2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) | 978-1-6654-1451-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/Confluence51648.2021.9377183

Comparative Analysis of Supervised Models for Diamond Price Prediction

Garima Sharma
Computer Science and Engineering
Graphic Era (Deemed to be) University
Dehra Dun, Uttarakhand

Vikas Tripathi
Computer Science and Engineering
Graphic Era (Deemed to be) University
Dehra Dun, Uttarakhand

Manish Mahajan
Computer Science and Engineering
Graphic Era (Deemed to be) University
Dehra Dun, Uttarakhand

Awadhesh Kumar Srivastava
Computer Science and Engineering
KIET Group of Institutions, Delhi NCR,
India

***Abstract*— Diamond is one of the expensive gemstones on the planet that occurs naturally in the form of minerals made of carbons. Precious stones like diamonds are always high in demand due to their monetary rewards. The price of such stones varies according to their features. Given this, we carried out a comparative analysis and implementation of various supervised models in predicting the price of the diamond. In our work, we evaluated eight different supervised models like linear regression, lasso regression, ridge regression, decision tree, random forest, ElasticNet, AdaBoost Regressor, and Gradient-Boosting Regressor and showcases the best suitable model with the more accurate result of all. This paper aims from data preprocessing, finding a correlation between the dataset attributes, training the above-given models, testing their accuracy, analyzing their outcomes, and in turn finding the best of them is the Random Forest Regression Model.**

Keywords*— *Supervised Machine Learning Models; Correlation Matrix, Model Selection

I. Introduction

Diamond, one of the rarest and naturally occurring substance mineral that is composed of carbon. It is the hardest substance which is known today. It has the highest thermal conductivity and also chemically resistant. These are the world's most popular gemstones. It is one of the gemstones on which more money is spent than any other combined gemstone. The diamond gains popularity because it has an optical property. Other factors include is its durability, custom, fashion, and aggressive marketing by diamond producers.[1] any other combined gemstones. The diamond gains popularity because it has an optical property. Other factors include is its durability, custom, fashion, and aggressive marketing by diamond producers.[1]Diamonds have the highest non-metallic luster - known as "adamantine." The high percentage of the light that strikes the surface of diamonds gets reflected due to this luster property. This is a property that gives diamond gemstones their "sparkle". Diamond is extremely hard (ten on Mohs scale), it is often used as an abrasive. Most of the industrial diamonds are used for this purpose. Saw blades, grinding wheels and drill bit are embedded with small particles of a diamond.

The principal motive of this research paper is to introduce supervised machine learning methods to predict the price of diamonds(that is given in US dollars($)). By using a diamond dataset from Kaggle and supervised machine learning methods to detect a precise outcome. Also, a comparison of outcomes of Linear regression, Decision tree, Lasso regression, Random Forest, Ridge regression, ElasticNet, AdaBoost-Regressor, and Gradient-BoostingRegressor model[2] is performed for detecting which one among them performs superior for the tasks. The work is arranged as follows: Section 2 showcases the review of previously done related work. Section 3 presents the experiments conducted for analysis. Section 4 consists of the results obtained from experiments performed in previous section. The conclusion of the paper is in the last section where the accuracy of other models of the same domain is compared.

II. Background Study

Many studies were proposed that have been done to predict the diamond price using different techniques. *José*[2] data mining techniques like M5P, linear regression, and neural networks. M5P data mining model shows that this model possesses a high capacity to predict the diamond price. Using Dimensionality Reduction by high correlation proves that this M5P model is a useful technique to apply on a diamond dataset. A study implemented by *Singfat*[3] using (MLR) Multiple linear regression to find the relationship between diamond price and diamond 4C's.]. The information that is given in the dataset, an MLR data mining model for this dataset is the right and the usual path to inspect the price of a diamond. Various Machine learning algorithms like linear regression, logistic regression, random forest regression, were used for predicting the gold and diamond prices but the gap lies in finding the best using preprocessing and correlation approach.[4]

More often, one would predict the price which is generally represented in US dollars of a stone. Though, the relationship shall not be linear because these heavy stones are expensive than lighter stones. To understand it better, a graphical representation to examine the Kaggle diamond dataset is achieved through scatter plot data visualization[5]. In Fig 1.the relation of carat against price is shown.

978-0-7381-3160-3/21/$31.00 ©2021 IEEE

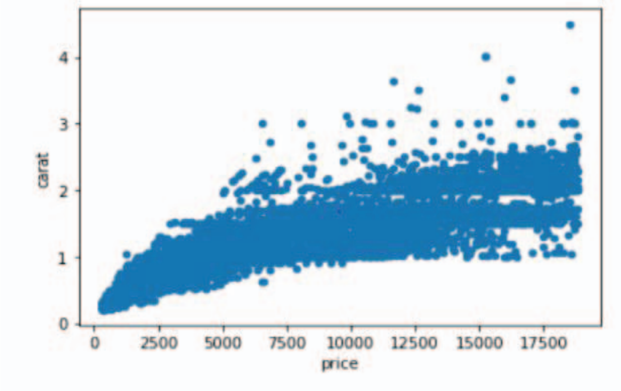


Fig 1. Scatterplot of Carat vs Price

The clear belief is that there is a strong relationship between carat and price, nevertheless, it can be recognized that this trend seems to be fade away, which means heavy Diamonds have higher price volatility.

Nonetheless, in the present study, a broad range of data analyzing methods and some more features are considered. Analyzing the dataset of 53,940 records adds more accuracy and robustness to this research.

III. PROPOSED METHODOLOGY

This Methodology Section is divided into 4 subparts that is Section A, B, C, and D. Section A will tell us about the tool used. In section B we will see that data is acquired. In the section, we will see a correlation between a different variable and in section D we see what can be evaluated with statistics.[5]

We use supervised learning techniques for analyzing the dataset. It provides us with a powerful tool for processing and classifying data using machine language. In supervised learning we use labeled data, which is a dataset that has been classified, to use for the learning algorithm. For the classification of other unlabeled data, this dataset is used as a base to predict data by using machine learning algorithms [7].

Linear Regression is used to determine the extent up to which there is a linear relationship among dependent and one or more independent variables.

In *Lasso regression* prediction error for a quantitative response, a variable is minimized by obtaining a subset of the predictors. The regression coefficients for some variables shrink toward zero because lasso imposes a constraint on model parameters.

Ridge regression is the Variation of Linear Regression. It is a method to generate a parsimonious model when the amount of predictor variables in the single set exceeds the number of observations[8].

Decision trees are mostly used to identify a strategy that is most anticipated to reach a particular goal. It is mainly used for analyzing the decision and also one of the popular tools to use in machine learning.

Random Forest consist of a group of decision trees. The final result for the random forest is found by aggregation of the decision tree result. This model is powerful because they limit the overfitting by not increasing the error substantially.[9]

ElasticNet uses penalties from both ridge and lasso regression for regularizing regression models. This technique combines both ridge and lasso methods to learn from the shortcomings to improve the regularization of statistical models.

AdaboostRegressor is a meta-estimator. It begins by fitting the original dataset with a regressor and then fits the same dataset with additional copies of the regressor but for the current prediction, the weight of instances is adjusted according to error.

GradientBoostingRegressor builds an additive model in forwarding stage-wise fashion. For arbitrary differentiable loss functions, it allows optimization. A regression tree at each stage is fit on the negative gradient of a given loss function.[10]

A. Tools Used

For analyzing the dataset Python 3 is used. It is an open source, interpreted, and high-level language. It gives a great approach to object-oriented programming. For a data scientist, it is one of the best languages to use for data science projects or applications. One of the main reasons for choosing Python as it is widely used in the research and by scientific communities' reason of being its easy to use and easily adaptable language syntax.

B. Data Acquisition

For analyzing the Kaggle repository is the data source for thousands of datasets. It is an online community for machine learning practitioners and data scientists and also it is a vigorous, attested, and adequate resource for investigating different data sources. Users can find and publish different datasets on Kaggle. They can explore datasets and build models in a web-based data-science environment[12]. The diamond dataset from Kaggle will provide the essential features of the Diamond Dataset. The features of diamonds are shown in Table 1 and Fig 2. The four feature of a diamond is represented to understand the term easily.

TABLE I DATA-FEATURES

Name	Values	Data Type
Price(USD)	326….18823	Numerical
Carat weight	0.2……5.01	Numerical
Cut quality	Fair, Good, Very Good, Premium, Ideal	Categorical
Color	J, I,H,G,F,E,D	Categorical
Clarity	I1,SI2,SI1,VS2,VS1, VVS2,VVS1,IF	Categorical
X(length in mm)	0….10.74	Numerical

Y(length in mm)	0….58.9	Numerical
Z(length in mm)	0…...31.8	Numerical
Depth percentage	z/mean(x,y)= 2*z/(x+y)(43…79)	Numerical
Table width	43…95	Numerical

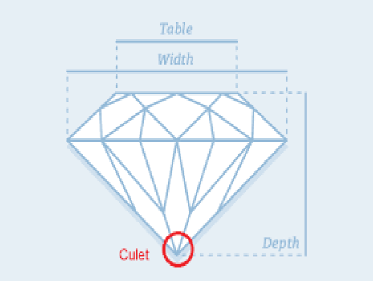


Fig 2. Diamond culet, width, depth, and table feature

C. *Correlation Matrix*

The correlation matrix is represented by the Table 01 shows the correlation between distinct variables. The correlation between the two variables is shown in each cell. For representing this correlation, heat map is used to analyze these relationships better. Fig 3. shows the correlation between distinct fields.

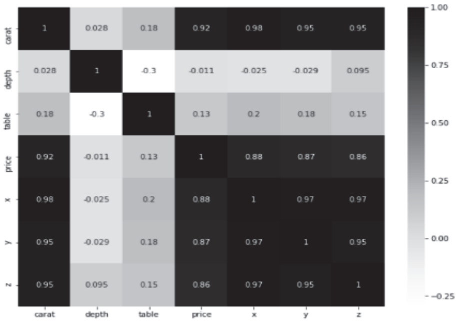


Fig 3. High Correlation Heat Map

From the above Heat map[13], we can find x, y, and z are correlated with price, and the price of diamond and carat(weight) of a diamond are highly correlated. Hence we can readily say that carat is one of the main features to predict the price of a diamond.

D. *Evaluating the Statistics*

To start evaluating stats we will split the dataset into Train set (80%) and Test set (20%). The test set allows our model to make predictions on values that it has never seen before. But taking random samples from our dataset can introduce significant ***sampling bias.*** Therefore, to avoid sampling bias, the data will be divided into different homogenous subgroups called strata. This is called ***Stratified Sampling***. The carat is the most important parameter to predict the price of the diamonds we will use it for Stratified sampling. We have used root mean square error(RMSE) to manage undesirable large errors in this huge dataset as mean absolute error will not efficiently work for this huge wild dataset. For particularly undesirable large errors, RMSE is most useful[14]. We will first find the RMSE and Cross Validation Scores(CV_scores) to check the performance. The function will plot a graph to show how well our algorithm has predicted the data.

CV_scores(Cross-validation Scores) starts by shuffling the data and splitting it into *k* folds. Then the *k* models are fit on k−1/k of the data and evaluated on 1/k of the data. The results that are obtained from each evaluation are averaged together to compute a final score, after finding the final score the final model is fit on the entire dataset to begin the processing[15]. The K Folds are Evaluated as shown in Fig 5.

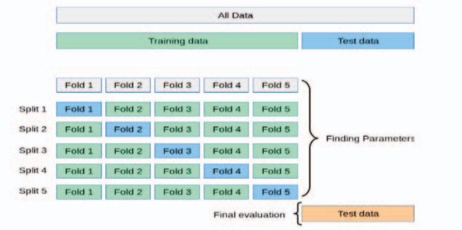
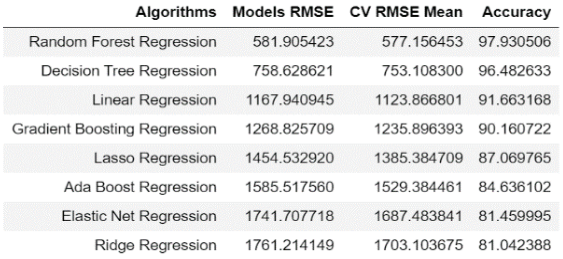


Fig 4. K Folds Evaluation

IV. Result Analysis

The accuracy alone with RMSE and CV score of linear regression, random forest, lasso regression, decision tree, ridge regression, ElasticNet, AdaBoostRegressor, and GradientBoostingRegressor, is shown in Fig 5.



Algorithms	Models RMSE	CV RMSE Mean	Accuracy
Random Forest Regression	581.905423	577.156453	97.930506
Decision Tree Regression	758.628621	753.108300	96.482633
Linear Regression	1167.940945	1123.866801	91.663168
Gradient Boosting Regression	1268.825709	1235.896393	90.160722
Lasso Regression	1454.532920	1385.384709	87.069765
Ada Boost Regression	1585.517560	1529.384461	84.636102
Elastic Net Regression	1741.707718	1687.483841	81.459995
Ridge Regression	1761.214149	1703.103675	81.042388

Fig 5. Model Accuracies

The *Random Forest* produced a better overall result than any other supervised learning algorithms. Orderly to acquire these results, we have to look in-depth into data. This led to the finding of outliers which is small in number that was harming the performance of several models, by eliminating these outliers, a decline in errors can be seen.

V. Comparative Analysis

The Random Forest Regression and Decision tree models show remarkably better accuracy than any model concerning root mean squared error, nevertheless, the Random Forest Regression model had top performance overall and with the highest accuracy and with the least errors. The comparison of the seven models used is shown in Fig 6.

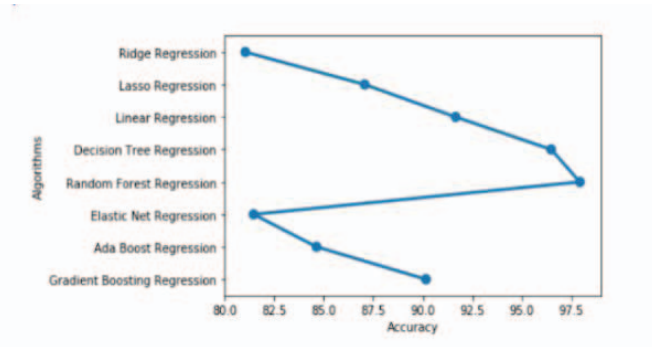


Fig 6. Accuracy Comparison of Models

VI. Conclusion and Future work

By experimenting and analyzing it is feasible to conclude that the Supervised learning methods such as AdaBoostRegressor, GradientBoosting-Regressor, linear regression, decision tree, ElasticNet, ridge regression, lasso regression, and random forest model can be utilized to evaluate diamond prices. Random Forest Regression model shows 97% accuracy It gives this much high accuracy because it possesses a high capacity to determine continuous numerical values. Future work shall include Unsupervised models orderly to extend the accuracy of predictions of a diamond dataset and also its robustness.

References

[1] 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, 2011, pp 2452–2459.

[2] José M., “Implementing data mining methods to predict Diamond prices” Peña Marmolejos Graduate School of Arts and Sciences, Fordham University, Int'l Conf. Data Science-ICDATA'18. https://csce.ucmss.com/cr/books/2018/LFS/CSREA2018/ ICD8070.pdf

[3] A. C. Pandey, S. Misra and M. Saxena, "Gold and Diamond Price Prediction Using Enhanced Ensemble Learning," 2019 Twelfth International Conference on Contemporary Computing (IC3), Noida, India, 2019, doi: 10.1109/IC3.2019.8844910, pp. 1-4.

[4] Singfat the Chu, “Pricing the C’s of diamond stones”, National University of Singapore, Journal of Statistics Education Volume. https://www.tandfonline.com/doi/full/10.1080/10691898.2001.11910 659

[5] Diamond-The most popular gemstone [online]- https://geology.com/minerals/diamond.shtml

[6] Waad Alsuraihi, Ekram Al-hazmi, Kholoud Bawazeer, Hanan AlGhamdi, “Machine Learning Algorithms for Diamond Price Prediction”, Publication: IVSP '20: Proceedings of the 2020 2nd International Conference on Image, Video and Signal Processing, March 2020.

[7] Alexandru Niculescu-Mizil, Rich Caruana, “Predicting good probabilities with supervise Learning”, Publication: learning August 2005.

[8] Supervised Machine Learning Models with sci-kit learn [online]- https://scikitlearn.org/stable/supervised_learning.html

[9] Linear, Ridge and Lasso regression with sci-kit learn [online]- https://www.pluralsight.com/guides/linear-lasso-ridge-regression-scikit-learn

[10] Decision tree and Random Forest regression [online]- https://towardsdatascience.com/decision-trees-and-random-forests.

[11] GradientBoostingRegressor with sci-kit learn [online]- https://scikitlearn.org/stable/modules/ generated/sklearn.ensemble. Gradient Boosting Regressor.html

[12] Datasets - Diamonds dataset, Kaggle datasets repository [online] https://www.kaggle.com/shivam2503/diamonds

[13] Tovi Grossman, George Fitzmaurice, “Patina: Dynamic heatmaps for visualizing application usage”, Publication: CHI April 2013. https://dl.acm.org/doi/abs/10.1145/2470654.2466442

[14] Chai T. “Root mean Square Error (RMSE) or Mean absolute error(MAE)”, (NOAA Air Resources Laboratory (ARL), NOAA Center for Weather and Climate Prediction, 5830 University Research Court, College Park, MD 20740, USA; https://ui.adsabs.harvard.edu/abs/2014GMDD....7.1525C/abstract

[15] Brownlee, J. (2018, May 22). “A Gentle Introduction to k-fold Cross-Validation”.Online – “https://machinelearningmastery.com/k-fold-cross-validation/ - Retrieved 21 October 2019,