Joseph Loss (loss2)

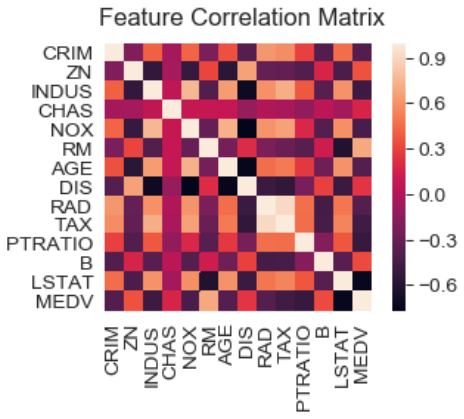
IE598 MLF F18

Module 4 Homework (Regression)

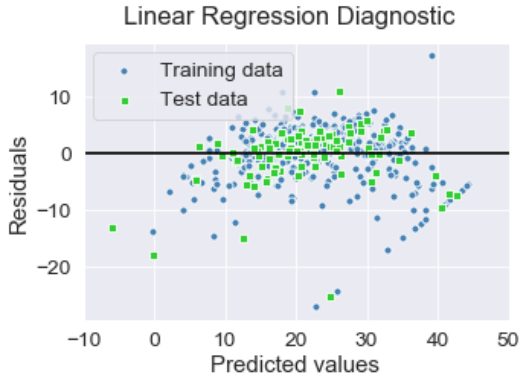
Part 1: Exploratory Data Analysis

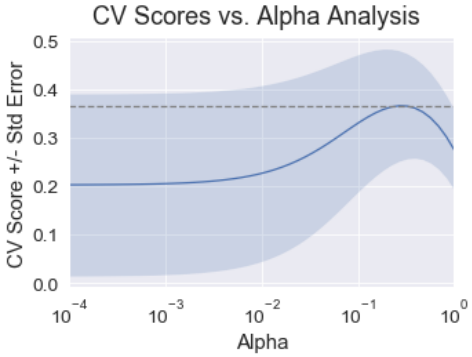
DF Head and Describe:

```
      CRIM    ZN  INDUS  CHAS    NOX  ...    TAX  PTRATIO       B  LSTAT  MEDV
0  0.00632  18.0   2.31     0  0.538  ...  296.0     15.3  396.90   4.98  24.0
1  0.02731   0.0   7.07     0  0.469  ...  242.0     17.8  396.90   9.14  21.6
2  0.02729   0.0   7.07     0  0.469  ...  242.0     17.8  392.83   4.03  34.7
3  0.03237   0.0   2.18     0  0.458  ...  222.0     18.7  394.63   2.94  33.4
4  0.06905   0.0   2.18     0  0.458  ...  222.0     18.7  396.90   5.33  36.2

[5 rows x 14 columns]
             CRIM          ZN  ...       LSTAT        MEDV
count  506.000000  506.000000  ...  506.000000  506.000000
mean     3.613524   11.363636  ...   12.653063   22.532806
std      8.601545   23.322453  ...    7.141062    9.197104
min      0.006320    0.000000  ...    1.730000    5.000000
25%      0.082045    0.000000  ...    6.950000   17.025000
50%      0.256510    0.000000  ...   11.360000   21.200000
75%      3.677082   12.500000  ...   16.955000   25.000000
max     88.976200  100.000000  ...   37.970000   50.000000

[8 rows x 14 columns]
```



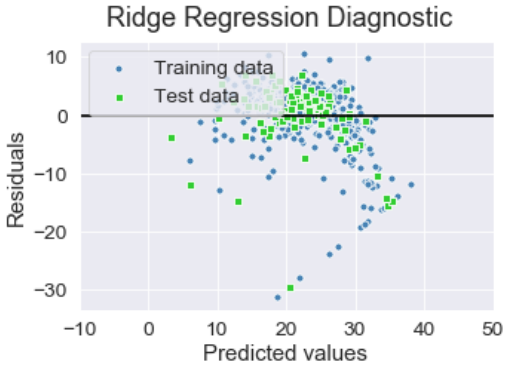
Part 2: Linear regression





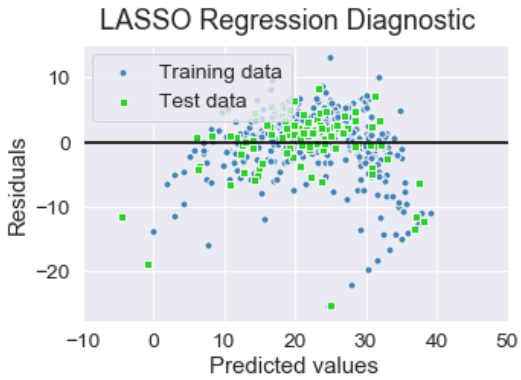
Slope: -0.113
y-Intercept: 30.247
MSE train: 21.641, test: 24.291
R^2 train: 0.751, test: 0.669

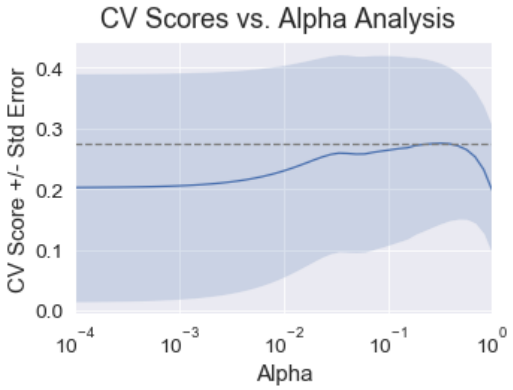
Part 3.1: Ridge regression



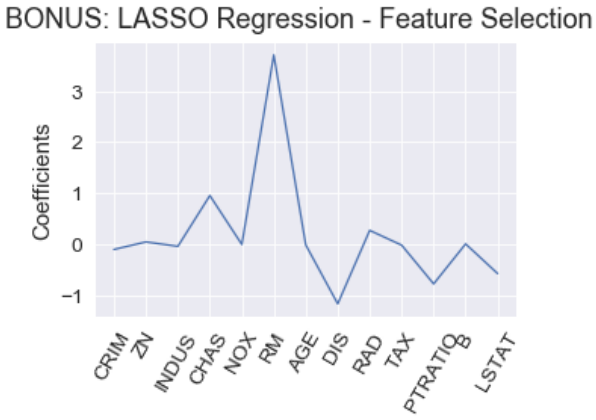
Slope: -0.061
y-Intercept: 19.319
MSE train: 30.794, test: 28.367
R^2 train: 0.646, test: 0.613
The optimal Ridge Alpha is: 0.26827

Part 3.2: LASSO regression

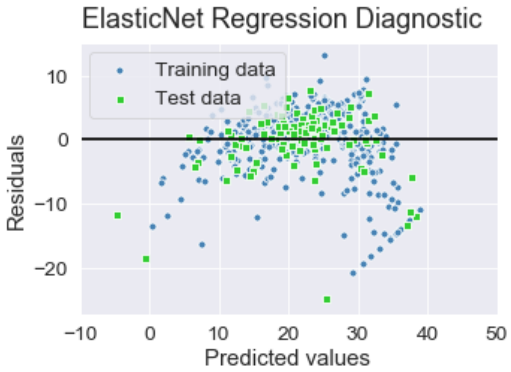


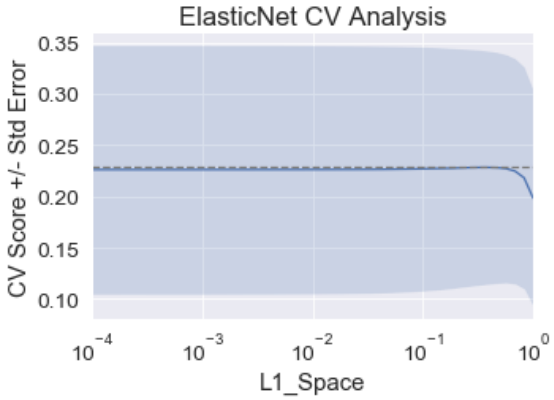


Slope: -0.077
y-Intercept: 34.936
MSE train: 26.417, test: 24.409
R^2 train: 0.696, test: 0.667
The optimal Lasso Alpha is: 0.32374



Part 3.3: Elastic Net regression





Slope: -0.093
y-Intercept: 39.911
MSE train: 26.861, test: 23.969
R^2 train: 0.691, test: 0.673
The optimal l1_ratio is: 0.39069

Part 4: Conclusions

I'm very interested to discuss my results as part of a class discussion and see what the other students found in their analysis. It seems that my models actually underfit the data, as each one returned a smaller MSE and R2 for the testing set than for the training set. I implemented the code as Raschka did in his book, but I believe that his results showed overfitting because he used a different size for his training and testing sets (in addition to a different random_state input, as we used 42 for our analysis).

Note 1: The sns.pairplot function was performed on all the features of the housing dataset, but was excluded from this report due to its extremely large size. Please execute my Python code to generate this graphic.

Note 2: The BONUS chart under the LASSO regression was something neat that I learned while reading about this type of regression online. Essentially, this graph is an indication of what features are most relevant/important to another (in this specific display, I used 'MEDV', and we can see that the 'RM' feature has the highest relation and is of the most importance to MEDV, versus the other variables of the dataset.

==========

My name is Joseph Loss
My NetID is: loss2
I hereby certify that I have read the University policy on Academic Integrity and that I am not in violation.

Part 5: Appendix

https://github.com/chicago-joe/IE598_F18_HW4