Advance Regression Assignment – Part 2

Name: Bharat Hegde

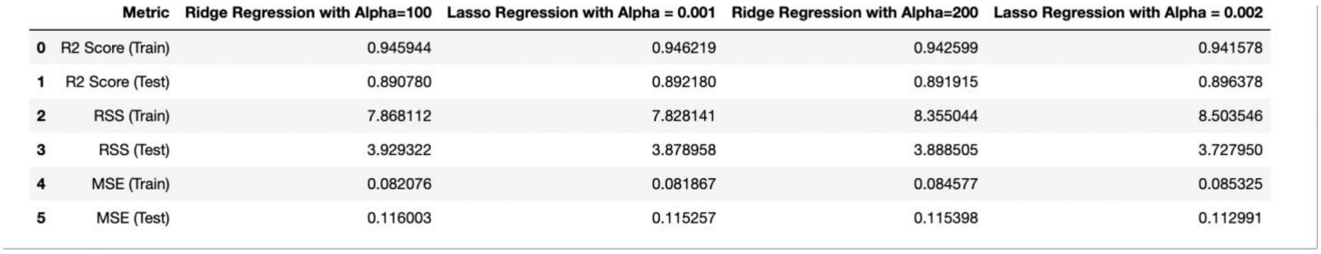
Question 1

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?

Ans:

The optimal value of Alpha for Ridge = **100**
The optimal value of Alpha for Lasso = **0.001**

The changes in the model if I double the alpha is explained in the below table:



	Metric	Ridge Regression with Alpha=100	Lasso Regression with Alpha = 0.001	Ridge Regression with Alpha=200	Lasso Regression with Alpha = 0.002
0	R2 Score (Train)	0.945944	0.946219	0.942599	0.941578
1	R2 Score (Test)	0.890780	0.892180	0.891915	0.896378
2	RSS (Train)	7.868112	7.828141	8.355044	8.503546
3	RSS (Test)	3.929322	3.878958	3.888505	3.727950
4	MSE (Train)	0.082076	0.081867	0.084577	0.085325
5	MSE (Test)	0.116003	0.115257	0.115398	0.112991

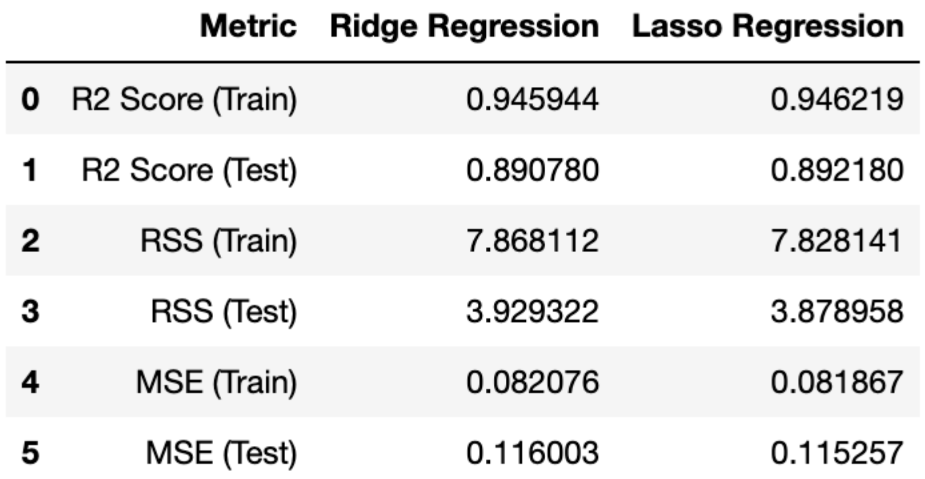
The most important factors after doubling the Alpha (considering Lasso coefficients)

GrLivArea	0.101276
OverallQual	0.055978
TotalBsmtSF	0.048721
OverallCond	0.035982
GarageArea	0.028575

Question 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?

Ans: Will choose the Lasso regression. The model metrics for Lasso is marginally better than Ridge. Also Lasso does the Feature selection, and coefficients for many features are made 0 and hence lesser features are used compared to Ridge.



	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.945944	0.946219
1	R2 Score (Test)	0.890780	0.892180
2	RSS (Train)	7.868112	7.828141
3	RSS (Test)	3.929322	3.878958
4	MSE (Train)	0.082076	0.081867
5	MSE (Test)	0.116003	0.115257

Question 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?

Ans: Before excluding the top 5 predictor variables, below were the top 5 variables.

```
GrLivArea                 0.084268
TotalBsmtSF               0.054251
OverallQual               0.049741
OverallCond               0.037623
2ndFlrSF                  0.028755
```

After excluding above from the model, the top 5 variables are:

```
1stFlrSF                  0.086184
GarageArea                0.037584
YearRemodAdd              0.037149
BsmtFinSF1                0.035614
Neighborhood_Crawfor      0.031107
```

Question 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?

Ans: To make sure that the model is robust and generalisable we need to make sure that model is neither overfitting or underfitting. That way, model will be able to provide good results on the unseen data in the future also.

If the model is too complex, the bias will be low, but the variance will be high, i.e any slight changes in the input there will be large changes in the output. If keep the model very simple then bias will be high then the accuracy of the model will be low on the training data itself. We will need to strike a balance between the bias and variance.

The accuracy will be high if the bias is low. But in order to make sure that model will not overfit, we will need to increase the bias to some extent, so that accuracy decreases. We will need to strike a fine balance between bias and accuracy.

We will use techniques like Lasso and Ridge regressions to make sure that we don’t overfit and thereby making the model more robust and generalisable.