CS F320 ASSIGNMENT-2

Group member	ID
Anjan Neelisetty	2019A8PS0367H
Praneet Sai Madhu Surabhi	2019A7PS0060H
Yarramsetty Sanjeeva Sai Preetham	2019A3PS0485H

Under the supervision of

Prof. N.L. Bhanu Murthy

SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS OF

CS F320: Foundations of Data Science

Assignment-2

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI (RAJASTHAN)

HYDERABAD CAMPUS

(December 2021)

CS F320 Assignment 2 Report

Linear Regression:

A linear relationship between a goal and one or more predictors can be found using linear regression. In the case of one predictor, it is known as simple regression if more than one predictor then it is known as multiple linear regression.

Preprocessing:

The given dataset has 13 column features named:
bedrooms bathrooms sqft_living sqft_lot floors waterfront view condition gradesqft_above sqft_basement sqft_living15 sqft_lot15 to predict the price and around 1188 rows of data. Then we create a random 70-30 split to aid in training and testing respectively. Data is standardized using the mean and variance method. This shuffled and normalised dataset then we perform further

1. Greedy forward feature selection
2. Greedy backward feature selection
3. Linear regression model without any pre-processing and feature selection.

Details about methods used

We have filled in the missing values, standardized the data, removed the outliers, split the data and performed Greedy forward feature selection and Greedy backward feature selection then we performed the linear regression model without pre-processing and feature selection. We have used the Batch gradient during linear regression

Batch Gradient Descent:

To perform a single step in Batch Gradient Descent, all of the training data is taken into account. We take the average of all of the training samples' gradients and

utilise that average gradient to update our parameters. So that's just one epoch's worth of gradient descent.

Batch Gradient Descent is ideal for error manifolds that are convex or somewhat smooth. Because we're averaging over all of the gradients of training data for a single step, the graph of cost vs. epochs is likewise fairly smooth. Over the epochs, the cost gradually decreases.

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

Standardization

Standardisation is a scaling technique that centres the results around the mean with a unit standard deviation. This means that the attribute's mean becomes zero, and the resulting distribution has a standard deviation of one unit.

$$X' = \frac{X - \mu}{\sigma}$$

Splitting and shuffling the data

We had shuffled and split the data in a ratio of 7:3 using the help of .sample function

```
a.sample(frac=0.7)
```

Handling Missing values

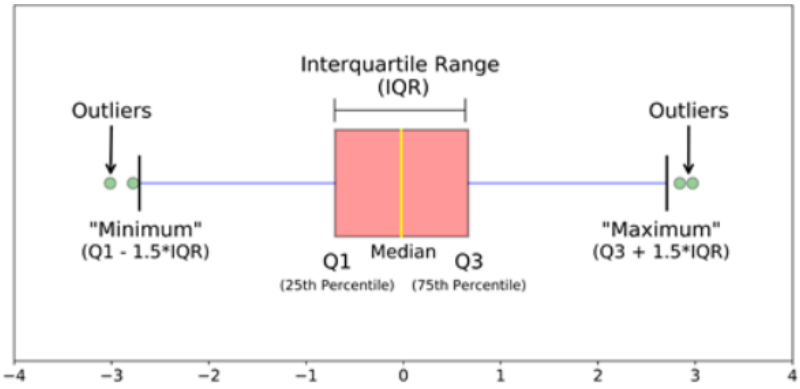
We have handled this issue by filling the mean values in the place of missing values

```
data['Age'].replace(np.NaN , data['Age'].mean())
```

Removing outliers

Outliers in input data can distort and mislead machine learning algorithms' training processes, resulting in longer training durations, less accurate models, and ultimately inferior outcomes.

We have handled this using Interquartile Range(IQR)



Greedy forward feature selection

Forward selection is an iterative technique in which no feature is included in the model at the start. We keep adding the feature that best improves our model in each iteration until adding a new variable does not increase the model's performance.

Greedy backward feature selection

Backward elimination begins with all of the features and eliminates the least significant feature at each iteration, improving the model's performance. We repeat this process until no improvement is noticed when characteristics are removed

Our Results

The subset of features which we got in Greedy forward feature selection are:

```
[9, 3, 7, 12, 8, 11, 10, 0, 6, 5]
```

Which are sqft_above, sqft_lot, condition, sqft_lot15, grade, sqft_living, sqft_basement, bedrooms, view, waterfront,.

The subset of features which we got in Greedy backward feature selection are:

```
[0, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]
```

Which are bedrooms, sqft_lot, floors, waterfront, view, condition, grade, sqft_above, sqft_basement, sqft_living, sqft_lot15.

Training and Testing error of feature selections and linear regression without preprocessing and feature selection

Method	Training error	Testing error
Linear regression without preprocessing and feature selection	NaN	NaN
Linear regression with filling NaN values	36901758826.93145	34056581134.96324
Greedy forward feature selection with pre-preprocessing	0.08994541686113564	0.12588324034000875
Greedy backward feature selection with pre-processing	0.09003832169101747	0.08827146946550876