**GOLD PRICE PREDICTOR**

**Submitted By:**

* **Kasfi Ahamed**
  + ID- 19101094

**Introduction:**

Gold is one of the precious metals. It is used as a medium for money or exchange because of its limited supply and high value. This metal’s scarcity and difficulty in extraction made it a valuable commodity. It plays a distinctive role not only as a financial asset in international currency reserves but also contributes significantly to the stabilization of the international money market. Due to its increasing value, many people considered gold as an attractive investment.

Investments and savings are an important element of everyone's life. Investments are meant to earn favorable returns in the future by employing our present funds in some trustworthy assets. In economic terms, it is considered to be the purchasing of assets that are kept low in the present and to be used in the future for creating our wealth. These investments can be sold for a higher price which in turn increases the profit. This particular study is concentrating on the gold price with respect to other factors in the economy. To analyze the data, three machine learning algorithms are employed- Decision Tree Regressor, XGB Regressor, and Random Forest Regressor. We can find accurate data by comparing these three algorithms under various circumstances.

**Motivation:**

Gold is considered an attractive investment due to its area of usage and increase in value. The price of gold depends on various factors such as currency value, political issues, transportation costs, etc. This increase in the value of gold and down prices in other markets led more investors to be attracted to gold. But at some point when the gold price went down, investing in gold became riskier. Hence, the prediction of the gold price is a vital issue in financial economics.

**Methodology:**

The study is carried out using Machine Learning. In order to acquire the best possible result, both data training and testing were undertaken. The Decision Tree Regressor, XGB Regressor, and Random Forest Regressor were among the machine learning methods employed to analyze the data in this study. Regression analysis is a statistical tool for determining the relationship between two or more variables. When one of the independent variables changes while the other variables remain constant, regression analysis is performed to see how the value of the dependent variable changes. Decision Tree Regression follows a predictive model. It uses a set of binary rules and then calculates the target value. Every individual tree has its own branches, nodes, and leaves. A decision tree is a type of tree that can be used to forecast and classify data. A Random Forest uses several decision trees with a technique called Bootstrap Aggregation to solve both regression and classification problems. The random forest algorithm can be used for feature engineering, which is the process of determining the most essential attributes from a set of options. Gradient boosting is a technique for selecting the best forecast from a large group. It also generalizes them by allowing the optimization of an object. Python is used to implement these machine learning methods (Decision Tree Regression, Random Forest Regression, and XGB Regressor) in this study.

**Dataset Description:**

**Source**: <https://www.kaggle.com/datasets/altruistdelhite04/gold-price-data> [Dataset]

A dataset, as the name suggests, is a collection of data. In Machine Learning projects, we always need a dataset. Firstly, we need the training dataset to train our model, to help it predict. Then, we use testing datasets to predict and check how accurate our model is.

This dataset is about gold price prediction. There are 6 columns and 2290 rows in this dataset. First, we need to import library functions to read the dataset and perform EDA.

There are 5 columns which are floating or numerical and the other one is an object.

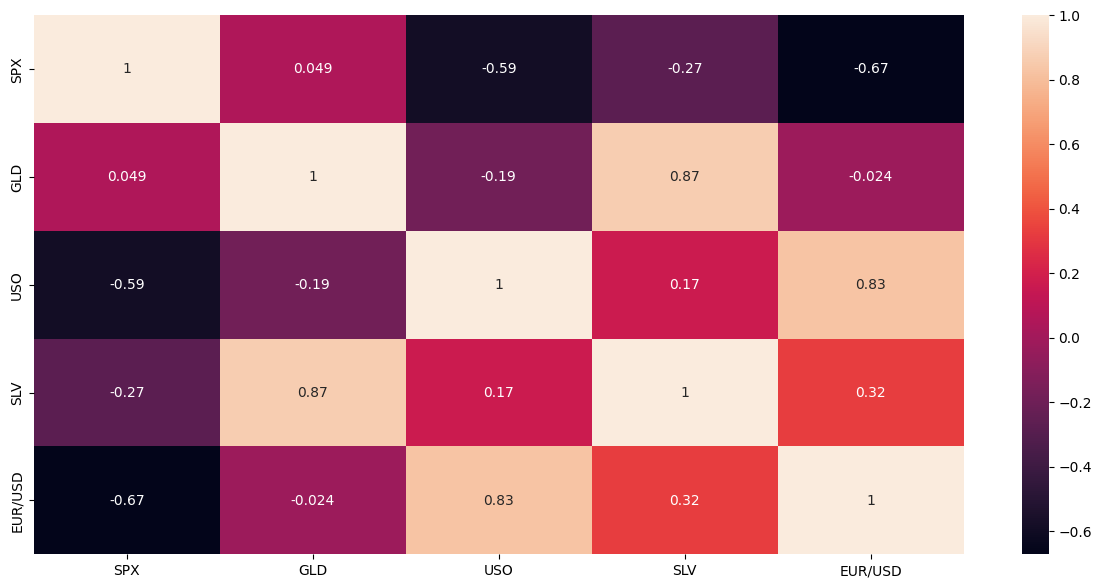
**Features:**

* Date - mm/dd/yyyy
* SPX - is a free-float weighted measurement stock market index of the 500 largest companies listed on stock exchanges in the United States.
* USO - United States Oil Fund - Not Sure of UOM
* SLV - Silver Price
* EUR/USD - currency pair quotation of the Euro against the US

This is the label that we are going to use

**Label:**

* GLD - Gold Price



**Biasness:**

All the unique classes do not have an equal number of instances.

**Dataset Preprocessing:**

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. Real-world data generally contains noises, and missing values, and may be in an unusable format that cannot be directly used for machine learning models. Data preprocessing is required for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

In machine learning data preprocessing, we divide our dataset into a training set and a test set. This is one of the crucial steps of data preprocessing as by doing this, we can enhance the performance of our machine learning model.

In order to perform data preprocessing using Python, we need to import some predefined Python libraries such as numpy, matlab, and pandas. Then we need to import the datasets which we have collected for our machine learning project.

Now for preprocessing, we followed these steps and got this takeaway.

1. **Finding Unwanted Columns**

Take-away: We won't consider the Date feature so we will drop this feature in the feature engineering section.

1. **Finding Missing Values**

Take-away: No missing value found.

1. **Find Features with One Value**

Take-away: No feature with only one value.

1. **Explore the Categorical Features**

Take-away: There is one categorical feature.

1. **Find Categorical Feature Distribution**

Take-away: N/A

1. **Relationship between Categorical Features and Label**

Take-away: N/A

1. **Explore the Numerical Features**

Take-away: There are 4 numerical features.

1. **Find Discrete Numerical Features**

Take-away: There are no Discrete Variables in the given dataset.

1. **Relation between Discrete numerical Features and Labels**

Take-away: N/A

1. **Find Continuous Numerical Features**

Take-away: There are 4 continuous numerical features

1. **Distribution of Continuous Numerical Features**

Take-away: It seems SPX, SLV, and EUR/USD are distributed normally, USO is heavily skewed towards the right and seems to have some outliers.

1. **Relation between Continuous numerical Features and Labels**

Take-away: It seems the SLV feature linearly progresses with GLD

1. **Find Outliers in numerical features**

Take-away: It seems USO and SLV have some outliers.

1. **Explore the Correlation between numerical features**

Take-away: It seems the SLV feature is heavily correlated with GLD

**Dataset Splitting:**

1. Train Set: 80%
2. Test Set: 20%

**Model Training:**

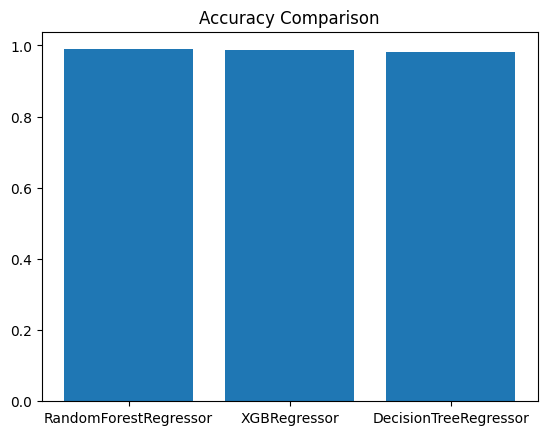
We have applied 3 models which are given below-

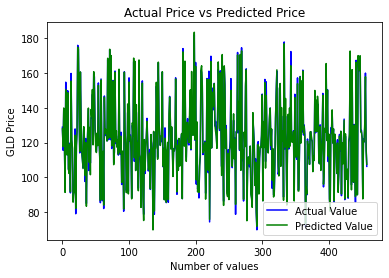
* **DecisionTreeRegressor**: A decision Tree is a decision-making tool that uses a flowchart-like tree structure or is a model of decisions and all of their possible results, including outcomes, input costs, and utility. Decision-tree algorithms fall under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables.
* **RandomForestRegressor**: Random Forest Regression is a supervised learning algorithm that uses ensemble learning methods for regression. It operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees.
* **XGBRegressor:** GBoost stands for "Extreme Gradient Boosting" and it is an implementation of gradient boosting trees algorithm. The XGBoost is a popular supervised machine learning model with characteristics like computation speed, parallelization, and performance.

**Result:**

Bar chart (Comparing f1-score):

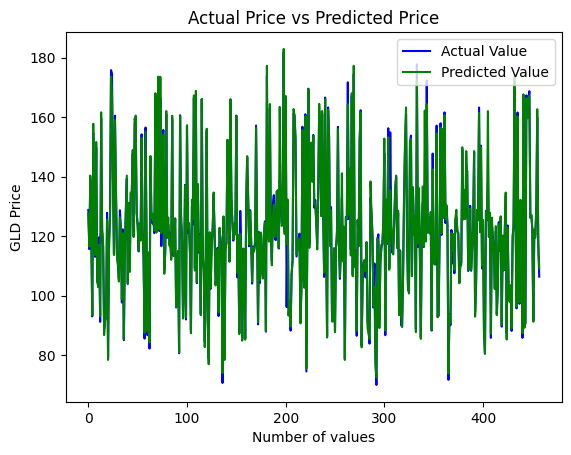
|  |  |  |
| --- | --- | --- |
|  | **Models** | **Accuracy Scores** |
| 0 | decision\_tree | 0.981362 |
| 1 | RandomForestRegressor | 0.989081 |
| 2 | XGBRegressor | 0.987071 |

****

**Actual price vs Predicted Price using XGB regressor**

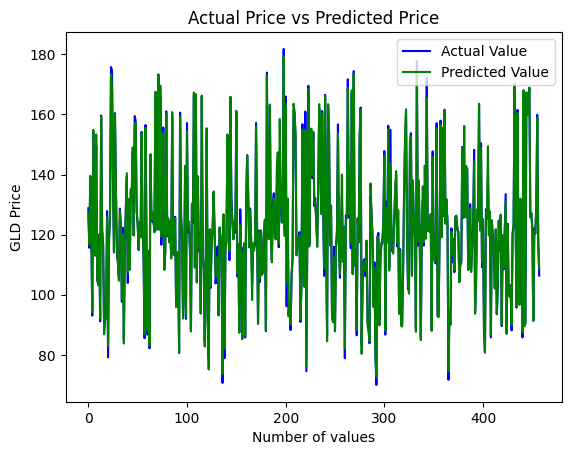
|  |  |  |
| --- | --- | --- |
|  | **Name** | **Result** |
| 0 | The mean absolute error | 1.6747 |
| 1 | The mean squared error | 0.989081 |
| 2 | The root mean squared error | 2.57895 |

**Actual Price vs Predicted Price using Decision Tree Regressor**

****

|  |  |  |
| --- | --- | --- |
|  | **Name** | **Result** |
| 0 | The mean absolute error | 1.46927 |
| 1 | The mean squared error | 9.58803 |
| 2 | The root mean squared error | 3.0964 |

**Actual Price vs Predicted Price using XGB Regressor**

****

|  |  |  |
| --- | --- | --- |
|  | **Name** | **Result** |
| 0 | The mean absolute error | 1.31940 |
| 1 | The mean squared error | 5.616758 |
| 2 | The root mean squared error | 2.369970 |

**Conclusion:**

Gold is considered an attractive investment due to its area of usage and increase in value. The price of gold depends on various factors such as currency value, political issues, transportation costs, etc. A particular study is concentrating on the gold price with respect to other factors in the economy. To analyze data we will use machine learning algorithms. By comparing these machine learning algorithms we can find accurate data under various circumstances.

In conclusion, it has been shown that machine learning algorithms are very effective in predicting gold prices. Linear regression, neural networks, and random forest these models have been applied to historical data to forecast future prices with reasonable accuracy. It is important to note that the accuracy of the prediction depends on the quality and quantity of the data which is also used by the model’s ability to capture complex patterns and relationships in the data. Geopolitical events, economic policies, and market sentiment these external factors can also impact the price of gold and can not always be accounted for in the models. While machine learning can provide valuable insights into gold price trends, it should be used as a tool in conjunction with other forms of analysis so that we can make informed investment decisions.

**Future Work:**There are several things we can work on in the future on this topic. For example:  
  
1)Time Series Modeling: Time-series data is generally used to model gold prices. While traditional regression models can be used to predict gold prices, more advanced time-series models such as ARIMA, SARIMA, or LSTM could be explored to improve the accuracy of the predictions.  
  
2)Deep learning: Deep learning models, for example, CNN or GAN could be used to learn features directly from raw data such as images of gold prices over time.  
  
3)Ensemble methods: Ensemble methods such as random forests, gradient boosting, or stacking could be used to combine multiple models and improve the overall prediction accuracy.

**References:**

1)Data set- <https://www.kaggle.com/datasets/altruistdelhite04/gold-price-data>

2)Aye G, Gupta R et al (2015) Forecasting the price of gold using dynamic model averaging. Int Rev Financ Anal 41:257–266

3)Liu, D., Li, Z. (2017). Gold Price Forecasting and Related Influence Factors Analysis Based on Random Forest. In: Xu, J., Hajiyev, A., Nickel, S., Gen, M. (eds) Proceedings of the Tenth International Conference on Management Science and Engineering Management. Advances in Intelligent Systems and Computing, vol 502. Springer, Singapore. https://doi.org/10.1007/978-981-10-1837-4\_59