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?

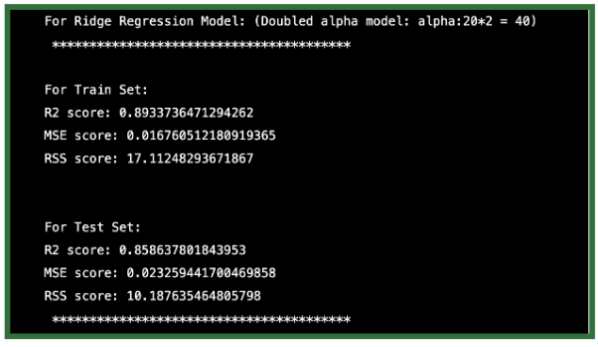
Answer:

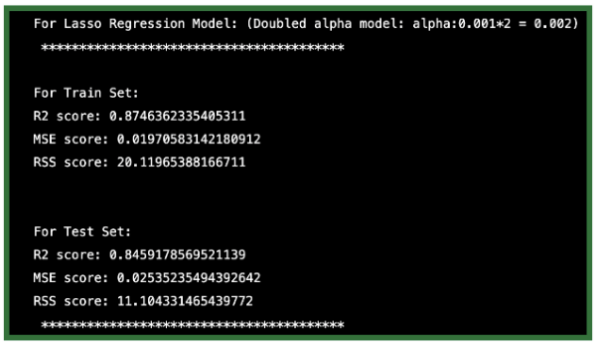
(i) Value of Alpha for Ridge: 20
Optimal Value of Alpha for Ridge: 0.001

Below is the actual values obtained from optimal alpha value:

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.961907	0.908017	0.891968
1	R2 Score (Test)	0.821533	0.864157	0.859201
2	RSS (Train)	6.113540	14.762358	17.338129
3	RSS (Test)	12.861657	9.789851	10.147052
4	MSE (Train)	0.077381	0.120244	0.130313
5	MSE (Test)	0.171361	0.149503	0.152206

(ii) The changes observed with the model for a doubled value of alpha:





Inference:

1. The R2 value becomes less with the doubled alpha value.
2. The penalty is not effectively applied to the cost function on both Lasso and Ridge

(iii) Important Predictor Variables Ridge and Lasso

```
For Ridge Regression (Doubled alpha model, alpha=20*2=40):
 ************************************************************************************************************************
The most important top10 predictor variables after the change is implemented are as follows:

 ['OverallQual_Excellent', 'CentralAir_Y', 'BsmtQual_Excellent', 'Neighborhood_Edwards', 'Neighborhood_Crawfor', 'OverallQual_V_Good', 'OverallQual_Below_Average', 'Neighborhood_NridgHt', 'Condition1_Norm',
'BsmtExposure_Gd']
 ************************************************************************************************************************
```

```
For Lasso Regression (Doubled alpha model: alpha:0.001*2 = 0.002):
 ************************************************************************************************************************
The most important top10 predictor variables after the change is implemented are as follows:

 ['OverallQual_Excellent', 'CentralAir_Y', 'BsmtQual_Excellent', 'OverallQual_V_Good', 'Neighborhood_Crawfor', 'Neighborhood_Somerst', 'Neighborhood_Edwards', 'OverallQual_Below_Average', 'Condition1_Norm',
'Neighborhood_NridgHt']
 ************************************************************************************************************************
```

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?

Alpha: Ridge: 20 and Lasso: 0.001

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.961907	0.908017	0.891968
1	R2 Score (Test)	0.821533	0.864157	0.859201
2	RSS (Train)	6.113540	14.762358	17.338129
3	RSS (Test)	12.861657	9.789851	10.147052
4	MSE (Train)	0.077381	0.120244	0.130313
5	MSE (Test)	0.171361	0.149503	0.152206

Inference:

1. R2 score for Ridge is slightly better than the Lasso as Ridge treated penalty effectively.
2. Train had a lot of variables. Hence it could be the reason of Ridge treatment produced better results than Lasso
3. RSS value is slightly higher in Lasso compared to Ridge.
4. MSE value is slightly higher in Lasso compared to Ridge

For this model, Ridge produced effective results as it had more features. When the model contains more features, ridge could be good for the prediction:

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?

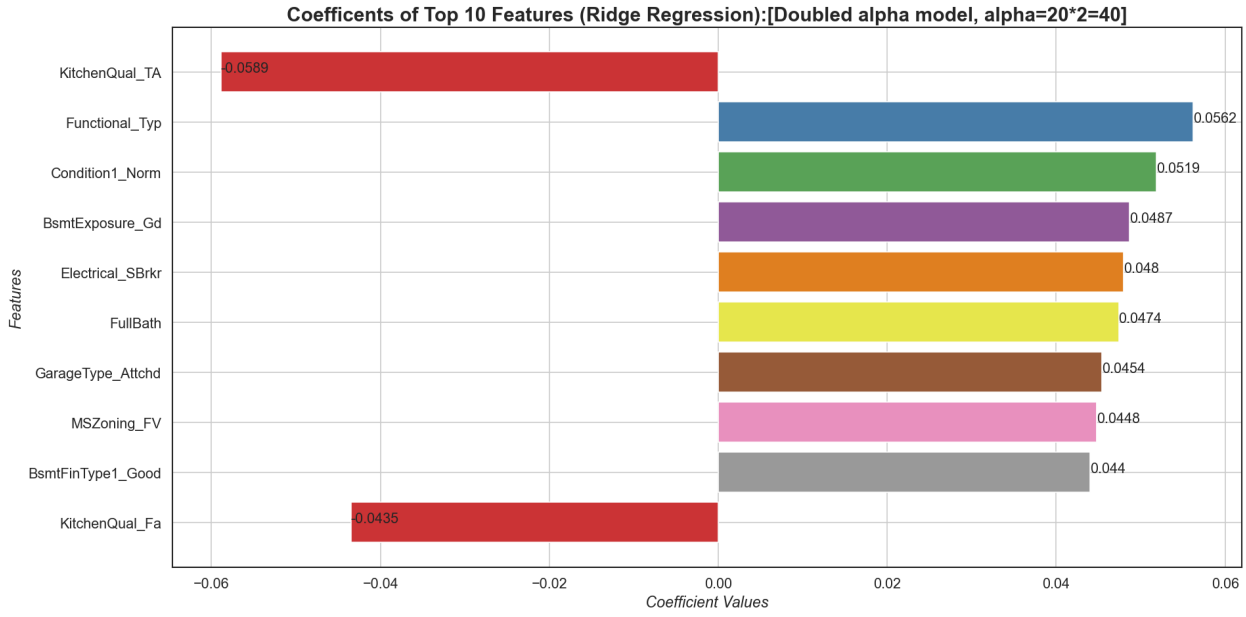
Please refer to the document Kishore_Advancd_Regression_Q3_ANSWER.ipynb

Before removing, Top 5 features in original Ridge model:

['OverallQual_Excellent', 'CentralAir_Y', 'BsmtQual_Excellent', 'OverallQual_V_Good', 'Neighborhood_Crawfor']

After removing the original variable, the new model gives the current five most important variables

['Functional_Typ','Electrical_SBrkr','Condition1_Norm','MSZoning_FV', 'BsmtFinType1_Good']



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?

1. Good models are always simple rather than simpler.
2. Simple models always have few errors in tests.
3. Simple model have high bias and low variance, where as complex ones are with low bias and high variance
4. Simple models are always less complex due to their high Alpha value.
5. The model becomes generalized if it is out of Underfitting and Overfitting. Underfitting is always the reason behind giving bad results in both train and test. Overfitting gives good results on the train and bad results in tests.
6. Hence the optimized model is the effective model which has the optimal value of alpha that minimizes the beta coefficients without any error term being high.
7. The Bias-Variance trade implements to have optimized model. Where it could trade one another to find the best fitted model.