University of the West of England

Machine Learning and Predictive Analytics

UFCFMJ-15-M

Coursework Assignment

Module Leader and Tutor:

Dr. Hisham Ihshaish

Produced and submitted by:

Name : Parashkumar Shah

Student ID: 21051422

Table of Contents

The below assignment is **2195** words long.

Diamond Price Prediction

using Machine Learning Techniques

1. Introduction

Diamonds hold impression of love-bearing crystals over the centuries, moreover diamond being one of the most expensive objects on earth because of its rarity and hardness makes it monetary valuable for investments, industrial uses, and jewelry studding. The global diamond market with overall valuation of $87.31 billion in 2018 has potential 3% CAGR (Anon., 2019). A rule of thumb states that a person should spend 3 months salary on diamond engagement ring, another belief suggests bigger the diamond more expensive it is (Lex). Well, it is a partial truth because the valuation of diamond depends upon many other factors commonly known as 4 C's – Carat, Color, Clarity and Cut (Slisha, 2020) introduced by Gemological Institute of America. Thus, it becomes very crucial for buyers and sellers to know the actual price of diamonds but merely, knowing these factors does not equip any person with ability to appraise diamonds as it needs months to utterly understand these factors and years to master the art.

Thanks to years old industry which constantly emphasis on upgraded smart technologies (Sakshi, 2019) which facilitates generation and collection of data of selling prices and several other aspects in very efficient manner. Vast data could be analyzed and transformed into knowledge that could help predict the prices using Machine learning [ML] techniques that enables identify hidden patterns and analyze the problems in more quick and efficient manner (Katarzyna, 2021), which can be statistically proven, which humans will be less capable of doing by themselves. The efficiency of ML improves for price prediction as it involves the study of price volatility, analyzing multiple data sets, effects of several dependent variables influencing the price.

This project attempts in accurate prediction of diamond price using ML algorithms and using statistical measures to prove their performance which will help industry and future buyers.

Jupyter Notebook with codes to support the research can be found on Gitlab: here

2. Data Exploration

The data set used for research has been taken from Kaggle, Diamonds (Shivam, 2017), which is originally a snapshot pricelist of Tiffany & Co. from 2017. It contains 11 attributes and 53940 records. Attributes include:

- Weight of diamond in carat
- Categorial information such as color, clarity and cut
- Information on parameters like Table %, Total Depth %, Diameter Length (mm), Diameter width(mm) and Total depth (mm)
- Diamond price in US Dollars

Below are the images for better understanding of the characteristics discussed above.

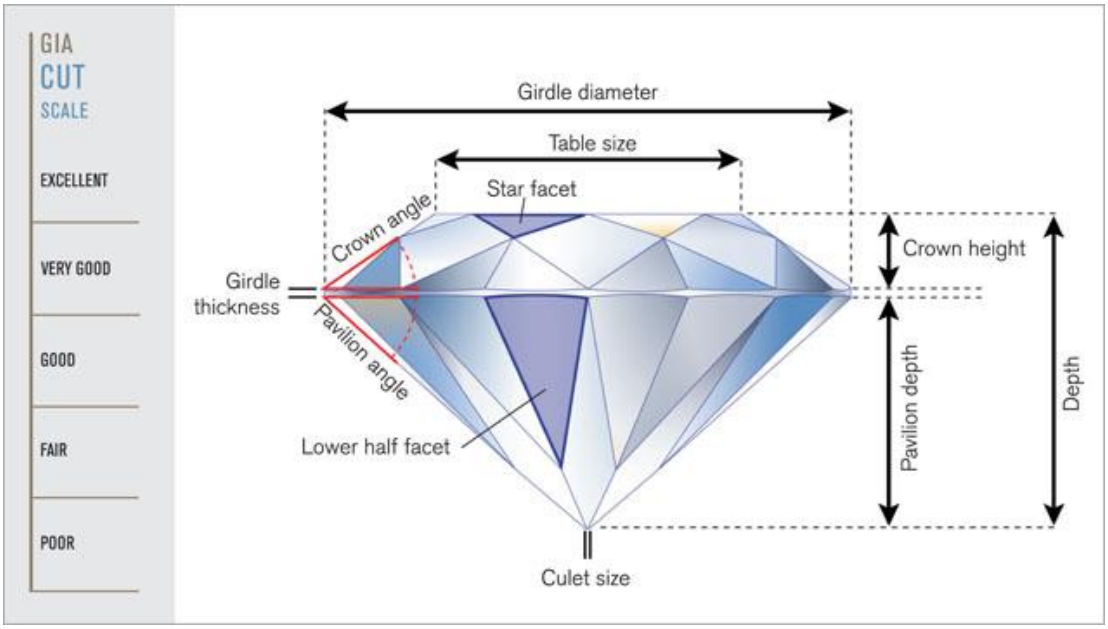


Figure 1: The parameters of Polished diamond. Source: yourdiamondguru.com

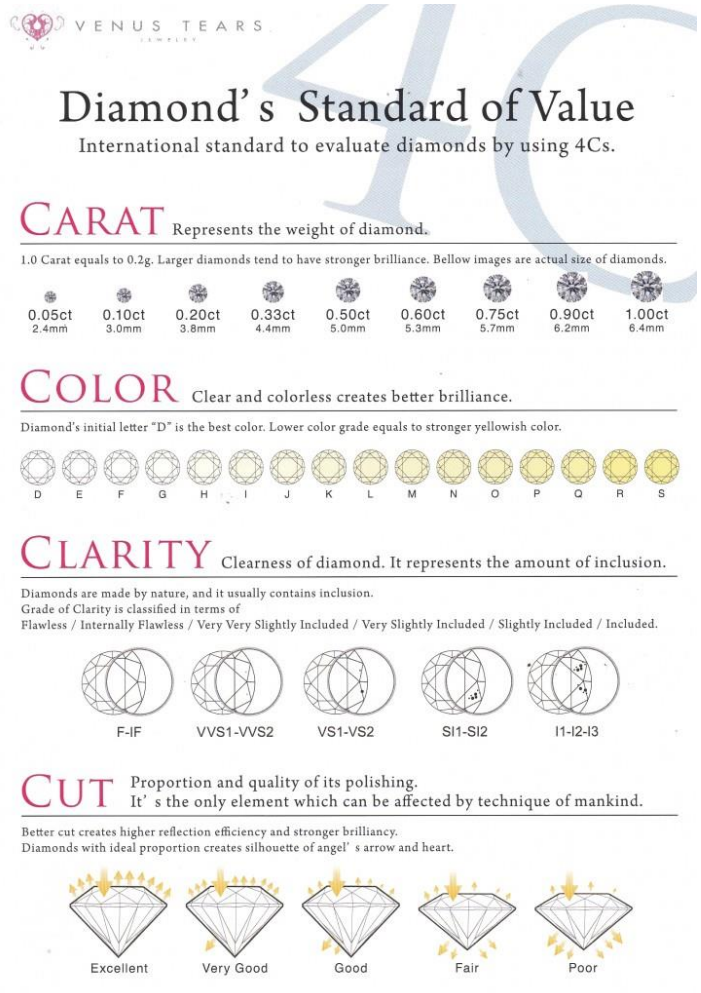


Figure 2: Detailed classification of 4 C. Source: [Venus-Tears]

Explanation:

The broadest part of the diamond is the diameter, which is measured in length and width. Apart from Round Brilliant diamonds, all other shapes have a big difference in length and width.

Table represents the top portion of the diamond.

Depth is the total height of diamond.

Carat is a measurement unit which is equal to 0.2 grams. It can vary from 0.01 to 100 carats.

Color ideally ranges from D to Z. Where D is colorless, and z is colored diamond. The colored diamonds are usually yellow but can be green and brown as well.
Note: For the pricing we are not considering Fancy colored Diamond.

The impurities in a diamond are clarity, Fl (Flowless) is the purest form without any impurities and I1, I2 and I3 are with maximum impurities.

Cut is the shape of diamond, mostly categorized as Round and Fancy. Further divided into Proportions, Polish and Symmetry.

Note: These characteristics are graded by laboratories like GIA, IGI, etc. and can differ because of difference in grading standards.

3. Data Cleaning

For ensuing that data was cleaned, I started with finding missing values as any value missing can bias the ML model (Nasima, 2021), luckily no missing values found. Another step was detecting anomalies by using the summary statistics of each variable, which showed x (length), y(width) and z(depth) columns shows 0 minimum, which cannot be technically possible, thus those records were removed.

	Unnamed: 0	carat	depth	table	price	x	y	z
count	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000
mean	26970.500000	0.797940	61.749405	57.457184	3932.799722	5.731157	5.734526	3.538734
std	15571.281097	0.474011	1.432621	2.234491	3989.439738	1.121761	1.142135	0.705699
min	1.000000	0.200000	43.000000	43.000000	326.000000	0.000000	0.000000	0.000000
25%	13485.750000	0.400000	61.000000	56.000000	950.000000	4.710000	4.720000	2.910000
50%	26970.500000	0.700000	61.800000	57.000000	2401.000000	5.700000	5.710000	3.530000
75%	40455.250000	1.040000	62.500000	59.000000	5324.250000	6.540000	6.540000	4.040000
max	53940.000000	5.010000	79.000000	95.000000	18823.000000	10.740000	58.900000	31.800000

Figure 3: Summary Statistics

Unwanted columns like sequence column were removed and additionally to decrease the variability and enhance statical power, removing the outliers becomes particularly important (statisticsbyJim). Finally, after all cleaning 53910 records remained for further analysis.

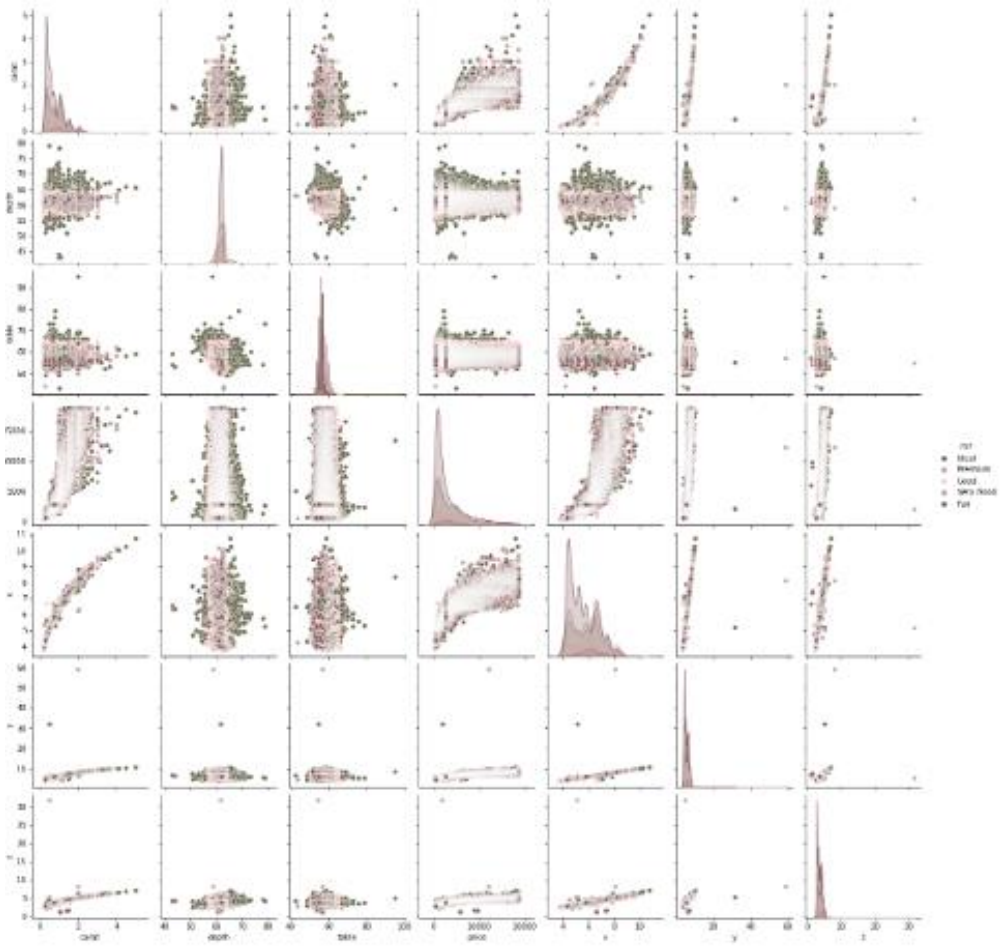


Figure 4: Scatterplot Matrix

4. Exploratory Data Analysis [EDA]

EDA is a crucial step before training models as it helps understand different patterns visually. Its features can determine whether to use supervised or unsupervised approach in ML (Shrimal, 2018). I undertook the EDA on all variables according to their data type to know the distribution of each variable and knowing the dependent and independent variables. As the project demands for prediction of price, I decided to choose price as dependent variable.

The next step was finding the relation of independent variables to price. The price is highest in Premium cuts. Also, all the clarities showed right skewed distribution.

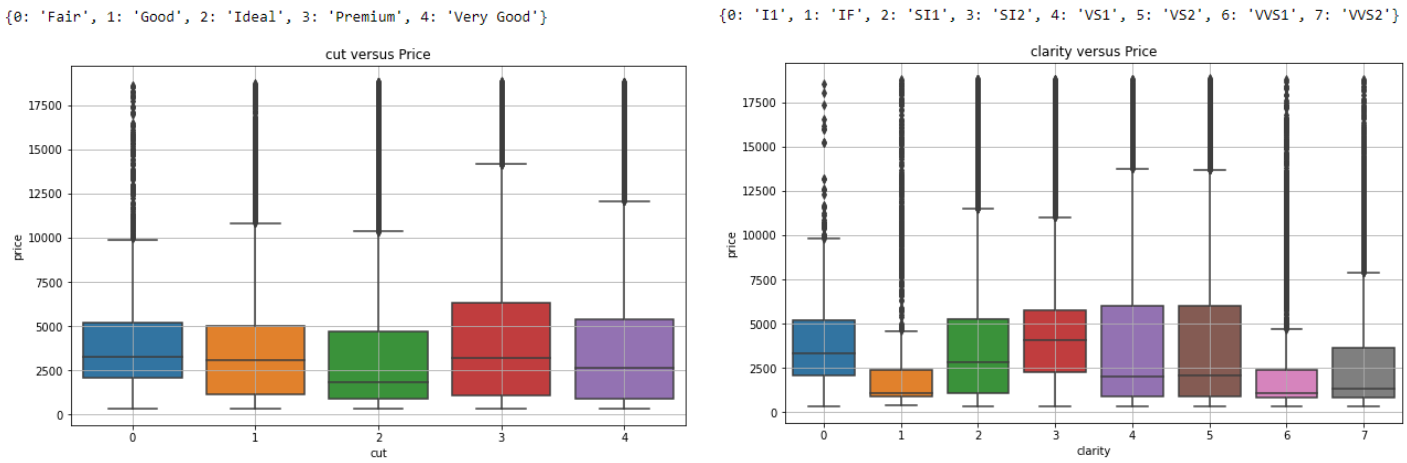


Figure 5: Categorial Box plot to understand the relation with price

While comparing carat with price it showed they have a positive linear correlation. And the distribution of the carat is near to normal. And the Majority of the data is in centre.

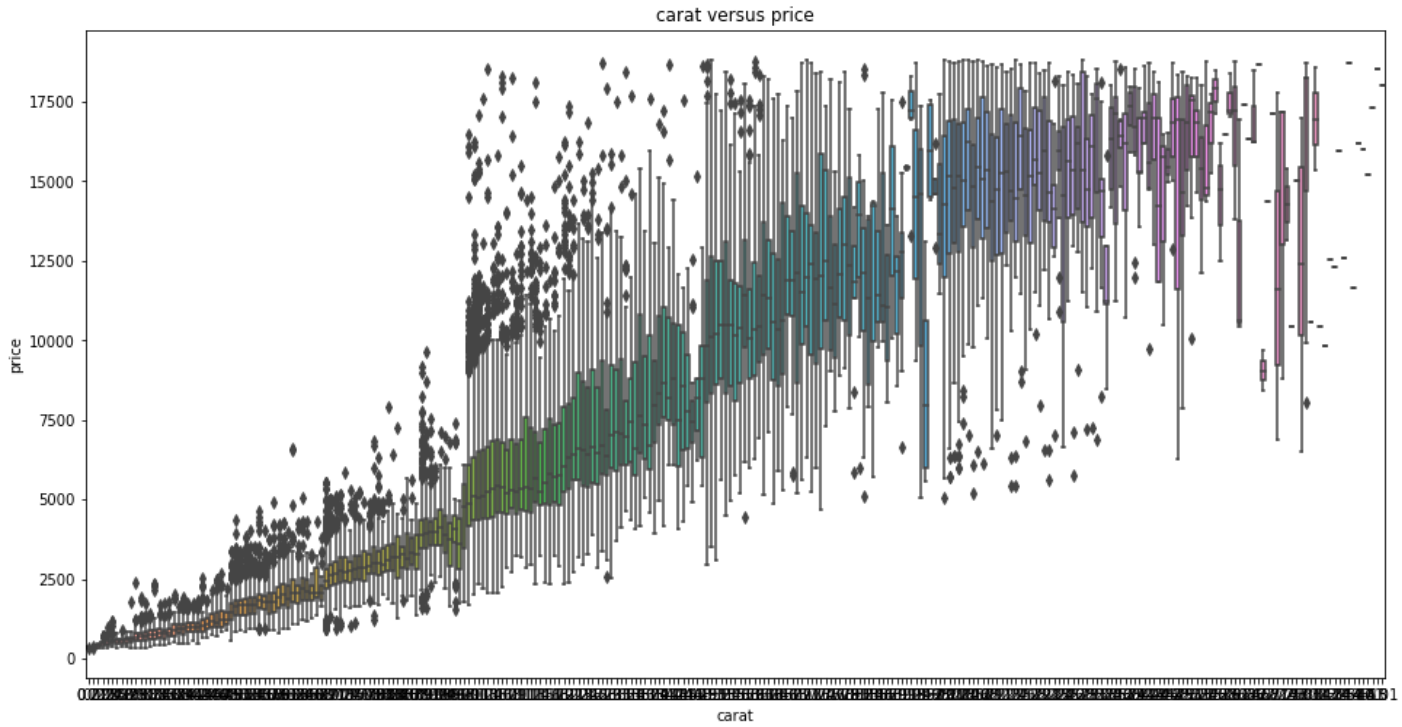


Figure 6: Relation of carat (weight) and price.

5. Feature Selection

I used correlation matrix to understand the relations between dependent variables as it becomes important to know that the dependent variables should not be highly correlated to avoid multicollinearity (Zach, 2020) and found the x, y, and z columns are highly correlated to carat. Its an obvious guess that with increase in size the diameter and height will increase. There was no point of keeping similar predictors, thus I removed those 3 columns (Ananya19b, 2021).

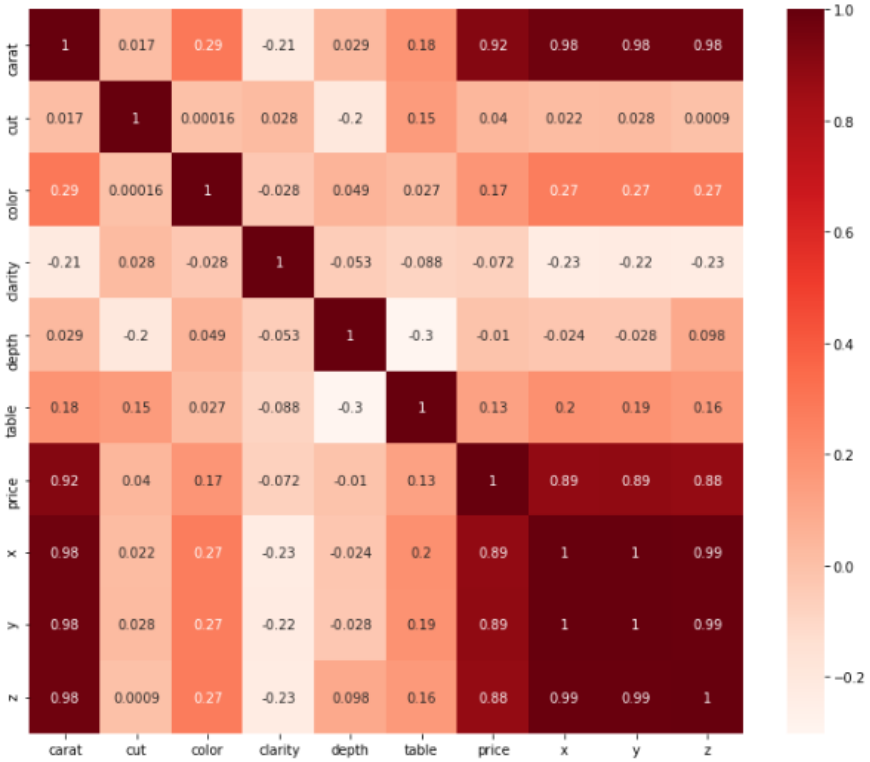


Figure 7: Correlation Matrix

6. Feature Scaling and normalization

I found the distribution of variables was not normal. “Price” and “Carat” were right skewed, to reduce the kurtosis, I transformed that variable with log to base 10 which helped in getting nearly normal distribution (Changyoung, et al., 2014). Log transformation log did not perform well for left skewed variables “cut” and “clarity,” thus it was normalized by using power transformation (Georgios, 2020). I scaled down other variables using StandardScaler.

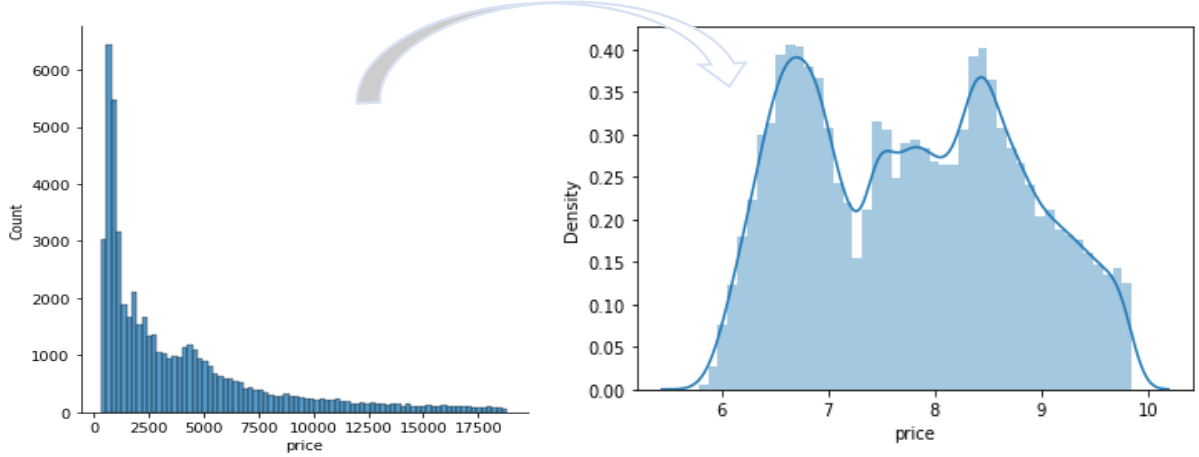


Figure 8: Log transformation of dependent variable “price”

7. Further Analysis

Finally, 6 independent variables were chosen with a dependent variable “price” for further training models. From figure 7 it was observed that “carat” has a high correlation with “price”, Choosing only a particular will not take account for the significance of other variables. For determining the price of a diamond, all the factors play significant role. Thus, more than one variable has been chosen to undertake **multivariate analysis**. As it performs better with larger datasets which will make computational cost slightly higher, but it will account for the proportion of influence of each variable (Harsha, 2020). The final chosen variables are as follows.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 53910 entries, 0 to 53939
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   log_carat   53910 non-null  float64
 1   pt_cut      53910 non-null  float64
 2   pt_clarity  53910 non-null  float64
 3   ss_color    53910 non-null  float64
 4   ss_depth    53910 non-null  float64
 5   ss_table    53910 non-null  float64
 6   log_price   53910 non-null  float64
dtypes: float64(7)
memory usage: 5.3 MB
```

Figure 9: Final Columns

8. Training ML Algorithms

I have split the independent variables and target variable into test and training sets with the ratio of 2:8. Splitting helps evaluate the model with true instances, which otherwise may become biased (Aylin, 2021). Being the ground truth is provided in the dataset we have both input and output values to train model, Thus I have used supervised approach of ML. Moreover, our prediction involves continuous variables thus Regression techniques are used. 3 algorithms and 3 ensembles for training the model to predict diamond price which are explained below. 10-fold cross validation had been used in all algorithms for model evaluation.

8.1 Multiple linear regression [MLR]

$$yi = \beta 0 + \beta 1xi1 + \beta 2xi2 + \ldots + \beta pxip + \epsilon$$

The base assumption of this model includes linear relationship between dependent and independent variables, dependent variables are not highly correlated, the residual variance are constant, and observations are independent. Meeting all assumptions after the cleaning process motivated me to use this model for my dataset; also, this model is simple to apply and understand. It also accurate prediction when there are complex relations (Andriy, 2022). Comparatively it was performing lower, thus rejected.

8.2 Decision tree regression [DTR]

The benefit of DTR include minimal preprocessing, it is amazingly easy in interpretation, it has very less influence of outliers and missing values, both categorial and numerical variables can be managed at same time. However, some disadvantages are It is computationally expensive and required more time in training model but gives better accuracy, likely to overfit if fully grown (Pavan, 2020). As I had both categorial and numerical variables I used DTR on my dataset.

8.3 K Nearest Neighbor regression [KNNR]

It is the simplest method yet effective and easy to interpret. As this method works on the concept of finding the distance amongst the neighbors (data points) it has been implemented assuming a better accuracy. Firstly, with the help of Grid search, optimum k value has been found and then the accuracy was evaluated based on k.

Ensembles

An ensemble works on concept of combining predictions from 2 or more models. Its main advantages are better performance and reduction in dispersion of predictions of models (Jason, 2020).

8.4 Ridge regression [RR]

It creates parsimonious model, mostly used when number of predictors increases number of observations or when data set has multi collinearity. As the collinearity in our data has been removed in pre-processing, RR showed same results as linear regression, thus rejected.

8.5 Lasso regression [LR]

It used the shrinkage technique. It is also known as penalized technique, it uses subset of variables and helps to increase model interpretation (Ajay, 2021). LR has further Normalized the dataset, due to the dual effect the performance was degraded, and thus it was rejected.

8.6 Random Forests Regressor [RFR]

It is ensemble of Decision Tree. The average of all predictions made by several individual trees created from different samples at each node is taken. It works on concept of a decision of 100 average men is better than 1 expert (Hisham, 2022). It gives highly accurate results and has a drawback of very heavy computational cost.

8.7 XG Boost regression [XGBR]

If efficiently implements gradient boosting. By combining the estimates of weaker models, it attempts to accurately predict a target variable. It uses regularization parameter to reduce prediction sensitivity. It performs well on tabular data and produces smaller trees. It cannot be easily interpreted.

9. Evaluation of model performance

It is a crucial part in process of model development (Divya, 2019) which helps to find best model and helps verify its statical significance.

9.1 Train Test accuracy

After the models were trained, train and test accuracy were applied for initial comparison between models.

	Model	Train_Accuracy	Test_Accuracy
0	Linear regression	0.96	0.96
1	Ridge Regression	0.96	0.96
2	Lasso Regression	0.90	0.91
3	Decission tree Regressor	1.00	0.98
4	Random Forests Regressor	1.00	0.99
5	XG Boost Regressor	0.94	0.94
6	K nearest Neighbor Regressor	0.96	0.94

Figure 10: Train-Test accuracy comparison

As per the results, DTR and RFR seemed to be performing best predictions, which is about 100% on train data and 98% and 99% respectively on test data. It is important to note that RFR is an ensemble of DTR which is successfully delivering 1% better result than the later. Additional tests were further carried out to determine model with statistical significance.

9.2 Mean Squared Error [MSE]

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

General formula for mean squared error.

MSE is a metric that measures the average of squared difference of actual and predicted values. The minimum value is 0, the lower MSE represents that the model fits good. While higher the MSE, the model's fitting degrades. In general practice, its square root is often used which is denoted by RMSE (Zach, 2021) and works on the same principle. A strong is its biasness for higher values (Akhilendra, 2019), but as our values had been normalized and scaled, MSE is assumed to perform good.

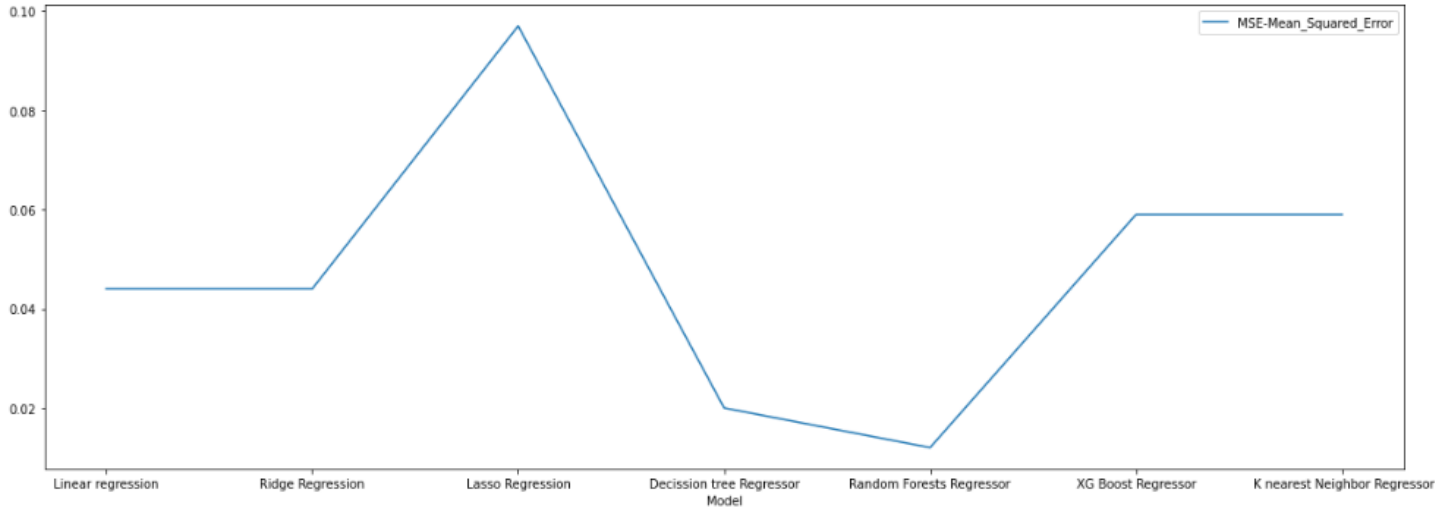


Figure 11: MSE comparison

Amongst the regression models, RFR seems to fit best on the data set with the MSE as low as 0.012, which is slightly better than DTR which has MSE of 0.020.

9.3 Coefficient of Determination (R^2)

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

Also known as the residual sum of square, it's the variance in the dependent variable. Measures how accurately the model makes the predictions. I used this metric to measure the further accuracy of my models. The value of R^2 lies between 1 and 100%. The more the R^2, the more accurate predictions are made by the model.

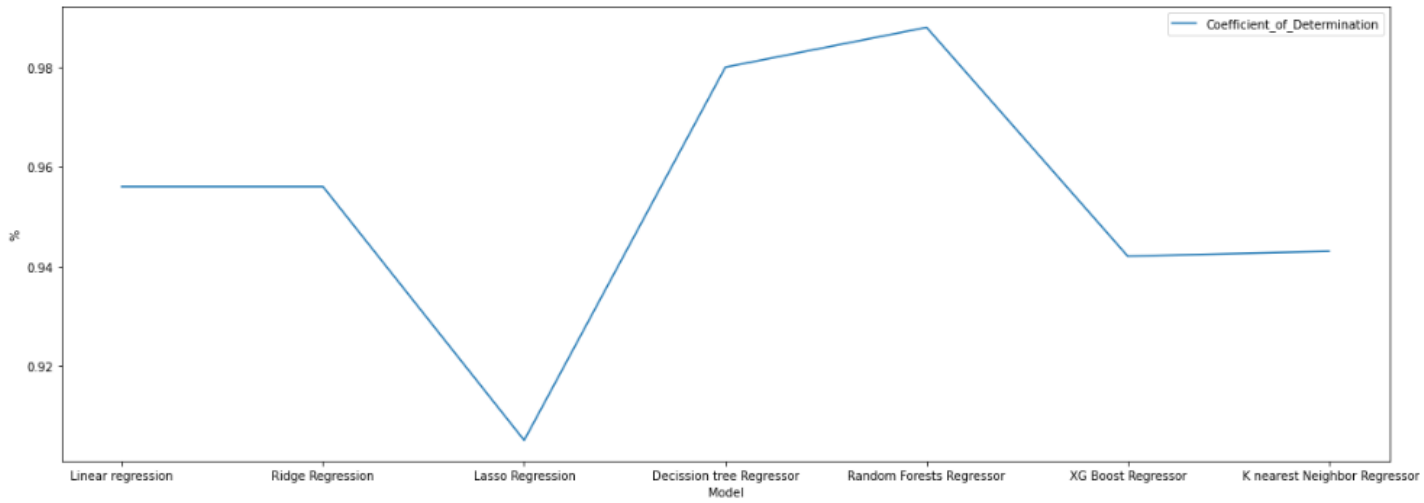


Figure 12: MSE comparison

RFR outperforms other algorithms in making predictions with the score of 98.8%.

9.4 Cost of Computation

Computational cost is the total time that a model takes for its execution on a real time hardware. It is important to know the cost when there are limited resources.

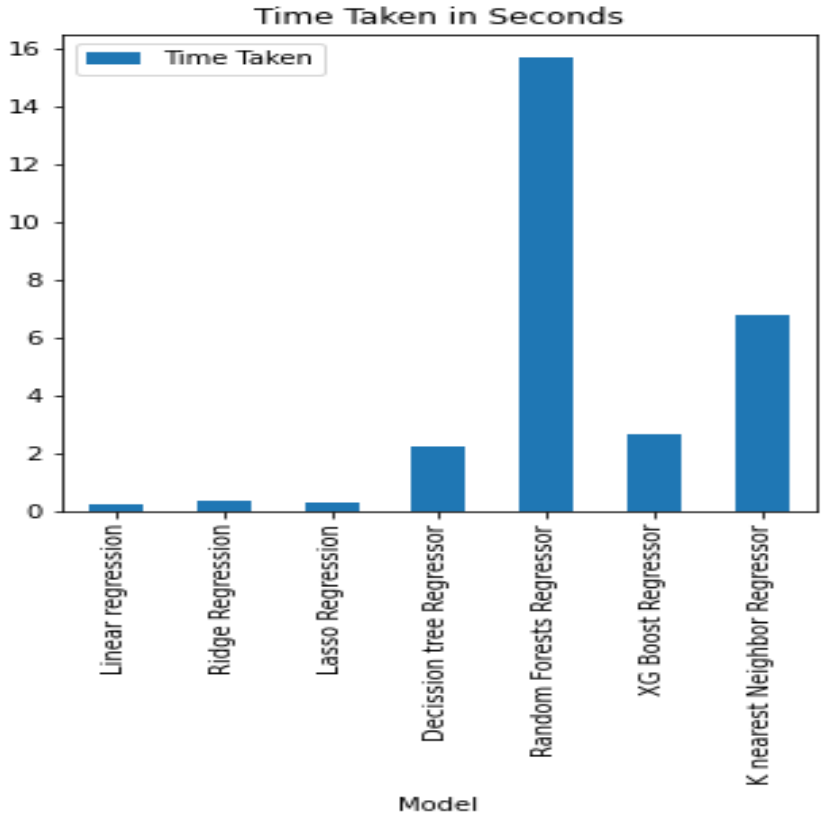
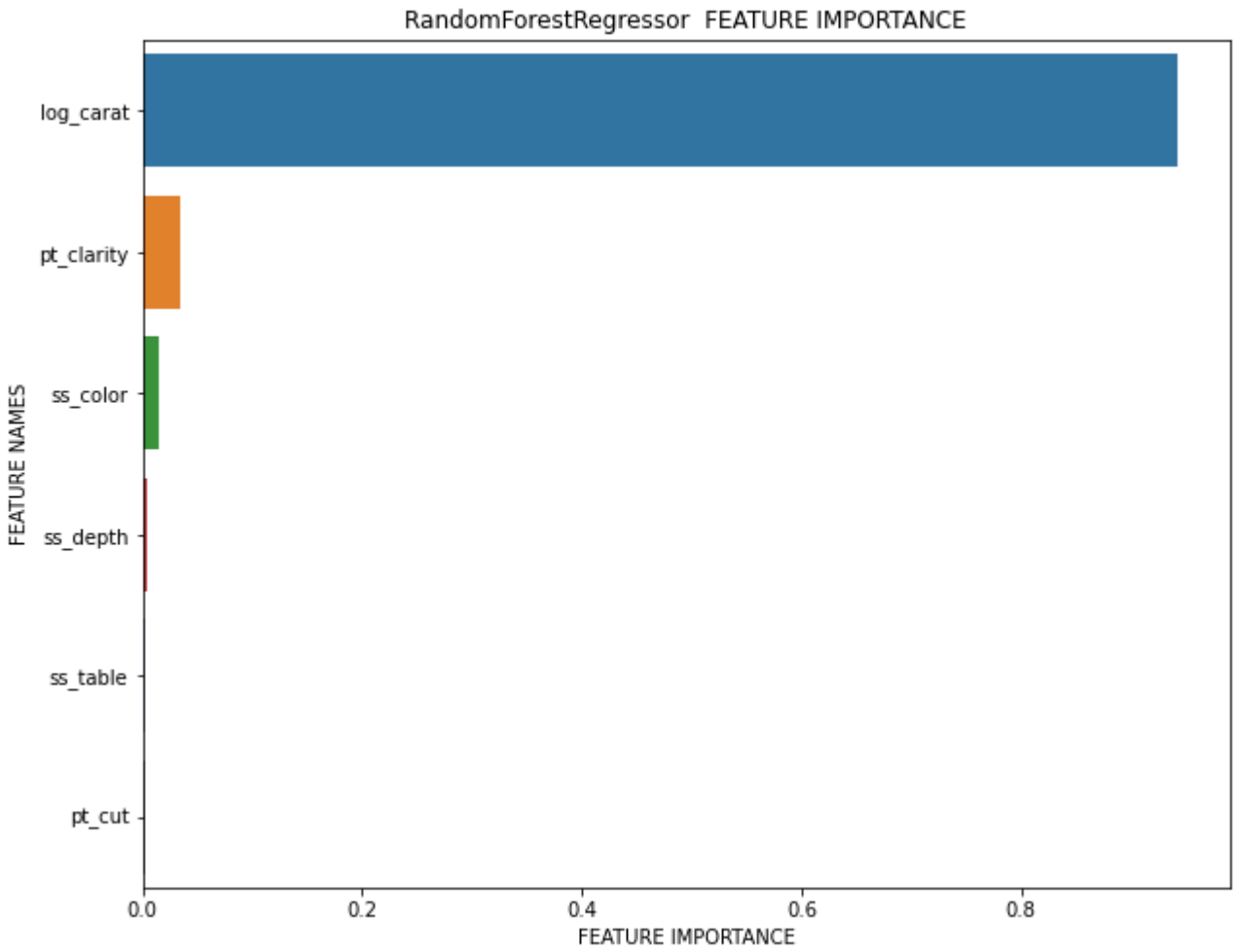


Figure 12: Computational Cost

Time taken by RFR is 15.7 Seconds which is much higher than other models. This could be expensive when there is a huge dataset, but for a commodity like diamond, we need to have at most precision, for which the computational can be traded off.

9.5 Feature Importance



10. Conclusion

Undoubtedly ML is highly accurate in predicting Diamond price. With our Best performing model with 99% accuracy on test data, 98.8% R^2 and just 0.012 MSE, Industry will benefit from these predictions. Overfitting could have been a suspicion, but related results were obtained in previous research, where M5P model made prediction at 99.03% accuracy (Harshvadan, et al., 2021) which helps build the confidence over RFR model. The model could be trained with few more features such as fluorescence, polish, symmetry, luster, etc. to make even more accurate prediction.

----------x----------x----------

11. Bibliography

1) Lex, A. If Diamonds Aren't Rare Why are They So Expensive?. *Rare Carat.com* [online]. Available from: https://www.rarecarat.com/blog/diamond-ring-tips/why-are-diamonds-so-expensive [Accessed 14 May 2022]

2) Slisha, K. (2020). Estimating a Diamond and Diamond Ring's Value. *Withclarity.com* [online]. 22 July. Available from: https://www.withclarity.com/blog/2020/07/22/estimating-diamond-value/ [Accessed 14 May 2022]

3) Anon. (2019). Market Analysis Report. *grandvewsearch.com* [online]. Nov. Available From: https://www.grandviewresearch.com/industry-analysis/diamond-market#:~:text=Industry%20Insights,Pacific%20like%20India%20and%20China. [Accessed 14 May 2022]

4) Sakshi, B. (2019). What are the Trending and Latest Technologies Used in the Diamond Industry. *Techprevue.com* [online]. 14 September. Available from: https://www.techprevue.com/latest-technologies-diamond-industry/ [Accessed 14 May 2022]

5) Katarzyna, R. (2021). Price Prediction: How Machine Learning Can Help You Grow Sales. *Dlabs.ai* [online]. 15 September. Available from: https://dlabs.ai/blog/price-prediction-how-machine-learning-can-help-you-grow-your-sales/#:~:text=ML%20improves%20the%20accuracy%20of%20price%20predictions&text=In%20truth%2C%20most%20conventional%20methods,businesses%20down%20a%20rabbit%20hole [Accessed 14 May 2022]

6) Nasima, T. (2021). All You Need To Know About Different Types Of Missing Data Values And How To Handle It. *Analyticsvidhya.com* [online]. 29 October. Available at: https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/ [Accessed 14 May 2021]

7) Abhigyan. (2020). Detecting and Removing Outliers. *Medium.com* [online]. 12 April. Available at: https://medium.com/analytics-vidhya/detecting-and-removing-outliers-7b408b279c9 [Accessed 15 May 2022]

8) Jim, F. Guidelines for Removing and handling Outliers in Data. *Statisticsbyjim.com* [online]. Available at: https://statisticsbyjim.com/basics/removeoutliers/#:~:text=Outliers%20increase%20the%20variability%20in,results%20to%20become%20statistically%20significant. [Accessed 15 May 2022]

9) Srimal, A. (2018). Why EDA is necessary for Machine Learning. *Medium.com* [online]. 29 July. Available at: https://medium.com/@srimalashish/why-eda-is-necessary-for-machine-learning-233b6e4d5083 [Accessed 15 May 2022]

10) Zach. (2020). How to Read a Correlation Matrix. *Statology.org* [online]. 27 January. Available at: https://www.statology.org/how-to-read-a-correlation-matrix/#:~:text=A%20correlation%20matrix%20conveniently%20summarizes%20a%20dataset.&text=It%20would%20be%20very%20difficult,between%20each%20pair%20of%20variables. [Accessed 15 May 2022]

11) Ananya19b. (2021). Multicollinearity: Problem, Detection and Solution. *Analyticsvidhya.com* [online]. 17 February. Available at: https://www.analyticsvidhya.com/blog/2021/02/multicollinearity-problem-detection-and-solution/ [Accessed 15 May 2022]

12) Changyoung, F et al. (2014). Log-Transformation and its implications for data analysis. *Shanghai Archives of psychiatry*. April. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4120293/ [Accessed 15 May 2022]

13) Georgios, D. (2020). When (& why) to use log transformation in regression?. *gdcoder.com* [online]. 14 February. Available at: https://gdcoder.com/when-why-to-use-log-transformation-in-regression/ [Accessed 15 May 2022]

14) Great, L. (2020). Overview of Multivariate analysis. *mygreatlearning.com* [online]. 29 July. Available at: https://www.mygreatlearning.com/blog/introduction-to-multivariate-analysis/ [Accessed 15 May 2022]

15) Aylin, A. (2021). Splitting a Dataset into Train and Test Sets. *Baeldung.com* [online]. 14 January. Available at: https://www.baeldung.com/cs/train-test-datasets-ratio [Accessed 16 May 2022]

16) Andriy, B. (2022). Linear Vs. Multiple Regression: What's the Difference?. *Investopedia.com* [online]. 31 March. Available at: https://www.investopedia.com/ask/answers/060315/what-difference-between-linear-regression-and-multiple-regression.asp [Accessed 16 May 2021]

17) Pavan, V. (2020). Pros and Cons of Decision Tree Regression in Machine Learning. *Upgrad.com* [online]. 24 December. Available at: https://www.upgrad.com/blog/pros-and-cons-of-decision-tree-regression-in-machine-learning/#:~:text=perfect%20decision%20tree%3F-,Advantages,can%20also%20be%20easily%20understood. [Accessed 16 May 2022]

18) Jason, B. (2021). Why Use Ensemble Learning?. *Machinelearningmastery.com* [online]. 26 October. Available at: https://machinelearningmastery.com/why-use-ensemble-learning/#:~:text=An%20ensemble%20is%20a%20machine,on%20the%20same%20training%20data. [Accessed 16 May 2022]

19) Ajay, O. (2021). Lasso Regression: A Complete Understanding. *Jigsawacademy.com* [online]. 9 March. Available at: https://www.jigsawacademy.com/blogs/ai-ml/lasso-regression#:~:text=Lasso%20regression%20is%20also%20called,helps%20to%20increase%20model%20interpretation. [Accessed 16 May 2022]

20) Divya, S. (2019). What is Predictive Model Performance Evaluation. *Medium.com* [online]. 19 March. Available at: https://medium.com/@divyacyclitics15/what-is-predictive-model-performance-evaluation-8ef117ae0e40 [Accessed 16 May 2022]

21) Zach, (2021). MSE vs. RMSE. Which Metric Should You Use?. *Statology.com* [online]. 30 September. Available at: https://www.statology.org/mse-vs-rmse/ [Accessed 17 May 2022]

22) Akhilendra. (2019). Evaluation Metrics for Regression Models – MAE Vs MSE vs RMSE vs RMSLE. *akhilendra.com* [online]. 20 March. Available at: https://akhilendra.com/evaluation-metrics-regression-mae-mse-rmse-rmsle/ [Accessed 17 May 2022]

23) Harshvadan, M et al. (2021). Diamond Price Prediction using Machine Learning. *IEEE*. 28 January. Available at https://ieeexplore.ieee.org/document/9689412 [Accessed 7 May 2022]