GOLD PRICE PREDICITION MODEL

MACHINE LEARNING Data Analysis

SUBMITTED By:

CLASS -- CSE_3(2nd YEAR)

- NANDINI SAWDAGAR University Roll No—500221020039
- SHIPU KUMAR University Roll No—500321010014
- NANDINI CHATTOPADHAY University Roll No—500221020038
- RISHI KUMAR University Roll No—500221010051
- Md SAHIL University Roll No—500221010025

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to all those who have contributed to the successful completion of this project. We are grateful to **GNIT** for providing us with the opportunity to develop this project. Thank you for your encouragement and guidance. This project would not have been possible without your support.

We extend our heartfelt appreciation to Mr. Moloy Dhar and Ms. Pallabi Das, Associate Professors at GNIT, Kolkata. Their unwavering support, valuable advice, and constant inspiration have truly motivated and guided us throughout this project. Without their encouragement and guidance, this endeavour would not have materialized. We are sincerely grateful for the pearls of wisdom they generously shared, which enriched our understanding and shaped the outcome of our work. Thank you for their invaluable contributions to our success.

We are deeply grateful for their valuable input, which played a crucial role in achieving our goals. Their contributions have made a significant impact, and we extend our heartfelt appreciation to each and every member involved in this project. Together, we have accomplished great things through teamwork and collaboration.

Abstract

Since ancient times, gold has been revered for its intrinsic value and has played a significant role in various aspects of human civilization. Initially, it served as a medium of exchange and a means of remuneration. However, its purpose has evolved, and it is now widely regarded as a valuable investment asset, reflecting the economic wealth of nations. Remarkably, gold's importance extends beyond the affluent, encompassing individuals from all walks of life.

Especially during unprecedented crises like the Covid-19 pandemic, investments in gold have proven to be prudent. The historical stability and universal acceptance of gold make it a reliable hedge against economic uncertainty. Consequently, predicting gold rates using live data and strategically investing in gold can yield substantial benefits.

To accomplish accurate gold rate predictions, various machine learning algorithms have been employed, including linear regression, decision trees, and random forests. By leveraging these advanced analytical techniques, researchers have reached a compelling conclusion: the random forest method consistently produces more precise results. Its ability to account for multiple factors and complex interactions enhances its predictive capabilities in the dynamic gold market.

The keywords associated with this domain include gold rates, prediction, live gold rates, random forest, decision tree, linear regression, and predictive analytics. By harnessing the power of these tools and staying updated with real-time gold data, investors can make informed decisions and seize profitable opportunities in the ever-changing landscape of gold investments.

1.Introduction

- Gold, the earliest renowned metal, played a crucial role in technological progress. It preceded iron and copper, which are commonly associated with economic and cultural advancement. Gold has maintained its value and acted as a measure of a nation's financial strength. Its historical significance is undeniable, making it a vital asset in assessing economies.

- Gold's allure attracts substantial investments, reflecting its foundational importance. Its price is influenced by supply, demand, and speculation, but unique factors such as savings and disposal play a significant role in shaping it.

- Gold's appeal extends to small-scale investors, who view it as a secure investment choice. Governments strategically invest in gold, considering their financial conditions and interest rates as indicators of economic strength. The United States exhibits notable activity, attracting capital to the gold market. Multiple factors contribute to gold rate fluctuations and influence its price.

- Gold spot prices are determined twice a day, considering the supply and demand dynamics of the gold market. Even a slight change in the price of gold can result in substantial gains or losses for both individual investors and government banks. The reasons behind the rise in gold rates are multifaceted, including its remarkable utility and rarity. Additionally, gold finds applications in finance, trade, mechanical, industrial, dental, and medical fields, accounting for approximately 12% of its overall demand.

- Gold exhibits remarkable properties, including high thermal and electrical conductivity, corrosion resistance, and resistance to bacterial colonization. Recent price fluctuations are influenced by the growing middle classes in emerging markets aspiring to Western lifestyles. Extracting small amounts of gold requires processing significant quantities of gold ore, demanding extensive labour. Lower-grade ores may yield as little as 5 grams of gold per ton

- In summary, gold's historical significance, enduring value, and multifaceted utility captivate investors and shape global economies, making it an exceptional and influential metal.

2. Methodology

2.1 Gold Investments

- Since ancient times, the value of gold has remained high despite economic crises. In recent years, its value has fluctuated due to currency markets, crude oil prices, and inflation. Investing in gold carries higher risks, but its global recognition as a core and liquid asset makes trading accessible. Forecasting gold rates aids in planning and safeguarding investments. However, predicting gold rates, like stocks, presents challenges due to their non-linearity. Shareholders seek protection from political and monetary instability and societal upheavals.

- Gold's stability in the market makes it a favourable investment for individuals and multinational corporations, leading to increased gold reserves. It has evolved into a form of currency, prompting governments to bolster their gold holdings. Daily fluctuations in gold prices are governed by government policies and regulators. Estimating monthly rate fluctuations assists financial professionals in making informed buying and selling decisions. Gold is also used globally as an alternative payment method in business transactions. Central banks hold precious metals to ensure debt payment, regulate inflation, and demonstrate financial strength.

- Alloyed gold demonstrates superior retention properties in the placement of ultrasonic aluminium wire connections compared to standard alloy metallization. It is crucial for prospective gold buyers to comprehend the factors influencing gold rates to predict interest rate patterns effectively and enhance profitability. Investors and researchers alike exhibit

significant interest in understanding the gold market for in-depth analysis and precise forecasting.

2.2 Dataset

- Data for this study is collected from Kaggle which is a live dataset that collects the data for **nearly 10 years** from the 2008-2018. Data for attributes, such as date, open and closing stock , high and low price of stock and adj.close and volume of stocks were gathered. Data of many other stocks in addition with gold are scrapped from Kaggle. Price of all the stocks during this period is also included in the analysis. Table below lists the attributes of data which are extracted for the past 10 years.

```
[ ] gold_data.head()
```

	Date	SPX	GLD	USO	SLV	EUR/USD
0	1/2/2008	1447.160034	84.860001	78.470001	15.180	1.471692
1	1/3/2008	1447.160034	85.570000	78.370003	15.285	1.474491
2	1/4/2008	1411.630005	85.129997	77.309998	15.167	1.475492
3	1/7/2008	1416.180054	84.769997	75.500000	15.053	1.468299
4	1/8/2008	1390.189941	86.779999	76.059998	15.590	1.557099

```
[ ] #Print the last 5 rows of the dataFrame
    gold_data.tail()
```

	Date	SPX	GLD	USO	SLV	EUR/USD
2285	5/8/2018	2671.919922	124.589996	14.0600	15.5100	1.186789
2286	5/9/2018	2697.790039	124.330002	14.3700	15.5300	1.184722
2287	5/10/2018	2723.070068	125.180000	14.4100	15.7400	1.191753
2288	5/14/2018	2730.129883	124.489998	14.3800	15.5600	1.193118
2289	5/16/2018	2725.780029	122.543800	14.4058	15.4542	1.182033

```
[ ] # number of rows and coloums
    gold_data.shape
```

```
(2290, 6)
```

- The above listed dataset consists of 2290 rows and 6 columns. Where the different columns are used are comparisons between the :-
    - S&P index of US Stock Market(SPX),
    - The rate of Gold in US dollars(GLD),
    - the Dollar rate at the particular time frame(USD)
    - The exchange-traded fund(ETF) for the Silver in the time frame(SLV)
    - Lastly the Euro(currency)-Usd exchange rates
- Here we do a lot of cleaning and pre-processing techniques to remove the unwanted attributes. Missing values are also treated to give the good prediction accuracy. Gold prices change day to day and the daily changes have been recorded in the analysis. The current gold rates higher in comparison with the previous years. As we see huge variations in price, we decided to split the dataset in a sequential split instead of random sampling. Therefore, the recent 30% of the data has been used as the test set, and the earliest 70% of data has been used for training. There is a major fluctuation in gold prices over the past years, so recent historical data would be more indicative of the future trend. This graph gives the trend of gold prices in the past 10 years.

2.3 Machine Learning Implementation

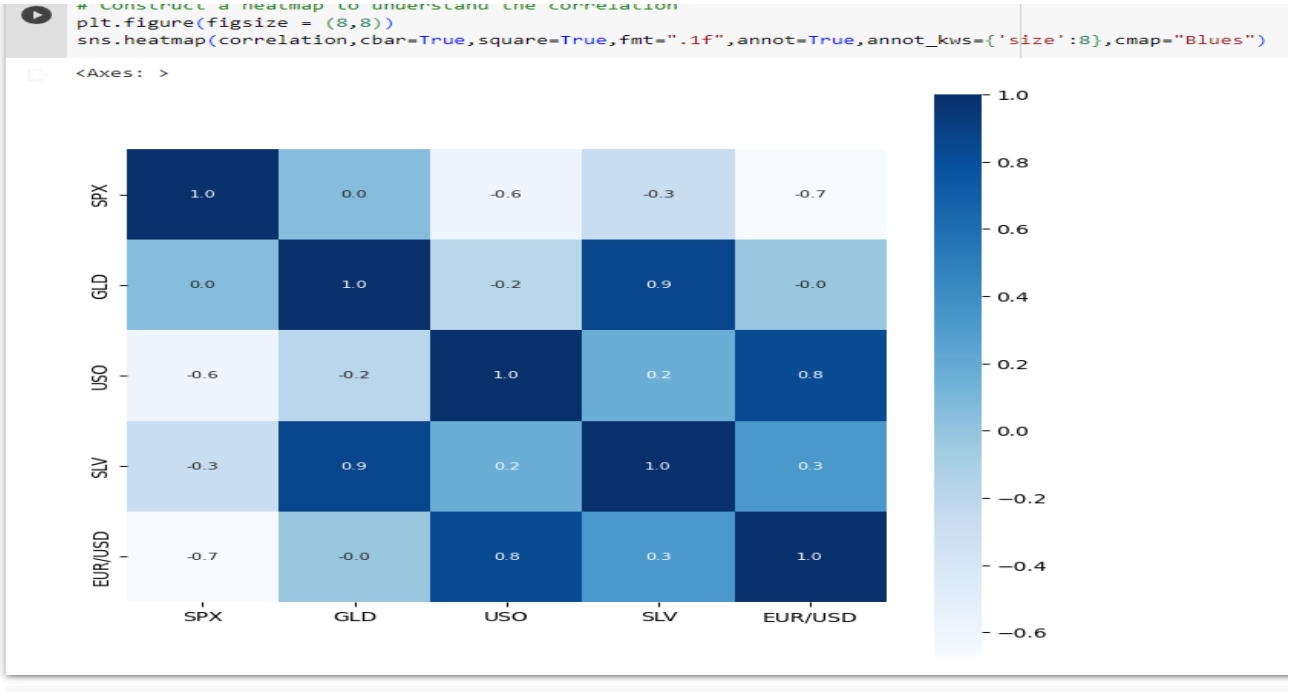
We get the input in the form of tickle data and fit the models and return back the prediction result in the form of tickle, so there will be no change in the prediction and the accuracy.

➢ Correlation

- Positive correlation refers to a relationship between two variables where they move in the same direction. When one variable increases, the other variable also tends to increase, and when one variable decreases, the other variable tends to decrease as well.

- Negative correlation, on the other hand, describes a relationship between two variables where they move in opposite directions. When one variable increases, the other variable tends to decrease, and vice versa.

In the context of investments, correlation is often used to assess the relationship between different assets or securities. A positive correlation between two investments means they tend to move in sync, while a negative correlation suggests they tend to move in opposite directions. Correlations can be useful in diversifying portfolios and managing risk.



##Heatmap to understand the Correlation

```
[ ]  # correlation values of GOLD(GLD)
     print(correlation['GLD'])
```

```
SPX         0.049345
GLD         1.000000
USO        -0.186360
SLV         0.866632
EUR/USD    -0.024375
Name: GLD, dtype: float64
```

```
[ ]  # correlation of the distribution of the GLD price
     sns.distplot(gold_data['GLD'],color='green')
```

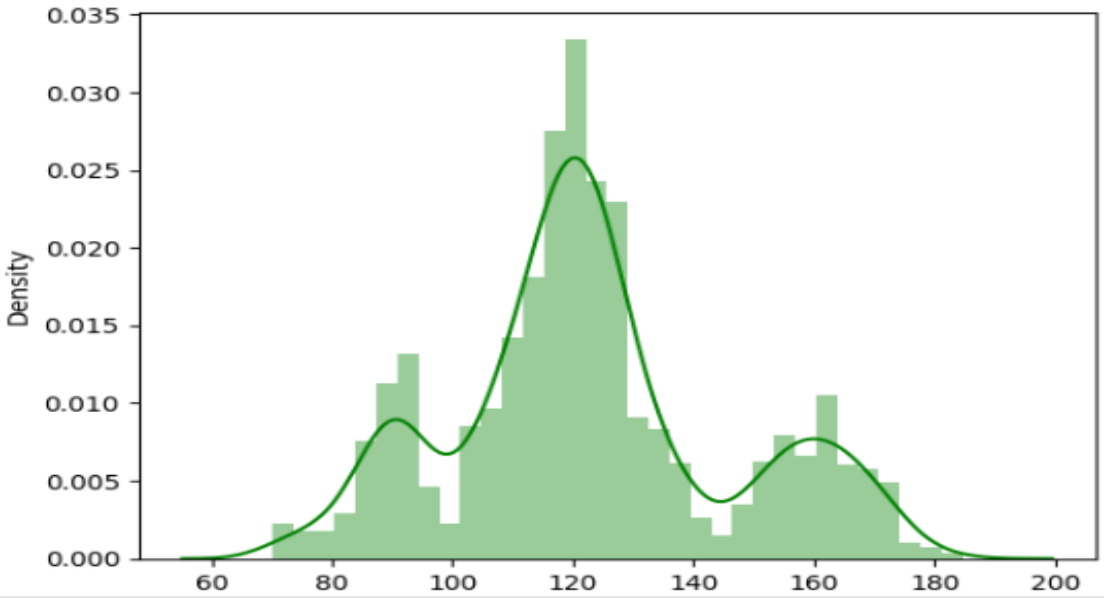
```
<ipython-input-15-fda1802da749>:2: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(gold_data['GLD'],color='green')
<Axes: xlabel='GLD', ylabel='Density'>
```



##correlation of the distribution of GOLD(GLD) price

➢ Splitting of training and testing data

With the history of the dataset, the training set has been divided into 70% and the testing set has been divided into 30% in order to give more accuracy. The following are the machine learning models that are used to predict almost the exact value in the future.

Splitting the Features and Target

```
[ ]  X = gold_data.drop(['Date','GLD'],axis=1)
     Y = gold_data['GLD']
```

```
     print(X)
```

```
                SPX        USO      SLV   EUR/USD
0      1447.160034  78.470001  15.1800  1.471692
1      1447.160034  78.370003  15.2850  1.474491
2      1411.630005  77.309998  15.1670  1.475492
3      1416.180054  75.500000  15.0530  1.468299
4      1390.189941  76.059998  15.5900  1.557099
...            ...        ...      ...       ...
2285   2671.919922  14.060000  15.5100  1.186789
2286   2697.790039  14.370000  15.5300  1.184722
2287   2723.070068  14.410000  15.7400  1.191753
2288   2730.129883  14.380000  15.5600  1.193118
2289   2725.780029  14.405800  15.4542  1.182033

[2290 rows x 4 columns]
```

```
[ ]  print(Y)
```

```
0        84.860001
1        85.570000
2        85.129997
3        84.769997
4        86.779999
           ...
2285    124.589996
2286    124.330002
2287    125.180000
2288    124.489998
2289    122.543800
Name: GLD, Length: 2290, dtype: float64
```

Splitting into training data and Test Data

```
[ ]  X_train,X_test,Y_train,Y_test = train_test_split(X,Y,test_size = 0.2,random_state=2)
```

➢ Random forest model for Model Training

- Random forest is a bagging model that builds multiple decision trees and merges them together to get a more accuracy in prediction.
- Random forest adds additional randomness to the model that we fit the training data, while growing the trees. In Accordance with our dataset, the random forest model gives the best accuracy.
- For the application in the stock market, the Random Forest algorithm can be used to identify a stock's behaviour and the expected loss or profit can be forecasted from the prediction model.

Model Training: Random Forest Regressor

```
regressor = RandomForestRegressor(n_estimators=100)


# training of the model
regressor.fit(X_train,Y_train)
```

```
RandomForestRegressor()
```

Model Evaluation

```
# prediction on Test Data
test_data_prediction = regressor.predict(X_test)

print(test_data_prediction)
```

```
[168.57109926  81.71429998 115.59790015 127.6656007  120.77320135
 154.76329757 150.38129978 126.2074006  117.48789883 126.0165008
 116.82960079 172.14870088 141.45009861 167.95419932 115.11190015
 117.60340034 138.74200272 170.32910131 158.89650387 160.57519964
 155.23780053 125.39170003 175.87860011 157.4320029  125.19980047
  93.75679977  77.41710005 120.47819989 119.21180005 167.50979982
  88.22460059 125.22629992  91.10210079 117.52600044 121.03919907
 136.08540105 115.39520096 115.0276008  147.47610029 107.52190112
 104.64150242  87.07099766 126.70050053 117.79730008 153.71759905
 119.5186002  108.35949983 108.08999846  93.01850018 127.15189771
  75.48770031 113.67289925 121.36349999 111.3020992  118.84679896
 120.53959915 158.89649976 168.08300091 146.93749659  85.86599849
  94.53470037  86.72989865  90.54669991 118.89240077 126.47740057
 127.50380001 170.59299959 122.25539932 117.52169891  98.4638001
 168.24740144 143.03709832 131.46760206 121.14540199 121.18259942
 119.9550085  114.54780166 118.27410036 107.49330083 127.9700007
 113.90999974 106.78330008 116.82030061 119.51609897  89.13880049
  88.29739881 145.83240219 127.04399965 113.54280019 110.28939858
 108.2962989   77.00619911 169.9570015  113.93699913 121.62299906
 128.20660205 154.83469801  91.85499912 135.32620134 158.66550334
 125.19120048 125.43040044 130.59570136 114.78520142 119.74290017
  92.0261998  110.39969895 168.71899903 157.39279873 114.05129934
 106.76610104  79.29189952 113.25170002 125.93710073 107.12389936
 119.34480099 156.20600309 159.73939862 120.29649974 134.69300275
 101.33469997 117.51949801 119.25690008 112.9848007  102.78149958
 160.07409725  99.10610027 146.80789891 125.56740102 169.95459906
 125.87509879 127.38059735 127.11160104 113.91009916 112.94470073
 123.60869913 102.18869868  89.26299985 124.62589932 101.32579938
 107.09359937 113.5463007  117.19980076  99.40319954 121.70010019
 163.34039923  87.27269869 106.66909968 117.13800077 127.74110137
 124.28880086  80.65639906 120.04040082 157.90779815  87.97409973
 110.42709952 118.90389933 172.7159986  102.98639893 106.14540028
 122.51330051 157.93269786  87.48129833  93.45070033 113.85590026
 176.74340061 114.53489978 119.1870001   94.65120068 125.64439975
 165.53360033 114.76130084 116.69050134  88.36809868 148.67760052
 120.27699917  89.69120019 111.66709978 117.37370013 118.70170123
  88.1914992   94.33830001 117.12770001 118.45930179 120.1765004
```

https://colab.research.google.com/drive/1PIIn7h55vpOiro7DvOoiSNRdMhWK6VyG?usp=sharin

```
      104.98520126 149.46890874 113.84440189 124.82540133 147.83438024
      119.55390173 115.28400081 112.53659986 113.60740187 142.31200049
      117.82549787 102.92820862 115.90390188 103.6296019   98.77520014
      117.4364004   90.69560017  91.84310033 153.95899925 102.70099978
      154.99040866 114.51930163 137.96890139  90.13049868 115.41849992
      114.29949946 122.90510024 121.84530003 165.46740066  92.89919935
      135.89300111 121.38459914 120.79690073 104.77230022 143.99080273
      121.71849941 116.58260027 113.7872005  127.24349723 122.83389923
      125.77019943 121.2654004   86.89369898 132.54560072 142.43780248
       92.83919955 159.63789946 158.19310285 126.33209942 165.15639949
      108.7234997  110.01000081 103.73109853  94.55450074 127.9584028
      107.1822084  162.09469981 121.81940011 131.81210081 130.7085015
      160.47699961  90.11599859 175.20350164 127.8165001  127.80929793
       86.35999893 124.51579956 150.28049741  89.70830023 106.7281997
      108.78619975  83.84869912 136.01939998 154.86750186 139.69920328
       74.25119999 152.14110001 126.00330014 126.70240027 127.36939915
      108.48529982 156.41700004 114.49630134 116.86270113 125.57369912
      153.95190168 121.23469991 156.40719925  92.88790047 125.53360154
      125.93000038  87.83370034  92.24069933 126.34439925 128.38730371
```

```
# R squared error
error_score = metrics.r2_score(Y_test, test_data_prediction)
print("R squared error : ",error_score*100)
```

```
      R squared error :  98.86198175511663
```

❖ Result and Discussion

With the help of the random forest classifier, we could get almost the exact predicted value. For example, here the input date is given as 2018/11/26. The predicted output by the random forest model is given as $1780.94. This prediction estimation will give the stock market users and even normal customers to easily use this interface.

- **R-squared Error**

  The R-squared error, also known as the coefficient of determination, is a statistical measure that represents the proportion of the variance in the dependent variable that is predictable from the independent variable(s) in a regression analysis. It is a measure of how well the regression model fits the observed data.

  R-squared error ranges from 0 to 1, with 0 indicating that the independent variable(s) cannot explain any of the variability in the dependent variable, and 1 indicating that the independent variable(s) perfectly explain the variability in the dependent variable.

  In simpler terms, the R-squared error tells us the percentage of the dependent variable's variance that can be explained by the independent variable(s) in a regression model. A higher R-squared value indicates a better fit of the model to the data, suggesting that the independent variable(s) are more effective in explaining the variability in the dependent variable.

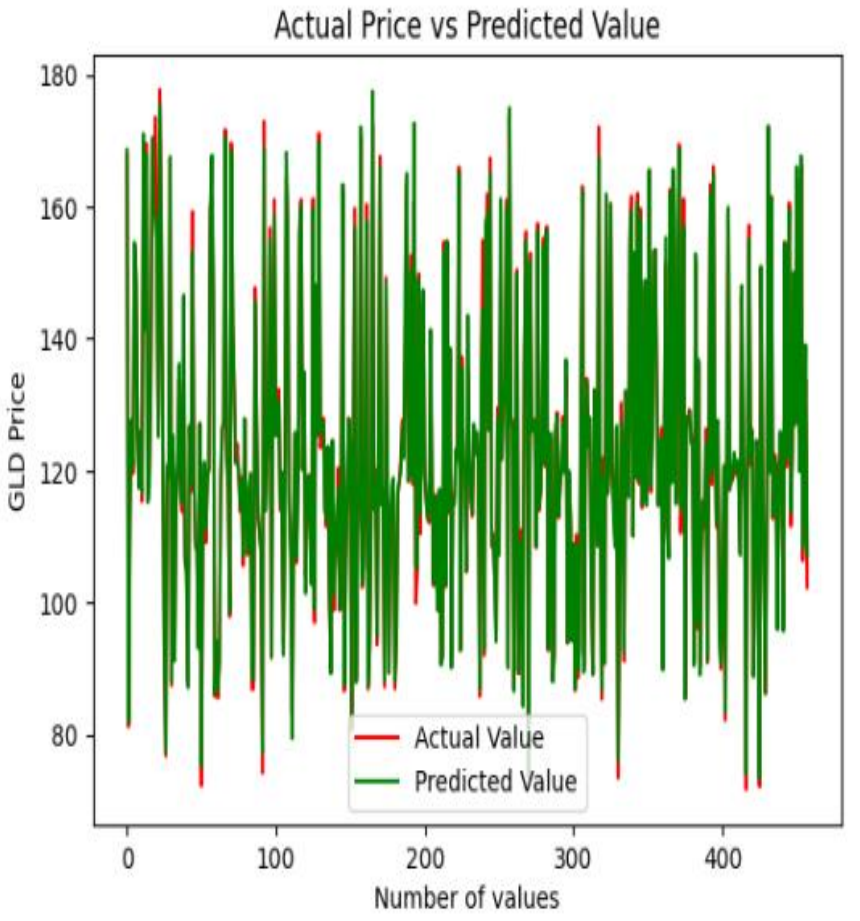
  R-squared Error: 98.86198175511663

➢ Compare actual values and predicted values in a plot

Compare the actual values and predicted values in a plot

```
[28] Y_test = list(Y_test)
```

```
plt.plot(Y_test,color='Red',label = 'Actual Value')
plt.plot(test_data_prediction,color='green',label="Predicted Value")
plt.title("Actual Price vs Predicted Value")
plt.xlabel("Number of values")
plt.ylabel('GLD Price')
plt.legend()
plt.show()
```



4. Conclusion

As gold proves to be a viable source of investment, the investment opportunities has been expanded to huge numbers, and it arose a need for predicting the future highs and lows that the commodity(gold) might hit. By testing with different machine learning prediction models such as linear regression, decision tree regression, random forest regression, it has been established that random forest comes out with better accuracy in predicting future gold rates. By building a web application that could possibly merit people who are interested with investing in gold, it could benefit more number of people worldwide as it has been hosted in heroku as well. Investors must also include technical and fundamental analysis of the commodity to arrive at better decisions.

5. Future Enhancements

Our first step would be updating our web application by including other ornamental share prices and enhancing our app with better UI, so as to make it more user friendly for the users. We've also planned to include various share prices of Nifty fifty companies' data and their historical behaviour. Incorporating blogging features for the traders and the users to share their experiences and ideas would be the next step. And to reach out to a greater number of audience, we've planned to launch it as an app in play store.

7. Reference

[1] D. G. Navin, "Big Data Analytics for Gold Price Forecasting Based on Decision Tree Algorithm and Support Vector," International Journal of Science and Research, pp. 2026-2030, 2015.

[2] M. S. S. S. S. Kumar Chandar, "Forecasting Gold Prices Based on Extreme Learning Machine," International Journal of Computers Communications & Control, pp. 372-380, 2016.

[3] Shafiee,Shahriar and Erkan Topal. An overview of global gold market and gold price forecasting. Resources Policy .Volume:35, 2010

[4] For DATASET-- https://www.kaggle.com/datasets/altruistdelhite04/gold-price-data