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 :
The optimal value for Ridge regression is 5 and for Lasso Regression is 0.0001.

On choosing double the value of alpha for ridge and lasso regression, the model still is a good fit. The only change is that the beta coefficients reduce further towards zero. Which means the variance dropped with bias remaining constant.

In ridge regression, the coefficients tends to zero and in lasso regression, they can be zero.

After implementing the change, the most important predictor variable for ridge regression is 'Overall Quality' and for lasso regression is 'Above ground Living area(GrLivArea)'.

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:
The optimal value of lambda i have chosen is 5 for ridge and 0.0001 for lasso regression as these values provide the most optimal bias-variance tradeoff.

As by regularisation, we are adding the additional error, yet the R2 score value is not much reduced,RSS and MSE are low which means these values are not compromising much on the bias.

Although when we increase the value of lambda, it will add more penalty to our model to make it generalised that is making the model simpler without thinking about fitting all data of the dataset.

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:
The five most important predictor variables after excluding the five most important predictor variables are:

1- 1stFlrSF ⇒ First Floor square feet.
2- 2ndFlrSF⇒ Second floor square feet.
3- RoofMatl_WdShngl ⇒ Roof material(Wood Shingles).
4- PoolArea⇒ Pool area in square feet.
5- Neighborhood_NoRidge⇒ Physical locations within Ames city limits(Northridge).

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 the model is robust and generalisable, it has to be in its simplest form. This can be explained well with the help of bias-variance tradeoff.

For the simple models, the bias will be relatively high but the variance will be low and we need to balance this tradeoff to make the model optimal.

This affects the accuracy as the simple model makes more mistakes on the test sets yet they can generalise on most of the data. On the other hand if the variance is high, the model will overfit and will perform extremely poor on the testing dataset.