FODS assignment-2

How we preprocess data:-

We replaced all the Nan values in the data with the mean value of the respective column

We normalized the data using z-score normalization

Here we are given 13 parameters of which we have to determine which features are more reliable to predict the target value

Here we consider generating all the possible subsets and determine the subset with the least error but this is computationally expensive so we consider 2 greedy approaches

Where we consider the best feature addition possible to the features which we are already considering and add it to the features which we currently consider(This is called a greedy forward selection)

Also, we can initially consider all the features as reliable and keep on deleting the feature that is currently the least reliable (This is called greedy backward elimination)

Note that in both approaches, we used gradient descent with no regularization for estimating the linear regression models in each iteration.

Features selected from the greedy forward selection is	Sqft_living Grade View Condition Sqft_basement Waterfront bathrooms

Features selected from the greedy backward elimination is	Sqft_living15 Grade View Condition Sqft_basement Waterfront Bathrooms sqft_above

Errors calculated in each approach

We used root mean square error in our model

Model	Training error	Testing error
Greedy forward selection	381.18	111.286
Greedy backward elimination	378.24	100.45
Error without preprocessing and feature selection	64458334953333.95	19706576803048.867