MMF Data Science

Diamond

Zhe Dong

Nov 22, 2022­­­­­­­­

1. Evaluating the problem

In this Kaggle project, we are presented with prices and data on features of diamonds, and

aim to predict diamond prices using data on their features.

The features involved are 6 continuous independents - x, y, z, carat, depth, table;

And 3 categorical independents – cut, color, and clarity.

The target is price.

1. Extracting data
2. Opened csv files and checked for number of sheets
3. Used open\_csv() to import both train and test data files

­

1. Data cleaning
2. Used shape, head(), info() to look at the names of the columns and dropped one column that contains only id numbers for the diamonds that should not be relevant to their prices.
3. Used describe() to check for zeros in x, y, z and drop them, since dimension of a diamond must be positive to make sense.
4. Used pairplot() and regplot() function to plot scatter plots of every two features to identify correlations and look for any outliers. I got rid of significant outliers in x, y, z, and reran the pairplots to double check.
5. Fitting and get test score
6. Used fit\_transform() fuction to apply label encoder to all columns for categorical data.
7. I used three methods, constructed pipelines for them respectively and compared the scores for my test set after doing standard scaling. The methods used are linear regression, decision tree, and random forest. Comparing the RMSE’s of these methods, I decided to use random forest which gave the lowest RMSE to be used for predicting the prices for the real test data.
8. Save model and apply on test set
9. Because the test data is separate from the train data already, separating our train data into smaller train and test data would limit the model performance by limiting the model’s access to all training data. So after confirming my model works, I alter my codes from using 80% of training data to using all training data, redid the model and saved it using pickle.dump(). Finally, load my model and apply to the test data (first column which was also just id was dropped) to get output of prices.
10. Manually checked number of results in csv file and it matches the number of rows in test file. Looking at diamond with higher carats and sizes, bigger diamonds are generally pricier, which makes sense in the real world.