CH2013 Lab - ML Assignment

Nohan Joemon
CH19B072

The analysis (step-by-step):

1: Reading the csv file and removing the column 'id' since 'id' is just for numbering and won't be useful for our analysis

2: Normalization of the dataset - Normalization of the dataset is important in ML because some features are numerically much higher than other features (eg: 'LotArea' values are very high compared to 'OverallQual', 'OverallCond')

3: Splitting the dataset into train, test and validation.
(Reason: If we use the entire dataset for training and predictions and use the accuracy of these predictions to report the best combination, our report can be wrong due to overfitting. Instead, we use different datasets for training and predicting)

4: Finding the best combination
Procedure:
We need to decide which 3-variable combination is the best.

(a) We consider all 3-variable combinations, one-by-one
(b) For each combination, we extract the corresponding features from train, validation and test data:
    (i) We train the model on the train data
    (ii) We use the trained model to make predictions on the validation data and report the accuracy of these predictions (using MSE)
(c) We choose the combination that gives the lowest MSE on validation data
(d) Finally, we use the test data to report the test accuracy of the best combination.

Best Combination of features:

features 2,4 and 5: (OverallQual, TotRmsAbvGrd, GarageArea)
(Note: Due to randomness of splitting, the best combination can sometimes change, but on repeated running, this combination was produced most of the time. 1,2,5 is another combination that appeared for some runs)

MSE of test data: (minor changes due to random splitting)

0.4866 (Note that this was taken on the normalized data, for a particular run)

Other results:

Refer to the code (attached as a separate Matlab file)