MMF Data Science 2022: Predicting prices of diamonds

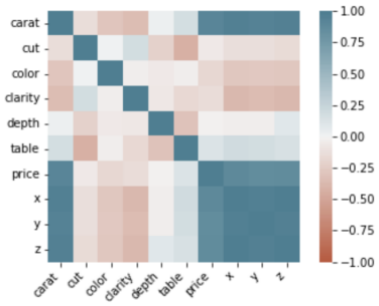
Shaohong Dong

1. Data Preparation

   I import both training and testing data and check for any missing or bad data. There is no missing value. However, there are few bad data in both the training and testing dataset, i.e., x (length), y (width) and z (depth) have 0 values, which does not make sense for a diamond in the real world (except the diamonds are in Metaverse). Since the testing dataset also has such bad data, I think it is better to train my model with the bad data.

2. Feature Engineering

   I apply Ordinal Encoding on “Cut”, “Color” and “Clarity”, since all these categorical features have ranks (Best to Worst) that needs to be encoded with ordinal numbers. In addition, I have checked the correlation between each feature. From the correlation plot below, it is obvious that the x (length), y (width), z (depth), and caret (weight) have correlations. Therefore, besides training the model on all of the features, I have also trained my model on features except x, y, and z.



3. Model Training

   I use the H2O AutoML to train my model on two different training datasets: all features ('carat', 'cut', 'color', 'clarity', 'depth', 'table', 'x', 'y', 'z') and all features except ‘x’, ‘y’ and ‘z’. The AutoML worked very well, and my best models were Stacked Ensemble model.

4. Model Selection and Result

   Comparing the RMSE of my two best models on two different training datasets. The one trained on all features is better. It is probably because we only have 9 training features in total (not much information), and the x (length), y (width), and z (depth) are very important and can provide some information to the model.

5. Future work

   There could be some improvements in feature engineering to somehow reduce the correlation between x (length), y (width), and z (depth) while still keeping their information to train the model.