ABSTRACT

When people buy a diamond, they are least bothered about the properties of the diamond and are more concerned about the overall design of the jewellery. We trust the retailer that the diamond he/she is giving is priced according to its value. After doing some research in the market and asking people about their procedure of buying a diamond, we came to know that most of the people don’t have an idea about the properties of the diamond on which the price is calculated. To help the retailers to set the suitable prices and customers to get the right prices, we created a model and a user-friendly website “Storia Lucet” that will calculate the right price for the given diamond.

Regression techniques are one of the most popular techniques used in data mining and predictive modelling. These can be integrated to determine or predict the correct prices of a diamond. Some of the regression algorithms used are Multiple Regression, Polynomial Regression, and Decision Tree Regression. The major objective of the report is to assess these 3 models to determine the right price of the diamond given all its properties. The dataset which is used in this project is a diamond dataset that is available on Kaggle with approximately 54000 rows and 10 properties including the price of the diamond. The results show the predicted price calculated by the regression algorithms. It also helps us to know which properties are more correlated with the price. Correlation Coefficient has also been calculated so that the accuracy of these algorithms can be determined and then the coefficient of all the models is compared to know which model is the most accurate. These findings may help both retailers and customers to find the correct prices of different diamonds.

Table of Contents

List of Figures...7
List of Tables..8

CHAPTER 1 9

1.1 Introduction……………………………………………………………….……………9

1.2 Properties Of Diamond………………………………………………………..………10

CHAPTER 2 12

2.1 Literature Review 12

CHAPTER 3……………………………………………………………………….…..22

3.1 Proposed Methodology of Models………………………………………..…..………22

3.2 Approach For Website……………………………..………..………………….……34

CHAPTER 4 38

4.1 Languages Used...38

4.2 Libraries Used 39

4.3 Software Used 41

CHAPTER 5 42

5.1 Results of the Models Used 42

5.2 Website Result...45

CHAPTER 6 ....**48**

6.1 Analysis of Results..48

CHAPTER 7 49

7.1 Conclusion...49

7.2 Future Work...50

CHAPTER 8 51

8.1 References 51

CHAPTER 9 52

9.1 Annexure 52

List of Figures

Figure 2.1 Color Grading Scale 13

Figure 2.2 Cut Evaluation... 13

Figure 2.3 Different Carat Diamonds.. 14

Figure 2.4 Inclusions in Diamond 15

Figure 2.5 Important Terms in Decision Tree 17

Figure 2.6 Decision Tree Regression 18

Figure 3.1 Unique Grading of cut... 23

Figure 3.2 Unique Grading of colour... 24

Figure 3.3 Unique Grading of clarity... 25

Figure 3.4 Price(X-axis) vs Carat(Y-axis) 28

Figure 3.5 Log Price(X-axis) vs Carat(Y-axis) 28

Figure 3.6 [Carat, Table and Depth] (X-axis) vs Log Price(Y-axis) ……………………29

Figure 3.7 [Carat, Table and Depth] (X-axis) vs Log Price(Y-axis) ……………………30

Figure 3.8 [Carat, Table and Depth] (X-axis) vs Log Price(Y-axis) ……………………30

Figure 3.9 Flowchart of Proposed Methodology..33

Figure 3.10 Response from the server.. 37

Figure 4.1 Data Manipulation in Pandas... 40

Figure 5.1 Correlation Coefficient... 42

Figure 5.2 Mean Absolute Error... 42

Figure 5.3 Root Mean Square Error.. 43

Figure 5.4 Correlation Coefficient... 43

Figure 5.5 Mean Absolute Error... 43

Figure 5.6 Root Mean Square Error.. 44

Figure 5.7 Correlation Coefficient... 44

Figure 5.8 Mean Absolute Error... 44

Figure 5.9 Root Mean Square Error.. 45

Figure 5.10 Header... 45

Figure 5.11 Carousal Images.. 45

Figure 5.12 Diamond Price Calculator.. 46

Figure 5.13 Overview.. 46

Figure 5.14 Cut Section.. 47

Figure 5.15 Clarity Section.. 47

Figure 5.16 Colour Section.. 47

Figure 5.17 Carat Section... 47

List of Tables

Table 3.1 Dataset……………………………………………………………......….22

Table 3.2 Cut Mapping……………………………………………...……......….23

Table 3.3 Colour Mapping ……………….……………………………......….24

Table 3.4 Clarity Mapping ……………….……………………………......….25

Table 3.5 Data Dictionary ……………….…………………….…………......….27

Table 3.6 Correlation Matrix……………….…………………………….....….31

Table 3.7 Updated Dataset ……………….……………………………......….32

Table 3.8 Cut Server Dictionary ……….……………………………......….35

Table 3.9 Colour Server Dictionary …….………...………………......….35

Table 3.10 Clarity Server Dictionary …….……………………………......….36

Table 6.1 Comparison of Models …….……………………...……….......….48

CHAPTER 1

1.1 INTRODUCTION

Diamonds are naturally occurring minerals made up of carbon. Every carbon atom is surrounded by 4 carbon atoms that form the strongest chemical bond. Diamonds are known for a variety of reasons other than its elegance and that includes their coarse nature, ability to scatter light and they are also the hardest occurring substance. It has the highest thermal conductivity and is chemically resistant. It also has some optical properties like high refractive index, high dispersion that enables white light to separate into component colours, and lustre that makes it a popular gemstone [1].

USES OF DIAMOND

Diamonds are considered a symbol of wealth. Some of the uses of diamond besides being made into jewellery are:

- Due to its high strength diamonds are used for cutting, polishing and drilling purposes in aerospace sectors.
- They are also used in the semiconductor industry because they have high heat conductivity and also acts as insulators.
- The small diamond particles called nano-diamonds when attached to chemotherapy drugs can be very effective in treating cancer.

GEM DIAMONDS AND INDUSTRIAL DIAMONDS

Gem diamonds have clarity and colour that make them appropriate for jewellery. They are known for their quality and beauty. These diamonds are rare and form a minor portion of the diamond production. These are known as the perfect diamonds with the least faults and impurities.

Industrial diamonds are used in different procedures such as grinding, cutting, and drilling. Here, properties such as cut, clarity, and colour that are relevant to gemstones are not that important. These diamonds are used to produce small size abrasive powders [1].

1.2 PROPERTIES OF DIAMOND

A standard method of evaluating the quality of a diamond was developed by GIA in 1950 and is known as "The 4Cs of Diamond Quality." The following properties are the major factors that should be considered to obtain an accurate price: cut, carat, colour, clarity, depth, width, length, and table width. The qualities of an individual diamond are described by the 4Cs and the price is based on these qualities. Retailers generally use all these terms to describe and assess the individual diamonds as each of the 4Cs are more accurate than other properties.

What are the responsibilities of a customer while buying a diamond?

Diamond is said to be a symbol of "enduring love and commitment" so it is very important to be very sure that customer is getting what he/she is paying for. A buyer should keep the following things in mind while buying a diamond.

1. Customers should understand all the 4Cs of a diamond to determine its quality. The basics of it will not only tell you about the diamond's quality but will also help in understanding its price. These 4Cs are Color, Clarity, Cut, Carat.

   Mentioned below are some of the price adjustment factors identical to these 4Cs that customers should keep in mind:

   a. Is the diamond cut precise?
   (check the diamond's optical symmetry)

   b. What are the polish and symmetry ratings?

   c. Check whether there is a presence of fluorescence or not?

   d. Does the diamond have inclusions, if yes then what is its nature?

   e. Does it have a culet?

Mentioned below are some price adjustment factors which are not related to diamond but should be kept in mind while buying a diamond:

a) Does the jeweller offer any kind of policies or warranties like free repairs or ring resizing?

b) Do you know your country's laws or taxes which are applicable for consumer goods?

2. A buyer should choose a jeweller like he would choose a doctor. Your jeweller should be trained and should be able to answer all your questions in a simplified language. A trained jeweller will not only explain the 4Cs of a diamond but will also tell the differences between diamonds that look similar. Their training generally comes from a prestigious professional program like GIA Graduate Gemologist or Applied Jewellery Professional diploma programs.

3. Buyers should ask for a Diamond Grading Report: A certification or a diamond grading system from GIA or AJS labs is more than just a piece of important information, it gives proof of what you are buying. It also gives assurance to buyers that their diamond is natural. It becomes difficult for the jewellers to recognize the diamonds without the lab verification.

4. Protect the purchase: Buyers should make sure that their diamonds are appraised and insured which generally relies on grading reports from GIA. Customers should also try their diamonds laser inscribed with the GIA report number as it will verify if your diamond is lost or stolen.

CHAPTER 2

2.1 LITERATURE REVIEW

Diamond is a Greek word which means unbreakable. It is a different form of carbon. Diamonds are formed under high temperatures and pressures roughly 100 miles beneath the earth’s surface. Most of the diamonds were delivered to earth’s surface by volcanic eruptions and some of them were delivered in meteorites. Africa is the only continent that has maximum natural deposits of diamonds.

FEATURES FOR DESCRIBING THE QUALITY OF A DIAMOND

GIA developed some standards for describing diamonds that are accepted globally. The 4Cs of a diamond describes the qualities of a diamond and is a universal method in evaluating an individual diamond. These 4Cs of diamond led to two important things; first, jewellers were able to describe the quality of a diamond in a universal language and second, it helped customers know what exactly they were purchasing.[2]

The 4Cs are:

1. Colour: A pure and structurally perfect diamond is like a drop of water with no hue. Its colour fluctuates from nearly transparent to a spectrum of yellow colour. The more yellow the diamond is, the more it is of poor quality. We can measure the degree of colour by the Gemological Institute of America’s grading system. It is the most accepted grading scheme which starts with the alphabet D, meaning colourless and increases to alphabet Z, meaning light yellow. Though there’s another category of diamonds that are popular and are called coloured diamonds. Some of the colour differences are so minute that they are not visible to customers but these differences result in a huge change in the diamond’s quality and price.

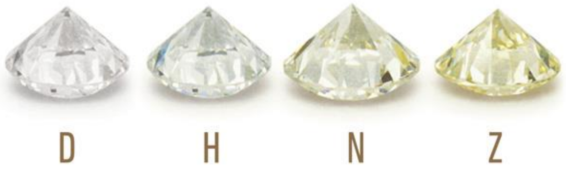


Figure 2.1: Colour Grading Scale

2. Cut: Diamonds are known for their ability to scatter light and people often think that cut means shape but the actual meaning of cut is the ability of diamond's faces to interact with light. The GIA's grading system for calculating the cut evaluates seven components which are; Brightness, scintillation, and fire which evaluates the top-view of the diamond and weight ratio, durability, polish, and symmetry considers its design. The cut of diamond varies from "Ideal" to "Fair". Ideal diamonds are flawlessly polished so that it can be reflective and emit maximum fire.

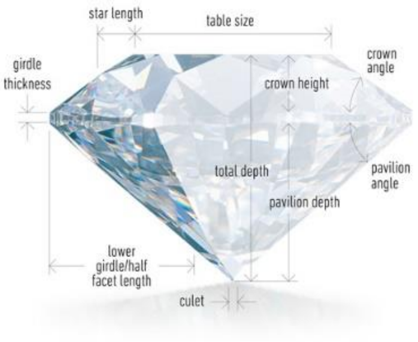


Figure 2.2: Cut Evaluation

3. Carat: Carat is basically the mass or weight of a diamond. We can define a carat as 200mg. As the carat of the diamond increases the price of diamond also increases because larger diamonds are rarely found and are more coveted. If we compare two diamonds with the same carat weight then they can have different prices as they will depend on the other three factors which are clarity, colour, and cut. Generally, diamonds are sold via the carat.

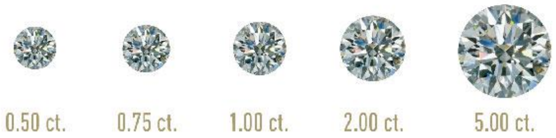


Figure 2.3: Different Carat Diamonds

4. Clarity: Diamonds have several internal characteristics which are called ‘inclusions’ and external characteristics called ‘blemishes. These inclusions and blemishes are very small and cannot be seen by inexperienced eyes. These blemishes decrease the transparency and price of a diamond and can also diminish the durability of the gem. GIA’s grading system for calculating clarity classifies on a scale from FL meaning flawless and perfect to I3 meaning Obvious Inclusions or imperfections. They are graded based on nature, position, and size.

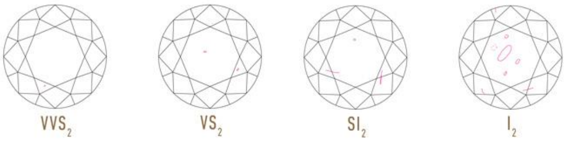


Figure 2.4: Inclusions in Diamond

Some more factors or features which may affect the price of a diamond are x, y, z dimensions, depth and table.

BACKGROUND INFORMATION OF THE ALGORITHMS USED IN THE PROJECT

This report presents different machine learning algorithms that are used to calculate the right price of a given diamond. After doing a lot of research and studying different algorithms like SVM, neural networks and many more, three different regression models are used for calculating the price. Also, the correlation coefficient for the accuracy and mean absolute error for calculating the error are used.

Regression in machine learning is a type of supervised learning method. It is one of the most popular techniques used for predicting value and in data mining. It also helps to find the relationship between variables. People generally know 3-4 types of regression that are common but there are more than 10 types of techniques that are used by analysts [3]. The difference between most of them depends on the number of independent variables and the relation between dependent and independent variables.

TYPES OF REGRESSION

a) Linear Regression: It is one of the simplest and common types of algorithm. In this technique, we predict the value of dependent variable Y that depends on independent variable X. The relation between the two variables is linear and the data is modelled using a straight line. Equation 1 represents the Linear Regression.

Linear Regression Use Cases: Sales forecasting, Resource consumption forecasting, Supply cost forecasting etc.

Linear Regression Assumptions:

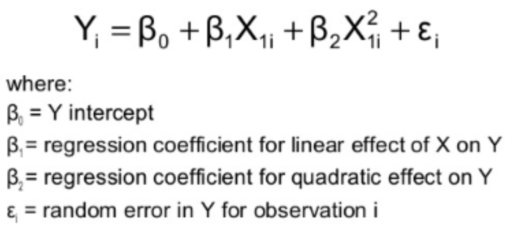
i. All variables should be numeric, not categorical.
ii. Data should be free of missing values.
iii. All predictors are independent of each other.
iv. Prediction errors are normally distributed.

$$y = \theta_1 + \theta_2.x$$

Equation 1

Where,
y= Dependent Variable
x= Independent Variable
θ_1= intercept
θ_2= coefficient of x

b) Polynomial Regression: It is a form of regression analysis in which the relationship between the dependent variable y and the independent variable x is modelled as an nth degree polynomial in x. In this technique, polynomial terms are added to multiple linear regression that is why it is a special case of multiple linear regression. The curve is quadratic instead of linear in polynomial regression. Equation 2 represents the Polynomial Regression.

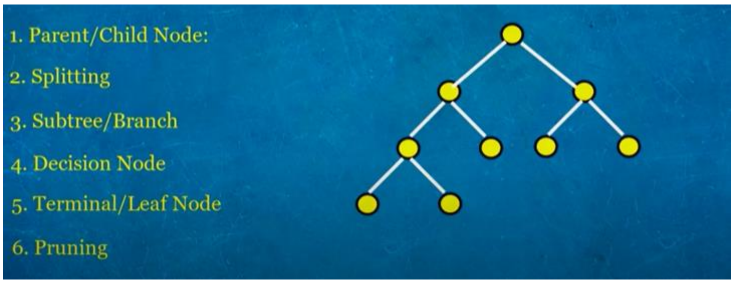


$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \varepsilon_i$$

where:
β_0 = Y intercept
β_1= regression coefficient for linear effect of X on Y
β_2= regression coefficient for quadratic effect on Y
ε_i = random error in Y for observation i

Equation 2

c) Decision Tree Regression: Decision Tree Regression is a non-linear regression technique. It can be used to predict both continuous and categorical target variable. Decision tree develops models in the form of a flowchart or tree. The result of the node is represented by edges and nodes have either decision nodes and end nodes. Various types of Decision Tree Algorithm are ID3, C4.5, CHAID, MARS.



1. Parent/Child Node:
2. Splitting
3. Subtree/Branch
4. Decision Node
5. Terminal/Leaf Node
6. Pruning

Figure 2.5: Important Terms in Decision Tree

In this model, we break down the information by deciding on asking a set of problems.

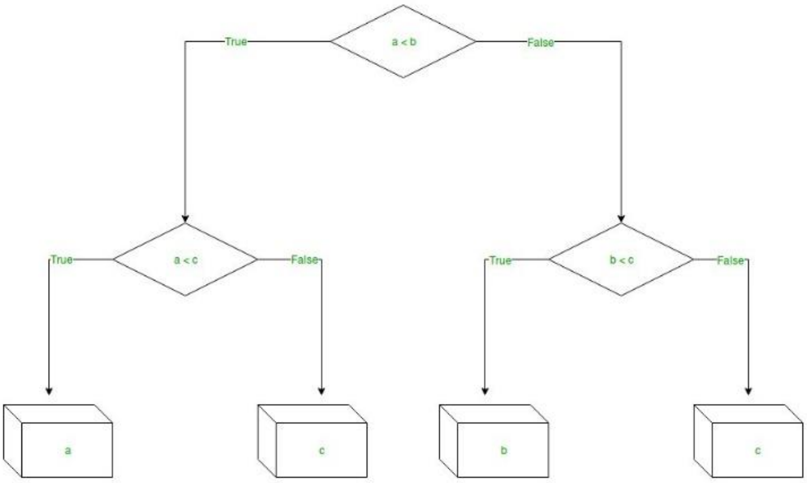


Figure 2.6: Decision Tree Regression

d) Logistic Regression: Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. This algorithm predicts y=1 as a function of x. It is one of the most popular technique to fit the model for categorical data. Equation 3 represents the Logistic Regression.

Logistic Regression Use Cases: Purchase propensity vs ad spends analysis, Customer churn prediction.

Logistic Regression Assumptions:

- Data is missing of free values
- All predictors are independent of each other.
- The predicted variable is binary or ordinal.

$$p = \frac{1}{1 + e^{-(b_0 + b_1 X_1 + b_2 X_2 + --- + b_k X_k)}}$$

Equation 3

e) Support Vector Regression: SVR is same as support vector machine as it has all the main characteristics of the algorithm but both are used to predict different types of variables. This technique can be used for both non-linear as well as linear models. It finds an optimal solution using non-linear kernel functions. It maximizes the margin and minimizes the error. Unlike in simple regression, SVR fit the error within a certain threshold. Equation 4 represents Linear SVR and equation 5 represents Non-Linear SVR.

$$y = \sum_{i=1}^{N} \left(\alpha_i - \alpha_i^*\right) \cdot \left\langle x_i, x \right\rangle + b$$

Equation 4

Polynomial

$$k(\mathrm{x}_i, \mathrm{x}_j) = (\mathrm{x}_i . \mathrm{x}_j)^d$$

Gaussian Radial Basis function

$$k(\mathrm{x}_i, \mathrm{x}_j) = \exp\left(- \frac{\left\| \mathrm{x}_i - \mathrm{x}_j \right\|^2}{2\sigma^2} \right)$$

Equation 5

APPLICATIONS OF REGRESSION

Some of the popular applications of different types of regression are;

i. A Linear regression model is used in salary prediction, financial portfolio, calculating the price of the house based on the number of bedrooms, washrooms etc.

ii. Logistic regression algorithm is used in the insurance sector, finance, and marketing (predicting whether the company will make a profit or not).

iii. SVR algorithms are generally used in oil industries, face recognition, and image classification.

EVALUATION METRICS

The comparison and performance of all the algorithms used in this project have been evaluated using different regression metrics. These metrics are correlation coefficient (R2 Score), root mean square error (RMSE) and mean absolute error (MAE).

1. Correlation Coefficient: It is the R_squared score that measures the accuracy of the model and helps to know which model is better. It is also known as the coefficient of determination. Equation 6 represents the formula of the correlation coefficient.

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

Equation 6

2. Mean Absolute Error(MAE): The difference between the predicted value and the actual value gives Mean Absolute Error (MAE). Equation 7 represents the formula of the Mean Absolute Error.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|$$

Equation 7

3. Root Mean Square Error (RMSE): It is the square root of mean square error. The sum of the square of the difference between the actual values and predicted values divided by the number of data points gives Mean square error (MSE). Equation 8 represents the formula of the Root Mean Absolute Error.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Equation 8

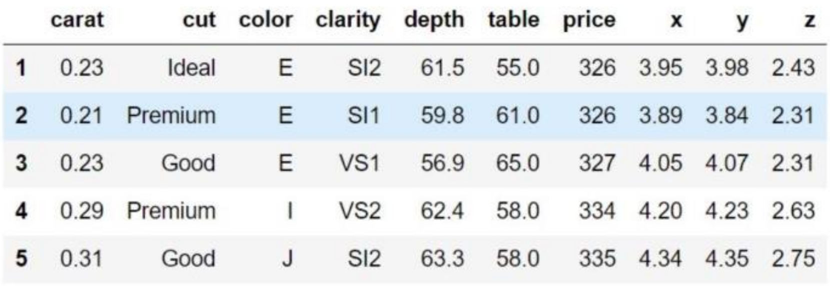
CHAPTER 3

APPROACH TO DESIGN/METHODOLOGY

This project makes a telling conclusion of predicting the accurate price of the given diamond using various tools and methodologies which includes different regression techniques and also a website for the end-user for one-step access to the data model. It calculates the price and then displays the comparative results by comparing the price of all the models.

3.1 PROPOSED METHODOLOGY OF MODELS

1. Data Acquisition: It’s very important to know dataset properly. We tried to get the live data from the internet but the consistency of the model was hampered as the websites were not reliable and those who were providing the live dataset were expensive. So, we acquired the dataset from Kaggle for better consistency and accuracy.



	carat	cut	color	clarity	depth	table	price	x	y	z
1	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
5	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75

Table 3.1: Dataset

To commiserate with the dataset properly, Pie charts are created in Python using the Matplotlib library for Cut, Clarity and Colour as values other than these are Integer values which are can be integrated into the algorithms.

a. The cut is represented in Figure 3.1

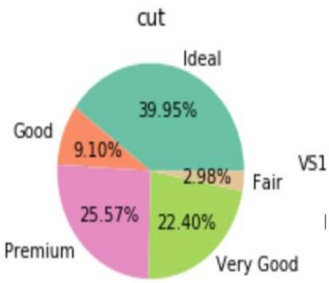


Figure 3.1: Unique Grading of Cut

From the Pie Chart (figure 3.1), an inference can be made that all the unique gradings from the dataset. The percentages of each grade embellish the pre-requisite knowledge helps to create a data dictionary for the training dataset.

In the dataset, these grades changed to numeric values as given in Table 3.2.

Actual	Updated
Fair	1
Good	2
Very Good	3
Premium	4
Ideal	5

Table 3.2: Cut Mapping

b. Colour is represented in Figure 3.2.

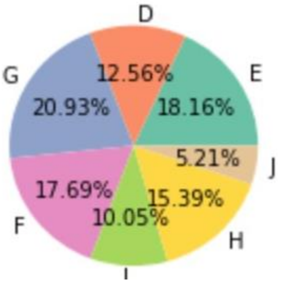


Figure 3.2: Unique Grading of Colour

The Pie chart indicates all the unique grades for the colour data entry. And so, the new data dictionary is created and shown in Table 3.3.

Actual	Updated
J	1
I	2
H	3
G	4
F	5
E	6
D	7

Table 3.3: Colour Mapping

c. Clarity is represented in Figure 3.3.

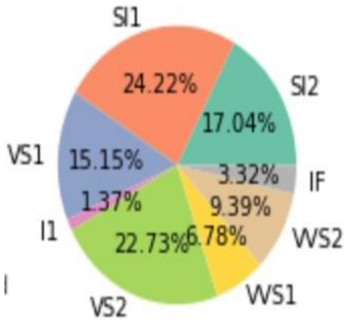


Figure 3.3: Unique Grading of Clarity

The Pie chart for clarity indicates the unique values of each grade in the dataset. With the unique values of the grades, conversion of these values and replacement of these values into integer values for the machine learning algorithms is done and shown in Table 3.4.

Actual	Updated
I3	1
I2	2
I1	3
SI2	4
SI1	5
VS2	6
VS1	7
VVS2	8
VVS1	9
IF	10
FL	11

Table 3.4: Clarity Mapping

The dataset contains a total of 54000 data entries that contains the details of 10 properties each. These 10 properties are listed as followed:

a. Carat: The carat is the measure of the weight of the diamond and the unit is carat only. However, the cut, clarity and colour of the diamond are measured quantitatively in terms of a universal grading system.
b. Cut: Cut is measured on a scale from "Ideal", "Premium", "Very Good", "Good" and lastly "Fair".
c. Colour: Color is graded on a scale starting from "D" is the best to "J" being the worst.
d. Clarity: Clarity is measured from a standard from "I1", "SI2", "SI1", "VS2", "VS1", "VVS2", "VVS1", and "IF".
e. Depth: depth is calculated by the Z-axis value to the average of its diameter at the farthest point.
f. X, Y and Z: coordinates are measured in millimetres.
g. Price: USD.

Detailed data distribution of these properties is shown below in Table 3.5

Columns	**Definitions**
Index	counter
Carat	Carat weight of the diamond
Cut	Describe the cut quality of the diamond. Quality in increasing order Fair, Good, Very Good, Premium, Ideal
Colour	Colour of the diamond, with D being the best and J the worst
Clarity	(in order from best to worst, FL = flawless, I3= level 3 inclusions) FL, IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1, I2, I3
Depth	depth %: The height of a diamond, measured from the culet to the table, divided by its average girdle diameter
Table	table%: The width of the diamond's table expressed as a percentage of its average diameter
Price	price of the diamond
X	length mm
Y	width mm
Z	depth mm

Table 3.5: Data Dictionary

2. Relationship of Properties with Price: It is very important to know the relationship of all the properties with the price as to how the values of these properties affect the price in total. To do so, the graphs are plotted for each of the properties concerning the price of the diamond from the data set as shown in figures below:

   a. The Carat vs Price graph is depicted in Figure 3.4 where the price is greater than 17000 USD. So, to understand the curve properly, the log of the price was calculated which is shown in Figure 3.5. The graph shown in Figure 3.5 indicates that Carat is linearly related to the price. It also indicates that as the value of Carat goes beyond 2 there are more variations in the price which indicates that other factors are more dominating in this region.

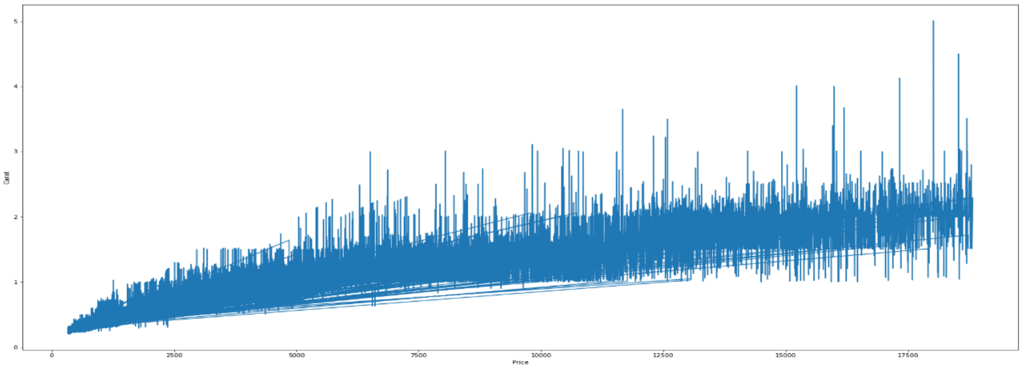


Figure 3.4: Price(X-axis) vs Carat(Y-axis)

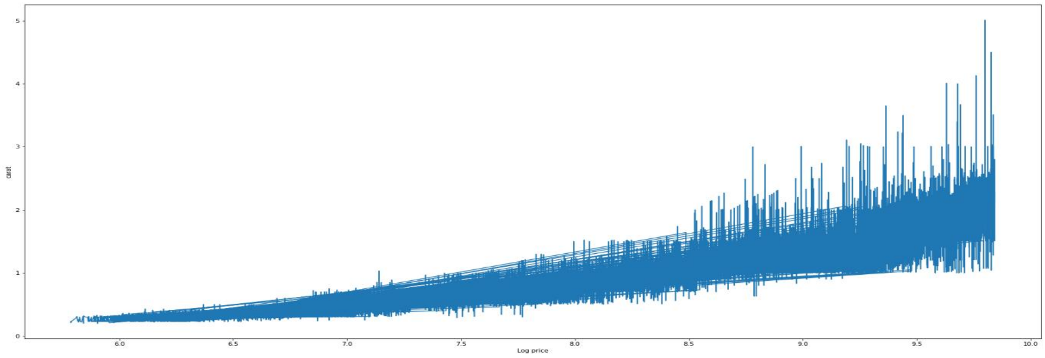


Figure 3.5: Log Price(X-axis) vs Carat(Y-axis)

b. The Graph shown in Figure 3.6 shows different Cuts on a graph (Carat Vs Log Price), (Table Vs Log Price) and (Depth Vs Log Price). It is observed that the Diamonds with Premium cut even tough have less Carat value are expensive indicated by the orange dots in the left-hand side of the graphs. This shows that the premium cut is the best. Although this is just for the sake of idea as, when these values are sent in the training algorithms, there is no need to put the coefficients on our own. It is just to check which values are more relevant for the dataset.

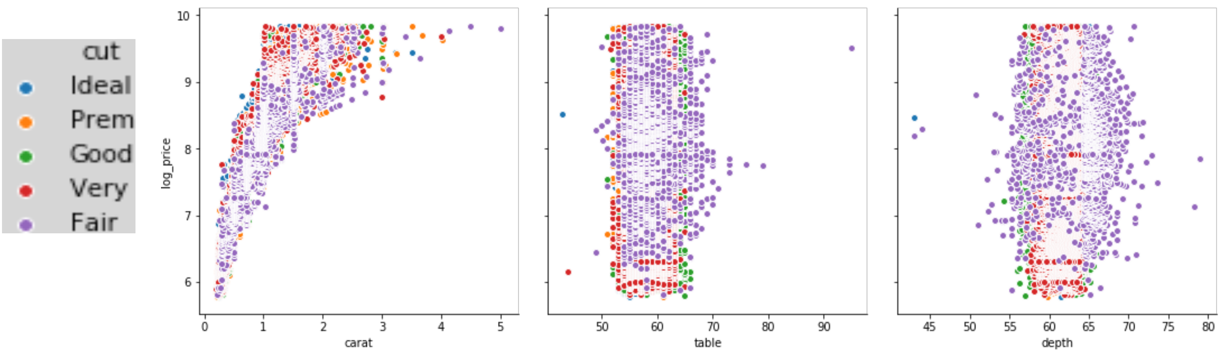


Figure 3.6: [Carat, Table and Depth] (X-axis) vs Log Price(Y-axis)

c. The graphs shown in Figure 3.7 help understand the dataset and how the different values of Clarity grades vary with the price. It is clear that that the IF Clarity stands on top of the grades indicated by the grey dots plotted on the left-side having low carat value but still being expensive.

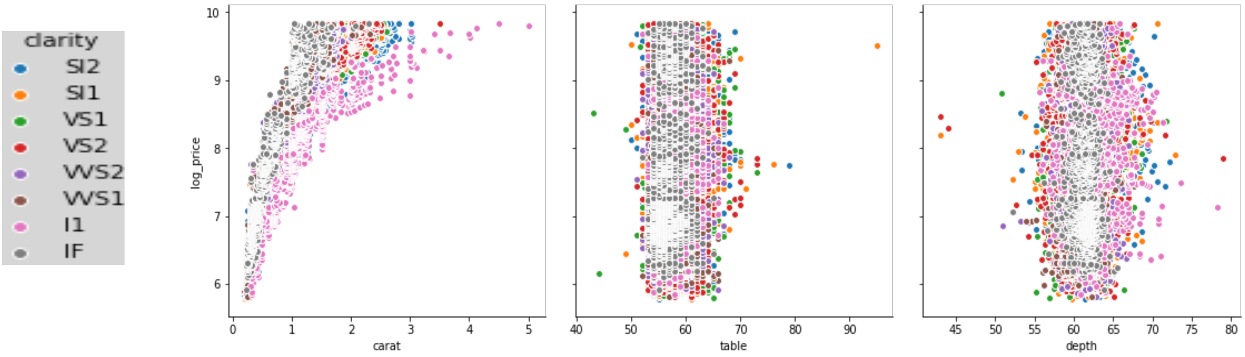


Figure 3.7: [Carat, Table and Depth] (X-axis) vs Log Price(Y-axis)

d. Similarly, in the graph shown in Figure 3.8 “D” grade is the best. As the dots representing the “D” colour are all on the left side having low carat value and high price.

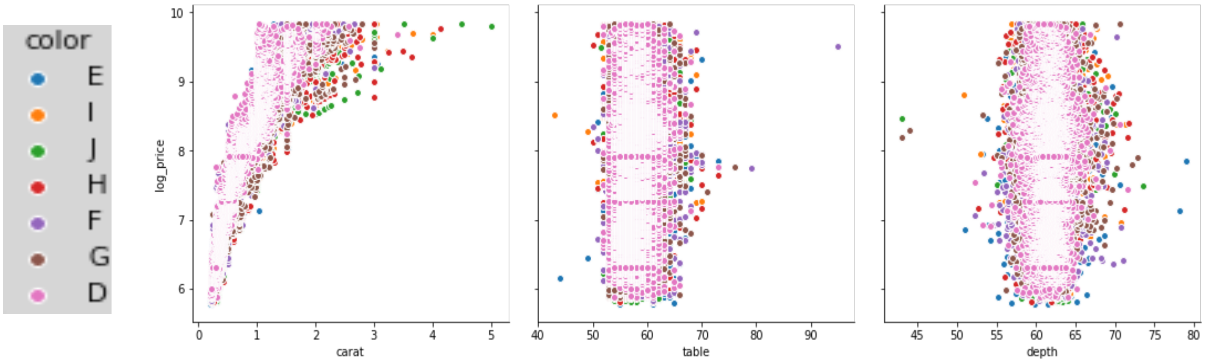
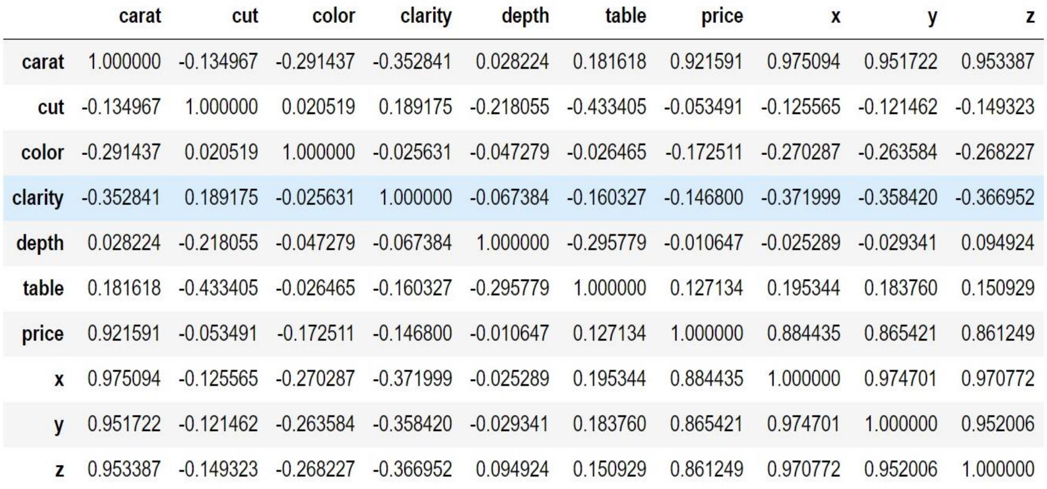


Figure 3.8: [Carat, Table and Depth] (X-axis) vs Log Price(Y-axis)

3. Correlation Matrix: Correlation matrix gives the relationship between all the properties and how much these properties are related.



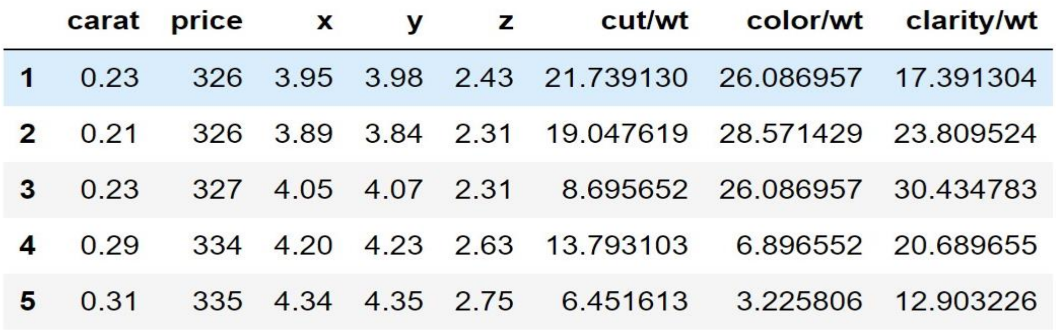
	carat	cut	color	clarity	depth	table	price	x	y	z
carat	1.000000	-0.134967	-0.291437	-0.352841	0.028224	0.181618	0.921591	0.975094	0.951722	0.953387
cut	-0.134967	1.000000	0.020519	0.189175	-0.218055	-0.433405	-0.053491	-0.125565	-0.121462	-0.149323
color	-0.291437	0.020519	1.000000	-0.025631	-0.047279	-0.026465	-0.172511	-0.270287	-0.263584	-0.268227
clarity	-0.352841	0.189175	-0.025631	1.000000	-0.067384	-0.160327	-0.146800	-0.371999	-0.358420	-0.366952
depth	0.028224	-0.218055	-0.047279	-0.067384	1.000000	-0.295779	-0.010647	-0.025289	-0.029341	0.094924
table	0.181618	-0.433405	-0.026465	-0.160327	-0.295779	1.000000	0.127134	0.195344	0.183760	0.150929
price	0.921591	-0.053491	-0.172511	-0.146800	-0.010647	0.127134	1.000000	0.884435	0.865421	0.861249
x	0.975094	-0.125565	-0.270287	-0.371999	-0.025289	0.195344	0.884435	1.000000	0.974701	0.970772
y	0.951722	-0.121462	-0.263584	-0.358420	-0.029341	0.183760	0.865421	0.974701	1.000000	0.952006
z	0.953387	-0.149323	-0.268227	-0.366952	0.094924	0.150929	0.861249	0.970772	0.952006	1.000000

Table 3.6: Correlation Matrix

It clearly shows how all the values are correlated with each other. It is prominently clear that the Carat is highly related to the price of a diamond also to embellish the fact that the values such as Cut, Colour, Clarity Depth and Table are having very less score, however, in the jewellery business, these values play a crucial role so it is important to consider these values. Earlier it was decided to use all the values for the dataset but with reference to Table 3.6 and knowing the importance of Carat, it was decided to divide the other values with Carat.

So, we are considering these values after dividing them by Carat.

Now the dataset looks as shown below in Table 3.7



	carat	price	x	y	z	cut/wt	color/wt	clarity/wt
1	0.23	326	3.95	3.98	2.43	21.739130	26.086957	17.391304
2	0.21	326	3.89	3.84	2.31	19.047619	28.571429	23.809524
3	0.23	327	4.05	4.07	2.31	8.695652	26.086957	30.434783
4	0.29	334	4.20	4.23	2.63	13.793103	6.896552	20.689655
5	0.31	335	4.34	4.35	2.75	6.451613	3.225806	12.903226

Table 3.7: Updated Dataset

4. Models Used: Before starting with the training phase, the dataset must be divided into 2 groups:
    a. Training set 60% of the dataset
    b. Testing set 40% of the dataset

For that, the model_selection library in Python is used to split the dataset.

After splitting the dataset into training and testing, it is needed to pre-process the values so that the data can be sent in for training phase. Pre-processing is necessary as the dataset contains values which are on different scales. What Pre-processing does it that it scales down each value in the dataset to a reduced range which creates an even ground for the data fitting algorithm to work on.

After pre-processing these sets were implemented into these 3 models mentioned below:

a. Linear Regression
b. Polynomial Regression
c. Decision Tree

In Linear Regression, the Sklearn library of python is used for training the dataset. Once the dataset is trained, the testing set is sent in the predict function of Sklearn library for testing the trained model.

The same procedure is followed for other 2 models and calculated the r2 score for comparison.

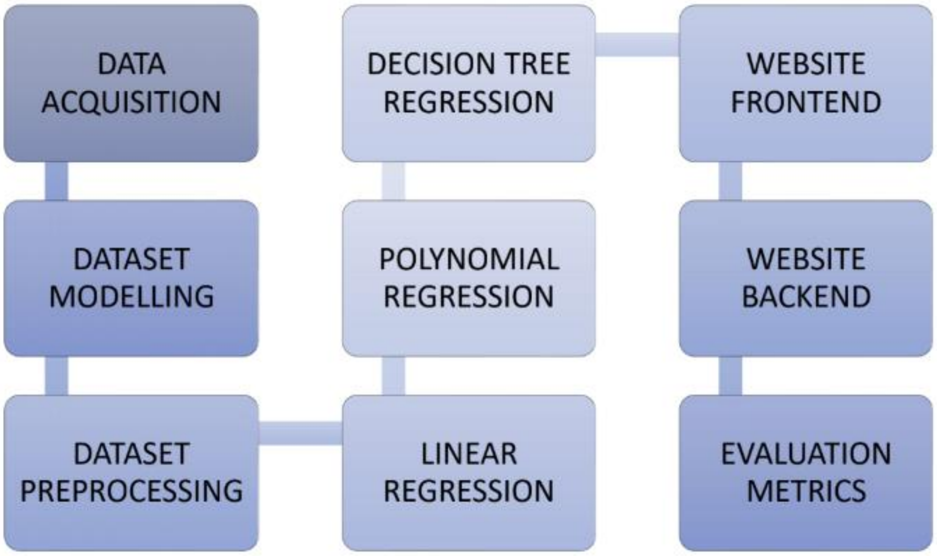


Figure 3.9: Flowchart of Proposed Methodology

3.2 APPROACH FOR WEBSITE

Purpose of the website is to help the novice individual to get the accurate price for the diamond in the comfort of their home or anywhere else. The website is developed keeping in mind those individuals who have less idea about the properties and the key properties of the diamond and also for helping them understand these properties at a glance. Along with predicting the Price, the website also educates the user about the 4C's of the diamond. After the research, it was found that not all the properties such as Depth, X length, Y Length, Z Length and table as mentioned in the dataset are known to the customer, but the properties such as cut, colour, clarity and carat are the properties which are necessary for the predicting the price are known to the customer.

A) FRONTEND: The frontend of the website is always created keeping in mind the customer's feasibility and ease of access. The frontend of the website is developed using Vue.js. To create a user friendly and interactive website we used Quasar Framework to make it more presentable. The name of the website is STORIA LUCET which is Greek for Crystal Jewels is written on the header conventionally. After the header, the Carousal or the slider shows some attractive images to lure the customer and to make it look attractive and appealing.

Data from the customer is acquired in a form and the price is displayed below it. It is important to convert the values from the user that are String into integer values on the frontend only because changing the Learning models will be a heavy toll.

The data dictionary is created in the Vue.js only, which will convert the String values of Cut and Clarity into an integer value. The API call for the server is done using the Axios library. In the API call from the frontend, data is sent as the payload in the body tag of the POST request. As these values must be kept as secret and should not be displayed in the URL.

The server accepts the values in a predefined dictionary as mentioned below:

i. The cut data dictionary is represented in Table 3.8.

Actual	**For Server**
Ideal	Z “String”
Premium	C “String”
Very Good	O “String”
Good	X “String”
Fair	R “String”

Table 3.8: Cut Server Dictionary

ii. For Colour, the values for the server were kept the same as to reduce scepticism. The data dictionary is mentioned below in Table 3.9.

Actual	**For Server**
D	D “String”
E	E “String”
F	F “String”
G	G “String”
H	H “String”
I	I “String”
J	J “String”

Table 3.9: Color Server Dictionary

iii. For Clarity, a different approach is used as the server required the data sent to be in integer values starting from 0 to 6. So instead of creating a data dictionary, we stored the data in an Array according to their values as requested by the server and sent their Array indices respectively. The data dictionary for the server is mentioned in table 3.10 below.

Actual	**For Server**
IF	0
VVS1	1
VVS2	2
VS1	3
VS2	4
SI1	5
SI2	6

Table 3.10: Clarity Server Dictionary

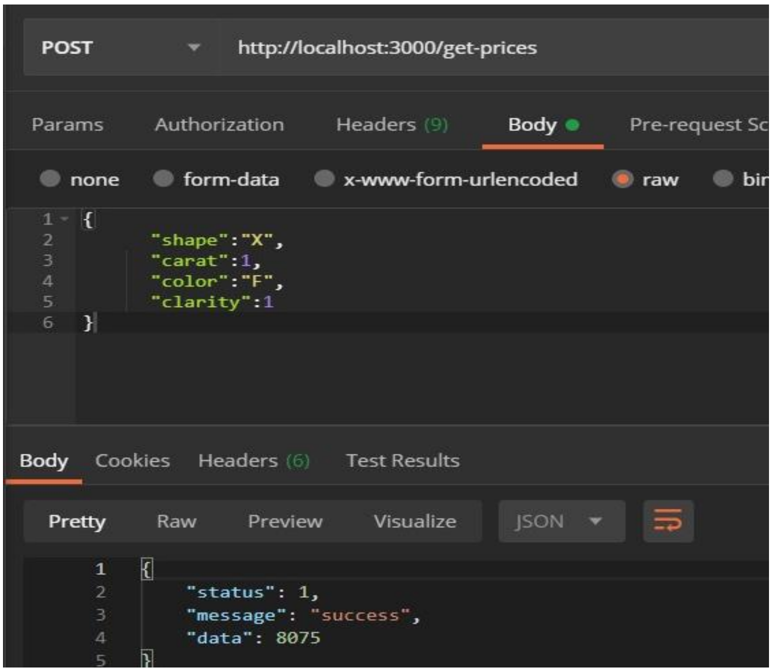
Instead of creating a data dictionary, we added them in an Array for clarity which is shown below:

```
Clarities: ['IF','VVS1','VVS2','VS1','VS2','SI1','SI2']
```

As it is known that array starts from 0 so it works properly. The value of the position of the element in an Array is saved in another variable, is the server value for the same.

iv. Carat is taken as an Integer, so it is sent as it is.

B) BACKEND: The backend of the project is written on Node.js. The node server gets the value as mentioned above and runs the Python script in which the model is running. It accepts the data using the GET request from the frontend at LocalHost://3000 (as the website is not published yet) and sends the values in the script which gives the integer value that is the Price which is emitted using the POST request. The payload of the POST request is shown in Figure 3.10 below.



http://localhost:3000/get-prices

```
{
        "shape":"X",
        "carat":1,
        "color":"F",
        "clarity":1
}
```

```
{
    "status": 1,
    "message": "success",
    "data": 8075
}
```

Figure 3.10: Response from the server

The website only takes the input of these 4 major values for predicting the price.

CHAPTER 4- TOOLS USED

4.1 LANGUAGES USED

1) Python: Its convenience has made it a popular language that lets you work quickly and combine methods more efficiently. Python has helped developers in web development and its flexibility has also allowed analysts to implement different machine learning models. It's simple, easy to learn and has a syntax similar to that of the English language. Some other characteristics of this language are:
   - It has a wide set of libraries like NumPy, Pandas, SciPy, Keras used for machine learning and AI.
   - Python is platform-independent which means you can execute code on one machine and can use it on another machine without any changes. It is supported by platforms like Windows and Linux.
   - Developers use python for different purposes like web development, Data analysis, DevOps, Machine learning, etc.

   This project uses Python 3.7 to build different regression models and to calculate the price of the diamond.

2) VUE.JS: VUE.JS: Vue.js is the latest and upcoming software technology for creating the frontend for a website. It has all the libraries built within, which helps in making the website faster than in comparison with conventional methods. It is a progressive web application framework. It is much easier to include libraries from other projects and makes it easier to use Vue.js into an existing project. Although in reality, it creates the conventional HTML, CSS and JavaScript files but the programmer does not have to write the exact code, Vue.js takes care of it on its own.

3) NODE.JS: NODE.JS: Node.js is used for writing the backend of a website. It is designed to handle the asynchronous events efficiently that is, it is designed to build a scalable network application. It can handle a large number of connections concurrently making it scalable.

4) Postman: Postman is used for checking HTTP requests outside our projects. It is a more efficient way of testing our GET and POST request that weather the request that is sent is giving the desired object or not.

4.2 LIBRARIES USED

1) NumPy: It is a python library that stands for "Numerical python". It was created by Travis Oliphant in 2005. It is a package that provides an array object called ndarray that is 50 times faster than lists in python. This array object is in the form of rows and columns. NumPy is partially composed in python, however, some parts that require fast calculation are written in c or c++. It is used in the domain of random number capability, Fourier transform, matrices, and linear algebra. We can also use lists but NumPy uses less memory and is more convenient as compared to lists. Some of the python distributions like Anaconda, Spyder has NumPy already installed.

2) Pandas: It is the most popular open-source library that stands for "Python Data Analysis Library". It was created by Wes McKinney in 2008 as he wanted to create a flexible tool for data analysis. Pandas is wholly written in C and python. It can analyze data using Series which is a one-dimensional array used to store different data types and DataFrames which is the 2-D data structure consisting of rows and columns. It is used in different domains like finance, analytics, and economics. This project used pandas to import the dataset from the excel. Pandas can be used to load, prepare, manipulate, model, and analyze the data.

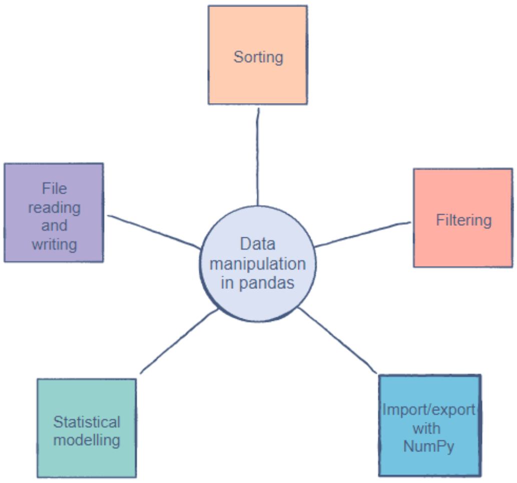


Figure 4.1: Data Manipulation in Pandas

3) Scikit-Learn: It is a library that is built upon SciPy (Scientific Python). It was created by David Cournapeau as a code project of google in 2007 and got published in January 2010. It is an open-source library and is build under a BSD license. Its also built upon some familiar libraries such as NumPy, Matplotlib, Sympy, and Pandas. It focuses more on data modelling rather than loading and manipulating data. This library is known for providing different types of supervised and unsupervised algorithms. Some of these models are:
    - Classification: Support vector machines (SVM), Random forest, Nearest neighbours
    - Regression: Ridge and Lasso Regression, Linear regression, Logistic regression.
    - Clustering: Mean-shift, K-means
    - Model Selection: Cross-validation, Metrics, Grid Search. [4]

4) Matplotlib: It is a popular plotting library of python. It was created by Michael Droettboom and was released in 2003. It is free and open-source and is build under the BSD license. It is an amazing library used for data visualization in python and creates two-dimensional plots of arrays. It is purely written in python. Some other uses where matplotlib is used are in python scripts, web application and other GUI toolkits. There are several plots which can be created using matplotlib library like Bar graph, Histogram, Scatter plot, Pie chart etc.

4.3 SOFTWARES USED

1. Jupyter Notebook: It is a web application that provides a computational environment to develop data science applications and they are called ipython notebooks. It helps you to develop and create code, equations, and text. Some of the uses of jupyter notebook are machine learning, data modelling, data visualization, etc. All the machine learning models used in this project were implemented on Jupyter notebook.

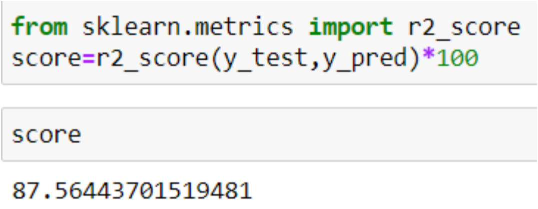
2. WebStorm: It's an Integrated Development Environment for advanced Java Script. It provides full support for JavaScript, Node.js, Vue.js, React, Angular. Making Project in this IDE is very efficient. It creates the project for a file at just on click and the auto-complete feature helps a lot while coding.

CHAPTER 5- DISCUSSION OF RESULTS

5.1 RESULT OF THE MODELS USED

The price of a given diamond was calculated using three different machine learning models and the performance of all the algorithms (Multiple Regression, Polynomial Regression, Decision Tree Regression) has been evaluated using the evaluation metrics. The metrics used are Correlation coefficient, Mean absolute error (MAE), and Root mean square error (RMSE).

a) Linear Regression: It was the first model that was used to calculate the price of a diamond. The figure shows the R2_score that calculates the accuracy of the model. As observed in figure 5.1, this model is 87% accurate.Figure 5.2 and 5.3 shows the calculated Mean absolute error and Root mean square error respectively.

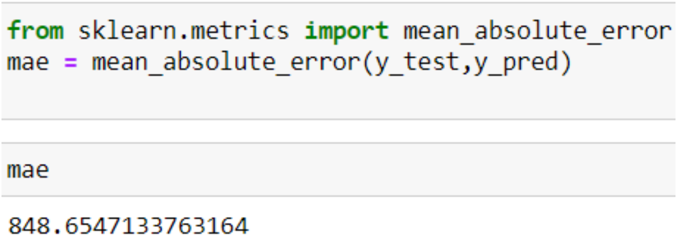


```
from sklearn.metrics import r2_score
score=r2_score(y_test,y_pred)*100
```

```
score
```

```
87.56443701519481
```

Figure 5.1: Correlation Coefficient

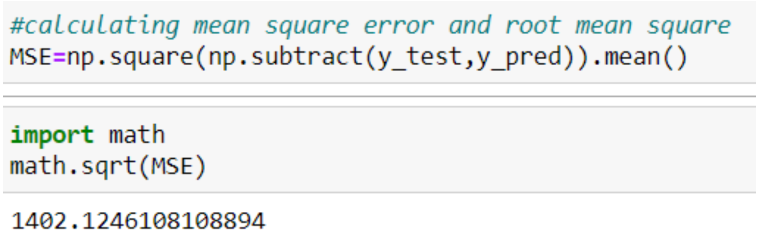


```
from sklearn.metrics import mean_absolute_error
mae = mean_absolute_error(y_test,y_pred)
```

```
mae
```

```
848.6547133763164
```

Figure 5.2: Mean Absolute Error



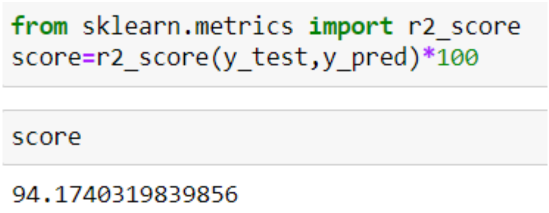
```
#calculating mean square error and root mean square
MSE=np.square(np.subtract(y_test,y_pred)).mean()
```

```
import math
math.sqrt(MSE)
```

```
1402.1246108108894
```

Figure 5.3: Root Mean Square Error

b) Polynomial Regression: It was the second model that was implemented to calculate the price as the previous model had a poor performance. Figure 5.4 shows the R2_score that calculates the accuracy of the model. As observed from the figure this algorithm is 94% accurate which makes it better than the Linear Regression algorithm. Figure 5.5 and 5.6 shows the calculated Mean absolute error and Root mean square error respectively.

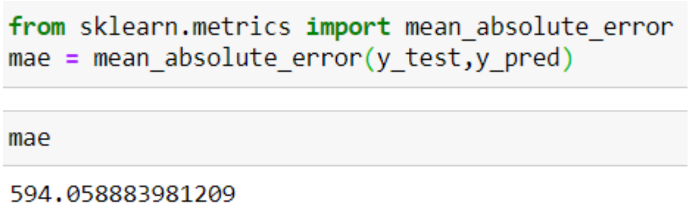


```
from sklearn.metrics import r2_score
score=r2_score(y_test,y_pred)*100
```

```
score
```

```
94.1740319839856
```

Figure 5.4: Correlation Coefficient

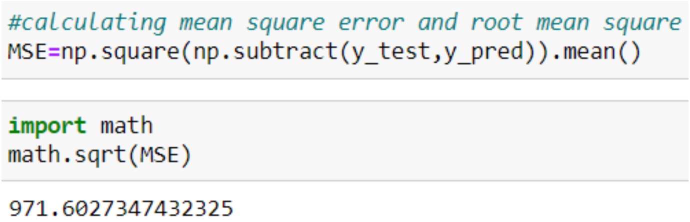


```
from sklearn.metrics import mean_absolute_error
mae = mean_absolute_error(y_test,y_pred)
```

```
mae
```

```
594.058883981209
```

Figure 5.5: Mean Absolute Error



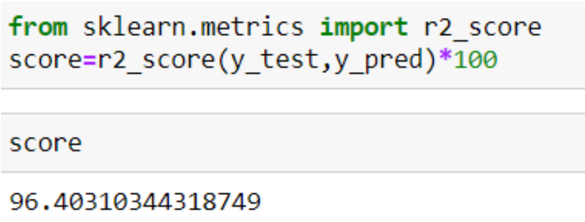
```
#calculating mean square error and root mean square
MSE=np.square(np.subtract(y_test,y_pred)).mean()
```

```
import math
math.sqrt(MSE)
```

```
971.6027347432325
```

Figure 5.6: Root Mean Square Error

c) Decision Tree Regression: It was the last and best model implemented to calculate the price of a given diamond. It had the best performance as compared to previous models. Figure 5.7 shows the R2_score that calculates the accuracy of the model. As observed from the figure this algorithm is 96% accurate which makes it better than the other two algorithms. Figure 5.8 and 5.9 shows the calculated Mean absolute error and Root mean square error respectively.

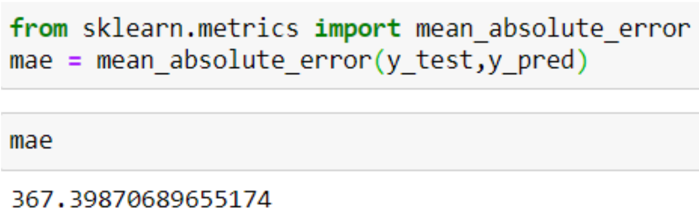


```
from sklearn.metrics import r2_score
score=r2_score(y_test,y_pred)*100
```

```
score
```

```
96.40310344318749
```

Figure 5.7: Correlation Coefficient

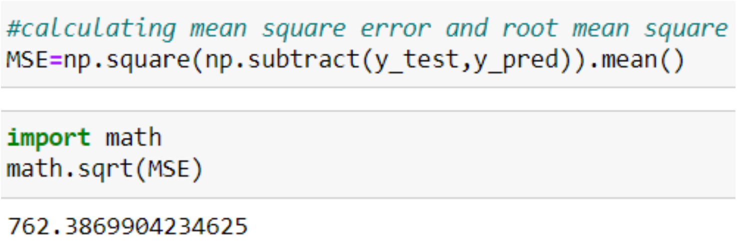


```
from sklearn.metrics import mean_absolute_error
mae = mean_absolute_error(y_test,y_pred)
```

```
mae
```

```
367.39870689655174
```

Figure 5.8: Mean Absolute Error



```
#calculating mean square error and root mean square
MSE=np.square(np.subtract(y_test,y_pred)).mean()
```

```
import math
math.sqrt(MSE)
```

```
762.3869904234625
```

Figure 5.9: Root Mean Square Error

5.2 WEBSITE RESULT

1. Header: The Header of the website shows the name of our website and also the buttons. The title is also a button which brings us back to the home page. The header is shown in figure 5.10.

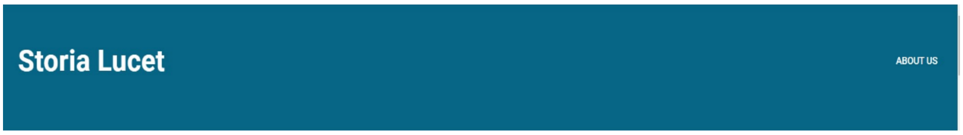


Figure 5.10: Header

2. Top Carousal: It is a slider with different pictures. There are 5 different images.

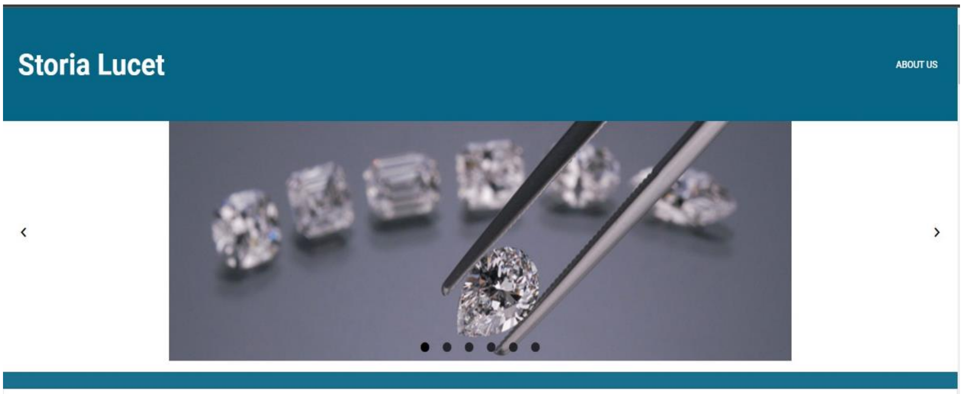
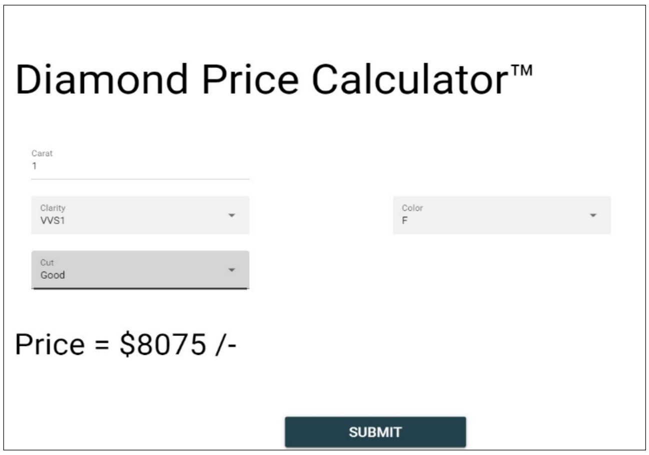


Figure 5.11: Carousal Images

3. Diamond Price Calculator Area: data acquisition section. Here all the values are taken from the customer and the price is calculated. It is shown in figure 5.12.



Diamond Price Calculator™

Carat
1

Clarity
VVS1

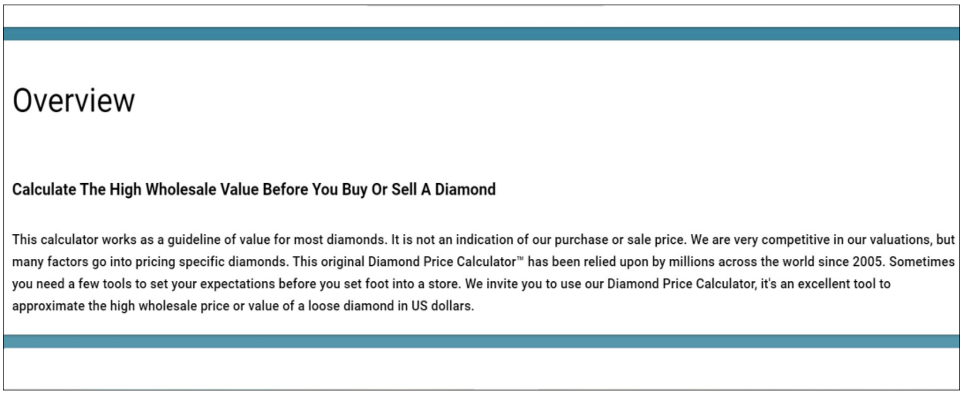
Color
F

Cut
Good

Price = $8075 /-

Figure 5.12: Diamond Price Calculator

4. Overview

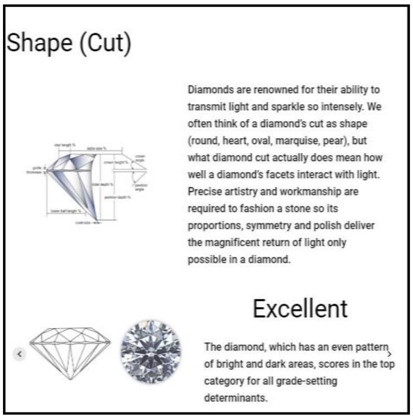


Overview

Calculate The High Wholesale Value Before You Buy Or Sell A Diamond

This calculator works as a guideline of value for most diamonds. It is not an indication of our purchase or sale price. We are very competitive in our valuations, but many factors go into pricing specific diamonds. This original Diamond Price Calculator™ has been relied upon by millions across the world since 2005. Sometimes you need a few tools to set your expectations before you set foot into a store. We invite you to use our Diamond Price Calculator, it's an excellent tool to approximate the high wholesale price or value of a loose diamond in US dollars.

Figure 5.13: Overview



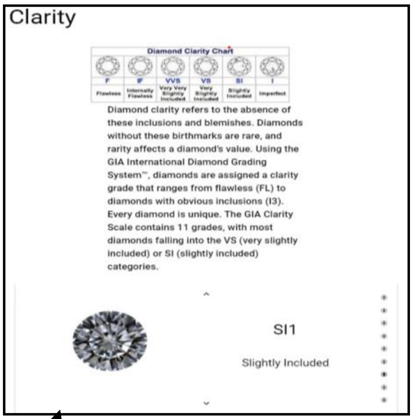
Shape (Cut)

Diamonds are renowned for their ability to transmit light and sparkle so intensely. We often think of a diamond's cut as shape (round, heart, oval, marquise, pear), but what diamond cut actually does mean how well a diamond's facets interact with light. Precise artistry and workmanship are required to fashion a stone so its proportions, symmetry and polish deliver the magnificent return of light only possible in a diamond.

Excellent

The diamond, which has an even pattern of bright and dark areas, scores in the top category for all grade-setting determinants.

Figure 5.14: Cut Section



Clarity

Diamond Clarity Chart

F	IF	VVS	VS	SI	I
Flawless	Internally Flawless	Very Very Slightly Included	Very Slightly Included	Slightly Included	Imperfect

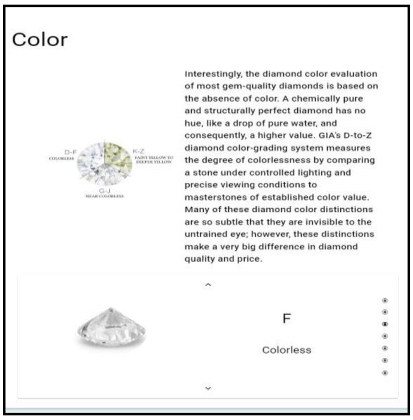
Diamond clarity refers to the absence of these inclusions and blemishes. Diamonds without these birthmarks are rare, and rarity affects a diamond's value. Using the GIA International Diamond Grading System™, diamonds are assigned a clarity grade that ranges from flawless (FL) to diamonds with obvious inclusions (I3). Every diamond is unique. The GIA Clarity Scale contains 11 grades, with most diamonds falling into the VS (very slightly included) or SI (slightly included) categories.

SI1

Slightly Included

Figure 5.15: Clarity Section

CAROUSAL



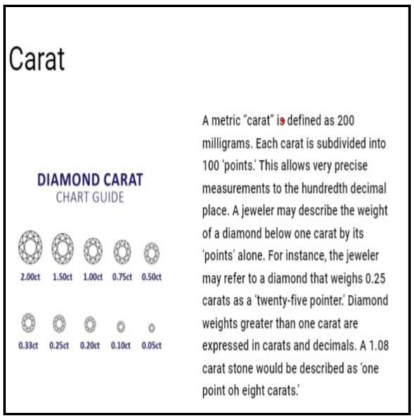
Color

Interestingly, the diamond color evaluation of most gem-quality diamonds is based on the absence of color. A chemically pure and structurally perfect diamond has no hue, like a drop of pure water, and consequently, a higher value. GIA's D-to-Z diamond color-grading system measures the degree of colorlessness by comparing a stone under controlled lighting and precise viewing conditions to masterstones of established color value. Many of these diamond color distinctions are so subtle that they are invisible to the untrained eye; however, these distinctions make a very big difference in diamond quality and price.

F

Colorless

Figure 5.16: Colour Section



Carat

A metric "carat" is defined as 200 milligrams. Each carat is subdivided into 100 'points.' This allows very precise measurements to the hundredth decimal place. A jeweler may describe the weight of a diamond below one carat by its 'points' alone. For instance, the jeweler may refer to a diamond that weighs 0.25 carats as a 'twenty-five pointer.' Diamond weights greater than one carat are expressed in carats and decimals. A 1.08 carat stone would be described as 'one point oh eight carats.'

Figure 5.17: Carat Section

CHAPTER 6

6.1 ANALYSIS OF RESULTS

The comparative results of all three algorithms are studied in detail and are compared with one another to depict the algorithm that best suits the problem domain. All the metrics used, helps us to evaluate the performance of the running algorithm. The dataset used is in xlsx format which includes 10 properties of the diamond and approximately 54000 entries.

The major objective of this report is to display the correct price of a given diamond. The results show the price of different diamonds predicted by three different Regression algorithms.

The following table 6.1 shows the combined performance of the three models by displaying all the evaluation metrics.

Method	**Correlation Coefficient (%)**	**MAE**	**RMSE**
Linear Regression	87.564	848.6547	1402.124
Polynomial Regression	94.174	594.0588	971.602
Decision Tree	96.403	367.3987	762.386

Table 6.1: Comparison of Models

As observed from Table 6.1, the Decision Tree Regression algorithm in diamond price prediction according to the coefficient of determination, mean absolute error, and root mean square error had the best performance as compared to linear regression and polynomial regression. More specifically, the decision tree model had a slight improvement in the correlation coefficient by 2%. It also had the least amount of mean absolute error and root mean square error.

CHAPTER 7

7.1 CONCLUSION

Regression is one of the most popular techniques used for predicting values and in data mining applications. In this report, various algorithms (Multiple Regression, Polynomial Regression, Decision Tree Regression) have been included which follow different techniques to calculate the price of the given diamond. From above calculations and observations, it can be concluded that Decision Tree Regression is the best algorithm among the algorithms used as it has the highest correlation coefficient and least amount of mean absolute error and root mean square error. This model has a high capacity to determine continuous numeric values. These results are recorded by all the algorithms and can be used for further studies. These findings may help both retailers to set a suitable price and customers to get the appropriate price of different diamonds.

A user-friendly website “Storia Lucet” has also been created so that the user can get the actual price of the given diamond. The website is scalable to handle more than 10,000 users at the same time. This website is faster than the other websites as it has been created with the latest and upcoming software such as Node and Vue.js. Once the website is loaded on the browser, no more network call is required for the user interface. A network call is made only when the user sends the request for calculation of the price. Rest all of the information is stored on the browser until and unless the user reloads or refreshes the page. Once the user has opened the page it remains on the browser until it has been shut down. We have created this website keeping in mind the scope of expansion in our project as discussed in the future work.

7.2 FUTURE WORK

1. In future work, we plan to run different algorithms to increase the robustness and precision of the result. Also, we can run these algorithms on different datasets.

2. Using this data set we intend to give a more accurate price for the diamond given the carat, colour, clarity, depth, table of the diamond.

3. Once our model is prepared, we plan to take it one step further. An automated process can be created that will capture the data of the diamond using x-rays, cameras, infrared cameras that will create a virtual model of the diamond and collected data can be used to predict the price of the diamond using our model.

CHAPTER 8

8.1 References

[1] Bhalla, D. (2018, march 25). *Listen Data*. Retrieved from listenData.com: https://www.listendata.com/2018/03/regression-analysis.html

[2] Garbade, D. M. (2018, september 26). *How to use the Scikit-learn Python library for data science projects*. Retrieved from opensource.com: https://opensource.com/article/18/9/how-use-scikit-learn-data-science-projects

[3] INC., G. I. (2002). *4Cs od Diamond Quality by GIA*. Retrieved from gia.edu: https://4cs.gia.edu/en-us/4cs-diamond-quality/

[4] King, H. M. (2005). *Diamond: A gem mineral with properties for industrial use*. Retrieved from Geology.com: https://geology.com/minerals/diamond.shtml

CHAPTER-9

9.1 Annexure

INTRODUCTION

PURPOSE OF PLAN:

As we all go with our family or on our own to buy diamond in any form be it a diamond studded jewelry or a solitaire from the jewelry shop. When we buy a diamond, we are least bothered about the properties of the diamond and are more focused on the over-all design of the jewelry. Once we decide the diamond which we want to buy on the basis of how it looks, we look at the carat of the diamond which we just check to show that we know a little about it. Basically, it's the trust. We trust the retailer that the diamond which the retailer is giving is priced according to its value. Nowadays the retailers are also giving the GIA certificate but still the certificate just tells the details about the diamond not the price of that diamond. After doing some research in the market and asking people about their procedure of buying a diamond, we came to know that most of people don't have an idea about the properties of the diamond on which the pricing is done. They decide to lay their trust on the retailer from which they are buying. The retailer has a complete advantage as whatever price he says the customer has to believe them. In order to help the customer buy the right diamond for the right price we try to create a system which will tell the right price for the given diamond.

BACKGROUND INFORMATION:

What are the uses of Diamonds?

Diamonds are known for variety of reasons other than its elegance and that includes their course nature, ability to scatter light and they are also the hardest naturally occurring substance. Some of the uses of diamond besides being made into jewelry are:

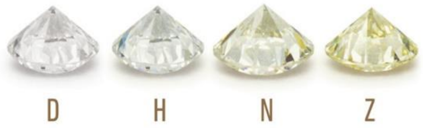
- Due to its high strength diamonds are used for cutting, polishing and drilling purpose in aerospace sectors.
- They are also used in semiconductor industry because they have high heat conductivity and also acts as insulators.
- The small diamond particles called nano-diamonds when attached to chemotherapy drugs can be very effective in treating cancer.

Features For Describing The Quality Of A Diamond

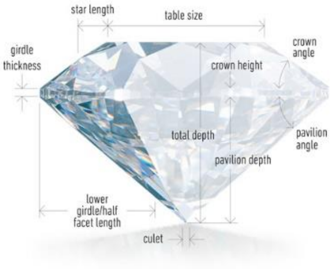
GIA developed some standards for describing diamonds which are accepted globally. The 4Cs of diamond describes the qualities of a diamond and is a universal method in evaluating an individual diamond. These 4Cs of diamond led to two important things; first, jewelers were able to describe the quality of a diamond in a universal language and second, it helped customers know what exactly they were purchasing.

The 4Cs are:

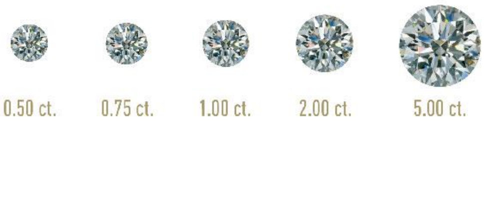
1. COLOUR: A pure and structurally perfect diamond is like a drop of water with no hue. Its colour fluctuates from nearly transparent to a spectrum of yellow colour. The more yellow the diamond is, the more it is of poor quality. We can measure the degree of colourlessness by Gemological Institute of America's grading system. It is the most accepted grading scheme which starts with the alphabet D, meaning colourless and increases to alphabet Z, meaning light yellow. Some of the colour differences are so minute that they are not visible to customers but these differences result in huge change in diamond's quality and price.



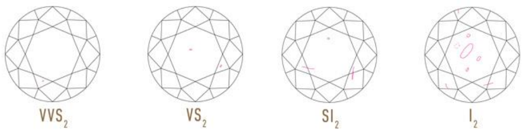
2. CUT: Diamonds are known for their ability to scatter light and people often think that cut means shape but cut actually means the ability of diamond's faces to interact with light. The GIA's grading system for calculating the cut evaluates seven components which are; Brightness, scintillation and fire which evaluates the top-view of the diamond and weight ratio, durability, polish and symmetry considers its design. The cut of diamond varies from "Ideal" to "Fair".



3. CARAT: Carat is basically the mass or weight of the diamond. We can define a carat as 200mg. As the carat of diamond increases the price of diamond also increases because larger diamonds are rarely found and are more desired. If we compare two diamonds with same carat weight then they can have different prices as they will depend on other three factors which are clarity, colour and cut.



4. CLARITY: Diamonds have several internal features which are called ‘inclusions’ and external features called ‘blemishes. These inclusions and blemishes are very small and cannot be seen by untrained eyes. These blemishes decrease the transparency and price of a diamond. GIA’s grading system for calculating clarity classifies on a scale from FL meaning flawless and perfect to I3 meaning Obvious Inclusions or imperfections. They are graded on the basis nature, position and size.



Some more factors or features which may affect the price of a diamond are x,y,z dimensions, depth and table.

What are the responsibilities of a customer while buying a diamond?

Diamond is said to be symbol of “enduring love and commitment” so it is very important to be very sure that customer is getting what he/she is paying for. A buyer should keep the following things in mind while buying a diamond.

1. Customer should understand all the 4Cs of diamond to determine its quality. The basics of it will not only tell you about diamond’s quality but will also help in understanding its price. These 4Cs are: Color, Clarity, Cut, Carat.

   Mentioned below are some of the price adjustment factors identical to these 4Cs that customers should keep in mind:

   - Is the diamond cut precise?
     (check the diamond’s optical symmetry)

   - What are the polish and symmetry ratings?

   - Check whether there is a presence of fluorescence or not?

- Does the diamond have inclusions, if yes then what is its nature?

- Does it have a culet?

Mentioned below are some price adjustment factors which are not related to diamond but should be kept in mind while buying a diamond:

- Does the jeweler offer any kind of policies or warranties like free repairs or ring resizing?

- Do you know your country's laws or taxes which are applicable for consumer goods?

2. Buyer should choose a jeweler like he would choose a doctor. Your jeweller should be trained and should be able to answer all your questions in a simplified language. A trained jeweler will not only explain the 4Cs of diamond but will also tell the differences between diamonds which look similar. Their training generally comes from a prestigious professional program like GIA Graduate Gemologist or Applied Jewellery Professional diploma programs.

3. Buyers should ask for a Diamond Grading Report: A certification or a diamond grading system from GIA or AJS labs is more than just an important information, it basically gives a proof of what you are buying. It also gives assurance to buyers that their diamond is natural. It becomes difficult for the jewellers to recognize the diamonds without the lab verification.

4. Protect the purchase: Buyers should make sure that their diamonds are appraised and insured which generally relies on grading reports from GIA. Customers should also try their diamonds laser inscribed with GIA report number as it will provide verification if your diamond is lost or stolen.

PROJECT GOALS AND OBJECTIVES:

At first, we must understand the problem at hand, what are the problems that we are facing right now and what are the consequences of the problem. In order to tackle the problem, we thought of all the possible solution to this problem and we finally came to the decision to solve this problem using new technologies in artificial intelligence. As artificial intelligence is the latest technology and it is one of the best ways of finding an optimal solution and it provides a method of inferring from the data. AI is a field which can be implemented to make inference from the data and learn from the earlier data set that is provided and find the solution for the problem.

So in this project we try to use Artificial intelligence methods such as Machine learning, Neural networks and its various methods to predict the price of the diamond and help the customer who is going to buy the diamond mostly on the basis of trust and the goodwill of that retailer in the market and make sure he doesn't get fooled and get the product for which the customer is making the payment for. The model that we are creating is universal such that it can be used by the retailer as well as by the customer. The retailer can set the price for the diamond as the retailer already have the given qualities of the diamond and with the properties the retailer can get the price for the given diamond.

SCOPE

SCOPE DEFINITION:

In this project we'll create an optimal model system which will predict the value of the diamond on the bases of the data we have already provided to it. In order to get the data first we did the research on the topic that what are the bases of price analysis of the diamond. We studied about the properties of the diamond and searched on the internet about the various factors that are considered during the pricing of the diamond. We found out that there are various certifications done by the laboratories which provide a certificate for the qualities of the diamond. So if the customer has that certificate then he/she can get the price of that diamond as well. We asked the people and surveyed in our areas that what are the factors that people consider while buying the diamond and we found out that most of the buyers are not aware of the properties that affect the prices. While buying people don't bother checking everything about diamond but when they want to sell the diamond, the diamond is again analyzed and sometimes they don't get the correct price as there were some flaws in the diamond then they have no other choice but to accept that value but also they have the option of second opinion.

First of all, in order to get started with artificial intelligence, first thing we need is the data set. It must include the data that is required and has all the major attributes that are considered while deciding the price. We explored a lot of websites and found a particular dataset that is available on the internet. We examined the dataset and applied our methods on the dataset to calculate the price of the diamond.

PROJECTED BUDGET:

We are aware that the pricing of the diamond is a very complex process and it requires a high-level expert to study and analyze the diamond and find out its properties. But that is the purpose of Artificial intelligence to replace the involvement of human experts with intelligent machines. Right now, we have all we need but if we decide to take it further in which we further propose to create a system that once given a diamond will calculate the properties and then give its actual value or price, which will require a weighing machine that is able measure the weight or carat in milligrams. A scope that will measure the cut of the diamond which can also be done by capturing the image of the diamond in controlled environment. With pictures taken with cameras we have a lot of horizons that we can explore and find out all the values that are required.

CONSTRAINTS

PROJECT CONSTRAINTS:

1. **COST:** Since the project needs a lot of data acquisition and nowadays data labeling is expensive. So, working inside the budget is too tall a task. So we have to calculate how much data will be needed and what are the expenses and if all the properties are known then there is no money involved as such.

2. **QUALITY:** We have to maintain the quality and functionality at any given point of time as it can directly affect user satisfaction.

3. **USER SATISFACTION:** User satisfaction is really a variable for any unceasing and long running job out there. All corporations do business for their prospects sooner or later. as, they are the source of profit plus if you cannot fulfill your clients wishes, you won't have the ability to withstand rivals throughout the time. That is why it is an important project constraint.

4. **TIME:** Time serves as money, so whenever commencing a project, there are closing dates that will be recommended respectively by project sponsor. They will determine when a tasks or project will start or finished.

PROJECT MANAGEMENT APPROACH

1. **PROJECT TIMELINE:**

   The project is divided in six following phases:

   - Requirements analysis phase.
   - Properties analysis phase.
   - Data acquisition phase.
   - Data understanding phase.
   - Model Evaluation phase.

2. **REQUIREMENT ANALYSIS PHASE:**

   During this phase we got to know and understand about various prospects of diamond and its properties

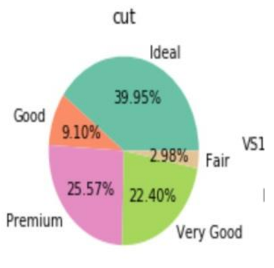
   - Studied the market and people's opinion.
   - Factors influencing the value of a diamond.
   - Is the project achievable or not?
   - Data learning and training set.

3. **PROPERTIES ANALYSIS PHASE:**
   - Analyze the diamond properties which are suitable for the price.
   - Select out the properties on which the model is to be prepared.
   - Acquire the data set including the all the properties.
   - Find the relationships of price with the properties.

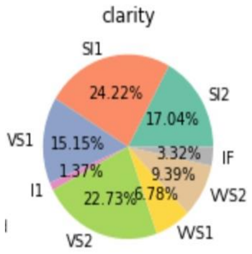
4. **DATA ACQUISITION PHASE:**
   - Research on the internet.
   - Select the Website from which we want the dataset.
   - Acquire the data and then validate it.

5. DATA UNDERSTANDING PHASE:

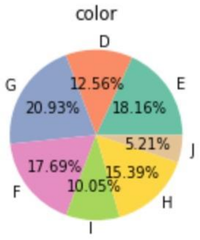
Cut



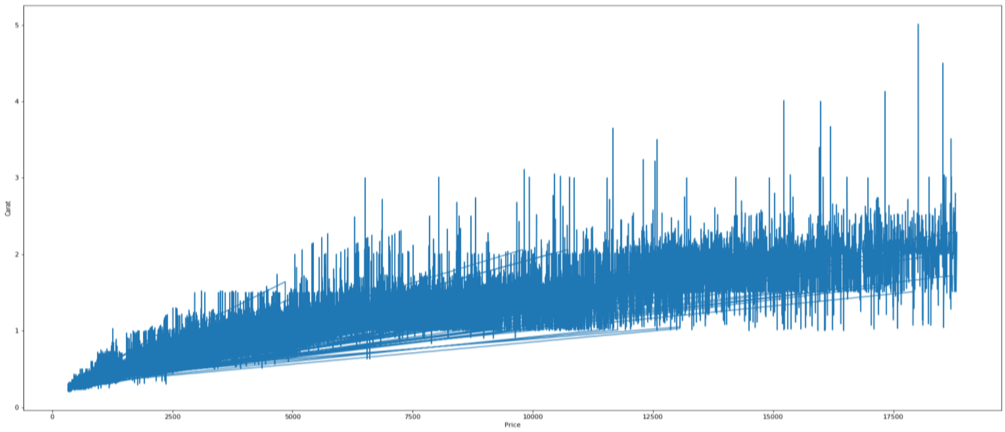
Clarity



Colour

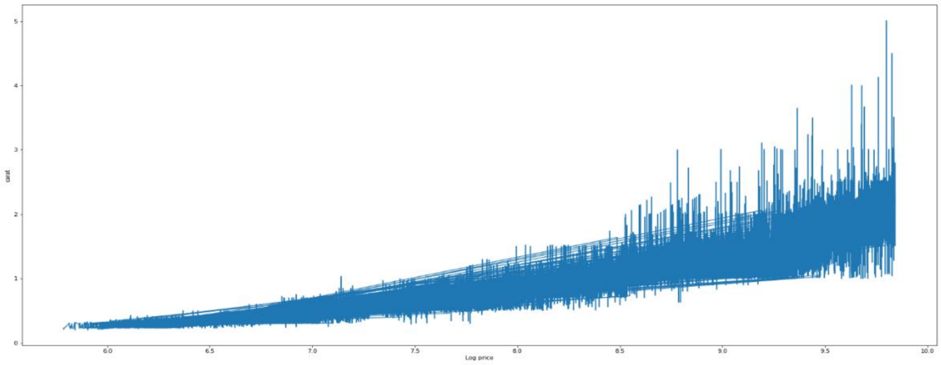


Price Vs Carat

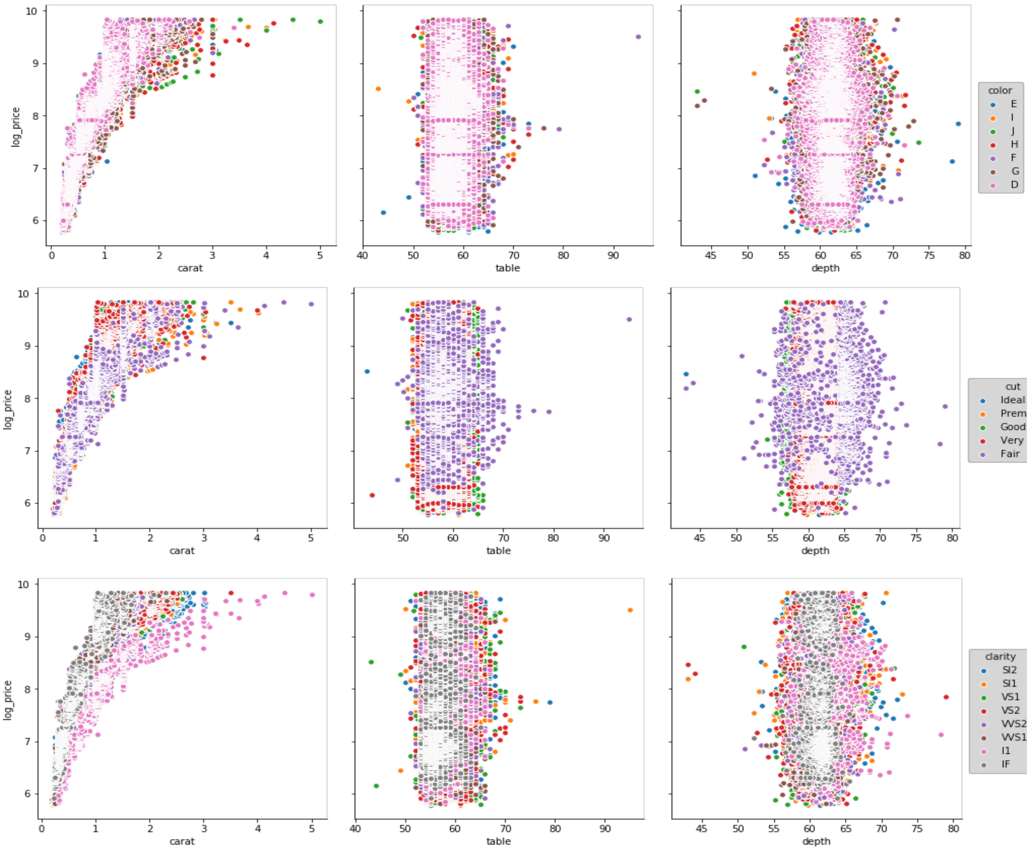


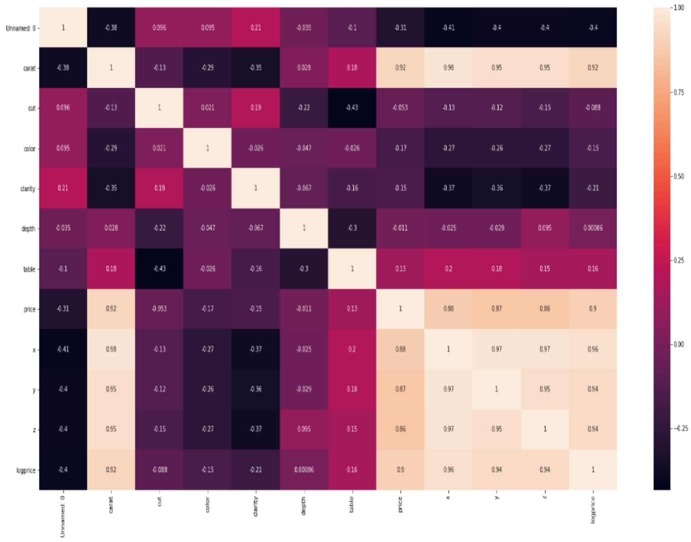
As these values are were difficult to analyze we took the log of the price and then ploted the graph:

Logprice Vs Carat



Heat Map representation of the values





6. TRAINING AND TESTING PHASE:

- We passed our data in the linear regresssion model but dividing the data set into training and testing.
- After that we used the sklearn library for linear regression model.
- Tested with training set of 200 entries.