Diamond Cut Evaluation

By Ibrahim Ahmad

01 Introduction

1. Problem definition
2. Dataset properties

Problem definition

Diamond’s quality as a luxury product can be defined by it’s cut, such definition is done be professionals to each diamond which is costly and tedious.

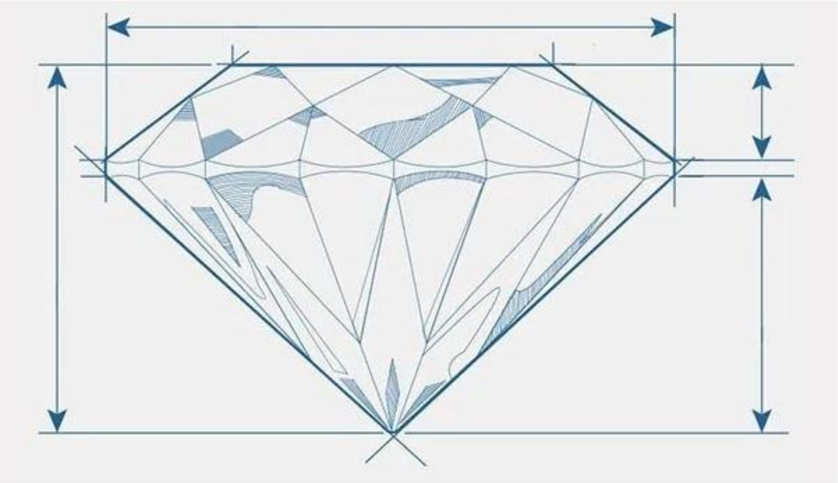
It’s important to be able to classify diamond’s cut based on a pre-existing set of attributes to understand the magnitude and impact of each attribute and to streamline the quality evaluation process.

In this case we will be using Logistic regression to classify the cut based on other scaler attributes.

Dataset Properties

We will be using the Diamonds dataset from Kaggle for our analysis. The dataset contains the attributes of 53940 diamond. While the dataset has many dimensions we will be using the following dimensions for out model:

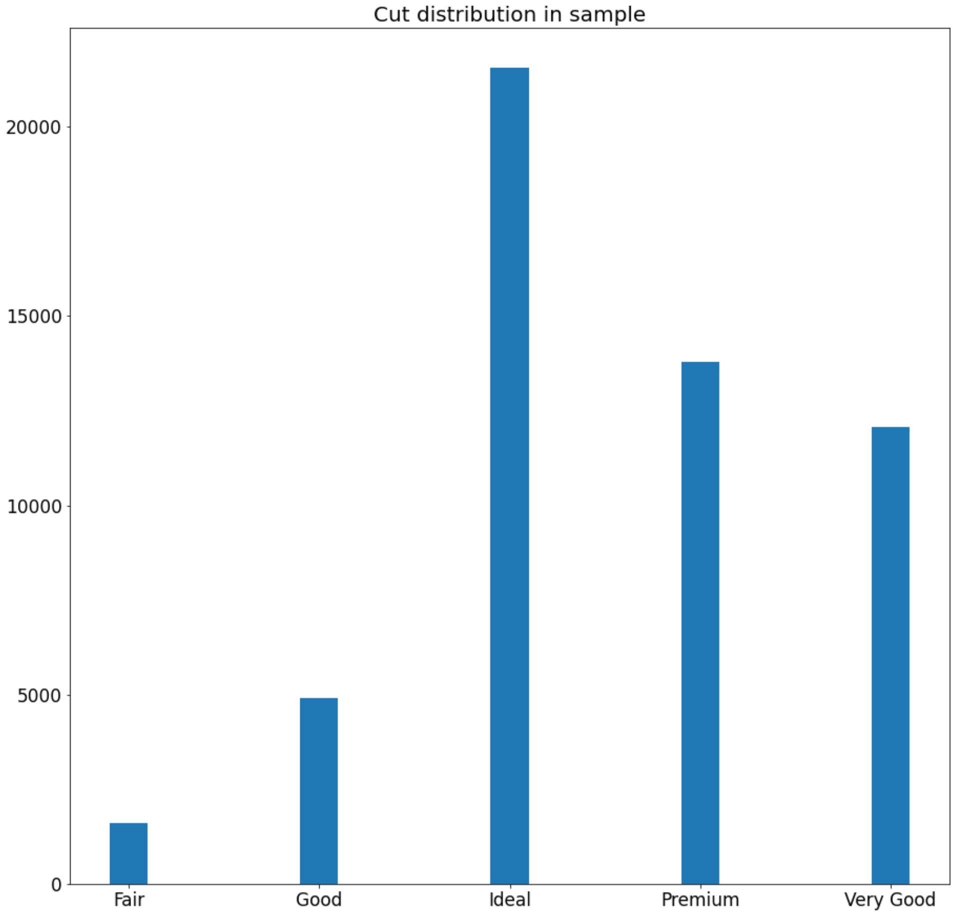
1. *Carat: it's the mass of the diamond using a specialized scale for diamonds and gemstones*
2. *Price: the price of the diamond in USD*
3. *The X, Y and Z coordinates of the diamond: these will allow us to evaluate the size of the diamond.*



Dataset Properties

Examining the dataset we notice that the 'Ideal' and 'Very good' cut represents the majority of the sample size so we randomly choose a sample out of these cuts and discarded the rest to limit the model's bias.

Additionally once we performed the model fitting we found that the chosen properties didn't have strong statistical significance for the 'Premium' cut. So it was removed and the model will only be used for the classification of the other cuts.



02 Model Creation

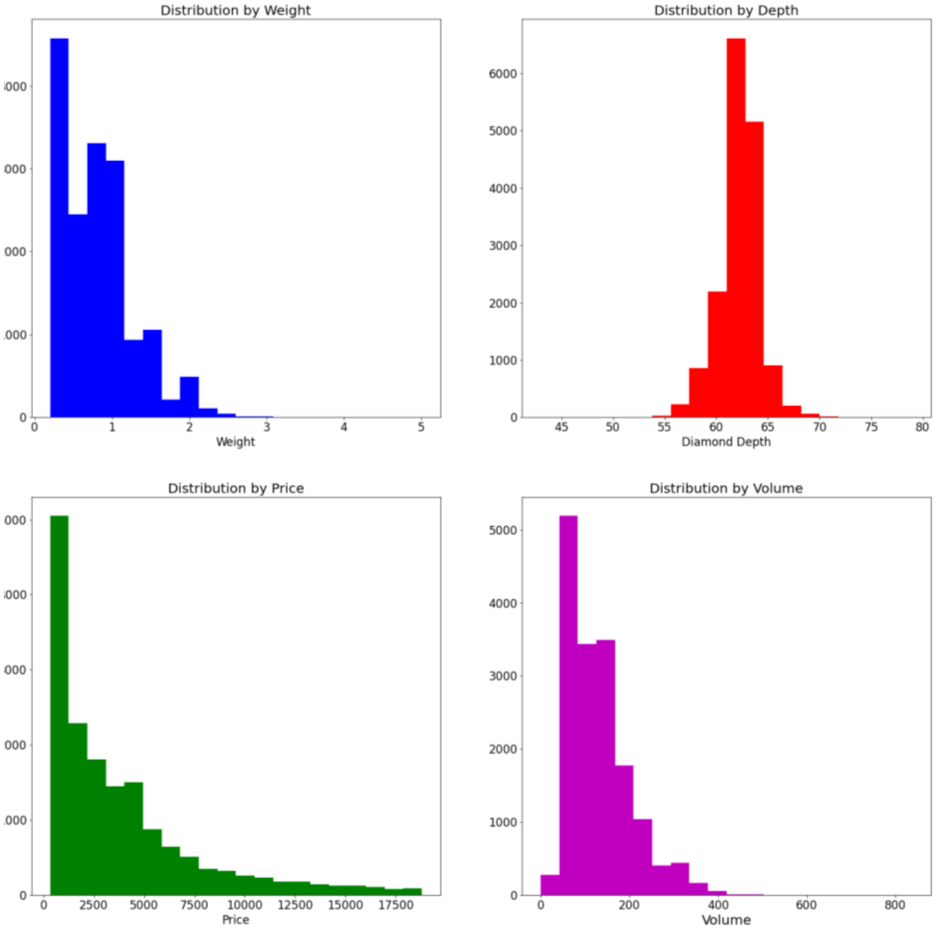
1. Data preparation
2. Model choice and fitting
3. Model’s good of fitness evaluation

Data preparation

We can create the 'Volume' attribute by multiplying the X,Y and Z attribute with each other, this is merely a simplification of the volume measurement process.

No outliers or missing values were found in the dataset.

Examining the distribution of the diamonds properties we notice that the majority of the diamonds in the dataset are small diamonds as such it's properties are heavily right skewed.



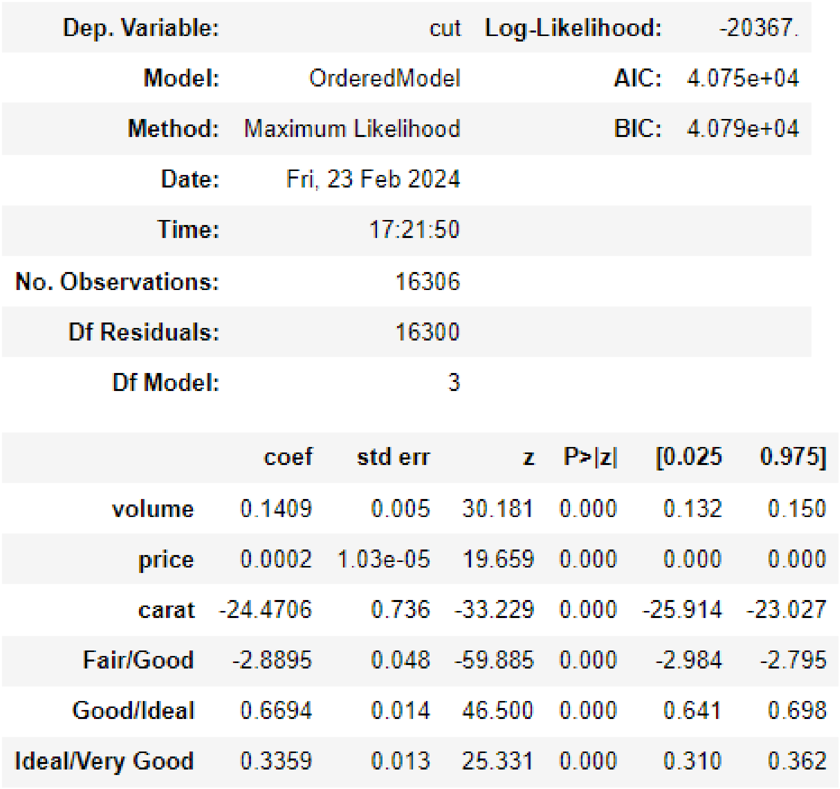
Model Choice

For classification purposes we will use ordinal logistical regression, this model will allow us to predict the ordinal quality of the diamond (Cut ranging from Fair to Very good) from the attributes Carat, Price and Volume.

After creating and fitting the model we can evaluate the impact of each attribute on the cut from the output:

The coef column represents the linear representation of the impact each variable has in respect to it's scale

While the P>|Z| column represents the significance test for each attribute



Dep. Variable:	cut	**Log-Likelihood:**	-20367.
Model:	OrderedModel	**AIC:**	4.075e+04
Method:	Maximum Likelihood	**BIC:**	4.079e+04
Date:	Fri, 23 Feb 2024		
Time:	17:21:50		
No. Observations:	16306		
Df Residuals:	16300		
Df Model:	3		

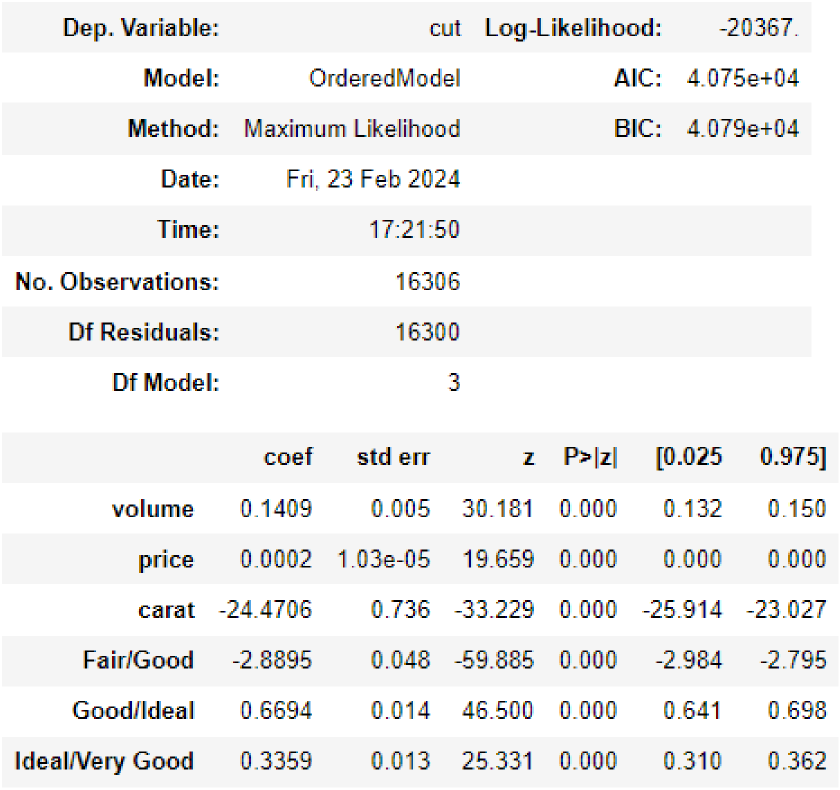
	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
volume	0.1409	0.005	30.181	0.000	0.132	0.150
price	0.0002	1.03e-05	19.659	0.000	0.000	0.000
carat	-24.4706	0.736	-33.229	0.000	-25.914	-23.027
Fair/Good	-2.8895	0.048	-59.885	0.000	-2.984	-2.795
Good/Ideal	0.6694	0.014	46.500	0.000	0.641	0.698
Ideal/Very Good	0.3359	0.013	25.331	0.000	0.310	0.362

Model Choice

We can see that the increase in volume is positively associated with the quality of the diamond while the carat has a negative association this indicates that the shape of the diamond and proportionality has a greater impact on the evaluation process of the cut.

Surprisingly the effect of the price is quite minimal, this indicate that an expensive diamond doesn’t necessitate a high quality diamond.

The zero P-Values indicates that the effect of the attributes is not attributed to chance and they are statistically significant.



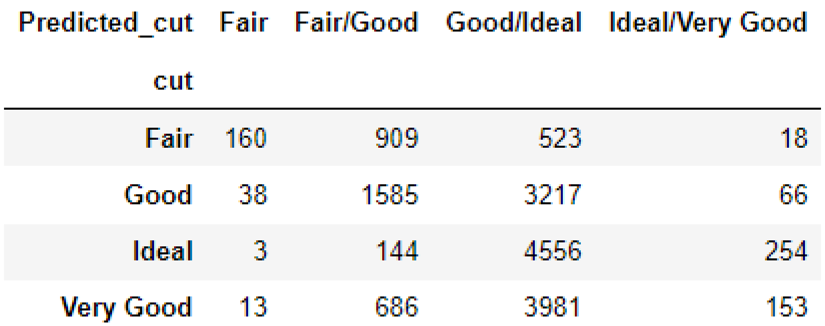
```
   Dep. Variable:                cut  Log-Likelihood:    -20367.
          Model:        OrderedModel             AIC:  4.075e+04
         Method:  Maximum Likelihood             BIC:  4.079e+04
           Date:    Fri, 23 Feb 2024
           Time:            17:21:50
No. Observations:              16306
    Df Residuals:              16300
        Df Model:                  3

                     coef   std err        z  P>|z|   [0.025   0.975]
         volume    0.1409     0.005   30.181  0.000    0.132    0.150
          price    0.0002  1.03e-05   19.659  0.000    0.000    0.000
          carat  -24.4706     0.736  -33.229  0.000  -25.914  -23.027
      Fair/Good   -2.8895     0.048  -59.885  0.000   -2.984   -2.795
     Good/Ideal    0.6694     0.014   46.500  0.000    0.641    0.698
Ideal/Very Good    0.3359     0.013   25.331  0.000    0.310    0.362
```

Model’s Good of fitness evaluation

In order to evaluate the performance of the model we must compare the predicted cut from the model Vs. the actual cut.

Note that Ordinal logistical regression works with the probability that a diamond will fall in a range of quality rather than a strict classification of the cut.

By identifying a strict Cutoff lines in the probability curve we can strictly classify the cut but this will lead to false classification in some cases.

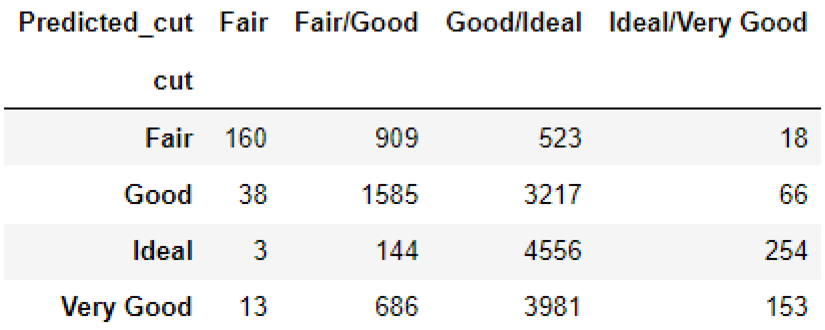


Predicted_cut	Fair	Fair/Good	Good/Ideal	Ideal/Very Good
cut				
Fair	160	909	523	18
Good	38	1585	3217	66
Ideal	3	144	4556	254
Very Good	13	686	3981	153

Model’s Good of fitness evaluation

From the following table we can note that there are some false negatives and positives by the model. As such it’s advised that based on the current parameters the model can be used as a preliminary stage of evaluation rather than a primary method.

It’s worth noting that including more dimensions may increase the accuracy of the model but at the cost of overfitting.



Predicted_cut	Fair	Fair/Good	Good/Ideal	Ideal/Very Good
cut				
Fair	160	909	523	18
Good	38	1585	3217	66
Ideal	3	144	4556	254
Very Good	13	686	3981	153

THANK

YOU