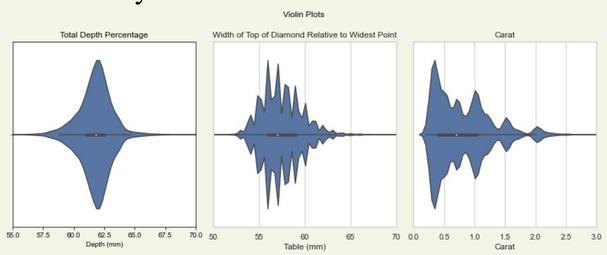
Diamond Pricing Analysis

Samantha Howard

Introduction

This data set contains almost 54,000 diamonds and 10 different features. The *Feature Description* section explains briefly the definition of each feature. The provided notebook goes more in depth, explaining the equations used to obtain carat and depth for clarification of the meaning of those features and to demonstrate the code used to produce the provided figures and results.

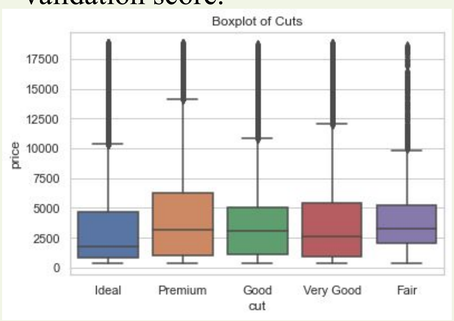
The choice behind this project was to reevaluate a previously used data set from one of my first semesters at Michigan State University in the CMSE department, as a means to demonstrate how my skills have developed over the course of my education.



Methods

For the Machine Learning aspect of this project the data was separated into training and testing data, the test size was 25% of the data and training being the remaining 75%.

Then a pipeline was created to use 5 different machine learning algorithms. These methods where then compared against each other to evaluate what one had the highest cross validation score.



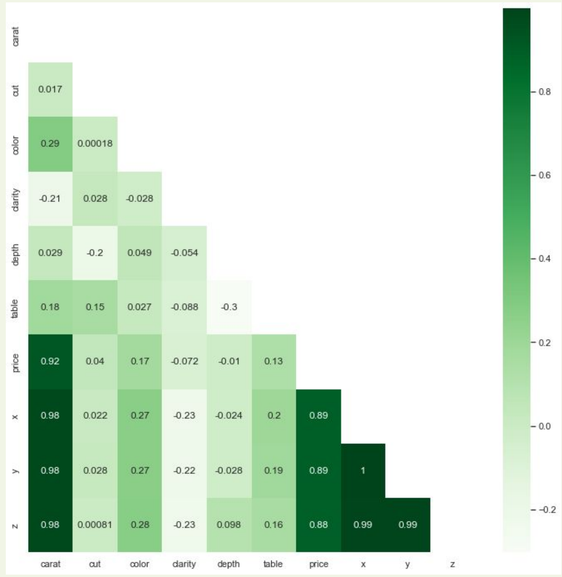
The 5 machine learning algorithms were:

- Linear Regression
- Decision Tree
- Random Forest
- XGB Regressor

Results

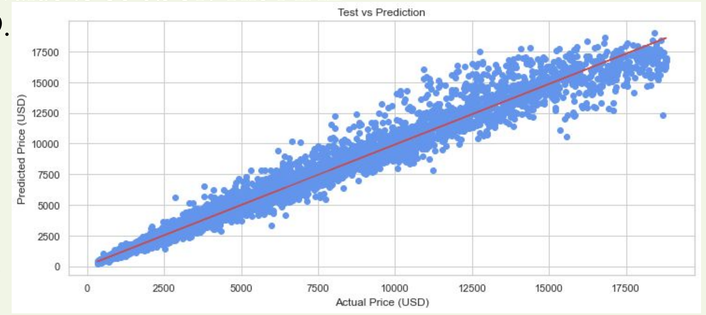
The figure below is a correlation plot that shows the coefficient calculated in relationship with one another. The darker green means there is stronger correlation between the two variables. When looking at the *price* column to see what affects the cost of a diamond the most; carat has a correlation coefficient of 0.92 to price, making that the most impactful to the price.

Notably the dimensions of X, Y and Z have a coefficient of 0.98 to carat. Meaning the larger these dimensions are in a diamond, the heavier it is and thus the larger carat. This is why in terms of price X, Y, and Z have coefficients of 0.88 or higher.



Prediction

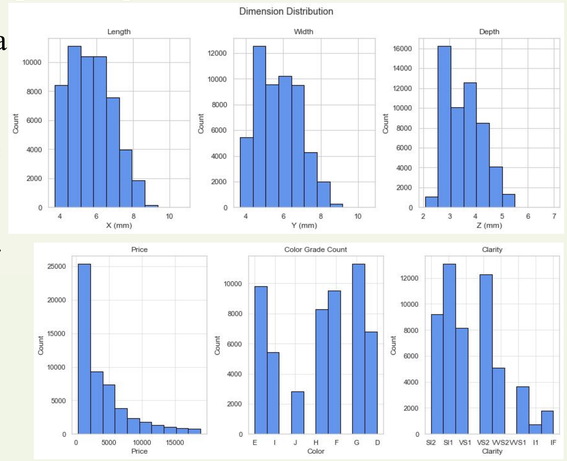
The model that had the highest cross validation score for this scenario was the XGB Regressor. Then calculated the R-squared and adjusted R-squared value to be 0.981. The MAE or Mean Absolute error was 278.09. The RMSE or Root-mean-square deviation was 544.736. To the right is a graph of the data plotted with the actual diamond price being the X-axis and the predicted price by this model being the Y-axis with the red line the line of regression for the data.



Feature Descriptions

Price: displayed in US dollars (USD)
Carat: weight of a given diamond
Cut: quality of the cut
Color: diamond color grade
Clarity: a measurement of how clear the diamond is
X: length in mm
Y: width in mm
Z: depth in mm
Table: width of top of diamond relative to widest point in mm
Depth: total depth percentage

To the right, is a few histograms that show the distribution of the variables. In the price distogram there is a mean of $3,932.80. Where the min price is $326 and with a max of $18,823.00.



In the methods section there is located a boxplot, this boxplot demonstrates that distributions of the *cut* of each diamond, as it can be seen, it is relatively evenly distributed evenly throughout the data.

Python Notebook:

for citations, code and more in-depth explanations

SCAN ME