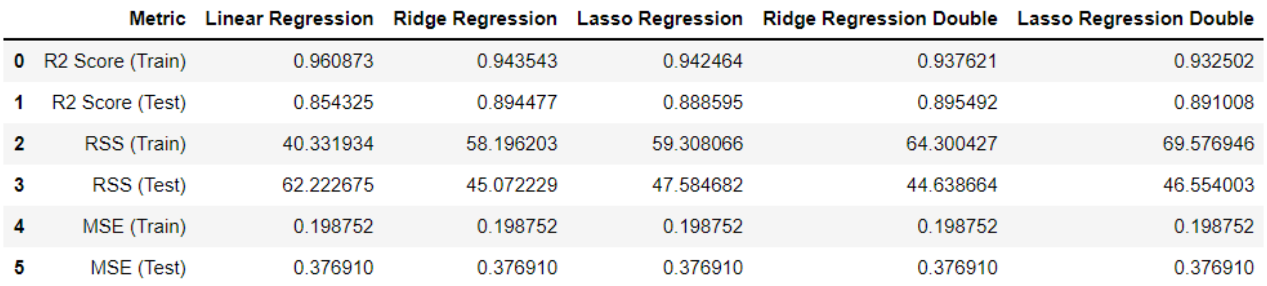
Subjective Questions - Suman Sourav Sahoo

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?

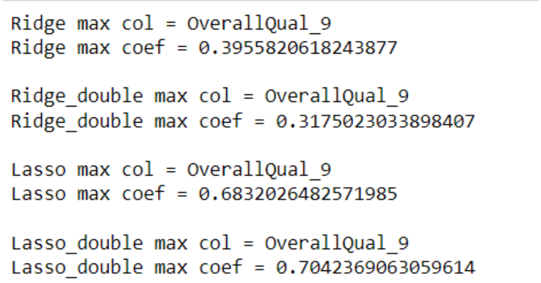
Ans :

- Optimal values of alpha for both ridge and lasso are:
  Ridge = 8
  Lasso = 0.001
- After doubling both the alphas i.e. 16 and 0.002,
  1. Ridge:
     a. Training R2 score decreases by a bit and test R2 score increases by a bit
     b. Train RSS increases and test RSS decreases
     c. Train and test MSE remain same
  2. Lasso:
     a. Training R2 score decreases by a but and test R2 score increases by a bit
     b. Train RSS increases and test RSS decreases
     c. Train and test MSE remain same



	Metric	Linear Regression	Ridge Regression	Lasso Regression	Ridge Regression Double	Lasso Regression Double
0	R2 Score (Train)	0.960873	0.943543	0.942464	0.937621	0.932502
1	R2 Score (Test)	0.854325	0.894477	0.888595	0.895492	0.891008
2	RSS (Train)	40.331934	58.196203	59.308066	64.300427	69.576946
3	RSS (Test)	62.222675	45.072229	47.584682	44.638664	46.554003
4	MSE (Train)	0.198752	0.198752	0.198752	0.198752	0.198752
5	MSE (Test)	0.376910	0.376910	0.376910	0.376910	0.376910

- Most important predictor variable for ridge and lasso before and after doubling the alpha is same i.e. “OverallQual_9”



```
Ridge max col = OverallQual_9
Ridge max coef = 0.3955820618243877

Ridge_double max col = OverallQual_9
Ridge_double max coef = 0.3175023033898407

Lasso max col = OverallQual_9
Lasso max coef = 0.6832026482571985

Lasso_double max col = OverallQual_9
Lasso_double max coef = 0.7042369063059614
```

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:

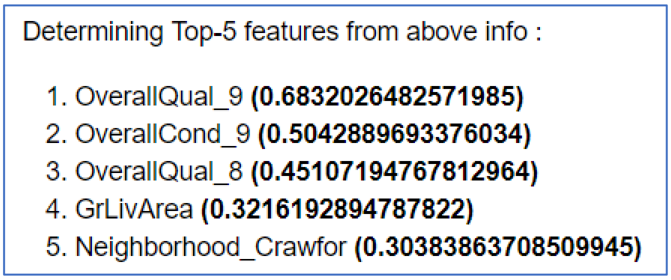
We use Lasso regression in this case because for a model having such a high number of features, feature selection becomes important and Lasso does that by equating the coefficients of many features to zero.

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:

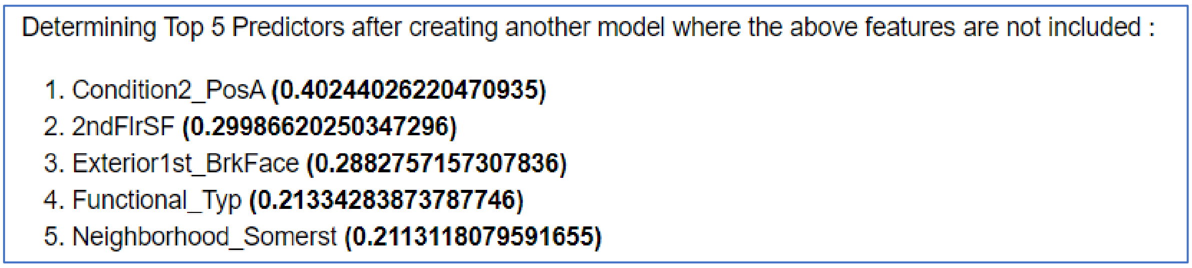
- Top 5 important variables in Lasso
  - OverallQual_9
  - OverallCond_9
  - OverallQual_8
  - GrLivArea
  - Neighborhood_Crawfor



Determining Top-5 features from above info :

1. OverallQual_9 **(0.6832026482571985)**
2. OverallCond_9 **(0.5042889693376034)**
3. OverallQual_8 **(0.45107194767812964)**
4. GrLivArea **(0.3216192894787822)**
5. Neighborhood_Crawfor **(0.30383863708509945)**

- Top 5 variables after creating another model where the above features are not included:
  - Condition2_PosA
  - 2ndFlrSF
  - Exterior1st_BrkFace
  - Functional_Typ
  - Neighborhood_Somerst



Determining Top 5 Predictors after creating another model where the above features are not included :

1. Condition2_PosA **(0.40244026220470935)**
2. 2ndFlrSF **(0.29986620250347296)**
3. Exterior1st_BrkFace **(0.2882757157307836)**
4. Functional_Typ **(0.21334283873787746)**
5. Neighborhood_Somerst **(0.2113118079591655)**

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 a model is robust and generalisable, we must ensure that our model is resistant to outliers and use more robust error metrics.
- To take care of the existing outliers in the data, we use various techniques to remove them
  - Capping the values at a certain threshold
  - Removing the outliers manually
  - Transforming certain values (exp, log etc.)

-END-