Data Science Assignment Write-up

Yi Zhou, 1004407545

Nov 20th, 2022

The goal of the model is to predict the price of diamonds according to information on their weight, cut quality, color, clarity, and dimension measurements. The entire dataset contains this information (including price) for around 54,000 diamonds. The dataset is randomly split into two set where about 70% of the data is used to train the model and the remaining 30% is used to test the accuracy of the model.

Before fitting the model, I used one-hot encoding to turn all the ordinal qualitative information (cut, color, and clarity) into number rankings for both train and test datasets. Cut quality is classified into five ratings – Fair, Good, Very Good, Premium, and Ideal – I changed the ratings to 1, 2, 3, 4, 5 respectively. Similarly for color, the classifications of J (worst), I, H, G, F, E, to D (best) are replaced by 1, 2, 3, 4, 5, 6, 7 respectively. For clarity, the ratings from I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, to IF (best) are replaced by 1, 2, 3, 4, 5, 6, 7, 8 respectively. All three numberings are consistent in that smaller numbers denote lower ratings and larger numbers represent better conditions. The independent variables I included in the model are weight (carats), cut (1-5 rating), color (1-7 rating), clarity (1-8 rating), depth (percentage out of 100), and table (percentage out of 100). Depth is the depth percentage calculated by depth divided by the mean of width and length of the diamond. Table is the width of the top of the diamond divided by the width of the widest point of the diamond.

I start with working with the train.csv data file only. I first used the LinearRegression function from the sklearn linear\_model package to fit a multilinear regression model with the training dataset. However, prediction results from a simple multilinear regression model have very large mean squared errors. So I used the PolynomialFeatures function from the sklearn preprocessing package to fit higher degree regressions. I tried fitting the training set to 2 to 5 degrees polynomial regression models. I decided to use a third-degree polynomial regression model because there were significant error reductions with the degrees changed from 1 to 2 to 3. There is still reduction in MSE from 3 to 4 to 5, but the reductions were less significant, and to not overfit the model I used third degree.

Next, I read in the test.csv data file and use the above model to predict diamond prices based on the same independent variables but from the test dataset. Then I save the predicted results in a separate csv file.