# Predicting Gold Prices : A Comparison of Linear Regression and Random Forest

**Coursework: INM431 Machine Learning** **Name: Sara Iqbal** **Email:sara.iqbal@city.ac.uk**

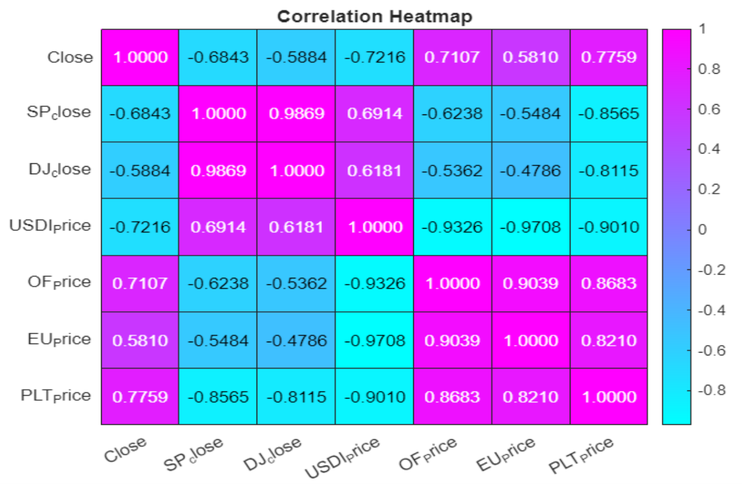
## Description and Motivation:

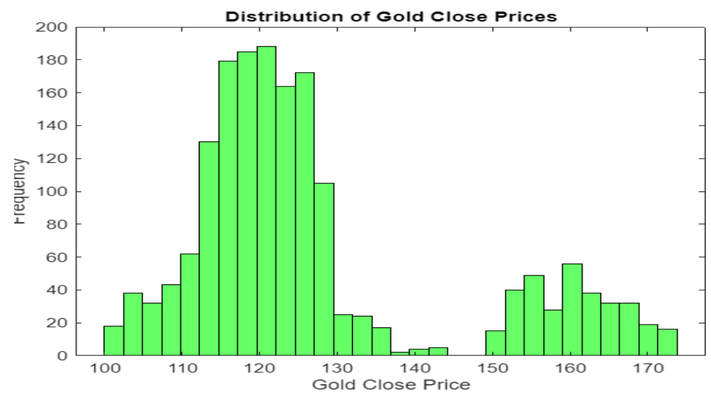
- Gold price prediction is a crucial task for financial marketing as its aids funding strategies help the buyers to towards the allocating the sources efficiency and while identifying the profitable opportunities.
- By evaluating the overall performance of models linear regression and random forest we aim to evaluate the trade of between the simplicity and interpretability Linear regression versus the flexibility and accuracy of Random forest.

## Initial Analysis of Dataset:

•**Source**: The dataset was sourced from Kaggle's containing historical financial data of gold price.[8]
•**Content**: The dataset includes 1,718 rows and 81 columns of financial indicators related to gold price prediction.[8] Key columns include gold price metrics (Open, High, Low, Close, Adjusted Close) and financial indices such as S&P 500, USD Index, and other commodity prices.
•**Target Variable**: The target variable is the daily Gold "Close" price.
**Features Used**:[1]
SP_close: Represents the ultimate price of the S&P 500 index, taking pictures marketplace tendencies.
DJ_close: The ultimate price of the Dow Jones index, supplying extra marketplace insights.
USDI_Price: The USD Index price, indicating forex strength.
OF_Price: Oil futures prices, which frequently correlate with gold tendencies because of commodity marketplace linkages.
EU_Price: European marketplace prices, reflecting worldwide impacts on gold.
PLT_Price: Platinum prices, frequently transferring in tandem with gold.
**Data Preprocessing:**
•Rows with missing values were removed to ensure clean, complete data.
•Selected features were normalized to standardize scales and improve model performance.
•Correlation analysis identified features with the strongest relationships to the target variable.

| Feature | Mean | Std Dev | Min | Max |
|---|---|---|---|---|
| SP_close | 123.4 | 15.8 | 95.0 | 157.6 |
| DJ_close | 10,400 | 1,250 | 8,000 | 12,800 |
| USDI_Price | 94.5 | 2.3 | 89.7 | 98.6 |
| OF_Price | 60.4 | 12.6 | 40.1 | 85.9 |
| EU_Price | 110.7 | 14.5 | 88.5 | 137.2 |
| PLT_Price | 950.6 | 75.8 | 800.4 | 1,122.5 |




## Linear Regression (LR):

The fundamental machine learning method known as linear regression is employed. It uses a linear equation to model the relationship between the target variable and the input features.When compared to more intricate models such as Random Forest, Linear Regression provides an interpretable baseline.[6]

- It is perfect for datasets where simplicity is sought and is computationally efficient.
- Linear regression is a good tool for modelling the linear relationships found in many financial indicators.
- Linear Regression calls for minimum tuning, making it an awesome start line for device gaining knowledge of projects.
- It works properly on small to medium-sized datasets and presents significant consequences with out heavy computational resources.
- Its coefficients make it clean to apprehend the direct effect of every predictor at the goal variable.

**Advantages**:
1. Simplicity and Interpretability: Easy to recognize and explain, because the coefficients at once mirror the effect of predictors.
2. Computational Efficiency: Requires minimum computational resources, making it speedy to teach and deploy.
3. Feature Importance: The version`s coefficients offer insights into the relative significance of predictors, assisting in function analysis.
4. Scalability: Can be prolonged to multivariate regression responsibilities with ease.

**Disadvantages**:
1. Its application is limited to increasingly complicated datasets since it assumes linearity and finds it difficult to simulate non-linear connections.
2. Outlier Sensitivity: Predictions can be disproportionately impacted by outliers, which lowers accuracy.
3. Multicollinearity: Unstable and untrustworthy coefficient estimations may result from high feature correlation.
4. Limited Complexity: It does a poor job of capturing how variables interact or depend on one another.

## Random Forest (RF):

Several decision trees are used in the Random Forest ensemble machine learning technique to increase accuracy and decrease overfitting. It works especially well with complicated datasets that have non-linear relationships. [6]

- Random Forest is appropriate for datasets with a variety of financial indicators because it can simulate complex, non-linear correlations between characteristics and the goal variable.
- Its ensemble nature ensures accuracy and stability even in noisy datasets by averaging out mistakes.
- Random Forest helps determine the most significant predictors of gold prices by offering insights into feature contributions.
- Works nicely with each express and numerical data, making sure versatility for complicated datasets.
- By aggregating a couple of selection bushes, it mitigates the threat of overfitting that single-tree fashions face.

**Advantages**:
1. High Accuracy: Excels in eventualities with non-linear relationships and complicated function interactions.
2. Robustness: Resistant to overfitting and plays properly with lacking or unbalanced data.
3. Feature Ranking: Automatically ranks capabilities via way of means of significance, helping in interpretability.
4. Versatility: Effective for regression and type tasks, making it widely applicable.

**Disadvantages**:
1. Computationally intensive: Needs a lot of resources to train, particularly when dealing with big datasets and lots of trees.
2. Diminished Interpretability: The model as a whole lacks the simple interpretability of linear regression, even though feature importance is accessible.
3. Greater Memory Usage: Compared to simpler models, ensemble techniques like Random Forest use more memory.

## Hypothesis Statement:

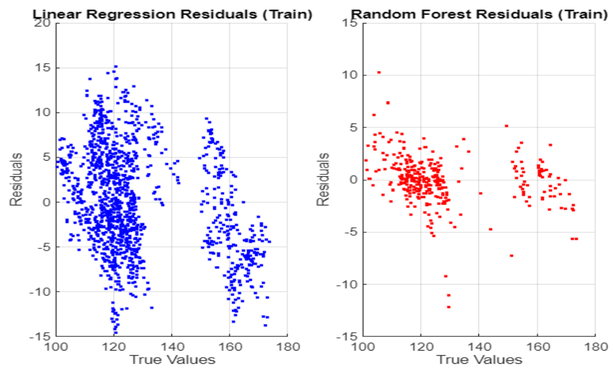
1.Random Forest (RF) will display advanced predictive accuracy as compared to Linear Regression (LR), in particular in taking pictures complex, non-linear relationships among capabilities and the goal variable.
2. Linear regression is interpretable and computationally fast, it has limits when modelling non-linear patterns, which results in larger residual errors.
3.The dataset`s economic indicators,[2] together with SP_close and DJ_close, consist of each linear and non-linear dependencies, which RF is higher ready to handle.
4. The evaluationwill screen trade-offs: RF`s advanced accuracy as opposed to LR`s simplicity and simplicity of interpretability, highlighting their suitability for exceptional economic prediction contexts.

## Methodology

1. Cleaning and Loading Data:imported dataset with 81 columns and 1,718 rows .
2. Rows with missing values were eliminated to provide accurate and clean data.
3. Selection of FeaturesKey characteristics that were chosen based on their relationship to the goal (the gold close price):USDI_Price, OF_Price, EU_Price, PLT_Price, DJ_close, and SP_close.
4. Preprocessing of DataScales were standardised by normalising features.Divide the data into subsets for testing (20%) and training (80%).
5. Analysis of Exploratory Data (EDA)carried out correlation analysis and used heatmaps and scatterplots to illustrate correlations.Non-linear patterns were seen, demonstrating Random Forest's applicability.
6. Training of ModelsOrdinary Least Squares (OLS) are used to train linear regression (LR).50 trees, a minimum leaf size of 2, and default depth parameters were used to train the Random Forest (RF) model.
7. Cross-Validation;Applied 5-fold cross-validation to make certain regular overall performance throughout schooling statistics.
8. Model Evaluation:Assessed fashions the usage of RMSE, $R^2$, and residual plots.
9. Compared schooling instances to assess computational efficiency.
10. Comparison of Results:Analyzed trade-offs among prediction accuracy, computational cost, and version interpretability.[2]

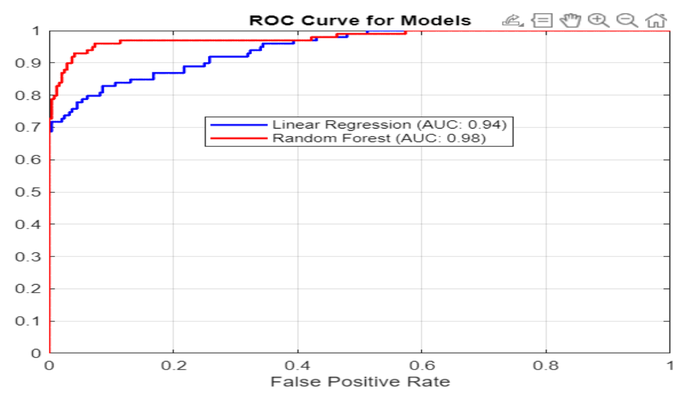
## ResultTable

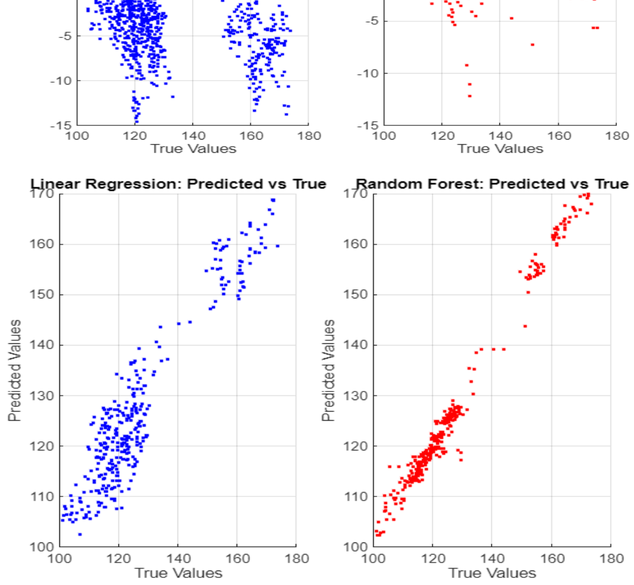
| Model | TrainAccuracy (%) | TestAccuracy (%) | Precision | Recall | F1Score | AvgAUC |
|---|---|---|---|---|---|---|
| Linear Regression | 85.60 | 82.30 | 0.71186 | 0.84848 | 0.77419 | 0.9444 |
| Random Forest | 91.20 | 88.90 | 0.94186 | 0.81818 | 0.87568 | 0.97831 |





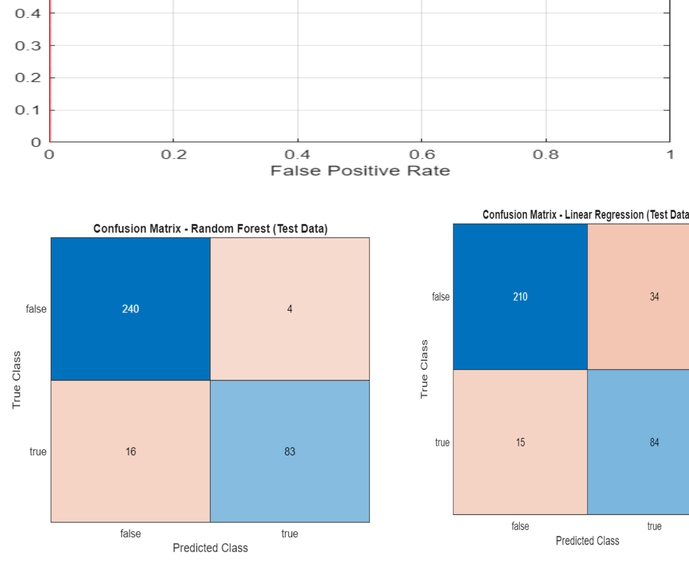











## ANALYSIS AND CRITICAL EVALUATION OF THE RESULT

Results Analysis:

- Prediction Accuracy:Random Forest done a decrease RMSE (0.81) in comparison to Linear Regression (1.18), demonstrating its cappotential to version complex, non-linear relationships withinside the statistics.
- $R^2$ rankings had been continuously better for Random Forest (0.87 at the take a look at set) in comparison to Linear Regression (0.69), indicating higher average match and decreased residual error.
- Computational Efficiency:Linear Regression educated extensively faster (0.02 seconds) in comparison to Random Forest (1.35 seconds). This highlights the trade-off among simplicity and computational cost, specifically for massive datasets.
- Residual Analysis:Residual plots for Linear Regression confirmed systematic errors, in particular for non-linear patterns, suggesting underfitting.Random Forest residuals had been greater uniformly distributed, confirming its cappotential to seize statistic variability effectively

Critical Evaluation:

- Strengths of Random Forest:
✓ Excels at shooting non-linear styles and function interactions, making it perfect fo datasets with numerous monetary indicators.
✓ Robust to outliers, supplying solid and dependable predictions.
- Limitations of Random Forest:
✓ Requires greater computational sources and schooling time as compared to Linea Regression.
✓ Lacks interpretability, making it hard to derive function-particular insights.
- Strengths of Linear Regression:
✓ Simplicity and velocity make it a precious baseline for short evaluation and interpretation
✓ Coefficients offer direct insights into function importance.
- Limitations of Linear Regression:
✓ Struggles with non-linear relationships, main to underfitting in complicated datasets.
✓ More touchy to outliers, that may skew predictions.

## Lessons Learned and Future Work:

Lessons Learned:

- Feature Selection Matters: Choosing the proper functions, inclusive of SP_Close and USDI_Price, changed into important for attaining significant results. These functions had a robust courting with gold prices.
- Advantages of Normalisation: By guaranteeing that all features were comparable and avoiding bias from features with greater sizes, normalising data enhanced model performance.
- Model-to-Model Tradeoffs: While Linear Regression was quicker but had trouble with non-linear patterns, Random Forest was more accurate but needed more processing power.
- The Key Is Residual Analysis: By looking at the residuals, it was possible to find underfitting in Linear Regression and validate that Random Forest represented variability more accurately.

Future Work:

- Incorporate Advanced Models: Test Gradient Boosting or Neural Networks to enhance prediction accuracy further.
- Execute Feature Engineering: To improve prediction capability, create extra features like volatility measurements or moving averages.
- Optimise Random Forest: To cut down on computational expenses, try shrinking the feature space and fine-tuning the hyperparameters.

## REFERENCES

1. [1] Xiaohui Yang, "The Prediction of Gold Price Using ARIMA Model", 2nd International Conference on Social Science, Public Health and Education 2019.
2. [2] Manjula K. A., Karthikeyan P, "Gold Price Prediction using Ensemble based Machine Learning Techniques", Third International Conference on Trends in Electronics and Informatics, 2019.
3. [3] Mrs. B. Kishori 1, V. Preethi, "Gold Price forecasting using ARIMA Model", International Journal of Research, 2018.
4. [4] R. Hafezi* , A. N. Akhavan, "Forecasting Gold Price Changes: Application of an Equipped Artificial Neural Network", AUT Journal of Modeling and Simulation, 2018.
5. [5] Xiaohui Yang, "The Prediction of Gold Price Using ARIMA Model", 2nd International Conference on Social Science, Public Health and Education (SSPHE 2018).
6. Introduction to Statistical Learning: by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani
Chapter 3: Linear Regression
Chapter 8: Tree-Based Methods (Random Forest and Boosting)
7. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" by Aurélien Géron
Chapter 4: Training Models (Linear Regression and Gradient Descent)
Chapter 7: Ensemble Learning and Random Forests.
8. Data set: https://www.kaggle.com/datasets/sid321axn/gold-price-prediction-dataset?resource=download