International Journal of Advance and Innovative Research
Volume 12, Issue 2: April - June 2025
ISSN 2394 - 7780

GOLD PRICE ANALYSIS AND FORECASTING OVER TIME

[1]Dr. Dipankar Misra, [2]Adrija Mondal, [3]Sudipto Das, [4]Priyanshu Dey, [5]Srabani Kairi and [6]Shreyan Ray
[1]Professor and [2, 3, 4, 5, 6]BCA, Department of Computer Science & Engineering, JIS University, Kolkata, India

ABSTRACT
The objective of this project is to forecast and predict the movement of gold prices with time series and machine learning models in Python. Since commodity prices, especially gold, are non-stationary and non-linear, this research examines the shortcomings of conventional models and presents advanced methodologies to more effectively capture the dynamic nature of gold. We worked on historical gold price data and performed data pre-processing, feature engineering, visualization, and predictive modelling. Our method involves the ***Random Forest Regressor*** *and comparison with other models. Accuracy of the model is measured with regard to* R^2 *score, which is a comparison measure between different models. The aim is to offer robust forecasting data and solve the biggest problems of gold price forecasting with machine learning models.*

INTRODUCTION

Overview
Gold remains an important commodity in the world of international finance. Gold prices are regulated by a very complex interplay between economic, geopolitical, and market-related factors. Accurate forecasting of gold prices is of extreme importance to investors, economists, and policymakers alike. However, traditional statistical models usually assume linearity as well as stationarity, which are not always true to the nature of actual world commodity data. Such a situation indicates the importance of the use of machine learning models that are capable of dealing with non-linear trends, volatility, and more advanced feature sets.

Objective
The primary objectives of this project are:

1. To collect and pre-process historical gold price information.
2. To illustrate and comprehend the trends, seasonal variation, and price action in gold.
3. To mitigate disadvantages of classical models such as linear regression or ARIMA in dealing with non-linearity and non-stationarity.
4. To create a machine learning model—in this case, a Random Forest Regressor—to predict gold prices.
5. To compare gold price movements and provide detailed forecasting analysis.

METHODOLOGY

1. Collecting Data
We used gold price information spanning 24 years, starting from 2000 to 2024, containing multiple factors, such as opening, high, low, closing prices and adjusted closing prices, from a publicly available data set in Kaggle, downloaded as CSV file sources.

2. Data Pre-processing
• Null values were identified and then resolved.

• Parsing of dates enabled timestamp conversion to set index values.

3. Feature Engineering
Correlation matrix was used to study the correlation among different features. Correlation analysis helps to understand how different indicators relate to each other and the overall gold price trends. If multiple features have a high positive correlation among themselves, including all of them can introduce redundancy. Adjusted closing price was chosen as the primary feature as it is more historically consistent and reliable for long-term analysis.

4. Model Building
Random Forest Regressor was applied due to its power in ensemble-based, non-linear modelling. Apart from that, it also prevents overfitting as it uses multiple decision trees, which themselves are trained on randomly split individual datasets and the predictions of each decision tree are averaged. The dataset was splitted into train and test datasets. Train data were used to train the model in order to predict the "Adj_close" gold price using engineered features. The trained model was evaluated on test data.

ISSN 2394 - 7780

5. Assessment Criteria

R^2 score was used to check the performance of the model. R^2 score is used to measure how well a regression model explains the variability of target variable. It quantifies the "goodness of fit" of a model. R^2 values range from 0 to 1 (or 0% to 100%). A value which is closer to 1 indicates that the model explains a large portion of the data's variability, suggesting a better fit. On the contrary, a value closer to 0 suggests that the model doesn't explain much of the variability and is no good than mean-based prediction model.

These steps provided insight into the accuracy and stability of the model.

LITERATURE SURVEY

1. Methods of Predicting Gold Prices

Time series forecasting methods such as ARIMA, Holt-Winters, and GARCH have traditionally been used in financial forecasting. However, these models are generally linear and stationary in nature, limiting their application to actual gold price data with regime changes, volatility clusters, and complex patterns.

2. Machine Learning in Commodity Forecasting

Research indicates that ensemble techniques, such as Random Forests and Gradient Boosted Trees, are good at capturing nonlinear relationships. Multiple study indicates the strength of Random Forest in dealing with noisy, multivariate data. Deep learning techniques, such as Long Short-Term Memory (LSTM) networks, are also promising but need vast amounts of data and careful parameter tuning.

3. Python Financial Forecasting Tools

- **Pandas:** Used for pre-processing and manipulation of data.
- **Scikit-learn:** Used for implementation of machine learning models.
- **Seaborn & Matplotlib:** Used for trend detection and visualization.

4. Limitations of Existing Models

- **Linearity Assumption:** Many models cannot detect nonlinear trends.
- **Stationarity Requirement:** Real-world data is often non-stationary.
- **Feature Simplicity:** Traditional methods ignore multivariate complexity.
- **Poor Generalization:** Overfitting in standard models lowers future accuracy.

Our research fills these gaps by integrating large-scale preprocessing, feature engineering, and non-linear modelling methods which also prevents overfitting.

FINDINGS AND ANALYSIS

1. Dataset Overview

The gold price dataset used in the project was collected from Kaggle, which is a reliable and popular data source. The dataset has comprehensive gold price information including multiple key indicators like open, close, adjusted close, high, low, and volume.

Adj_Close	Close	High	Low	Open	Volume
273.899994	273.899994	273.899994	273.899994	273.899994	0
278.299988	278.299988	278.299988	274.799988	274.799988	0
277.000000	277.000000	277.000000	277.000000	277.000000	0
275.799988	275.799988	275.799988	275.799988	275.799988	2
274.200012	274.200012	274.200012	274.200012	274.200012	0

2. Dataset Description

The dataset used in the project is a large one, containing gold prices over the period of 24 years, from August 30,2000 to October 30,2024. The dataset reflects gold price trends of certain influential time periods like 2008 global financial crisis and COVID-19 pandemic etc. The dataset includes following columns:

- **Date:** The specific date on which gold prices are recorded.

ISSN 2394 - 7780

- **Open:** The starting price of a particular date.
- **High:** Highest price recorded on a particular date.
- **Low:** Lowest price recorded on a particular date.
- **Close:** The ending price of a particular date.
- **Adjusted Close:** Adjustments made on closing price keeping in mind dividends, stock splits and other corporate actions.
- **Volume:** Total amount of gold traded during a trading session.

The dataset comprehensively reflects gold price movements over the period, capturing various market conditions, trends, and impactful events that have influenced gold prices.

3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is an essential process for understanding the features of a dataset, detecting patterns, trends, and outliers, and preparing the data for predictive modeling. In this section, we conduct several analyses and visualizations to derive insights from global gold price data between August 30, 2000, and October 30, 2024.

	Adj_Close	Close	High	Low	Open	Volume
count	6064.000000	6064.000000	6064.000000	6064.000000	6064.000000	6064.000000
mean	1150.724818	1150.724818	1156.481085	1144.814528	1150.766623	4262.785455
std	569.143985	569.143985	572.181256	565.977679	569.078431	24284.937764
min	255.100006	255.100006	256.100006	255.000000	255.000000	0.000000
25%	628.274994	628.274994	630.875000	626.000000	628.700012	20.000000
50%	1231.199951	1231.199951	1237.450012	1226.049988	1231.049988	105.000000
75%	1605.024963	1605.024963	1614.274994	1592.724976	1604.999969	395.000000
max	2788.500000	2788.500000	2789.000000	2774.600098	2774.600098	386334.000000

4. Statistics

Key statistics are included:

- **Mean:** The average value of each feature.
- **Standard Deviation:** It is a process of measure of the amount of variation.
- **Minimum and Maximum:** The range of values.
- **25th, 50th (Median), and 75th Percentiles:** Quartiles that Provide insights into the distribution of the data.

5. Correlation Matrix

There is a significantly high positive correlation among the "Open," "High," "Low," "Close" and “Adjusted Close” prices. So, including all of them into the model will inevitably introduce redundancy and make the model less efficient. So, adjusted closing price was chosen as the primary feature for the model for its reliability compared to other prices.

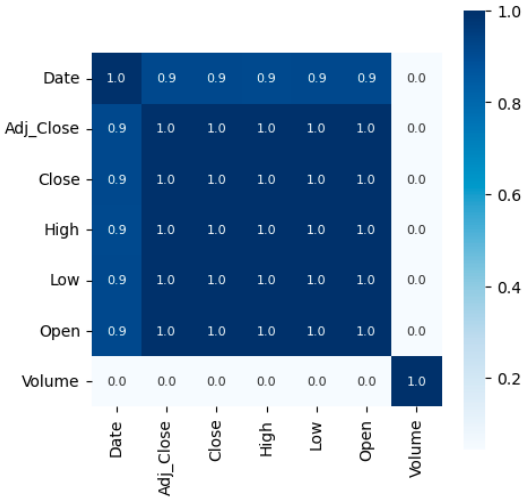
6. Visualization

Visualizations help in understanding the temporal dynamics of the gold prices and identifying trends, seasonality, and outliers.

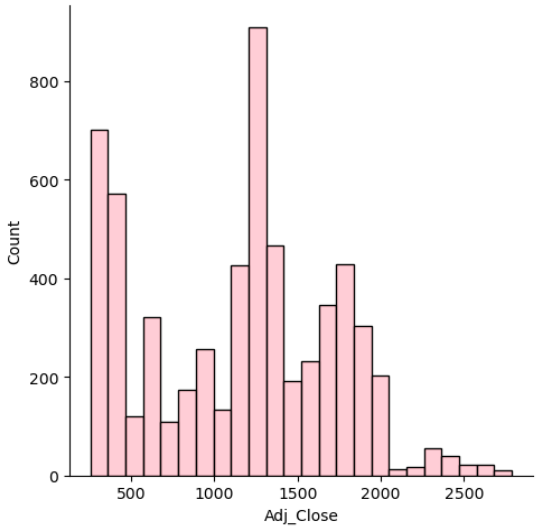
- **Correlation Heatmap**

The following heatmap shows the correlation between different indicators like open, high, low, close, adjusted close, volume to understand their relationship with each other and price trends. Correlation analysis is also essential for feature selection.

ISSN 2394 - 7780



The distribution of adjusted closing prices throughout the years is shown with the help of the following graph.



7. Model Selection and Implementation

We chose Random Forest Regressor for gold price forecasting as it deals non-linear data like financial data well compared to traditional statistical models or linear regression models. It also prevents overfitting by implementing multiple decision trees. Each decision tree is assigned a randomly split subset from the main dataset. After each decision tree makes an individual prediction, all of them are averaged to make a uniform prediction.

The dataset was split into train and test sets. The model was trained on the train dataset and evaluated using the test dataset.

8. Model Assessment

R^2 score was used to check the efficiency of the model. A score closer to 1 means that that the model is well suited for prediction and a closer score to 0 means the model does no better than mean-based prediction. As our score was 0.9999964241664631, we can certainly say that our model is good fit for predicting gold prices.

The Random Forest model captured the short-term and long-term trends extremely well. The feature importance plot revealed that the "Adjusted Close" values were the best predictors.

CONCLUSION

This research demonstrated an integrated approach to gold price forecasting that combines financial insight, time series techniques, and machine learning models. The Random Forest Regressor model outperformed simple models by accurately predicting price movements and reacting to volatility.

Key Takeaways:

- Gold price is driven by multi-factor dynamics that need non-linear modelling.

ISSN 2394 - 7780

- Conventional models are inadequate in unstable, non-stationary environments.
- Random Forest-based machine learning techniques offer enhanced accuracy and interpretability.
- Feature engineering plays a crucial role in boosting predictive performance.

Future Work:

- Macro indicators (e.g., inflation, interest rates) can be added as additional features improve the efficiency of the model. Comparison with other commodities like crude oil or silver prices can be made to find out how those factors influence gold prices.
- We can compare our outcomes with deep learning models such as LSTM and Prophet.
- Interactive dashboards for real-time forecasting can be developed with live data accessed via web scrapping.

REFERENCES

- Box, G. E. P., & Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day.
- Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
- Bhunia, A., & Mukhuti, S. (2013). A study on dynamic relationship between gold price and exchange rate in India. *Asian Journal of Research in Banking and Finance, 3*(9), 1–12.
- Cameron, A. C., & Windmeijer, F. A. G. (1997). An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics, 77*(2), 329–342.
- Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*, 6. https://doi.org/10.1186/s12864-019-6413-7
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica, 50*(4), 987–1007. https://doi.org/10.2307/1912773
- Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research, 15*(1), 3133–3181.
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering, 9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55
- Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice* (2nd ed.). OTexts. https://otexts.com/fpp2/
- Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research, 259*(2), 689–702. https://doi.org/10.1016/j.ejor.2016.10.031
- McKinney, W. (2010). Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, 51–56.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.
- Shah, D., Isah, H., & Zulkernine, F. (2019). Stock market analysis: A review and taxonomy of prediction techniques. *International Journal of Financial Studies, 7*(2), 26. https://doi.org/10.3390/ijfs7020026
- Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software, 6*(60), 3021. https://doi.org/10.21105/joss.03021
- Yuan, Y., & Zhang, L. (2019). Forecasting gold prices with a novel hybrid model. *Resources Policy, 62*, 588–595. https://doi.org/10.1016/j.resourpol.2019.04.001
- Zhang, G., Eddy Patuwo, B., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting, 14*(1), 35–62. https://doi.org/10.1016/S0169-2070(97)00044-7

ISSN 2394 - 7780

- **Vidya G S and Hari V S**, "Gold Price Prediction and Modelling using Deep Learning Techniques", 2020 IEEE Recent Advances in Intelligent Computational Systems (RAICS) | December 03-05,2020 | Trivandrum
- **Hajiyani Aneri Anishbhai and Dr. Dhaval Maheta,** "A STUDY ON FORECASTING GOLD PRICES IN INDIA
- USING ARIMA MODEL", Veer Narmad South Gujarat University
- **"Gold as a strategic asset: 2024 edition Potential risks and challenges"**, World Gold Council
- **Rushikesh Ghule, Abhijeet Gadhave, Manish Dubey, Jyoti Kharade**, "Gold Price Prediction using Machine Learning", August 2022, International Journal of Environmental Engineering 6(6)
- https://github.com/IamMayankThakur/gold-price-analysis
- **Shahriar Shafiee ,Erkan Topal,** "An overview of global gold market and gold price forecasting", Elsevier
- **Manjula Ka, Karthikeyan P,** "Gold Price Prediction using Ensemble based Machine Learning Techniques", April 2019, International Conference on Trends in Electronics and Informatics(ICOEI)