Research on the Gold Price Forecasting Based on Machine Learning Models

Yiqi Xin

CUHK Business School, The Chinese University of Hong Kong, Shatin, Hong Kong, 999077, China

1155157055@link.cuhk.edu.hk

Abstract. In recent years, time series data analysis has been attracting much attention and can be applied to financial forecasting. As gold market can be used to manage investment risk, many methods that focus on time characteristics have been introduced to predicting the gold price. This study attempts to use auto regressive integrated moving average model (ARIMA), Decision tree model and Multi-Linear Regression model to predict the close price of gold AU99.99. The study uses root mean square error (RMSE) and R-sq to evaluate the practicability of the model. The results indicate that ARIMA (2,1,2) is not suitable to predict the price of AU99.99. Moreover, the Muti-Linear Regression model is the most suitable model for forecasting next day's close price. The effective model of this study is important to investors to understand and forecast the trend of gold market in time which raises the yield on the trade.

Keywords: Machine Learning, Gold Price, ARIMA, Decision tree, Multi-Linear Regression

1 Introduction

Gold is a kind of metal which can be very sensitive to the price change [1]. When studying changes in world finance, changes in gold prices are a good form of expression. Other assets like equities and currencies, in Corti and Holliday's view, are frequently correlated with gold prices [2]. Moreover, some indices, such as the dollar index and the Shanghai Composite Index, also have the potential to influence gold prices [3]. For investors, they should always grasp the changes in gold prices and make reasonable predictions to reduce risks [4]. There has been several research in gold prices. Megan Potoski has attempted to predict how the current day's price fix affects the London PM price fix of gold the following day with machine learning models [5].

This study is trying to figure out whether the ARIMA, Decision tree and MLR model is suitable to predict AU99.99's close price and which one has the best performance. For ARIMA Model, Dr. M. Massarrat Ali Khan demonstrated how to predict the price of gold with the traditional Box Jenkins approach [6]. Meyler and his team developed a semi-automated algorithm to fit an ARMA model to stablize time series data. By introducing the new algorithm, they predicted the Irish inflation rate [7]. According to

V. Escudero et al. (Eds.): BADS 2022, AHCS 11, pp. 1339–1346, 2023.
https://doi.org/10.2991/978-94-6463-102-9_138

Tripathy, ARIMA (0,1,1) can be the most accurate model to forecast gold prices in India [8]. Nyoni also used the ARIMA approach, which was based on minimum AIC, to study inflation in Kenya between 1960 and 2017 [9]. For the Decision tree Model, Rady and his team demonstrated how the tree-based model works in time series forecasting [10]. There have been many studies to predict precious metal price movements in this way. Navin and Vadivu introduced an implementation of using Decision tree Regression to predict gold price [11]. For Multi-Linear Regression Model, a multiple linear regression model was established for the stocks of China Citic Bank by Chen and the opening price of the stock was predicted successfully [12]. Shokri and his team estimated and predicted the global price of silver using a combined multiple linear regression (MLR) which provides the inspiration to use MLR to predict gold price [13].

Section 2 introduces data sources and model principles. Section 3 is the analysis of the three models' performance based on R-sq and RMSE. Section is the conclusion of the most suitable model.

2 Data and Model

2.1 Data

All the data comes from google websites. The daily close price of gold AU99.99 over a period of 15years from July 2007 to July 2022 was selected as the original data of study. In the cross-validation of MLR model, the first 70% of the original data was set to training data and the last 30% to testing data.

2.2 Model

ARIMA Model.

To predict time series data, the auto regressive integrated moving average model (ARIMA) is frequently utilized. The advantage of ARIMA model prediction is that it represents various types of time series, including autoregressive (AR), moving average (MA), and combinations of AR and MA(ARMA). The ARIMA model is represented by ARIMA (p,d,q), where "p" denotes the Autoregressive process, "d" denotes in which the order that the data are stationary, "q" denotes the order in which moving average is applied. The ARIMA model can be summarized in the following formula

$$\hat{y}_t = \mu + \phi_1 y_t - 1 + \cdots + \phi_p y_t - p - \theta_1 e_t - 1 - \cdots - \theta_q e_t - q \quad (1)$$

Where, *"y_t"* stands for the actual value and *"et"* stands for the random error of time period *"t"*; The model parameters are explained by ϕ_i and θ_j. Integers p and q implies the orders of the model. In the event that either q or p are both 0 then the model changes to an AR or MA model, respectively. Identification, parameter estimation, and diagnostic testing are the three processes that are always involved in developing an ARIMA model.

Decision tree Model.

A Decision tree is considered as a form of visualization containing two types of nodes: root nodes and leaf nodes. The leaf nodes contain the results. For predicting gold prices, normally two types of Decision trees are used in studies: the classification tree and the regression tree. When the category which the data belongs to is the predicted result, then it will be regarded as classification tree analysis. When the true value or number is being predicted, then it will be regarded as regression tree analysis. The Decision Tree is always a way to dig out the expected data, also can be applied for checking possible trends among various branches. While using the decision tree analysis, reducing the compellability, and selecting the features can be solved by presenting all the instances as attribute values automatically [11].

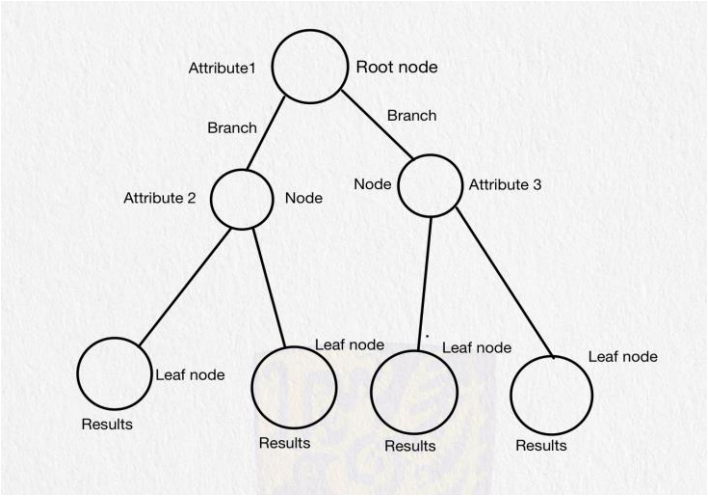

Fig. 1. Decision tree architecture

In this study, the close price of AU99.99 for the following day is predicted using decision tree regression. The specific experimental design is as follows:

1. Use the previous day's price to predict the next day's price
2. Conducting a 5-fold time series cross-validation with the training set to get average RMSE and R-sq
3. The model accuracy was evaluated by RMSE and R-sq

Multilinear Regression.

In a complex regression model called Multilinear Regression, the relationship between one component and two or more independent variables is estimated. The equation is

$$Y = a + b_1 X_1 + b_2 X_2 \tag{2}$$

"a" is a constant intercept, "b_1" and "b_2" are the coefficient. In this study, the traditional multi-linear regression is not enough for it cannot fully explain the relationship between future values and time series, so that a new version of MLR model is introduced.

The new version brings the concept of moving average in the MLR model. The model design is as follows:

1.Using the moving average of the last 3days to be the independent variable X1

2.Using the moving average of the last 7days to be the independent variable X2

3.Conducting Cross-validation which the training set includes the 70% of the original data.

4.The model accuracy was evaluated by RMSE and R-sq

3 Result Analysis

3.1 Results Based on ARIMA Model

To begin with, the original data need to be plotted to see if there exist any trends or stationarity. Time series must be changed to be stationary if stationarity is not immediately apparent by taking the first order difference of non-stationary series values [8]. Figure 1 shows the trend of the original data, and Figure 2 shows the data after first-order splitting, which proves that the original data is not a stable data, but the data looks stable after the first difference.

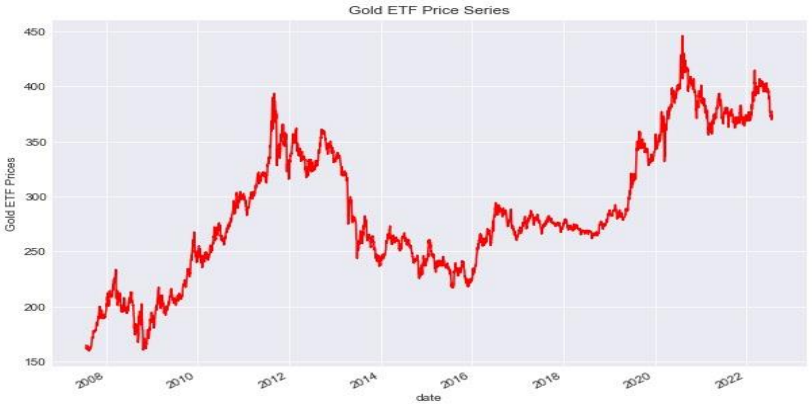

Fig. 2. Original series of AU99.99 close price

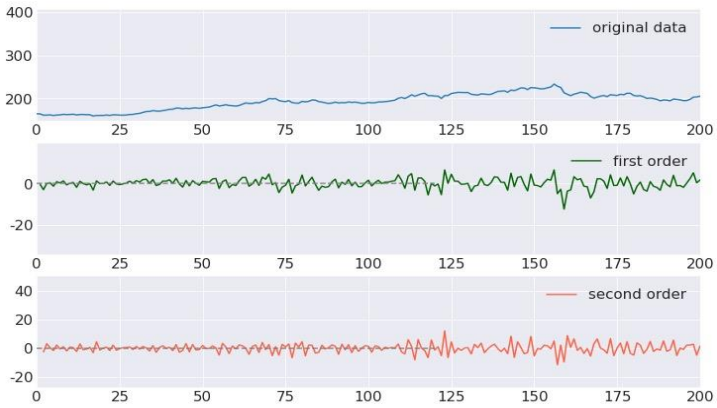

Fig. 3. First difference series data

After visualizing the stationary of the first order difference, an ADF test was conducted to make sure that the first order difference is stationary. The results show that in the hypothesis test the P-value of the first difference data is 1.6099e-28, which means that we can reject the H0 and the first order difference data is stationary.

After determining the value of d, the value of p and q need to be decided. By judging the AIC and BIC, the most suitable model for prediction can be found. The model in this experiment is chosen based on having the minimum AIC. Table 1 lists all models of AIC and BIC and their values. The lowest AIC ARIMA model (2,1,2) was selected.

Table 1. Different ARIMA models with their AIC and BIC value

MODEL	AIC	BIC	MODEL	AIC	BIC
ARIMA(0,1,0)	12560.02	12565.87	ARIMA(2,1,3)	12544.34	12579.42
ARIMA(0,1,1)	12559.47	12571.16	ARIMA(2,1,4)	12545.00	12585.93
ARIMA(0,1,2)	12558.50	12576.04	ARIMA(3,1,0)	12560.63	12584.01
ARIMA(0,1,3)	12560.50	12583.88	ARIMA(3,1,1)	12561.38	12590.61
ARIMA(0,1,4)	12560.48	12589.72	ARIMA(3,1,2)	12543.33	12578.41
ARIMA(1,1,0)	12559.30	12570.99	ARIMA(3,1,3)	12546.64	12587.56
ARIMA(1,1,1)	12560.08	12577.62	ARIMA(3,1,4)	12547.67	12594.45
ARIMA(1,1,2)	12560.50	12583.89	ARIMA(4,1,0)	12560.56	12589.79
ARIMA(1,1,3)	12561.33	12590.57	ARIMA(4,1,4)	12562.56	12597.64
ARIMA(1,1,4)	12562.42	12597.50	ARIMA(4,1,2)	12545.45	12586.38
ARIMA(2,1,0)	12558.71	12576.25	ARIMA(4,1,3)	12547.79	12594.57
ARIMA(2,1,1)	12560.68	12584.07	ARIMA(4,1,4)	12549.55	12602.17
ARIMA(2,1,2)	12542.49	12571.73			

By using the ARIMA (2,1,2) model, Figure 5 produced this model training group and test group fitting image, the fitting degree of training group is very high, with RMSE of only 4.30, but the results of the test set are not ideal, the predicted data shows

a slight decline. The reason is that the AU99.99 closing price movements had no obvious up or down trend, and it has been fluctuated in a certain range. Moreover, ARIMA is only sensitive to the linear relationship, for nonlinear relationship, it is important to use other models to make predictions.

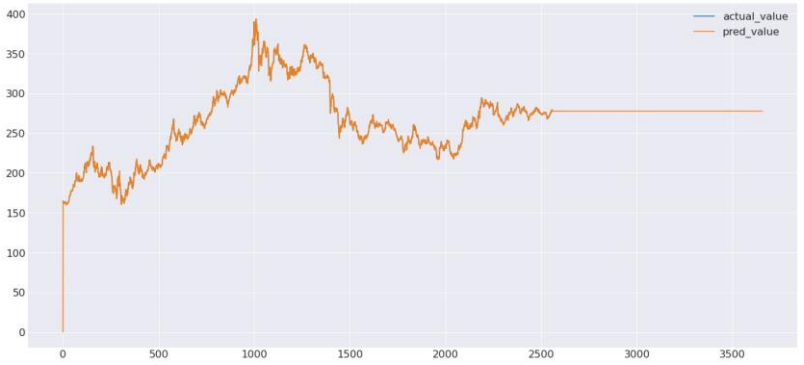

Fig. 4. Auto regressive integrated moving average (2,1,2)

3.2 Results Analysis Based on Decision tree Model

Figure 6 plots the predictions and the original data. After the cross-validation, RMSE is 1.08 and R-sq is 0.86, which proves that although the errors are small, the model can only explain 86% of the data. It can be seen from the figure that when the rising trend of data is stable, the fitting degree is high, and the error is small. The Decision tree model cannot predict the fluctuations when the data have drastic fluctuations. If prediction with smaller error is required, some other variables need to be added to the model.

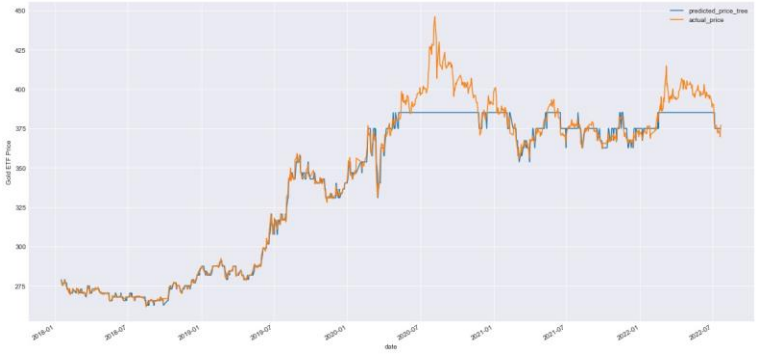

Fig. 5. Original Data and Prediction of Decision tree Regression

3.3 Results Analysis Based on Multi-Linear Regression Model.

Figure 7 shows the result of prediction and the original data. After the cross-validation, the average RMSE is 3.81 which it shows that the error in the prediction is very small. Also, the average R-sq is 0.97, which implies that the model is highly adaptive to this

kind of data, and most of the data can be accurately fitted. But due to the nature of the moving average, there always lags between the real value and the expected value, which means that when the two kinds of values are aiming to attain the same value, the predicted value always arrives later than the real value.

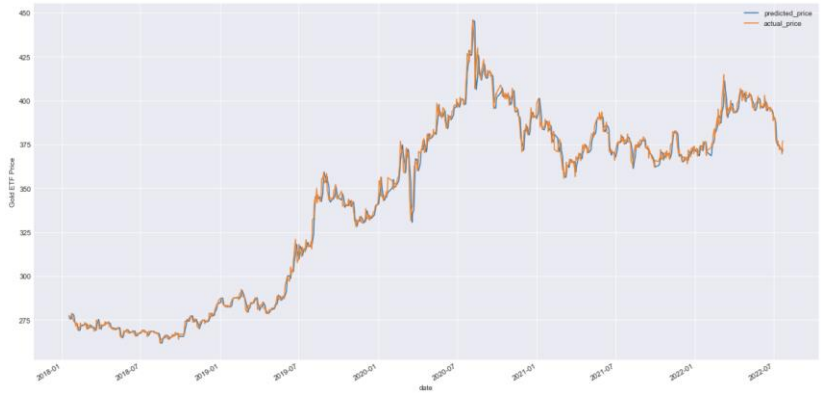

Fig. 6. Original Data and Prediction of Multi-Linear Regression Model

4 Discussion

Through the analysis and comparison of the results of the three models, the most suitable model for predicting the close price of AU99.99 is the MLR model. Although ARIMA fits the historical data perfectly, but it is not suitable to forecast the accurate price. The experimental results show that ARIMA is not ideal in predicting the closing price. The reasons are (1) The upward or downward trend of the data is not obvious. (2) ARIMA can only predict the linear regression problem, so that more variables need to be added. The Decision tree model shows the lowest RMSE in all the models, which means the value it could predict is accurate. Since there are serval gaps between the prediction line and the original data line, which implies that it is less sensitive to the fluctuations, it is still risky for making investing decision with this model. The MLR is from now on the best model and it can be used in the practical situations. With an average R-sq of 0.97, it can prove that 97% of the test data can be explained by the model which is sufficient to show the prediction accuracy and stability of the model.

5 Conclusion

The study conducts three machine learning models to predict the closing price of gold the following day and to determine its accuracy by comparing RMSE with R-sq. The experimental results show that ARIMA is not ideal in predicting the close price. The MLR model is the best model. The combination of the MA and MLR model provides a new and effective way for predicting the price of gold. However the new version MLR also has limitations, that is, because of the moving average algorithm, the final predicted value will have lag problems. Where these models are not perfect, they need

to be improved. During the pandemic, it is significant to predict the gold price accurately when investors are doing their trade, even the new version of the MLR model is not enough for quasi-transactions. In the future, ANN can be applied to financial forecasting to fit better models.

References

1. Ranson, D., Wainright, H.C. (2005), Why Gold, not Oil, is the Superior Predictor of Inflation. Gold Report, World Gold Council, November.
2. Corti, C.H., Holliday, R. (2010), Gold Science and Applications. USA: Taylor and Francis Group, LLC.
3. Chang Pengyuan, Yao Hongxin. Analysis of international gold price forecast [J]. Business Situation,2019(7):64. (in Chinese) DOI:10.3969/j.issn.1673-4041.2019.07.055.
4. Shafiee, S., & Topal, E. (2010). An overview of global gold market and gold price forecasting. Resources policy, 35(3), 178-189.
5. Potoski, M. (2013). Predicting gold prices. CS229, Autumn.2
6. Khan, M. M. A. (2013). Forecasting of gold prices (Box Jenkins approach). International Journal of Emerging Technology and Advanced Engineering, 3(3), 662-670.
7. Meyler A, Kenny G, Quinn T. Forecasting Irish inflation using ARIMA models[J]. 1998.
8. Tripathy N. Forecasting gold price with auto regressive integrated moving average model[J]. International Journal of Economics and Financial Issues, 2017, 7(4): 324-329.
9. Nyoni T. Modeling and forecasting inflation in Kenya: Recent insights from ARIMA and GARCH analysis[J]. Dimorian Review, 2018, 5(6): 16-40.
10. Rady, E. H. A., Fawzy, H., & Fattah, A. M. A. (2021). Time series forecasting using tree based methods. J. Stat. Appl. Probab, 10, 229-244.
11. Navin G V. Big data analytics for gold price forecasting based on decision tree algorithm and support vector regression (SVR)[J]. International Journal of Science and Research (IJSR), 2015, 4(3): 2026-2030.
12. Lulu Chen. Stock price forecast based on multiple linear regression analysis -- A Case study of China Citic Bank [J]. Economic Research Guide,2016(19):75-76. DOI:10.3969/j.issn.1673-291x.2016.19.032.
13. Shokri, B. J., Dehghani, H., & Shamsi, R. (2020). Predicting silver price by applying a coupled multiple linear regression (MLR) and imperialist competitive algorithm (ICA). Metaheuristic Computing and Applications, 1(1), 1.