risk planning model:

$$max\, z = \beta(z_1 r_1 + z_2 r_2) - (1 - \beta)(z_1 q_1 + z_2 q_2) \tag{5}$$

$$\begin{cases} z_1 + z_2 = 1 \\ z_1, z_2 \ge 0 \\ \dfrac{n_i z_i}{n_i x_i} + \dfrac{n_i - m_i}{10 n_i} + \dfrac{\gamma}{10}, \forall i = 1,2 \\ \gamma = 0.44\eta + 0.55\theta \\ \eta = \dfrac{\sum_{90-days-before-trading} \dfrac{n_1 - m_1}{m_1}}{90} \\ \theta = \dfrac{\sum_{90-days-before-trading} \dfrac{m_1 - \overline{m_1}}{\overline{m_1}}}{90} \\ \overline{m_1} = \dfrac{\sum_{15-days-before-trading} m_1}{15} \end{cases} \tag{6}$$

We get the best daily trading strategy, which is, reach $18536 finally.

4. Model Validation

To prove that our model provides the best trading strategy, we assign a small perturbation to the trading strategy, such as randomly changing the trading volume of gold and bitcoin on certain days, and find that the final return decreases.

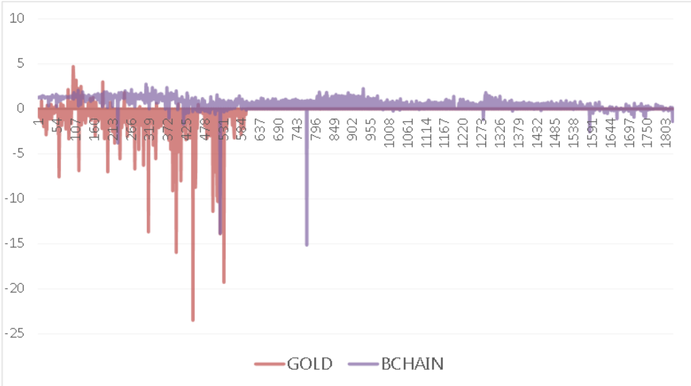

Fig. 9 Daily trading volume of people with prudent personality

At the same time, we can also see that the final income is continuous with the change of the strategy. This is because we set a high risk threshold in the process of building the model, so small changes in trading strategies have little impact on the final returns, and the model has good sensitivity.

5. Conclusion

In this paper, we used ARIMA to establish the linear model and LSTM to establish the nonlinear model for predicting the price of gold and Bitcoin. According to this forecast, investors can make decisions about the future strategies to obtain more profit. However, the data of price about gold and Bitcoin is nonlinear, which means it is a high noise series and this is why the traditional time series analysis model such as ARIMA can't reflect perfectly about the trend of price changes. Whereas LSTM is finished according to the Neural Network, it has a ability about computing the index of dependence among each searching data. In this case, LSTM can adapt quickly when the price fluctuates and this is why LSTM can work efficiently among the fluctuating time series. After determining the predicted model, we build the optimization object and the optimal policy.

The concept of quantitative investing, which uses powerful computer technology combined with efficient mathematical models to help investors to make trading strategies, has become more popular in recent years. Starting from the value of gold and bitcoin, this paper constructed an excellent model only using the value data of five years. It can be predicted that if we can add more types and quantities of data to train this model, a more perfect model will be obtained, which is filled with enthusiastic economists.

References

[1] Syed Jawad Hussain Shahzad, Elie Bouri, David Roubaud, et al. Safe haven, hedge and diversification for G7 stock markets: Gold versus bitcoin[J]. Economic Modelling. Vol. 87 (2020), p. 212-224.

[2] Baur Dirk G, Hoang Lai. The Bitcoin gold correlation puzzle. Journal of Behavioral and Experimental Finance. Elsevier. Vol. 42 (C) (2021) .

[3] P. Kayal, P. Rohilla. Bitcoin in the economics and finance literature: a survey. SN Bus Econ 1, Vol. 1 (2021), No. 7, p. 1-21.

[4] Information on: www.bitcoin.org

[5] Information on: www.120btc.com

[6] Fang Yiqiu, Lu Zhuang, Ge Junwei. Forecasting stock prices with a combined RMSE loss LSTM-CNN model[J/OL]. Computer Engineering and Applications. Vol 9 (2022), p. 294-302.

[7] Mengya Li, Juan He, Rong Zhou, et al. Research on prediction model of mixed gas concentration based on CNN-LSTM network[C]. Proceedings of 2021 3rd International Conference on Advanced Information Science and System (AISS 2021). Vol. 12 (2021) No. 1, p. 344-348.

[8] Yujie Fang, Juan Chen, Zhengxuan Xue. Research on Quantitative Investment Strategies Based on Deep Learning[J]. Algorithms. Vol. 12 (2019) No. 2, p. 1-35.

[9] Wang Zhijin, Su Qiankun, Chao Guoqing, et al. A multi-view time series model for share turnover prediction[J]. Applied Intelligence, 2022 (prepublish).

[10] Song Shuang, Li Shugang, Zhang Tianjun, et al. Research on the Multi-Parameter Modeling of Submarine Sediment Prediction[J]. Applied Mechanics and Materials. Vol. 462-463 (2013), p. 13-16.