



Historical Data-Based Gold Price Prediction using Intelligent Algorithms

Dhanush N¹, Raghavendra R²

PG Student, Department of MSc CS-IT, Jain (Deemed-to-be University), Bangalore, India¹

Assistant Professor, School of CS & IT, Jain (Deemed-to-be University), Bangalore, India²

Abstract: Gold's price is always fluctuating, either rising or falling. Given that gold is a major element of the financial market, gold price prediction is an essential area of finance. Many machine-learning methods have been used in published studies to anticipate gold prices. Several classification techniques, including random forest, decision tree, logistic regression, and linear regression, are used in this work. This article's topic originates from study done to understand the worth of gold. There is currently a constant market for gold. The gold price trend shows that gold is one of the best investment strategies. It is, therefore, prudent to forecast the direction of the gold rate. Numerous statistical models can be used to forecast and model data. The price of gold is consistently shown to be nonlinear. Price prediction is key to sound financial and investing strategy. The price fluctuation of gold can be represented as an exponential curve. Convolutional neural networks are among the best tools for resolving nonlinearities in data, and RNNs are especially useful for time series forecasting and estimation. Using data from the World Gold Council, it is found that the suggested design is among the most effective financial forecasting techniques.

Keywords: Regression, linear regression, logistic regression, decision tree, random forest, Machine Learning and Prediction.

I. INTRODUCTION

Exploring historical gold prices serves as a valuable exercise for gaining insights that can inform strategic decisions in the buying or selling of this precious metal. In the broader context, the realms of savings and investment play pivotal roles in individuals' lives, with investments typically geared towards long-term wealth accumulation rather than immediate consumption. The dollar exchange rate, inflation, and monetary policy are just a few of the numerous variables that affect the price of gold. Both domestic and foreign scholars. Regression models are typically built using these factors. The historical data of the financial time series is modelled and analysed using the time series analysis method in this work [1]. This holds especially true in the context of India, which boasts one of the world's fastest-growing economies, presenting a landscape rich with surplus capital and diverse business prospects. Investors in such an environment have an array of financial instruments at their disposal, including stocks, deposits, commodities, and real estate, each carrying its unique set of risks and potential returns.

Among the myriad investment options, gold emerges as an appealing choice due to its appreciating value and versatile applications. Investors increasingly perceive gold as a safeguarding commodity, often turning to it as the "asset of last resort" during periods of uncertainty in established capital markets and foreign exchange arenas. The yellow metal, in this sense, becomes a defensive tool against the fluctuations and uncertainties prevalent in other markets.

The primary goal of it is to provide a thorough analysis that clarifies how well different intelligent algorithms predict gold prices. This research aims to advance the field by providing insights into the dynamics and predictive power of these algorithms in the context of gold price prediction by utilising historical data and state-of-the-art approaches. These data can be analysed using three machine learning algorithms: gradient boosting regression, random forest regression, and linear regression. It is discovered that there is a strong correlation during period I and a weak correlation during period II between the variables. Period I data fits these models well, but period II data does not support these models as well. Gradient boosting regression is found to provide better accuracy for the two periods taken separately, while random forest regression is found to have better prediction accuracy for the entire period [2].

Since gold is a commodity that can be traded, its value is inversely correlated with the US dollar. Demand is harmed when the US dollar appreciates relative to other currencies, making gold more expensive. Conversely, when the USD declines, the metal becomes more affordable for buyers in other countries, which raises the price of gold. Additionally, gold is used to make jewellery, which is particularly well-liked for festivals and weddings in China and India, two of the world's largest consumers [3].



The goal of it is to deepen our understanding of the complex interactions between economic variables and gold prices, opening the door to more precise forecasts and well-informed choices in the fields of financial markets and gold investments.

The prediction of gold prices has been a subject of extensive research owing to the significant economic implications and investment interests associated with this precious metal. The application of sophisticated machine learning methods and clever algorithms has drawn interest recently as potential methods for predicting gold prices from past data. Research on stock market analysis is highly popular. It is extremely difficult to forecast the stock markets with accuracy. Cascade statistical models are used to forecast future stock markets. We apply the model to the MCX commodity (Gold) in order to represent gold [4].

It explores the use of intelligent algorithms to forecast gold prices through the use of historical data. Using a variety of machine learning models and algorithms for precise prediction and analysis, it seeks to investigate the complex link between many economic conditions and gold prices. Investor's preference for gold as a protective asset increases due to their negative expectations concerning the situation in the developed foreign exchange markets and the capital markets [5]. Due to their pessimistic expectations about the state of the developed capital and foreign exchange markets, investors' preference for gold as a protective asset is growing. Notably, the dynamics of supply and demand, inherent to any commodity, also exert influence on gold prices.

However, gold stands apart in its resilience to annual production fluctuations. Its capacity for long-term storage and the accumulation of substantial quantities over the years mean that yearly production variations have limited impact on its prices. The rising value of gold, particularly when juxtaposed against the declining fortunes of real estate and financial markets, has catapulted it into the spotlight as a preferred investment option.

Nevertheless, recent market dynamics have introduced a level of volatility to gold prices, thereby amplifying the associated risks. Uncertainty looms over the duration of these elevated prices and the eventual downturn. While extensive research has explored the multifaceted impact of economic variables on gold prices, this article endeavors to delve into the specific relationships between economic and market factors and the price of gold. The essay's structure encompasses an introductory section, a comprehensive review of relevant literature, segments dedicated to model planning, data exploration, and model building, culminating in a conclusive section.

Crucially, the historical gold price dataset under scrutiny comprises seven essential columns: Date, Open, Close, Low, High, Volume, and Currency, providing a robust foundation for the detailed analysis and examination undertaken in the subsequent sections of the article.

II. LITERATURE REVIEW

The survey of literature encompasses a diverse range of research papers that examine the relationship between gold prices and various economic, financial, and market factors. Together, these studies show how complex the underlying dynamics are, providing a detailed picture of the variables affecting gold prices.

Various factors influencing gold values have been explored in literature, with a consistent observation that gold values evolve in conjunction with the dollar and general market returns. Socioeconomic factors and their impact on gold rates have also been extensively studied, including the relationship between gold prices and other commodities, such as crude oil. However, findings from these studies appear to be somewhat contradictory, prompting a need for further exploration.

The factors affecting gold prices have been a focal point in literature, with different studies proposing various approaches to analyze these relationships. For instance, Manjula K. A. and Karthikeyan highlight the solid relationship between gold and factors like the cost of petroleum and dollar-rupee conversion based on their literature review. Historical perspectives, such as gold being used as a mode of payment and representing a country's financial strength, are also explored in studies like the one by Iftikharul Sami and Khurum Nazir Junejo.

In a model proposed by R. Hafezi and A. N. Akhavan, artificial neural networks, specifically the BAT-Neural Network, are suggested as effective tools for predicting future gold prices. Xiaohui Yang, on the other hand, recommends the ARIMA model as the most reliable among various models for predicting gold prices. Deep multiple kernel learning (DMKL) models have been employed by Shian-Chang Huang and Cheng-Feng Wu to project oil prices using data from oil, gold, and currency markets [6].



Within this expansive landscape of research, economic indicators emerge as pivotal determinants of gold prices. Factors such as petroleum costs, currency conversions, and inflation rates have garnered significant attention for their observed correlations with fluctuations in gold values. Researchers, exemplified by Manjula K. A., Karthikeyan, emphasize the robust relationship between these economic indicators and gold prices, highlighting their importance in shaping the yellow metal's market trajectory. In the realm of predictive modeling, various methodologies have been employed to forecast gold prices [7].

Studies by R. Hafezi, A. N. Akhavan, Xiaohui Yang, and others advocate for sophisticated models like artificial neural networks, ARIMA, and Deep Multiple Kernel Learning. These models, validated through empirical analysis, have demonstrated promise in accurately predicting future gold prices, offering insights into potential trends and movements in the market. However, amidst the research consensus lie divergent viewpoints regarding the correlation between socioeconomic measures, inflation rates, and gold returns. Perspectives vary, as evidenced by contrasting opinions from Lawrence, Dr. Scassiavilanni, and Hanan Naser, underscoring the complexity and perhaps the subtlety of these relationships.

Moreover, comprehensive analyses by Ismail et al., Khaemasunun, Ai et al., Malik and Ewing, Ghosh et al. delve into a multitude of economic factors. These factors span across currency exchange rates, energy costs, and various market indicators, collectively showcasing their substantial impact on gold prices [8].

The depth of these studies sheds light on the intricate interdependencies between different markets and their influence on the value of gold. Within the analytical landscape, multivariate regression models, especially multiple linear regression, have emerged as key tools for understanding the sensitivity of gold prices to multifaceted variables. Toraman's insights underscore the wide adoption and efficacy of these models in evaluating the complex relationships shaping gold prices [9].

Lastly, the indication of employing advanced algorithms for historical data-based gold price prediction signifies an evolving trend in research methodologies. This suggests a shift towards leveraging sophisticated computational techniques to unravel the intricate patterns and trends governing gold price fluctuations, reflecting a dynamic and evolving landscape in the study of gold markets.

III. MODEL PLANNING

Following a review of existing literature, this study has identified five key elements believed to influence the price of gold. The factors under consideration encompass the stock market, petroleum oil prices, the exchange rate between the rupee and the dollar, inflation, and interest rates.

To accurately depict stock values, the Nifty 500 indicator values are utilized, representing the top 500 firms on the National Stock Exchange. The goal of this research was to create a forecasting model that would be able to predict gold prices based on a variety of economic variables, including changes in currency prices and inflation. As a result of the US dollar's decline, investors are now placing their money in gold since it has a significant stabilising effect on investment portfolios [10].

Inflation is assessed using the Consumer Price Index with a reference year, while interest rates are represented by accounts opened for durations exceeding one year and their corresponding term savings. The current gold price in rupees per unit serves as a reflection of the gold price. Monthly statistics for each of these factors were systematically collected.

Datasets from the Centre for Observing Indian Economy were employed to obtain 228 samples for each variable. The dataset was split, allocating 20% for testing purposes and utilizing the remaining 80% to train the algorithm. The machine learning techniques employed in this study involve linear regression and gradient boosting regression.

a. Data Discovery:

1. Learning about domain:

The focus is on understanding the Gold price trends over time, involving details such as Date, Open, Close, Low, High, and Currency. Each record in the dataset corresponds to a specific date, providing information on the opening and closing amounts, high and low values, and the currency associated with the gold amount.



2. Identifying resources:

Kaggle is utilized as the platform for the dataset, capturing records based on date and calculating currency based on opening and closing amounts.

3. Framing the problem:

The goal is to predict the currency value for a given amount of gold on a specific day, considering factors such as opening and closing amounts.

4. Identifying key stakeholders:

Primary stakeholders include shareholders, gold miners, investors, traditional authorities, and non-governmental/community-based organizations.

5. Interviewing the Analytics Sponsor:

Questions involve exploring whether the gold price is low or high, understanding reasons for changes in gold prices, and the ability to predict gold price fluctuations.

6. Developing Initial Hypothesis:

Utilizing classification and regression models, such as decision trees, logistic regression, linear regression, and random forest, to forecast the currency value of gold within a specific time period.

7. Identifying Potential Data Sources:

Leveraging information about gold prices and reviewing raw data related to historical gold prices.

b. Data Preprocessing: Data preprocessing is a crucial step in data preparation for subsequent analysis. The steps involved include:

1. Preparing systematic sandbox:

Importing and loading the dataset, "Daily Gold Price Historical Data," into the R environment.

2. Executing ETLT (Extract Transform Load Transform):

Extracting and transforming data from the "Daily Gold Price Historical Data" dataset, preparing it for cleaning and analysis.

3. Learning about the data:

The dataset comprises multivariable data, consisting of 5774 rows and 7 columns related to Gold Prices, including Open, Close, High, Low, and Volume.

4. Data conditioning:

Cleaning and transforming the "Daily Gold Price Historical Data" dataset in the R environment based on specified formulas for classification and regression analysis.

5. Survey and visualize:

Analyzing and reviewing the dataset for negative values, ensuring that any discrepancies are addressed and the data is well-prepared for subsequent analysis.

This comprehensive model planning approach ensures a clear understanding of the problem, the relevant stakeholders, and the necessary preprocessing steps to facilitate effective analysis and modeling of gold price trends.



IV. ARCHITECTURE

This architecture states the process that we are going to build to predict gold prices.

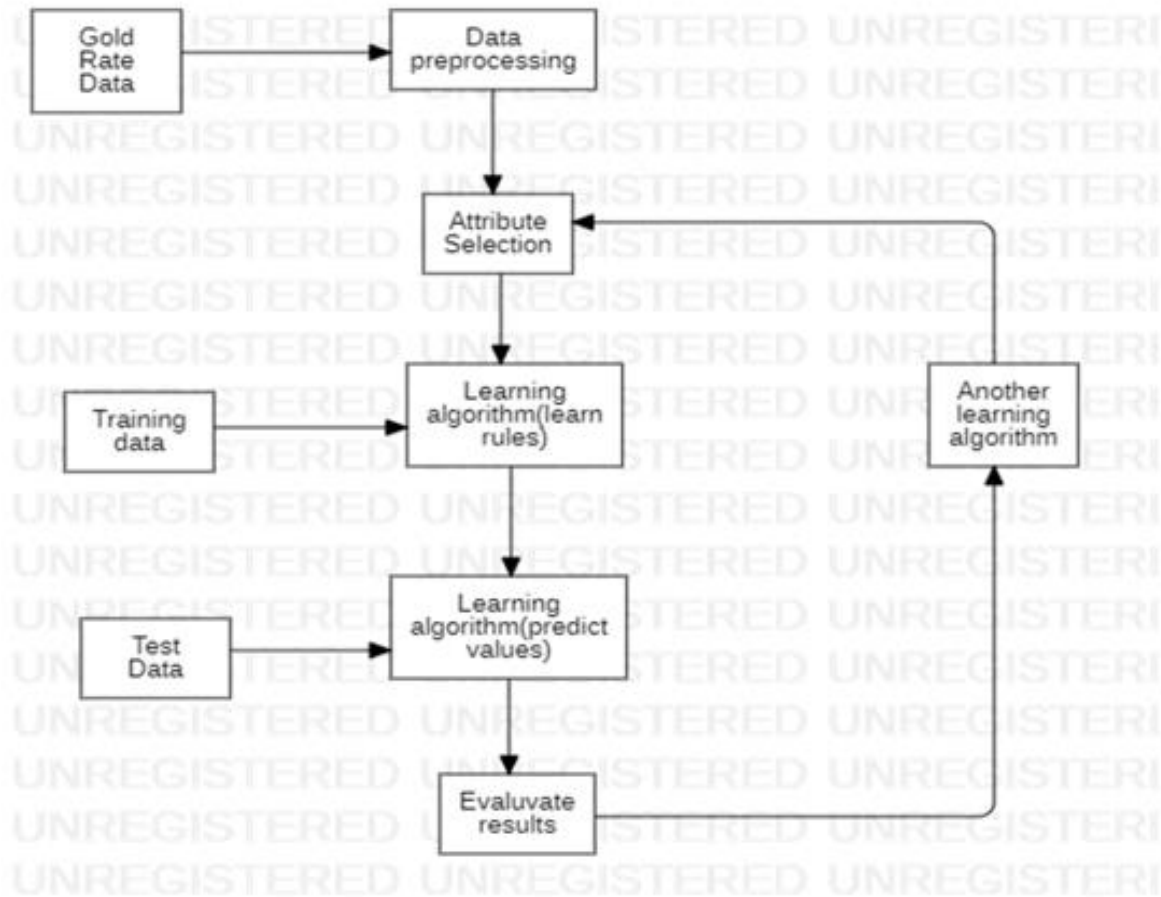


Fig-1 Architecture Process Flow Of The Prediction.

We use a variety of machine learning approaches to train and develop the model using the given data. The majority of the data is used to prepare the models, with the remaining amount set aside to evaluate their effectiveness. In this work, we employ various machine learning techniques, such as the Random Forest Regression, the Linear Regression, and the LSTM Model.

Steps Involved are :-

- **Defining Explanatory Variables**

Since explanatory factors determine the value of the Gold EFT price, we specify them. We can easily use these factors to anticipate the price of the Gold ETF. The explanatory variables in this strategy are the moving averages for the last three and nine days. We use the `dopna()` function to remove the NaN values and store the feature variable.

We must increase X by other elements in order to predict the price of the Gold ETF. Technical indicators, such as those seen in the US financial data or the Gold Miners ETF (GDX) or Oil ETF (USO), might be one of these causes.

- **Divide The Data Into Training And Testing Dataset**

This is the critical step when the historical dataset is split into training and testing data subsets. In predictive modelling, this division is essential. The training set is important since it is the foundation upon which the linear regression model is built. The process of creating a model relies on combining expected outcomes and using past trends to help the model understand underlying linkages.



In the meanwhile, the testing subset improves the accuracy and dependability of the model by validating its performance on unobserved data. During this procedure, the model is trained based on the expectation of aggregated results, making sure the model recognises and understands predicted patterns in the dataset.



Fig-2 Storage Of Historical Data.

• **Create A Linear Regression Model**

Regression analysis is the statistical technique used to determine how numerous variables are related to one another. This study explains why certain independent values vary while other values stay the same and why the dependent values fluctuate.

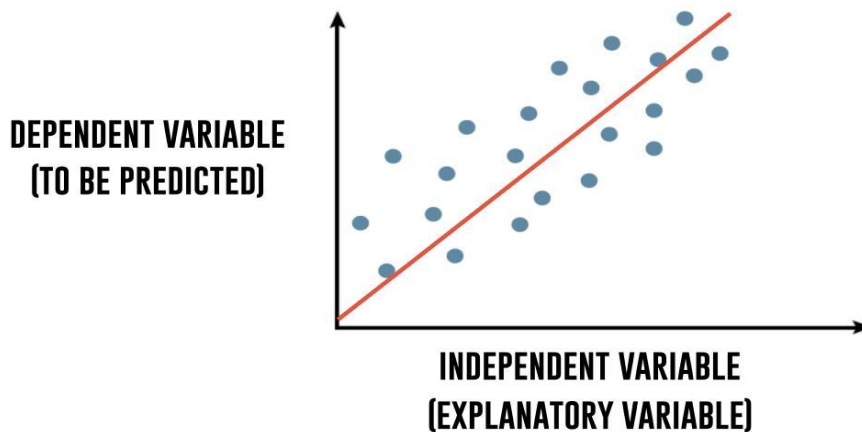


Fig-3 Regression variable model.

Multiple independent variables in a linear regression are called multiple linear models. A multiple linear regression is shown below, using Y as the dependent variable and X1, X2, and X2 as the independent variable values.

$$Y = a + (b1)*(X1 + (b2*)X2)+(bp*Xp) \dots\dots\dots$$

• **Evaluate And Predict The Gold ETF Prices**

Now, it's time to test the model using the test dataset to predict gold prices based on the knowledge acquired from the training dataset. Using the associations that were learnt during training, this predictive technique determines the expected gold prices ('y') given the explanatory factors ('X'). In essence, the model predicts how gold prices would behave in hypothetical data points within the test dataset by using its comprehension of the relationships between these factors and gold prices. This assessment stage assists in determining the model's precision and its capacity to generate trustworthy forecasts outside of the training set of data.



Model Building

• Model Planning

A model that anticipates expenses with the highest degree of precision would be picked to power a tool or programme.

In the pursuit of developing a model with the utmost precision in predicting expenses, the selection of a robust tool or program is imperative. The chosen model must exhibit accuracy and efficiency to power applications effectively.

1. Decision Tree:

A decision tree, often referred to as a selection tree, functions as a tool for both classification and prediction. It is structured like a flowchart, with each internal node representing the result of a particular check and each leaf node representing the check's consequence. This method, which bases its predictions on taught data, is similar to supervised learning. The selection tree is a common data mining tool because it provides a visual representation of decisions, which facilitates comprehension.

Our two packages of choice for processing these decision tree models are RPART and PARTY. By using these packages, we may make predictions based on decision trees, which helps us understand the choices that were made during the forecasting process.

2. Random Forest:

In supervised learning, the Random Forest technique is useful for both classification and regression tasks. It's an ensemble learning technique that combines several classifiers to improve model performance and solve challenging issues. In contrast to a single Random Forest decision tree, this method provides variety by using several characteristics in the model construction process, yielding reliable results.

For regression and classification tasks, Random Forest is a collective technique that uses several decision trees with Bootstrap Aggregation (bagging). The fundamental idea is to use numerous decision trees rather than just one in order to minimise bias and oscillations that come with utilising individual Decision Tree models. The ultimate goal of this method is to combine several trees. The key benefit of Random Forest over other algorithms is that it minimises bias while lowering variance. Additionally, it facilitates feature engineering by selecting the most important features from the available dataset.

3. Linear Regression:

Linear regression is a modeling technique employed to elucidate the connection between one or several independent variables and a continuous dependent variable. In any instance of linear regression, the objective is to determine a line that best fits a set of data points, often achieved through the least squares method. To construct a linear regression, the 'lm()' function is commonly utilized.

Multiple linear regression models, a variation of linear regression, encompass more than one independent variable. In these instances, Y represents the dependent variable, while X1, X2, and so forth denote the independent variables.

$$Y = (a + b_1) * (X_1 + b_2) * (X_2 + \dots + b_p) * (X_p) \dots\dots$$

It currently functions as both a statistical method and a machine learning tool.

4. Logistic Regression:

Predictive analytics and classification activities make considerable use of logistic regression, a fundamental mathematical model. The primary function of this algorithm is to forecast event probabilities by using a predetermined set of independent variables present in the data. Most of the time, the dependent variable's outcome falls between 0 and 1. Applying a logic transformation to these probabilities is known as logistic regression. In order to create a logistic regression model for data, one frequently uses the 'glm()' programme.

These diverse modeling approaches offer a range of tools to cater to different aspects of predictive analytics, classification, and relationship modeling between variables. The selection of the appropriate model depends on the nature of the data and the specific goals of the prediction task.



- **Model Building**

Dataset: Daily Gold Price Historical Data

Decision tree execution (using party package):

Algorithm 1: Decision Tree on Trained dataset

```
set.seed(i,ii,iii,iv)
i<-sample(2,nrow(gold2),replace=T,
prob=c(0.7, 0.3))
      tD <- gold2[i==1,]
      testD <- gold2[i==2,]
      library(party)
      MF <- Open~High+Low+Close+Volume
      Gctree<-ctree(MF, data=trainData)
      table(pred1(gold_ctree),TrainData$Open)
      print(gold_ctree)
      plot(gold_ctree)
      tPred<pred1(gold_ctree,newdata=testData)
      table(tPred, testData=gold2$Open)
```

In the above algorithm decision tree algorithms is applied on the trained dataset for the prediction of the gold.

Algorithm 2: Decision tree execution (using rpart package)

```
MF <- Open~High+Low+Close+Volume
gold_rpart <- rpart(myFormula, data = gold.train,
control = rpart.control(minsplit =10))
attributes(gold_rpart)
pt(gold_rpart$scptable)
pl(gold_rpart)
te(gold_rpart, use.n=T)
opt <- which.min(gold_rpart$scptable[,"xerror"])
cp <- gold_rpart$scptable[opt, "CP"]
gold_prune <- prune(gold_rpart, cp = cp)
pl(gold_pru)
te(gold_pru, use.n=T)
```

The algorithm described above uses the rpart package to predict gold by applying decision tree algorithms to a trained dataset.

Algorithm 3: Random Forest execution:

```
in<-sample(2,nrow(gold2),replace=T,          prob=c(0.6, 0.2))
trData <- gold2[in==1,]
teData <- gold2[in==2,]
library(rForest)
rf <- rForest(Open ~ ., data=trData,  ntree=100, proxi=TRUE)
tab(pred(rf), trData$Open)
plot(rf)
varImpPlot(rf)
```

The training dataset is used in the technique above to apply Random Forest, which is then used to forecast the gold. Similar to this, the historical data—which is regarded as the training data—is subjected to the logistic and variable significance algorithms.



V. RESULTS

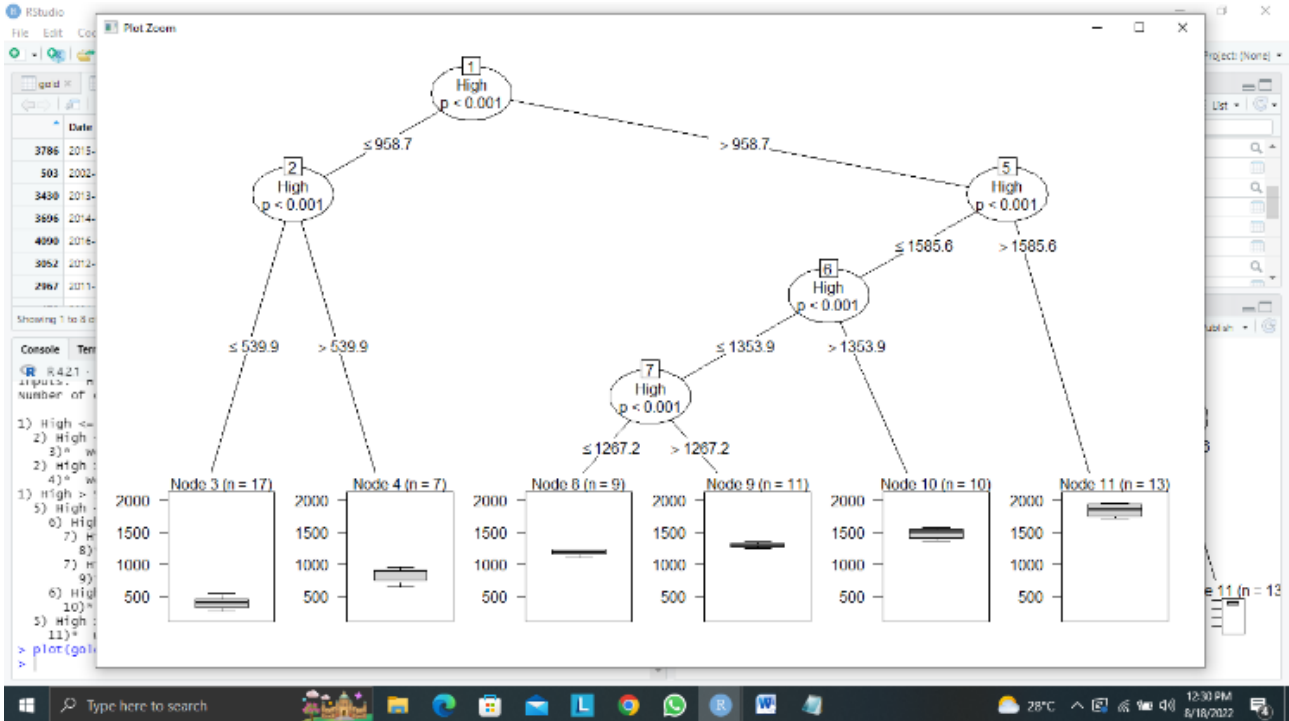


Fig-4 Decision Tree with Package Party

It provides the decision tree procedure together with the package party for the gold prediction.

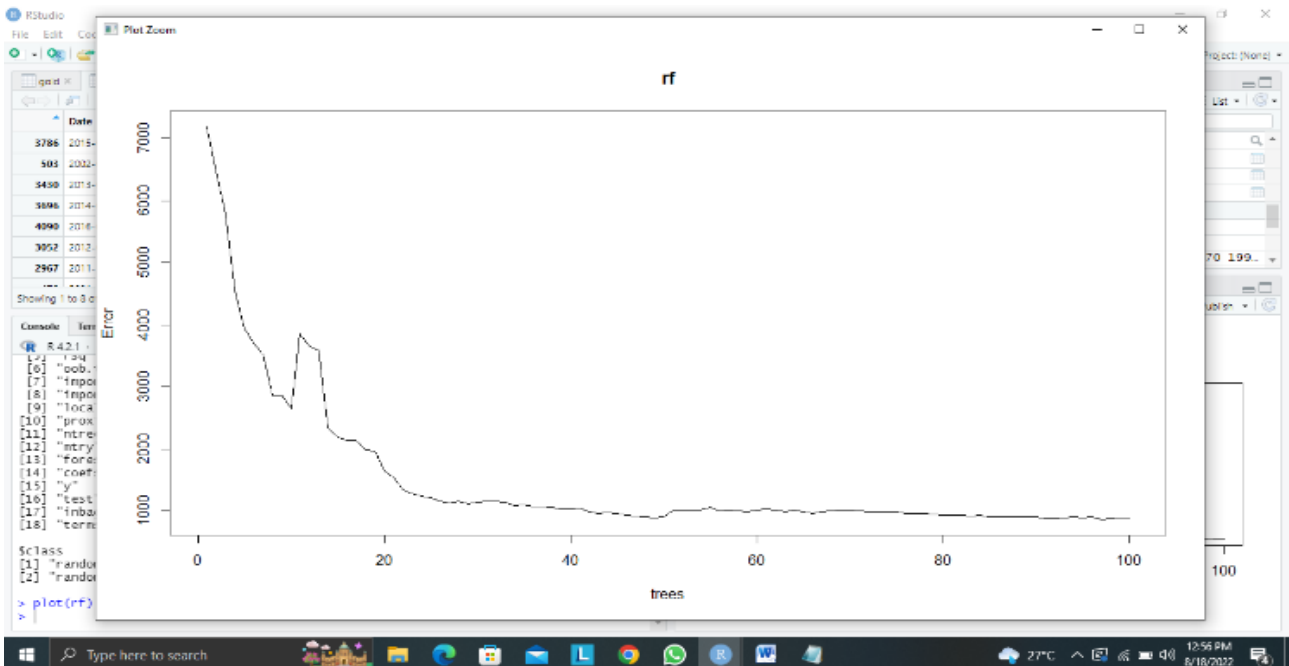


Fig-5 Random Forest

It provides the Random Forest algorithm for gold prediction using training data.

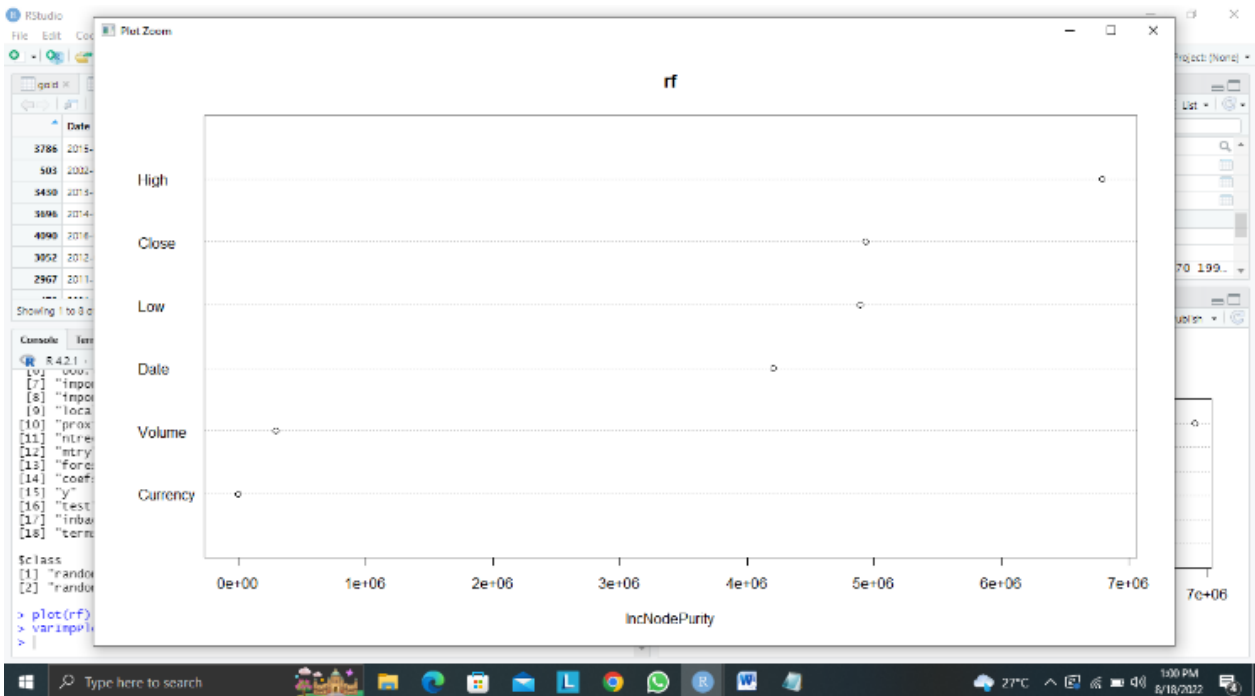


Fig-6 Variable Importance

It provides the variable importance technique for gold prediction using training data. It indicates the precise days of significance on which price increases occur.

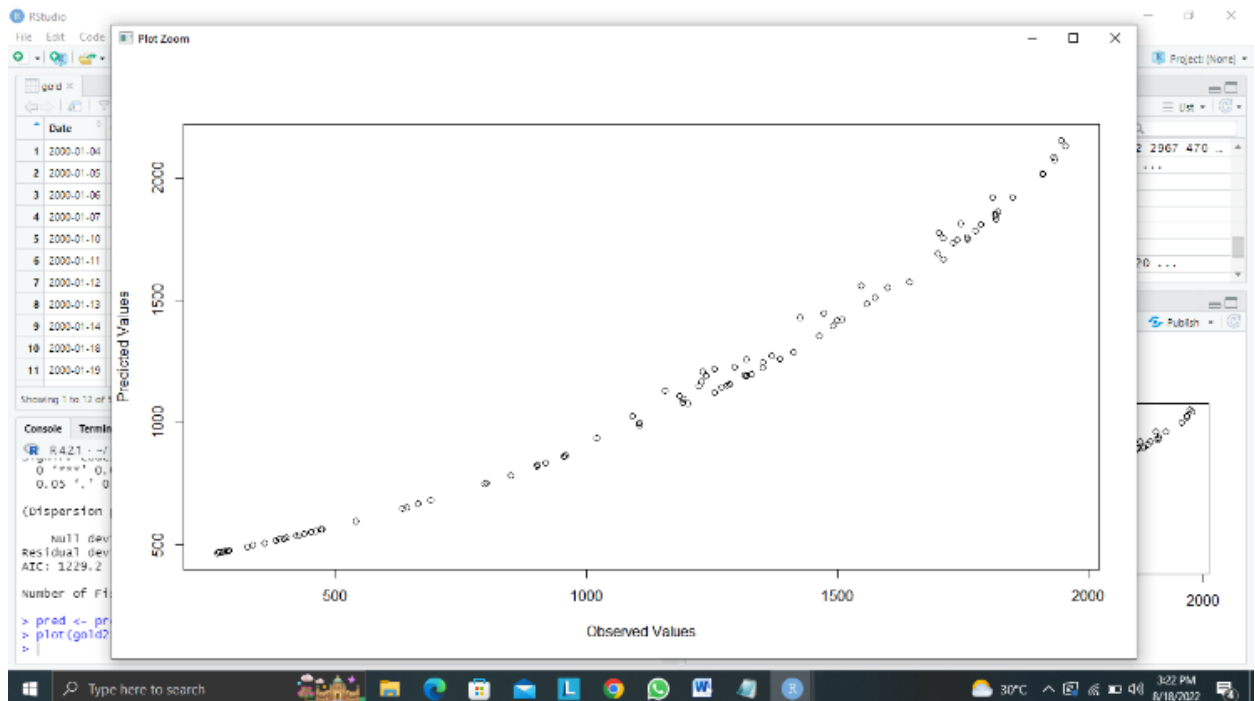


Fig-7 Logistic Regression

It provides the logistic regression approach for gold prediction using training data.



VI. CONCLUSION

Using a dataset of daily gold price records, the research presented explores the use of several classification algorithms for gold price prediction. Of them, the decision tree algorithm stands out as being very good at prediction because of how simple it is and how well it can adjust to the features in the dataset.

A key component of the research is examining the complex relationship between gold prices and several important variables. A number of variables are carefully looked at, including interest rates, inflation rates, rupee-to-dollar ratios, stock market performance, and petroleum oil costs. Through the examination of monthly price information, discernible patterns in the variations of gold prices surface, providing insight into periods of notable fluctuations in gold prices. The study investigates the correlation between gold prices and influential factors such as the stock market, petroleum oil costs, the rupee-to-dollar exchange rate, inflation, and interest rates. Monthly price statistics are analysed, revealing trends like fluctuations in gold prices during certain periods.

Three different machine learning methods are used in the study to examine and determine the association between these important characteristics and gold prices: gradient boosting regression, random forest regression, and linear regression. The results demonstrate a strong association between these factors, and gradient boosting regression outperforms random forest regression in terms of prediction accuracy. This result emphasises how well sophisticated regression methods capture the intricate relationships present in the dataset. Essentially, the study highlights the critical role that machine learning algorithms play in interpreting the fluctuations in gold prices with respect to different economic aspects. It highlights how important it is to modify these algorithms to fit the particular features of the dataset in order to get predictions that are more accurate. In addition, the paper recommends greater investigation and in-depth study to improve and maximise the use of these machine learning techniques in gold price forecasting in the context of changing economic environments.

In summary, the research underscores the significance of machine learning algorithms in understanding the dynamics of gold prices in relation to economic factors. It emphasizes the need to consider the unique characteristics of the dataset for accurate predictions. Further exploration and research are recommended to enhance understanding and application of these machine learning approaches in forecasting gold prices.

REFERENCES

- [1]. Xiaohui Yang, "The Prediction of Gold Price Using ARIMA Model", 2nd International Conference on Social Science, Public Health and Education 2019 <https://doi.org/10.2991/ssphe-18.2019.66>
- [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 - <https://doi.org/10.1109/ICOEI.2019.8862557>
- [3]. Mrs. B. Kishori I, V. Preethi, "Gold Price forecasting using ARIMA Model", International Journal of Research, 2018.
- [4]. K. R SekarManav Srinivasan, K. S. Ravichandran and J. Sethuraman, "Gold Price Estimation Using A Multi Variable Model", International Conference on Networks & Advances in Computational Technologies, 2017 <https://doi.org/10.1109/NETACT.2017.8076797>
- [5]. J. Jagerson and S. W. Hansen, "All about investing in gold", McGraw-Hill Publishing, 2011.
- [6]. R. Hafezi* , A. N. Akhavan, "Forecasting Gold Price Changes: Application of an Equipped Artificial Neural Network", AUT Journal of Modeling and Simulation, 2018.
- [7]. Manjula K. A., Karthikeyan P, "Gold Price Prediction using Ensemble based Machine Learning Techniques", Third International Conference on Trends in Electronics and Informatics, 2019 - <https://doi.org/10.1109/ICOEI.2019.8862557>
- [8]. P Khaemasunun, "Forecasting Thai gold prices," Available Http://www Wbiconpro Com3-Pravit. Pdf Acss, vol. 2, 2014.
- [9]. C. Toraman, Ç. Basarir, and M. F. Bayramoglu, "Determination of factors affecting the price of gold: A study of MGARCH model," Bus. Econ. Res. J., vol. 2, no.4, p. 37, 2011.
- [10]. Z. Ismail, A. Yahya, and A. Shabri, "Forecasting gold prices using multiple linear regression method," Am. J. Appl. Sci., vol. 6, no. 8, p. 1509, 2009. <https://doi.org/10.3844/ajassp.2009.1509.1514>