A Report

On

Forward and Backward Feature Subset Selection using Greedy Approach

BY

Aaryan Gupta	**2018B1A70775H**
Abhinava M	**2018A8PS0844H**
Shubham Singla	**2019A3PS0392H**

Under the supervision of

Dr. 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)

Introduction

In this assignment, we use linear regression to predict house prices using the other 13 attributes (bedrooms, bathrooms, sqft_living etc.) of the dataset.

The steps followed are:-

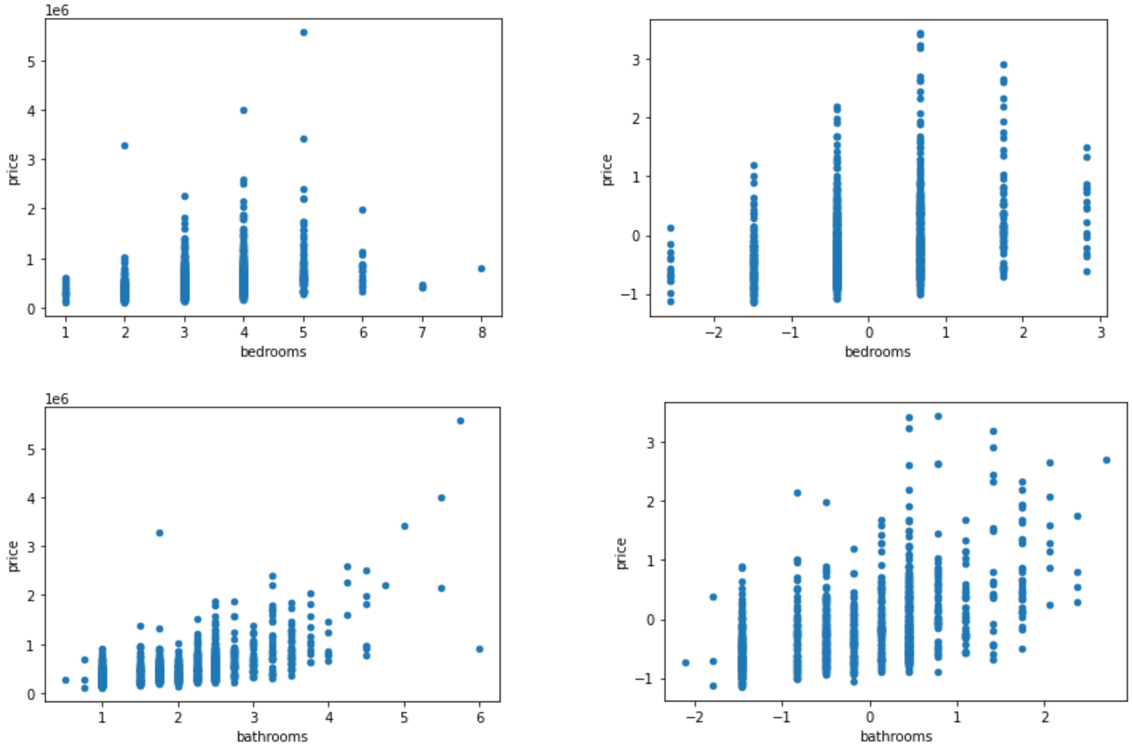
1. Data Preprocessing using techniques like standardization, normalization, detecting outliers and handling missing values.
2. Performing greedy forward and backward feature selection to find the subset of features that provide the optimal regression model and to find the minimum training and testing error of the optimal model obtained.

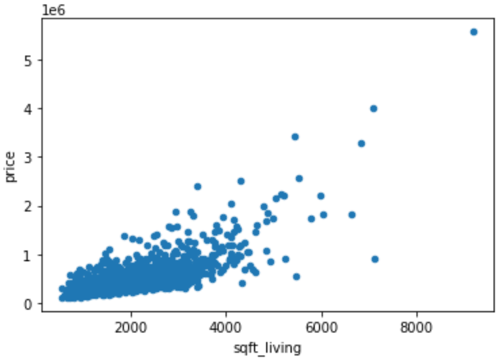
The dataset has 1188 rows and 14 columns, in which 1-13 are features and the 14th column is for the price.

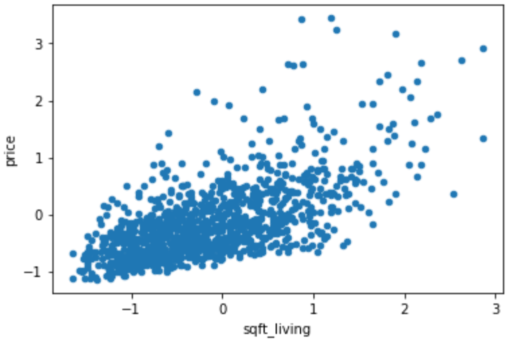
Removal of outliers: An outlier is an observation that diverges from well-structured data. We removed all outliers with z score >3 that removes outliers not in the 99.7% range, there were 107 outliers.

Fig1. (Scatter Plot Price vs features)

Raw Data **Preprocessed Data**







The z-score method for removal of outliers:

For each observation (Xn), it is measured how many standard deviations the data point is away from its mean ($\bar{X}$). Following a common rule of thumb, if z > C, where C is usually set to 3, the observation is marked as an outlier. This rule stems from the fact that if a variable is normally distributed, 99.7% of all data points are located 3 standard deviations around the mean rule of thumb, if z > C, where C is usually set to 3, the observation is marked as an outlier. This rule stems from the fact that if a variable is normally distributed, 99.7% of all data points are located 3 standard deviations around the mean

$$z_n = \frac{X_n - \overline{X}}{SD_X}$$

Normalization: Mean normalization was performed, which is a way to implement Feature Scaling by calculating and subtracting the mean for every feature and dividing this value by the standard deviation.

Handling Missing Values: All the missing values were replaced with mean.

A. Greedy Forward Feature Selection Algorithm

It is a feature selection method which:

1. **Begins** with a model that contains no features(called the *Null Model*)
2. The most important feature S1 = fi is selected first using some criterion.
3. Then pairs of features are formed with fi and the best pair is selected as S2 = {fi, fi}.
4. Set of 3 features are formed using S2 and the best set of 3 features is selected as S3 = {fi, fj, fk} so on..
5. This process is repeated until a predefined number of features are selected.

Disadvantage of Greedy forward feature selection algorithm:

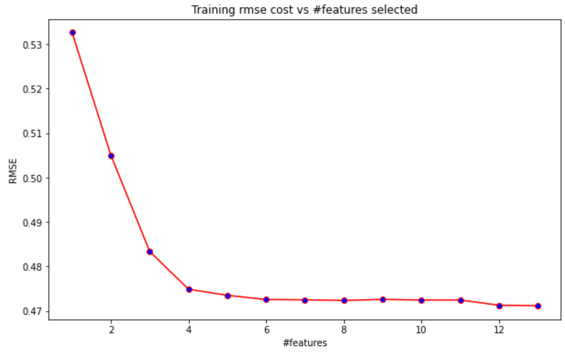
The new features are added continuously in the selected features set. It does not give flexibility to remove the features that have been already added in case they have become obsolete after the addition of new features.

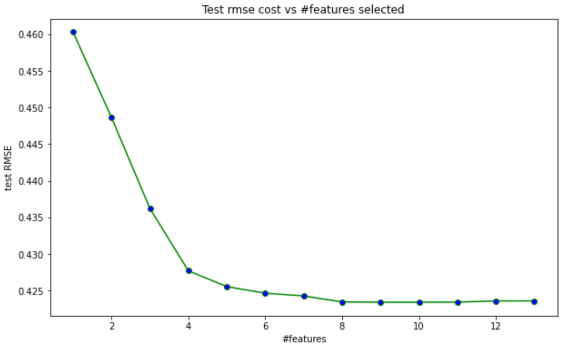
Criteria for choosing the Most significant features to add in each step:

It has the smallest rmse than all other combinations of features

When to use forward selection?

Unlike backward elimination, forward stepwise selection can be used when the number of variables under consideration is very large, even larger than the sample size. This is because forward selection starts with a null model and proceeds to add variables one at a time.





B. Greedy Backward Feature Selection Algorithm

It is a feature selection method which:

1. Begins with a model that contains all features under consideration. (Full model)
2. Then starts removing the least significant features one after the other
3. Until a pre-specified stopping rule is reached or until no variable is left in the model.

Disadvantage of Greedy backward feature selection algorithm:

Working with the full model or models close to the full model in size is computationally more expensive.

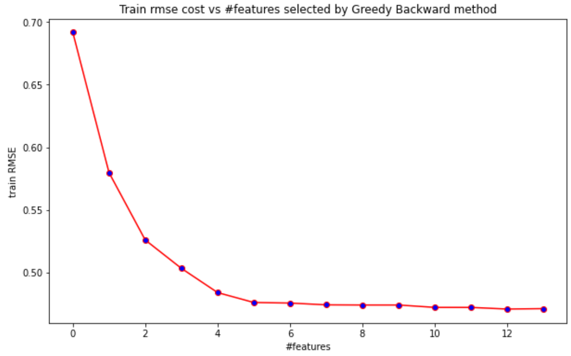
Criteria to choose the least significant variable to remove at each step:-

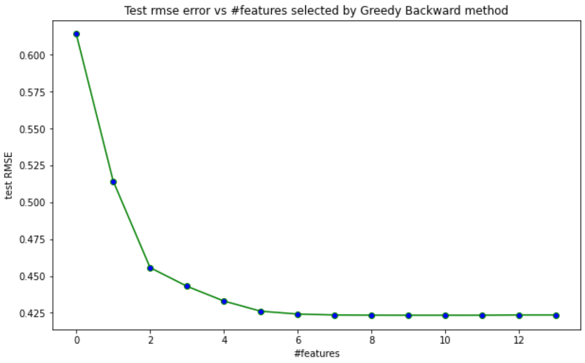
It has the maximum rmse than all other combinations of features

When to use backward selection?

If a pair of important variables are not significant marginally but are jointly significant (Correlated), then forward selection tends to miss both variables whereas backward elimination has higher chance of selecting them

Starting with the full model has the advantage of considering the effects of all variables simultaneously.





Subset of optimal features from Greedy Forward Selection Algorithm

```
[2, 8, 6, 7, 12, 10, 9, 0, 3, 5, 1]
```

```
['sqft_living', 'grade', 'view', 'condition', 'sqft_lot15',
'sqft_basement', 'sqft_above', 'bedrooms', 'sqft_lot', 'waterfront',
'bathrooms']
```

```
Training cost:  0.4731683044855858
```

```
Testing cost:  0.4226553544872862
```

Subset of optimal features from Greedy Backward Selection Algorithm

```
[0, 1, 2, 3, 4, 6, 7, 8, 9, 10, 12]
```

```
['sqft_living', 'grade', 'view', 'condition', 'sqft_lot15',
'sqft_basement', 'sqft_above', 'bedrooms', 'sqft_lot', 'waterfront',
'bathrooms']
```

```
Training cost:  0.4739056308247379
```

```
Testing cost:  0.4226179179360664
```

```
Least training error for raw data =  472291.0103533441
```

```
Testing error for raw data =  526336.3189498015
```

Minimum Training and Testing Error in 3 different Scenarios:-

	Algorithms	Minimum Training Error	Minimum Testing Error
1	Greedy Forward	0.4731683044855858	0.4226553544872862
2	Greedy Backward	0.4739056308247379	0.4226179179360664
3	Linear regression without preprocessing or feature selection	472291.0103533441	526336.3189498015

From the testing error, it seems that both the methods are showing nearly the same minimum testing error, with **greedy backward being slightly better than Greedy forward**.

And, we also know that the backward method is generally the preferred method, because the forward method produces the suppressor effects. These suppressor effects occur when predictors are only significant when another predictor is held constant.