ISSN: 2037-4445

Gold Price Prediction Using an Ensemble of Random Forest and XGBoost

Dinesh Kumar Kushwaha, Dhananjay Kumar Sharma, Shanal Singh Khullar, Satyam Shukla, Tarun Kumar Pandey, Surendra Pal
Computer Science & Engineering
United Institute of Technology
Prayagraj, India
dskkushwaha658@gmail.com, dhananjay7862@gmail.com
, shanalkhullar@gmail.com, surendrapal5038@gmail.com, tarunpandeyuptu19@gmail.com,
shuklasatyammm7839@gmail.com

Abstract:
This research paper presents an optimized ensemble approach for gold price prediction by combining the Random Forest and XGBoost algorithms. The proposed methodology incorporates a meta-model to enhance the ensemble's predictive performance. Using data from Google Finance (2013-2023), we train and test the individual models before leveraging the meta-model to fuse their predictions into a unified ensemble prediction. Evaluation using multiple performance metrics, such as MAE, MSE, RMSE, R2, MAPE and Max AE, demonstrates the superior predictive capabilities of our ensemble approach compared to the individual algorithms. This study contributes to the advancement of gold price prediction by leveraging ensemble learning techniques and showcasing their effectiveness in capturing the complex dynamics of gold market trends. The proposed approach holds significant potential for improving financial decision-making and risk management strategies in the domain of gold investments.

Keywords: gold price prediction, random forest, xgboost, machine learning, ensemble learning, meta-model

1. Introduction

The value of gold as a safe asset and an economic indicator has gained immense significance globally, particularly in times of economic and political uncertainty. The COVID-19 pandemic has amplified economic instability [1], further emphasizing the importance of a secure investment such as gold. However, this has also resulted in out-dated prediction models struggling to cope with unprecedented changes in the market.

Accurately determining the price of gold holds utmost importance for investors and financial analysts seeking to make well-informed decisions. Recognizing the potential of machine learning techniques in this domain [6], we present a novel ensemble model that combines Random Forest and XGBoost algorithms to forecast gold prices. Additionally, we introduce a meta-model approach to optimize the ensemble's predictive performance.

We procured data on gold prices from Google Finance spanning from 2013 to 2023 to train and test our models. The Random Forest model, known for its ability to handle non-linear relationships [9], and the XGBoost model, a powerful gradient boosting algorithm [13], were individually trained on the dataset. Subsequently, we employed a meta-model, specifically a Linear Regression model, to combine the predictions from Random Forest and XGBoost into a unified ensemble prediction.

Through rigorous experimentation, our ensemble model achieved remarkable results, surpassing the performance of each algorithm individually. The Mean Absolute Error (MAE) for the ensemble was 7.56, the Root Mean Squared Error (RMSE) was 11.86, and the R-squared value was 0.99, indicating a strong correlation between the predicted and actual gold prices. Furthermore, the Mean Absolute

Percentage Error (MAPE) for the ensemble was 0.5%, with a maximum absolute error (Max AE) of 80.

To gain insights into the complex interplay between gold prices and various economic indicators, we utilized visualization techniques such as Heat maps and Pair plots. Our findings revealed significant correlations between gold prices and economic factors such as the S&P 500 index [10], Dow Jones Industrial Average, currency exchange rates, and the prices of other precious metals.

The presented research contributes to the growing body of literature on gold price prediction using machine learning techniques. By incorporating an ensemble model and the novel meta-model approach, we enhance the accuracy and effectiveness of gold price prediction, enabling investors and financial analysts to make informed decisions in the volatile market. Our work opens avenues for further research and development in this field, including the exploration of additional ensemble techniques and the incorporation of advanced deep learning models.

2. Literature Review

Gold price prediction has been a subject of extensive research in the fields of finance and economics. Conventional approaches, like the Autoregressive Integrated Moving Average (ARIMA) models, have been commonly employed for forecasting gold prices. To illustrate, Guha et al. [5] (2016) presented a predictive model based on ARIMA for gold price estimation .

ML methods such as linear regression, RF, and GB have gained popularity in recent years for predicting gold prices. A study by M.Shivani et al. [6] (2021) compared the performance of linear regression, Random Forest, and Gradient Boost models for gold price prediction. They found that the RF model outperformed the other two models.

Deep learning methods, such as Long Short-Term Memory (LSTM) networks, a form of recurrent neural network, have also been utilized for gold price forecasting, especially to evaluate long-term trends. For example, A study by Laor Boongasame et al. [7] (2022) proposed gold price forecasting method using Long Short-Term Memory and the Association Rule .

As highlighted, there are multiple ML and statistical methods that have been trailed for gold price prediction, but each has its own strengths and limitations. While some may work well in certain specific or broad scenarios, they may not be quite suitable for others. For example, individual ML models such as Random Forest or Gradient Boosting can be very prone to overfitting, and demand careful tuning of their hyperparameters.

3. Data and Methodology

Data Collection

We collected daily gold prices, as well as data of variables such as S&P 500 index, Dow Jones Industrial Average, US Oil Fund, currency exchange rates, Newmont Corporation stocks, and prices of other precious metals such as Silver from Google Finance from January 1, 2013, to April 28, 2023. This data was used for the creation of our dataset.

Data Preparation

We utilized the Pandas Library in Python to pre-process and clean the data. We first checked for missing values in the dataset and found that some dates were missing. To fill the missing values, we used the interpolate method, which estimates the missing values based on the values of neighbouring dates.

We then split the created dataset into training and testing sets, utilizing 80% of the data for training purposes, while 2% for testing.

Data Analysis

We performed exploratory data analysis to gain insight into the data and visualize the relationships between the features and the target variable. We used Seaborn to create Jointplots to show the relationship of gold prices with each independent variable in individual graphs. We also created a Heat map and Pair plot to better visualize and understand the correlation in a clear and concise manner.

Model Training

In the model training phase, we employed two machine learning algorithms, Random Forest[9] and XGBoost[13], to predict gold prices.

Once both models were trained, we proceeded to create an ensemble prediction using a meta-model approach. We combined the predictions from Random Forest and XGBoost models to form an ensemble feature matrix. Each model's prediction was considered as a separate feature.

Subsequently, we trained the meta-model, in this case, a Linear Regression model, using the ensemble features and the actual gold prices. The model learned to combine the predictions from the individual models, optimizing the ensemble's performance.

Model Evaluation

To evaluate the performance of our model, we utilized various metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-Squared (R2),.Mean Absolute Percentage Error (MAPE), and Maximum Absolute Error (Max AE). We also visualized the predicted and actual gold prices using line plots and scatter plots.

4. Results

To obtain the gold price data for the period between 2013 and 2023, we collected information from reliable sources such as Google Finance. The dataset includes daily gold prices, allowing us to analyze the trends and patterns over the specified timeframe. Fig. 1 displays the historical gold prices, showcasing the significant rise during the COVID-19 pandemic from 2020 onwards. This period witnessed a surge in gold prices due to increased economic uncertainty and investors' shift towards safe-haven assets.

The graph illustrates the rapid increase in gold prices during the pandemic, highlighting the volatility and impact of global events on the precious metal market. This dataset serves as a valuable foundation for training and evaluating our proposed ensemble model for gold price prediction.

Fig. 1: Monthly Variation of Gold Price

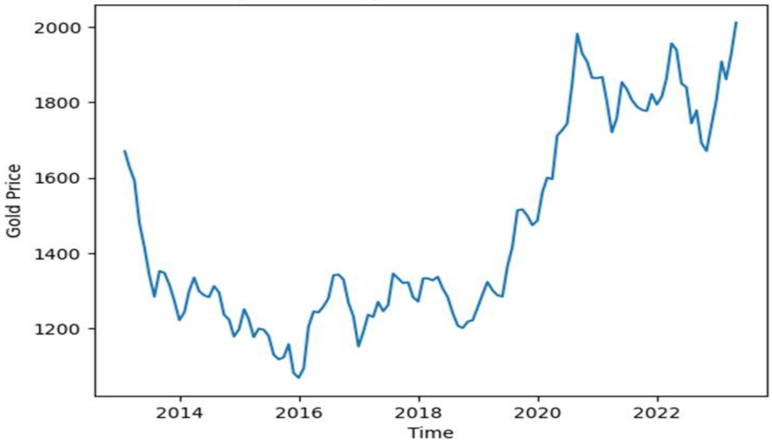

Fig. 2, the heat-map visualization, provides valuable insights into the relationship between various attributes and the price of gold.

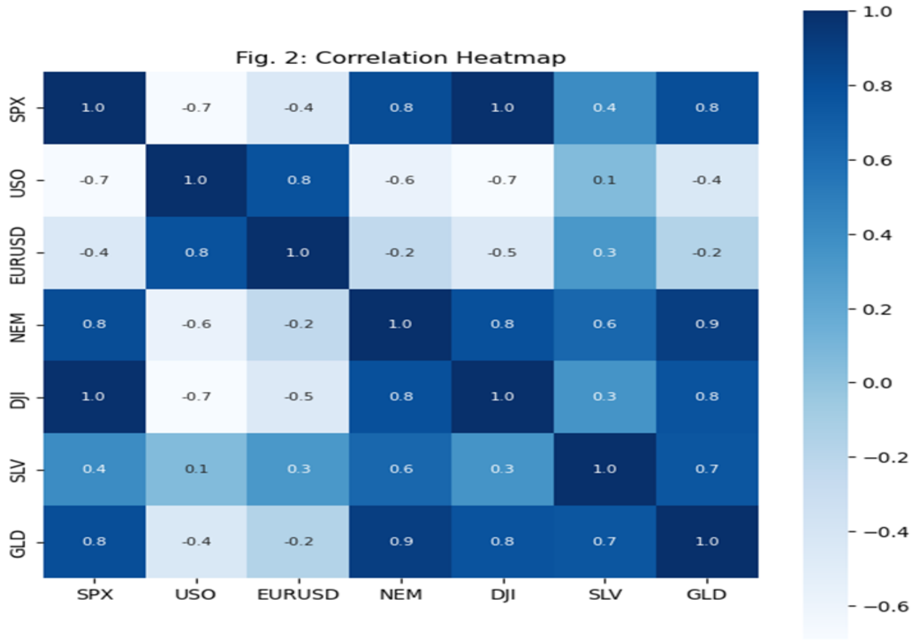

Fig. 2: Correlation Heatmap

	SPX	USO	EURUSD	NEM	DJI	SLV	GLD
SPX	1.0	-0.7	-0.4	0.8	1.0	0.4	0.8
USO	-0.7	1.0	0.8	-0.6	-0.7	0.1	-0.4
EURUSD	-0.4	0.8	1.0	-0.2	-0.5	0.3	-0.2
NEM	0.8	-0.6	-0.2	1.0	0.8	0.6	0.9
DJI	1.0	-0.7	-0.5	0.8	1.0	0.3	0.8
SLV	0.4	0.1	0.3	0.6	0.3	1.0	0.7
GLD	0.8	-0.4	-0.2	0.9	0.8	0.7	1.0

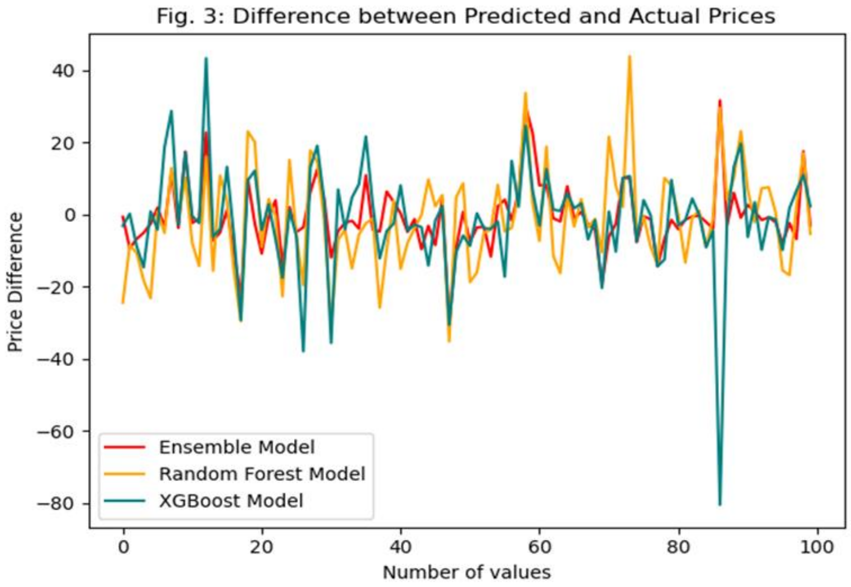

Fig. 3: Difference between Predicted and Actual Prices

Figure 1 Correlation Heatmap

Table 1 presents the evaluation metrics of our ensemble model, which include MAE, MSE, RMSE, R-squared, MAPE, and Max AE. Notably, the ensemble model (ENB) outperforms both the individual Random Forest (RND) and XGBoost (XGB) models across the board. Furthermore, we conducted comparisons with other commonly employed models,[15,16,17,19] including Gradient Boosting (GB), Decision Tree (DT), Adaptive Boosting (ADA) and Linear Regression (LR).

Table 1: Evaluation of Models

Model	MAE	MSE	RMSE	R^2	MAPE	Max AE
RND	10.65	226.3	15.04	0.99	0.74	84.35
XGB	8.47	158	12.57	0.99	0.57	80.50
GB	8.38	142.2	11.92	0.99	0.57	53.25
LR	63.30	6528	80.79	0.99	4.22	250
DT	13.54	413.2	20.32	0.99	0.94	144.2
ADA	23.91	936.1	30.59	0.98	1.66	136.3
ENB	**7.56**	**140.8**	**11.86**	**0.99**	**0.51**	**79.59**

In order to visually compare the performance of our models, we generated Fig. 3, which illustrates the differences between the predicted and actual gold prices. The plot showcases the difference values for the ensemble model, Random Forest model, and XGBoost model over a subset of test data for clarity, further providing compelling evidence of the ensemble model's superior performance compared to the individual models

These findings provide valuable insights for financial decision-making and highlight the potential of such ensemble model in predicting gold prices with greater accuracy.

5. Conclusion

Our study provides a comprehensive analysis of the factors influencing gold prices and offers a robust ensemble-based machine learning approach for accurate price prediction. Our evaluation metrics, including MAE, MSE, RMSE, R-squared, MAPE, and Max AE, validate the superior performance of our ensemble model compared to the individual Random Forest and XGBoost models. The ensemble model consistently achieves lower MSE and maximum absolute deviation, indicating higher accuracy and precision in gold price predictions. The incorporation of the meta-model, which leverages the ensemble features to train a Linear Regression model, adds an additional layer of sophistication and further improves the prediction accuracy.

The significance of our research lies in the potential impact it can have on the financial decision-making process in the precious metals market. By providing valuable insights and predictions, our ensemble model empowers investors and traders to navigate the market with confidence and optimize their investment strategies. Looking ahead, future research efforts will focus on refining and expanding our model, exploring additional input variables, and perhaps incorporating the same for predictions regarding other precious metals as well.

In conclusion, our research significantly advances the field of gold price prediction by conducting a thorough analysis, developing an advanced ensemble model, and providing valuable visualization tools. We expect our findings to have a meaningful impact on the financial industry and inspire further progress in accurately forecasting the prices of precious metals.

References

[1] P. K. Sahu, D. P. Bal, and P. Kundu, “Gold price and exchange rate in pre and during Covid-19,” Resources Policy, vol. 79, December 2022.

[2] M. Al-Ameer, W. Hammad, A. Ismail, and A. Hamdan, “The Relationship of Gold Price with the Stock Market,” IJEEP, Vol. 8, September 2018

[3] S. Coronado, R. Jiménez-Rodríguez, and O. Rojas, "An Empirical Analysis of the Relationships between Crude oil, Gold and Stock Markets," IAEE, Vol. 39, 2018.

[4] J. Sami, "Has the long-run relationship between gold and silver prices really disappeared? Evidence from an emerging market," Resources Policy, Vol. 74, December 2021.

[5] B. Guha and G. Bandyopadhyay, “Gold Price Forecasting Using ARIMA Model,” Journal of Advanced Management Science, vol. 4, no. 2, March 2016 Oxford.

[6] M. Shivani, Ch.Abhilash, T.Divya, Ch.Vasthav, and D.Priyanka, “Gold price prediction,” IJCRT, vol. 9, 6 June 2021.

[7] L. Boongasame, P. Viriyaphol, K. Tassanavipas, and P. Temdee, “Gold price forecasting method using Long Short-Term Memory and the Association Rule,” Journal of Mobile Multimedia, vol. 19_1, 2022..

[8] Zolzaya Luvsandorj, “Comparing Random Forest and Gradient Boosting,” towardsdatascience.com. https://towardsdatascience.com/comparing-random-forest-and-gradient-boosting-d7236b429c15 (accessed Nov. 17, 2022).

[9] Philip Wilkinson, “Non-Linear Regression with Decision Trees and Random Forest,” towardsdatascience.com. https://towardsdatascience.com/non-linear-regression-with-decision-trees-and-random-forest-afae406df27d (accessed Nov. 17, 2022).

[10] I.Akgül, M.Bildirici, and S. Özdemir, " Evaluating the Nonlinear Linkage between Gold Prices and Stock Market Index Using Markov-Switching Bayesian VAR Models,” Procedia - Social and Behavioral Sciences, volume 210, 2 December 2015.

[11] C. Toraman, Ç. Başarır, and M. F. Bayramoglu, "Determination of Factors Affecting the Price of Gold: A Study of MGARCH Model," BERJ, Vol. 2, 2011.

[12] M. Arfaoui, Ay. B. Rejeb, "Oil, Gold, US Dollar and Stock Market interdependencies: a global analytical insight," EJMBE, August 2017.

[13] C. Bentéjac, A. Csörgő, G. Martínez-Muñoz, "A Comparative Analysis of XGBoost," Artificial Intelligence Review, Volume 54, 2019.

[14] Z. Ismail, A. Yahya, and A. Shabri, “Forecasting Gold Prices Using Multiple Linear Regression Method,” American Journal of Applied Sciences, 6 (8): 1509-1514, 2009, ISSN 1546-9239.

[15] Perry Sadorsky, "Predicting Gold and Silver Price Direction Using Tree-Based Classifiers," Journal of Risk and Financial Management, Apr 2021.

[16] A. Wagh, S. Shetty, A. Soman, and D. Maste, "Gold Price Prediction System," IJRASET, Vol. 10, Apr 2022.

[17] A. K. Agarwal, and S. Kumari, "Gold Price Prediction using Machine Learning," IJTSRD, Vol. 4, August 2020.

[18] Radhamani V et al, "Gold Price Prediction Using ML Algorithms," YMER, Vol. 21, July 2022.

[19] P. Baser, J. R. Saini, and N. Baser, "Gold Commodity Price Prediction Using Tree-Based Prediction Models," IJISAE, Vol. 11, January 2023.

[20] Tianqi Chen, and Carlos Guestrin, “XGBoost: A Scalable Tree Boosting System,” presented at the KDD, San Francisco, CA, USA, Aug 16, 2016.

[21] K. A. Manjula, and P. Karthikeyan, “Gold Price Prediction using Ensemble based Machine Learning Techniques,” presented at the ICOEI, Tirunelveli, TN, India, 23-25 April, 2019.