Exploratory Data Analysis and Price Prediction On Diamond Dataset.

Apurva Srivastava -10800119012

Anurag Sharma - 10800119018

Subhamoy Banerjee - 10800119044

Sumona Mondal - 10800119021

Project Guide - Smita Chourasia

Project Goal

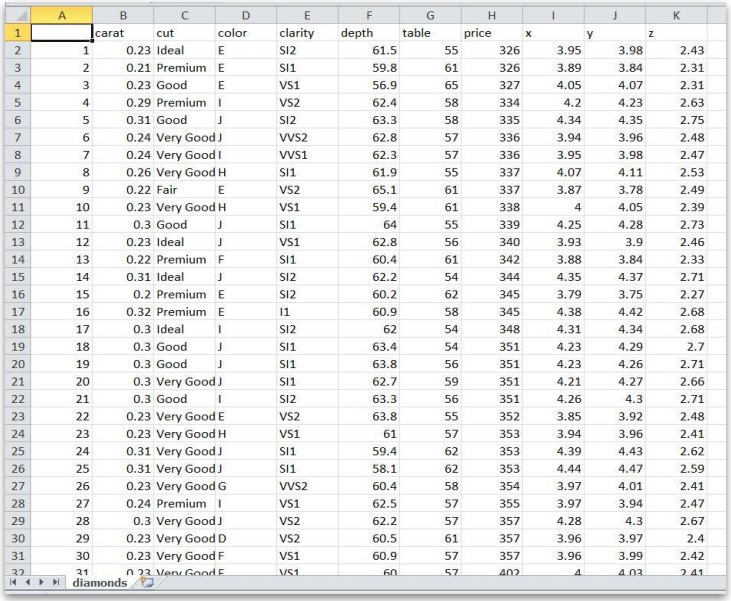
The socioeconomic and political history of the diamond industry is fascinating.

Understanding diamonds are important because each diamond is unique in its way. Even an expert cannot incorporate as much information about price as a picture of the entire market without analysing the characteristics of the diamonds. In this project,we perform an exploratory data analysis using Deep Learning on the diamond dataset to understand the diamond market trends, quality and price by analysing factors for Market Research.

Dataset

- Dataset downloaded from kaggle.com
- A data frame with 53940 rows and 10 variables.

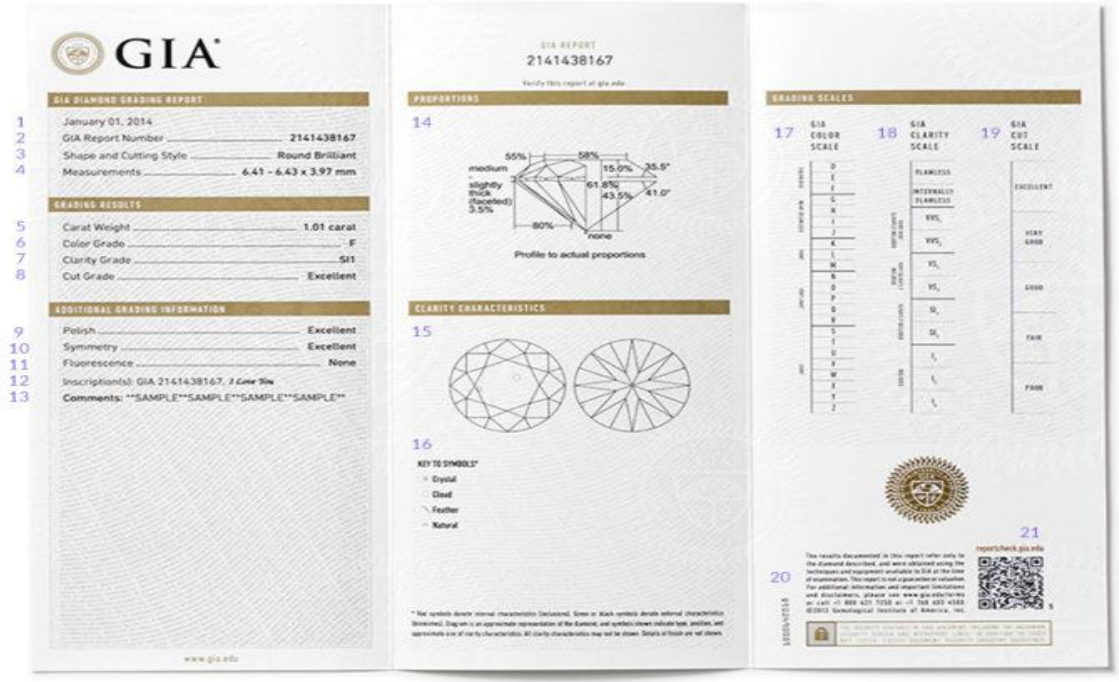
1. Price
   price in US dollars (\$326--\$18,823)
2. Carat
   weight of the diamond (0.2--5.01)
3. Cut
   quality of the cut (Fair, Good, Very Good, Premium, Ideal)
4. Color
   diamond colour, from J (worst) to D (best)
5. Clarity
   a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
6. X
   length in mm (0--10.74)
7. Y
   width in mm (0--58.9)
8. Z
   depth in mm (0--31.8)
9. Depth
   total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43--79)
10. Table
    width of top of diamond relative to widest point (43--95)



	A	B	C	D	E	F	G	H	I	J	K
1		carat	cut	color	clarity	depth	table	price	x	y	z
2	1	0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
3	2	0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
4	3	0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
5	4	0.29	Premium	I	VS2	62.4	58	334	4.2	4.23	2.63
6	5	0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
7	6	0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48
8	7	0.24	Very Good	I	VVS1	62.3	57	336	3.95	3.98	2.47
9	8	0.26	Very Good	H	SI1	61.9	55	337	4.07	4.11	2.53
10	9	0.22	Fair	E	VS2	65.1	61	337	3.87	3.78	2.49
11	10	0.23	Very Good	H	VS1	59.4	61	338	4	4.05	2.39
12	11	0.3	Good	J	SI1	64	55	339	4.25	4.28	2.73
13	12	0.23	Ideal	J	VS1	62.8	56	340	3.93	3.9	2.46
14	13	0.22	Premium	F	SI1	60.4	61	342	3.88	3.84	2.33
15	14	0.31	Ideal	J	SI2	62.2	54	344	4.35	4.37	2.71
16	15	0.2	Premium	E	SI2	60.2	62	345	3.79	3.75	2.27
17	16	0.32	Premium	E	I1	60.9	58	345	4.38	4.42	2.68
18	17	0.3	Ideal	I	SI2	62	54	348	4.31	4.34	2.68
19	18	0.3	Good	J	SI1	63.4	54	351	4.23	4.29	2.7
20	19	0.3	Good	J	SI1	63.8	56	351	4.23	4.26	2.71
21	20	0.3	Very Good	J	SI1	62.7	59	351	4.21	4.27	2.66
22	21	0.3	Good	I	SI2	63.3	56	351	4.26	4.3	2.71
23	22	0.23	Very Good	E	VS2	63.8	55	352	3.85	3.92	2.48
24	23	0.23	Very Good	H	VS1	61	57	353	3.94	3.96	2.41
25	24	0.31	Very Good	J	SI1	59.4	62	353	4.39	4.43	2.62
26	25	0.31	Very Good	J	SI1	58.1	62	353	4.44	4.47	2.59
27	26	0.23	Very Good	G	VVS2	60.4	58	354	3.97	4.01	2.41
28	27	0.24	Premium	I	VS1	62.5	57	355	3.97	3.94	2.47
29	28	0.3	Very Good	J	VS2	62.2	57	357	4.28	4.3	2.67
30	29	0.23	Very Good	D	VS2	60.5	61	357	3.96	3.97	2.4
31	30	0.23	Very Good	F	VS1	60.9	57	357	3.96	3.99	2.42
32	31	0.23	Very Good	F	VS1	60	57	402	4	4.03	2.41

diamonds

GIA Report



GIA

GIA DIAMOND GRADING REPORT

1 January 01, 2014
2 GIA Report Number 2141438167
3 Shape and Cutting Style Round Brilliant
4 Measurements 6.41 - 6.43 x 3.97 mm

GRADING RESULTS

5 Carat Weight 1.01 carat
6 Color Grade F
7 Clarity Grade SI1
8 Cut Grade Excellent

ADDITIONAL GRADING INFORMATION

9 Polish Excellent
10 Symmetry Excellent
11 Fluorescence None
12 Inscription(s): GIA 2141438167.
13 Comments: **SAMPLE**SAMPLE**SAMPLE**SAMPLE**

www.gia.edu

GIA REPORT
2141438167

PROPORTIONS

14

Profile to actual proportions

CLARITY CHARACTERISTICS

15

16

KEY TO SYMBOLS*
Crystal
Cloud
Feather
Natural

* Red symbols denote internal characteristics (inclusions). Green or black symbols denote external characteristics (blemishes). Diagram is an approximate representation of the diamond, and symbols shown indicate type, position, and approximate size of clarity characteristics. All clarity characteristics may not be shown. Details of finish are not shown.

GRADING SCALES

17 GIA COLOR SCALE

18 GIA CLARITY SCALE

19 GIA CUT SCALE

20

21 reportcheck.gia.edu

Data Pre-processing

- Data Type Verification
- Outliers Resolution
- Missing Values

Data Type Verification

- Cut, color and clarity as nominal data type
- Changed them to ordinal as quality has natural order
- Create indicator columns as needed
- Drop dimensionless diamonds data

Outlier Resolution

- Identity potential outliers using Explore Outliers tool
- Manually resolve the outliers based in business knowledge.

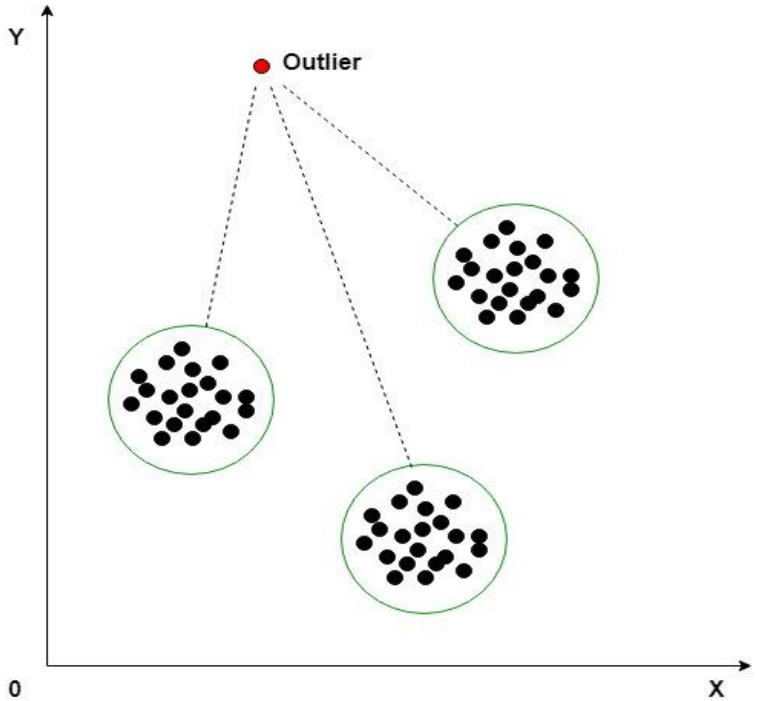
What is an Outlier?

An **outlier** is an object that deviates significantly from the rest of the objects. They can be caused by measurement or execution error. The analysis of outlier data is referred to as outlier analysis or outlier mining.

Detecting Outlier

Clustering based outlier detection using distance to the closest cluster:

In the K-Means clustering technique, each cluster has a mean value. Objects belong to the cluster whose mean value is closest to it. In order to identify the Outlier, firstly we need to initialize the threshold value such that any distance of any data point greater than it from its nearest cluster identifies it as an outlier for our purpose. Then we need to find the distance of the test data to each cluster mean. Now, if the distance between the test data and the closest cluster to it is greater than the threshold value then we will classify the test data as an outlier.



Before Removing the Outliers

click here for the image

Here, we are using *pairplot* in seaborn library to show the relationship between two numerical variables. Here, we plot all the properties of the diamond vs cut of the diamond.

From the *pairplot* we can see that there are many data points which are deviating from the usual clusters of the data points.

After Removing the Outliers.

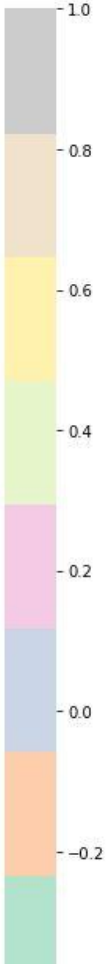
click here for the image.

After removing the outliers which were causing noise in the dataset, we are left with a more clean and useful data which is going to help us in performing the analysis on the Prediction of the Diamond price.

Plotting Heat map for visualization of Data

Heatmaps are a great tool for creating beautiful figures and can provide us with insights on tre
and allow us to easily identify potential outliers within a dataset.
Here, we are using a heatmap- the correlation matrix heatmap, to visualize the data.

A correlation matrix allows us to identify how well, or not so well, features within a dataset
correlate with each other as well as whether that correlation is positive or negative.
The returned value in our case is between -0.4 and +1.0



Heat map.

click here for the image.

- There is strong +ve correlation between carat and price and dimensions(x, y, z) and price.
- There is +ve correlation between color and price and table(width of top of diamond relative to widest point (43--95)) and price.
- There is little to no correlation between cut and price and depth(total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43--79)) and price.

Separating Dependent and independent variable

Here, we are separating the dependent variable i.e., price of the diamond from the all the independent variables for making the dataset ready to trained on various models of regressions.

We are building pipelines of standard scaler and model for various regressors.

1. Linear Regression
2. Decision Tree Regression
3. Random Forest Regression
4. K-nearest neighbor Regression
5. XGBoost Regression

Finding which model is better!

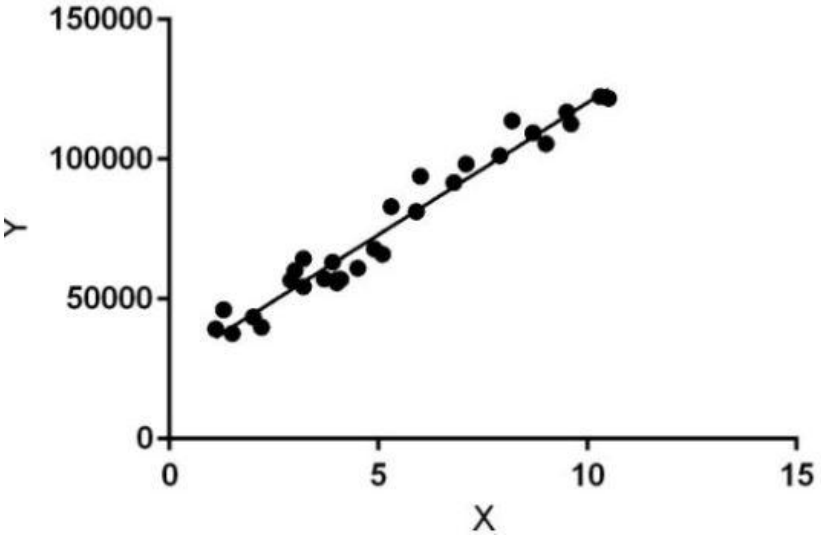
To find the better model we are working the various pipelines of the models on the training data and calculating the cross validation score.
Then, using that we are calculating the negated root-mean- squared error.
Closer the RMSE is to 0, better is the fit.
RMSE is a good estimator for the standard deviation σ of the distribution of our errors!

The RMSE is directly interpretable in terms of measurement units, and so is a better measure of goodness of fit than a correlation coefficient.

About each algorithms.

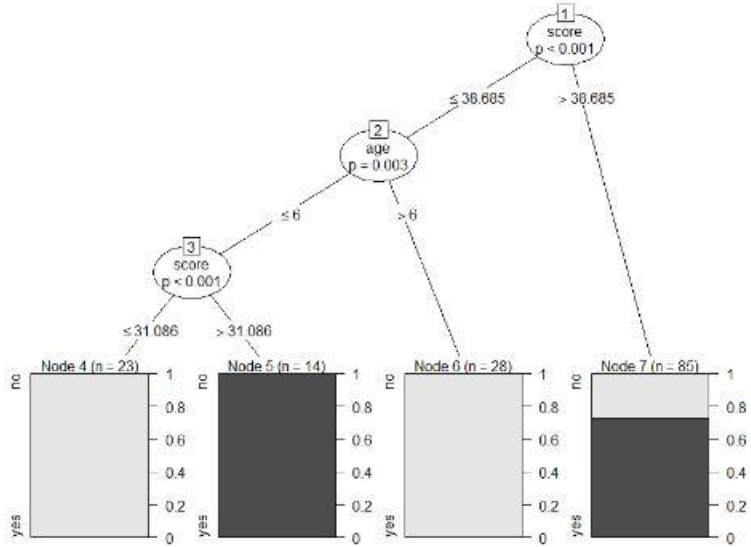
Linear Regression.

Linear Regression is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.



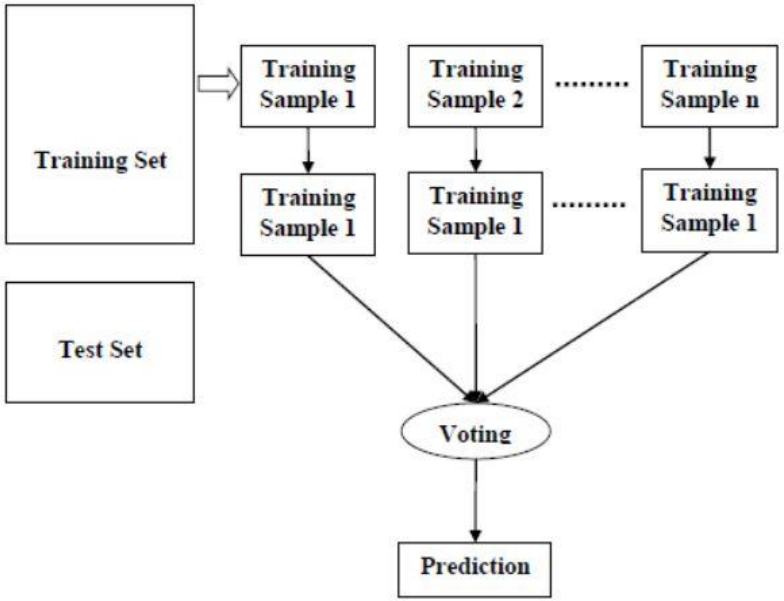
Decision Tree.

Decision Trees are useful supervised Machine learning algorithms that have the ability to perform both regression and classification tasks. It is characterized by nodes and branches, where the tests on each attribute are represented at the nodes, the outcome of this procedure is represented at the branches and the class labels are represented at the leaf nodes. Hence it uses a tree-like model based on various decisions that are used to compute their probable outcomes.



Random Forest.

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest.
Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

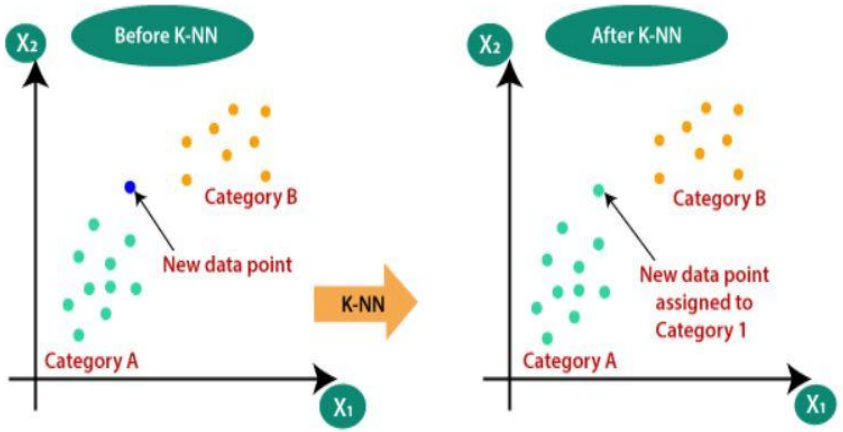


K-NN algorithm.

K-Nearest Neighbours is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection.

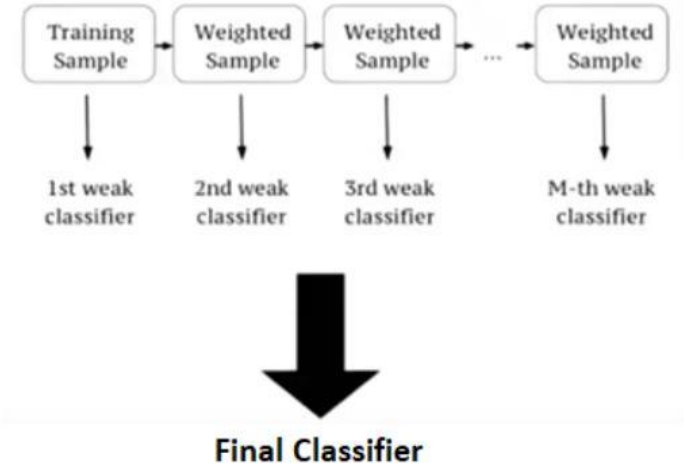
It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as GMM, which assume a Gaussian distribution of the given data).

We are given some prior data (also called training data), which classifies coordinates into groups identified by an attribute.



XGBoost

XGBoost is an implementation of Gradient Boosted decision trees. This library was written in C++. It is a type of Software library that was designed basically to improve speed and model performance. It has recently been dominating in applied machine learning.
In this algorithm, decision trees are created in sequential form.
Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results.
The weight of variables predicted wrong by the tree is increased and the variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model.
It can work on regression, classification, ranking, and user-defined prediction problems.



Negated Root Mean Squared Error of models.

- Linear Regression = -1344.798387
- Decision Tree = -51.745803
- Random Trees = -36.272998
- K-NN algorithm = -666.216132
- XGBoost Regressor = -205.509411

Here, we see that regressors like Linear Regression and K-NN algorithm is giving very high RMSE.

So, we are selecting other regressors for testing on unseen data,

Testing the Xgboost Model on Unseen Data

Checking how many times the diamond price is Overestimated or Underestimated and how many times it is exact on point.

OverEstimated -7366

UnderEstimated - 6072

Exactly Estimated - 39

Testing the Decision Tree Model on Unseen Data

Checking how many times the diamond price is Overestimated or Underestimated and how many times it is exact on point.

OverEstimated -2001

UnderEstimated - 2582

Exactly Estimated - 8894!!

Testing the Random Forest Model on Unseen Data

Checking how many times the diamond price is Overestimated or Underestimated and how many times it is exact on point.

OverEstimated -1180

UnderEstimated - 6236

Exactly Estimated - 6061

Conclusion

As we can see that in the Decision Tree gives us better estimation of price of diamonds.

So, we will be using **Decision Tree** as our Model!!