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 for ridge: 100
Optimal value of alpha for ridge: 10

After make the double alpha for ridge and lasso i.e. 20 and 200

For Lass: As alpha value increased more feature removed from model. But r2score is also dropped by 1% in both test and train data

For Ridge: Coeff values are increasing as alpha will increase . r2_score of train data is also drop from .807 to 0.45

Top feature: Neighborhood_NoRidge, Neighborhood_NridgHt, OverallQual, overallQual, Neighborhood_veenkar

Question2: 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:

If we have too many variables and one of our primary goal is feature selection, then we will use Lasso.
The model we will choose to apply will depend on the use case.
If we don't want to get too large coefficients and reduction of coefficient magnitude is one of our prime goals, then we will use Ridge Regression.

We observe that Both Ridge and Lasso give very similar results in terms of performance. Although Ridge model performs slightly(1%) better than lasso on the test dataset, we still decide to choose the Lasso model to apply finally. Lasso helps with feature elimination and as our dataset has over 130+ columns , so feature elimination can be an advantage in realising the most important predictor variables. Hence, our final model is Lasso with r2 score of 88 on train and 84 on test datasets respectively.

Question3: 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 in our Lasso model are –
'GrLivArea','GarageType_Attchd','MSZoning_RM','SaleType_New','TotalB

smtSF'. If we remove these and rebuild the model, the five most important predictor variables now are–
MasVnrArea ,Neighborhood_StoneBr ,Neighborhood_NridgHt,Fireplaces ,GarageArea

Question4: 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: By making sure the model is not over fitting and is as simple as possible, we are ensuring that it is robust and generalizable. The accuracy of the model will go up if we try to over fit the model but that no longer makes it generalizable. When the model is generalized the accuracy should be pretty good on both the training and the testing dataset making the model robust.

To make model robust and generalisation 3 feature are required
1. Model accuracy should be > 70-75% : I Our case it's comming 80% (Train) and 81%(Test) which is correct.
2. P-value of all the feature is < 0.05
3. VIF of all the feature are < 5

Thus we are sure that model is robust and generalisable.