Report: Gold Price Prediction using Random Forest and Gradient Boosting

1. Introduction

- **Objective**: This project aims to predict daily gold prices (GLD) using historical financial data and various economic indicators.
- **Data**: The dataset contains daily records of SPX (S&P 500 Index), GLD (Gold Price), USO (Crude Oil Prices), SLV (Silver Prices), and EUR/USD exchange rates.
- **Models Used**: We implemented and evaluated two machine learning models, Random Forest Regressor and Gradient Boosting Regressor, both selected for their suitability in handling non-linear relationships in time series data.
- **Evaluation Metrics**: The performance of each model was measured using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R2R^2R2 Score.

2. Data Preprocessing

- **Handling Missing Data**: No missing values were present in the dataset after data preprocessing.
- **Date Conversion**: The 'Date' column was converted from object format to datetime format to allow for time-based feature engineering.
- **Feature Engineering**:
    - **Lagged Features**: Created lagged versions of GLD, SPX, and USO to capture short-term dependencies.
    - **Rolling Averages and Standard Deviations**: Calculated moving averages and standard deviations for 7, 14, and 30-day windows to capture trends and volatility.
    - **Percentage Change**: Calculated for SPX to represent short-term momentum.
    - **Cumulative Metrics**: Calculated cumulative averages for SPX to reveal long-term trends.
    - **Ratio Features**: Created SPX-to-GLD ratios to highlight relative performance.
- **Scaling**: Numerical features were standardized to improve model performance and convergence speed.

3. Feature Importance and Correlation Insights

The correlation matrix provides insights into the relationships between the gold price (GLD) and other financial indicators:

- **GLD and SLV (Silver Price)**: Strong positive correlation (0.87), indicating silver prices are closely tied to gold prices, making SLV a significant predictor.
- **GLD and USO (Oil Price)**: Moderate positive correlation (0.18), suggesting that rising oil prices may slightly influence gold prices, potentially due to inflation concerns.
- **GLD and EUR/USD (Euro/USD Exchange Rate)**: Weak correlation (-0.02), indicating minimal direct impact of currency exchange rates on daily gold price predictions.
- **GLD with Lag Features (GLD_lag1, GLD_lag3, GLD_lag7)**: High correlations (0.99) with past gold prices, showing that historical GLD values are strong predictors of future prices.
- **GLD and SPX (S&P 500 Index)**: Very low positive correlation (0.05), suggesting that stock market trends have limited influence on gold prices in this dataset.

Key Observations:

- **Safe-Haven Asset**: Gold often rises during crises, inversely correlated with stock market trends.
- **Oil Impact**: Higher oil prices may slightly boost gold prices due to inflationary concerns but can also favor stocks in economic growth phases.
- **Gold-Silver Link**: Gold and silver prices often move together, though silver is also influenced by industrial demand.

Conclusion: Features like SLV and lagged values of GLD (GLD_lag1, GLD_lag3, GLD_lag7) are valuable predictors of gold prices. Indicators such as SPX and EUR/USD, however, are less impactful. Non-linear models, like Random Forest and Gradient Boosting, are well-suited for capturing these complex relationships.

4.Correlation Analysis of GLD with Other Assets

Here we examine the correlations between GLD (Gold ETF) and various financial instruments, highlighting the nature and strength of these relationships.

Asset	**Correlation Type**	**Correlation Strength**
GLD vs. SPX	Non-linear	Weak negative correlation
GLD vs. USO	Non-linear	Weak positive correlation
GLD vs. SLV	Mostly linear	Strong positive correlation
GLD vs. EUR/USD	Non-linear	Weak negative correlation

Insights

- The correlation between **GLD and SPX** is characterized as non-linear with a **weak negative correlation**, indicating that as the SPX index increases, GLD may not consistently decrease and vice versa.

- The relationship between **GLD and USO** is also non-linear but exhibits a **weak positive correlation**, suggesting a mild tendency for both to move in the same direction.
- A **strong positive correlation** exists between **GLD and SLV**, indicating that these two assets tend to move together, particularly in linear terms.
- The correlation between **GLD and EUR/USD** is non-linear with a **weak negative correlation**, reflecting a complex relationship where fluctuations in one may not directly impact the other.

5. Model Development and Rationale

Why RandomForestRegressor?

- The Random Forest Regressor's ensemble method of multiple decision trees is suitable for capturing the complex, non-linear relationships between gold prices and financial indicators. This robustness against overfitting makes it ideal for volatile time series data like gold prices.

Why GradientBoostingRegressor?

- Gradient Boosting builds trees sequentially, correcting errors from previous trees, making it effective for modeling non-linear interactions and handling gold's volatile patterns.

Model Hyperparameter Tuning

- Both models were tuned with GridSearchCV for optimal parameters:
    - **Random Forest**: n_estimators, max_depth
    - **Gradient Boosting**: n_estimators, learning_rate, max_depth

6. Model Evaluation and Comparison

Performance Metrics

- **Metrics Used**: MAE, MSE, RMSE, and R2R^2R2 score.

Model	MAE	MSE	RMSE	R2R^2R2
Random Forest	0.943	1.819	1.349	0.9964
Gradient Boosting	0.883	1.427	1.194	0.9972

Key Insights

- **Accuracy**: Gradient Boosting has lower MAE and MSE, indicating more accurate predictions overall.
- **Error Handling**: Lower RMSE in Gradient Boosting suggests it handles larger errors better than Random Forest.
- **Variance Explained**: Gradient Boosting's higher R2R^2R2 (0.9972) indicates it explains more variance in the target variable than Random Forest.

Conclusion: The optimized Gradient Boosting model outperforms Random Forest on all metrics, making it the preferred model for this dataset.

7. <u>Feature Importance and Residual Analysis</u>

Feature Importance

- **Key Features:** Both models relied heavily on recent gold price data (GLD_lag1, GLD_lag7, GLD_14d_avg), indicating that recent GLD trends are primary predictors.
- **Other Factors:** SPX and SLV have moderate importance, while USO is less significant.

Residual Plot Analysis

- **Gradient Boosting:** Residuals are centered around zero with a slight spread, suggesting good generalization with minor overfitting.
- **Random Forest:** Residuals are also centered around zero, with minimal overfitting and close alignment of training and cross-validation scores.

Conclusion: Both models show unbiased predictions, but Gradient Boosting demonstrates slightly better robustness against large prediction errors.

8. <u>Error Distribution Analysis</u>

- **Shape:** Both models' error distributions are roughly bell-shaped and centered around zero, indicating normal, unbiased prediction errors.
- **Spread:** Gradient Boosting's error distribution has a tighter spread, showing that it handles prediction errors more effectively.

9. Model Prediction Performance Comparison

In this analysis, we evaluated the performance of two machine learning models: Random Forest and Gradient Boosting, using the Total Absolute Error (TAE) as our metric of choice.

Model	Total Absolute Error
Random Forest	430.74
Gradient Boosting	403.70

Insights

- Gradient Boosting demonstrates a lower Total Absolute Error (TAE) of 403.70, indicating more accurate predictions compared to Random Forest, which has a TAE of 430.74.
- While both models exhibit similar performance levels, Gradient Boosting shows a superior ability to manage errors, particularly in instances of larger deviations from actual values.

10.Overall Comparison and Final Conclusion

- Our analysis found that **recent gold price trends (GLD_lag1, GLD_lag7)** and **silver prices (SLV)** are the most influential features in predicting daily gold prices. The lagged values of GLD highlight the importance of short-term trends, while SLV's strong positive correlation with GLD reflects their shared roles as safe-haven assets.
- **Oil prices (USO)** contribute moderately, indicating inflationary effects, while **SPX (S&P 500 Index)** and **EUR/USD exchange rates** show minimal direct impact. These findings confirm that recent price data and precious metal correlations are key drivers in forecasting gold prices.

- **Gradient Boosting outperformed Random Forest across all key metrics (MAE, MSE, RMSE, and R2R^2R2), suggesting it captures complex, non-linear relationships in the data more effectively.**
- Random Forest demonstrated good generalization, though it was slightly outperformed in accuracy and error handling.

Final Recommendation: The Gradient Boosting Regressor is the better model for predicting gold prices due to its precision and minimized prediction error, especially on larger deviations.