Assignment-based Subjective Questions

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 to 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 is 8 and the optimal value of lasso is 0.001. If we double the alpha value for both ridge and lasso, the change in r2 score is not much but there is a visible change in their coefficients. R2 score on the train dataset of ridge and lasso were `0.9299162769875283 and 0.9227862283066279 respectively but after doubling the alpha value those changed to 0.9248157635562784 and 0.903281307612036.`

	Ridge Co-Efficient
Neighborhood_StoneBr	0.108729
Neighborhood_Crawfor	0.078849
Condition2_Norm	0.073274
Exterior1st_BrkFace	0.065115
OverallQual	0.065053
Neighborhood_NridgHt	0.060942
MSZoning_FV	0.055821
SaleCondition_Normal	0.053481
Condition1_Norm	0.044741
GrLivArea	0.041793

	Ridge Co-Efficient
Neighborhood_StoneBr	0.073322
OverallQual	0.066650
Neighborhood_Crawfor	0.060807
Condition2_Norm	0.049950
Exterior1st_BrkFace	0.049756
Neighborhood_NridgHt	0.047900
SaleCondition_Normal	0.046873
Condition1_Norm	0.040418
GrLivArea	0.040376
OverallCond	0.040285

	lasso Co-Efficient
Neighborhood_StoneBr	0.113231
Neighborhood_Crawfor	0.079119
GrLivArea	0.078501
OverallQual	0.073781
Neighborhood_NridgHt	0.056145
Exterior1st_BrkFace	0.055713
SaleCondition_Partial	0.054588
MSZoning_FV	0.052616
SaleCondition_Normal	0.046321
OverallCond	0.041819

	lasso Co-Efficient
OverallQual	0.081071
GrLivArea	0.073107
OverallCond	0.039755
TotalBsmtSF	0.035033
SaleCondition_Partial	0.034597
Neighborhood_Crawfor	0.028388
Condition1_Norm	0.027592
GarageArea	0.024143
SaleCondition_Normal	0.023059
TotRmsAbvGrd	0.022747

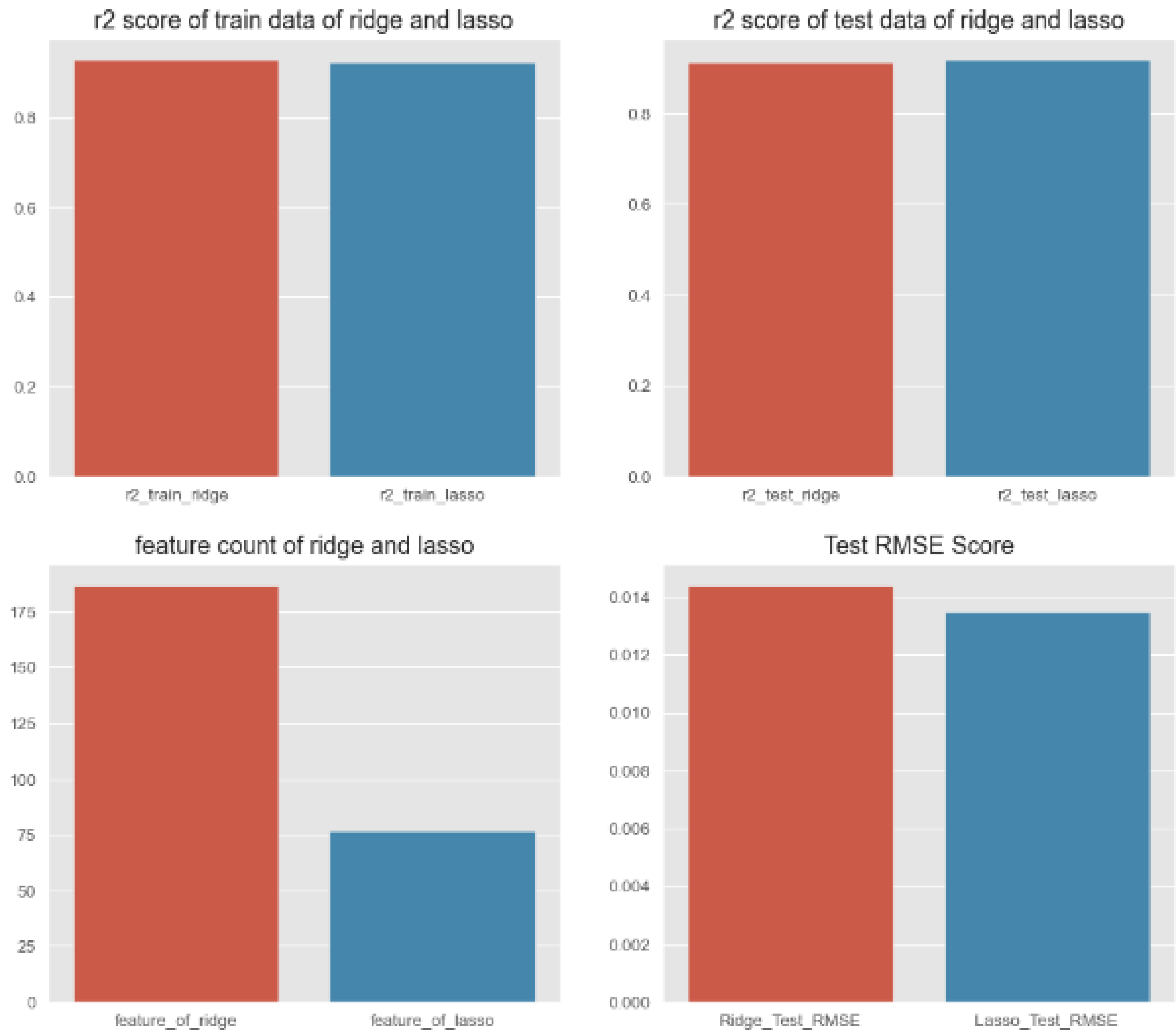
Question 2

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:

we can see the r2 score of train and test for both ridge and lasso is almost the same. but the number of features taken for the lasso is quite less compared to the lasso.

thus, **lasso regression** is the one we will choose.



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 removing the five most important predictors. We see a certain drop in r2 score value on the test data set and MSE also increased.

```
The R2 Score of the model on the test dataset is 0.9165618009972352
The MSE of the model on the test dataset is 0.014036923967351869
The most important predictor variables are as follows:
```

	Lasso Co-Efficient
MSZoning_FV	0.441142
MSZoning_RL	0.386218
MSZoning_RH	0.378404
MSZoning_RM	0.340910
Condition2_PosA	0.214520

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?

Answer:

To make the model more robust and generalizable, make the model simple but not simpler which will not be of any use.

As Per, Occam's Razor— given two models that show similar 'performance' in the finite training or test data, we should pick the one that makes less on the test data due to the following reasons:-

▪ Simpler models are usually more 'generic' and are more widely applicable

▪ Simpler models require fewer training samples for effective training than the more complex ones and hence are easier to train.

▪ Simpler models are more robust.

o Complex models tend to change wildly with changes in the training data set

o Simple models have low variance, and high bias, and complex models have low bias, high variance

o Simpler models make more errors in the training set. Complex models lead to overfitting — they work very well for the training samples but fail miserably when applied to other test samples.

Regularization can be used to make the model simpler. Regularization helps to strike the delicate balance between keeping the model simple and not making it too naive to be of any use.

Accuracy of the model can be maintained by keeping the balance between Bias and Variance as it minimizes the total error as shown in the below graph.