Q1. 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?

Ans1.

The optimal value for alpha for:

Ridge : 10.0
Lasso : 0.001

After doubling the value of alpha for both ridge and lasso :

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 0.92

Changes in Lasso metrics:

R2 score of train set decreased from 0.92 to 0.90
R2 score of test set decreased from 0.92 to 0.91

Note: The coding part that will lead to the answer after doubling value of alpha is available in the jupyter notebook

Q2. You have applied the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Ans2.

The model we will choose here is Lasso as we are getting optimal value for alpha and the feature selection is happening as well, which would reduce multicollinearity. However, if the focus is that we don't want to get too large coefficients and reduction in coefficient magnitude, then we will use Ridge Regression.

Q3. 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 not have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Ans3.

After dropping our top 5 lasso predictors, we get the following new top 5 predictors:-

“2ndFlrSF” , “Functional_Typ”, “1stFlrSF”, “MSSubClass_70”, “Neighborhood_Somerst”

Note: The coding part that will lead to the answer after removing top 5 lasso predictors and considering next 5 is available in the jupyter notebook

Q4. 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?

Ans4.

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

II. 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.

III. 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.

IV. With 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.

V. We have to find strike some balance between model accuracy and complexity. This can be achieved by Regularization techniques like Ridge Regression and Lasso.