A Project Report

on

GOLD PRICE PREDICTION USING RANDOM FOREST REGRESSION

Submitted in partial fulfillment of the
requirement for the award of the degree of

BACHELOR OF TECHNOLOGY

Under The Supervision of
Name of Supervisor : Dr Shiv Kumar Verma
Designation : Professor

Submitted By

Name of Student : Prashant Kumar Pal
Enrollment/Admission No : 18SCSE1180012

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA
12, 2021

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the project, entitled **"GOLD PRICE PREDICTION"** in partial fulfillment of the requirements for the award of the Bachelor of Technology submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of 07, 2021 to 12, 2021 under the supervision of Dr Shiv Kumar Verma, Professor, Department of Computer Science and Engineering of School of Computing Science and Engineering , Galgotias University, Greater Noida

The matter presented in the project has not been submitted by me for the award of any other degree of this or any other places.

Prashant Kumar Pal, 18SCSE1180012

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Supervisor Name
Designation

CERTIFICATE

The Final Project Viva-Voce examination of Prashant Kumar Pal : 18SCSE1180012 has been held on_____and his work is recommended for the award of Bachelor of Technology

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Project Coordinator **Signature of Dean**

Date: November, 2021
Place: Greater Noida

Acknowledgement

Place:

Date:

In the accomplishment of completion of my project on **GOLD PRICE PREDICTION USING RANDOM FOREST REGRESSION** I would like to convey my special gratitude to my guide **Dr Shiv Kumar Verma** Sir, of School of Computer Science and Engineering and as well as **Dr Preeti Bajaj** Mam, Vice Chancellor of Galgotias University.

Your valuable guidance and suggestions helped me in various phases of the completion of this project. I will always be thankful to you in this regard.

I am ensuring that this project was finished by me and not copied.

Name: Prashant Kumar Pal

Signature

Abstract

This project is based on preparing machine learning model random forest regression to understand the relationship between gold price and selected factors influencing it, namely stock market, crude oil price, dollar/euro ratio, gold price and silver price.

All the operations to train the model are performed using Google Colaboratory. Monthly price data used for period was used for the study the dataset is collected in csv file format from kaggle.

The data was further split into two periods, training data and testing data using sklearn library and also use it to import random forest regressor to predict the price of gold using different factors that are influencing gold price.

Machine learning algorithms, random forest regression was used in analyzing these data. It is found that the correlation between the variables is strong during the period I and weak during period II. While these models show good fit with data during period I, the fitness is not good during the period II.

While random forest regression is found to have better prediction accuracy for the entire period. We will get the the accurate data of price after training and testing of model.

Contents

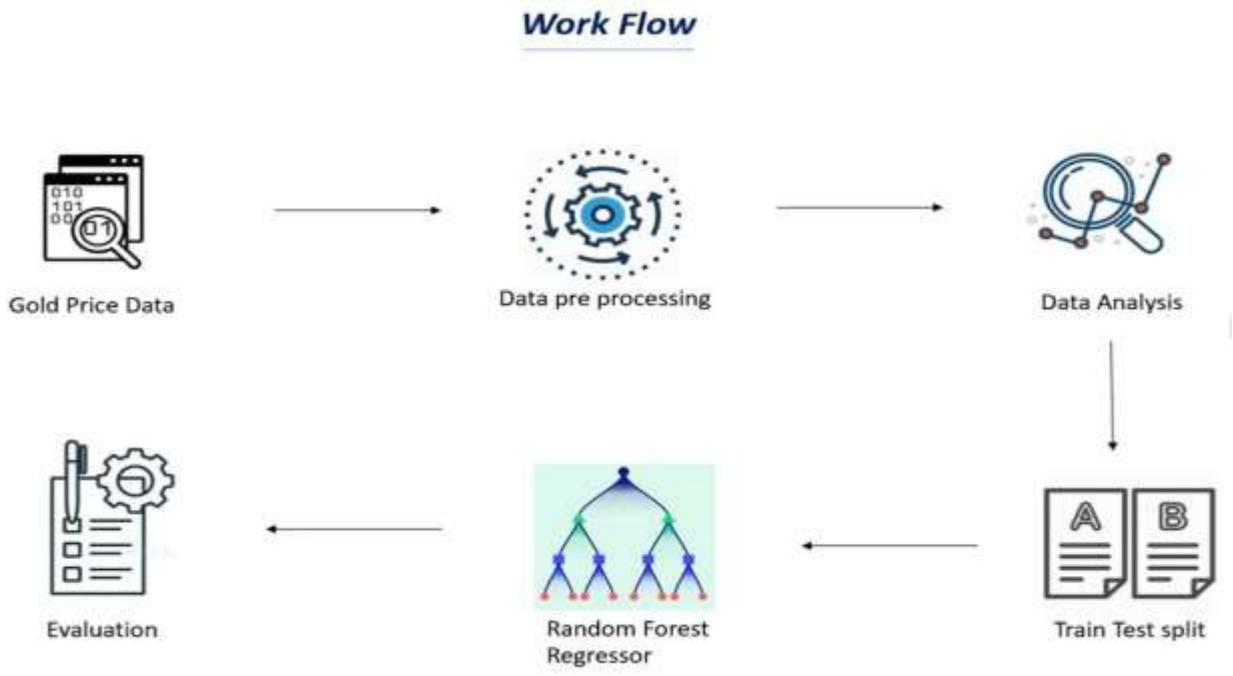
CHAPTER-1 Introduction

Savings and Investments form an integral part of everyone's life. Investments refer to the employment of present funds with an objective of earning a favourable return on it in future. In an economic sense, an investment can be considered as the purchase of assets that are not consumed today but are used in the future to create wealth. In finance, an investment is purchase of a monetary asset with the idea that the asset will provide income in the future or will later be sold at a higher price for a profit. Gold is a precious metal, so like any other goods, gold's price should depend on supply and demand. But, since gold is storable and the supply is accumulated over centuries, this year's production has little influence on its prices. Gold is used both as a commodity and as a financial asset. This raising value of gold coupled with the volatilities and fall in prices of other markets like capital markets and real estate markets has attracted more and more investors towards gold as an attractive investment. Understanding such relationship will be helpful not only to monetary policymakers but also to investors, fund managers and portfolio managers to take better investment decisions in the market. Further this study uses three machine learning algorithms, linear regression, random forest regression and gradient boosting regression in analyzing these data. Comparison of these three methods will help us in identifying the accuracy of these methods under various conditions.

Chapter 2 Literature Survey

There are many studies dealing with the price of gold in the literature. Although various different variables are used in these studies, it is observed that gold prices are regressed against USA dollar and stock return in general. The relationship between other macroeconomic variable and gold prices has also been studied by many researchers. The relationship between gold price and prices of other commodities especially crude oil has also been extensively studied. But the results from these studies are found to be contradicting. Some of the studies on the factors influencing gold price and various techniques used for studying these relationships are discussed in the following sections. We have forecasted gold prices based on multiple economic factors such as commodity research bureau future index, USD/Euro foreign exchange rate, inflation rate, money supply, New York Stock Exchange Index; Standard and Poor 500 index, Treasury bill and USD index. The study finds that Commodity Research Bureau future index, USD/Euro foreign exchange rate, Inflation rate and money supply have a significant impact on gold price. We have used multiple linear regression (MLR) model for forecasting the gold prices and are of the opinion that MLR model appeared to be useful for predicting the gold price. From the review of literature, it can be seen that multiple linear regression is widely used technique for understanding relationship among such variables.

Chapter 3 Functionality and Working of Project



1. Data Collection: The first thing required while building a machine learning model is the data. The data is collected from kaggle website consisting of 2290 records and 6 attributes.
2. Data Preprocessing: Data preprocessing is required when the data is incomplete, inconsistent or noisy. The data collected was noisy, so we performed outlier analysis and removed the noisy data. The data transformation is also done by performing normalization in which the data in each attribute is scaled between the range 0 to 1.
3.Data Analysis: Analysis of gold price and all other factors on which the gold price depends on like share market, crude oil price, silver price and dollar/euro ratio etc through various machine learning algorithms we can analyse data through plotting different charts and graph and also perform many calculations using numpy.
4.Train and Test split: The data was divided into training and testing data we have used 60% of data as training data and 40% data data to test the accuracy and error of models.

5. Training the model: The model is trained by importing the required model and by passing the training data to it. The dataset is splitted into train and test data. The linear model is imported from sklearn and the Random forest regressor module are imported from sklearn.ensemble. These models are trained by passing the train data. While conducting training, it is also important to record the metrics of each training process. The metrics that are tested are mean absolute error, root mean square error and r2 score.
6. Prediction and Evaluation: The trained model is checked by predicting the test data of the dependent variable using the test data of the independent variables.

A. Dataset: The data was sourced from the kaggle website consisting of ten years data from January 2008 to Decmeber 2018. It consists of the variables date, silver price, stock profit exchange, gold price ,US dollar rate and united states oil ETF. The dataset consists of 2290 records.
B. Machine Learning Algorithm
   Random forest is a supervised learning algorithm which performs both classification and regression tasks. This algorithm operates by constructing multiple decision trees during training time and outputting the mean prediction of individual trees.

Importing the Libraries

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn import metrics
```

Data Collection and Processing

```
# loading the csv data to a Pandas DataFrame
gold_data = pd.read_csv('/content/gld_price_data.csv')
```

```
# print first 5 rows in the dataframe
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 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 columns
gold_data.shape
```

```
(2290, 6)
```

```
# getting some basic informations about the data
gold_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2290 entries, 0 to 2289
Data columns (total 6 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   Date     2290 non-null   object 
 1   SPX      2290 non-null   float64
 2   GLD      2290 non-null   float64
 3   USO      2290 non-null   float64
 4   SLV      2290 non-null   float64
 5   EUR/USD  2290 non-null   float64
dtypes: float64(5), object(1)
memory usage: 107.5+ KB
```

```
# checking the number of missing values
gold_data.isnull().sum()
```

```
Date       0
SPX        0
GLD        0
USO        0
SLV        0
EUR/USD    0
dtype: int64
```

```
# getting the statistical measures of the data
gold_data.describe()
```

	SPX	GLD	USO	SLV	EUR/USD
count	2290.000000	2290.000000	2290.000000	2290.000000	2290.000000
mean	1654.315776	122.732875	31.842221	20.084997	1.283653
std	519.111540	23.283346	19.523517	7.092566	0.131547
min	676.530029	70.000000	7.960000	8.850000	1.039047
25%	1239.874969	109.725000	14.380000	15.570000	1.171313
50%	1551.434998	120.580002	33.869999	17.268500	1.303296
75%	2073.010070	132.840004	37.827501	22.882499	1.369971
max	2872.870117	184.589996	117.480003	47.259998	1.598798

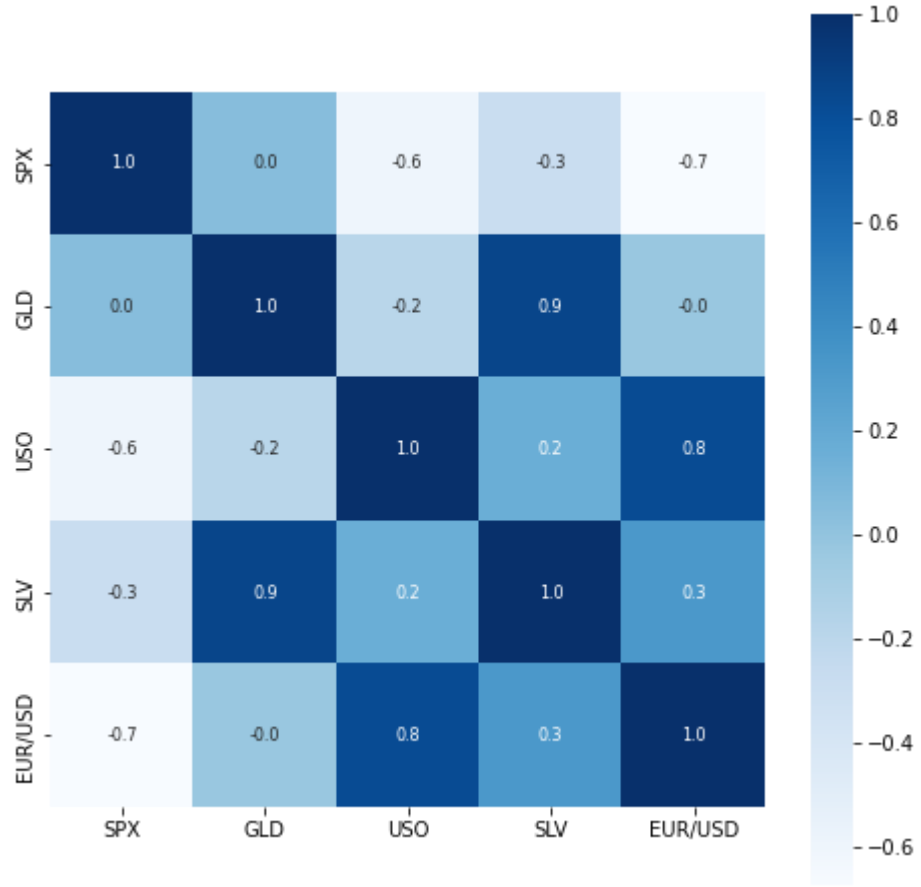
Correlation:

1. Positive Correlation
2. Negative Correlation

```
correlation = gold_data.corr()
```

```
# constructing a heatmap to understand the correlatiom
plt.figure(figsize = (8,8))
sns.heatmap(correlation, cbar=True, square=True, fmt='.1f',annot=True, annot_kws={'size':8
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f92ad962fd0>
```

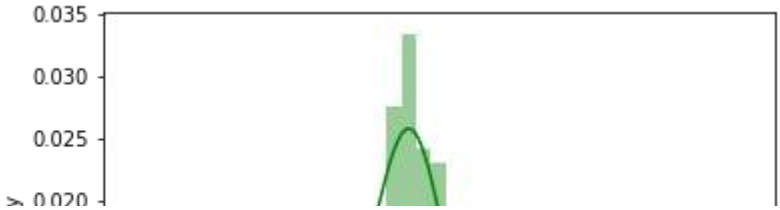



```
# correlation values of GLD
print(correlation['GLD'])
```

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

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

```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarnin
  warnings.warn(msg, FutureWarning)
<matplotlib.axes._subplots.AxesSubplot at 0x7f9297286350>
```



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)
```

Model Training: Random Forest Regressor

```
regressor = RandomForestRegressor(n_estimators=100)
```

```
# training 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)
```

```
[166.34190031 114.87020105 116.78880118  88.23559844 148.66910079
 120.28279984  89.45949996 112.02280007 117.13650017 118.59070116
  88.16739993  94.12839996 116.98330026 118.42540199 120.2474003
 126.7868979  121.96539988 150.64899994 165.81420075 118.55369948
 120.20110103 150.90880139 118.2849991  172.37969867 105.43979935
 105.00160135 148.69780104 113.92020023 124.93160104 146.80649978
 119.45590127 115.50540079 112.10680012 113.43430215 142.88960165
 117.65459777 103.06800034 115.7833011  103.46590183  99.07970035
 117.23780087  90.81320021  91.59140069 153.41199925 102.69529956
 154.05570096 114.33810166 138.18150197  90.0757985  115.46239954
 113.49589972 122.79950085 121.88320027 165.58670058  92.89779952
 135.65320186 121.2189001  120.78100062 104.3423004  140.48210299
 121.35559927 116.5920005  113.65110117 126.93729754 122.76909929
 125.79509949 121.165901    86.77289903 132.08500205 145.67620178
  92.76969944 157.66139921 159.33490239 126.12599896 165.30579942
 108.88129941 109.83940093 103.66579838  94.23860087 127.66520286
 106.9322004  161.61230022 121.61730057 131.82250081 130.59140154
 160.65840112  90.12089882 174.41400182 127.64419993 126.74889831
  86.62149885 124.68519911 150.22599694  89.60360027 106.89729958
 109.12889996  84.04339885 135.84869897 155.25950211 138.28130359
  73.46350032 152.34160216 126.00499958 126.69879987 127.43999893
 108.79699995 156.3194993  114.49650127 116.93150174 125.11079971
 154.05260099 121.36489975 156.4147992   93.00700048 125.4991015
 125.89250049  88.0403007   92.15129948 126.24309891 128.50960349
 113.14990063 117.84939765 120.82739976 127.03059768 119.32400083
 136.07140031  93.83849935 119.74740041 113.33280101  94.35609928
 108.71309969  87.35809915 109.22679946  89.57599943  92.48290036
 131.56300278 162.41270029  89.3742997  119.60130073 133.33950176
 123.69180009 128.4286019  102.0516985   89.07299874 131.40360083
 119.85520033 108.64649994 169.1223015  115.08020056  86.55719907
 118.81200057  91.05769976 161.72399992 116.46000051 121.59290006
 160.29999798 120.14439942 112.95339932 108.41869871 126.7433998
  75.81530051 102.98509988 127.44960232 121.77709943  92.64680011
 132.28500051 118.0059012  116.20759997 154.56600274 158.89590089
 109.98499985 155.23009819 119.42610074 160.72320098 118.73610067
 158.00309904 115.06869949 116.52240023 149.91059893 114.74460089
 125.19179846 166.92789997 117.65450002 124.84919922 153.27820393
 153.46810238 132.07930084 114.73780019 121.25600227 124.71840065
  89.82690007 123.13620023 154.55900163 111.85810027 106.64109972
 162.15400139 118.45649968 165.80289991 133.90630086 114.57059969
 152.99209897 168.64710005 115.0984     114.0736014  158.00899905
```

```
 85.34349857 127.09780061 127.80680086 128.87259925 124.39300108
123.91590045  90.58570073 153.46319974  97.25369967 136.59519974
 88.85249908 107.70269975 115.10590077 112.75050125 124.22079925
 91.39809888 125.39400138 162.44679884 119.68029952 165.128701
126.73849794 112.62610028 127.41489917  94.97919939  90.90329966
103.20939916 120.92109978  83.10579954 126.37920001 160.49340478
117.52330078 118.42789998 119.93470011 122.7189995  120.05360136
121.53429982 118.06680044 107.01169995 148.19999997 126.34809877
115.66420107  74.03309996 127.84000097 154.01420113 122.38499963
125.59260045  88.71740048 103.19989868 125.06390033 120.24210037
 73.37350094 151.83730028 121.03040024 104.69180042  86.67939779
115.10049926 172.23619779 120.03259999 160.29159699 113.19219969
121.14860055 118.28510085  95.98149981 118.70680066 125.32670042
118.57079958  96.11560079 153.62120172 122.20050006 147.00410029
159.0927024  113.67480001 122.65129909 148.97669821 127.21140018
165.77830021 136.0184996  120.04729926 166.95789882 108.35839947
121.66439883 139.14350057 106.91219884]
```

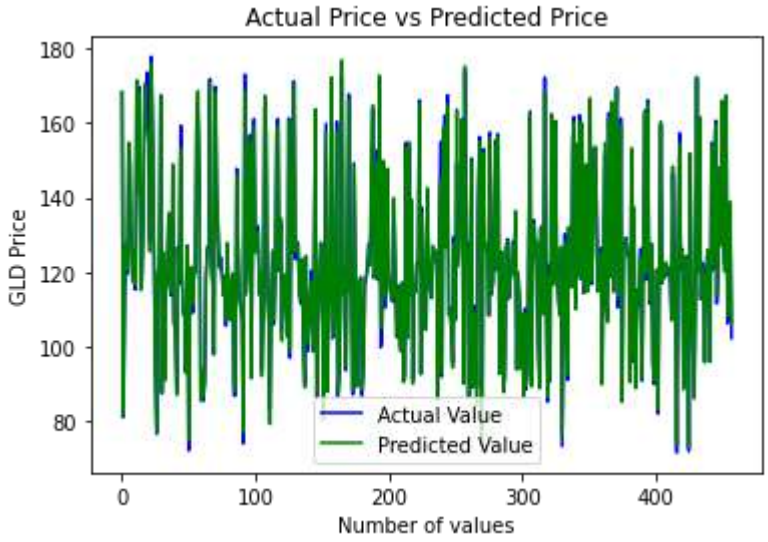
```
# R squared error
error_score = metrics.r2_score(Y_test, test_data_prediction)
print("R squared error : ", error_score)
```

```
R squared error :  0.9895391826941908
```

Compare the Actual Values and Predicted Values in a Plot

```
Y_test = list(Y_test)
```

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

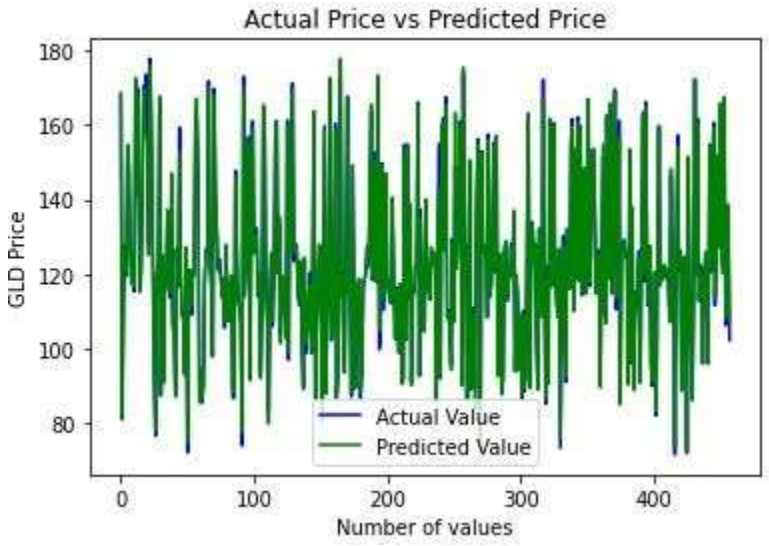


Chapter 4 Result and Discussions

After applying random forest regression techniques on the data, the results are given

R squared error : 0.9887647199316348

In random forest model, the accuracy obtained for training data is 99.83% and the accuracy obtained for test data is 99.77%. The accuracy difference is very less.



The accuracy is higher for random forest and the accuracy difference is also very less compared to gradient boosting. Hence, random forest regressor model is considered.

Chapter 5 Conclusion and Future Scope

This study was conducted to understand the relationship between gold price and selected factors influencing its price, namely stock market, crude oil price, silver price, euro and dollar ratio. Monthly price data for the period January 2008 to January 2018 was used for the study. The data was further split into two periods, period I from January 2008 to October 2015 during which period the gold price exhibits a raising trend and period II from November 2015 to January 2018 where the gold price is showing a horizontal trend. The machine learning algorithms, random forest regression was used in analyzing these data. It is found that the correlation between the variables is strong during the period I and weak during period II. While these models show good fit with data during period I, the fitness is not good during the period II. Random forest regression is found to have better prediction accuracy for the entire period. It is concluded that machine learning algorithms are very useful in such analysis, but the characteristics of the data influences their accuracy. Further research with such data and different techniques may be conducted for better understanding of the performance of these techniques.

For future work, we can improve the results and predict the price more accurately by incorporating the other factors such as gold production, crude oil price, platinum price, inflation to the data and by using deep learning.

REFERENCE

[1] V. K. F. B. Rebecca Davis, "Modeling and Forecasting of Gold Prices on Financia Markets," American International Journal of Contemporary Research, 2014.

[2] Iftikharul Sami and Khurum Nazir Junejo, "Predicting Future Gold Rates using Machine Learning Approach", International Journal of Advanced Computer Science and Applications, 2017.

[3] D Makala and Z Li, “Prediction of gold price with ARIMA and SVM”, Journal of Physics: Conference Series, 2021.

[4] Navin, Dr. G. Vadivu, "Big Data Analytics for Gold Price Forecasting Based on Decision Tree Algorithm and Support Vector Regression (SVR)", International Journal of Science and Research (IJSR), 2013.

[5] P. V. M. Vasava, P. G. M. Poddar, Sima P Patel, "Gold Market Analyzer using Selection based Algorithm", International Journal of Advanced Engineering Research and Science (IJAERS), 2016.