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:

Optimal value of alpha for ridge regression : 3.0
Optimal value of alpha for lasso regression : 0.0001
If we double the value of alpha for ridge and lasso regression and take it as 6.0 & 0.0002 then
For Ridge :
R2 score for training set changes from 0.822633 to 0.829878
R2 score for testing set changes from 0.815670 to 0.821800
MSE for train data reduced from 0.046234 to 0.045280
MSE for test data is reduced from 0.047829 to 0.047027

We can see very minor drop in the MSE for both train and test data set when ridge model is built using doubled alpha i.e. 6.0
The see that r2 score for test data is very close to training data when ridge model is built using doubled alpha
The predictor variable having highest importance is GrLivArea with coefficient value as 0.101399

For Lasso:

R2 score for training set changes from 0.824725 to 0.817995
R2 score for test data changes from 0.818436 to 0.810379
MSE for training data changes from 0.045961 to 0.046835
MSE for test data reduced from 0.047469 to 0.048511
We see that r2 score for training and test data is very close when lasso model is built using doubled alpha i.e. 0.0002
We see a slight increase in MSE when model is built using doubled alpha
The predictor variable having highest importance is GrLivArea having coefficient value as 0.431760

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:

For ridge we would use the doubled Alpha i.e. 6.0 as it is showing
R2 score being close for both training and test dataset
MSE is dropping for both train and test dataset when doubled Alpha is used to
build the model

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:

The top 5 important variables in the lasso model with lambda was 0.0001 (best
alpha param given by GridSearch) are

GrLivArea	0.432738
Neighborhood__NridgHt	0.088422
Neighborhood__NoRidge	0.083849
GarageCars	0.079940
Neighborhood__StoneBr	0.068230

When lasso model is built again using lambda 0.0002 then the variable
TotalBsmtSF is ranked 5th compared to previous one where it is ranked 6th.

GrLivArea	0.431760
Neighborhood__NridgHt	0.085579
Neighborhood__NoRidge	0.082048
GarageCars	0.079180
TotalBsmtSF	0.060317

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:

The model built using Lasso is robust and generalisable as we can see from
Linear Regression that even though it is giving good r2 square we see that the
residual analysis does not meet the assumptions of linear regression model.
The plot of residuals vs predictions shows that there is a pattern in residual

which indicates overfitting.
The regularisation addresses this, as we have used lasso, we can see that in the second model where we have build the lasso model using alpha 0.0002 we see that More coefficients are reduced to 0.