MMF1922HF - 2022 Fall – Final Project

Harry (Zhihao) Zuo

1003213088

This is a 1-page write-up for the description and methodology used in the project.

**Project Objective**

The purpose of the project is the evaluate prices of diamonds.

**Data Preparation**

The Data Step starts with searching for bad data points, including missing data, unreasonable values, outliers, etc. Then, one-hot encoding is used to convert few categorical variables into numerical ones. With a 2:8 ratio, data is split into test and training sets.

**Model Building & Evaluation**

The RMSE (Root Mean Square Error) is chosen to evaluate model performance.

Since this is a regression problem, linear regression is the first model to try. After training, fitting model on the test set gives an RMSE of more than 1200, which suggests linear regression might not be a good model. This aligns with my expectation that in practice, linear regression does not usually yield satisfactory results, as it assumes a linear dependence between variables.

Then, a more complicated ensemble method - XGBoost - is chosen as the model. It is a distributed gradient-boosted tree machine learning algorithm, predicting targets by combining the results from sets of weaker models, without assuming any linear dependence.

Using the default XGBRegressor and setting n\_estimators to be 1000, the trained model has an RMSE of 542 on the test set.

**Final Result**

With the same parameters, the model is trained over all the data from “train.csv”, and the scores form the “test.csv” are predicted, eventually yielding an RMSE of 524, a decent score that beats the benchmark.