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?

Answer:

Optimal value of alpha for ridge: 7
Optimal value of alpha for ridge: 100
After making the double alpha for ridge and lasso i.e. 14 and 200
For Ridge: Coeff values are increasing as alpha will increase. r2_score of train data is also a slightly dropped
For Lasso: As alpha value increased more features were removed from model. But r2score is also dropped slightly in both test and train data
The most important variable after the changes has been implemented for ridge regression are as follows: -

1. MSZoning_FV
2. MSZoning_RL
3. Neighborhood_Crawfor
4. MSZoning_RH
5. MSZoning_RM
6. SaleCondition_Partial
7. Neighborhood_StoneBr
8. GrLivArea
9. SaleCondition_Normal
10. Exterior1st_BrkFace

The most important variable after the changes has been implemented for lasso regression are as follows: -

1. GrLivArea
2. OverallQual

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:

We will choose Lasso as its giving feature selection option also. It has removed unwanted features from model without affecting the model accuracy. Which makes are model generalized and simple and accurate.

It is important to regularize coefficients and improve the prediction accuracy also with the decrease in variance and making the model interpretably. Ridge regression uses a tuning parameter called lambda as the penalty is square of magnitude of coefficients which is identified by cross validation. Lasso regression uses a tuning parameter called lambda as the penalty is absolute value of magnitude of coefficients which is identified by cross validation. As the lambda value increases Lasso shrinks the coefficient towards zero and it make the variables exactly equal to 0. Lasso also does variable selection. When lambda value is small it performs simple linear regression and as lambda value increases, shrinkage takes place and variables with 0 value are neglected by the model.

Question 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 3:

Top 5 features are Neighborhood_NoRidge, Neighborhood_NridgHt, 2ndFlrSF, OverallQual, Neighborhood_Veenker. After dropping them model accuracy reduced from 80 and 81% to 55% and 58%. Now topmost features are: Next top 5 features after droping 5 main predictors 1stFlrSF, MSSubClass_90, MSSubClass_120, TotalBsmtSF, HouseStyle_1Story

Question 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 4:

To make model robust and generalizable 3 features are required:

1. Model accuracy should be > 70-75%: I our case its coming 80%(Train) and 81%(Test) which is correct.
2. P-value of all the features is < 0.05

3. VIF of all the features are < 5

Thus, we are sure that model is robust and generalizable.
The model should be as simple as possible, though its accuracy will decrease but it will be more robust and generalizable. It can be also understood using the Bias-Variance trade-off. The simpler the model the more the bias but less variance and more generalizable. Its implication in terms of accuracy is that a robust and generalizable model will perform equally well on both training and test data i.e., the accuracy does not change much for training and test data. Bias: Bias is error in model when the model is weak to learn from the data. High bias means model is

unable to learn details in the data. Model performs poor on training and testing data. Variance: Variance is error in model when model tries to over learn from the data. High variance means model performs exceptionally well on training data as it has very well trained on this of data but performs very poor on testing data as it was unseen data for the model. It is important to have balance in Bias and Variance to avoid overfitting and under-fitting of data