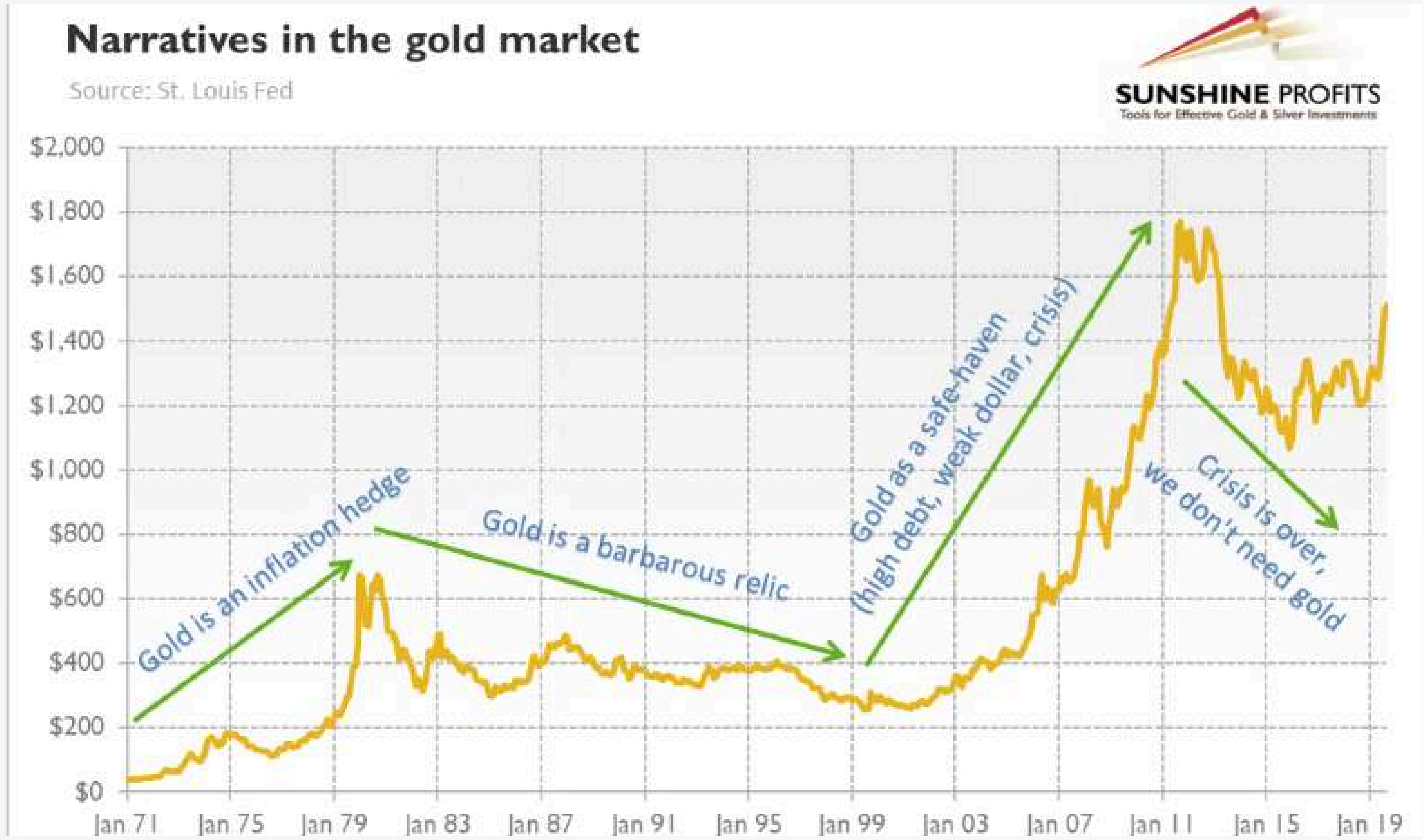
DATA ENGINEERING PROJECT
GOLD PRICE PREDICTION

PROBLEM STATEMENT



Problem:

Predicting gold prices is a challenging task due to the dynamic and volatile nature of the market. Gold prices are influenced by a variety of factors, including global economic conditions, currency exchange rates, oil prices, interest rates, and geopolitical events. These factors make accurate prediction a complex and crucial task for investors, economists, and analysts.

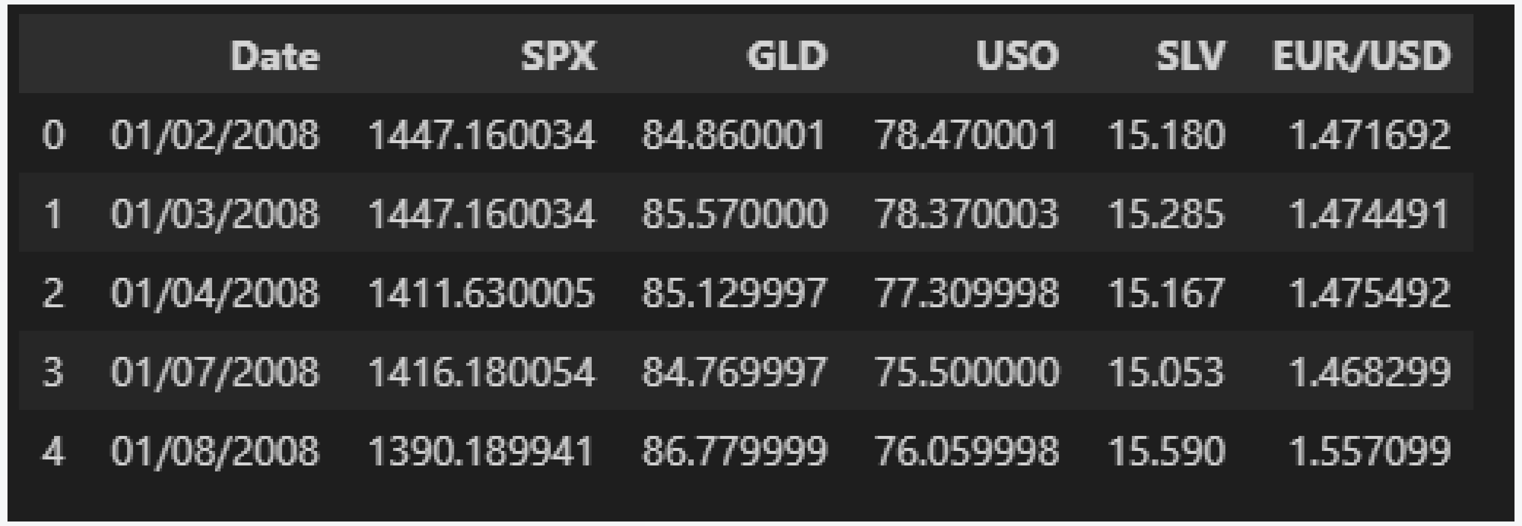
Goal:

The primary goal of this project is to develop a reliable machine learning model to predict future gold prices. This model leverages historical data and key influencing variables to offer insights into price trends and assist in decision-making processes.

DATA OVERVIEW

Datasets Used:

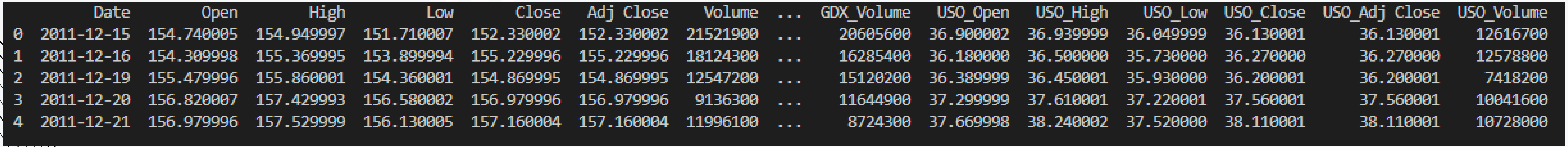
1. **Kaggle Dataset:**
   - Contains historical gold price data with features like date, open, high, low, and close prices.
2. **USO Dataset:**
   - Provides data on crude oil prices and other key economic indicators that influence gold prices.



	Date	SPX	GLD	USO	SLV	EUR/USD
0	01/02/2008	1447.160034	84.860001	78.470001	15.180	1.471692
1	01/03/2008	1447.160034	85.570000	78.370003	15.285	1.474491
2	01/04/2008	1411.630005	85.129997	77.309998	15.167	1.475492
3	01/07/2008	1416.180054	84.769997	75.500000	15.053	1.468299
4	01/08/2008	1390.189941	86.779999	76.059998	15.590	1.557099

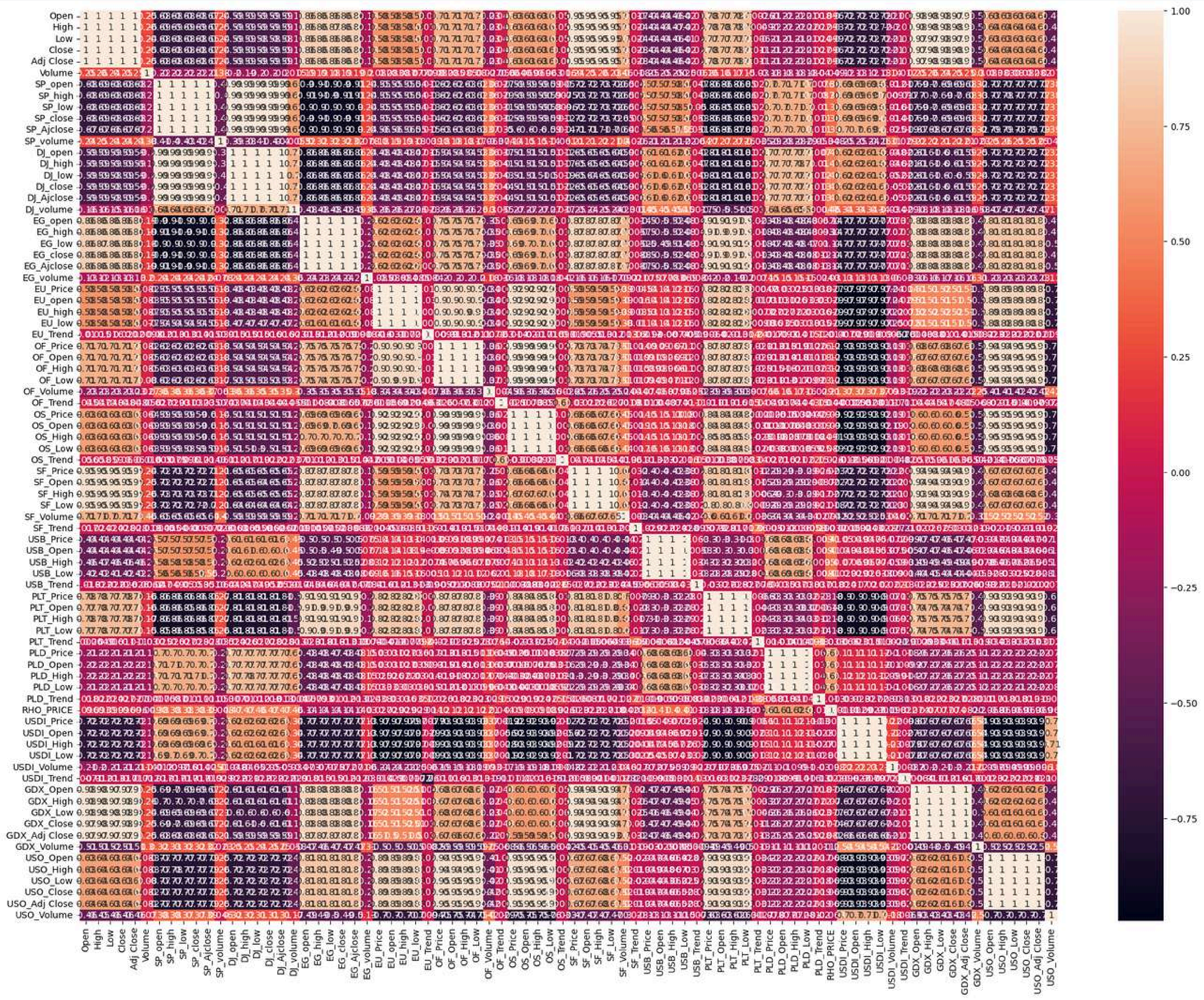
Key Features:

- Gold Prices: Historical values of gold (open, high, low, close).
- Oil Prices: Data from the USO dataset to study correlations.
- Currency Exchange Rates: Exchange rate data for USD and other major currencies.
- Interest Rates: Global interest rates as an economic indicator.
- Additional economic variables influencing gold price trends.

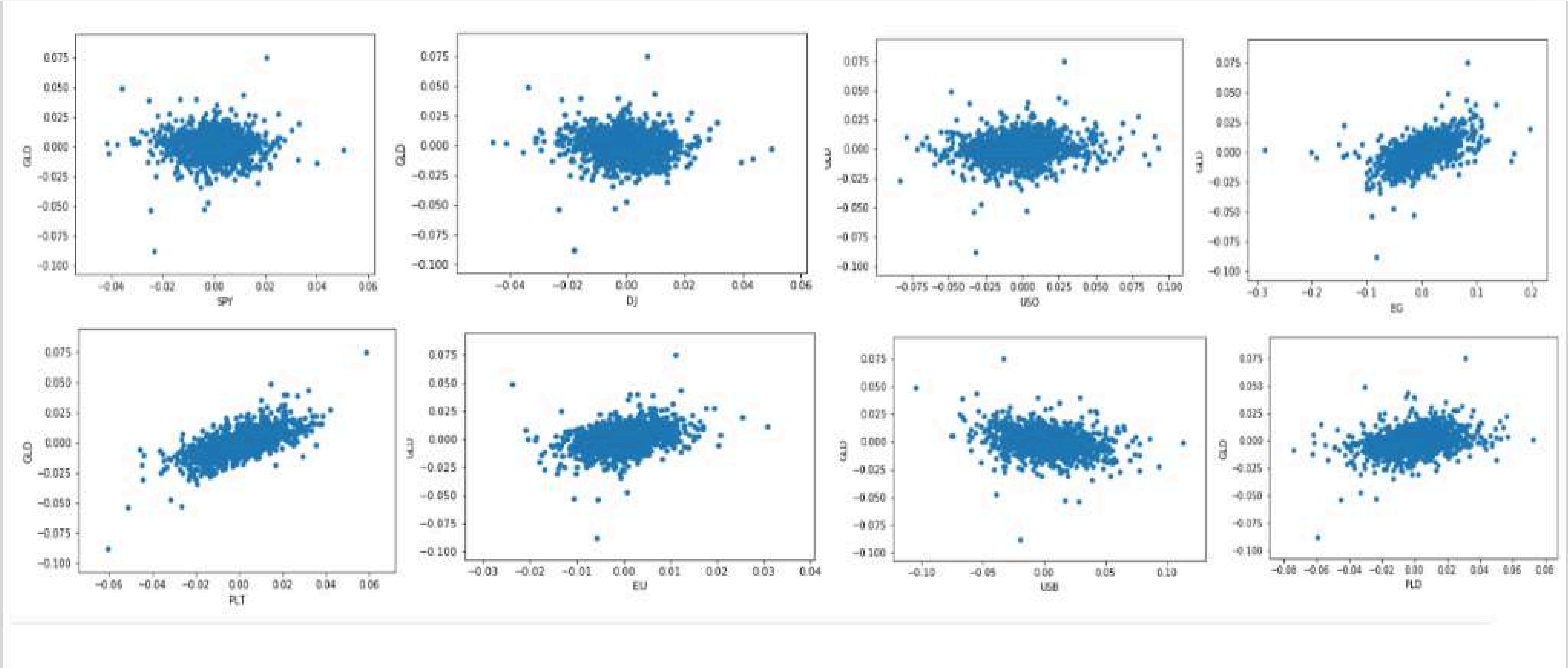


```
         Date        Open        High         Low       Close   Adj Close    Volume  ...  GDX_Volume   USO_Open   USO_High    USO_Low  USO_Close  USO_Adj Close  USO_Volume
0  2011-12-15  154.740005  154.949997  151.710007  152.330002  152.330002  21521900  ...    20605600  36.900002  36.939999  36.049999  36.130001      36.130001    12616700
1  2011-12-16  154.309998  155.369995  153.899994  155.229996  155.229996  18124300  ...    16285400  36.180000  36.500000  35.730000  36.270000      36.270000    12578800
2  2011-12-19  155.479996  155.860001  154.360001  154.869995  154.869995  12547200  ...    15120200  36.389999  36.450001  35.930000  36.200001      36.200001     7418200
3  2011-12-20  156.820007  157.429993  156.580002  156.979996  156.979996   9136300  ...    11644900  37.299999  37.610001  37.220001  37.560001      37.560001    10041600
4  2011-12-21  156.979996  157.529999  156.130005  157.160004  157.160004  11996100  ...     8724300  37.669998  38.240002  37.520000  38.110001      38.110001    10728000
```

DATA ANALYSIS CORRELATION



DATA ANALYSIS SCATTER PLOTS

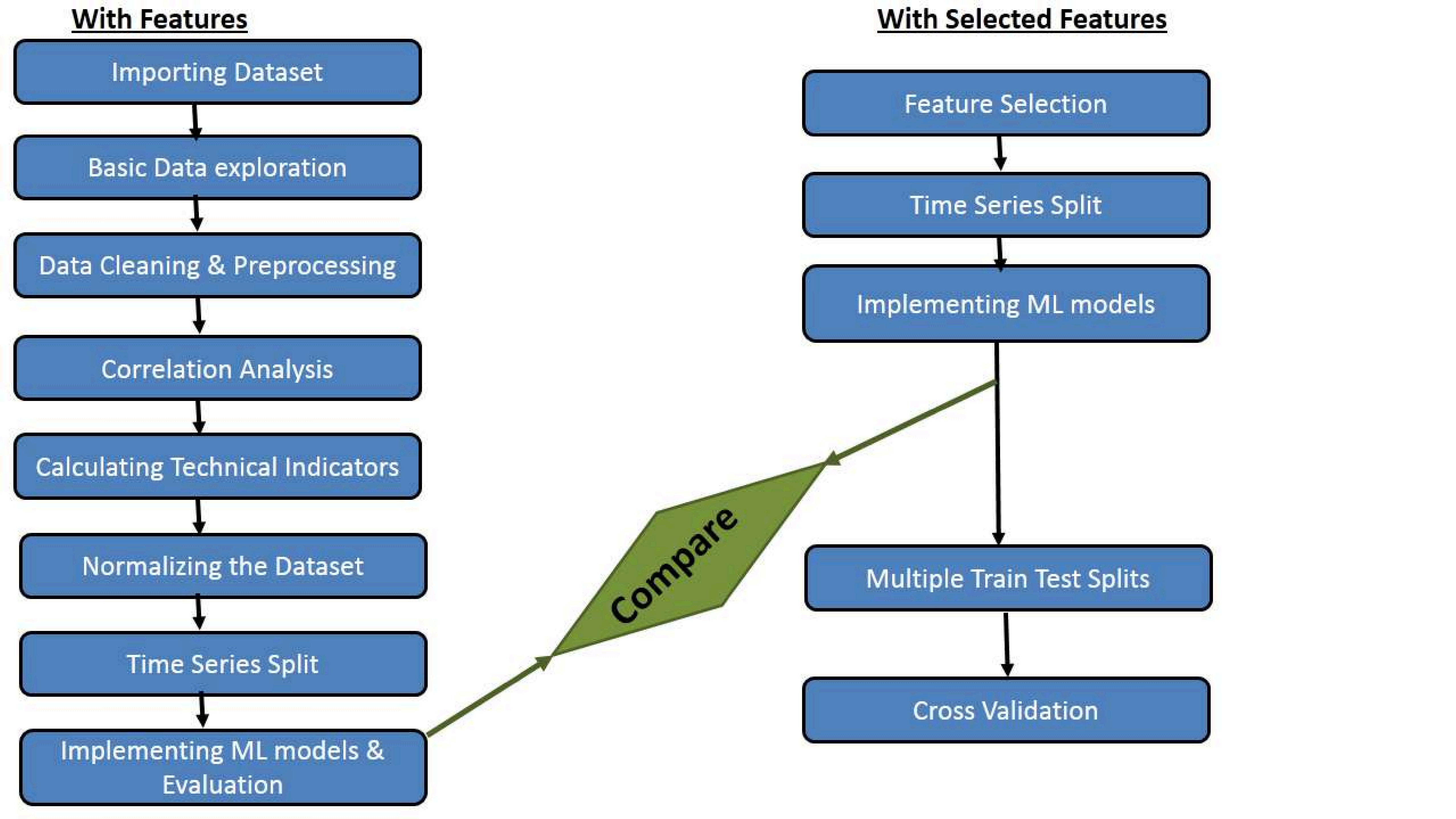


DATA ANALYSIS INSIGHTS

Key Observations:

1. **Correlations:**
   - Strong positive correlation observed between gold prices and currency exchange rates.
   - Moderate correlation found with oil prices and interest rates.
   - Weak or no correlation with certain other economic variables.
2. **Trends:**
   - Seasonal Trends: Gold prices typically peak during certain periods, such as global economic uncertainties or geopolitical tensions.
   - Historical Trends: An overall upward trend observed over the past decades, with occasional sharp fluctuations.
3. **Insights:**
   - A rise in oil prices often corresponds to an increase in gold prices, suggesting potential hedging behavior.
   - Currency depreciation or inflation shows a significant impact on gold prices.

PROJECT WORKFLOW



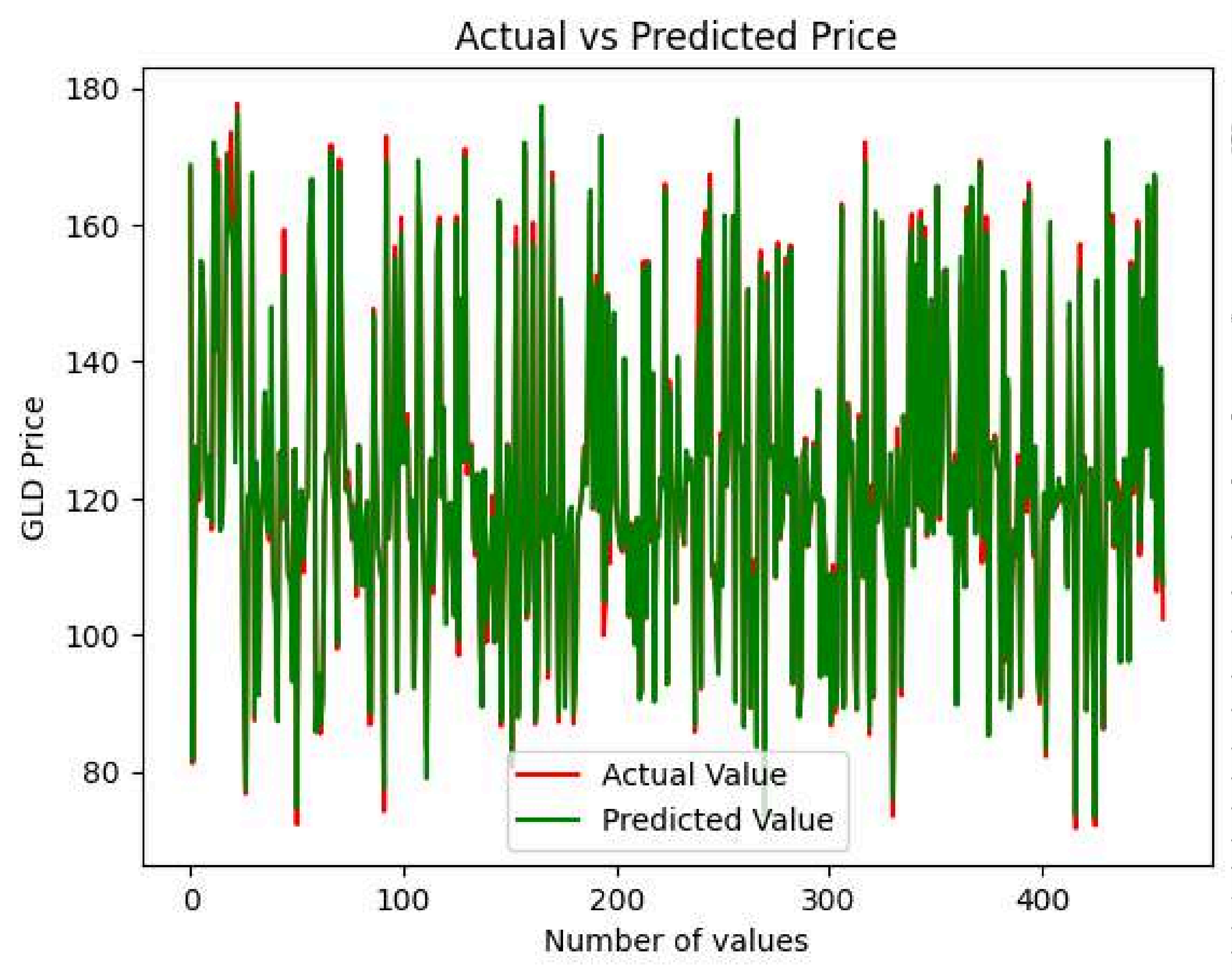
KAGGLE DATASET

Machine Learning Techniques Applied

RANDOM FOREST REGRESSOR

How Random Forest Regressor Helps in Predicting Gold Prices?

1. Handles Complexity:
   - Gold prices are influenced by multiple factors, such as oil prices, currency rates, and interest rates.
   - Random Forest handles complex relationships and interactions between these variables effectively.
2. Ensemble Learning:
   - It combines predictions from multiple decision trees, reducing the risk of overfitting and improving accuracy.
   - Each tree focuses on different parts of the data, capturing diverse patterns.
3. Feature Importance:
   - Random Forest identifies which factors (features) have the most influence on gold prices, helping in better model interpretability.
4. Robustness:
   - It is less sensitive to noise and missing data, making it reliable for volatile markets like gold.
5. Non-Linearity:
   - Captures non-linear relationships between inputs (e.g., oil prices) and output (gold prices), which are common in financial markets.
6. Prediction Accuracy:
   - By aggregating predictions, Random Forest provides more stable and accurate gold price forecasts.



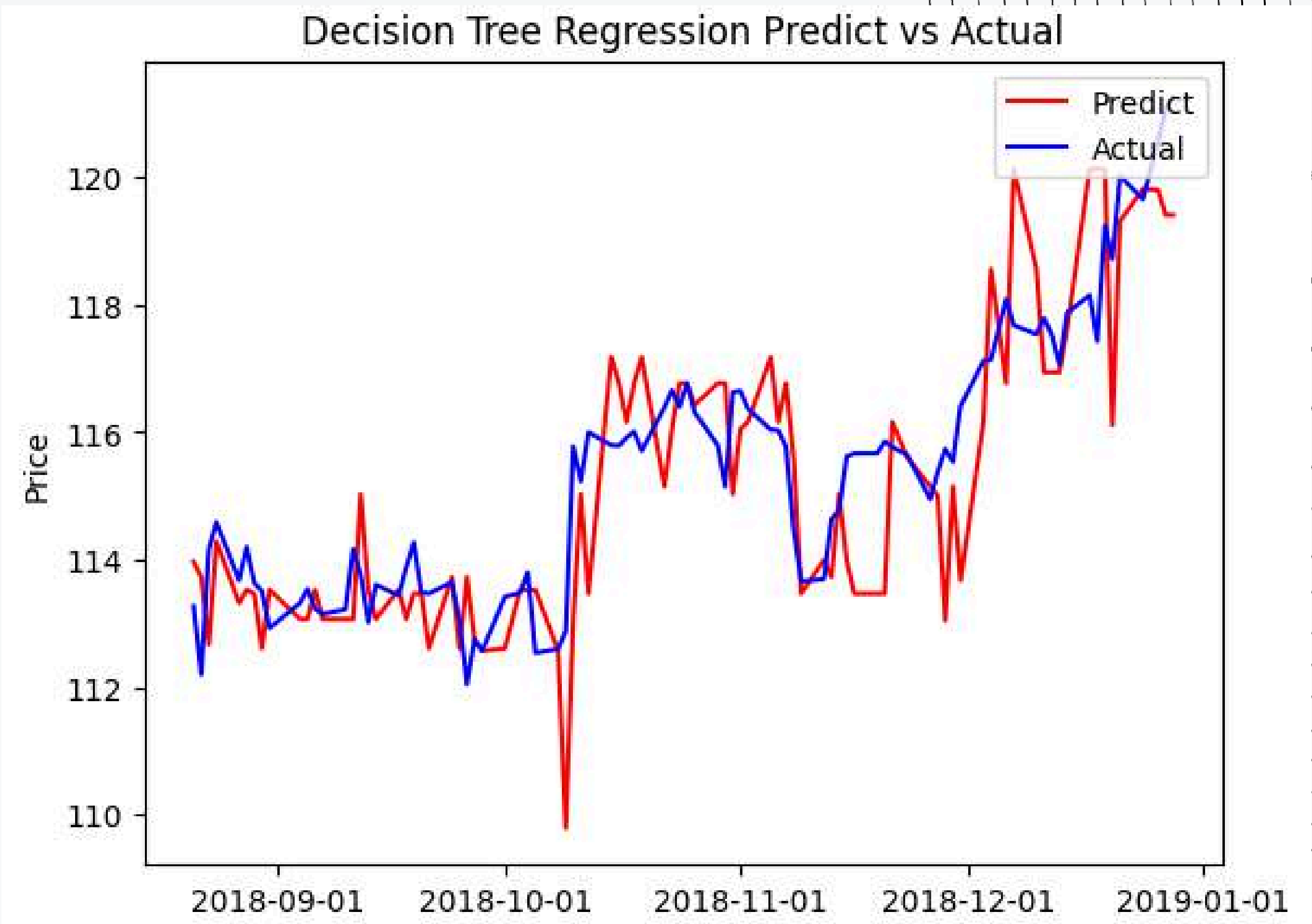
FINAL USO DATASET

Machine Learning Techniques Applied

DECISION TREE REGRESSOR(BENCHMARK MODEL)

How Decision Tree Regressor Helps in Predicting Gold Prices?

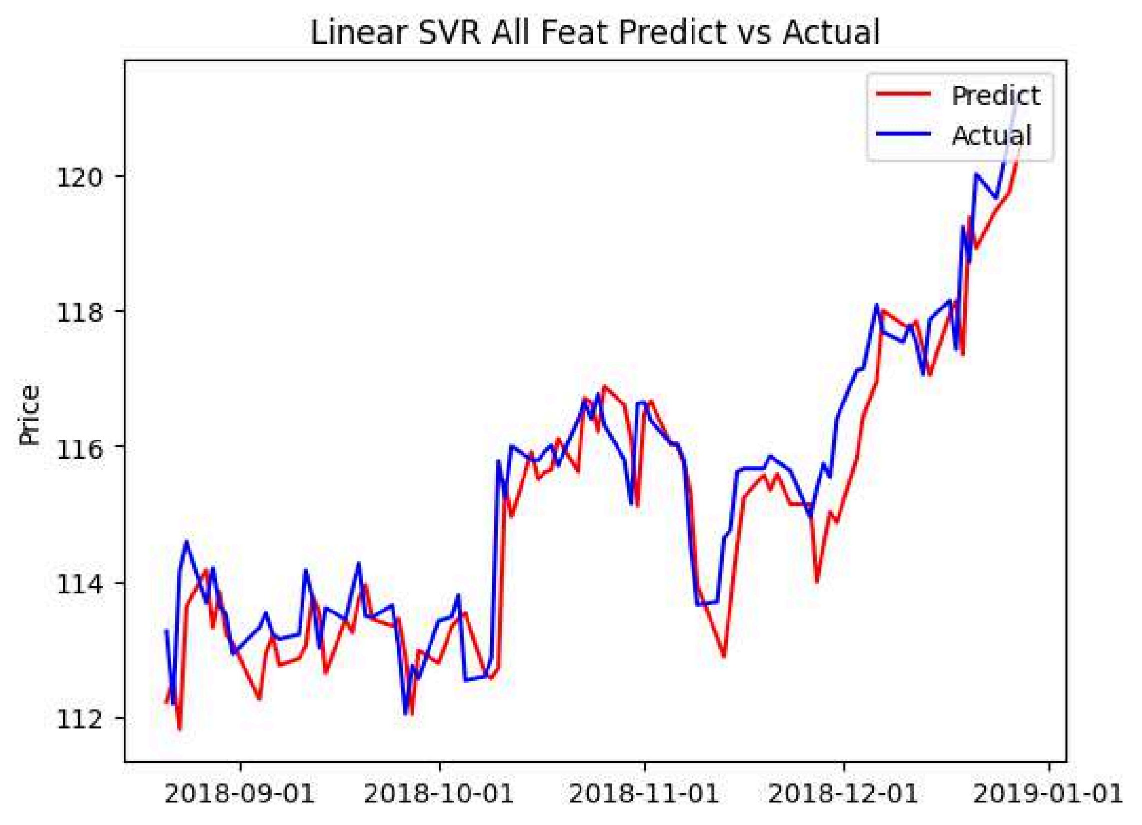
1. **Simple and Intuitive**:
   - A Decision Tree Regressor splits data into smaller, homogeneous groups based on features like oil prices or currency rates.
   - It predicts gold prices by learning simple decision rules inferred from the data.
2. **Handles Non-Linearity**:
   - Gold price fluctuations often exhibit non-linear relationships with influencing factors. Decision Trees can effectively capture these patterns.
3. **Feature Splitting**:
   - It identifies the most critical features (e.g., interest rates, oil prices) at each split, helping understand their impact on gold prices.
4. **Works with Mixed Data**:
   - Handles both numerical (gold price, oil price) and categorical (market trends) data without requiring extensive preprocessing.
5. **Fast and Interpretable**:
   - Decision Trees are computationally efficient and provide clear, interpretable models, which are useful for understanding the factors driving predictions.
6. **Localized Predictions**:
   - By dividing data into regions (e.g., low vs. high oil prices), it specializes in predicting gold prices within specific contexts.

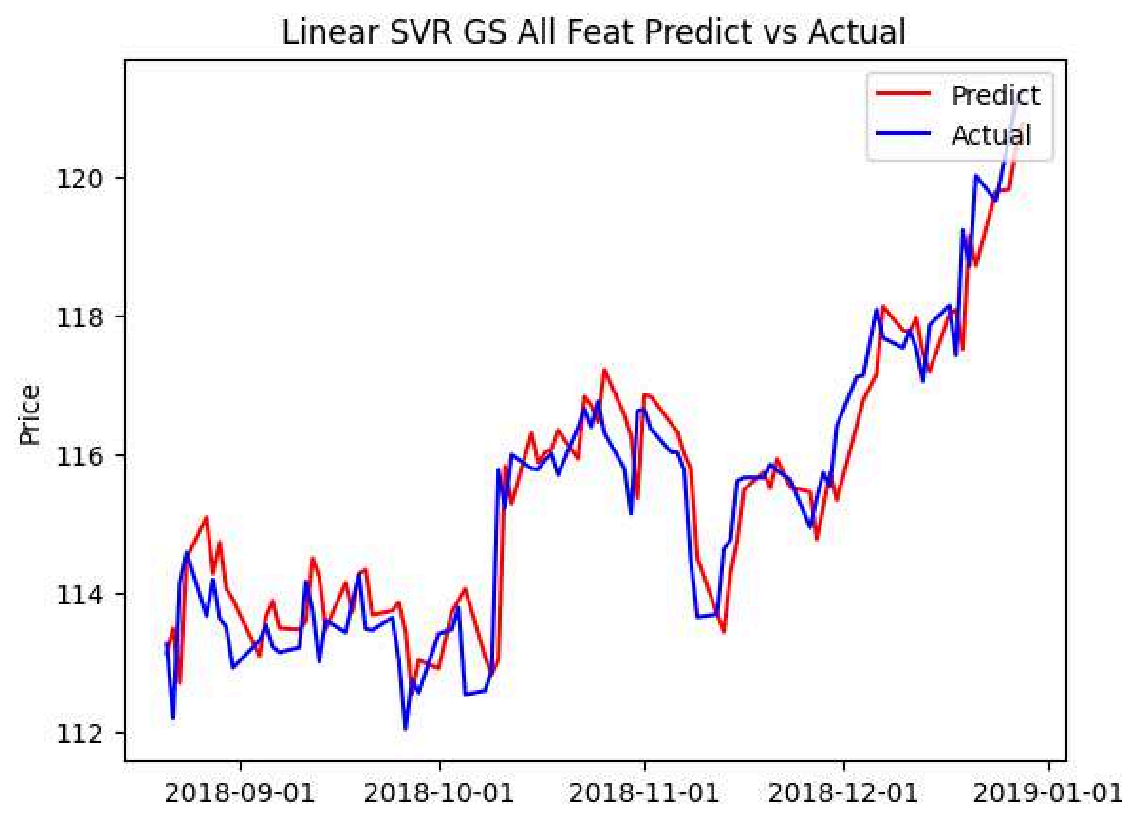


SUPPORT VECTOR REGRESSOR

How Support Vector Regressor (SVR) Helps in Predicting Gold Prices?

1. **Effective for Complex Relationships:**
   - SVR models non-linear relationships between gold prices and features (e.g., oil prices, currency rates) using kernel functions like RBF (Radial Basis Function).
2. **Margin-Based Predictions:**
   - SVR predicts gold prices by finding a function that fits the data within a margin of tolerance (epsilon), avoiding overfitting and focusing on significant trends.
3. **Robust to Outliers:**
   - By focusing on points within the margin, SVR reduces the impact of outliers in the data, which are common in volatile markets like gold.
4. **Kernel Trick:**
   - The ability to use kernels allows SVR to map input features (e.g., interest rates, currency rates) to higher-dimensional spaces, capturing complex patterns.
5. **Balances Bias and Variance:**
   - SVR works well with smaller datasets and avoids overfitting by carefully controlling complexity through regularization (C parameter).
6. **Handles Multi-Feature Dependencies:**
   - SVR considers interactions among various features, making it effective in markets where multiple factors influence gold prices simultaneously.

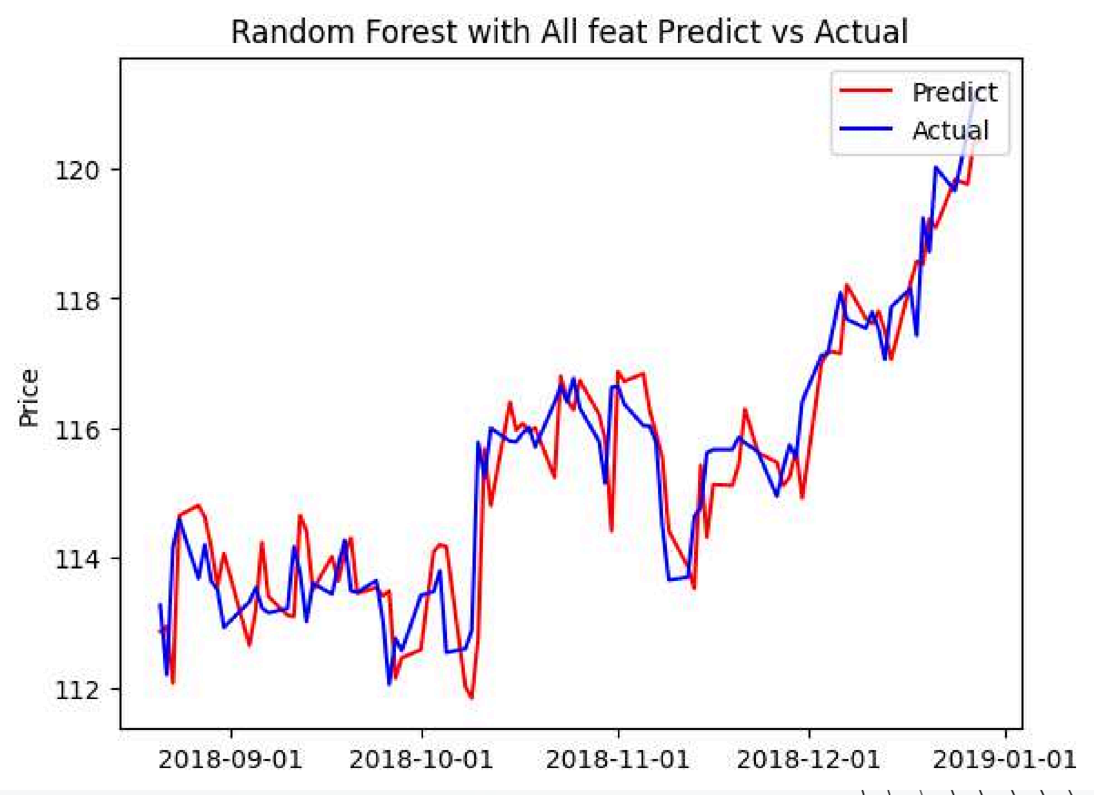


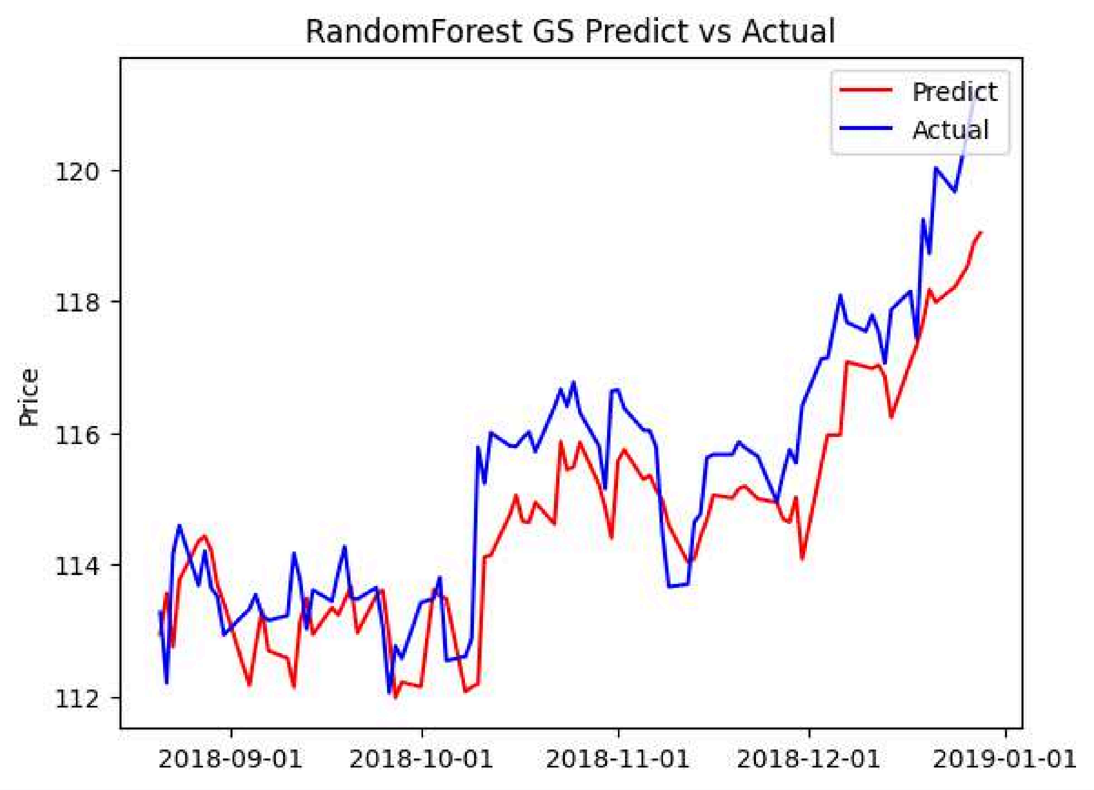


RANDOM FOREST REGRESSOR

How Random Forest Regressor Helps in Predicting Gold Prices?

1. Handles Complexity:
    - Gold prices are influenced by multiple factors, such as oil prices, currency rates, and interest rates.
    - Random Forest handles complex relationships and interactions between these variables effectively.
2. Ensemble Learning:
    - It combines predictions from multiple decision trees, reducing the risk of overfitting and improving accuracy.
    - Each tree focuses on different parts of the data, capturing diverse patterns.
3. Feature Importance:
    - Random Forest identifies which factors (features) have the most influence on gold prices, helping in better model interpretability.
4. Robustness:
    - It is less sensitive to noise and missing data, making it reliable for volatile markets like gold.
5. Non-Linearity:
    - Captures non-linear relationships between inputs (e.g., oil prices) and output (gold prices), which are common in financial markets.
6. Prediction Accuracy:
    - By aggregating predictions, Random Forest provides more stable and accurate gold price forecasts.

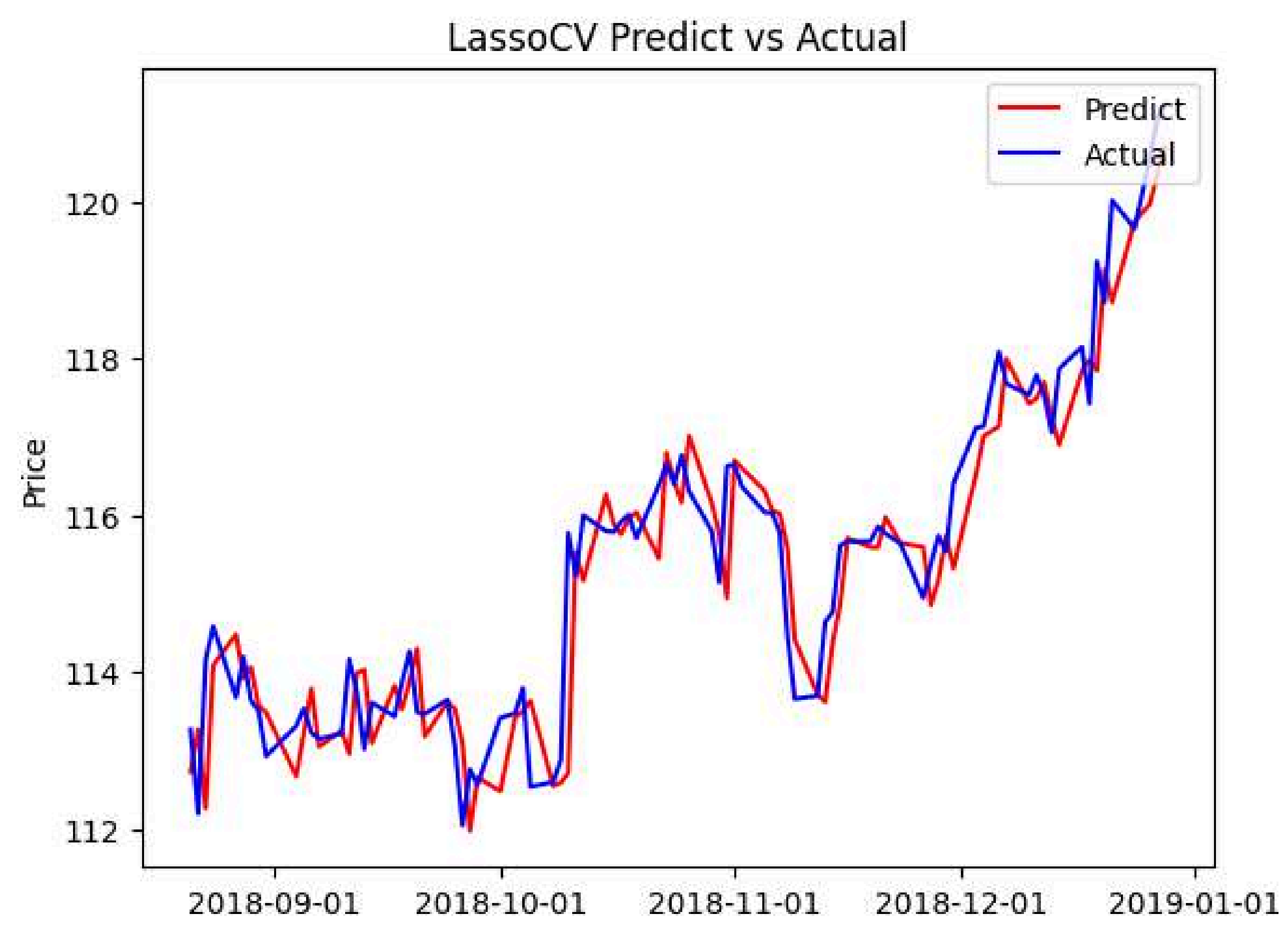

Random Forest with All feat Predict vs Actual


RandomForest GS Predict vs Actual

LASSOCV

LassoCV (Least Absolute Shrinkage and Selection Operator with Cross-Validation):

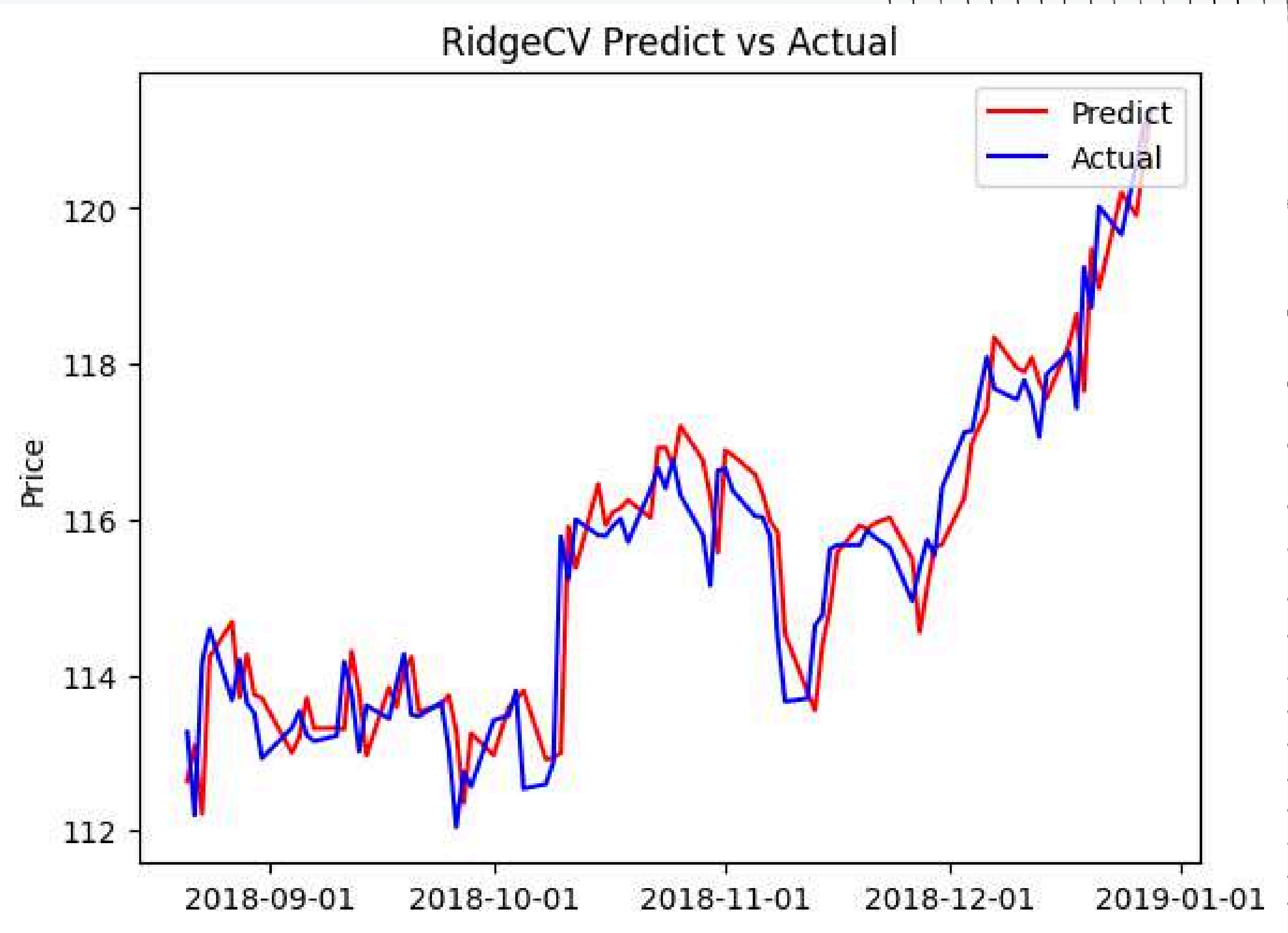
1. **Feature Selection:**
   - LassoCV performs automatic feature selection by shrinking some coefficients to zero, eliminating less important predictors (e.g., irrelevant economic indicators).
2. **Simplicity in Models:**
   - By reducing the number of active features, LassoCV creates simpler, more interpretable models.
3. **Robust to Multicollinearity:**
   - Effectively handles highly correlated features, which is common in economic datasets.
4. **Cross-Validation:**
   - Automatically finds the optimal regularization parameter (α\alphaα) using cross-validation, improving model performance without manual tuning.
5. **Prediction Stability:**
   - Well-suited for datasets with many variables, ensuring a stable and accurate prediction of gold prices.



RIDGECV

RidgeCV (Ridge Regression with Cross-Validation):

1. **Regularization for Stability:**
   - ○ RidgeCV adds a penalty for large coefficients, reducing the risk of overfitting in models with many predictors.
2. **Handles Multicollinearity:**
   - ○ Unlike Lasso, Ridge does not exclude features but instead reduces their impact, making it ideal for datasets where all features provide some predictive value.
3. **Cross-Validation Optimization:**
   - ○ Automatically selects the best regularization parameter (α\alphaα) using cross-validation, ensuring balanced bias-variance trade-off.
4. **Works Well with Noisy Data:**
   - ○ RidgeCV is effective for predicting gold prices in noisy datasets by reducing the impact of minor fluctuations in data.



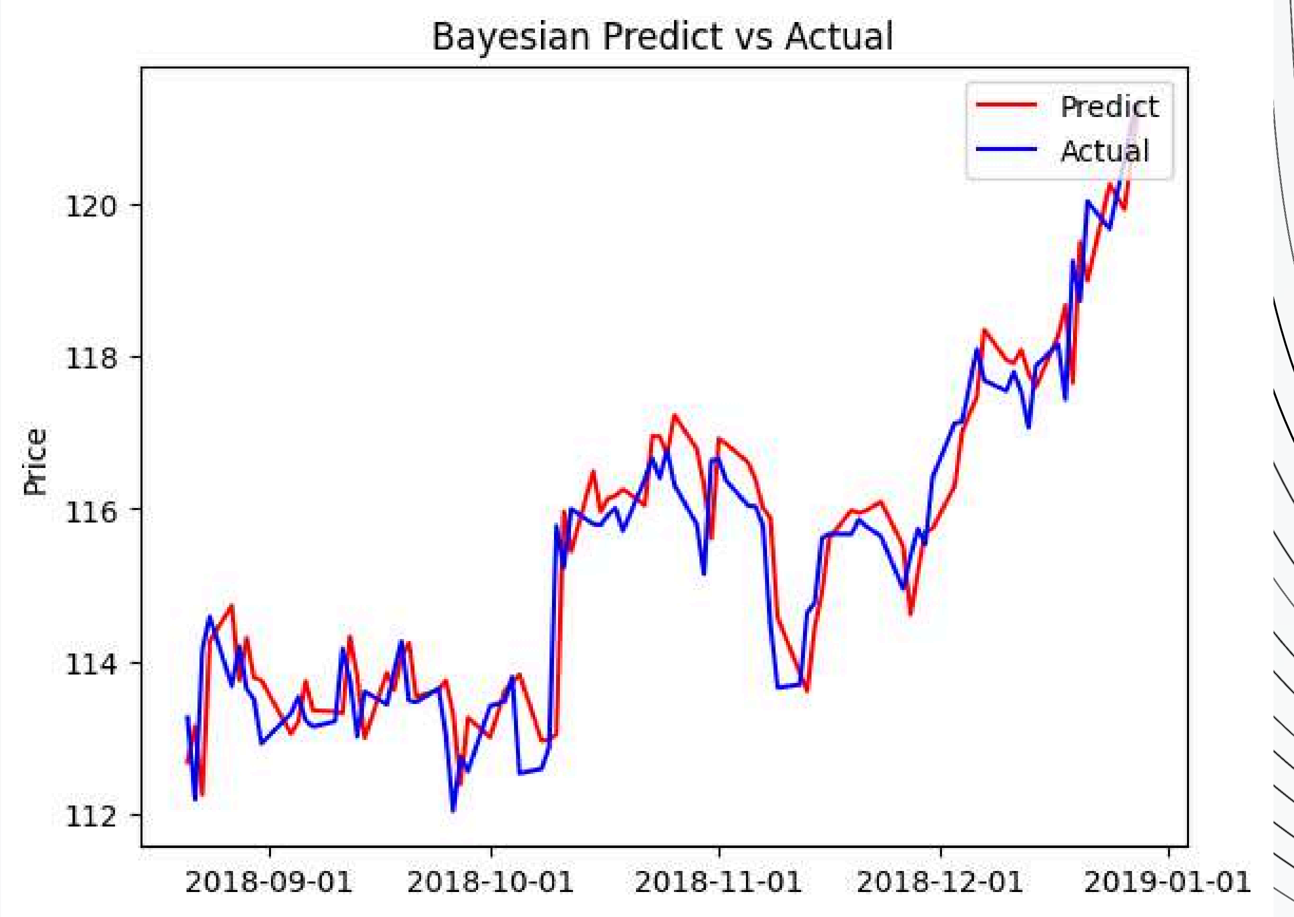
BAYESIAN RIDGE

Overview of Bayesian Ridge Regression:

Bayesian Ridge Regression is a probabilistic approach that extends linear regression by incorporating Bayesian inference. It predicts gold prices while quantifying uncertainty in the predictions.

Key Advantages for Gold Price Prediction:

1. **Regularization:**
   - Similar to Ridge Regression, Bayesian Ridge adds a penalty to large coefficients, preventing overfitting in models with many features like gold prices, oil prices, and exchange rates.
2. **Probabilistic Predictions:**
   - Unlike standard methods, it provides a probability distribution for predictions, offering insights into the confidence level of price estimates.
3. **Automatic Hyperparameter Tuning:**
   - Uses Bayesian inference to estimate the regularization parameters (α\alphaα and λ\lambdaλ) directly from the data, eliminating the need for manual cross-validation.
4. **Handles Multicollinearity:**
   - By modeling coefficients probabilistically, it effectively handles highly correlated features, common in economic datasets.
5. **Adaptability to Noisy Data:**
   - Bayesian Ridge is robust to noisy or uncertain data, making it well-suited for gold price predictions influenced by volatile economic factors.



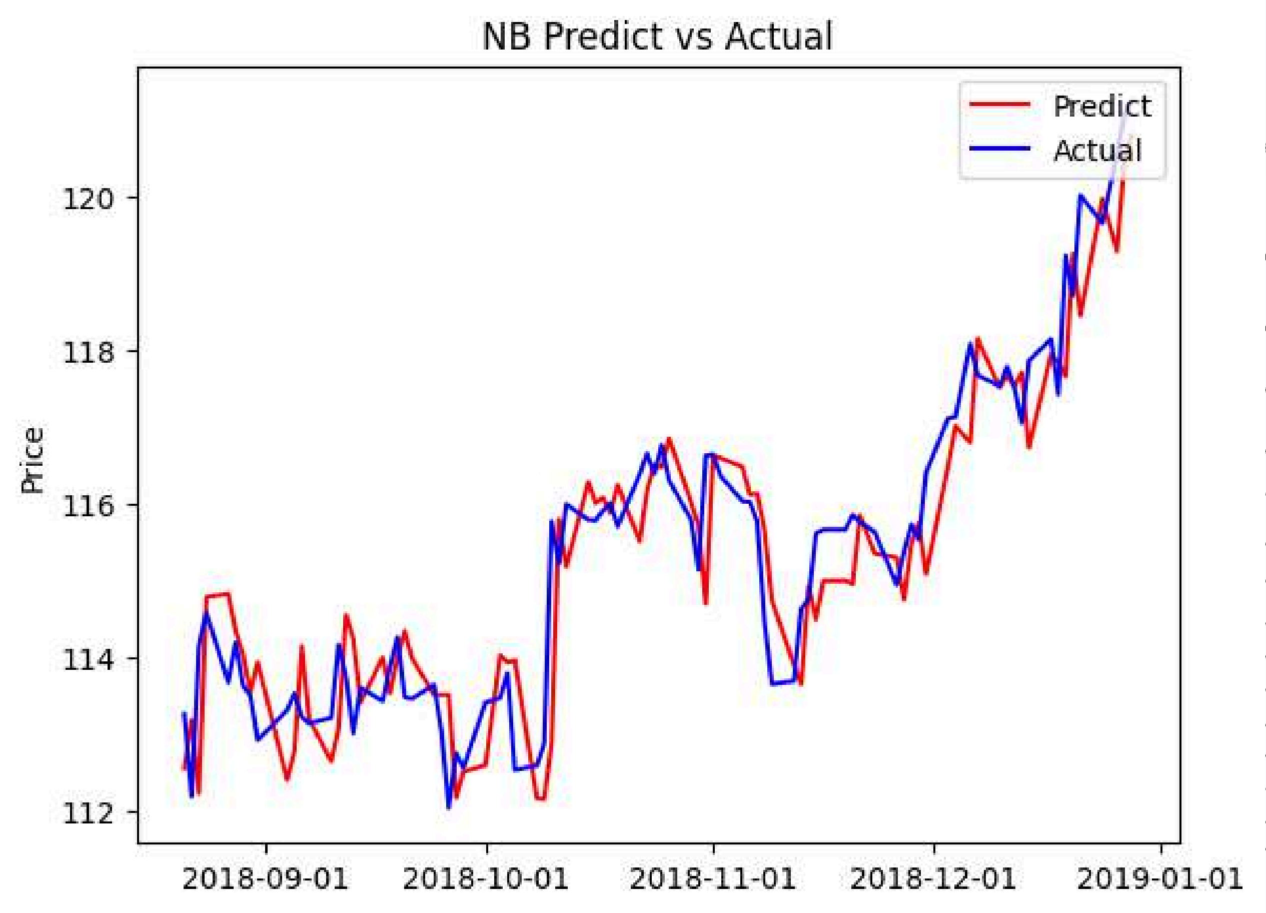
GRADIENT BOOSTING REGRESSOR

Overview of Gradient Boosting Regressor:

Gradient Boosting Regressor (GBR) is a powerful machine learning technique that builds an ensemble of weak prediction models, typically decision trees, to create a strong predictive model. It optimizes model accuracy by iteratively correcting errors from previous models.

Key Advantages for Gold Price Prediction:

1. **Captures Nonlinear Relationships:**
   - GBR excels at capturing complex patterns and nonlinear relationships between gold prices and features like oil prices, interest rates, and currency exchange rates.
2. **Boosting Mechanism:**
   - It sequentially trains decision trees where each subsequent tree focuses on minimizing the errors of the previous ones, leading to high accuracy in predictions.
3. **Feature Importance:**
   - GBR provides insights into which features (e.g., oil prices or currency rates) have the most significant impact on gold prices, aiding interpretability.
4. **Handles Complex Data:**
   - It performs well on datasets with mixed feature types and interactions, common in financial data used for predicting gold prices.
5. **Regularization Techniques:**
   - Techniques like learning rate, tree depth limits, and subsampling prevent overfitting, ensuring robustness in predictions for volatile markets.



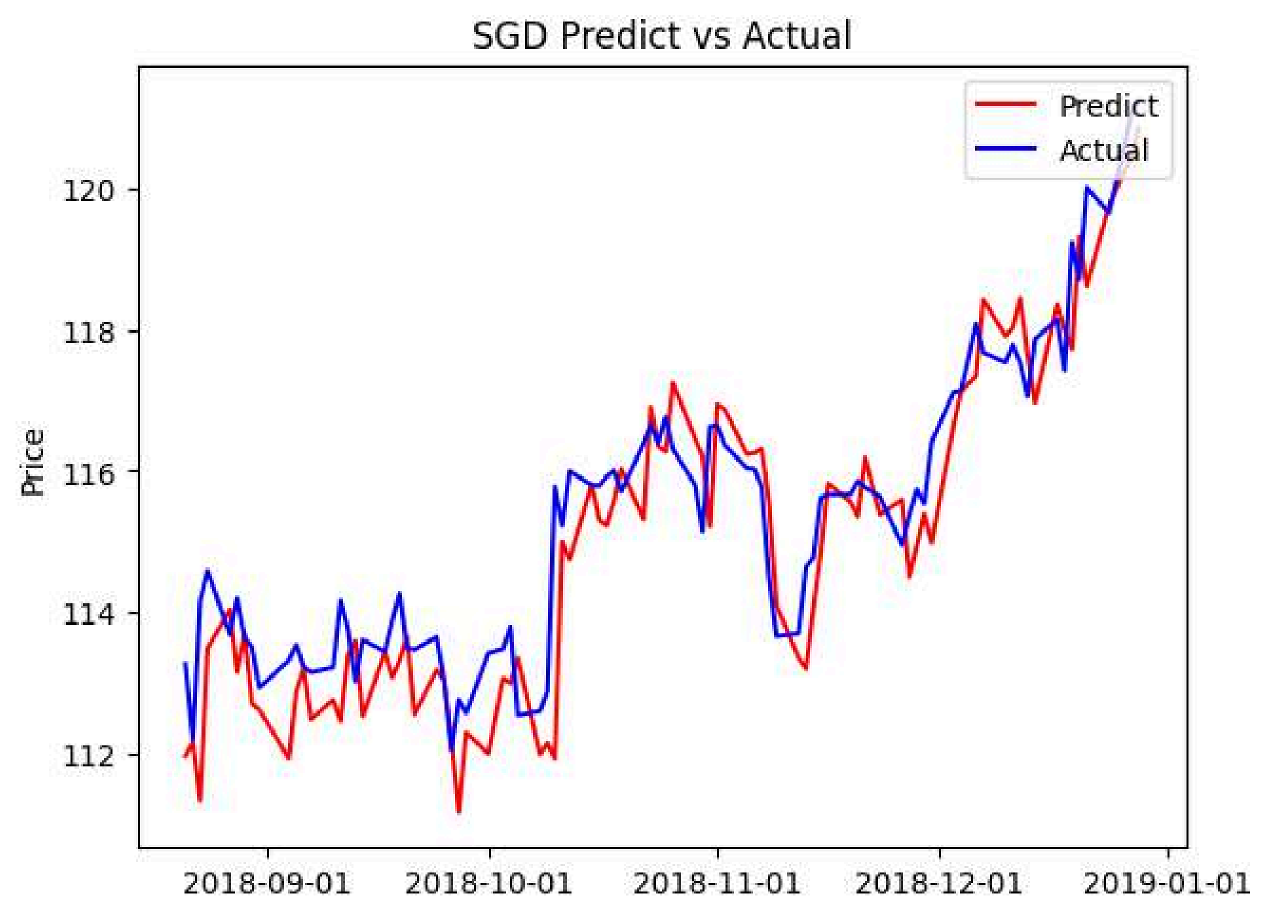
STOCHASTIC GRADIENT DESCENT

Overview of Stochastic Gradient Descent:

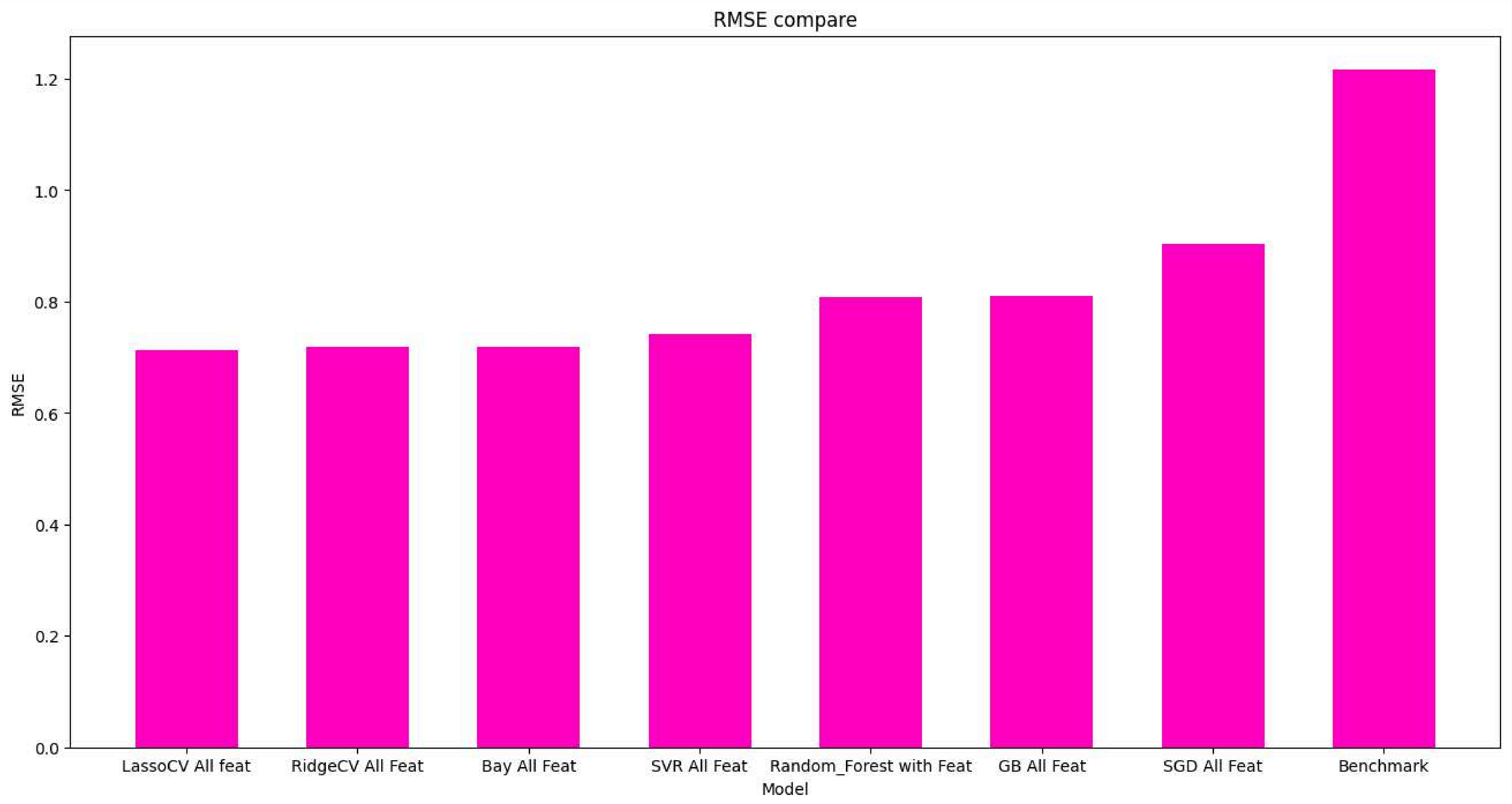
Stochastic Gradient Descent (SGD) is an optimization algorithm used to train machine learning models by minimizing the error between predictions and actual values. It works by iteratively adjusting model parameters (weights) using gradients calculated from the loss function.

Key Advantages for Gold Price Prediction:

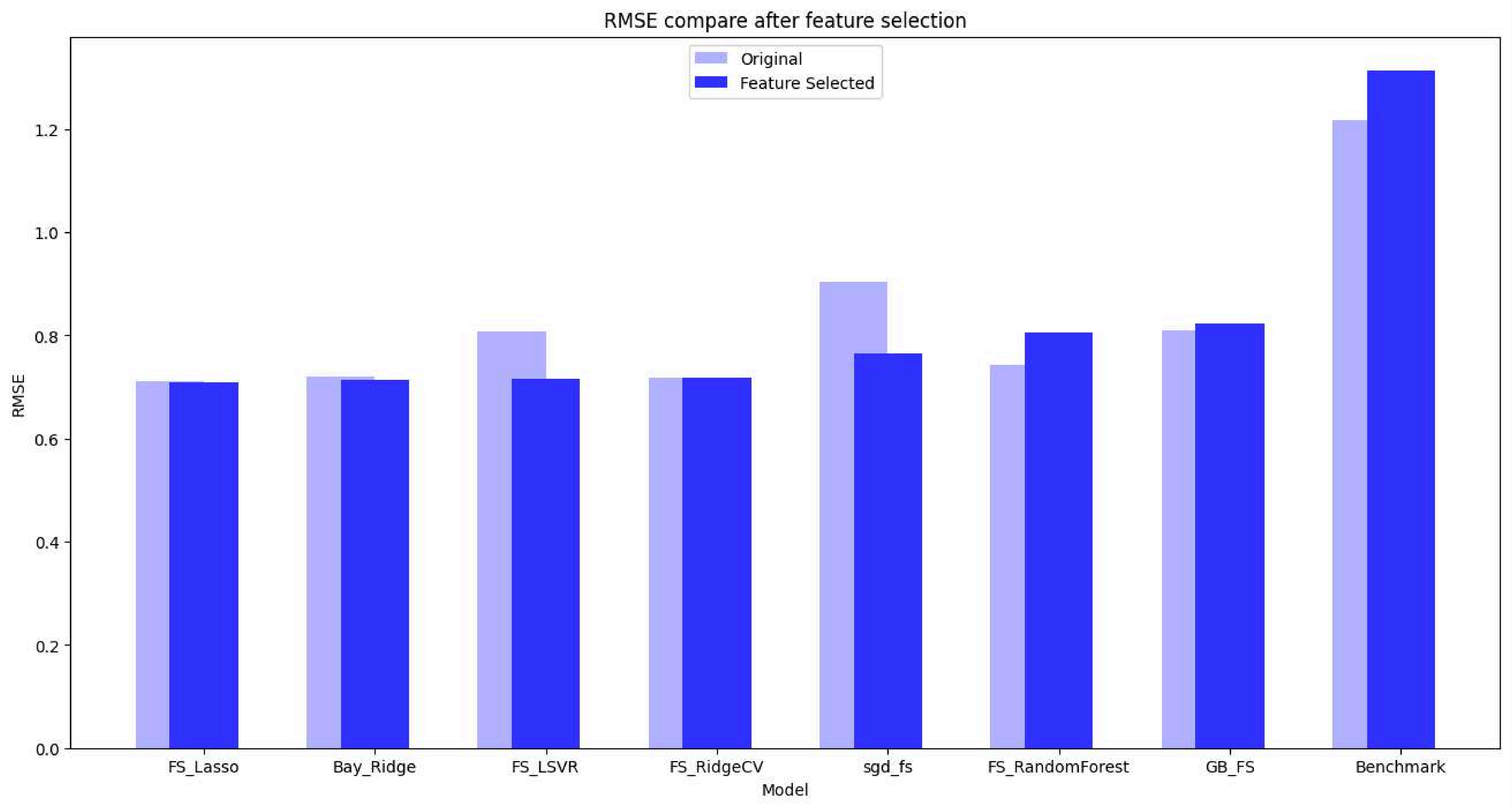
1. **Efficient Training on Large Datasets:**
   - SGD updates model parameters incrementally using small batches of data or even single data points. This makes it efficient for handling large datasets containing gold price histories and economic indicators.
2. **Scalability:**
   - Suitable for complex models with numerous features (e.g., interest rates, oil prices, currency exchange rates), ensuring quick convergence even in data-rich environments.
3. **Flexibility with Regularization:**
   - Supports techniques like L1 (Lasso) and L2 (Ridge) regularization, which help prevent overfitting and improve generalization for volatile data like gold prices.
4. **Adaptive Learning:**
   - Variants like Mini-Batch SGD and momentum-based approaches adaptively adjust learning rates, making SGD robust against noisy financial data.
5. **Simplicity and Speed:**
   - The algorithm's simplicity and computational efficiency make it ideal for iterative refinement of predictive models, especially when frequent updates to the dataset (e.g., daily gold prices) are needed.



COMPARISON OF RMSE OF ALL MODELS

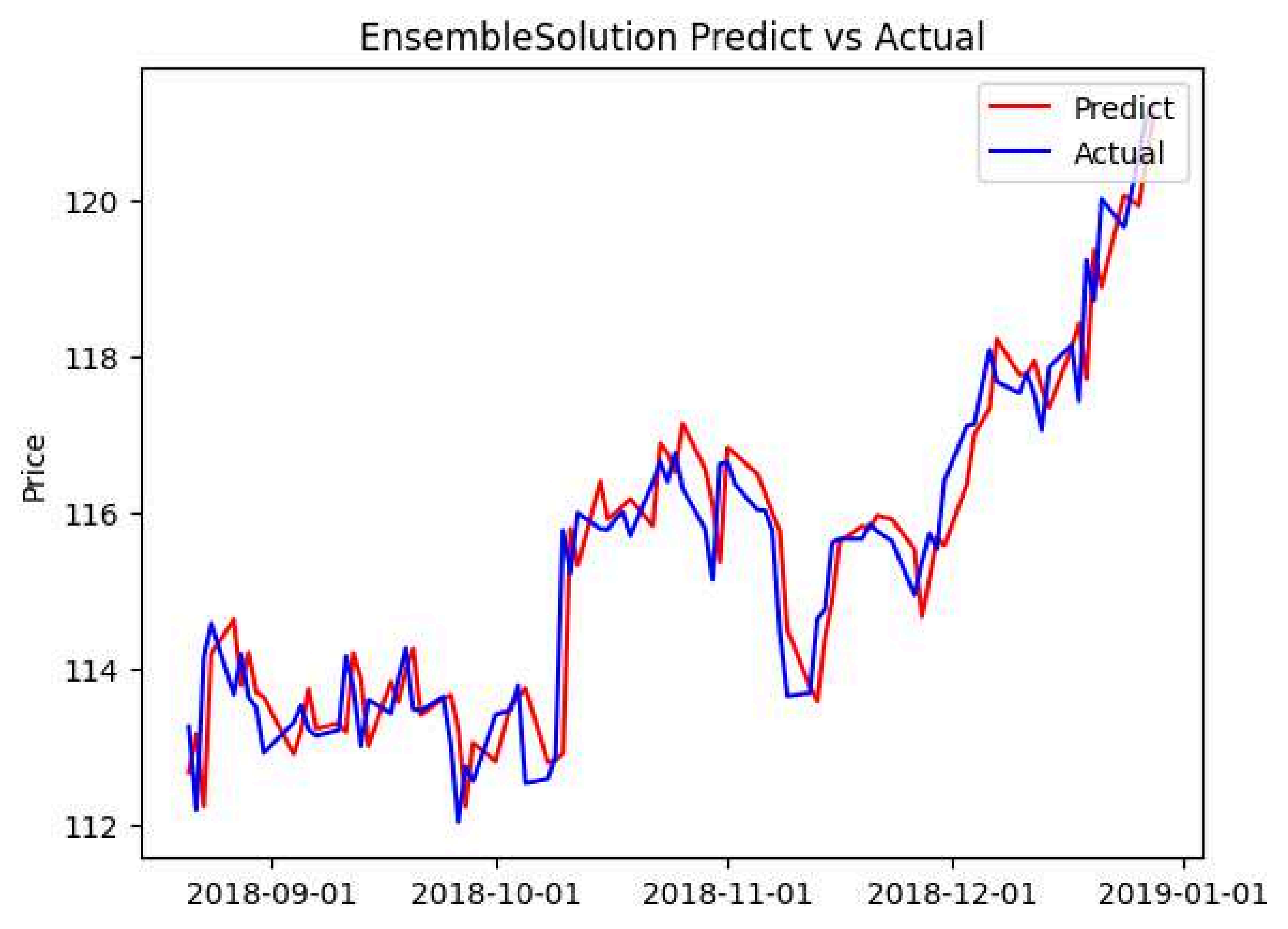


RMSE COMPARISON AFTER FEATURE SELECTION



ENSEMBLE SOLUTION (BEST PERFORMING MODEL)

So now we will ensemble top three performing models i.e, in case of all the features model Lasso,Bayesian ridge and Ridge are the best performing models so we will ensemble these three models while in case of feature selected models we will combine Lasso,Bayesian Ridge and Linear SVR and will compare all the feature ensemble models with feature selected ensemble models.



Results:

Ensemble Solution Model with Original features

RMSE: 0.7007271848739193

R2 score: 0.8869036929481462

DOCKERIZATION

Steps Taken to Dockerize the Project:

1. **Base Image Selection:**
   - Used a lightweight Python base image (python:3.9-slim) to create the container.
   - Ensures compatibility with the required Python version and libraries.
2. **Environment Setup:**
   - A Dockerfile was created to automate the container build process.
   - It specifies the dependencies required for the project, including libraries for data analysis (e.g., pandas, numpy), machine learning (e.g., scikit-learn, gradient boosting), and visualization (e.g., matplotlib).
3. **Application Files and Dependencies:**
   - The project files (e.g., Gold_Price_Prediction.ipynb) and datasets were copied into the container's /app directory.
   - A requirements.txt file was used to list all Python dependencies, which were installed within the container.
4. **Jupyter Lab Integration:**
   - Jupyter Lab was installed and configured in the Docker container to enable interactive exploration of the project.
5. **Exposing Ports:**
   - Port 8888 was exposed to allow access to the Jupyter Lab interface from the host machine.
6. **Simplified Execution:**
   - Once built, the container can be run with a single command, making it easy for anyone to replicate the environment without manual setup.

Benefits of Dockerizing This Project:

1. **Consistency:**
   - Eliminates the "it works on my machine" issue by ensuring the same environment across all platforms.
2. **Portability:**
   - The project can be easily shared and run on any system with Docker installed.
3. **Simplified Setup:**
   - All dependencies and configurations are bundled, reducing setup time for collaborators or deployment.
4. **Isolation:**
   - The container isolates the project from the host system, avoiding conflicts with other applications.
5. **Scalability:**
   - Containers can be deployed to cloud platforms or orchestrated using tools like Kubernetes for scaling predictions.

How to Run the Dockerized Project:

1. **Build the Docker Image:**

```
docker build -t gold-price-prediction .
```

1. **Run the Container:**

```
docker run -p 8888:8888 gold-price-prediction
```

1. **Access Jupyter Lab:**

   Open a web browser and navigate to http://localhost:8888 to interact with the project.

By dockerizing, the project becomes easily reproducible, robust, and ready for collaborative or production use.

LINKS

- **Project Website:** https://goldpricepredictiondeproject.netlify.app/
- **Dockerhub Repository:** https://hub.docker.com/repository/docker/arjunbhattad/gold-price-prediction/general
- **Github Repository:** https://github.com/ArjunBhattad2004/Gold_Price_Prediction

GROUP 3
GROUP MEMBERS

Arjun Bhattad	B22AI051
Dev Pandya	B22AI016
Chirag Kumar	B22AI056

THANK YOU!