Gold Prediction Based on XGBoost and OLS

Baixi Jiao *

Ira A. Fulton Schools of Engineering at Arizona State University, Tempe, USA

* Corresponding Author Email: bjiao1@asu.edu

Abstract. As a matter of fact, the value of gold is not only reflected in how much it is worth, but its unique nature also makes it determine economic and political policies. In reality, Gold is often seen as a hedge against inflation, influencing the direction of countries and investors. Gold prices from 2012-2018 from various websites were chosen as the dataset, including opening and closing prices and so on. With this in mind, using XGBoost and linear regression models successfully predicted the price of gold more accurately and it is not much different from the actual price. However, these methods are outdated, and geopolitical factors, national economic policies, and so on are some of the major factors affecting the price of gold in an era of increasingly up-to-date information. According to the analysis, the usage of integrated forecasting models combining Random Forests, MLR, SVM and ANN is a more futuristic way to predict the price of gold.

Keywords: Gold Prediction, XGBoost, linear regression, forecasting models.

1. Introduction

Forecasting gold prices dates back to the ancient world, when gold was a symbol of riches and power. However, scientific procedures as one knows them now began to emerge in the twentieth century. The early research was primarily qualitative, focusing on geopolitics, central bank policy, and gold mining production. Researchers began employing quantitative approaches for predicting with the emergence of modern computing in the mid-twentieth century. The Box-Jenkins ARIMA models were among the first to be utilized for forecasting time series such as gold prices. These models were built on the premise that future prices may be predicted using historical data and mistakes [1]. Machine learning and artificial intelligence were first employed in gold price predicting in the early twenty-first century. Neural networks, support vector machines, and random forests were examples of early machine learning models. These models have the advantage of being highly accurate and capable of comprehending intricate, non-linear linkages [2]. More advanced models become conceivable with the advent of big data. Forecasting began to employ deep learning, a form of machine learning. Long short-term memory (LSTM), bidirectional long short-term memory (Bi-LSTM), and gated recurrent unit (GRU, another type of RNN) models are capable of processing enormous volumes of data and learning sophisticated patterns [3]. Researchers have started including other forms of data into their models. Social media sentiment analysis, Google trends data, and other unusual data sources, for example, began to be employed alongside standard financial data [4]. Today's gold price forecasting is increasingly complex, employing a wide range of algorithms and data sources. Hybrid models, which blend the best features of numerous types, are becoming more popular. Advances in artificial intelligence, big data, and computer power are anticipated to have an impact on gold price predicting in the future. Reinforcement learning and other sophisticated machine learning techniques, in particular, may become increasingly common. The history of gold price forecasting has been one of ongoing innovation and improvement. As technology advances, one may anticipate ever more complex and accurate models in the future [5].

For thousands of years, gold has been regarded as a valuable metal. For thousands of years, people have used gold as a reserve currency, barter, and jewelry. A vital commodity in both the manufacturing and financial markets. The price of gold is therefore closely watched on the international financial markets. The global financial markets are closely monitored. Gold is regarded as one of the most essential investing instruments. To keep their connections to the outside world, nations and multinational businesses rely on exchange rates, one of the most important economic

elements, along with gold reserves. access to the outside world. The gold index and gold market are thus among the biggest and most important financial marketplaces in the world. As a result, the gold index can be immediately influenced positively or adversely by the numerous changes that might occur in the market, economic, and governmental policies [6]. Gold price forecasting is an essential element of financial study with considerable consequences for investors, policymakers, and economists. Oil and gold are essential commodities for the global economy, and investors typically include them in their stock portfolios. Gold is typically seen as a safe-haven asset during times of crisis. Gold and crude oil had an impact on international stock markets during the financial crisis and COVID-19, as well as the real economy. Gold is essential for currency trading and hedging, but rising gold price volatility is linked to riskier investment profiles, whereas falling gold price volatility linked, as well. Understanding gold price volatility is essential for financial markets, the entire economy, hedging decisions, and derivatives valuation [7]. The importance of gold as an inflation hedge is rising. The Consumer Price Index (CPI) and the price of gold are interrelated. Due to its dual function as a financial asset (used as a store of value) and a commodity (used in the production of jewelry and other industrial applications), gold is recognized as a unique asset [8].

Because of its function in preserving value and hedging against inflation, gold has become an essential asset in global financial markets. Forecasting the gold price is so critical for investors, financial institutions, and policymakers in making decisions and managing risk. The XGBoost and Ordinary Least Squares (OLS) models are two options for this problem, each with its own set of benefits. Extreme Gradient Boosting, or XGBoost, is an integrated machine learning approach that constructs decision trees using the gradient boosting framework. There are several advantages to adopting XGBoost for gold price prediction. XGBoost captures intricate nonlinear correlations between input characteristics, which is effective when predicting gold prices with interconnected and nonlinear impacts. (Working with Complex Nonlinear Relationships) Even if the data contains numerous characteristics or the connections in the data are complicated, XGBoost avoids overfitting and remains dependable. (XGBoost is parallelizable, which means it can train using many cores on the computer's CPU, making it quicker and more scalable. A linear regression model called Ordinary Least Squares seeks to minimize the sum of squared differences between observed and predicted values. The appeal of utilizing OLS for gold price forecasting stems from its ease of use and interpretability. OLS models are simple to comprehend and apply. They are appropriate for situations in which the independent and dependent variables have a linear connection. An OLS model's coefficients directly assess the effect of a one-unit change in the independent variable on the dependent variable. As a result, OLS models are highly interpretable, which is crucial for understanding the variables influencing gold price movements. OLS supports statistical inference, so one may test hypotheses, compute confidence intervals, and learn more about the relationships in the data.

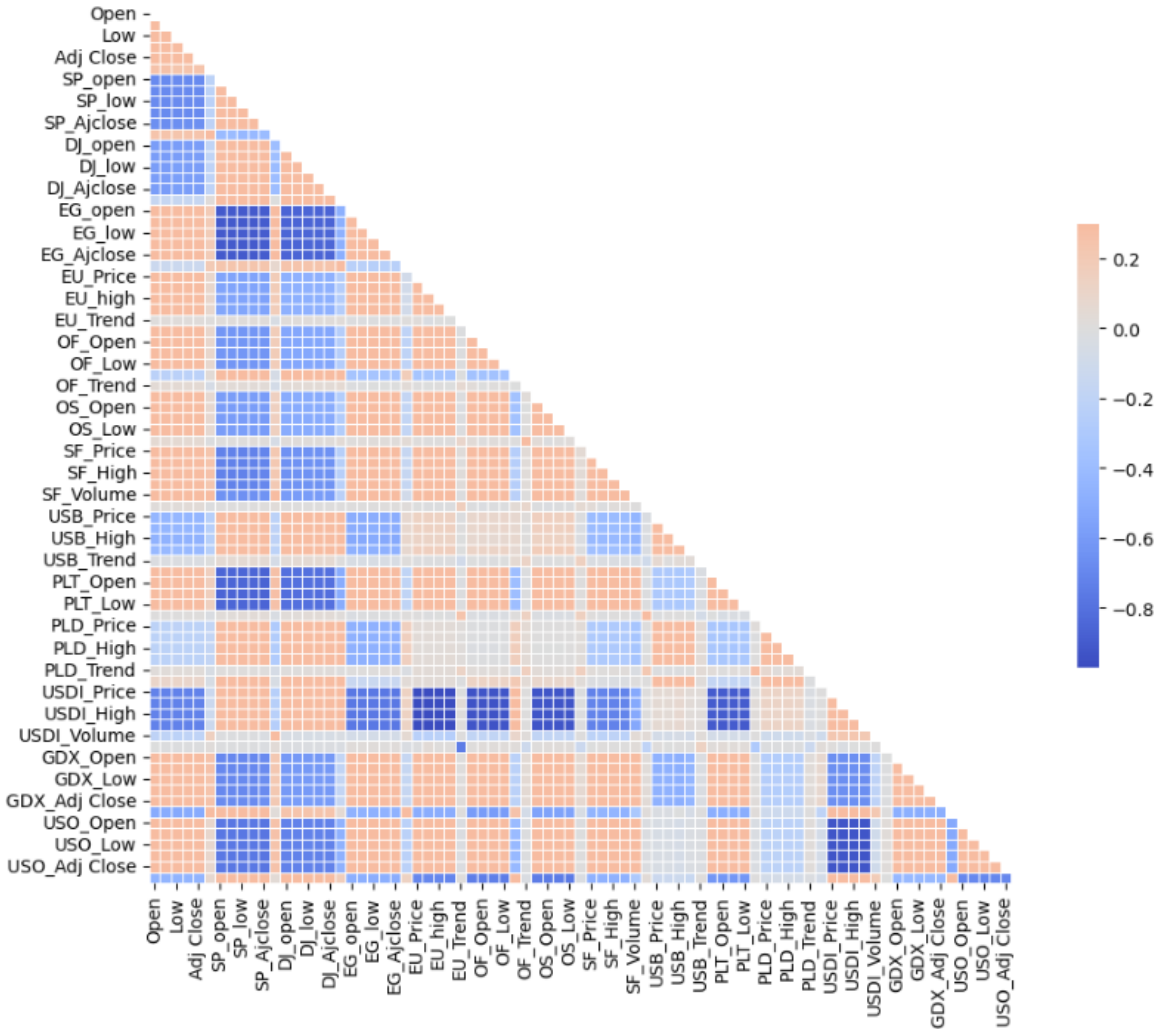

Figure 1. Feature analysis [7].

2. Data and Method

Between November 18, 2011 and January 1, 2019, information for this study was obtained from a variety of sources. 80 columns and 1718 rows make up the data set. The dataset includes 1718 rows and 80 columns of attribute data, such as the price of oil, the Standard & Poor's (S&P) 500 Index, the Dow Jones U.S. Bond Rate (10-year), the Euro-Dollar exchange rate, the price of precious metals such as silver, platinum, and palladium, the U.S. Dollar Index, the Eldorado Gold Corporation, and Gold Miners, and ETF price summaries. Date, open, high, low, close, adjusted close, and volume are the seven columns that make up the historical data for the Gold ETF that was obtained from Yahoo Finance. The difference between an adjusted close and a close is that a stock's closing price is its price at the conclusion of the trading day. Contrarily, the adjusted closing price evaluates value by considering dividends, stock splits, and new stock offerings. The dependent variable for the whole dataset is the spot price of gold. The primary source of data for this project is the daily opening and closing prices, which serve as the dependent variable. The time variable is the independent variable. The gold price is determined as an average of the weekly gold price [9]. The correlation analysis of the features is shown in Fig. 1. Some of the features are selected for prediction.

To estimate the price of gold, this research employed two separate models to perform regression analysis on the dataset: XGBoost and Linear Regression, with hyperparameters tailored to the XGBoost model using a grid search. The XGBoost model, which is the initial model and the main model in this research, employs the following parameters: Maximum Depth: 3, 5, 7, 10; Minimum Child Weight: 1, 2, 3; Learning Rate: 0.01, 0.1, 0.2, 0.3; Subsample: 0.5, 0.7, 1.0; Colsample bytree:

0.5, 0.7, 1.0. Colsample bytree, Number of Estimators, and Objective are all set to "reg: squared error". Linear regression, unlike XGBoost, does not contain hyperparameters. It optimizes during training to obtain the optimum coefficients for the features. Two common regression metrics are used in this project for model evaluation: Mean Square Error (MSE) and R-squared (R2) score. MSE is a metric that evaluates the average squared difference between the real target value and the projected value, with lower MSE values indicating better model performance. The degree of variation in the target variable that can be anticipated from the independent variable is measured by the R2 metric. Higher numbers denote greater fit; the scale goes from 0 to 1. The value increases with fit quality.

3. Results and Discussion

A simple machine learning model called linear regression has the response represented by a linear combination of predictors. Initial weights for the features are set to zero or a random number throughout construction. By multiplying these weights by the appropriate feature values and summing them, predicted values for the target variable are calculated. Using a loss function like Mean Squared Error (MSE), the difference between the actual and predicted values is computed and combined into a single result. The weights are then changed based on this mistake using an optimization approach such as Gradient Descent, and the procedure is repeated until the error cannot be decreased any more. XGBoost (Extreme Gradient Boosting) is a more sophisticated version of gradient boosting machines. It starts by assigning equal weights to all observations and building a basic model, generally a decision tree, that forecasts all instances as the target variable's average. For regression tasks, the loss function, such as MSE, is determined. The residuals (errors) from the previous model are then fitted to a new model, which seeks to repair the errors caused by the prior model. This new model is integrated with the old model sequence, and the process is repeated until a certain number of models are formed, or the loss cannot be decreased any longer. The final projections are calculated by averaging the predictions from all models. A Hyperparameter search was conducted. The best performing set of hyperparameters is recorded in the Table 1 and Table 2. Based on the results, the MSE (Mean Square Error) and R2 (R-squared) scores indicate that both models performed very well on the data set. Given that the R2 score is 1.0 and the MSE is close to 0, the linear regression model appears to have a perfect fit, indicating that it has captured all patterns in the data. A flawless score in a real-world circumstance, however, can be an indication of overfitting. The XGBoost model, on the other hand, performs exceptionally well, with an R2 value of about 0.999, suggesting that the model can account for 99.9% of the variation in the dependent variable. Though XGBoost's MSE is larger than that of linear regression, it is still relatively low, indicating that the model provides a strong fit to the data.

Table 1. Summary of metrics for two models

MSE	XGBoost	0.21884125418844
MSE	Linear Regression	5.89986664566e-16
R2	XGBoost	0.999356203641910
R2	Linear Regression	1.0

Table 2. Summary of the factors.

colsample_bytree		learning_rate		max_depth		min_child_weight		n_esimator		subsample	
0.5	3.596	0.01	3.508	3	3.474	1	3.329	100	3.559	0.5	3.568
0.7	3.456	0.1	3.634	5	3.429	2	3.749	200	3.694	0.7	3.442
1	3.413	0.2	3.425	7	3.531	3	3.495	500	3.672	1.0	3.683
		0.3	3.362	9	3.486						

4. Limitations and Prospects

Predictive modeling techniques, e.g., XGBoost and linear regression are frequently utilized because of their effectiveness in handling complex information. However, when used to predict gold

prices, these models have intrinsic flaws. The dependent and independent variables must have a linear relationship in order to linear regression to work, which may not always be the case for gold prices. A variety of variables impact gold prices, many of which may have non-linear or complicated correlations with the price. Furthermore, linear models may not properly represent the possible volatility and unpredictable variations in gold prices, limiting their prediction accuracy. XGBoost, a gradient boosting framework, on the other hand, is capable of capturing non-linear correlations and has demonstrated effectiveness in a variety of machine learning problems. The intricacy of XGBoost, however, makes it vulnerable to overfitting, which is when a model performs well on training data but poorly on unknown data. Both models are also sensitive to outliers, which can affect forecasts when present in the data.4. Both models are also sensitive to outliers, which can skew forecasts when present in the data. Furthermore, the performance of both models is significantly dependent on the model's attributes. If essential gold price predictors such as economic factors, geopolitical events, or market emotion are not included, the projections may be inaccurate.

Several possible paths might improve the prediction capacities of gold price models in the future. The growth and advancement of machine learning and AI technologies point to a future in which predictions may become more accurate. The incorporation of increasingly complex and diversified elements is predicted to be important. More detailed data, such as global economic indicators, geopolitical events, market sentiment research, and even social media trends, might give useful prognostic insights. The Random Forest algorithm and the XGBoost algorithm have been combined to provide an efficient ensemble strategy for gold price prediction. It uses metamodeling to improve the ensemble's forecasting performance. The project trains and tests individual models on data from Google Finance (2013-2023), then combines their forecasts into a uniform ensemble prediction using meta-models. Multiple performance measures, including MAE, MSE, RMSE, R2, MAPE, and Max AE, reveal that the ensemble technique outperforms individual algorithms in terms of predictive power. This work advances gold price forecasting by proving the efficiency of ensemble learning approaches in capturing complicated dynamic patterns in the gold market. The methodologies provided have a high potential for improving financial decision-making and risk management tactics in the gold investing arena [10].

Forecasting the Gold Price Index (XAU/USD) is recommended in this study. The XAU/USD movement is examined by gathering data from July 2019 to July 2020. Included is a feature set. The data set includes the opening and closing prices, as well as the high and low prices. This is part of the data set. The dataset now includes the US Dollar Index variable, which has an influence on gold. Technical indicators including Bollinger Bands, the relative strength index, and simple moving averages were introduced to expand the feature set. Then, five different regression models were created: support vector regression, random forest regression, decision tree regression, and linear regression. Finally, by merging the preceding four regression models, a voting regression model and a stacking regression model were used to provide a more robust gold price. To assess and predict the price of gold, academics have used a range of models based on linear regression (MLR), support vector machines (SVM), artificial neural networks (ANN), and others. To the best of the knowledge, this is the first-time regression models have been combined to forecast the XAU/USD index. The experimental results indicate that utilizing a combined regression model instead of separate estimations results in more reliable gold price forecasts [6].

5. Conclusion

To sum up, this study investigates the effectiveness of combining machine learning models, especially the combination of linear regression and XGBoost, in forecasting gold prices. These models' strong predicted accuracy illustrates their potential value in financial forecasting. The study provides a flawlessly fitted linear regression model with an R2 score of 1.0 and an MSE near zero, but the XGBoost model, while not perfect, explains 99.9% of the variability in the dependant variable. These models, however, have intrinsic limitations. The dependent and independent variables must

always have a linear relationship for linear regression to be valid, which may not always be the case with gold prices. With xGBoost, a framework for gradient augmentation, overfitting is a possibility, and both models are prone to outliers. Despite these drawbacks, this research offers a significant advancement in the field of financial forecasting. The effective implementation of these integrated machine learning models in forecasting the price of gold, such as Random Forest, MLR, SVM, and ANN, is predicted to enhance financial decision making and risk management techniques in the field of gold investing. Forecasting powers might be improved in the future by combining more complicated and diverse aspects such as global economic data, geopolitical events, and market sentiment analyses.

References

[1] B. Guha and G. Bandyopadhyay, Journal of Advanced Management Science 4, 2 (2016).

[2] S. Patalay, and M. R. Bandlamudi, "Gold Price Prediction Using Machine Learning Model Trees," International Conference on Changing Business Paradigm (ICCBP), Murshidabad (2021) pp. 154 - 188.

[3] M. Yurtsever, Avrupa Bilim ve Teknoloji Dergisi 31, 341 - 347 (2021).

[4] E. Bouri, R. Gupta, S. Hosseini, and C. K. M. Lau, Emerging Markets Review 34, 124 - 142 (2018).

[5] S. Das, T. P. Sahu and R. R. Janghel, Resources Policy 79, 103109 (2022).

[6] Z. H. Kilimci, Journal of Emerging Computer Technologies 2 (1), 7 - 12 (2022).

[7] K. Yen-Ku, A. Maneengam, P. T. Cong, et al., Resources Policy 79, 103024 (2022).

[8] S. S. Sharma, Economic modelling 55, 269 - 278 (2016).

[9] Gold Price Prediction Dataset. (2021, July 20). Kaggle. Retrieved from: https://www.kaggle.com/datasets/sid321axn/gold-price-prediction-dataset.

[10] D. K. Kushwaha, D. K. Sharma, S. S. Khullar, et al., Rivista Italiana di Filosofia Analitica Junior 14 (1), 116 - 121 (2023).