CS F320 Foundation of Data Science

Assignment-2 Report

RAHUL JAUHARI – 2020A7PS0106H

Alankar Agarwal - 2020A7PS1698H
Divyan Goyal - 2020A7PS0042H

2-A Correlation coefficients and Principal Component Analysis

1. Pearson correlation coefficient.

   The Pearson correlation coefficient measures the linear association between variables. Its value can be interpreted like so:

   - +1 - Complete positive correlation
   - +0.8 - Strong positive correlation
   - +0.6 - Moderate positive correlation
   - 0 - no correlation whatsoever
   - -0.6 - Moderate negative correlation
   - -0.8 - Strong negative correlation
   - -1 - Complete negative correlation

   In this we first select those features which have highest correlation coefficient with respect to target feature. Then iteratively we add more independent features into consideration and see among which set we find the maximum correlation with target feature. Then we include one more feature in the already selected pool and compute the correlation formula and store it in a list. And we select those features in which there is a naximum correlation.
   We repete this process until we have included all of the features.

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

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

Feature	Selected feature	RMSE training Error	RMSE testing Error
1	['RH_out']	0.12744293807079896	0.12377512398095444
2	['RH_out', 'RH_2']	0.12280256178320602	0.12067375920442275

3	['RH_out', 'RH_2', 'RH_1']	0.12152331174996302	0.12057399109008163
4	['RH_out', 'RH_2', 'RH_1', 'RH_5']	0.12125058320723521	0.12027354159242656
5	['RH_out', 'RH_2', 'RH_1', 'RH_5', 'RH_3']	0.12117263345076451	0.1202532686321593
6	['RH_out', 'RH_2', 'RH_1', 'RH_5', 'RH_3', 'T2']	0.12019799715607825	0.11953770913826667
7	['RH_out', 'RH_2', 'RH_1', 'RH_5', 'RH_3', 'T2', 'T3']	0.12004159182079739	0.1193459051818575
8	['RH_out', 'RH_2', 'RH_1', 'RH_5', 'RH_3', 'T2', 'T3', 'Press_mm_hg']	0.11861842075783625	0.11771256507916784
9	['RH_out', 'RH_2', 'RH_1', 'RH_5', 'RH_3', 'T2', 'T3', 'Press_mm_hg', 'T1']	0.11860495887610788	0.11772187755289303
10	['RH_out', 'RH_2', 'RH_1', 'RH_5', 'RH_3', 'T2', 'T3', 'Press_mm_hg', 'T1', 'T_out']	0.11859219466708536	0.11762955844630468
11	['RH_out', 'RH_2', 'RH_1', 'RH_5', 'RH_3', 'T2', 'T3', 'Press_mm_hg', 'T1', 'T_out', 'Windspeed']	0.11754152386122085	0.11657782334376934
12		0.11612830484359445	0.11405011691078575

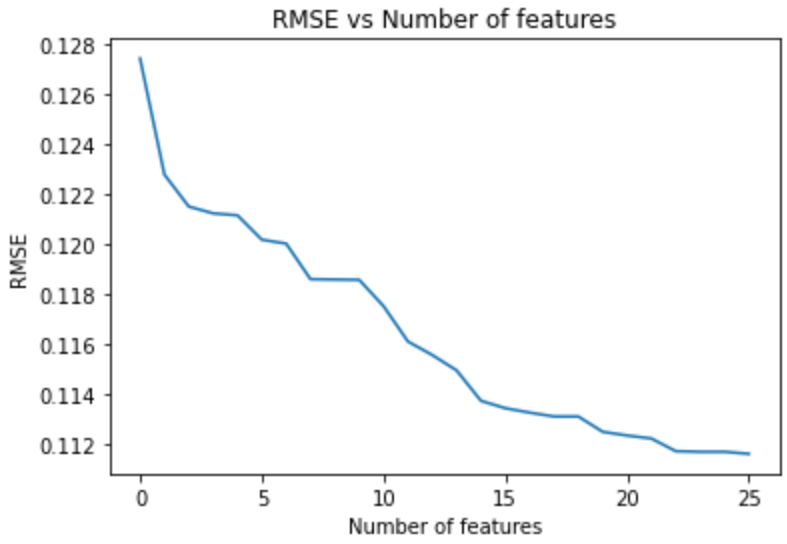
	['RH_out', 'RH_2', 'RH_1', 'RH_5', 'RH_3', 'T2', 'T3', 'Press_mm_hg', 'T1', 'T_out', 'Windspeed', 'RH_9']		
13	['RH_out', 'RH_2', 'RH_1', 'RH_5', 'RH_3', 'T2', 'T3', 'Press_mm_hg', 'T1', 'T_out', 'Windspeed', 'RH_9', 'T6']	0.11558631625110742	0.11335317343710932
14	['RH_out', 'RH_2', 'RH_1', 'RH_5', 'RH_3', 'T2', 'T3', 'Press_mm_hg', 'T1', 'T_out', 'Windspeed', 'RH_9', 'T6', 'T5'] RMSE train: RMSE test:	0.1149736815731303	0.11294166071703092
15	['RH_out', 'RH_2', 'RH_1', 'RH_5', 'RH_3', 'T2', 'T3', 'Press_mm_hg', 'T1', 'T_out', 'Windspeed', 'RH_9', 'T6', 'T5', 'RH_8'] RMSE train: RMSE test:	0.11375670021475669	0.1113739638735952
16	['RH_out', 'RH_2', 'RH_1', 'RH_5', 'RH_3', 'T2', 'T3', 'Press_mm_hg', 'T1', 'T_out', 'Windspeed', 'RH_9', 'T6', 'T5',	0.1134594127604545	0.11084434125817659

	'RH_8', 'T8'] RMSE train: RMSE test:		
17	['RH_out', 'RH_2', 'RH_1', 'RH_5', 'RH_3', 'T2', 'T3', 'Press_mm_hg', 'T1', 'T_out', 'Windspeed', 'RH_9', 'T6', 'T5', 'RH_8', 'T8', 'Visibility']	0.1132839770680946	0.11070592197101511
18	['RH_out', 'RH_2', 'RH_1', 'RH_5', 'RH_3', 'T2', 'T3', 'Press_mm_hg', 'T1', 'T_out', 'Windspeed', 'RH_9', 'T6', 'T5', 'RH_8', 'T8', 'Visibility', 'RH_7']	0.1131291135408054	0.11045292279215342
19	['RH_out', 'RH_2', 'RH_1', 'RH_5', 'RH_3', 'T2', 'T3', 'Press_mm_hg', 'T1', 'T_out', 'Windspeed', 'RH_9', 'T6', 'T5', 'RH_8', 'T8', 'Visibility', 'RH_7', 'T4']	0.11312861465624412	0.11045210018469301
20	['RH_out', 'RH_2', 'RH_1', 'RH_5', 'RH_3', 'T2', 'T3', 'Press_mm_hg', 'T1', 'T_out', 'Windspeed', 'RH_9', 'T6', 'T5', 'RH_8', 'T8',	0.1125169498215868	0.11028105388297123

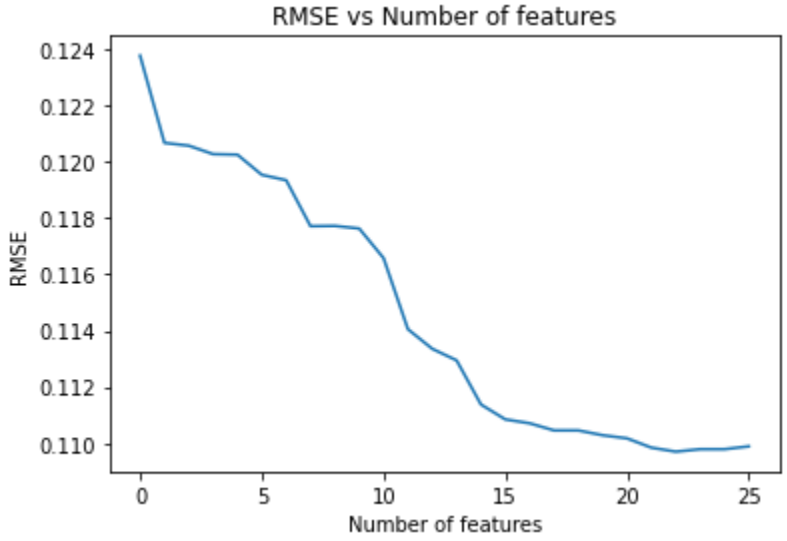
	'Visibility', 'RH_7', 'T4', 'Tdewpoint']		
21	['RH_out', 'RH_2', 'RH_1', 'RH_5', 'RH_3', 'T2', 'T3', 'Press_mm_hg', 'T1', 'T_out', 'Windspeed', 'RH_9', 'T6', 'T5', 'RH_8', 'T8', 'Visibility', 'RH_7', 'T4', 'Tdewpoint', 'RH_4']	0.11236982299757121	0.11017018716757675
22	['RH_out', 'RH_2', 'RH_1', 'RH_5', 'RH_3', 'T2', 'T3', 'Press_mm_hg', 'T1', 'T_out', 'Windspeed', 'RH_9', 'T6', 'T5', 'RH_8', 'T8', 'Visibility', 'RH_7', 'T4', 'Tdewpoint', 'RH_4', 'T7'] RMSE train: RMSE test:	0.11224831907253865	0.10983972916544918
23	['RH_out', 'RH_2', 'RH_1', 'RH_5', 'RH_3', 'T2', 'T3', 'Press_mm_hg', 'T1', 'T_out', 'Windspeed', 'RH_9', 'T6', 'T5', 'RH_8', 'T8', 'Visibility', 'RH_7', 'T4', 'Tdewpoint', 'RH_4', 'T7', 'T9']	0.11174199936971925	0.10969803979115604
24	['RH_out', 'RH_2', 'RH_1', 'RH_5', 'RH_3', 'T2', 'T3', 'Press_mm_hg', 'T1', 'T_out', 'Windspeed',	0.11171720547513804	0.10978496881221952

	'RH_9', 'T6', 'T5', 'RH_8', 'T8', 'Visibility', 'RH_7', 'T4', 'Tdewpoint', 'RH_4', 'T7', 'T9', 'rv2']		
25	['RH_out', 'RH_2', 'RH_1', 'RH_5', 'RH_3', 'T2', 'T3', 'Press_mm_hg', 'T1', 'T_out', 'Windspeed', 'RH_9', 'T6', 'T5', 'RH_8', 'T8', 'Visibility', 'RH_7', 'T4', 'Tdewpoint', 'RH_4', 'T7', 'T9', 'rv2', 'rv1']	0.11171720547513804	0.10978496881221919
26	['RH_out', 'RH_2', 'RH_1', 'RH_5', 'RH_3', 'T2', 'T3', 'Press_mm_hg', 'T1', 'T_out', 'Windspeed', 'RH_9', 'T6', 'T5', 'RH_8', 'T8', 'Visibility', 'RH_7', 'T4', 'Tdewpoint', 'RH_4', 'T7', 'T9', 'rv2', 'rv1', 'RH_6']	0.11163636250654298	0.10988616869654728

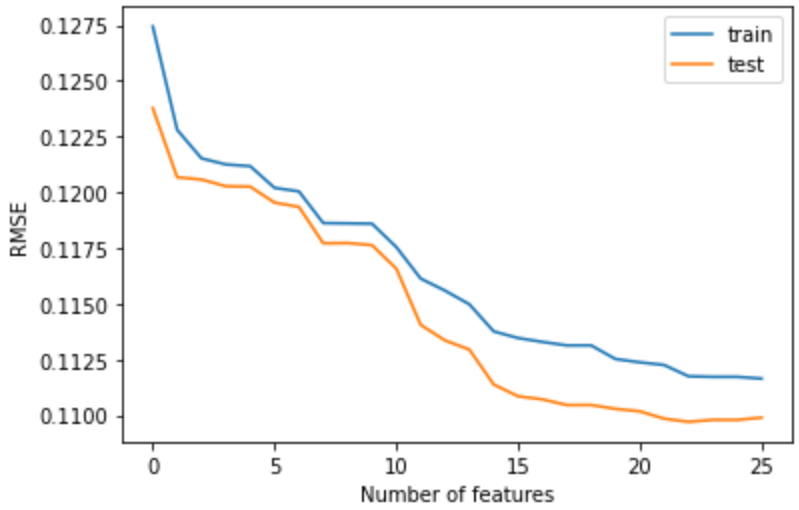
Training



Testing



Comparative



Best model:

Best model is obtained when we included all the features except ‘rv2', 'rv1', 'RH_6' with

Training error :- 0.11174199936971925

Testion error :- 0.10969803979115604

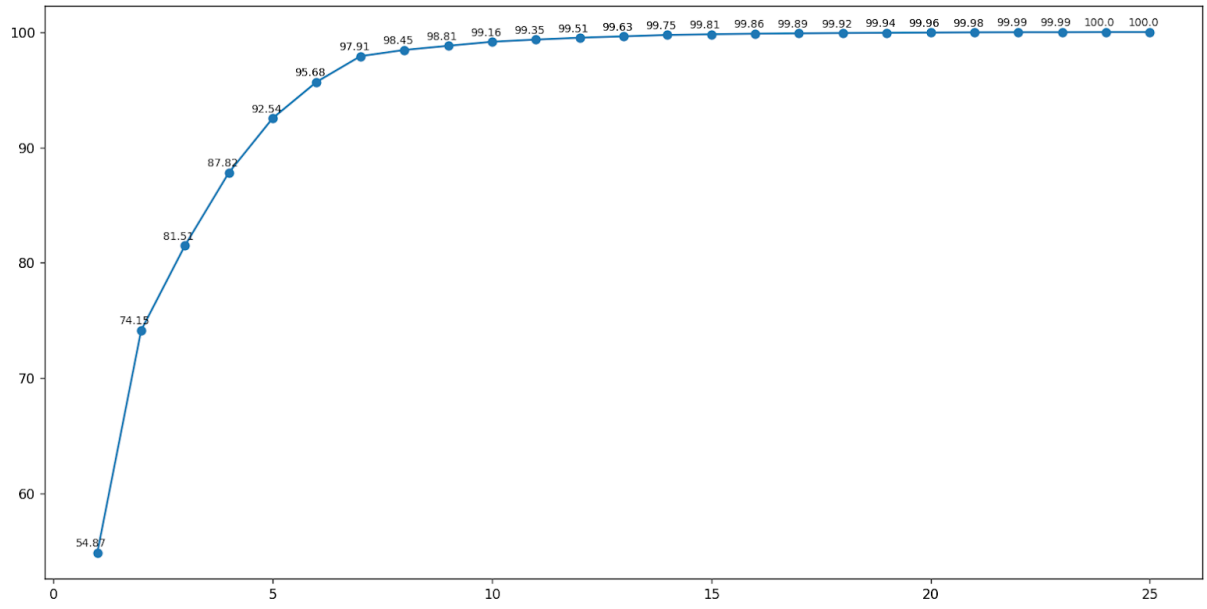
Observation:

We observed that there is decrease in rmse value for training and testing error as we introduced features based on their correlation with target feature.

2. Principal component analysis

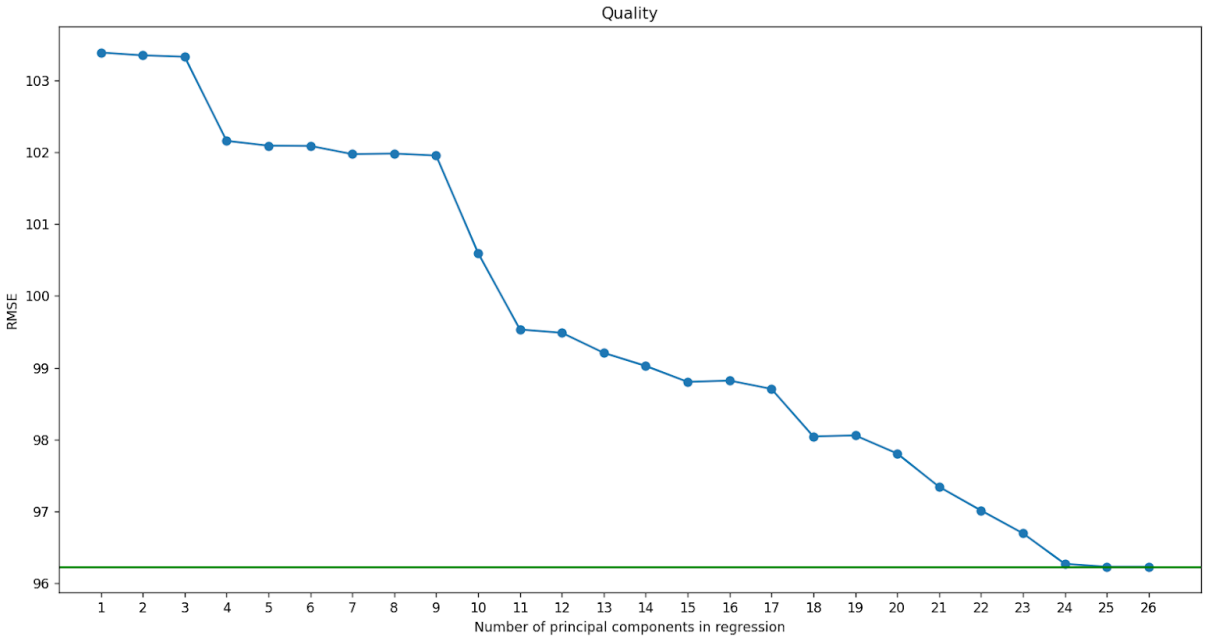
Principal component analysis is is a way to reduce number of features defining a data while retaining most of the information. It transforms a number of variables that may be correlated into a smaller number of uncorrelated variables, known as principal components. The principal components are linear combinations of the original variables weighted by their variances (or eigenvalues) in a particular orthogonal dimension. The main objective of PCA is to simplify your model features into fewer components to help visualize patterns in your data and to help your model run faster. Using PCA also reduces the chance of overfitting your model by eliminating features with high correlation.

Below is graph of sum of %variance of each of 26 feature. They are plotted in decreasing order of their importance in deriving the resultant value. In the program we can find the important features like RH_6 comes out to be the most important feature and then rv_1 and so on.



Using RH_6, rv2 and Visibility we can generate a model that describes more than 80% of our data reducing 26 features to just 3 principal components. That's a huge computational advantage for little drop in accuracy of data.

Using RH_6, rv2, Tdewpoint, Visibility, RH_out, RH_5 and this accuracy an be increased to more than 95%.



In this graph we have plotted the root mean square error of the linear regression model generated using x principal componenets of the dataset.

This graph tells us about the reduction in root mean squared error when we start using more and more feature for our model. When using just one feature the error is understandably more than using more than one feature. When the linear regression model is designed using all the 26 features, the error from the testing data is least.

2-B Greedy Forward and Backward Feature Selection

We used closed form to calculate theta value it is given by the following function:-

$$\theta = (X^T X)^{-1} X^T \vec{y}.$$

Without using an iterative algorithm, it is possible to solve for the optimal values of the parameter theta and then proceed directly to the global optimum using an algorithm known as the normal equation. It works only for Linear Regression and not any other algorithm.

Derivation:-

MSE error formula:-

$$\sum_{i=1}^{m} \left(y^{(i)} - \theta^T x^{(i)} \right)^2$$

To calculate theta , we take the partial derivative of the MSE loss function (equation 2) with respect to theta and set it equal to zero. Then, do a little bit of linear algebra to get the value of theta.

$$J(\theta) = ((X\theta)^T - y^T)(X\theta - y)$$

$$J(\theta) = (X\theta)^T X\theta - (X\theta)^T y - y^T (X\theta) + y^T y$$

$$\frac{\partial J}{\partial \theta} = 2X^T X\theta - 2X^T y = 0$$

$$X^T X\theta = X^T y$$

$$\theta = (X^T X)^{-1} X^T \vec{y}.$$

1. Greedy forward feature selection

   Greedy forward selection is a popular technique for feature subset selection.
   The first step in the Forward Feature Selection process is to train n(indepent features) models using each feature separately and evaluate the results. And among these models we select the one which gives the least RMSE error.
   Then we will include another feature and make a model with these input features and we will repeat this process until we have included all the features.

Feature	Selected feature	RMSE training Error	RMSE testing Error
1	['Press_mm_hg']	0.12207202008521323	0.12728079372937962
2	['Press_mm_hg', 'T1']	0.12076475782676556	0.12563287457674824
3	['Press_mm_hg', 'T1', 'T6']	0.11997824796228031	0.12466982388355759
4	['Press_mm_hg', 'T1', 'T6', 'RH_1']	0.11946340712913418	0.12402061985170326
5	['Press_mm_hg', 'T1', 'T6', 'RH_1', 'Windspeed']	0.11909646361696528	0.12375247029409626
6	['Press_mm_hg', 'T1', 'T6', 'RH_1', 'Windspeed', 'T_out']	0.11885565918925203	0.12338185608021972
7	['Press_mm_hg', 'T1', 'T6', 'RH_1', 'Windspeed', 'T_out', 'T3']	0.11871093509592891	0.1231138060887015
8	['Press_mm_hg', 'T1', 'T6', 'RH_1', 'Windspeed', 'T_out', 'T3', 'RH_3']	0.11860547389078217	0.12294585701025215

9	['Press_mm_hg', 'T1', 'T6', 'RH_1', 'Windspeed', 'T_out', 'T3', 'RH_3', 'T2']	0.11854181651205462	0.12278763016496988
10	['Press_mm_hg', 'T1', 'T6', 'RH_1', 'Windspeed', 'T_out', 'T3', 'RH_3', 'T2', 'RH_5']	0.11850584765874644	0.12277380638139229
11	['Press_mm_hg', 'T1', 'T6', 'RH_1', 'Windspeed', 'T_out', 'T3', 'RH_3', 'T2', 'RH_5', 'Visibility']	0.1184896704351273	0.12266223499391006
12	['Press_mm_hg', 'T1', 'T6', 'RH_1', 'Windspeed', 'T_out', 'T3', 'RH_3', 'T2', 'RH_5', 'Visibility', 'T8']	0.11852118650040139	0.12266617802355023
13	['Press_mm_hg', 'T1', 'T6', 'RH_1', 'Windspeed', 'T_out', 'T3', 'RH_3', 'T2', 'RH_5', 'Visibility', 'T8', 'rv1']	0.11858494150807482	0.12253502567247826
14	['Press_mm_hg', 'T1', 'T6', 'RH_1', 'Windspeed', 'T_out', 'T3', 'RH_3', 'T2', 'RH_5', 'Visibility', 'T8', 'rv1', 'RH_4']	0.11866175577812442	0.12257032182598443
15	['Press_mm_hg', 'T1', 'T6', 'RH_1', 'Windspeed', 'T_out', 'T3', 'RH_3', 'T2',	0.11873910580705099	0.12256447954137069

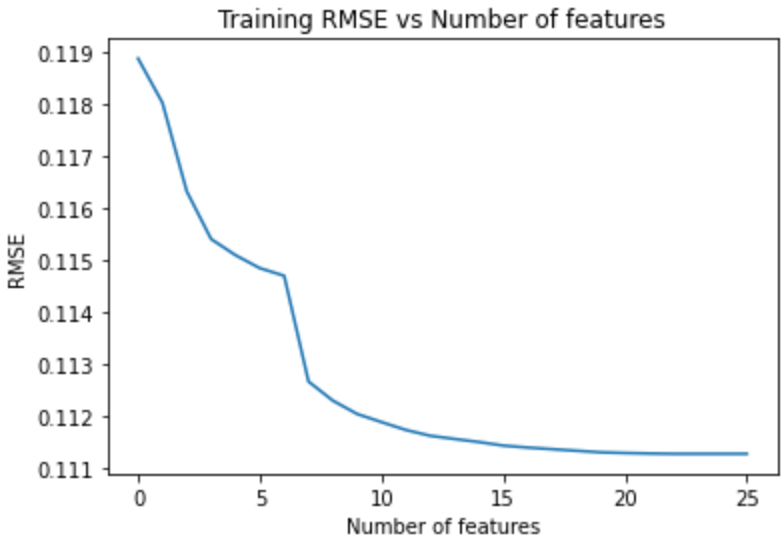
	'RH_5', 'Visibility', 'T8', 'rv1', 'RH_4', 'T4']		
16	['Press_mm_hg', 'T1', 'T6', 'RH_1', 'Windspeed', 'T_out', 'T3', 'RH_3', 'T2', 'RH_5', 'Visibility', 'T8', 'rv1', 'RH_4', 'T4', 'T7']	0.11882648452803729	0.1226444122422815
17	['Press_mm_hg', 'T1', 'T6', 'RH_1', 'Windspeed', 'T_out', 'T3', 'RH_3', 'T2', 'RH_5', 'Visibility', 'T8', 'rv1', 'RH_4', 'T4', 'T7', 'RH_6']	0.11890622820479854	0.12277997132175732
18	['Press_mm_hg', 'T1', 'T6', 'RH_1', 'Windspeed', 'T_out', 'T3', 'RH_3', 'T2', 'RH_5', 'Visibility', 'T8', 'rv1', 'RH_4', 'T4', 'T7', 'RH_6', 'rv2']	0.11898141871174933	0.12274422998643396
19	['Press_mm_hg', 'T1', 'T6', 'RH_1', 'Windspeed', 'T_out', 'T3', 'RH_3', 'T2', 'RH_5', 'Visibility', 'T8', 'rv1', 'RH_4', 'T4', 'T7', 'RH_6', 'rv2', 'T5']	0.11904889451190027	0.12279371693359176
20	['Press_mm_hg', 'T1', 'T6', 'RH_1', 'Windspeed', 'T_out', 'T3', 'RH_3', 'T2', 'RH_5', 'Visibility',	0.1191270226899001	0.12289488964472448

	'T8', 'rv1', 'RH_4', 'T4', 'T7', 'RH_6', 'rv2', 'T5', 'RH_2']		
21	['Press_mm_hg', 'T1', 'T6', 'RH_1', 'Windspeed', 'T_out', 'T3', 'RH_3', 'T2', 'RH_5', 'Visibility', 'T8', 'rv1', 'RH_4', 'T4', 'T7', 'RH_6', 'rv2', 'T5', 'RH_2', 'T9']	0.11920887315724217	0.12296081211050286
22	['Press_mm_hg', 'T1', 'T6', 'RH_1', 'Windspeed', 'T_out', 'T3', 'RH_3', 'T2', 'RH_5', 'Visibility', 'T8', 'rv1', 'RH_4', 'T4', 'T7', 'RH_6', 'rv2', 'T5', 'RH_2', 'T9', 'Tdewpoint']	0.11928875156120662	0.12301272282747999
23	['Press_mm_hg', 'T1', 'T6', 'RH_1', 'Windspeed', 'T_out', 'T3', 'RH_3', 'T2', 'RH_5', 'Visibility', 'T8', 'rv1', 'RH_4', 'T4', 'T7', 'RH_6', 'rv2', 'T5', 'RH_2', 'T9', 'Tdewpoint', 'RH_7']	0.11937962346349577	0.12309399084805682
24	['Press_mm_hg', 'T1', 'T6', 'RH_1', 'Windspeed', 'T_out', 'T3', 'RH_3', 'T2', 'RH_5', 'Visibility', 'T8', 'rv1', 'RH_4', 'T4', 'T7', 'RH_6', 'rv2', 'T5', 'RH_2',	0.11946513654035422	0.12317740755397719

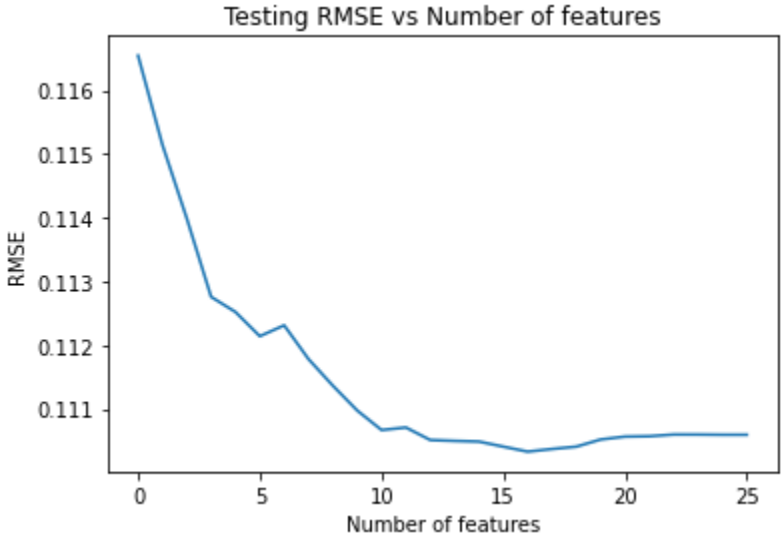
	'T9', 'Tdewpoint', 'RH_7', 'RH_9']		
25	['Press_mm_hg', 'T1', 'T6', 'RH_1', 'Windspeed', 'T_out', 'T3', 'RH_3', 'T2', 'RH_5', 'Visibility', 'T8', 'rv1', 'RH_4', 'T4', 'T7', 'RH_6', 'rv2', 'T5', 'RH_2', 'T9', 'Tdewpoint', 'RH_7', 'RH_9', 'RH_8']	0.11955320099143922	0.12326559613507934
26	['Press_mm_hg', 'T1', 'T6', 'RH_1', 'Windspeed', 'T_out', 'T3', 'RH_3', 'T2', 'RH_5', 'Visibility', 'T8', 'rv1', 'RH_4', 'T4', 'T7', 'RH_6', 'rv2', 'T5', 'RH_2', 'T9', 'Tdewpoint', 'RH_7', 'RH_9', 'RH_8', 'RH_out']	0.11969155992759786	0.12344995546061983

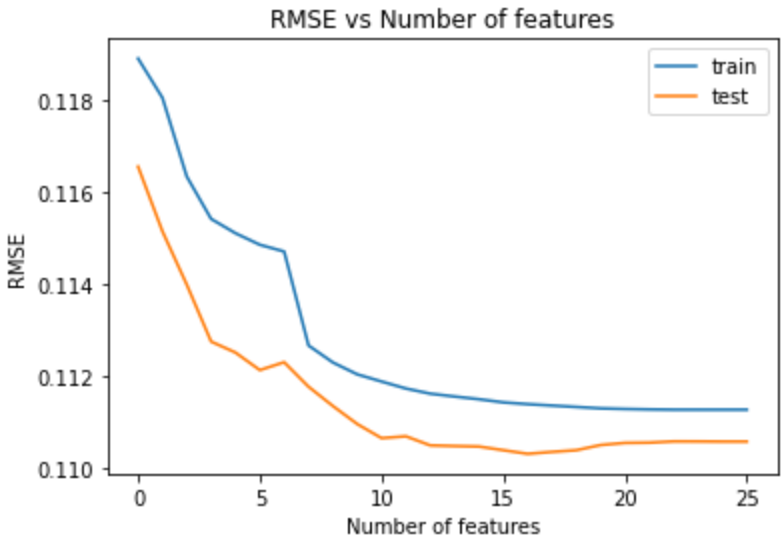
RMSE Error:

Training error



Testing error





Best model

Best model for greedy forward is obtained when we use the following features:-
['Press_mm_hg', 'T1', 'T6', 'RH_1', 'Windspeed', 'T_out', 'T3', 'RH_3', 'T2', 'RH_5', 'Visibilit y']

Rmse training error:- 0.1184896704351273
Rmse testing error:- 0.12266223499391006

Observation:-
We observed that both training and testing error decreases as we introduce new features into linear regression.

2. Greedy backward feature selection

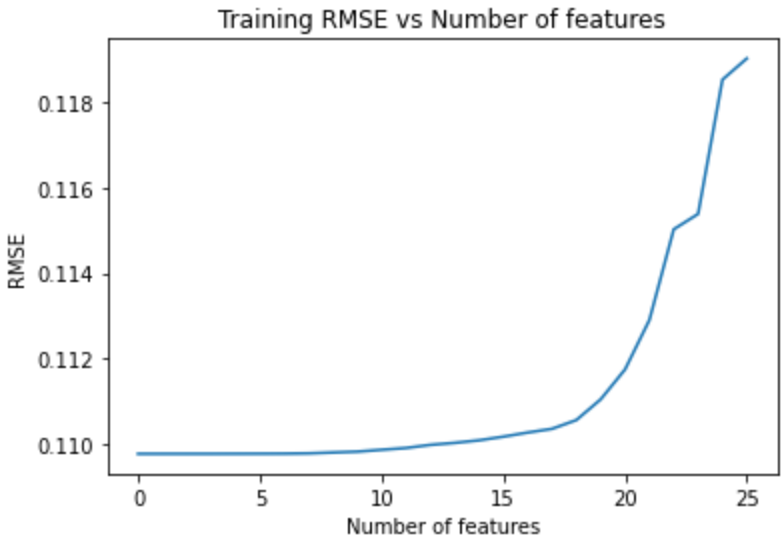
   In greedy backward feature selection we iteratively remove features which have least impact(worst feature) on the model(after removing a feature there is not much difference in training error and testing error).

Feature Removed	Removed feature	RMSE training Error	RMSE testing Error
1	rv1	0.10978422151330516	0.11658266023600317
2	RH_out	0.1097843954407821	0.11659247342131872
3	Tdewpoint	0.10978455238348325	0.1165920166057929
4	rv2	0.10978476304468833	0.1165988778953707
5	T1	0.10978585521094467	0.11660135230423369
6	T7	0.10978690959072593	0.11658262473577718
7	T5	0.10978859360909723	0.11660392445232096
8	Press_mm_hg	0.10979487517802797	0.11656908472753301
9	RH_5	0.10981546384708281	0.11668565948651782
10	Visibility	0.10983472638140189	0.11668521123276836
11	RH_9	0.10987843735276503	0.11661176193341322
12	RH_3	0.10991971559333386	0.11673887465045595
13	RH_7	0.10999311268937104	0.11689557370866377

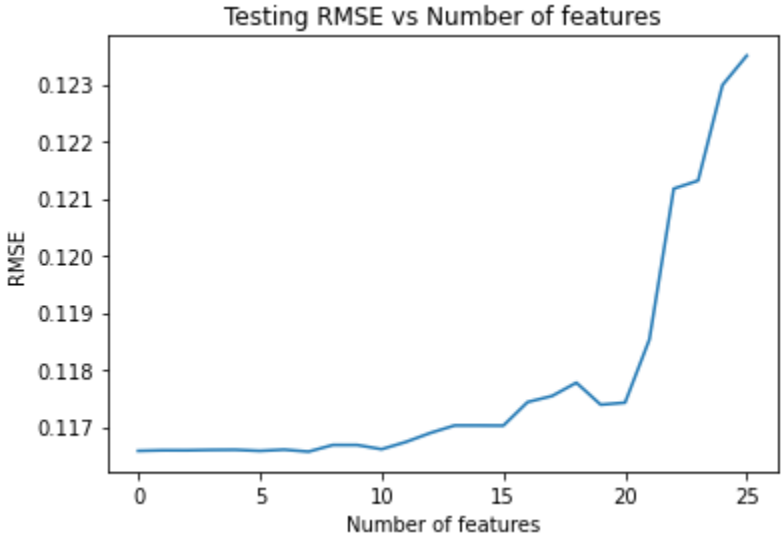
14	RH_4	0.11003976689936489	0.1170273826048694
15	T4	0.11009916335420968	0.11702705236149484
16	RH_6	0.1101843153460488	0.11702443496370371
17	T_out	0.11028107453896595	0.1174375614779414
18	Windspeed	0.1103674741512249	0.11754290002054954
19	T6	0.11057263689204798	0.1177760143908223
20	T8	0.1110560713379486	0.11739106749383814
21	T2	0.1117549500218775	0.11742683729797185
22	RH_2	0.11290832754577598	0.11853287238582794
23	T3	0.1150259810925781	0.12117269000652776
24	T9	0.11538398729981067	0.12131308968619414
25	RH_1	0.11851943587706132	0.12298529922789211
26	RH_8	0.11901326838980962	0.12350168724794332

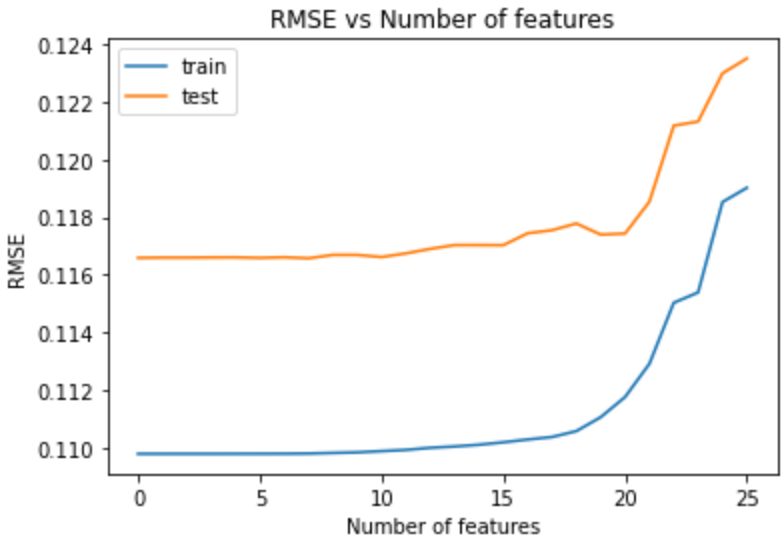
RMSE Error:
Rmse training error:- 0.10978422151330516
Rmse testing error:- 0.11658266023600317

Training error



Testing error





Observation:-
We observed that training error and testing error remained almost remains constant even if we removed the following features

- rv1
- RH_out
- Tdewpoint
- rv2

This implies that these features have very less contribution while calculating target feature.

Best model:-
Best model is the one in which we remove only the first feature rv1.

2-C Comparative Analysis

Errors obtained by including all features:-
Training Error With all features = 0.11035070991745938
Testing Error With all features = 0.11430279002283553

Part A

Pearson correlation :-
RMSE training error :- 0.11174199936971925
RMSE testing error :- 0.10969803979115604

Principal Component Analysis :-
RMSE training error :- 0.09623469677620804
RMSE testing error :-0.10185136627642967

Part B

Greedy forward :-
Rmse training error:- 0.1184896704351273
Rmse testing error:- 0.12266223499391006

Greedy backward:-
Rmse training error:- 0.10978422151330516
Rmse testing error:- 0.11658266023600317

As we can see that RMSE error values for Pearson correlation and greedy backward are very much similar to that of linear reggresson obtained by including all the features.

Best 6 features as per pearson correlation coefficient are 'RH_out', 'RH_2', 'RH_1', 'RH_5', 'RH_3', 'T2' whereas the PCA returns RH_6, rv2, Tdewpoint, Visibility, RH_out, RH_5 as it's principal components. Greedy backward gives T2, RH_2, T3, T9, RH_1 and RH_8 and forward gives Press_mm_hg, T1, T6, RH_1, Windspeed, T_out. There are some differences between these methods and their impotant features, which is natural as they all have different methods to evaluate their importance. Still we see that features like RH_1, T2 and RH_2 are common amongst these methods showing their similar purpose.