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 value of alpha:

Ridge: 20

Lasso: 0.001

When we double the alpha value, the error metrics get slightly worse. In general, increasing alpha increases bias and reduces variance of the model.

After increasing alpha, the important predictor variables are:

Ridge:

'Neighborhood_NridgHt', 'Neighborhood_NoRidge', 'BsmtQual_Ex', 'GarageCars', 'OverallQual', 'GrLivArea', 'Neighborhood_StoneBr', 'BsmtExposure_Gd', 'Condition1_Norm', '2ndFlrSF'

Lasso:

'Neighborhood_NoRidge', 'Neighborhood_NridgHt', 'Neighborhood_StoneBr', 'RoofMatl_WdShngl', 'BsmtQual_Ex', 'GrLivArea', 'KitchenAbvGr_1', 'Neighborhood_Crawfor', 'BsmtExposure_Gd', 'Neighborhood_Somerst'

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:

Since both Ridge and Lasso have similar error scores, it makes more sense to select Lasso as it helps reduce the number of variables and makes the model simpler.

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 top 5 predictor variables with the new model are:

'PoolQC_Ex', 'Street_Pave', 'GrLivArea', 'KitchenAbvGr_1', 'Exterior2nd_ImStucc'

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:

The simplest way to check for model generalizability is to score the model on a held-out test set and make sure the training and test scores are similar. This might mean that the training accuracy might decrease because an overfit model will always have better scores than a generalizable model on the training data.