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 values of alpha for both ridge and lasso are:
> Ridge = 8
> Lasso = 0.001

After doubling both Ridge and Lasso to 16 and 0.002
1. Ridge:
> a. Training R2 score decreases by a bit and test R2 score increases by a bit
> b. Train RSS increases and test RSS decreases
> c. Train and test MSE remain same

2. Lasso:
> a. Training R2 score decreases by a but and test R2 score increases by a bit
> b. Train RSS increases and test RSS decreases
> c. Train and test MSE remain same

Most important predictor variable for ridge and lasso before and after doubling the alpha is same i.e. "OverallQual_9"

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:
We use Lasso regression in this case because for a model having such a high number of features, feature selection becomes important and Lasso does that by equating the coefficients of many features to zero.

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:

Top 5 important variables in Lasso

1. RoofMatl_WdShngl 1.00188133088509
2. OverallQual_9 (0.7766964919635191)
3. Neighborhood_NoRidge (0.4182886990030101)
4. OverallQual_8 (0.37732673696292596)
5. GrLivArea   (0.3216192894787822)

Top 5 variables after creating another model where the above features are not included:

1. 2ndFlrSF (0.40133411433892)
2. Exterior1st_BrkFace (0.2613702219719933)
3. Exterior2nd_ImStucc (0.24125063719210066)
4. BsmtExposure_Gd (0.21461857660102496)
5. Functional_Typ (0.18922538668468167)

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:

To make sure that a model is robust and generalisable, we must ensure that our model is resistant to outliers and use more robust error metrics.
To take care of the existing outliers in the data, we use various techniques to remove them

- Capping the values at a certain threshold
- Removing the outliers manually
- Transforming certain values