Hanhua Zhang

This assignment is designed and asks for price prediction on diamonds base on specific parameters using machine learning models.

I used python packages like pandas, numpy, math, sklearn, lightgbm, xgboost. I encountered a problem as MAC seems like having problem with lightgbm installed by pip and I connot run it. I solved it by install another one using homebrew. I am given 2 sets of data, one is the training set that includes the price for the machine learning model to study, the other is the test set that does not include the price. I first input the 2 csv files onto a notebook then create a copy so I won’t make any change to the original data. Then I discovered that there is an unnamed parameter which is id, so I added it.

I then discovered that some parameters like cut, color and clarity are strings, I used builtin function .map to convert them into integers. Then I combined the 2 sets using an extra parameter “train”, I let it = 1 then it is the training sets, otherwise it is the test sets. I clean the data afterwards by dropping the price section as the test target, then re-seperate the 2 data sets. Finally I use train\_test\_split from sklearn to have my xtrain, train, xtest, test for model price prediction.

I use two machine learning models, xgbregressor and lgbmregressor, this two are similar in use while lgbm generally take shorter time to produce similar result. I combined the prediction together with 0.3 of xgb and 0.7 of lgbm to obtain my final result.

I appreciate my instructor Charles Tsang for all the knowledges, I acknowledge and thank my classmates for the logical discussion. Below are some references to my work.

<https://stackoverflow.com/questions/44937698/lightgbm-oserror-library-not-loaded>

<https://thecleverprogrammer.com/2022/09/26/diamond-price-analysis-using-python/>

<https://github.com/search?q=mmf1922>