Min Jae (Ian) Lee

Data Science 1922 – Final Assignment

Objective

* To analyze what kind of variables or factors influence diamonds’ prices.
* To determine the most accurate and suitable model for this analysis.

Data Understanding, Data Management, and Feature Engineering Methods:

* Explanatory analysis: Before handling data or fitting a model to data, it is important look how data is collected. Data types, columns, rows, and descriptive statistics are good ways to check and understand the data. For instance, there are 10 different variables with different types such as float, integer, and object. Also, since we have our objective, we can use “price” as a dependent variable. It is very useful to see correlations between variables.
* Imputation: Based on data information, there are no Null , NaN, or missing values in the data. However, there exists 0 values in “x”, “y”, “z” columns, so they are dropped from dataset.
* Outliers: By using pairplots, we can also identify some outliers in the dataset. We can also use boxplots, scatterplots or heatmap to identify possible outliers. These outliers are removed to increase accuracy. (It is probably not a good idea so remove every outlier in the dataset because it can cause overfitting, so only obvious outliers are removed in this project.)
* Feature Engineering: Hot-Encoding: there are three variables need to be converted to numerical numbers: cut, color, and clarity. They are categorical variables. For the “cut” variable, cut has 'Good', 'Very Good', 'Fair', 'Ideal','Premium', so we convert them into numerical numbers or orders. For example, ({'Good':1, 'Very Good':2, 'Fair':3, 'Ideal':4,'Premium':5}). “color” and “clarity” are also converted in the same methods.
* In addition, “x”, “y”, “z” columns look like they are related to caret size. Thus, I use them to create two additional variables as volume and surface which show some linear relationship with caret.

Model Selection

* For machine learning, to train the model, it is necessary to split the dataset into a train set and a test set. In this project, 80% of the dataset is used as a train set and 20% are used as a test set.
* We can run linear regression, but high RMSE suggests that the linear regression approach does not capture importance of the dataset.
* We can use other machine learning techniques such as decision tree, random forest regression and XBG regression are used to lower RMSE.

Evaluation

* Among many different machine learning techniques, random forest regression and XBG regression provided the lowest RMSE. In this project, random forest regression is used to train and test the model because it provides the lowest RSME. Thus, this suggests that random forest regression provides more accurate prediction for diamond’s prices.