COMPENG 4SL3 Assignment 2

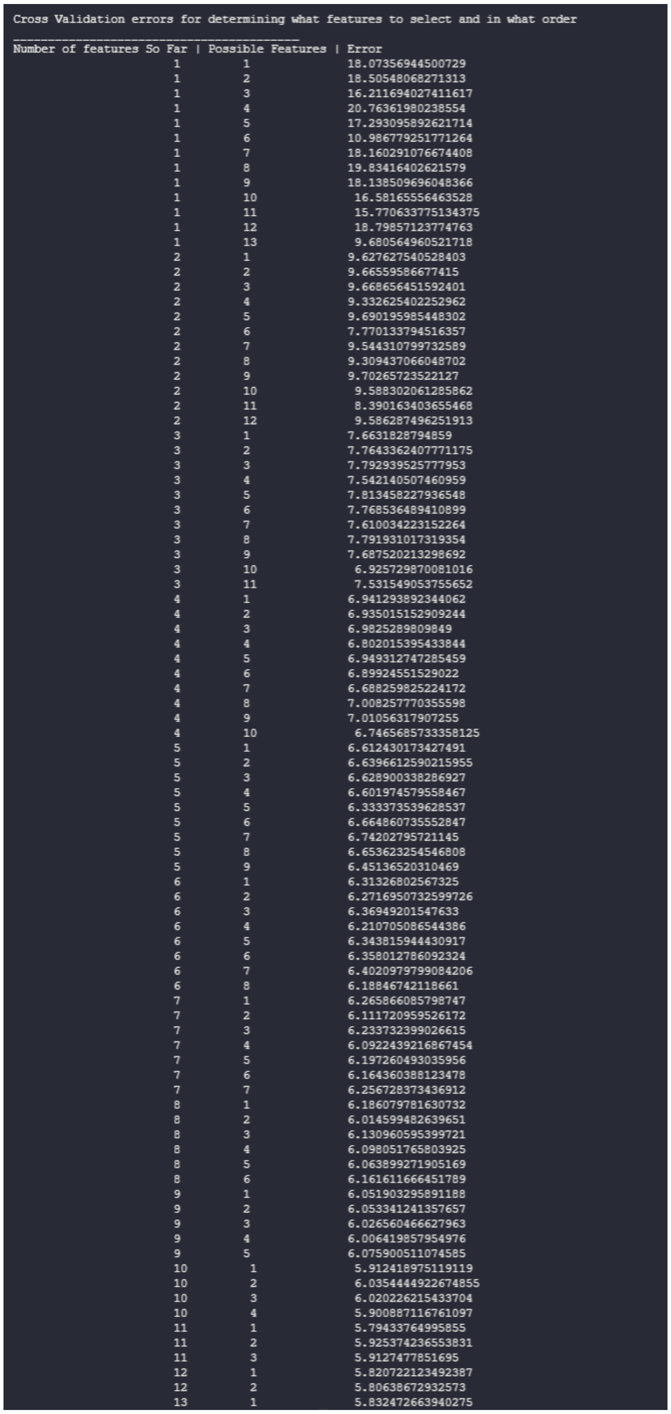
Uday Sharma

Dr. Sorina Dumitrescu

COMPENG 4SL3 Assignment 2
Uday Sharma
400139246

To build up a model using multiple features, we use a greedy approach and pick the next best feature to add, see the chart below which highlights all of the features currently in our data set, as well as the resultant error from adding the next data set. The optimal path for adding in the features is:

[12, 5, 10, 7, 4, 11, 3, 1, 8, 9, 0, 6, 2]



```
Cross Validation errors for determining what features to select and in what order

_________________________________________
Number of features So Far | Possible Features | Error
                       1        1             18.07356944500729
                       1        2             18.50548068271313
                       1        3             16.211694027411617
                       1        4             20.76361980238554
                       1        5             17.293095892621714
                       1        6             10.986779251771264
                       1        7             18.160291076674408
                       1        8             19.83416402621579
                       1        9             18.138509696048366
                       1        10             16.58165556463528
                       1        11             15.770633775134375
                       1        12             18.79857123774763
                       1        13             9.680564960521718
                       2        1             9.627627540528403
                       2        2             9.66559586677415
                       2        3             9.668656451592401
                       2        4             9.332625402252962
                       2        5             9.690195985448302
                       2        6             7.770133794516357
                       2        7             9.544310799732589
                       2        8             9.309437066048702
                       2        9             9.70265723522127
                       2        10             9.588302061285862
                       2        11             8.390163403655468
                       2        12             9.586287496251913
                       3        1             7.6631828794859
                       3        2             7.7643362407771175
                       3        3             7.792939525777953
                       3        4             7.542140507460959
                       3        5             7.813458227936548
                       3        6             7.768536489410899
                       3        7             7.610034223152264
                       3        8             7.791931017319354
                       3        9             7.687520213298692
                       3        10             6.925729870081016
                       3        11             7.531549053755652
                       4        1             6.941293892344062
                       4        2             6.935015152909244
                       4        3             6.9825289809849
                       4        4             6.802015395433844
                       4        5             6.949312747285459
                       4        6             6.89924551529022
                       4        7             6.688259825224172
                       4        8             7.008257770355598
                       4        9             7.01056317907255
                       4        10             6.7465685733358125
                       5        1             6.612430173427491
                       5        2             6.6396612590215955
                       5        3             6.628900338286927
                       5        4             6.601974579558467
                       5        5             6.333373539628537
                       5        6             6.664860735552847
                       5        7             6.74202795721145
                       5        8             6.653623254546808
                       5        9             6.45136520310469
                       6        1             6.31326802567325
                       6        2             6.2716950732599726
                       6        3             6.36949201547633
                       6        4             6.210705086544386
                       6        5             6.343815944430917
                       6        6             6.358012786092324
                       6        7             6.4020979799084206
                       6        8             6.18846742118661
                       7        1             6.265866085798747
                       7        2             6.111720959526172
                       7        3             6.233732399026615
                       7        4             6.0922439216867454
                       7        5             6.197260493035956
                       7        6             6.164360388123478
                       7        7             6.256728373436912
                       8        1             6.186079781630732
                       8        2             6.014599482639651
                       8        3             6.130960595399721
                       8        4             6.098051765803925
                       8        5             6.063899271905169
                       8        6             6.161611666451789
                       9        1             6.051903295891188
                       9        2             6.053341241357657
                       9        3             6.026560466627963
                       9        4             6.006419857954976
                       9        5             6.075900511074585
                       10        1             5.912418975119119
                       10        2             6.0354444922674855
                       10        3             6.020226215433704
                       10        4             5.900887116761097
                       11        1             5.79433764995855
                       11        2             5.925374236553831
                       11        3             5.9127477851695
                       12        1             5.820722123492387
                       12        2             5.80638672932573
                       13        1             5.832472663940275
```

We then perform basis expansion for the features that we had chosen and retrieve the cross-validation errors for those as well as the error without basis expansion. Unfortunately, due to unknown issues, performing basis expansion for subsets with more than 6 features resulted in an error pertaining to generating a singular matrix. Therefore, I was unable to determine a cross-validation error that would be less than the ones determined without basis expansion. The functions used for basis expansion were:

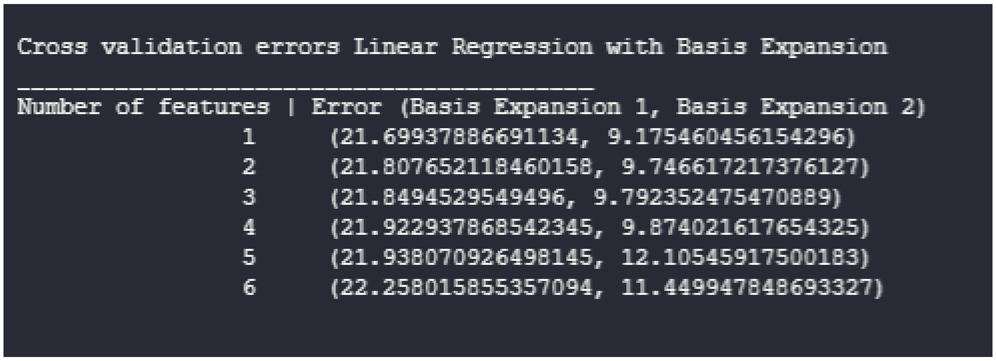
Basis Expansion 1:

$$\begin{pmatrix} 1 & \cdots & F\hat{}2 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & F\hat{}2 \end{pmatrix}$$

Basis Expansion 2:

$$\begin{pmatrix} 1 & \cdots & F^2 & \cdots & F\hat{}3 \\ \vdots & \ddots & F^2 & \cdots & \vdots \\ 1 & \cdots & F^2 & \cdots & F\hat{}3 \end{pmatrix}$$

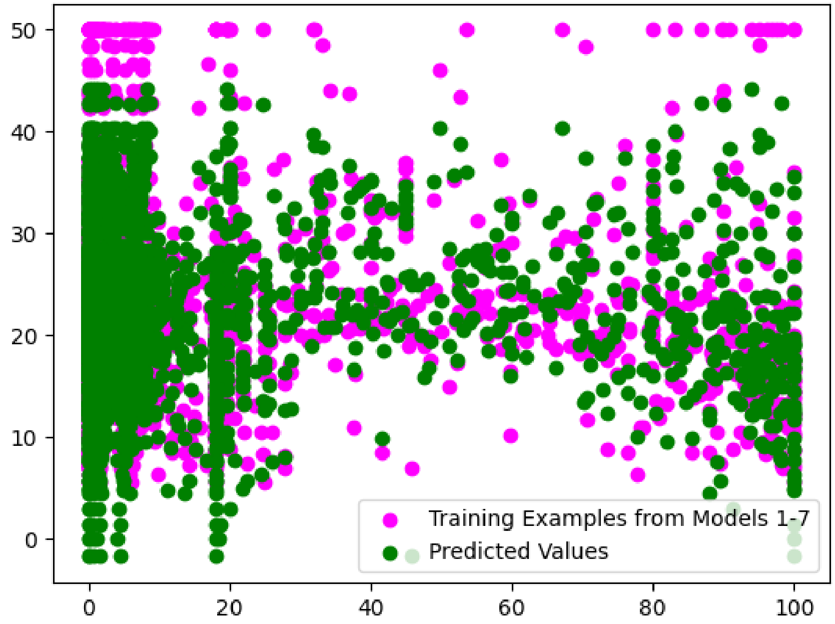
For every selected feature. This was done through some experimentation with regards to what would achieve and error close to the cross-validation error without Basis Expansion as due to the singular matrix issues that I was facing in this section of the code.



```
Cross validation errors Linear Regression with Basis Expansion

________________________________________
Number of features | Error (Basis Expansion 1, Basis Expansion 2)
                1     (21.69937886691134, 9.175460456154296)
                2     (21.807652118460158, 9.746617217376127)
                3     (21.8494529549496, 9.792352475470889)
                4     (21.922937868542345, 9.874021617654325)
                5     (21.938070926498145, 12.10545917500183)
                6     (22.258015855357094, 11.449947848693327)
```

The number of features I chose to pick was 7 as that was representative of point where I achieved a local minimum in test and cross-validation errors and I did not have data for any point past that specific number of features. See the plot below:



(Selected model with K=7 features)

As was mentioned before, the optimal path for adding in the features was:

[12, 5, 10, 7, 4, 11, 3, 1, 8, 9, 0, 6, 2]

This mean that the first feature added in was the 12th, the second was the 5th and the third was the 10th respectively. This means that the features used to train the model above were *[12,5,10,7,4,11,3]* in that order. Furthermore, the parameter vector for the subset of features chosen was:

[3.21254083e+01, -5.35384296e-01, 4.11731919e+00, -9.64232038e-01

-1.19352313e+00, -1.73636446e+0,1 9.08568890e-03, 3.15634022e+00]

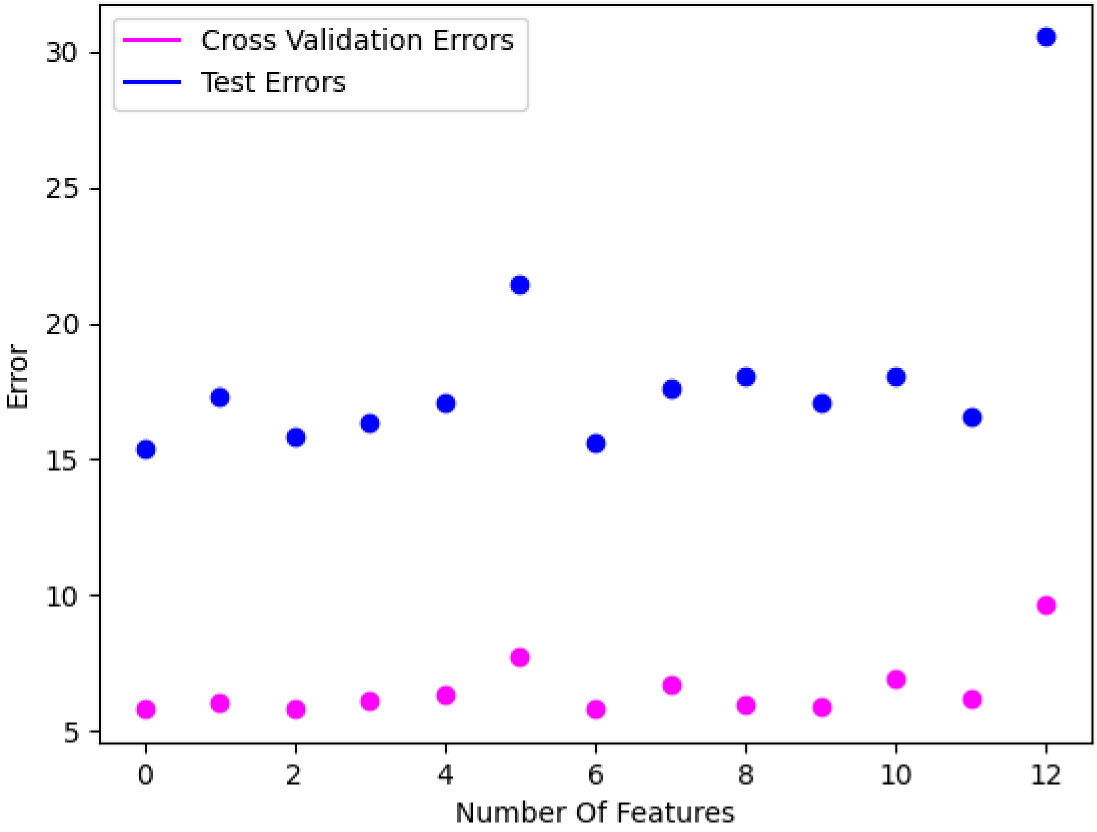
Furthermore, the parameter vector for basis expansions for this selected subset were:

[-1.4340133256905858, -12.847520339064431, 5.808831631218254, -1.4739453948291157, 75.92044801025622, 0.03571734207840116, 0.024759702199866673, 1.332522600534162, -0.19017352116995312, 0.08093694696438547, -66.99022073021843, -5.702511989363543e-05]

[-1.9948943239548953, -18.705107653106097, 10.027909272583202, -3.2658102232226156, 249.49722670391202, -0.005461431006494877, 0.06622974491356004, 1.6516762827191087, -0.6483927654448962, 0.35344318374893646, -413.35557091468945, 0.00012466020322809968, -0.000854154888053904, 0.009944074983820883, 0.012844515746031960, -0.01428328739217477, 206.83283338067122, -2.298521019213854e-07]

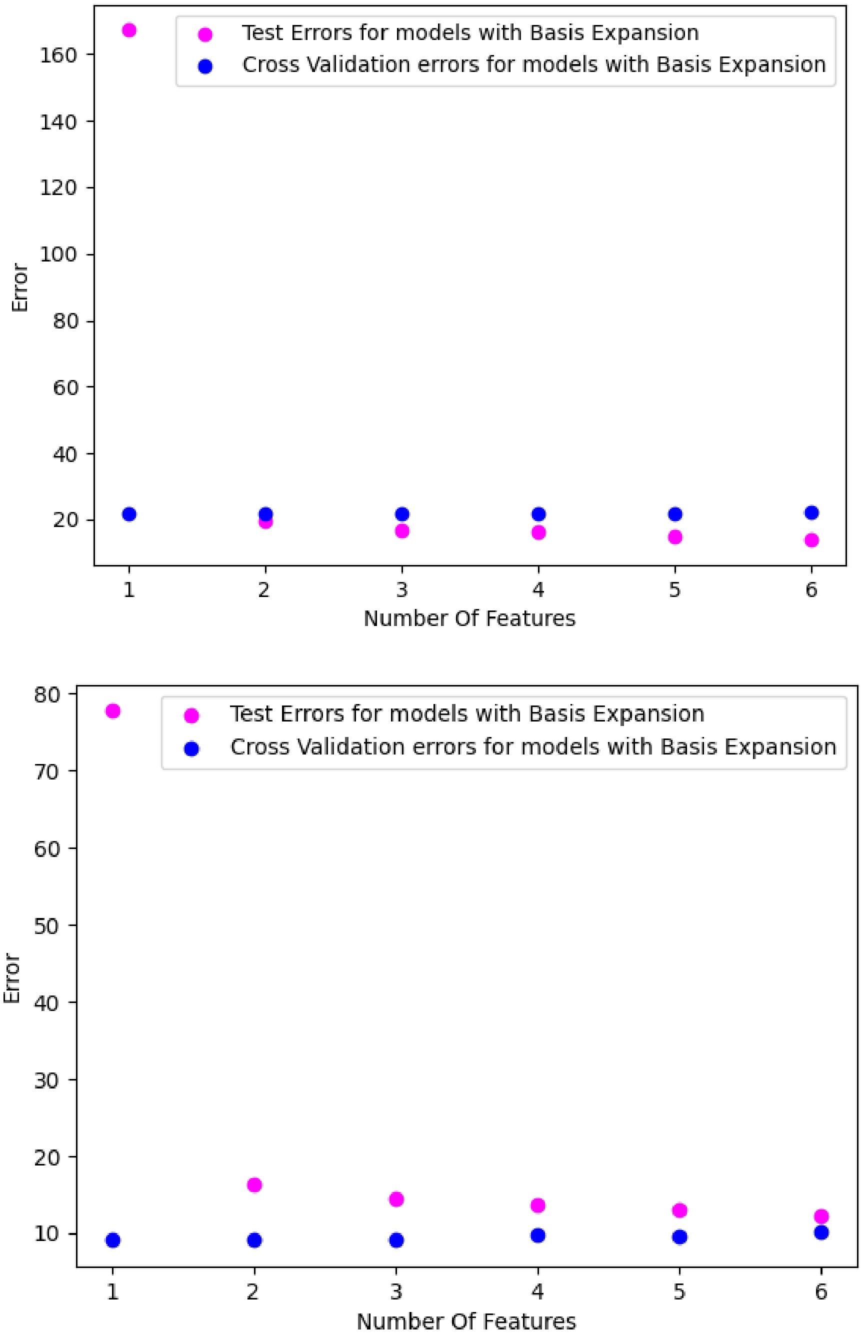
COMPENG 4SL3 Assignment 2
Uday Sharma
400139246

See below the plot of the cross-validation errors and test errors for the first 13 models without basis expansion:



(Cross validation and test errors against the number of features)

We can see from the plot above that at 6 features we see the lowest cross-validation error and this corresponds with the test errors as well, that at 6 features we also see the lowest test error. We can also see that the test error is consistently larger than the validation error for all models. The trend of cross-validation errors follows the same pattern as the test errors excepts they are much lower. Furthermore, we can take a look at this for the data we have for basis expansion as well.



(Basis expansion plot of cross validation and test errors vs number of features)

It is difficult to see a pattern here due to the issues that were faced with regards to singular matrices. However, for the second basis expansion function, we can see that the cross validation error is consistently lower than the test error for all models.

COMPENG 4SL3 Assignment 2
Uday Sharma
400139246

These basis functions were chosen in regard to getting results as opposed to what may or may not have been better as per the previous discussion with regard to the singular matrix error. Experimentation was done between different functions for Basis Expansion in order to determine what resulted in a singular matrix and what did not.

COMPENG 4SL3 Assignment 2
Uday Sharma
400139246

References:

1. “API Reference.” *Scikit*, https://scikit-learn.org/stable/modules/classes.html.
2. *YouTube*, YouTube, 5 Feb. 2018, https://www.youtube.com/watch?v=e8bBlie6N58. Accessed 24 Oct. 2021.