Joshua (Ha Rim) Kim
1004391339

MMF1922H – Course Project Report

The aim of the report is to present how I employed machine learning techniques to predict diamond prices. The discussion is primarily on two topics: data processing and model selection.

Data Processing

Although the provided datasets were already well-structured in tables, well-formatted in CSV files, and reasonably clean without any missing values, there were still improvements which could be applied to ensure the quality of the data before training a prediction model:

1. *faulty data removal*: one could simply remove data points with zero "x", "y", or "z" values as a typical diamond would have non-zero measurements in any of the three dimensions
2. *outlier removal*: by observing the pairplot (generated using Python *seaborn* package) of the numeric variables, I detected and removed the following outliers which deviate significantly from the general trends between features:
    a. data points containing values greater than 30 in any dimensions or less than 2 in "z"
    b. data points with values greater than 75 or less than 45 in "depth"
    c. data points with values greater than 80 or less than 40 in "table"
3. *duplicate data removal*: there were 97 duplicate rows in the training data which were removed to prevent the model from overfitting to the repeated examples
4. *ordinal encoding*: since all three categorical variables ("cut", "color", and "clarity") have inherent orderings, I applied ordinal encoding by mapping categories to their corresponding integers

Model Selection

Following the data preparation phase, the model selection process was facilitated by *H2O*, a Python package offering an open-source automated machine learning framework, i.e. AutoML. Since the project boils down to a classic regression problem using tabular data, a heavily studied topic in the field of supervised machine learning, such task is where AutoML excels. The pool of machine learning algorithms from which *H2O* AutoML selects are:

- generalized linear model with regularization (GLM)
- distributed random forest (DRF)
- extremely randomized trees (XRT)
- gradient boosting machine (GBM)
- extreme gradient boosting (XGBoost)
- fully-connected multi-layer artificial neural network
- stacked ensemble of all the base models
- stacked ensemble using subset of the base models

A convenient feature of *H2O* AutoML is the hyperparameter optimization using random grid search. The library documentation lists the hyperparameters of each base model, along with all potential values that can be randomly chosen in the search. I ran the experiment for 5 hours, followed by the model selection stage where the AutoML ranks the trained models based on 5-fold cross-validation RMSE. Assuming the cross-validation score translates to performance on the test dataset, I chose the model ranked first by the AutoML as my final algorithm. The final model is a stacked ensemble using the best version of each base models with training RMSE of 373.45, cross-validation RMSE of 509.05, and test RMSE (calculated on Kaggle) of 497.57.