Garrett McNally

MMF1922HF - Data Science

Course Project

The goal of the course project is to take a dataset of diamond specifications including physical dimensions, colour and clarity to construct a model to predict price. The approach applied was to loosely follow the steps of the CRISP-DM model. Steps taken were as follows:

Data and Business Understanding

The first step was to understand how variables contained in the dataset were represented, how they might and would affect price and how they interact with one another. In researching diamonds, both colour, cut and clarity were characteristics where there was a clear preference ordering. A diamonds "table" and "depth" were both manipulations of the height, width and length of the diamond. Otherwise in general, the larger the diamond by either physical size or weight, intuitively should see an increase in price. This could be seen in plotting each variable vs the price.

Data Preparation and Modelling

After understanding the variables within the dataset, and seeing their plots there were multiple steps applied to attempt to prep the data. First, an ordinal scale was applied to the colour, cut and clarity of the diamonds based on the knowledge gained from the research in the last section. Attempting to model at this stage was naïve, as many steps were missing from the data preparation despite having checked for any missing or NA values. Using this first approach with a basic OLS model yielded a RMSE score of 5128.17802. By going back and analysing the data once more, it was clear that there were outliers present in the data. By restricting the domain to remove these data points, a new OLS model was made. This did not prove efficient enough and a new model was selected, Gradient Boosted Trees.

Evaluation

After the first submission, the RMSE was calculated on subsequent models to give an estimate on how the model might perform, with an increase in error to be expected when used on the testing set. By using this as a proxy for the test set RMSE, the model could be iterated and improved to attempt to achieve a better score. The OLS model with the data prepped was scoring over 1200 RMSE and as such a new model using GB Trees was chosen and was able to score ~380 on the training set. This was used for the submission which beat the benchmark with a final RMSE of ~540. This achieve the model objectives as the only requirement was to beat the benchmark.

Deployment

To deploy the solution, the data manipulation used to calculate the model was recreated to apply the same model once more. After this, the predict price scores were calculated and then written to CSV for submission. If this tool was required for a customer, the model would need to be implemented into a user-friendly software and it would be most effective if more data points were added over time as they became available, and the model is re-trained.