iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 9 **Issue:** VI **Month of publication:** June 2021

**DOI:** https://doi.org/10.22214/ijraset.2021.36028

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429*
*Volume 9 Issue VI June 2021- Available at www.ijraset.com*

# Prediction and Analysis of Gold Prices using Ensemble Machine Learning Algorithms

Gudipally Chandrashakar[1], Talapalli Vineeth Kumar[2], Sai Samhith Thatikonda[3], Yashwanth Chennu[4], Bejjam Vasundhara Devi[5]
[1,2,3,4]*Student,* [5]*Assistant Professor, Computer Science and Engineering,*
*Sreenidhi Institute of Science and Technology, Hyderabad, India*

***Abstract— In this article, we used historical time series data up to the current day gold price. In this study of predicting gold price, we consider few correlating factors like silver price, copper price, standard, and poor's 500 value, dollar-rupee exchange rate, Dow Jones Industrial Average Value. Considering the prices of every correlating factor and gold price data where dates ranging from 2008 January to 2021 February. Few algorithms of machine learning are used to analyze the time-series data are Random Forest Regression, Support Vector Regressor, Linear Regressor, ExtraTrees Regressor and Gradient boosting Regression. While seeing the results the Extra Tree Regressor algorithm gives the predicted value of gold prices more accurately.***

***Keywords— Gold Prices, Forecasting, Linear Regression, Gradient Boosting Regression, Random Forest Regressor, Support Vector Regressor, Extra Tree Regressor.***

## I. INTRODUCTION

Machine learning has also been used to forecast financial variables, but typically with an emphasis on forecasting stocks rather than commodities. I chose to be interested in my project. Supervised learning is applied to gold price forecasting to see what kind of achievement I could accomplish. I have developed this as a regression problem: given historical worth knowledge up to a given day, my algorithmic rule tries to predict whether or not the gold worth tomorrow is going to be higher or lower than it's nowadays.

From the point in time, I expected that feature definition would be the foremost difficult element of this project, as the price knowledge alone lends very little insight into future value movement. The goal so becomes to seek out and to calculate the options that best capture the movement of gold costs and supply information on not solely their past and current movements, but also their future movements. Investments and savings are an important part of everyone's lives. Investments refer to the use of gift money with the intention of earning a profit in the future. In a monetary context, investment can be described as the purchase of property that isn't utilized right now but can be used to produce wealth in the future. In financing, funding is the purchase of a capital product with the expectation of benefit in the future or later resale at a higher price. The Indian economy is among the fastest growing in the world, which has resulted in higher disposable income levels and a multitude of financing options.

There are some funding avenues to be had for traders, which incorporate shares, commodities, and actual property. In terms of chance and return characteristics, each of them is distinct. Because of its rising price and ease of use, gold is another currency that many traders are considering as a viable financing option. Investor's preference for gold as a shielding asset will increase because of their negative expectancies regarding the state of affairs inside the advanced foreign change markets and the financial markets [1]. Gold is also known as "the currency of very last choice," meaning that it is the asset that traders depend on when the mature global financial markets are unable to have adequate returns [2].

As a result, traders may regard gold as a means of protecting themselves against market volatility. Gold is a precious commodity, its price is determined by supply and demand, much as any other product. However, since gold is storable and the delivery is spread out over decades, this year's production has no impact on its prices. Gold is seen as both a commodity and a currency. Gold looks much less like an asset than shares or stocks, which are long-term investments. The gold price is determined by several interconnected factors, including inflation rates, currency fluctuations, and political unrest[3]. The rising price of gold, along with the volatility and falling prices of other markets such as stock markets and real estate markets, has drawn a growing number of traders to gold as a viable source of finance.

However, the overdue charge of gold is still experiencing huge uncertainty, making gold investments riskier. There is concern about whether or not those excessive costs are sustainable and while the costs could reverse. Despite the fact that some study has been done on the relationship between the price of gold and a few monetary attributes. It remains taken into consideration that observation to expose the effect and effect of diverse macro-monetary elements on the charge of gold inside the gift state of affairs might be beneficial in figuring out the dynamic results of those relationships. Thus this paper is aimed toward reading the

connection among gold charge and decided on monetary and marketplace variable. Understanding such courting might be beneficial now no longer most effective to financial policymakers however additionally to fund managers, traders and portfolio managers to make more market-based funding decisions. In addition, in reading the data, this study employs five system learning algorithms, including linear regression, extra tree regression, gradient boosting regression, support vector regression, and random forest regression. The comparison of those five strategies will assist us in figuring out the accuracy of those strategies below diverse conditions. This research paper is established with literature evaluation with inside the subsequent phase accompanied with the aid of using sections on information, effects and conclusion.

## II. PROPOSED METHODOLOGY

### *A. Data Set*

This data is collected from Yahoo finance which is the most reliable website for the collection of data. This dataset has data from January 2008 up to the existing date, previous data before January 2008 is not considered due to the abrupt Financial Crisis which occurred in the year 2007-2008. The Financial Crisis of 2007-08 was a worse economic crisis than the Great Depression of 1929. The Financial Crisis of 2007-08 was regarded as "The Great Recession". So, it's better to consider the data after The Great Recession of 2007-08 up to the existing date.

The price of gold that we have a tendency to be attempting to predict is taken in US Dollar. Tons of improvement and pre-processing were performed on the dataset. The matter of missing values, noisy values were handled inappropriate manner to finish the dataset. The values in attributes are rounded up to two decimals. Attributes that are considered in the dataset are silver price, copper price, standard and poor's 500 value, dollar-rupee exchange rates, Dow Jones Industrial Average Value. Few Precious metals like Silver and Copper are being considered in this attributes list because of their high correlating factor.

The current gold prices train-test dataset was split into a fashion of random sampling instead of a sequential fashion. As the data set contains in total of 3303 records where training data contains 2642 records which are 80% of total data records and testing data contains 661 records which is 20% of total data records.

### *B. Correlation Analysis*

Correlation analysis is used to evaluate the strength of bonding between two attributes. The correlation value of the two variables ranges between -1 to +1. Two types of Correlation analysis are present namely Strong Correlation Analysis and Weak Correlation Analysis. Strong Correlation mainly emphasizes the value of more than 0.5. Weak Correlation mainly emphasizes the value less than 0.5. In This Case study, the prediction attribute is gold value to predict that gold value few correlating attributes are considered like silver price, USD to INR change price, Dow Jones Industrial Average price, Copper Price, Standard, and poor's 500 value are taken into consideration of training the models.

TABLE I
CORRELATION ANALYSIS BETWEEN ATTRIBUTES

| Attributes | Correlating w.r.t Gold Price |
|---|---|
| Simple Moving Avg(3) | 0.997409 |
| Simple Moving Avg(9) | 0.994515 |
| Silver Price | 0.704504 |
| USD/INR Price | 0.407761 |
| S&P 500 Price | 0.403195 |
| DOW Price | 0.394015 |
| Copper Price | 0.328643 |

### *C. Data Pre-Processing*

The data which is obtained from Yahoo Finance is taken separately for each attribute then merged all the attributes into one data frame. Data initially obtained from Yahoo Finance consists of Date attribute is in String data type and it is converted to Date object data type using to date time method of pandas module. Every dataset obtained from yahoo finance consists of 5 attributes namely Open, High, Low, Close, and Volume of that particular day where the close of that particular attribute if that day is considered If there are any null values in any one of the attributes then the total record is not considered. Rather than replacing the null values in the Dataset, removing two or three records will be more efficient. Dates of Gold Data are taken into consideration and the dates of all other attributes should be the same if not the dates are removed. The data of date

attribute is further splitter into 4 attributes namely 'Year', 'Month', 'Date', 'Week Day'. Furthermore in Week-Day, every day is represented in a numerical representation like Sunday as '0', Monday as '1', Tuesday as '2', Wednesday as '3', Thursday as '4', Friday as '5', Saturday as '6'.After splitting the Date attribute into 4 other attributes the date attribute was removed from the working dataset.

The modern-day gold expenses train-test dataset used to be break up into a trend of random sampling alternatively of sequential trend. As the records set carries incomplete of 3303 files the place education records carries 2642 files which are 80% of complete facts archives and trying out records consist of 661 archives which are 20% of whole statistics records.

The Simple Moving Average attribute is taken into consideration where this moving average considers the values of previous 'n' No. of days and its mean value. Here, we have taken two Simple Moving Averages of 3 days and 9 days we have considered those two because of their high correlating values.

## III.MACHINE LEARNING MODELS

In this Study, few algorithms of machine Learning were used to train the model using a training dataset which is 80 percent of the total data set and the models were tested against the test dataset which is 20 percent of the total dataset. The few algorithms of Machine Learning used are Linear Regression, Gradient Boosting Regression, Random Forest Regressor, Extra Tree Regressor, and Support Vector Regressor.

Statistical Procedure for determining the correlation between the Dependent target variable to its Independent variables is called Regression. Regression analysis is more of studying the extent of how much the dependent variable changes for a small change in the independent variable.

### *A. Linear Regression Model*

Linear regression is one of the best and most famous Machine Learning algorithms. It is a statistical technique that is used for predictive analysis. Linear regression algorithm indicates a linear relationship between an established (y) and one or extra impartial independent (x) variables, as a result, known as linear regression. Since linear regression indicates the linear relationship, which potential it finds how the fee of the based variable is altering in accordance to the cost of the unbiased variable.

$$Y = A_1*X_1+A_2*X_2+A_3*X_3 \ldots\ldots\ldots +A_N*X_N+ \text{Constant} \quad (1)$$

*Y is known as Dependent Target variable $X_1, X_2, X_3, \ldots X_N$ is known as Independent Variables*
*$A_1, A2, A3 \ldots A_N$ is known as Coefficient of Independent Variables*

### *B. Gradient Boosting Regression Model*

In machine learning, "boosting" refers to the process of integrating several simple models into a single composite model. Since simple models (also known as weak learners) are introduced one at a time, while existing trees in the model remain unchanged, boosting is known as an additive model. The full final model becomes a better predictor as we integrate more and more basic models. The term "gradient" in "gradient boosting" refers to the algorithm's use of gradient descent to reduce loss. Gradient boosting Regression creates a rudimentary model that maps features to the residual. This residual is applied to the current model data, nudges the model towards the correct goal. The overall model prediction increases when this step is repeated many times.

### *C. Support Vector Regression Model*

Support Vector Machine can also be used as a regression tool while retaining all of the algorithm's key characteristics (maximal margin). With a few minor exceptions, the Support Vector Regression (SVR) uses the same rules as the SVM for classification. For starters, since production is a real number, predicting the information at hand becomes extremely difficult. Which has an infinite number of possibilities In the case of regression, a margin of tolerance (epsilon) is set as a rough approximation to the SVM that the problem may have already requested. But, aside from this, there is a more complicated reason: the algorithm is more complicated, so it must be taken into account. The key principle remains the same, though: to minimize the error by personalizing the hyper-plane that maximizes the margin while keeping in mind that some error is tolerable.

### *D. Random Forest Regression Model*

Random forests or random decision forests a square measure Associate in a nursing ensemble learning technique for classification, regression, and alternative tasks that operate by constructing numerous call trees at coaching time and outputting the category that's the mode of the categories (classification) or mean/average prediction (regression) of the individual trees.. Random forest could be a versatile, straightforward to use machine learning algorithmic program that produces, even while not hyper-parameter calibration, an excellent result most of the time. It's additionally one of the foremost used algorithms, owing to

its simplicity and variety (it will be used for each classification and regression task). During this post, we'll learn the way the random forest algorithmic program works, however it differs from alternative algorithms and the way to use it. Another great advantage of the random forest algorithm is that determining the relative significance of each feature on the prediction is very easy. Sklearn has a great tool for measuring the value of a function by looking at how often the tree nodes that use it minimize impurity across the entire forest. After practicing, it calculates this score for each feature automatically.

*Decision Trees and Random Forests: what's the difference?*

There are some variations between a random forest and a group of decision trees. When you give a decision tree a training dataset with features and labels, it will generate a collection of rules that will be used to make predictions. To predict whether an individual would click on an online advertisement, for example, you could collect the following data advertisements that the individual has previously clicked on, as well as some characteristics that characterize his or her decision. When you combine the features and labels in a decision tree, you'll get some laws that will help you predict whether or not the advertising will be clicked. The random forest algorithm, on the other hand, chooses observations and features at random to construct multiple decision trees and then averages the results. Another distinction is that "deep" decision trees can be prone to over fitting. Random forest usually avoids this by generating random subsets of the features and using those subsets to create smaller trees. It then joins the sub-trees together. It's important to remember that this doesn't always work, and it also slows down the calculation depending on how many trees the random forest generates.

*E. Extra Trees Regression Model*

According to the classic top-down technique, the Extra-Trees algorithm creates an ensemble of un-pruned decision or regression trees. Its two key distinctions from other tree-based ensemble approaches are that it breaks nodes at random and that it grows trees using the entire learning sample (rather than a bootstrap replica).The Extra Trees Forest's Decision Trees are all made from the initial training sample. Then, at each test node, each tree is given a random sample of k features from the feature set, from which it must choose the best feature to divide the data according to some mathematical criteria (typically the Gini Index).Multiple de-correlated decision trees are generated from this random sample of features.

This concept is particularly useful in the sense of many problems with a large number of numerical features that differ more or less continuously: it also contributes to increased precision due to smoothing while also greatly reducing computational burdens associated with determining optimal cut-points in standard trees and random forests.

Python is used to implement these machine learning algorithms (Linear Regression, Random Forest Regression, and Gradient Boosting Regression) in this research. The regression methods' prediction accuracy was assessed using Mean Squared Error (MSE), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

## IV. RESULTS AND DISCUSSION

The correlation table (Table-1) results show that there is a good correlation between the gold and silver attributes over the entire study period. Furthermore, gold has a weak correlation with crude oil. The gold price was plotted from 2008 to 2021 (Figure-6). As can be seen, the gold price exhibits a variety of characteristics trends during this time period. When the prediction accuracy of the five models is compared (Table-2), the R squared value for the entire period is very high (more than 95 percent). This means that all five algorithms' models fit the data very well. While Extra Trees Regressor is found to have the best fit over the entire period. Extra Trees Regressor has the lowest values for the entire period in terms of prediction accuracy using Mean Squared Error (MSE), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). While on the other hand, for a longer period of time. Extra Trees Regressor is found to be more effective.

Based on the findings of these analyses, it is possible to conclude that the trend in the gold price has shifted during the time period under consideration for this study. When the trend of the dependent variable changes but there are no significant changes in the trend of the independent variables, the accuracy of various methods may differ. As a result, the model used should be determined by the relationship between the variables in the study.

Below there are graphs of Actual Vs Predicted graphs of all five algorithms. For Linear Regression (Figure-1), for Gradient Boosting Regression (Figure-2), for Support Vector Regression (Figure-3), for Random Forest Regression (Figure-4), for Extra Trees Regression(Figure-7).


Fig.1 Actual Vs Predicted Graph for Linear Regression


Fig.2 Actual Vs Predicted Graph for Gradient Boosting Regression


Fig.3 Actual Vs Predicted Graph for Support Vector Regression


Fig.4 Actual Vs Predicted Graph for Random Forest Regression

| | Gold_Price |
|---|---|
| Gold_Price | 1.000000 |
| Silver_Price | 0.708007 |
| USD_Price | 0.347704 |
| Crudeoil_Price | -0.010718 |
| Copper_Price | 0.440338 |
| SP_Price | 0.440372 |
| Day | -0.008591 |
| Month | 0.009678 |
| Year | 0.475009 |
| Weekday | 0.000999 |
| S_3 | 0.997473 |
| S_9 | 0.994659 |

Fig5. Co relation Between every Attribute used in this Model


Fig.7 Actual Vs Predicted Graph for Extra Trees Regression

TABLE II

| Prediction Accuracy For Five Algorithms | | |
|---|---|---|
| Linear | $R^2$ | 0.9965114173099057 |
| | MAE | 12.27330511535097 |
| | MSE | 264.93153688262447 |
| | RMSE | 16.27671763233068 |
| RandomForest | $R^2$ | 0.9971992912924559 |
| | MAE | 10.600704490382494 |
| | MSE | 214.80888617327417 |
| | RMSE | 14.656359922343412 |
| GradientBoosting | $R^2$ | 0.9964235373672302 |
| | MAE | 12.395481092076112 |
| | MSE | 272.4477901252618 |
| | RMSE | 16.48045374218739 |
| Extra Tree | $R^2$ | 0.9976136827864989 |
| | MAE | 9.788193527779926 |
| | MSE | 178.99477383388637 |
| | RMSE | 13.37889284783634 |
| Support Vector Machine | $R^2$ | 0.9944935018659862 |
| | MAE | 12.81131804800291 |
| | MSE | 384.71432576919494 |
| | RMSE | 19.614135865981833 |

From the results of table (Table–2) we can justify that Extra Tree Regressor has lower RMSE, MSE, MAE compared with other algorithms and has a higher R SQUARE value compared with all other algorithms mentioned above.

## V. CONCLUSION

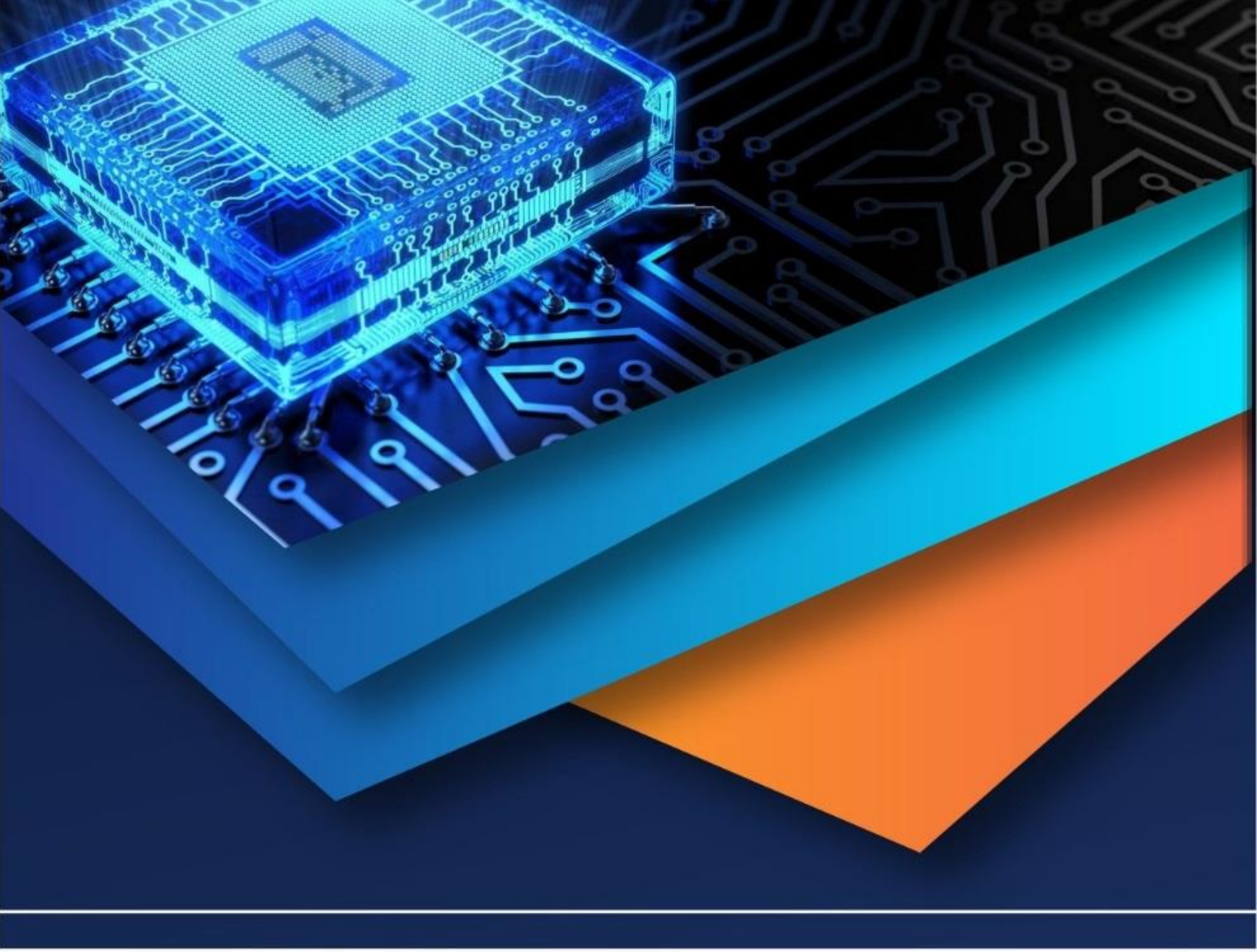
The objective of this project was to determine the correlation between gold price and a number of attributes that influence it, such as crude oil prices, the dollar-rupee exchange rate, silver, the S&P 500, Dow 300,Copper and so on. The data were analyzed using five algorithms of Machine Learning are : Linear regression, Extra Tree Regression, Random forest regression,

support vector regressor, and gradient boosting regression. These models have a good fit with the results. It has been discovered that extra tree regression has higher prediction accuracy. Machine learning algorithms are found to be very useful in this type of study, but the characteristics of the data have an impact on their accuracy. For a better interpretation of the efficiency of these methods, more exploration with such data and different techniques may be performed.

## REFERENCES

[1] Ka, Manjula & .P, Karthikeyan. (2019). Gold Price Prediction using Ensemble based Machine Learning Techniques. 1360-1364. 10.1109/ICOEI.2019.8862557.
[2] J. Jagerson and S. W. Hansen, "All about investing in gold", McGraw-Hill Publishing, 2011.
[3] W. Du and J. Schreger, "Local Currency Sovereign Risk," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 2976788, Dec. 2013.
[4] 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.
[5] L. A. Sjaastad and F. Scacciavillani, "The price of gold and the ex- change rate," J. Int. Money Finance, vol. 15, no.6, pp. 879–897, 1996.
[6] A. Munoz, "Machine Learning and Optimization", Courant Institute of Mathematical Sciences, New York, NY.
[7] Shunrong Shen, Haomiao Jiang, and Tongda Zhang. Stock market forecasting using machine learning algorithms. Stanford University, 2012.
[8] S.F.M. Hussein, M.B.N. Shah, M.R.A. Jalal, S.S. Abdullah, Gold price prediction using radial basis function neural network, in: International Conference on Modeling, Simulation and Applied Optimization, 2011, pp. 1–11.
[9] A. Parisi, F. Parisi, D. Daz, Forecasting gold price changes: Rolling and recursive neural network models, J. Multinatl. Financ. Manage. 18 (5) (2008) 477–487.
[10] V. K. F. B. Rebecca Davis, "Modeling and Forecasting of Gold Prices on Financia Markets," American International Journal of Contemporary Research, pp. Vol 4, No 3, 2014.
[11] 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.
[12] P. V. M. V. P. G. M. P. Sima P Patel, "Gold Market Analyzer using Selection based Algorithm," International Journal of Advanced Engineering Research and Science (IJAERS), vol. 3, no. 4, pp. 55-102, 2016.
[13] Swaminathan, Saishruthi. "Linear regression—detailed view." Medium, Towards Data Science 26 (2018).
[14] 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.
[15] Iftikharul Sami and KhurumNazirJunejo, "Predicting Future Gold Rates using Machine Learning Approach", International Journal of Advanced Computer Science and Applications, 2017
[16] NalinipravaTripathy, "Forecasting Gold Price with Auto Regressive Integrated Moving Average Model", International Journal of Economics and Financial Issues, 2017.
[17] 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.
[18] ZurianiMustaffa and NurulAsyikin Zainal, "A Literature Review On Gold Price Predictive Techniques", 4th International Conference on Software Engineering and Computer Systems (ICSECS), 2015.
[19] Dr. Abhay Kumar Agarwal | Swati Kumari "Gold Price Prediction using Machine Learning" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-4 | Issue-5, August 2020, pp.1448-1456, URL: www.ijtsrd.com/papers/ijtsrd33143.pdf
[20] Jan Ivar Larsen.Predicting stock prices using technicalanalysis and machine learning. NTNU, 2010.
[21] Vatsal H. Shah.Machine learning techniques for stockprediction. NYU, 2007.
[22] Shunrong Shen, Haomiao Jiang, and Tongda Zhang.Stockmarket forecasting using machine learning algorithms.Stanford University, 2012
[23] C. Lawrence, "Why is gold different from other assets? An empirical investigation," Lond. UK World Gold Council., 2003.
[24] S. A. Baker and R. C. Van Tassel, "Forecasting the price of gold: A fundamentalist approach," Atl. Econ. J., vol. 13, no. 4, pp. 43–51, 1985.
[25] S. M. Hammoudeh, Y. Yuan, M. McAleer, and M. A. Thompson, "Precious metals–exchange rate volatility transmissions and hedging strategies," Int. Rev. Econ. Finance, vol. 19, no. 4, pp. 633–647, 2010.
[26] H. Naser, "Can Gold Investments Provide a Good Hedge Against Inflation? An Empirical Analysis," Int. J. Econ. Financ. Issues, vol. 7, no. 1, pp. 470–475, 2017.

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429*
*Volume 9 Issue VI June 2021- Available at www.ijraset.com*



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY