MMF1922HF - 2022 Fall
Data Science
Mingming (Alice) Yin
Nov, 25th 2022

1. Getting information from the training dataset

Excluding the id Column, we have 10 columns in the dataset, and the target variable is price. The minimum value is 0 for x,y, and z, which seems unreasonable and needs to be fixed.

2. Data preparation

a) I created a new variable, vol, representing the volume of the diamond, calculated as vol = x*y*z. Then, I dropped the columns x, y, and z.

b) For categorical variables: cut, color, and clarity, I created dummy variables that represent the categories of your categorical independent variable. Then, I dropped the original columns of cut, color, and clarity

3. Model Fitting

I split the dataset into a training set and a testing set using a test size of 0.2 and fitted a random forest model with 100 trees. The calculated Root Mean Square Error for the training and testing sets are 228.50 and 237.38, respectively. Also, I've manually checked the results and see that higher carats and larger volumes are associated with higher prices, which looks reasonable in the real world.

4. Predicting

Then, I feed the test data into the model and write the predictions into a CSV file.