SUBJECTIVE QUESTIONS

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 of alpha for ridge regression is 10 and for lasso is 0.001.

If we choose to double the value of alpha, then in Ridge regression, the r2 score of train data reduces from 87.82% to 81.32%. Similarly, the r2 score for test data also drops from 85.74% to 81.31%, and even the coefficients move more towards 0.

In Lasso regression, the r2 score for train data drops from 88.52% to 87.19%, and r2 for test data drops from 85.83% to 85.2%. Also, the number of non zero coefficient drops from 23 to 18.

Significant predictors remain the same after the implementation. Only the coefficient drops for each predictor in both the regression.

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 would go with the lasso regression because of the following reasons:

   a. r2 score for test and train is better in lasso regression than in ridge regression.
   b. In Lasso regression, few of the coefficients were equal to 0, which reduced the number of feature predictors and made the model less complex.

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: After removing the top 5 variables from both the model, the top predictors we obtain by ridge regression are as follows:

1) LowQualFinSF: -0.183
2) GrLivArea: -0.18
3) KitchenAbvGr: -0.155
4) GarageCars: -0.139
5) MSSubClass_1-STORY 1945 & OLDER: -0.138

For Lasso regression the top 5 predictors are:

1) LowQualFinSF: -0.258
2) GrLivArea: -0.242
3) KitchenAbvGr: -0.228
4) GarageCars: -0.176
5) MSSubClass_1-STORY 1945 & OLDER: -0.173

The r2 score also reduces in both cases. In Ridge Regression, the score for train data becomes 78.75% and for test data is 78.86%. In Lasso Regression, the score for train data is 85.15% and for test data is 82.21%.

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 outliers do not impact them in the training data. The model should also be generalizable 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 give 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. We can also use a more robust error metric which reduces the influences of outliers This would help increase the accuracy of the predictions made by the model.