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 to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

The optimal value of alpha for the ridge is 10.

The optimal value of alpha for the ridge is 100.

If we choose to double the alpha for the ridge and lasso i.e., 20 and 200

Coeff values for Ridge are rising as alpha rises, and r2 score of train data is likewise a decrease of.807 to 0.45

For Lasso, more features were deleted from the model as the alpha value rose. However, r2score in both test and train data fell by 1%.

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?

Lasso will be our choice for the feature selection option. Without impacting the accuracy of the model, it has eliminated undesirable features. Which makes our model accurate, straightforward, and general.

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?

Neighborhood NoRidge, Neighborhood NridgHt, 2ndFlrSF, OverallQual, and Neighborhood Veenker are the top 5 features. Model accuracy decreased from 80 and 81% to 55% and 58% when these features were dropped. Top features following the removal of the top five primary predictors are TotalBsmtSF, 1stFlrSF, MSSubClass 90, MSSubClass 120, HouseStyle 1Story.

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?

Three features must be present for the model to be reliable and generalizable:

1. Model accuracy should be greater than 70–75%, and in our instance, it is 80% (for the train) and 81%. (Test)

, which is accurate.

2. P-value for each feature is less than 0.05.

3. The VIF for all features is 5.

As a result, we are certain that the model is reliable and generalizable.