Diamond Price Prediction

Linear Regression

By
Reem Alsuqaih
Njood Alqahtani

TABLE OF CONTENTS

INTRODUCTION

Have you ever asked yourself, how are diamonds priced? Well, this project talks about the diamonds price prediction based on their cut, color, clarity & other attributes and it also covers the building a simple linear regression model using Python

Methodology

Split the Data

03

Simple EDA

02

03

Experiments

01

Dataset
Understanding

04

Result

Dataset Understanding

Column	Description
Price	price in US dollars .
Carat	weight of the diamond .
Cut	quality of the cut (Fair, Good, Very Good, Premium Ideal)
Color	diamond color, from J (worst) to D (best).
Clarity	a measurement of how clear the diamond.
X , Y , Z	Length in mm , width in mm and depth in mm.
Depth	Total Depth percentage .

Pre-processing the dataset & Simple EDA

- To get started, we need to import some useful libraries that will help us import the dataset into our python environment, manipulate and analyze the same and later help us to visualize it.

- Outliers was observed across the dataset, so a good approach it to either remove it or correct it .

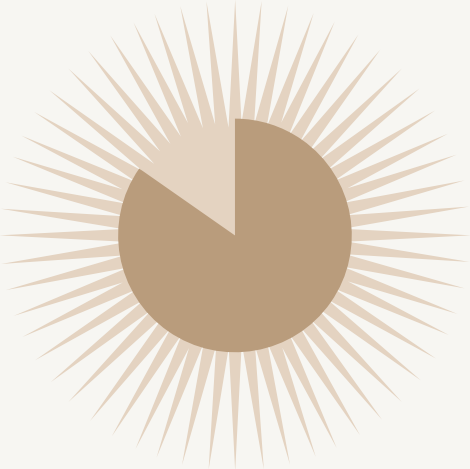
- The Model does not accept null value .

Experiments

Baseline Model

```
Validation R^2 score was: 0.889491407710001
Feature coefficient results:

carat : 11101.09
cut : 73.19
color : -267.77
clarity : 288.00
depth : -98.38
table : -92.39
x : -729.81
y : 125.86
z : -968.85
```



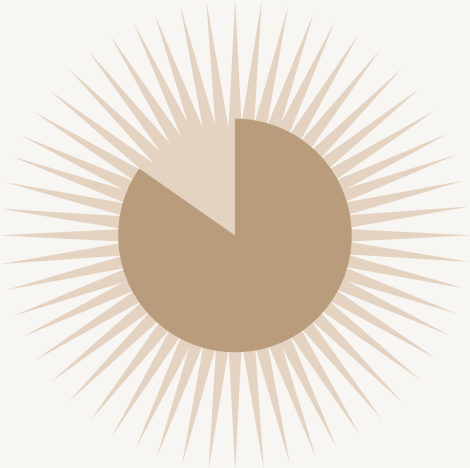
88%

It shows the score persentage of R squere

Feature Engineering

```
Validation R^2 score was: 0.8781979520306361
Feature coefficient results:

carat : 2280.56
cut : 68.08
color : 151.73
clarity : 277.08
depth : -91.96
table : -63.26
x : -6816.43
y : -17.74
z : 2136.68
O : -77.63
G : 670.28
```



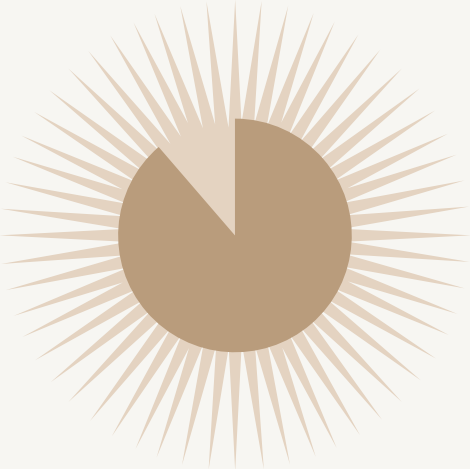
87%

It shows the score persentage of R squere is less than the previous experiment

Feature Engineering (adding new column)

```
Validation R^2 score was: 0.8919634792883033
Feature coefficient results:

carat : 9844.92
cut : 75.08
color : -268.08
clarity : 288.02
depth : -101.37
table : -89.93
x : -308.35
y : -425.07
z : -807.23
shape : 8.14
```



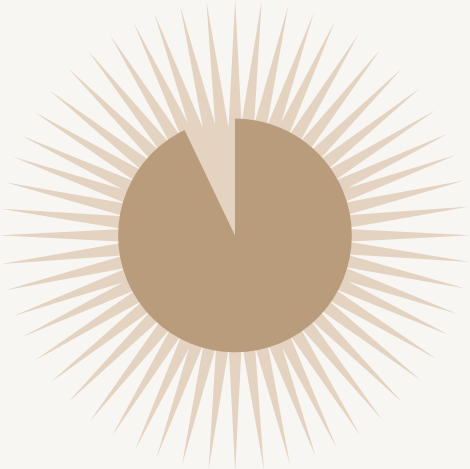
89%

It shows the score persentage of R squere is higher than the previous experiment

Log Experiment

```
Validation R^2 score was: 0.920735981955819
Feature coefficient results:

carat : -1.01
cut : 0.01
color : -0.06
clarity : 0.07
depth : -0.00
table : -0.01
x : 0.92
y : -0.02
z : 0.72
```



92%

It shows the score persentage of R squere is higher than the previous experiment

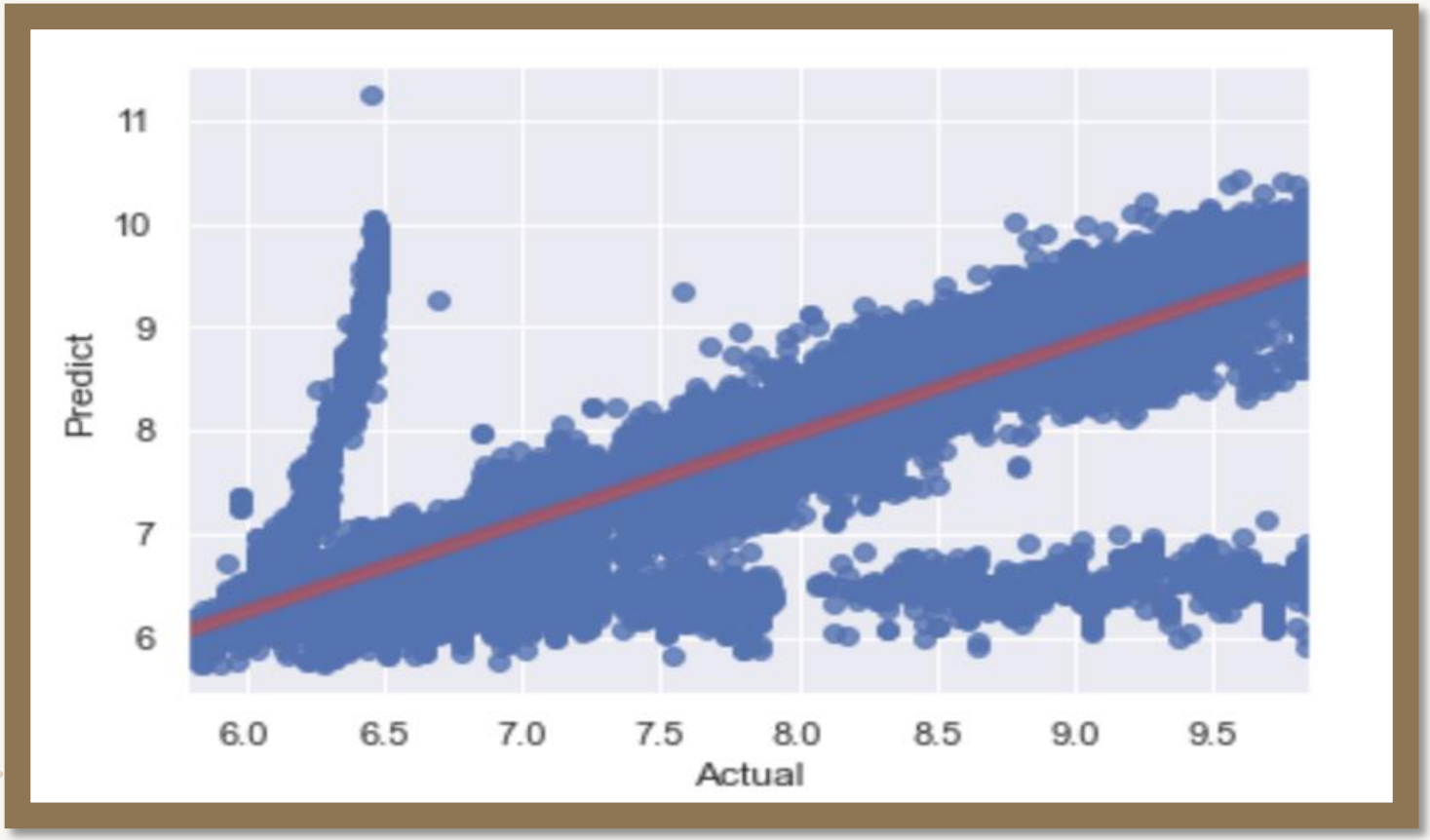
Result

Conclusion

	Predict	Actual
0	6.024215	5.786897
1	5.788030	5.786897
2	5.950484	5.789960
3	6.194198	5.811141
4	6.182943	5.814131
5	5.996040	5.817111
6	5.988527	5.817111
7	5.930407	5.820083
8	6.132733	5.820083
9	5.850855	5.823046
10	6.059965	5.826000
11	5.783058	5.828946
12	5.724973	5.834811
13	6.192540	5.840642
14	5.739189	5.843544

After choosing the best R square result, the prediction result shows that, the actual price is very close to the prediction price

Fit Line Visualization



Thank you