**CS F320 - FOUNDATIONS OF DATA SCIENCE**

**Semester I, 2021-2022**



**BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI HYDERABAD CAMPUS**

# ASSIGNMENT-2

S.Varshith Reddy - 2018B3A70022H

Manideep Bandaru - 2019A7PS0016H

**Dataset:**

The given dataset has 14 columns namely

1. Bedrooms
2. Bathrooms
3. Sqft-living
4. Sqft-lot
5. floors
6. waterfront
7. view
8. condition
9. grade
10. Sqft-above
11. Sqft-basement
12. Sqft-living
13. Sqft-lot15
14. Price

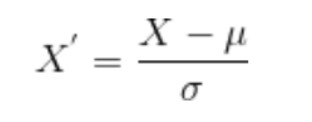
After Pre-processing we get 1164 examples from 1188 examples out of which we take 815 examples for training and 349 for testing. We aim to predict the Price using the 13 features mentioned above.

**Data Pre-Processing:**

The following pre-processing techniques have been used to clean and pre-process the data.

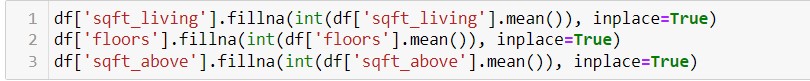
1. **Standardization:**

We standardized the data by columns. This is a scaling technique which is helpful in dealing with high valued data as the data gets centred around zero with mean and unit variance.



1. **Missing values:**

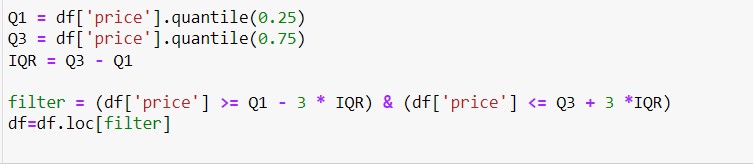
After standardizing we found some missing values and we replaced them with the mean of the column.



1. **Detecting outliers:**

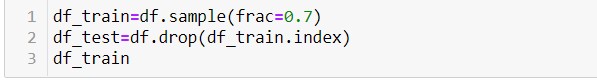
Machine learning algorithms' training processes can be distorted and misled by outliers in input data, resulting in longer training times, less accurate models, and ultimately inferior results.

We used Inter-quantile range to deal with the outliers.



1. **Data shuffling and splitting:**

Data is shuffled and split in the ratio of 70:30 which is used for training the model and testing the model respectively.



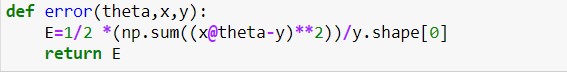
**Model and algorithms:**

Linear regression with gradient descent is used to predict the output price. Both Greedy Forward feature selection and Greedy backward feature selection is used to find out optimal features that help to predict the output. After getting the model for predicting the output, we run the test data that we obtained by 70:30 split and compare both training and testing errors.

1. **Error function:**

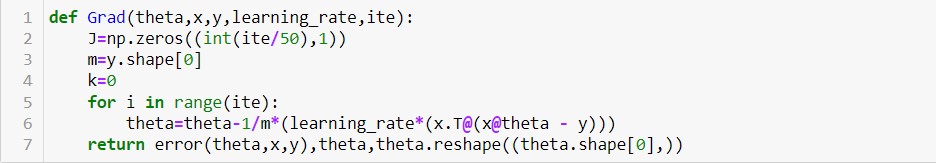
The error function we used to calculate the errors is given below:

Half of mean squared error with respect to our predictions is calculated. Here ‘theta’ is the parameters matrix, ‘x’ is the input matrix, y is the expected output matrix.



1. **Gradient descent:**

We minimize the error function by updating the parameters after every iteration as shown below. In this we take all the parameters in the partial derivation.



1. **Greedy forward feature selection:**

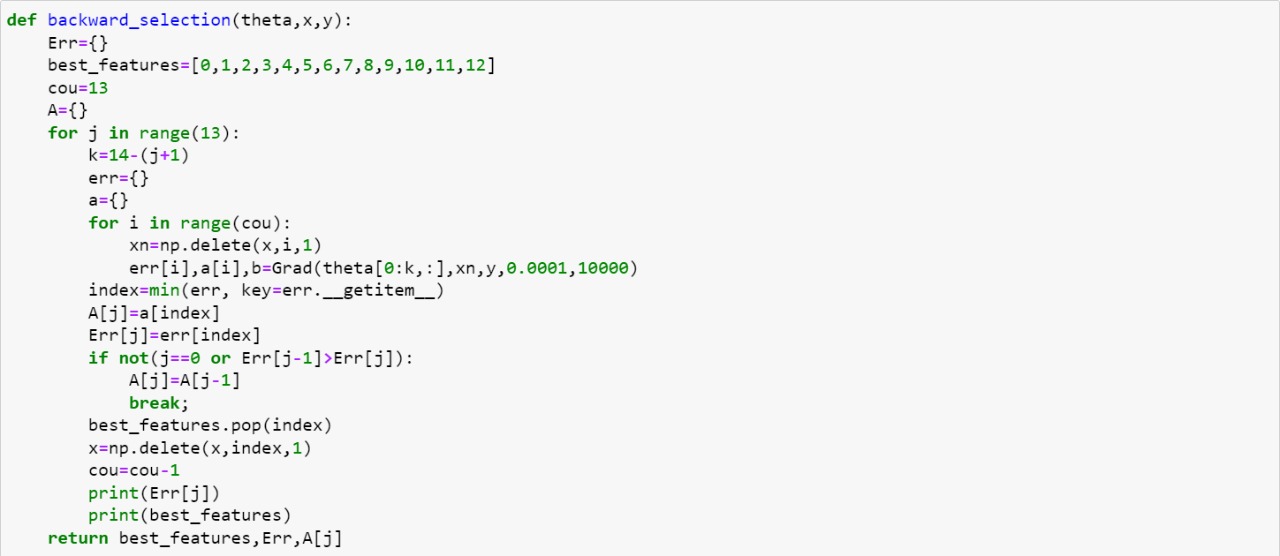
This algorithm uses elimination to select the best features available for the algorithm. We start with 1 feature and select the best feature that gives the least error and then take two features with the above selected one and get the best two features and we continue until the error of the previous iteration is lesser than the current iteration. The same concept is coded below:



1. **Greedy backward feature selection:**

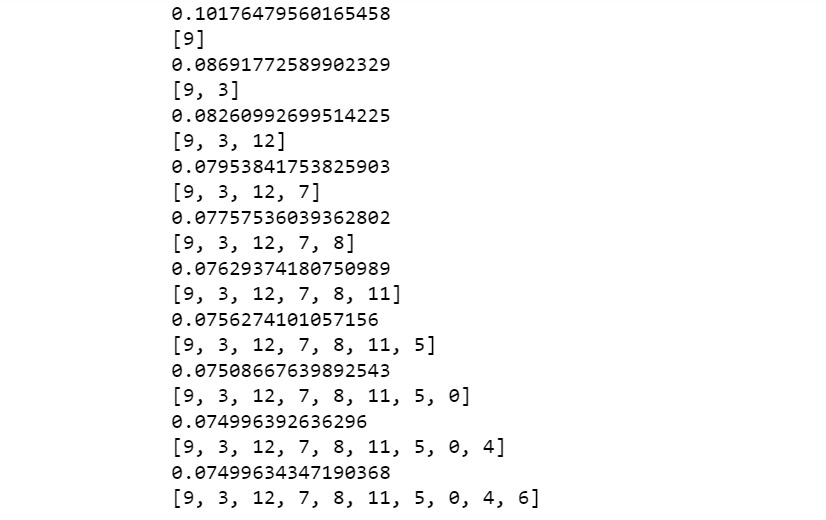
This algorithm is similar to forward feature selection and uses elimination to select the best features. In this algorithm we start by leaving one feature and training the data. The model with least error is selected and the feature which is left in the model is eliminated and we continue the same for the remaining features.

The code for the algorithm above is shown below:



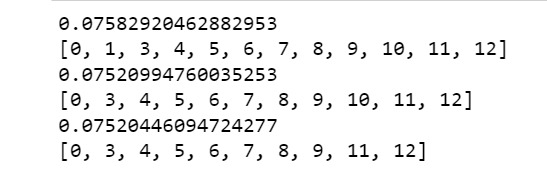
**Outputs obtained:**

We printed the output after each iteration in the forward and backward feature selection methods. The outputs are given as follows: For forward feature selection:



Here 0 corresponds to the first feature and 12 corresponds to the 13th feature similarly. The algorithm gave us 10 best features and eliminated 3 features. The eliminated features are Bathrooms, sqft-living and waterfront.

For backward feature selection:



The algorithm gave us 11 features and features and eliminated bathrooms and Sqft-living.

**Comparison of Best Models:**

Here non-pre-processed data contains the removed rows which have missing values because otherwise we would get no output and the output will be NaN.

Best model in forward feature selection had: 10 features which are -Bedrooms, sqft-lot, floors, waterfront, view, condition, grade, Sqft-basement, Sqft-living, Sqft-lot15

Best model in backward feature selection had:

1. features which are -Bedrooms, sqft-lot, floors, waterfront, view, condition,grade, Sqft-basement, Sqft-living, Sqft-lot15

**Comparison**

