Problem Statement - Part II

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?

Answer: The Optimal value for ridge and lasso regression is

- Ridge : 20
- Lasso : 0.001

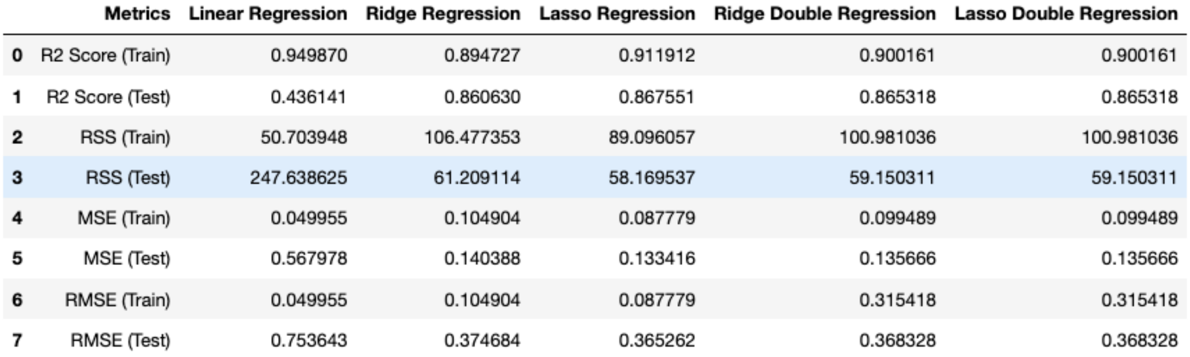
After changing the alpha values to doubled, i.e., Ridge: 40 and Lasso: 0.002

- Ridge R2 value changed from 0.89 to 0.90
- Lasso R2 value changed from 0.91 to 0.90
- And there is change in RSS, MSE and RMSE as well (given below).
- There is also a slight change in the most important top 5 predictors

The most important predictor after change is –

```
Ridge - GrLivArea (0.26028895194150586)
Lasso - OverallQual_10 (1.190015671994599)
```

Changes in metrics:



	Metrics	Linear Regression	Ridge Regression	Lasso Regression	Ridge Double Regression	Lasso Double Regression
0	R2 Score (Train)	0.949870	0.894727	0.911912	0.900161	0.900161
1	R2 Score (Test)	0.436141	0.860630	0.867551	0.865318	0.865318
2	RSS (Train)	50.703948	106.477353	89.096057	100.981036	100.981036
3	RSS (Test)	247.638625	61.209114	58.169537	59.150311	59.150311
4	MSE (Train)	0.049955	0.104904	0.087779	0.099489	0.099489
5	MSE (Test)	0.567978	0.140388	0.133416	0.135666	0.135666
6	RMSE (Train)	0.049955	0.104904	0.087779	0.315418	0.315418
7	RMSE (Test)	0.753643	0.374684	0.365262	0.368328	0.368328

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?

Answer:

I will choose the Ridge regression model as it performs better with the least RMSE and RSS when comparing to other models and due to its advantage of selecting features.

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?

Answer:

After dropping the top 5 features in the lasso model, when trained a new model with lasso(alpha=0.001), below are the five most important predictor variables with their coefficients.

```
RoofMatl_WdShngl (0.47399248143624517)
Exterior2nd_ImStucc (0.4116382781992408)
GrLivArea (0.349337291412178)
RoofMatl_CompShg (0.25448745970111936)
Exterior1st_BrkFace (0.23799742048083855)
```

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?

The model should be simple with less number of features as possible.
A generalized model will perform better on both test and training datasets. The implications of accuracy is that there would not be much difference between the training and test accuracy of a generalized model. In order to make sure that a model is robust, the outliers should be removed and the anomalies should be imputed. The data should be scaled and the model should be tuned. with the best hyperparameters