Diamond Price Prediction Assignment

November 25th, 2022

Huiwen(Wendy) Zhang

- **Data Preparation**

Diamond price prediction is considered a classic evaluation problem. The given data have main features such as carat, cut, color, and clarity. I start with the data preparation:

1. I removed the outliers of the x, y, and z parameters since those parameters have some extreme values like 0, which is unreasonable.
2. I replaced x,y, and z with a new variable volume equal to the multiplication of x, y, and z.
3. For the categorical variables, such as cut, color, and clarity, I created dummy variables for those variables.

- **Modeling and Evaluation**

I have split the data into train and test sets with a test ratio of 0.33. Then, I applied the random forest with 100 trees for this data set. For Evaluation, I computed the mean absolute error, mean squared log error, and root mean square error. Specifically, the root mean square error is 570.86. Then, I applied this model to the test data set and saved the prediction prices into the results csv file. Besides that, I manually checked the result diamond price, and the results seem reasonable since the diamond with a larger volume and higher carat is usually more valuable.