Diamond Price Prediction using Machine Learning

Harshvadan Mihir
Department of Electronics and Communication Engineering
Institute of Technology, Nirma University
Ahmedabad, India
harshvadanmihir@gmail.com

Soham Jani
Department of Electronics and Communication Engineering
Institute of Technology, Nirma University
Ahmedabad, India
soham.jani@outlook.com

Manish I. Patel
Department of Electronics and Communication Engineering
Institute of Technology, Nirma University
Ahmedabad, India
manish.i.patel@nirmauni.ac.in

Ruchi Gajjar
Department of Electronics and Communication Engineering
Institute of Technology, Nirma University
Ahmedabad, India
ruchi.gajjar@nirmauni.ac.in

2021 2nd International Conference on Communication, Computing and Industry 4.0 (C2I4) | 978-1-6654-2013-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/C2I454156.2021.9689412

***Abstract*— Diamond is one of the strongest and the most valuable substances produced naturally as a form of carbon. However, unlike gold and silver, determining the price of a diamond is very complex because many features are to be considered for determining its price. This paper aims to come up with the most efficient algorithm for the price prediction of diamonds. The algorithms such as Linear regression, Support Vector regression, Decision trees, Random Forest regression, K-Neighbors regression, CatBoost regression, Huber regression, Extra tree regression, Passive Aggressive regression, Bayesian Regression and XGBoost Regression are used to train the particular machine learning models on the diamond dataset for the prediction of diamond prices based on various attributes. The comparative analysis of various Machine Learning Regression models is done for the price prediction of any diamond. From the performance parameter values and analysis, it was found that the CatBoost Regression algorithm proved to be the most optimal algorithm having an R2 score of 0.9872 and formidable training and testing accuracies of 98.74% and 98.72% respectively. Hence, the CatBoost algorithm has been implemented for the price prediction of a diamond specimen with the help of the values of attributes extracted from an image of a diamond certificate.**

Keywords— Comparative Analysis, Diamond Price Prediction, Machine Learning, Regression

I. INTRODUCTION

Diamonds fall under the top tier category in jewelry and other expensive things in the world as it has the incredible ability to disperse light. Besides that, diamonds are also used in various machines and other types of equipment for cutting and slicing because it is one of the toughest substances in the world. As it is considered the most rigid element obtained from naturally sourced minerals, diamond is used for cutting metals, glass and other strong materials. In real-world applications, most diamonds are used for industrial applications like surgery equipment, high-quality drill machines, and aerospace and auto sectors. In addition, diamonds are also used in some expensive semiconductor products due to their high heat conductivity.

Generally, precious elements like gold and silver are valued based on their weights, but in the case of diamonds, the factors involved in determining the price of the diamonds are more than just the weight. These factors are carat, cut and many more. Since the price of diamonds is very high, a slight change in these factors would result in a significant variation in the price of the diamond. Like any other product, diamonds also undergo various steps before finally reaching the retail store, in which each step adds some value to the retail price of the diamond. After mining the raw diamond and cutting it, the price of the diamond is determined. In the case of the diamond, where it is used in ornaments, the diamonds undergo various processes to make it look better and transform it into beautiful ornaments which are used as personal adornment by people.

Even though the introductory price of the diamond is determined by the factors like polishing, cutting and mining, other features are also essential to determine the correct price of the diamonds. These features are clarity, carat weight, cut, width, length, color, percentage of depth and table width. Among these, the four main features, cut, clarity, carat weight, and color, are considered the most crucial factors for determining the price of a diamond and these four features are also known as the 4Cs.

We came across various researchers in this field who introduced and exercised an exhaustive list of Machine Learning algorithms to find the most suitable algorithm to train a model used for price prediction of diamonds. As our target variable, i.e., the 'price', is continuous, various Regressive Algorithms have been implemented.

This paper focuses on various machine learning regression algorithms such as Linear Regression, Support Vector Regression, Random Forest Regression, Decision Tree, Huber Regression, Passive Aggressive, Bayesian, Extra Tree, K-neighbor, XGBoost and CatBoost to predict the price of the diamonds using the diamond dataset.

The rest of the paper is organized in the following way: Section II focuses on the literature review explaining the work already done in other papers. Section III explains the proposed methodology for implementing the GUI for price prediction and selection of the appropriate algorithm for the same and describes the dataset along with the types of regression models in machine learning. Section IV discusses the results and analysis of the regression models from the values of performance parameters. Section V concludes the paper while also discussing the future work to be carried out for the following application of diamond price prediction.

II. LITERATURE REVIEW

The comparative analysis of performed in paper [1] for various supervised machine learning models such as Linear

Authorized licensed use limited to: VIT University. Downloaded on September 19,2023 at 09:54:18 UTC from IEEE Xplore. Restrictions apply.

Regression, Decision Tree, Ridge, Regression, Lasso Regression, Elastic Net, Random Forest Model, and so on was used to train the machine learning models for diamond price prediction. The work presented in the paper indicates that the random forest algorithm produces the best results among the other supervised learning algorithms. The paper split dataset as 80% for training and 20% for testing, the Random forest method can achieve the R2 score of 97.93%. The conclusion obtained from this analysis is that random forest is the best algorithm for predicting diamond prices using the particular dataset having an approximate accuracy of 97%.

The comparative analysis for the three algorithms linear regression, neural network, and M5P is carried out for predicting diamonds' price in [2]. The paper emphasizes the process of dimensionality reduction and the problem of correlation between the dataset features. Multiple highly correlated features in the dataset tend to decrease the accuracy of the model. For example, in the diamond dataset, the parameters height, weight and depth (x, y and z) are highly correlated with other. Therefore, the accuracy of the model can be increased by removing some of the correlated features. The accuracy of the M5P model for the diamond dataset is 98.7% on the original dataset and after performing dimensionality reduction, the accuracy of the M5P algorithm increased to 99.03%. From the analysis done in the paper, it can be concluded that the M5P algorithm is the best suitable algorithm for diamond price prediction using the particular dataset using dimensionality reduction.

After reviewing various research works and the performance of various algorithms for the price prediction of diamonds using the same dataset, we identified some of the regression algorithms which would perform well and can be used for diamond price prediction. Therefore, we have compared these methods to find out the suitable algorithm for the purpose of designing a Graphical User Interface (GUI) to scan the image of a diamond certificate and predict the diamond price in INR.

III. PROPOSED METHODOLOGY

In order to implement the GUI application of diamond price prediction, the overall implementation is divided into various parts, including pre-processing of the dataset and training the regression models to find the best regression model for the price prediction depending upon various factors like R2 score, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). In the end, the designing of GUI and feature extraction using Optical Character Recognizer (OCR) is done. Fig 1. represents the flowchart of the proposed methodology and the steps we followed for implementing the whole model.

As shown in the flowchart of Fig. 1. the dataset undergoes pre-processing and optimization to make it efficient for training. The dataset is split into two parts: training data is used to train the models, and testing data is used to test the models and obtain the performance parameters' values. After training the models and using the test data, the performance parameter values are obtained for all the models. The best regression model is identified based on these values and the same model is also used for price prediction. OCR is integrated into the GUI part of the programming section to extract the values of the attributes from the diamond certificate image provided as input by the user. After extraction, these values are supplied to the Regression model, and the price prediction is made in INR.

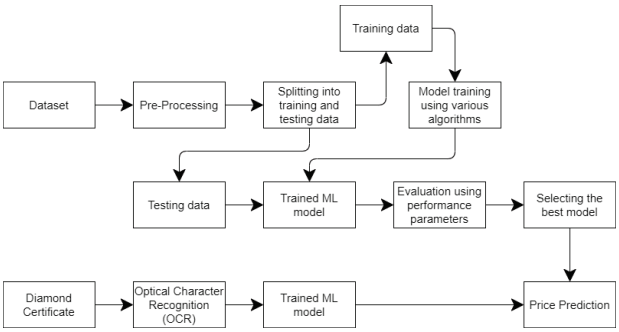

Fig. 1. Flowchart of the implementation process

A. Dataset description and Preprocessing

The diamond dataset consists of 53,940 total unique samples. The dataset attributes are carat, cut, color, clarity, depth, table, price, x, y and z. The attributes of the data are essential for predicting the best-estimated price of diamonds. The carat weight of all the diamonds in the datasets ranges from 0.2Kg to 5.01Kg. The cut is a categorical variable with five unique values: ideal, premium, very good, good, and fair. The color of the diamonds ranges from J to D, where J represents the worst and D represents the best. The clarity attribute is categorical, having eight unique values: IF, VVS1, VVS2, VS1, VS2, SI1, SI2 and I1, where I1 is the worst clarity and IF is the best clarity. Depth, table, price and x, y and z are continuous attributes having various integer and floating-point values. The measurement of the Table, Width and Depth of the diamond is done as shown in Fig. 2.

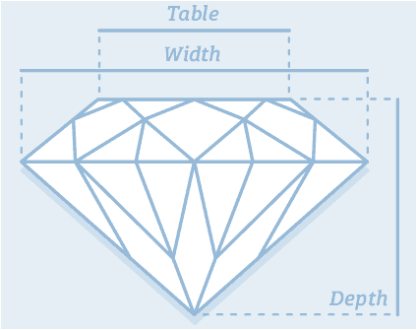

Fig. 2. Measurement of Table, Width and Depth in Diamond [10]

Among these attributes, the four attributes: carat, cut, color and clarity, are the most critical attributes because the price of the diamond mainly depends upon these four values. These four attributes are known as 4Cs of the diamond, a prevalent terminology in the diamond industry for jewelry professionals. Among these 4Cs, clarity, carat, and color have been used for over 2000 years and are the basis of the first grading system for diamonds.

Before training the various machine learning models using the actual dataset, the dataset is modified using various pre-processing techniques like Label Encoding to make the dataset more efficient for the training process.

Label Encoding is performed on the dataset to represent various categorical attributes as numeric values, which is more efficient in training for the machine learning models. The dataset's attributes are converted from alphanumeric form into appropriate numeric labels using Label Encoding.

Authorized licensed use limited to: VIT University. Downloaded on September 19,2023 at 09:54:18 UTC from IEEE Xplore. Restrictions apply.

For n number of categories, the labels 0,1,2,3,… up to n are given to these variables. In the diamond dataset, Label Encoding is applied on the categorical variables such as cut, color and clarity. Furthermore, the dataset is split into two segments: train and test, by keeping 70% samples for training and 30% for testing. Finally, the random state is kept as '42' so that the same random distribution of samples takes place throughout the analysis.

B. Machine Learning Regression Models Types

Since the price of the diamond is a continuous variable in the diamond dataset, we have selected various types of regression techniques to train the model for price prediction of diamond using the dataset and compare the performance from the accuracy of the respective models. The regression techniques are divided into various types based upon the implementation and working. Fig. 3. shows the types of regression techniques and their respective algorithms.

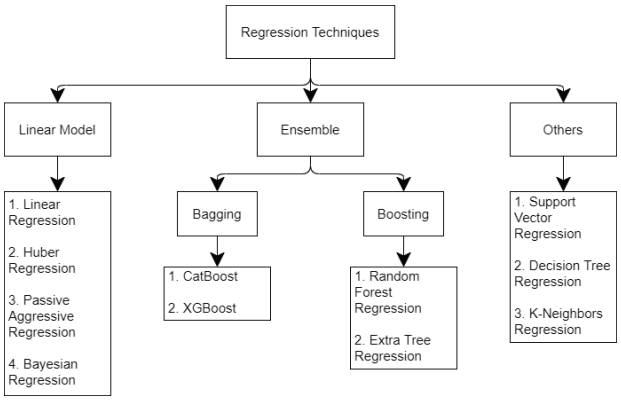

Fig. 3. Types of Regression Techniques

Linear Model regression techniques are simple and commonly used to predict continuous values. This is done by training the model for single or multiple parameters, finding weights and bias values to form a linear equation containing the dependent parameters, and predicting the values based on them. The algorithms such as Linear Regression, Huber Regression, Passive Aggressive Regression, and Bayesian Regression fall into this category.

In Ensemble regression, various regression models are trained individually for the same dataset, and the individual models are combined to achieve better accuracy for prediction. Ensemble regression is divided into two types are Bagging and Boosting. In Bagging, the individual homogeneous regression models are trained in parallel using their respective datasets, which are generated by dividing the actual dataset into smaller parts. Random Forest regression and Extra Tree regression are examples of bagging types of regression. In Boosting, the model's training is done sequentially where each regression model improves the results obtained from the previous model, resolves the errors, and increases stability in each phase. Algorithms such as CatBoost and XGBoost use Boosting regression technique.

The regression algorithms like Support Vector Regression, Decision Trees Regression and K-Neighbors Regression belong to the 'others' category because the implementation and working of these algorithms is different. In all, these are the types of regression techniques used for the price prediction of diamonds in our work.

IV. RESULTS AND ANALYSIS

A. Performance Comparison of Models

For the performance comparison, various regression algorithms were trained to identify the most suitable algorithm depending upon the performance metrics of the models. Figure 4 represents the graph of cross-validation (CV) Error vs K values for the K-Neighbor algorithm. We can analyze the model's error depending upon the value of K which is the number of nearest neighbors and determine its most optimal value for the slightest error. In this case, we can observe in Figure 4 that the error decreases with the increasing value of K, and for K=15, the error is the least. So from this graph, we have selected the value of K as 15 to achieve the best possible results for the K-Neighbor algorithm.

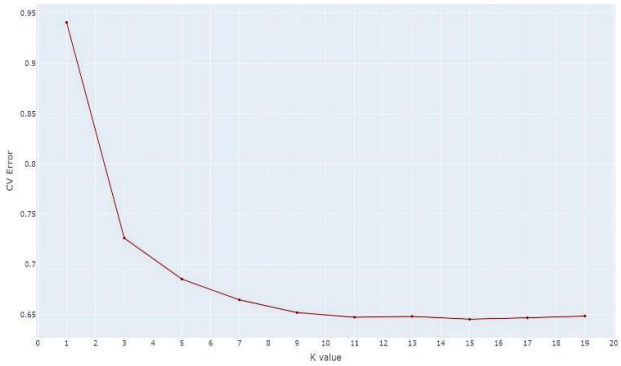

Fig. 4. CV Error vs K value graph for K-Neighbor algorithm

To find the most suitable algorithm, we have trained all the models for the same dataset and compared various parameters of all the models. Fig. 5. represents the R2 score comparison graph containing the R2 score of all the algorithms used for diamond price prediction on the same dataset. From Fig. 5. we can determine that CatBoost has the highest R2 score of 0.9823. The other models, such as XGBoost, K-Neighbor, Random Forest, and Extra Trees, also have close accuracy scores. Since CatBoost has the most efficient performance compared to others, we have used it for the diamond price prediction depending based on the input given by the user.

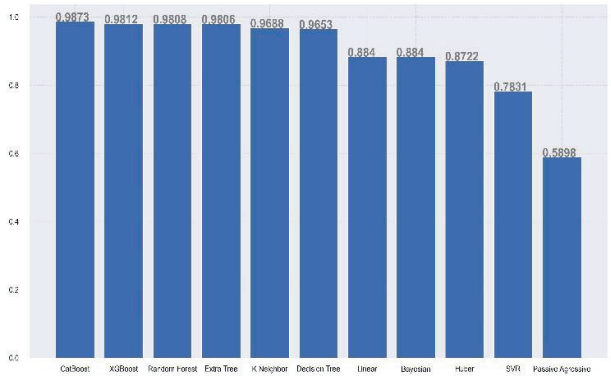

Fig. 5. R2 Score Comparison graph

The performance parameters such as training accuracy, testing accuracy, Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) are also calculated and taken into consideration. Table 1 contains the values of the R2 score, percentage training accuracy, percentage testing accuracy, RMSE and MAE of all the models. The reason

Authorized licensed use limited to: VIT University. Downloaded on September 19,2023 at 09:54:18 UTC from IEEE Xplore. Restrictions apply.

behind selecting CatBoost algorithm is because its R2 score is the highest, the training & testing accuracy are very high, and the values of RMSE and MAE and also comparatively low.

TABLE 1. VALUES OF R2 SCORE, %TRAINING ACCURACY AND %TESTING ACCURACY FOR VARIOUS REGRESSION TECHNIQUES

Algorithm Name	R2 Score	Training Accuracy (in %)	Testing Accuracy (in %)	RMSE	MAE
Linear Regression	0.8839	88.21	88.39	1345.26	864.39
Support Vector Machine	0.7831	77.84	78.31	1839.21	1072.6
Decision Trees	0.9653	99.99	96.53	735.50	365.34
Random Forest Regression	0.9808	99.72	98.08	546.60	272.53
K-Neighbors Regression	0.9687	96.79	96.87	697.83	358.56
CatBoost Regression	0.9872	98.74	98.72	525.81	271.10
Huber Regression	0.8722	86.98	87.22	1411.70	770.26
Extra Tree Regression	0.9805	99.99	98.05	548.98	270.86
Passive Aggressive Regression	0.5897	59.95	58.97	1488.79	963.01
Bayesian Regression	0.8839	88.21	88.39	1345.26	864.45
XGBoost Regression	0.9811	99.13	98.11	542.09	276.84

B. *GUI Implementation*

The GUI implementation component of this work requires a diamond certificate which has to be provided by the user in an image file. A diamond certificate is a document that contains the values of various types of attributes of a given diamond. These diamond certificates are used as a standard in the diamond industries to check the authenticity of the diamond and determine the price of the particular diamond from the values. The image of a typical diamond certificate is shown in Fig. 6.

Fig. 6. Typical diamond certificate

The diamond certificate given in Figure 6 is provided as the input image to the python program. Then, OCR is used to process the image file and extract the information from the image. After the data extraction, the values of various attributes of the particular diamond is stored and applied to the machine learning model. Finally, the machine learning model predicts the price of a particular diamond based on the values of the attributes, and the predicted price is displayed on the screen. Fig. 7. shows the resultant image displayed on the screen after the prediction, and the predicted price is in INR.

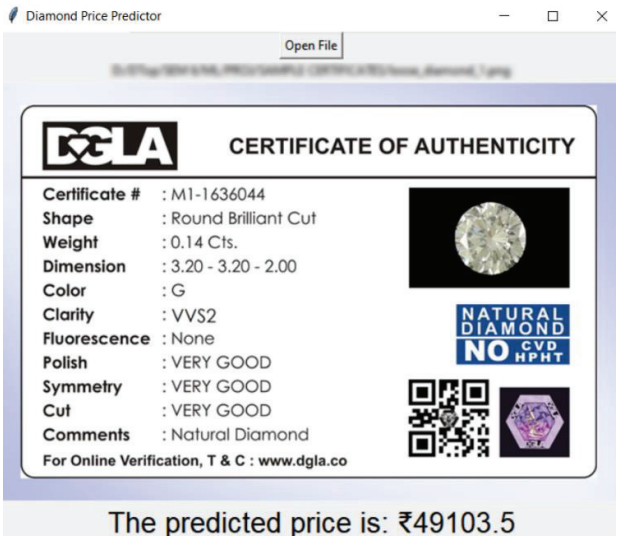

Fig. 7. Output Image with Predicted Price of Diamond

V. CONCLUSION AND FUTURE WORK

After training various regression models and analyzing the results of the algorithms such as Linear Regression, Support Vector Regression, Random Forest Regression, Decision Tree, Huber, Passive Aggressive, Bayesian, Extra Tree, K-neighbor, XGBoost and CatBoost, it can be concluded that CatBoost Regression is the most suitable algorithm for diamond price prediction having the highest R2 score of 0.9872 with comparatively lower RMSE and MAE values with training and testing accuracies of 98.74% and 98.72% respectively. The model is capable to extract the values of different parameters of the diamond from the image file of diamond certificate correctly and thus predicting the price of that particular. Talking about the future prospects, further efforts will be put into solving difficulties like blurry image of certificate, incorrect values of attributes, etc. We would also try to introduce the number of features such as shape, table value, polish, symmetry, etc. to obtain more accurate results.

REFERENCES

[1] G. Sharma, V. Tripathi, M. Mahajan, and A. K. Srivastava, "Comparative analysis of supervised models for diamond price prediction," in *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. IEEE, 2021, pp. 1019–1022.

[2] Marmolejos, José M. Peña. "Implementing Data Mining Methods to Predict Diamond Prices." in 2018 conference of Data Science ICDATA'18.

[3] A. C. Pandey, S. Misra, and M. Saxena, "Gold and diamond price prediction using enhanced ensemble learning," in *2019 Twelfth International Conference on Contemporary Computing (IC3)*. IEEE, 2019, pp. 1–4.

Authorized licensed use limited to: VIT University. Downloaded on September 19,2023 at 09:54:18 UTC from IEEE Xplore. Restrictions apply.

[4] S. Agrawal, “Diamonds dataset”, May 25, 2017, Kaggle. Kaggle datasets repository. [Online]. Available: https://www.kaggle.com/shivam2503/diamonds, Accessed on: Mar. 25, 2021.

[5] H. Drucker, C. J. Burges, L. Kaufman, A. Smola and V. Vapnik, "Support vector regression machines", *Advances in neural information processing systems*, vol. 9, pp. 155-161, 1997.

[6] Wen-jian Wang, "A redundant incremental learning algorithm for SVM," 2008 International Conference on Machine Learning and Cybernetics, 2008, pp. 734-738, doi: 10.1109/ICMLC.2008.4620501.

[7] V. G. S and H. V. S, "Gold Price Prediction and Modelling using Deep Learning Techniques," 2020 IEEE Recent Advances in Intelligent Computational Systems (RAICS), 2020, pp. 28-31, DOI: 10.1109/RAICS51191.2020.9332471.

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

[9] 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.

[10] M. Fried, "Diagram of Diamond Depth and Table" Aug. 8, 2020, *The Diamond Pro*. Montgomery, Alabama, USA: The Diamond Pro, Aug. 8, 2020. [Online]. Available: https://www.diamonds.pro/education/diamond-depth-and-table/, Accessed on: Oct. 8, 2021.

[11] S. Chu, “Pricing the c’s of diamond stones, “Journal of Statistics Education, vol. 9, 2001.

[12] J. M. P. Marmolejos, “Implementing data mining methods to predict diamond prices,” 2018.

[13] T. Doan and J. Kalita, "Selecting Machine Learning Algorithms Using Regression Models," 2015 IEEE International Conference on Data Mining Workshop (ICDMW), 2015, pp. 1498-1505, doi: 10.1109/ICDMW.2015.43.

Authorized licensed use limited to: VIT University. Downloaded on September 19,2023 at 09:54:18 UTC from IEEE Xplore. Restrictions apply.