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:

For Ridge, the Best Alpha value = 5.0

For Lasso, the Best Alpha value = 0.001

If we double the values of alpha in case of :

Ridge: It will reduce the value of coefficients, but the count of coefficients remains the same.

Lasso: It will weed out the less important features by making their coefficients values to 0.

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 to and why?

Answer:

For Ridge, the Best Alpha value = 5.0

For Lasso, the Best Alpha value = 0.001

Our R2 scores were good in both of them, but we would choose Lasso because by doing so we will have a lean model, as Lasso removes less significant features by making their coefficients 0.

Question 3

After building the model, you realized 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:

Exterior1st_BrkFace	: 0.570317
Neighborhood_MeadowV	: 0.438211
RoofMatl_Metal	: 0.437051
MSZoning_RL	: 0.430528
BsmtFinSF2	: 0.391922

Question 4

How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

Answer: A model needs to be made robust and generalizable so that they are not impacted by outliers in the training data. The model should also be generalisable so that the test accuracy is not lesser than the training score. The model should be accurate for datasets other than the ones which were used during training. Too much weightage should not given to the outliers so that the accuracy predicted by the model is high. To ensure that this is not the case, the outlier analysis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which it does not make sense to keep must be removed from the dataset. This would help increase the accuracy of the predictions made bythe model. Confidence intervals can be used (typically 3-5 standard deviations). This would help standardize the predictions made by the model. If the model is not robust , it cannot be trusted for predictive analysis