Gold and Diamond Price Prediction Using Enhanced Ensemble Learning

Avinash Chandra Pandey
Jaypee Institute of Information
Technology, Noida
avinash.pandey@jiit.ac.in

Shubhangi Misra
Jaypee Institute of Information
Technology, Noida
misrashubhangi98@gmail.com

Mridul Saxena
Jaypee Institute of Information
Technology, Noida
32k9mridul@gmail.com

***Abstract*—Precious metals like diamond and gold are in high demand due to their monetary rewards. Therefore, various techniques are generally employed to forecast prices of diamonds and precious metals with the aim of fast and accurate results. The prices fluctuate daily making it difficult to predict the next future value. Hence, by examining the pattern of previous prices we can apply regression models for future prediction. This paper aims at forecasting the future prices of precious metals like gold and precious stones like diamond, using ensemble techniques, aiming to get the most accurate result of all. Ensemble models are used for increasing the accuracy of prices. Later, feature selection methods are also used and results are compared.**

***Index Terms*—Ensemble models; Feature Selection; Boosting; Forecasting;**

I. Introduction

Price forecasting is the process of using historical data on a given product to predict the long-term trends of the market. Historically, Gold and Diamond have found its applications in almost every field. Most countries use Gold in the form of currency and for investments also. Banks prefer investing in precious metals due to its unique properties and high demand in the market. As a customer, there is always ambiguity about the correct time to invest, purchase or sell precious items like gold and diamond. It is of utmost significance to buyers and investors when it comes to making maximum profit out of investment and least expense out of a purchase made for the above-mentioned items. There are numerous models and applications currently in the market which are used for predicting the future price of these metals [1]. However, the price of these metals reflect non-linearity and dynamic time-series behavior. Therefore, forecasting prices of these metals is a difficult process. Many researchers have tried predicting future prices of these metals using various machine learning algorithms [2].

Machine learning algorithm are generally classified into two main categories, namely supervised and unsupervised machine learning methods. Supervised machine learning generally outperforms the other methods [3]. Some of the popular machine learning models which have been used for forecasting the future prices of Gold and Diamond are regression models like linear regressor, Random Forest, and ensemble techniques [4]. Moreover, some of the models namely AdaBoost regressor, Lightgbm and, XgBoost regressor are used to enhance the accuracy of basic models [5], [6]. Accuracy of predication also depends on the decency of features therefore, several feature selection methods are used for improving the efficiency and accuracy of the state-of-the-art approaches [7], [8], [9]. However, supervised models generally suffer from over-fitting and under-fitting problem and also shows poor performance for imbalance datasets [10], [11]. Therefore, to avoid these issues, this paper introduces a hybrid model based on the strength of random forest and principal component analysis (PCA). This paper attempts to first equalize, then better the paradigms in order to obtain a more reliable outcome.

The main contribution of this study can be encapsulate as follows.

- First, PCA, recursive feature elimination, and Chi-square test have been used to eliminate the correlation among features and obtain the best subset of features
- Second, ensemble models based the strength of random forest and linear regression are used for predicting the future prices of Gold and Diamond

The remainder of this paper is arranged as follows. Section 2, briefs the related work in field of Gold and Diamond price prediction. Preliminaries are discussed in Section 3 and the proposed method is given in Section 4. Experimental results are presented in Section 5 followed by conclusion in Section 6.

II. Background Study

According to Pradeep [4], forecasting prices of Gold and Diamond is a complicated task due to its non-linearity and fluctuating time series behavior, constrained with many factors like economic, financial etc. This paper examined different ensemble models for determining the future momentum of gold and silver stock price for the upcoming days relative to current day stock price. Machine learning algorithms like linear regression, logistic regression, random forest regression, were used for predicting the gold and silver price. Ling et al. [12] have used different ensemble models by explaining the types of classifiers contained in it and used the model for prediction. Ensemble models use the notion of voting, bagging and stacking techniques to enhance the prediction accuracy. Hafezi and Akhavan [13] analyzed the performance

978-1-7281-3591-5/19/$31.00 © 2019 IEEE

of various ensemble models and found that these models show better classification accuracy than other models. Moreover, these models are also used for prediction of stock price. Classification Ensemble learning strategies, especially boosting and bagging decision trees have improved the prediction accuracy of base learning algorithms[14]. Web[14] investigated that the betterment in accuracy of multi-strategy methods to ensemble learning is due to an rise in the diversification of ensemble members that are formed. From the experimentation it is found that the multi-strategy ensemble learning techniques are more correct than ensemble learning approaches. In multi-strategy ensemble learning technique base learning algorithms are repeatedly used with random training data sets [14]. Multi-strategy ensemble learning techniques reduce test error in the data set by investigating the link in multi-strategy ensemble learning and generation of diversity in ensemble membership.

Many researchers believe that before applying any machine learning algorithm data should be pre-processed and noisy data should be removed. Using data without any pre-processing may lead to inaccurate results. Many factors such as redundant and irrelevant features affect the success of machine learning on a given task. If the irrelevant features are used for training of machine learning models then models may suffer from the under fitting problem [15]. Hence, before training machine learning model feature selection methods can be used to obtain the optimal features from the dataset. Feature selection is the process of identifying and removing irrelevant features from the training data set. Feature selection is used to reduce the dimensionality of data set and to enable regression algorithms to operate faster.

Mahato and Attar [4] proposed a machine learning model to predict the future price of Gold and to identify the correct time of investment to gain more profit. Since, there are still plenty of ignored issues that need to be handled for improving the prediction accuracy. Therefore, a novel forecasting model has been introduced to enhance the forecasting accuracy [16]. Sivalingam et al. [17] used various forecasting techniques including basic forecasting approaches such as linear regression, multiple linear regression, ensemble models for the Gold price prediction. Further, Guha and [18] Bandyopadhyay used Arima model for predicting future prices of Gold. However, Arima model only forecasts the immediate future and not the value of longer time period. Moreover, it also requires long time series and doesn't guarantee perfect forecasting while machine learning models can predict the value of longer time period. This paper has reviewed the different forecasting approaches of gold prices up to 2018.In spite of a lot of methods had been proposed for gold price forecasting, several limitations that have been emphasized makes room for investigation and enhancement of these approaches open as there are still no ideal solutions to predict the gold price. After analyzing the results of all the papers, it can be concluded that stock price predictions and price predictions for precious metals aren't a very complex task. Prediction accuracy can be enhanced by employing appropriate machine learning algorithms and feature selection methods.

III. Preliminaries

A. Feature Selection

Feature Selection method calculates the correlation value between different groups of features and the features with less value is removed. Correlation coefficient value ranges between -1 to 1. The correlation between each attribute was obtained with color, depth, and clarity having a negative relationship with the price attribute. These three attributes were dropped from the dataset. In the proposed model, recursive feature elimination methods including Chi-Square Test and Principal Component Analysis are used for feature selection. Recursive Feature Elimination (RFE) technique removes the weak features recursively and finds the best features which can predict the results. After each iteration, it removes the weak features and ranks them according to their ability to predict. The proposed method employs Linear regression, Chi-Squared test, and Principal Component Analysis (PCA) to select the five best features from dataset. Chi-Squared Test is a statistical test mainly used on categorical data for finding the dependency of two features. Principal Component Analysis is a feature selection method which extracts important features from the set of uncorrelated features from the dataset. PCA constructs new features from the original data set and looks for features that show as many variations as possible resulting in the removal of the redundant properties/features. The new feature would have “maximum variance" and “minimum error".

B. Ensemble Model

Random Forest Regressor, Bagging, Adaboost, Lightgbm, Xgboost are ensemble which has been incorporated in the model for training and prediction [19]. Random Forest algorithm is an ensemble model that uses multiple decision trees for final prediction. In this, each tree is trained on a subset of training data that is arbitrarily selected. This subset is obtained by randomly selecting features. Each tree predicts its own result and then the final result is obtained by majority voting of the trees. The forest it builds is an ensemble of decision Trees which is trained with “Bagging" method. An and Meng [20] used a number of trees for price prediction. Bagging also known as bootstrap aggregating are trained by different regression models and the result is then calculated by using simple majority voting of each output [21]. For the same, a decision tree was used as an input parameter where each node represents a decision and each leaf gives an outcome.

Adaboost Regressor also Known as adaptive boosting uses weighing instances technique to predict the result. Initially, it gives equal weights to each training dataset and predicts the result. In the next iteration, it gives higher weightage to the observations which have been incorrectly classified. This process is iterated till the accuracy is reached or maximum iterations are reached. Lightgbm is a tree-based learning algorithm which makes vertical trees as compared to other models which make horizontal trees. It mainly increases the accuracy. In Gradient Boosting Regressor, it ensembles weaker prediction models and each newly generated model minimizes the loss function using gradient descent. It combines weak models into a single strong model in an iterative manner. XgBoost delivers improved prediction accuracy and is said to handle missing values efficiently. It also gives results 10 times faster than the other gradient boosting algorithms.

IV. Proposed Work

The proposed model combines the strength of ensemble model with different feature selection methods to enhance the classification accuracy. The proposed model uses the following three phases-

1) First, dataset is prepossessed to remove noise and blank entries.
2) Second, Feature selection methods such as Chi-square test, principal component analysis, and recursive feature selection are used for selecting the best features from data.
3) Regression models namely, linear and random forest are used for prediction.

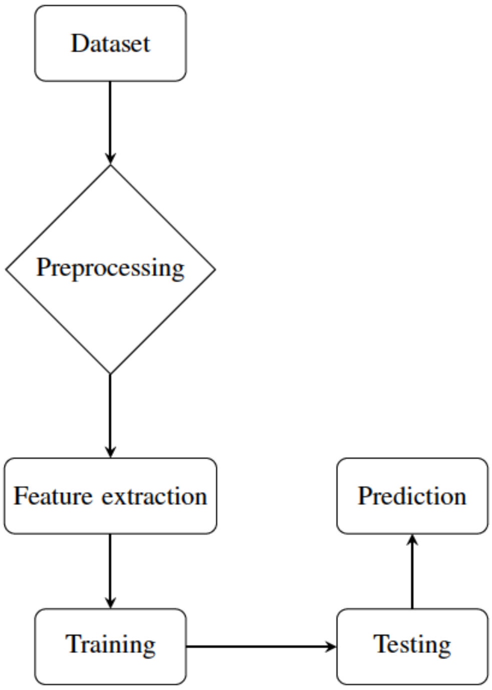

Fig. 1: Flow Chart of Proposed Model

The complete flow of proposed model has been depicted in Fig.1. From the figure it can be perceive that the proposed model employs the feature selection before training of ensemble model. Hence, it reduces the possibility of under-fitting and over-fitting problem. The proposed model uses linear regression to checks the extent to which the dependent and independent variables are related. This method provides results for binary variables where there is one dependent variable and another an independent variable. Hence, working with multivariate variable may not provide the exact expected results. Random Forest regressor solves the problem of accuracy by building decision trees according to the input given by the user and provides the most accurate results. Bagging can also be used for improving the accuracy of the dataset however, it requires other regression models as input for prediction. Later it combines the results of all the models used and predicts the results. Similarly, other ensemble models like Lightgbm, Adaboost improve the accuracy but could not handle missing data properly. Xgboost known as extreme gradient boosting handles missing values efficiently. It also gives results 10 times faster than the other gradient boosting algorithms.

V. Experimental Results:

The performance of proposed model has been evaluated on benchmark datasets taken from Kaggle [22]. It consists of total 10 features: Carat, cut, color, clarity, depth, table, price, x, y, z with a total of 53940 rows. Initially, all the features were used for prediction. Later, three feature selection techniques have been applied in order to reduce the redundancy and size of the dataset. The Chi-square test, RFE, and PCA select 9, 5, and 6 features respectively. To analyze the performance of proposed model k-fold cross validation is used and value of k is chosen to 5.

The performance of proposed model is compared in the terms of mean, best and worst accuracy and depicted in Table I. From the table, it can be analyzed that the random forest model outperforms the linear regression model. Further, to essence of feature selection, proposed models have used along with different feature selection methods namely, Chi-Square, REF, and PCA. The accuracy of proposed ensemble model with different feature selection methods have been depicted in Table in II. From the table, it can be visualize that between the two regression model, linear and random forest the later showed greater accuracy. Random forest regression with Chi-Square feature selection showed the best accuracy with 5 best features. There is a decrease in the accuracy after the application of Recursive Feature Extraction due to the removal of important features from the data set providing the five best features using a linear model. Among the ensemble models, Bagging regressor gave the highest accuracy whereas Ada Boost Regressor gave the least accuracy. The decrease in Ada Boost Regressor model's accuracy is due to overfitting of the training dataset.

TABLE I: Comparison of Regression models in terms of mean, best, and worst accuracy

Regression model	**Mean Accuracy**	**Best Accuracy**	**Worst Accuracy**
Linear	0.8695	0.8745	0.8456
Random Forest	0.9730	0.9821	0.9614

Moreover, the performance of proposed model has also been compared with state-of-the-art approaches including gradient boosting, ada boost, and bagging. The result of all the considered methods has been represented in Table III. If the result

TABLE II: Models with Feature Extraction

Regression Model	Feature Selection Model	Accuracy	Best Accuracy	Worst Accuracy
Linear	Chi-Square Test	0.8663	0.8812	0.8318
	PCA	0.8663	0.8832	0.8414
	RFE	0.8538	0.8794	0.8403
Random Forest	Chi-Square Test	0.9754	0.9803	0.9702
	PCA	0.9754	0.9781	0.9604
	RFE	0.9306	0.9713	0.9342

of Table II, II, and III are compared then it obtained that the performance of existing ensemble models can be enhanced by incorporating the feature selection and preprocessing steps. the training dataset.

TABLE III: Ensemble Models

Regressor model	Accuracy
Gradient Boosting	0.9584
Ada Boost	0.8814
Bagging	0.9658
Proposed Hybrid Model	0.9754

VI. Conclusion

Price prediction of precious metals is always a complex task but machine learning algorithms have made this task easier and more efficient for data analysts and researchers. In this paper, performance of various ensemble models have been compared. Further, various feature selection models like PCA, RFE etc. that play significant roles in determining the efficiency of the algorithm are also employed to reduce the chances of underfitting and enhance the prediction accuracy.

References

[1] M. M. A. Khan, "Forecasting of gold prices (box jenkins approach)," *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, no. 3, pp. 662–670, 2013.

[2] I. ul Sami and K. N. Junejo, "Predicting future gold rates using machine learning approach."

[3] Y. Zhu and C. Zhang, "Gold price prediction based on pca-ga-bp neural network," *Journal of Computer and Communications*, vol. 6, no. 07, p. 22, 2018.

[4] P. K. Mahato and V. Attar, "Prediction of gold and silver stock price using ensemble models," in *Advances in Engineering and Technology Research (ICAETR), 2014 International Conference on*. IEEE, 2014, pp. 1–4.

[5] C.-F. Tsai, Y.-C. Lin, D. C. Yen, and Y.-M. Chen, "Predicting stock returns by classifier ensembles," *Applied Soft Computing*, vol. 11, no. 2, pp. 2452–2459, 2011.

[6] G. I. Webb, "Multiboosting: A technique for combining boosting and wagging," *Machine learning*, vol. 40, no. 2, pp. 159–196, 2000.

[7] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *Machine Learning Proceedings 1994*. Elsevier, 1994, pp. 121–129.

[8] D. Banerjee, A. Ghosal, and I. Mukherjee, "Prediction of gold price movement using discretization procedure," in *Computational Intelligence in Data Mining*. Springer, 2019, pp. 345–356.

[9] A. C. Pandey* and D. S. Rajpoot, "Feature selection method based on grey wolf optimization and simulated annealing," *Recent Patents on Computer Science*, vol. XX, pp. XX–XX, 2019.

[10] S. Santiso, A. Casillas, and A. Pérez, "The class imbalance problem detecting adverse drug reactions in electronic health records," *Health informatics journal*, p. 1460458218799470, 2018.

[11] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial intelligence*, vol. 97, no. 1-2, pp. 245–271, 1997.

[12] X. Ling, W. Deng, C. Gu, H. Zhou, C. Li, and F. Sun, "Model ensemble for click prediction in bing search ads," in *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 2017, pp. 689–698.

[13] R. Hafezi and A. Akhavan, "Forecasting gold price changes: Application of an equipped artificial neural network," *AUT Journal of Modeling and Simulation*, vol. 50, no. 1, pp. 71–82, 2018.

[14] G. I. Webb and Z. Zheng, "Multistrategy ensemble learning: Reducing error by combining ensemble learning techniques," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 8, pp. 980–991, 2004.

[15] M. Karagiannopoulos, D. Anyfantis, S. Kotsiantis, and P. Pintelas, "Feature selection for regression problems," *Proceedings of the 8th Hellenic European Research on Computer Mathematics & its Applications, Athens, Greece*, vol. 2022, 2007.

[16] Z. Ismail, A. Yahya, and A. Shabri, "Forecasting gold prices using multiple linear regression method," *American Journal of Applied Sciences*, vol. 6, no. 8, p. 1509, 2009.

[17] K. C. Sivalingam, S. Mahendran, and S. Natarajan, "Forecasting gold prices based on extreme learning machine," *International Journal of Computers Communications & Control*, vol. 11, no. 3, pp. 372–380, 2016.

[18] B. Guha and G. Bandyopadhyay, "Gold price forecasting using arima model," *Journal of Advanced Management Science Vol*, vol. 4, no. 2, 2016.

[19] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Machine learning*, vol. 40, no. 2, pp. 139–157, 2000.

[20] K. An and J. Meng, "Optimal-weight selection for regressor ensemble," in *Computational Intelligence and Software Engineering, 2009. CiSE 2009. International Conference on*. IEEE, 2009, pp. 1–4.

[21] P. Langley *et al.*, "Selection of relevant features in machine learning," in *Proceedings of the AAAI Fall symposium on relevance*, vol. 184, 1994, pp. 245–271.

[22] Kaggle, "Data Set," https://www.kaggle.com/omdatas/historic-gold-prices/version/1, 2018, [Online; accessed 19-July-2018].