# Assignment 2 - Report

## 2-A Correlation coefficients and Principal Component Analysis

### Problem Statement:

- The given dataset is used to predict application energy usage (in Wh) (in the Application column of the dataset) using various attributes such as temperatures, relative humidity in various rooms in the house, windspeed, visibility, etc. (in the columns: T1, RH_1, T2, RH_2, Visibility,rv1,rv2 of the dataset).
- Dataset: Link: https://drive.google.com/file/d/1hkc5d67JY1eHQ6EsidRMe3rPS3v3LeGs/view?usp=share_link

### Models:

All the models are trained using the ordinary least square method, if the inverse of the covariance matrix exists. In case, it does not exist, the models are trained using Gradient Descent with learning rate 0.01 with 5000 iterations.

Ordinary Least Square Method:

$$\hat{\beta} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$

$\hat{\beta}$ = ordinary least squares estimator

$X$ = matrix regressor variable X

$\top$ = matrix transpose

$y$ = vector of the value of the response variable

All the data points of the attributes are scaled between 0 and 1.
All the errors plotted and calculated are Root Mean Square (RMS) error.

## I. Regression Model using Pearson Correlation Coefficient

The correlation between the target value and the attributes were found using Pearson correlation coefficient.

Formula:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$r$ = correlation coefficient
$x_i$ = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
$y_i$ = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

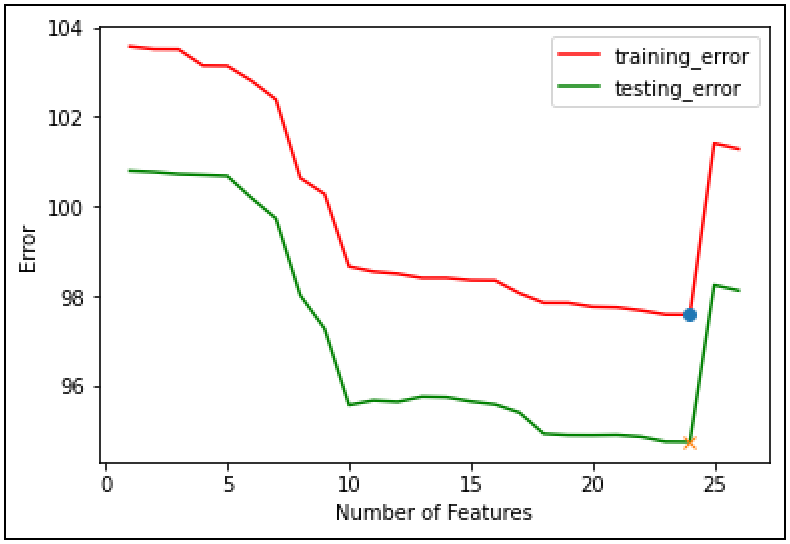
Then the absolute value of the correlation obtained is sorted in decreasing order of correlation and then selecting 1 to 26 features based on the sorted correlation value, the model is trained. Then minimum testing error is found to get the best model from the trained models using the correlation coefficients.
The Correlation value of each attribute with the target variable is:

```
RH_out: 0.1603333387234299
T6: 0.11930539008047673
T2: 0.10646831604893144
T_out: 0.10062515834755807
RH_8: 0.09206785956452258
Windspeed: 0.08725734570030218
RH_6: 0.08577211037256047
RH_1: 0.07502414297531407
T3: 0.07462762777047857
RH_2: 0.07052679526722619
RH_7: 0.06327234633190461
RH_9: 0.05505280796840861
T1: 0.04377197355001583
Press_mm_hg: 0.03654494149344637
T4: 0.03254899926733223
T8: 0.025696466335812754
RH_3: 0.023675031328264855
T7: 0.016838838400584033
Visibility: 0.014966859739902435
RH_4: 0.011582602101178528
Tdewpoint: 0.010087406069626737
T5: 0.007899586459726595
RH_5: 0.005100660540351886
rv1: 0.0019044139770705909
rv2: 0.0019044139770705909
T9: 0.0017395770003927047
```

The training and testing errors obtained for each of the subsets of attributes taken:

| | Training Error | Testing Error |
|---|---|---|
| 1 | 103.573525 | 100.796798 |
| 2 | 103.514707 | 100.768866 |
| 3 | 103.511715 | 100.724206 |
| 4 | 103.147414 | 100.705670 |
| 5 | 103.141255 | 100.683671 |
| 6 | 102.807578 | 100.189782 |
| 7 | 102.389867 | 99.732906 |
| 8 | 100.639181 | 98.018046 |
| 9 | 100.278667 | 97.271238 |
| 10 | 98.660973 | 95.568440 |
| 11 | 98.547456 | 95.672560 |
| 12 | 98.501142 | 95.640579 |
| 13 | 98.407219 | 95.755056 |
| 14 | 98.407064 | 95.743049 |
| 15 | 98.356232 | 95.649730 |
| 16 | 98.350944 | 95.583659 |
| 17 | 98.064060 | 95.400340 |
| 18 | 97.851146 | 94.927455 |
| 19 | 97.849490 | 94.896534 |
| 20 | 97.762358 | 94.893448 |
| 21 | 97.749700 | 94.900505 |
| 22 | 97.680530 | 94.860721 |
| 23 | 97.589790 | 94.751151 |
| 24 | 97.586401 | 94.746877 |
| 25 | 101.411465 | 98.248891 |
| 26 | 101.285656 | 98.124950 |

The Graph plotted on the above value:




It can be seen from the above graph that the minimum testing error occurs when 24 features out of the given 26 features is selected. This is because of the fact that 2 features - rv1 and rv2 have correlation value nearly 1, giving rise to dependence of between the assumed independent variables. Hence due to the same value being repeated twice, the value is slightly dominates the result. It can also be seen the impact is not that overpowering as the correlation of the variables with the required value is less.

The features selected in the best model:

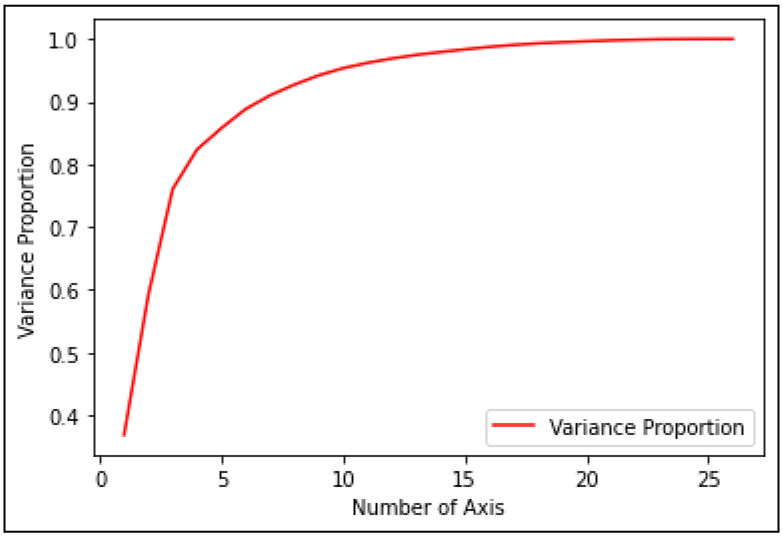
```
Features Selected:
RH_out
T6
T2
T_out
RH_8
Windspeed
RH_6
RH_1
T3
RH_2
RH_7
RH_9
T1
Press_mm_hg
T4
T8
RH_3
T7
Visibility
RH_4
Tdewpoint
T5
RH_5
```

II. Regression Model using Principal Component Analysis

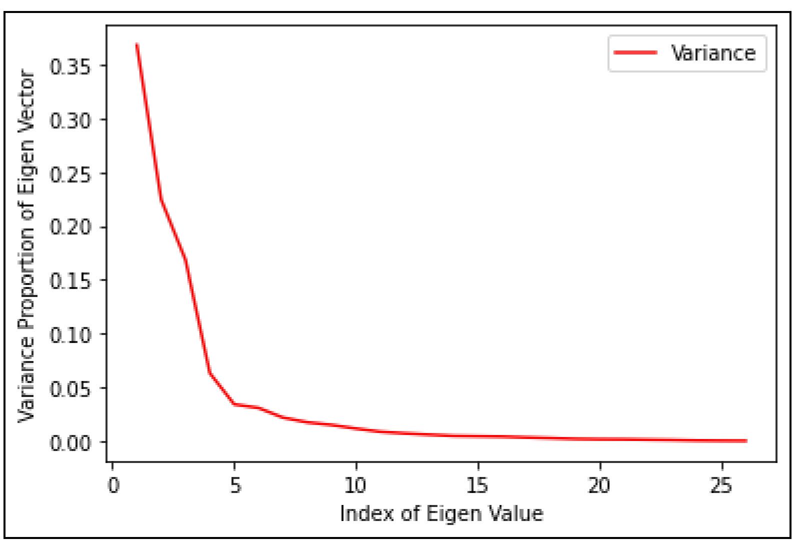
The eigenvalues and the corresponding eigenvectors of the covariance matrix of the given data are found. These eigenvalues show the proportion of variance captured when the data points are projected onto the eigenvector. The eigenvectors obtained are orthogonal and take multiple eigenvectors from the axis of a transformed coordinate system.

Now, the projected data points are taken to train the model using PCA.

The proportion of variance captured by a subset of eigenvectors sorted by eigenvalues:



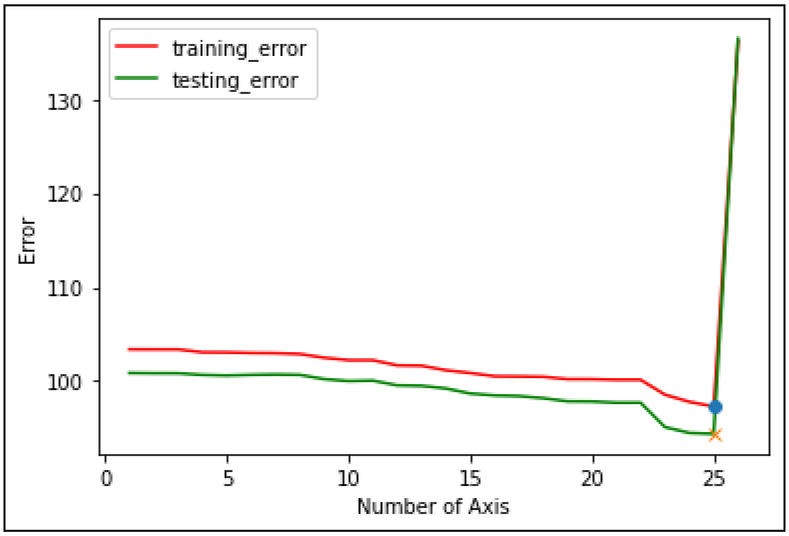

Individual eigenvalue capturing proportion of variance



Based on the above values, the model is trained by varying the number of axes, and the training and testing error is obtained.

| | Training Error | Testing Error |
|---|---|---|
| 1 | 103.375793 | 100.849868 |
| 2 | 103.368774 | 100.826635 |
| 3 | 103.368413 | 100.819763 |
| 4 | 103.060534 | 100.645151 |
| 5 | 103.051905 | 100.582704 |
| 6 | 102.987658 | 100.651769 |
| 7 | 102.967505 | 100.690310 |
| 8 | 102.878519 | 100.652480 |
| 9 | 102.480250 | 100.199107 |
| 10 | 102.221493 | 99.992605 |
| 11 | 102.217479 | 100.042222 |
| 12 | 101.665818 | 99.533960 |
| 13 | 101.626647 | 99.482702 |
| 14 | 101.146558 | 99.204567 |
| 15 | 100.850547 | 98.662800 |
| 16 | 100.504564 | 98.457602 |
| 17 | 100.484893 | 98.390786 |
| 18 | 100.446107 | 98.162782 |
| 19 | 100.194803 | 97.828445 |
| 20 | 100.190093 | 97.811517 |
| 21 | 100.127120 | 97.682017 |
| 22 | 100.127114 | 97.683477 |
| 23 | 98.535367 | 95.078389 |
| 24 | 97.775494 | 94.457046 |
| 25 | 97.276179 | 94.351092 |
| 26 | 136.187006 | 136.501923 |

Graphing the above-tabulated data:




The minimum testing error occurs when 25 axes are chosen (first 25 eigenvectors based on eigenvalues sorted in descending order), giving the best model using PCA.

The eigenvalues selected are:

```
Eigen Values Selected:
0.36137201362390436
0.22001467810004366
0.1649067323567587
0.06187161017572615
0.033352129448324275
0.030075064682935065
0.021264978606802713
0.016913862417986283
0.014517517436937218
0.01120912492354119
0.0082912493957791
0.006820527444507616
0.005645681032044566
0.0045077662455042244
0.004197847931099375
0.0038805365864463736
0.0030892031442502662
0.0024770141969162737
0.0017650979161468235
0.00147305220894227
0.0013579389852605957
0.001022293770081716
0.0007491469199514631
```

# 2-B Greedy Forward and Backward Feature Selection

## Problem Statement:

Perform i) greedy forward feature selection and ii) greedy backward feature selection to find the subset of features that make the optimal regression model. Find the minimum training and testing error of the optimal model (using 1,2,3,...26 features).

## Models:

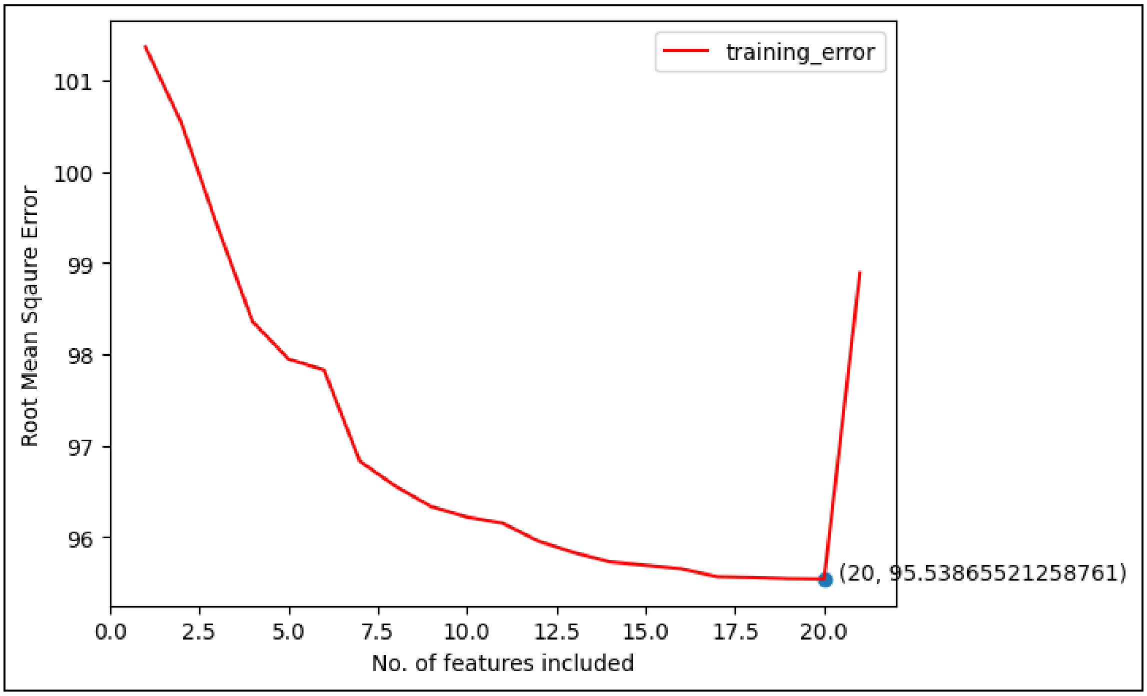
I. <u>Regression Model using Forward Feature Selection</u>

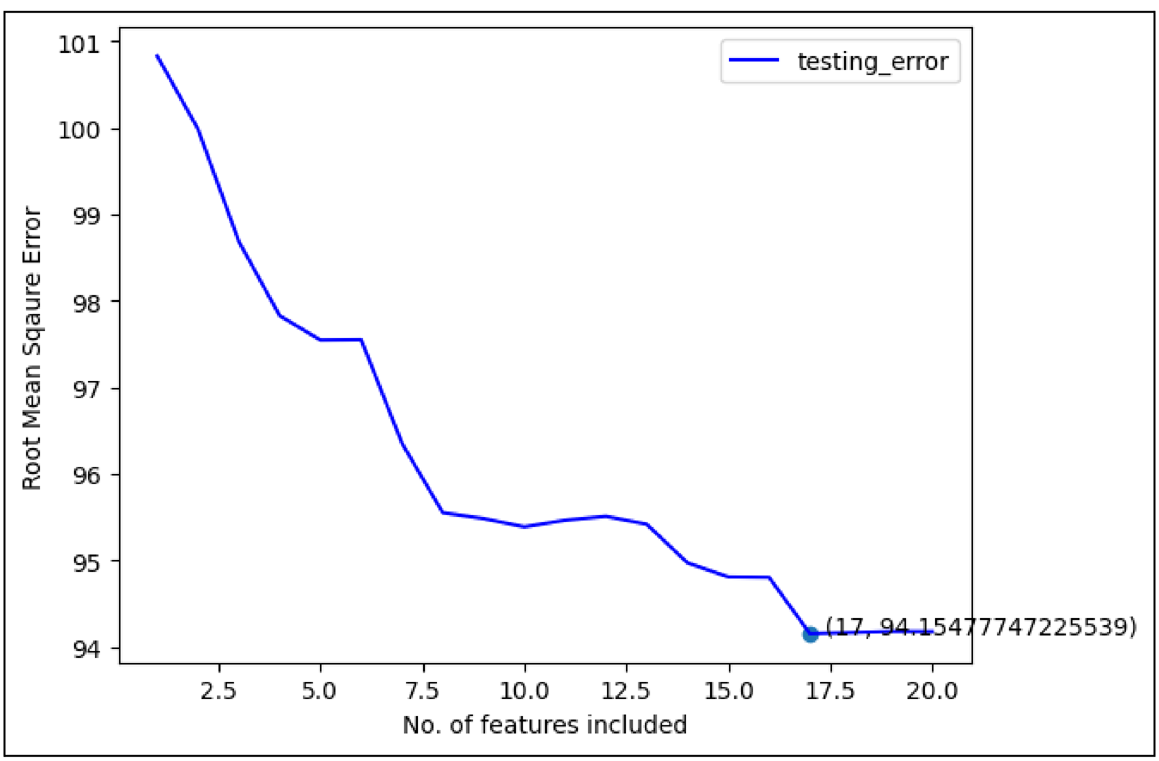
In the greedy forward heuristic, the features are selected based on the minimum validation error/testing error taking n features at a time. The best model obtained is then compared with a model trained by adding a feature in the subset of n features already taken. The two are compared. If the model with n+1 features predicts better, then the above process is continued until there are no more features that can be selected. If the model with n features is better than the model with n+1 features, then the model with n features is the best, and no more iterations of feature addition/selection are performed.

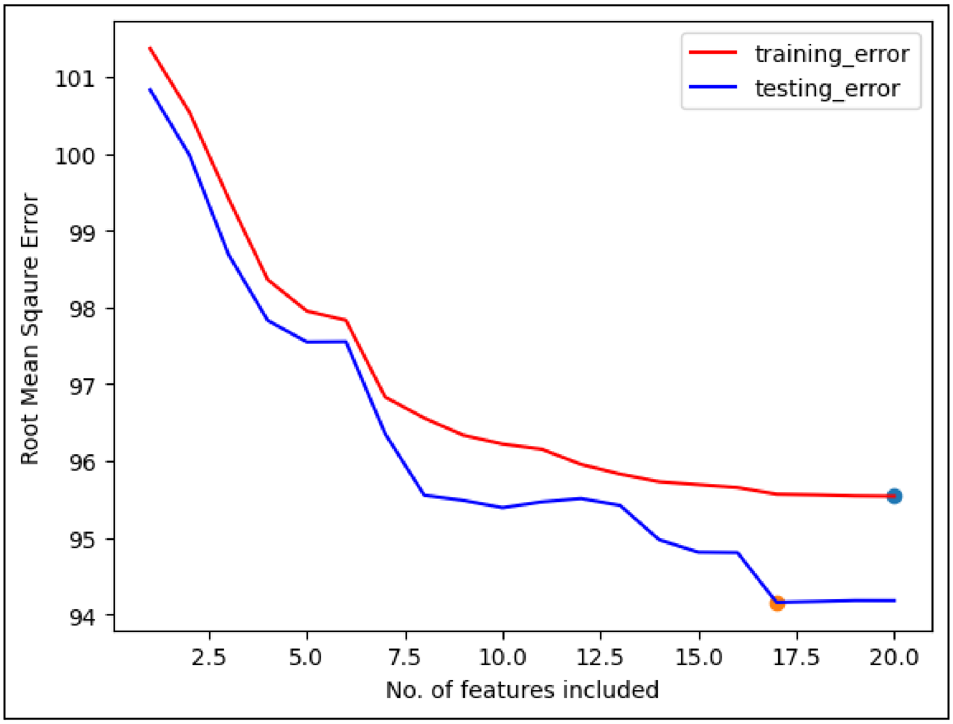
The tabulated training and testing error for the greedy forward approach:

| | Training Error | Testing Error |
|---|---|---|
| 1 | 101.365302 | 100.829199 |
| 2 | 100.540350 | 99.982648 |
| 3 | 99.417249 | 98.685852 |
| 4 | 98.358938 | 97.828126 |
| 5 | 97.949752 | 97.546865 |
| 6 | 97.829592 | 97.550532 |
| 7 | 96.830701 | 96.354096 |
| 8 | 96.555934 | 95.551726 |
| 9 | 96.332769 | 95.484140 |
| 10 | 96.218242 | 95.388117 |
| 11 | 96.150856 | 95.464914 |
| 12 | 95.954213 | 95.509738 |
| 13 | 95.827826 | 95.418654 |
| 14 | 95.727259 | 94.972024 |
| 15 | 95.689237 | 94.809872 |
| 16 | 95.650942 | 94.804854 |
| 17 | 95.564946 | 94.154777 |
| 18 | 95.556018 | 94.166222 |
| 19 | 95.543176 | 94.179559 |
| 20 | 95.538655 | 94.178460 |

Graphing the above training and testing errors:






When a set of 20 features were selected, the model performed better on the validation and training data as compared to all the models with 21 features selected. Hence the iterations were stopped at 20 features.
On checking the testing error, a set of 17 features had the least error, hence the best model using the greedy forward approach.

The minimum training and testing errors obtained are 95.5386 and 94.1547.

The features selected are:

```
Features Selected:
RH_out
RH_1
RH_8
RH_2
Windspeed
T7
T3
T2
T1
RH_4
T6
T_out
RH_3
T8
RH_6
T4
T9
```

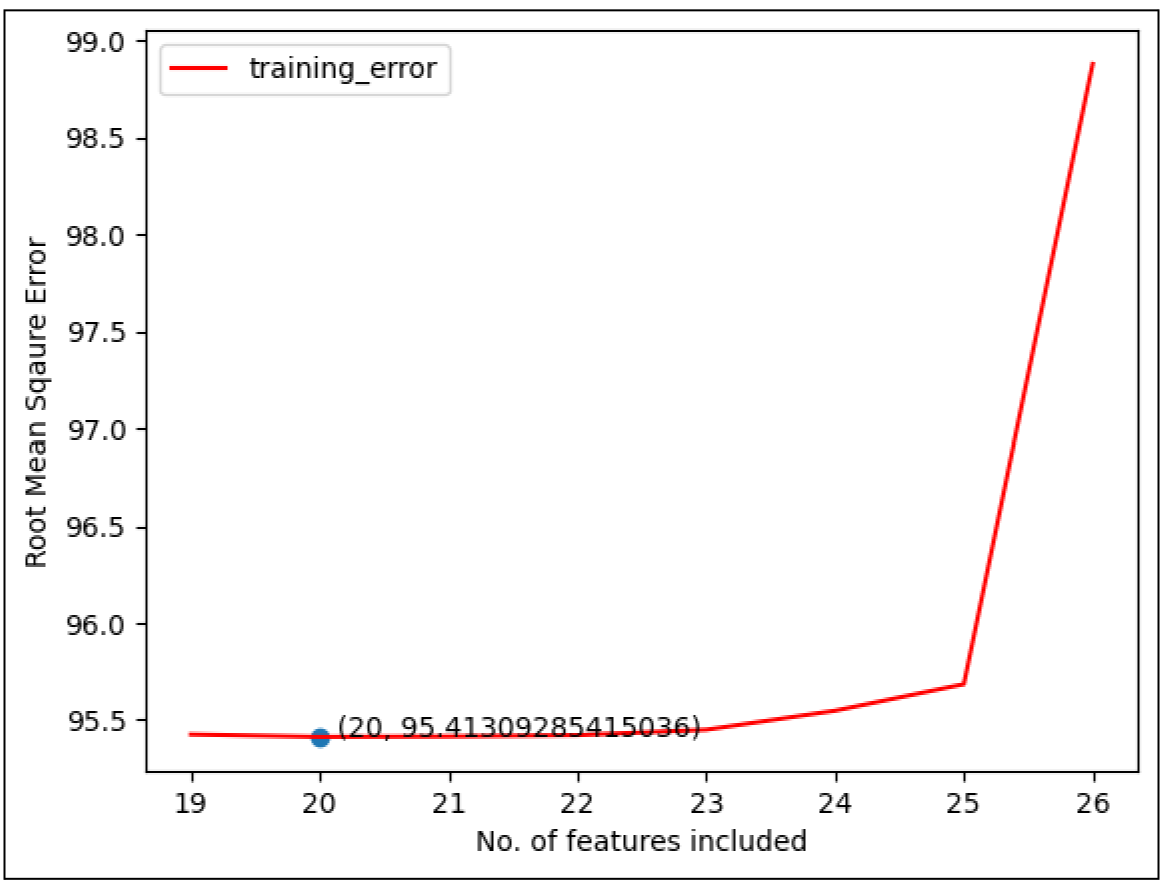
II. <u>Regression Model using Backward Feature Selection</u>

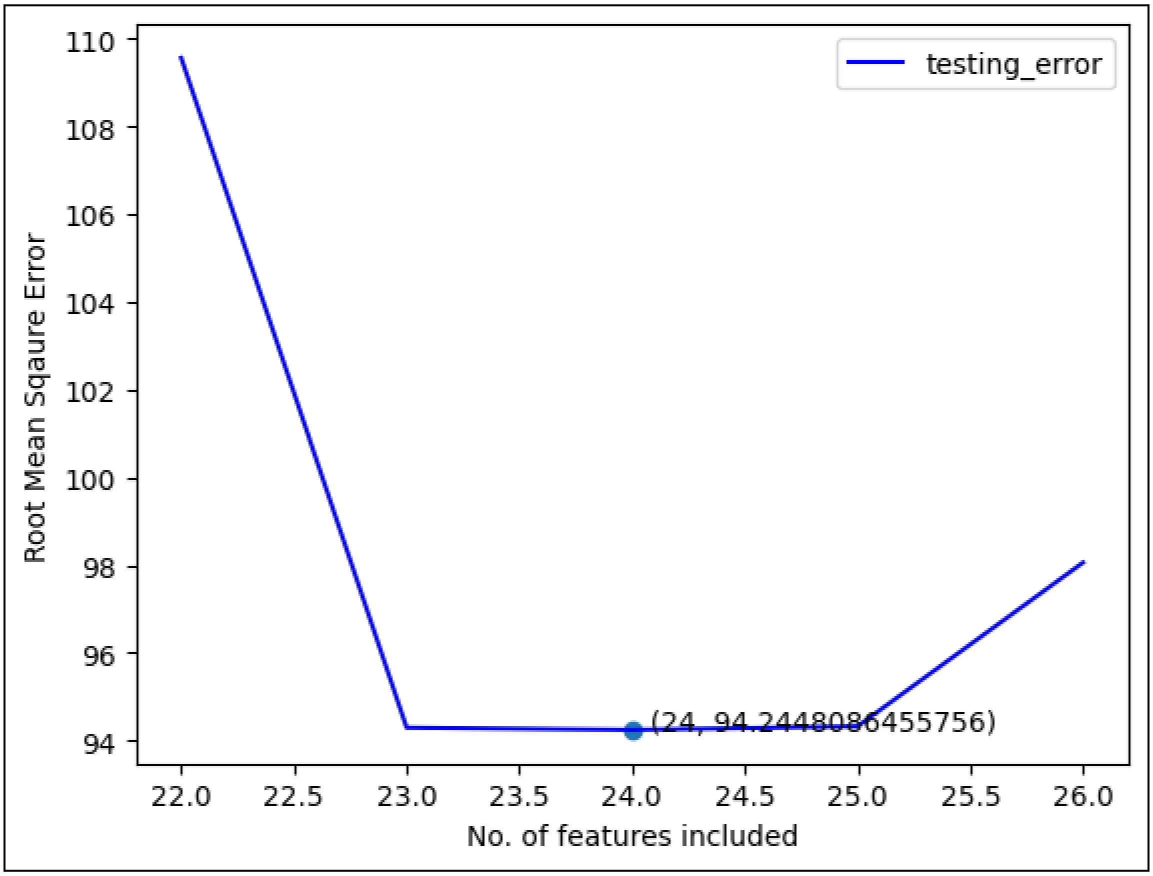
In the greedy backward heuristic, the features are selected based on the minimum validation error/testing error taking n features at a time. The best model obtained is then compared with a model trained by removing a feature in the subset of n features already taken. The two are compared. If the model with n-1 features predicts better, then the above process is continued until there are no more features that can be removed. If the model with n features is better than the model with n-1 features, then the model with n features is the best, and no more iterations of feature deletions are performed.

The tabulated training and testing error for the greedy backward approach:

| | Training Error | Testing Error |
|---|---|---|
| 20 | 95.413093 | 106.630224 |
| 21 | 95.415608 | 108.379569 |
| 22 | 95.421875 | 109.564684 |
| 23 | 95.449757 | 94.295111 |
| 24 | 95.547902 | 94.244809 |
| 25 | 95.684780 | 94.325599 |
| 26 | 98.879665 | 98.060106 |

Graphing the above training and testing errors:






When a set of 20 features were selected, the model performed better on the validation and training data as compared to all the models with 19 features selected. Hence the iterations were stopped at 20 features.
On checking the testing error, a set of 24 features had the least error, hence the best model using greedy backward approach.

The features dropped are:

```
Features Removed:
rv1
T7
```

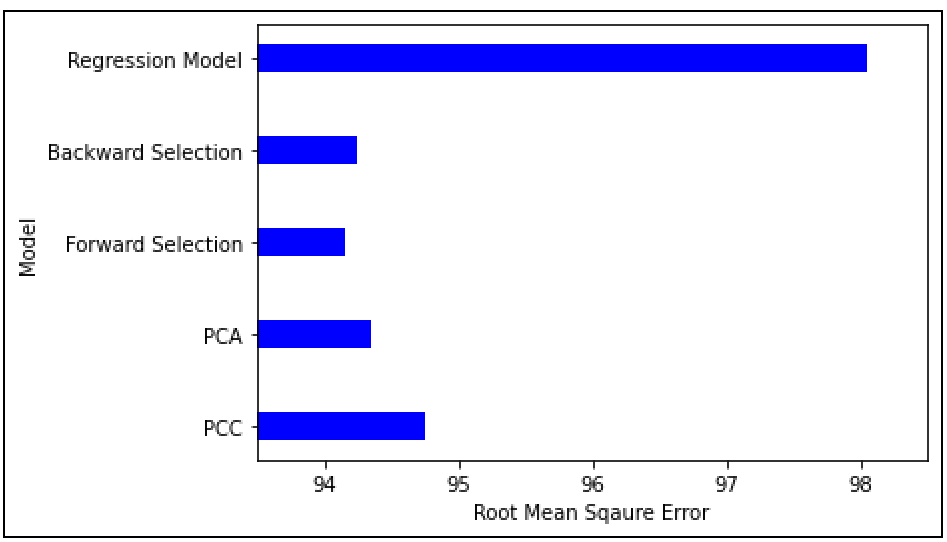
# 2-C Comparative Analysis

The following is the error analysis of the best models obtained using Pearson Correlation Coefficient, Principal Component Analysis, Greedy Forward Selection, Greedy Backward Selection as compared to the classical regression model.

It can be observed that the error in each of the models obtained above is lesser than the error obtained in regression model.

| | Error |
|---|---|
| PCC | 94.746877 |
| PCA | 94.351092 |
| Forward Selection | 94.154777 |
| Backward Selection | 94.244809 |
| Regression Model | 98.048048 |

Plotting the above testing error values:




The best model out of all the compared model is obtained using Greedy Forward Heuristic Approach with 17 features mentioned earlier, giving the minimum testing error.

The other models also have a similar error values, depicting higher dependence on a few attributes as compared to other attributes.

## Group Members:

Abhinav Tyagi - 2020A7PS2043H
Rishiraj Datta - 2020A7PS2075H
Ritvik - 2020A7PS1723H