Assignment-based Subjective Questions

Note: Two approaches were taken to solve this assignment.

Approach 1: I deployed l1 and l2 regularization right after completing EDA, cleaning up the data and RFE.

Approach 2:In the other approach, after RFE, I manually eliminated certain features looking at p values. After this, I deployed regularization.

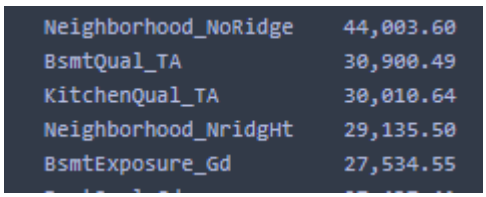
1. What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Solution:

- In both the approaches followed, optimum value of lambda(alpha) for ridge came out to be 6 whereas lasso gave 100.
- Impact of doubled lambda(alpha) on ridge: The training r2 score decreased by 0.006, whereas the r2 score on test data improved by 0.0009. There was minimal impact on the results. But what can be said is that since r2 score went down slightly on training, it can be said that the regularization would've been a little too harsh on some of the coefficients, reducing their impact on the output. But the test r2 score was good implying that the model is able to generalize well, but the impact is negligible. The other important observation was looking at the magnitudes of the coefficients:
- Impact of doubled lambda(alpha) on lasso: Here the r2 score decreased by 0.007 on the training data and improved by 0.0002 on test data. Since l1 norm does some feature selection, it would have resulted in certain coefficients' values being pulled towards zero. May be those features' impact had to be suppressed a bit more to generalize well on data. This helped its performance ever so slightly(0.0002) on test data. But this is not a noticeable, and we can say that accuracy of the model is pretty much unaffected.
- From a model complexity standpoint, with the added advantage of feature elimination of l1 regularization, a more aggressive lambda makes the model simpler. Our r2 score doesn't really change much either way, therefore a simpler model can always be chosen in the end.
- Impact on coefficients:

    Coefficients with optimum alpha:

- Ridge:



Neighborhood_NoRidge	44,003.60
BsmtQual_TA	30,900.49
KitchenQual_TA	30,010.64
Neighborhood_NridgHt	29,135.50
BsmtExposure_Gd	27,534.55

- Lasso:

Neighborhood_NoRidge	51,215.54
BsmtQual_TA	34,925.29
KitchenQual_TA	34,391.40
BsmtQual_Fa	32,742.32
KitchenQual_Fa	32,556.15

- Coefficients with doubled alphas:

- Ridge:

Neighborhood_NoRidge	36,912.33
Neighborhood_NridgHt	26,340.31
BsmtQual_TA	26,259.93
BsmtExposure_Gd	26,072.20
KitchenQual_TA	25,053.06

- Lasso:

Neighborhood_NoRidge	47,747.79
BsmtQual_TA	30,915.21
KitchenQual_TA	28,857.48
Neighborhood_NridgHt	28,591.48
BsmtExposure_Gd	27,088.64

Note: I have sorted the absolute value of coefficients. As you can see that there is more suppression on the coefficients with increased alphas. But since the performance was

2. You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Solution:

- In my runs, I was observing similar r2 scores on train and test data for both l1 and l2 regularization.
- But I would choose l1 regularization here, because as explained in the solution of the previous question, l1 regularization was able to identify few more features that are redundant. This allows the model complexity to go down without comprising on the performance of the model.

3. After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Solution:

- Before dropping the 5 most important features were:

```
Neighborhood_NoRidge   51,215.54
BsmtQual_TA            34,925.29
KitchenQual_TA         34,391.40
BsmtQual_Fa            32,742.32
KitchenQual_Fa         32,556.15
Name: Lasso_opt, dtype: float64
```

- After dropping the above 5 features we get:

```
Neighborhood_NridgHt   40,192.72
BsmtExposure_Gd        31,470.78
Neighborhood_Somerst   27,635.16
LandContour_HLS        24,033.60
Neighborhood_Edwards   23,217.29
```

Please note that I've sorted the coefficients after taking absolute value.

4. How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Solution:

- Firstly, we clean up the data and removed features that are not needed manually and via RFE. During clean up we also make sure that outliers are removed from the data. IQR analysis can help in this step. However, we must also pay attention to the amount of samples available to train our model on. If the data is less, dropping a lot of rows can give bad performance as well.
- A model can be claimed to be robust and generalizable, if train and test accuracy metric(here r2_score) are close to each other i.e. there isn't much of a deviation.
- Moreover, if there are outliers, its impact on the output gets bumped up noticeably. To address this problem(along with feature elimination), we employ l1/l2 regularization. These coefficients' magnitude gets reduced. l1 norm comes with an added advantage of feature elimination too.
- Therefore, EDA + data clean up + addressing outliers + removing redundant features and finally with regularization, we can make a model as generalizable and robust as possible.