Machine Learning Models for Gold Price Prediction: A Comparative Analysis and Evaluation

Ran Kong

Department of COMP, Hong Kong Baptist University, Hong Kong, China

22257888@life.hkbu.edu.hk

Abstract. The prediction of gold prices is meaningful for multiple industries. Gold is a relatively stable store of value, and its price is closely linked to stocks, exchange rates, and monetary policy. Observing the trend of gold prices can promote the stable development of financial businesses. In this article, the daily opening and closing prices of gold over the past ten years were chosen as the data support. Seven different models were trained to predict prices on a yearly basis with a half-year interval. These predictions were then compared to actual data, and evaluations and analyses were conducted to identify the most suitable model for gold price forecasting. The machine learning models used included Ridge Regression, Linear Regression, Decision Tree Random Forest, Lasso Regression, Support Vector Machine, and K-Nearest Neighbors. In the evaluation process, Mean Squared Error (MSE) was used as the criterion to assess the accuracy of the model predictions.

Keywords: Machine Learning, Gold Price Prediction, MSE, Random Forest.

1. Introduction

Gold has long been regarded as a universally accepted currency. In modern economics, it is commonly seen as a hedge against inflation, a source of wealth, and to some extent, a safe investment, especially during periods of stock market volatility [1]. Therefore, economic analysts increase their holdings of gold during inflation, anticipating further inflation [2]. During economic recessions, bulk commodities like gold create a sense of security and opportunity for investors, as they are positively correlated with inflation [3]. Therefore, predicting the price of gold holds practical value as it is crucial for trading strategy optimization and financial risk management for institutions engaged in gold trading. It helps improve profits and reduce risks [4]. Additionally, gold price volatility also impacts financial institutions and enterprises.

One of the more advanced prediction models currently available for gold price prediction is the one proposed by Fong-Ching et al. In their research, they utilized Genetic Algorithm (GA) optimization to train and predict future gold prices using Least Squares Support Vector Regression (LSSVR). The performance of the model was evaluated using Mean Absolute Percentage Error (MAPE) and demonstrated its effectiveness [5]. When it comes to predicting the prices of specific investment commodities like stocks, oil, and Bitcoin, their prediction methods overlap significantly with those used for gold. For instance, in Shruthi's stock price prediction model, machine learning methods such as Support Vector Machines (SVM) and Decision Trees were employed, which are similar to the methods used to predict gold prices [6]. At the same time, there's also LSTM-based method to predict stock price movement [6]. In fact, as Ali mentioned, there are strong correlations between price fluctuations among these commodities, often exhibiting simultaneous upward or downward trends [7].

However, there are far fewer predictive models available for gold compared to the commodities. Existing research on gold price prediction is limited and mostly focuses on proposing and evaluating individual models, lacking comparisons between multiple models. Therefore, there is a need to develop more models for predicting gold prices and the followings are the reasons why it is necessary to innovate new methods of gold price predication. While there are many Machines Learning (ML) and Artificial Intelligence (AI) methods in the financial domain, investors are constantly competing and attempting to find new approaches that outperform the market [8, 9]. For instance, Qiu et al. delves into an in-depth examination of statistical models, particularly emphasizing the use of extreme

value mixture approaches in the finance and insurance sectors. If a large group of traders apply the same technique, the market becomes saturated, and the model loses its predictive ability. Hence, there is a continuous need for new methods, as their uniqueness may make them effective in specific market conditions [10].

Previous and existing research gaps include the lack of comparison between different models and the possibility of having more applicable predictive models. Additionally, previous research only focused on predicting the direction of gold prices, rather than accurately predicting the prices themselves.

To address this, this paper proposes an idea that involves comparing other price prediction models, establishing models not previously used for gold prediction, and evaluating the accuracy of each model to determine their potential value for further development and improvement. Furthermore, this study chooses to predict the precise prices rather than just the upward or downward trends, as it is more advantageous for traders in formulating business plans. In this regard, this study has designed seven different gold prediction models, each with its own algorithmic advantages. These seven models include Ridge Regression, Linear Regression, Lasso Regression, Decision Tree, Random Forest, Support Vector Regression (SVR), and K-Nearest Neighbor (KNN). Then, Mean Squared Error (MSE) was employed, a relatively straightforward method, to assess their fit and accuracy.

2. Method

2.1. Dataset Preparation

The dataset used in this research institute comes from Kaggle [11]. The dataset includes international gold prices from August 18, 2013, to August 18, 2023. The information includes the date, volume, open price, close price, high price, and low price of gold, recorded daily. It provides important data support for analyzing trends in the gold market. In terms of data preprocessing, there are missing values in the volume column of this dataset. Since the volume column has missing values and the data is in a time series format, this study does not use the median or mean to fill the missing values. Instead, this paper uses forward filling to fill the missing values.

Next, the dataset is divided into the training set and the test set, with the data from 2023 serving as the test set, and the data before 2023 serving as the training set. Features (X_train and X_test) and target variables (y_train and y_test) are extracted from the training set and test set. The features are obtained by removing the date and closing price columns from the original data, which are used to describe the time and price of gold trading. The target variable is the closing price, which is the value the model aims to predict. Multiple regression models are iterated through, and each model is trained and used for predictions. In each iteration, the model is trained using the training set (X_train and y_train), and then it makes predictions on the test set (X_test), resulting in predicted values (y_pred).

2.2. Machine Learning Models

This study uses Ridge Regression model, Linear Regression, Decision Tree Random Forest, Lasso Regression, Support Vector Machine K-Nearest Neighbors. This study adopts MSE as a measure of the potential of being a gold price prediction model. The basic formula for MSE is as follows:

$$MSE = (1/n) * \Sigma(yi - \hat{y}i)\text{^}2 \quad (1)$$

In this formula, n represents the number of samples, yi represents the actual value of the i-th sample, and ŷi represents the predicted value of the i-th sample. The Σ symbol denotes the summation, which sums up the values for all samples. A smaller value of MSE indicates a smaller difference between the predicted values and the actual values, indicating better performance of the model. MSE is a non-negative value, meaning it is always greater than or equal to zero.

There are several reasons for choosing MSE. Firstly, it has intuitiveness as an evaluation criterion, where smaller values indicate better prediction results. In assessing the usability of gold price prediction models, intuitiveness is crucial as it can save time. Secondly, MSE is a differentiable

function, allowing for the determination of model parameters that bring the predicted results closest to the true values in many optimization algorithms. Additionally, MSE is sensitive to outliers, enabling it to focus on samples with larger errors that may be unstable in gold price prediction.

2.2.1 Ridge Regression

Ridge regression introduces an additional L2 regularization term into the minimized objective function to address multicollinearity issues. The objective function becomes:

$$minimize: 1/2 * ||Y - X\beta||\text{^}2 + \alpha * ||\beta||\text{^}2 \quad (2)$$

Where Y is the observed target variable vector, X is the input feature matrix, where each row represents a data sample and each column represents a feature, β is the regression coefficient vector, representing the weights assigned to each feature. Ridge regression helps stabilize the estimation of regression coefficients by adding a regularization term, which helps prevent the model from excessively fitting the training data and improves its ability to generalize to new samples.

2.2.2 Linear Regression

The basic formula for linear regression is as follows:

$$y = \beta_0 + \beta_1 x_1 + \varepsilon \quad (3)$$

Where y is the target variable, x_1 is the feature variable, β_0 and β_1 are the regression coefficients, and ε is the error term. Linear regression has the advantage of simplicity and high interpretability, making it practical for evaluating models for gold price prediction.

2.2.3 Decision Tree and Random Forest

A decision tree is a machine learning algorithm that makes decisions based on a tree-like structure, commonly used for classification and regression tasks. Decision trees, as chosen models, have the advantages of interpretability and visualization. The decision process of a decision tree model can be intuitively displayed through its tree structure. Random forests can be composed of multiple decision trees and provide high accuracy.

2.2.4 Lasso Regression

Lasso regression adds an L1 regularization term to the loss function, which can drive some feature coefficients to become zero, achieving feature selection and model sparsity. The mathematical expression for Lasso regression is as follows:

$$Minimize: 1/2 * ||Y - X\beta||\text{^}2 + \alpha * ||\beta||\_1 \quad (4)$$

The reason for choosing Lasso is that in this regression model, only a few features have a significant impact on the target variable, and the coefficients of other features approach zero, allowing for the exclusion of many irrelevant factors.

2.2.5 Support Vector Machine and K-Nearest Neighbors

Both models are machine learning methods. The main goal of SVM is to find an optimal hyperplane that separates samples of different classes and maximizes the margin between the two classes. K-Nearest Neighbors, on the other hand, classifies an unlabeled sample based on the majority class among its K nearest labeled samples, measured by distance.

3. Results and Discussion

After training seven models based on the data, the actual closing prices of the training set and the predicted closing prices of the test set were shown in Fig. 1 and Fig. 2. The performance evaluation results of the given regression models on the gold price prediction task show significant differences in predictive performance. The following figure displays the curves of actual data and predicted data.

Ridge MSE: 45.39

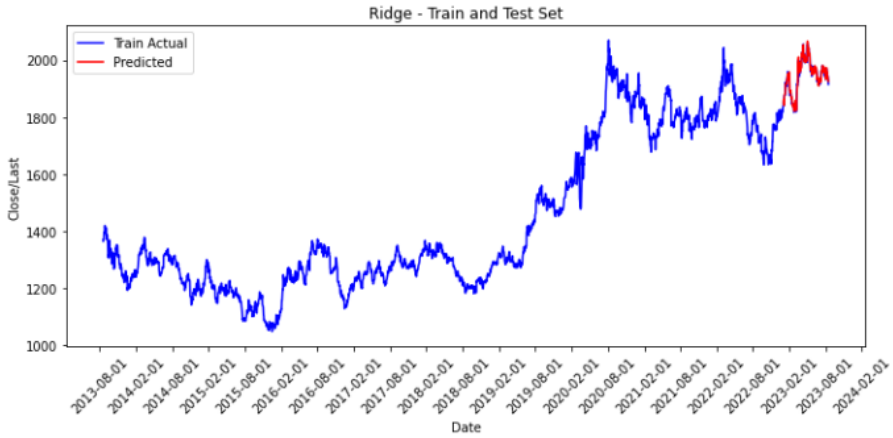

Figure 1. Price curves based on Ridge Regression (Picture credit: Original)

Linear Regression MSE: 45.39

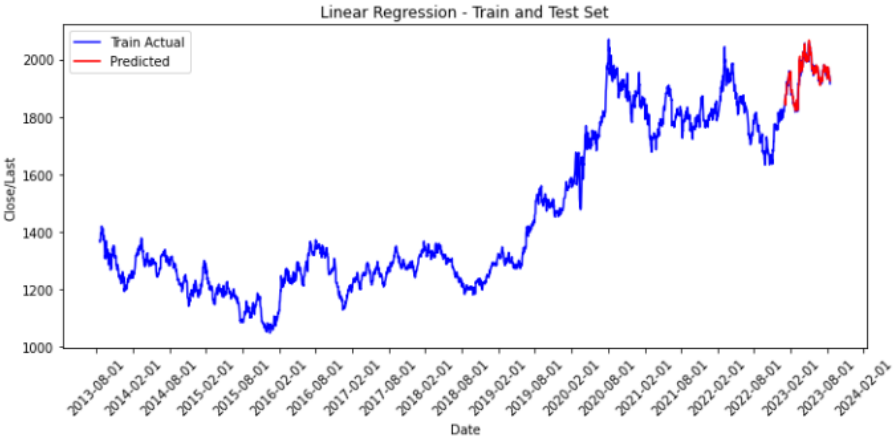

Figure 2. Price curves based on Linear Regression (Picture credit: Original)

The Ridge regression model and the linear regression model exhibit low mean squared error (MSE) values of 45.39 and 45.39 shown in Fig. 3 and Fig. 4, respectively, indicating good accuracy in predicting the gold price.

Lasso MSE: 74.68

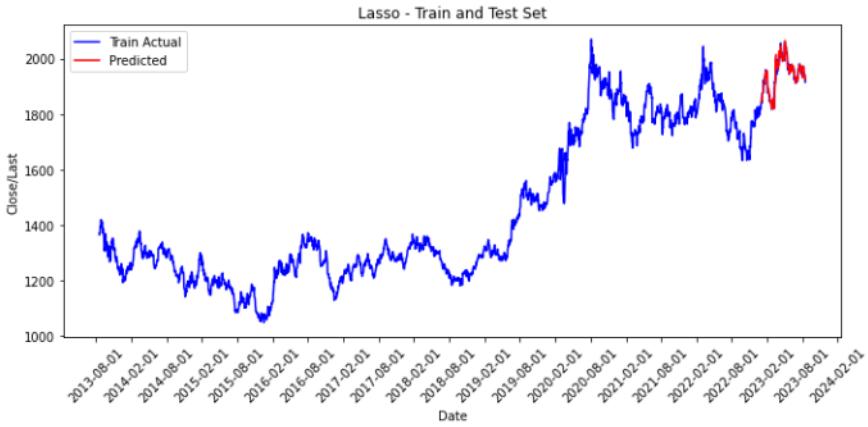

Figure 3. Price curves based on Lasso Regression (Picture credit: Original)

The Lasso regression model has a slightly higher MSE of 74.68 compared to Ridge and linear regression models shown in Fig. 4 and Fig. 5, but it is still relatively low.

Random Forest MSE: 90.31

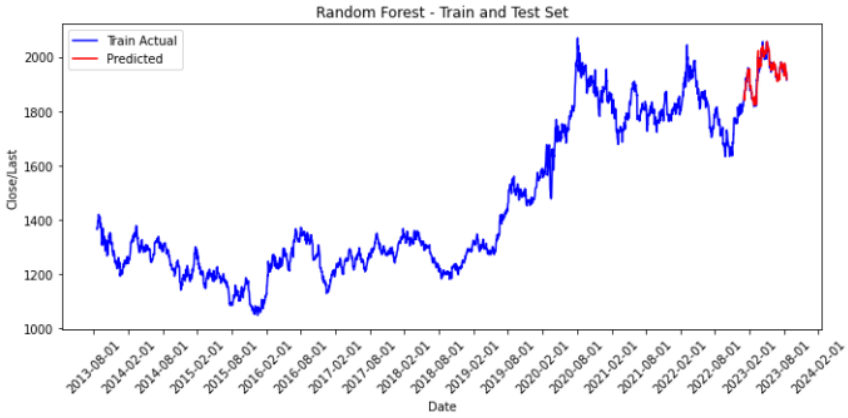

Figure 4. Price curves based on Random Forest (Picture credit: Original)

Decision Tree MSE: 121.59

Figure 5. Price curves based on Decision Tree (Picture credit: Original)

The decision tree model and the random forest model have MSE values of 121.59 and 90.31 shown in Fig. 6 and Fig. 7, respectively, indicating average predictive performance.

Support Vector Machine MSE: 407690.66

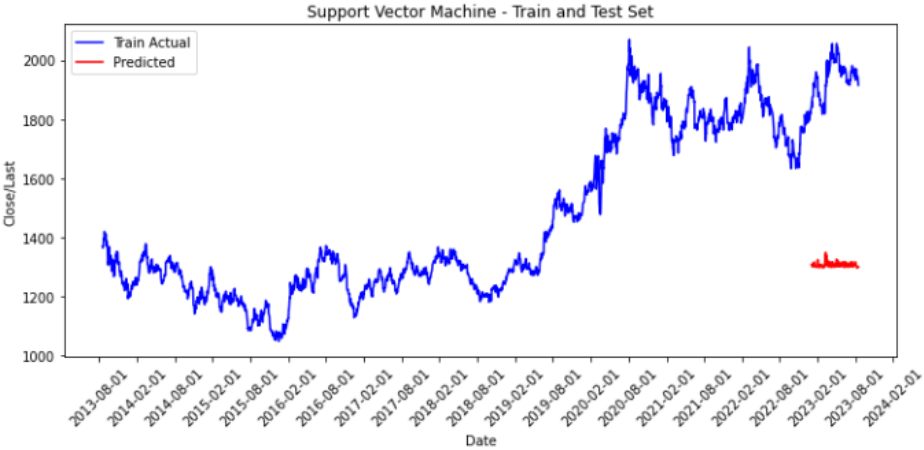

Figure 6. Price curves based on Support Vector Machine (Picture credit: Original)

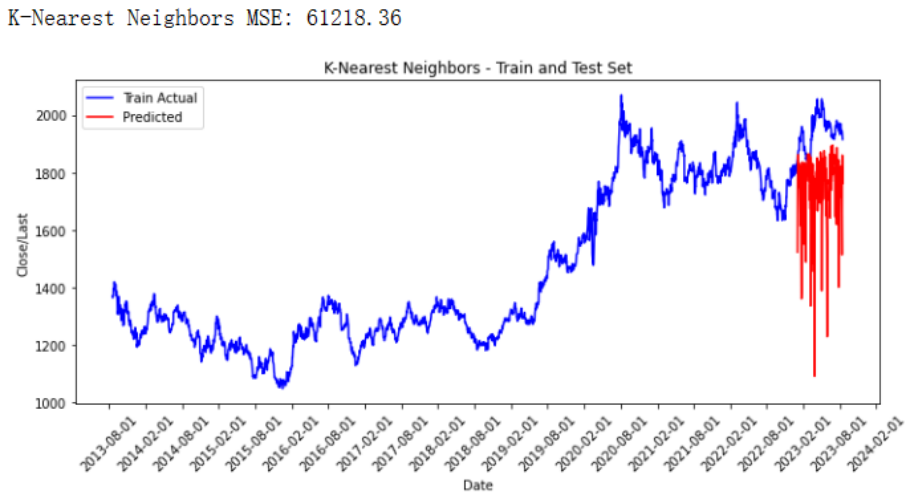

Figure 7. Price curves based on K-Nearest Neighbors (Picture credit: Original)

On the other hand, the support vector machine model and the K-nearest neighbors model have MSE values of 407,690.66 and 61,218.36, respectively, indicating relatively poor performance.

Based on the experimental results, this study considers the Ridge regression model, linear regression model, Lasso regression model, decision tree model, and random forest model as potential models for gold price prediction, as they exhibit relatively small and comparable MSE values. Lasso regression can compress feature coefficient to 0 in feature selection, excluding features that have relatively little impact on the price of gold. Ridge regression can fit linear relationship and prevent overfitting and linear regression can fit linear relationship, and both the regression is relatively simple among the machine learning models. As for decision tree, it can capture nonlinear relationships in data, to deal with irregular fluctuations in the price of gold. A smaller MSE value indicates that the actual data and predicted data are closer. In this study, no threshold was set because this was only a relative test, and the MSE value alone is not sufficient to establish a formal prediction model. Further research is needed to continue reducing this value, which will be addressed in more in-depth studies.

In further research, there are several methods to reduce the MSE value, such as considering more factors, it is suggested to incorporate additional predictive indicators, such as irregular and unexpected factors like inflation data. Inflation levels are typically closely related to gold prices, including metrics like inflation rate, consumer price index, monetary policy, real interest rates, and others, as these data have a significant impact on gold price trends. Governments often use interest rate adjustments and money supply control to manage inflation levels, which in turn affect gold prices. Other fixed factors, such as geopolitical factors, should also be considered, as economic conditions vary across different regions, and companies should analyze specific local data accordingly. The accuracy of the results can be further improved by incorporating techniques such as time series analysis and short-term memory [12]. In addition, some advanced machine learning or deep learning models can be also considered for further performance improvement [13-16].

Finally, there are regular factors, such as supply and mining data. Tracking gold production levels and financial reports of mining companies can provide insights. Additionally, trading volumes and open interest in gold exchanges can reflect market demand and supply conditions.

4. Conclusion

This paper proposes testing seven different models using historical gold price data from the past decade to obtain predictive data. A comparison is made between the predicted data and actual data to determine whether a particular machine method can serve as a gold price prediction model, thereby enabling further in-depth research. MSE is used as the evaluation criterion, with lower values indicating better suitability as a gold price prediction model.

The research findings indicate that among the seven models used in this study, the Ridge regression model, linear regression model, Lasso regression model, decision tree model, and random forest

model show potential as gold price prediction models. On the other hand, the support vector machine model and K-nearest neighbors model are not suitable for gold price prediction due to their significantly higher MSE values and considerable divergence from the aforementioned models. In the future, these five models can be further optimized and fine-tuned by businesses to serve gold price prediction purposes. Investing in gold as an asset storage option or as an investment can enhance enterprise stability, thereby promoting robust socio-economic development.

References

[1] Arouri MEH, Lahiani A, Nguyen DK. World gold prices and stock returns in China: Insights for hedging and diversification strategies. Economic Modelling, 2015, 44: 273-282, ISSN 0264-9993.

[2] Baur DG, McDermott TK. Is gold a safe haven? International evidence. Journal of Banking & Finance, 2010, 34 (8): 1886-1898, ISSN 0378-4266.

[3] Bampinas G, Panagiotidis T. Are gold and silver a hedge against inflation? A two century perspective. International Review of Financial Analysis, 2015, Volume 41: 267-276, ISSN 1057-5219.

[4] Nguyen QN, Bedoui R, Majdoub N, Guesmi K, Chevallier J. Hedging and safe-haven characteristics of Gold against currencies: An investigation based on multivariate dynamic copula theory. Resources Policy, 2020, Volume 68: 101766, ISSN 0301-4207.

[5] Fong-Ching Y, Chao-Hui L, Chiu C. Using market sentiment analysis and genetic algorithm-based least squares support vector regression to predict gold prices. International Journal of Computational Intelligence Systems, 2020, 13 (1): 234-246.

[6] Rammurthy S, Patil S. An LSTM-based approach to predict stock price movement for IT sector companies. International Journal of Cognitive Informatics & Natural Intelligence, 2021, 15 (4): 1-12. https://doi.org/10.4018/IJCINI.20211001.oa3.

[7] Ali R, Mangla IU, Rehman RU, Xue W, Naseem MA, Muhammad IA. Exchange rate, gold price, and stock market nexus: A quantile regression approach. Risks, 2020, 8 (3): 86.

[8] Qiu Y, Wang J. A Machine Learning Approach to Credit Card Customer Segmentation for Economic Stability. InProceedings of the 4th International Conference on Economic Management and Big Data Applications, ICEMBDA 2023, October 27–29, 2023, Tianjin, China 2024 Jan 19.

[9] Qiu Y. Estimation of tail risk measures in finance: Approaches to extreme value mixture modeling (Doctoral dissertation, Johns Hopkins University).

[10] Hossain E, Hossain MS, Zander P-O, Andersson K. Machine learning with Belief Rule-Based Expert Systems to predict stock price movements. Expert Systems with Applications, 2022, Volume 206: 117706, ISSN 0957-4174.

[11] Kaggle. Gold and silver prices. https://www.kaggle.com/datasets/kapturovalexander/gold-and-silver-prices-2013-2023?resource=download, 2024.

[12] Livieris IE, Emmanuel P, Panagiotis P. A CNN–LSTM model for gold price time-series forecasting. Neural Computing & Applications, 2020, 32 (23): 17351-17360.

[13] Liu Y, Liu L, Yang L, Hao L, Bao Y. Measuring distance using ultra-wideband radio technology enhanced by extreme gradient boosting decision tree (XGBoost). Automation in Construction. 2021 Jun 1; 126: 103678.

[14] Zhou Y, Osman A, Willms M, Kunz A, Philipp S, Blatt J, Eul S. Semantic Wireframe Detection.

[15] Qiu Y, Yang Y, Lin Z, Chen P, Luo Y, Huang W. Improved denoising autoencoder for maritime image denoising and semantic segmentation of USV. China Communications. 2020 Mar;17 (3): 46-57.

[16] Wang H, Zhou Y, Perez E, Roemer F. Jointly Learning Selection Matrices for Transmitters, Receivers and Fourier Coefficients in Multichannel Imaging. arXiv preprint arXiv: 2402.19023. 2024 Feb 29.