Gold Price Prediction Report

In this project, various regression models are used to predict gold prices based on historical financial indicators such as the S&P 500 index (SPX), oil prices (USO), silver prices (SLV), and the EUR/USD exchange rate.

Dependencies

To run this project, the following Python libraries installed:

- numpy: for numerical operations
- pandas: for data manipulation
- matplotlib and seaborn: for data visualization
- scikit-learn: for machine learning algorithms and metrics

Dataset

dataset

The dataset, gold_price_data.csv, contains daily data for various financial indicators along with the gold price (GLD). Each entry includes:

- **Date**: The date of the entry.
- **SPX**: The S&P 500 index value.
- **USO**: United States Oil price.
- **SLV**: Silver price.
- **EUR/USD**: The exchange rate between the Euro and US Dollar.
- **GLD**: The gold price, which is the target variable.

Data Preprocessing

1. Data Overview

The data loaded from gold_price_data.csv comprises historical gold prices and associated predictors, such as financial indicators and date information. The initial inspection included checking the data's size, structure, and presence of null or duplicated values.

- **Dataset Size**: Retrieved using date.size, indicating the total number of records.
- **Data Types and Null Values**: Using dataset.info() and dataset.isnull().sum(), confirmed that no significant null values were present, ensuring data completeness.
- **Duplicate Records**: Checked with dataset.duplicated().sum(), revealing no duplicate entries.

2. Date Parsing and Feature Engineering

Parsed the Date column into distinct features representing the year, month, day, and quarter.

- **Datetime Conversion**: Converted the Date column to a datetime object, which allows for straightforward time-based feature extraction.
- **Feature Extraction**: Created new columns:
    - **Year** (year): Useful for capturing year-to-year trends.
    - **Month** (month): To examine seasonality or monthly trends.
    - **Day** (day): Although less critical for long-term trends, this provides fine-grain detail.
    - **Quarter** (Quarter): Helps assess quarterly trends.

The Date column was subsequently dropped, as it was redundant after feature extraction.

3. Reordering Columns

To streamline the dataset and enhance interpretability, the columns were reordered, placing the target variable, GLD, at the end of the dataset

4. **Data Exploration and Visualization**

Boxplots: Created boxplots for each numeric feature to identify any outliers and to examine the spread of values. This helps in spotting skewed distributions or extreme values that could impact model performance.

Scatter Plots: Scatter plots were generated for each feature against GLD, the target variable. This helps visually inspect correlations, particularly linear and nonlinear relationships, which can inform model selection.

Correlation Analysis: Calculated the correlation matrix for all numerical features, displaying it as a heatmap to highlight the strength of relationships between predictors and GLD

5. Feature Selection for Modeling

The final step was selecting features (excluding GLD) for the independent variable matrix X and setting GLD as the dependent variable y.

- **X Matrix**: Contains all predictors except for GLD.
- **y Vector**: Contains the target variable, GLD.

This structured and processed data is now ready for model training and evaluation, with potential insights gained from the data exploration phase helping in informed model selection.

Models

1. Linear Regression

- Captures linear relationships.
- **Performance**:
    - R^2 Score: 0.902
    - Mean Absolute Error (MAE): 5.44
    - Mean Squared Error (MSE): 51.58

2. Polynomial Regression

- Captures polynomial relationships.
- **Performance**:
    - R^2 Score: 0.986
    - MAE: 1.03
    - MSE: 1.93

3. Support Vector Regression (SVR)

- Uses an RBF kernel for non-linear relationships.
- **Performance**:
    - R^2 Score: 0.987
    - MAE: 1.95
    - MSE: 7.02

4. Decision Tree

- Non-linear, based on recursive splitting.

- **Performance**:
    - R^2 Score: 0.993
    - MAE: 1.13
    - MSE: 3.68

5. Random Forest

- An ensemble model of decision trees.
- **Performance**:
    - R^2 Score: 0.996
    - MAE: 0.97
    - MSE: 2.21

Cross-Validation

Each model was evaluated with 5-fold cross-validation to confirm consistent performance. This step validated the robustness of our results and minimized overfitting.

Results and Analysis

The Random Forest model demonstrated the best performance, achieving the highest R^2 score (0.996) and lowest error rates:

- **R^2 Score**: Measures variance explained by the model. Random Forest's score of 0.996 indicates it explains nearly all variance.
- **MAE**: Shows the average error magnitude, with Random Forest's MAE (0.97) indicating near-accurate predictions.
- **MSE**: Penalizes larger errors. The Random Forest model's low MSE (2.21) highlights its robustness.

Model	**R^2 Score**	**MAE**	**MSE**
Linear Regression	0.902	5.44	51.58
Polynomial Regression	0.986	1.03	1.93
Support Vector Regression (SVR)	0.987	1.95	7.02
Decision Tree	0.993	1.13	3.68
Random Forest	0.996	0.97	2.21

Conclusion

The Random Forest model proved most accurate, making it the preferred choice for this dataset. This model effectively captures complex patterns within the features and explains nearly all observed variance in gold prices.

Hyperparameter Tuning

The following models were tuned for performance:

- **Polynomial Regression**: Degree 2 was chosen after experimentation to balance complexity and overfitting.
- **SVR**: Used the RBF kernel, which performs well on data with non-linear relationships. Hyperparameters like C and gamma were tuned, with cross-validation ensuring optimal performance.
- **Random Forest**: Number of trees (n_estimators) was set to 10, balancing training time and performance.

Feature Importance in Predicting Gold Prices

1. **S&P 500 Index (SPX)**: Reflects overall market performance and investor sentiment. Typically, there's an inverse relationship between SPX and gold prices; when the stock market declines, gold often increases as a safe-haven asset.
2. **Oil Prices (USO)**: Rising oil prices can lead to higher inflation, prompting investors to buy gold as a hedge. Thus, USO serves as an inflation indicator that can correlate positively with gold prices.
3. **Silver Prices (SLV)**: As a complementary precious metal, silver prices tend to follow gold trends. Increases in SLV often indicate rising gold demand, making it a relevant predictor.
4. **Euro/US Dollar Exchange Rate (EUR/USD)**: A weaker dollar increases gold demand for international buyers, typically leading to higher gold prices. Thus, fluctuations in EUR/USD directly impact gold's market value.