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 regression, optimal value of lambda is 10.

For Lasso regression, optimal value of lambda is 0.001.

After the alpha value is doubled the ridge regression R2 score of train set data reduces from 0.94 to 0.93 and test set data's R2 score remains same at 0.93.

After the alpha value is doubled the lasso regression R2 score of train set data reduces from 0.92 to 0.91 and test set data's R2 score reduced from 0.93 to 0.91.

Most important predictor variables after the alpha value is doubled are as below:

- OverallQual_8
- GrLivArea
- TotalBsmtSF
- OverallQual_9
- CentralAir_Y
- Neighborhood_Crawfor
- Exterior1st_BrkFace
- Functional_Typ

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:

Selection of the model depends on the requirement. Suppose there are many number of variables our key focus is selection of feature; in that case we will use Lasso. And in cases where our main aspect is coefficient magnitude reduction and we want to avoid too large coefficients then Ridge Regression can be used.

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:

- After building a new model dropping top 5 important predictor variables , the new most important predictor variables are `2ndFlrSF`, `Functional_Typ`, `1stFlrSF`, `MSSubClass_70` and `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?

Answer:

We can decide that a model is robust based on below conditions.

- When performance is not much impacted by the variation in data.
- When a model is capable of adapting itself to new, previously unseen data which is drawn from the same distribution that is used to create the model, then the model is said to be a Generalizable one. It should not overfit.
- An overfitting model has high variance and even the smallest data change can result in affecting the model heavily. Such model definitely identify all the training set data points but it fails in picking up the patterns of unseen test data.

- Also a model can be considered robust and generalizable it should not be too complex. Also a highly complex model will have higher accuracy. So for a robust and generalizable model the variance should be decreased which in turn leads to some bias. When bias is added accuracy decreases.