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 only if it is not affected by small changes in data. When a model overfits the training data its variance becomes very high, and it does not perform well while predicting unseen data. Also, a model which has high bias will underfit on unseen data. So, we need to strike a balance between the variance and the bias of the model so that we can achieve the minimum possible prediction error. Regularization techniques like Ridge and Lasso regression help us to prevent model from overfitting and from becoming too complex.

Implications on Accuracy after regularization.

When we do regularization, we compromise a little bias for better performance on the test data. The training accuracy is decreased due to this approach, but model performs better on test data than a model which overfits.