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:

Alpha(Ridge) : 10.0

Lambda(Lasso) : 0.001

If we double the values of lambda in both cases,

Changes in Ridge Regression metrics:

R2 score of train set decreased from 0.94 to 0.93

R2 score of test set remained same at 0.93

Changes in Lasso metrics:

R2 score of train set decreased from 0.92 to 0.91

R2 score of test set decreased from 0.93 to 0.91

Important predictor variables if alpha values are doubled(Ridge)

GrLivArea

OverallQual_8

OverallQual_9

Neighborhood_Crawfor

Functional_Typ

Exterior1st_BrkFace

OverallCond_9

TotalBsmtSF

CentralAir_Y

OverallCond_7

Important predictor variables if alpha values are doubled(Lasso)

GrLivArea

OverallQual_8

OverallQual_9

Functional_Typ

Neighborhood_Crawfor

TotalBsmtSF

Exterior1st_BrkFace

CentralAir_Y

YearRemodAdd

Condition1_Norm

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:

- The model we chose will depend upon the use case
- If our primary goal is variable selection then we will use Lasso
- If we don't want large coefficients then we can use Ridge regression.

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: If we remove the top 5 predictor variables and rebuild the model again the new top 5 predictor variables would be.

2ndFlrSF

Functional_Typ

1stFlrSF

MSSubClass_70

Neighborhood_Somerst

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: A model is robust when any variation in the data does not affect its performance much.

A generalizable model is able to adapt properly to new, previously unseen data, drawn from the same distribution as the one used to create the model.

To make sure a model is robust and generalizable, we have to take care it doesn't overfit. This is because an overfitting model has very high variance and a smallest change in data affects the model prediction heavily. Such a model will identify all the patterns of a training data, but fail to pick up the patterns in unseen test data.

In other words, the model should not be too complex in order to be robust and generalizable.

If we look at it from the prespective of Accuracy, a too complex model will have a very high accuracy. So, to make our model more robust and generalizable, we will have to decrease variance which will lead to some bias. Addition of bias means that accuracy will decrease.

In general, we have to find strike some balance between model accuracy and complexity. This can be achieved by Regularization techniques like Ridge Regression and Lasso.