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?

Soln:

Optimal alpha for Lasso regression : 0.001
Optimal alpha for Ridge regression : 20.0

If the alpha values are doubled then,

For lasso regression,

1. R2_score on test dataset changes from 0.825 to 0.837
2. R2_score on train dataset changes from 0.915 to 0.891
3. 5 Most important predictors now are Neighborhood_Crawfor,OverallQual,Neighborhood_NridgHt,Condition1_Norm, Neighborhood_Somerst

In this case, we observe improvement in the r2_score on the test set but degradation in r2_score on the train set.

For lasso regression,

1. R2_score on test dataset changes from 0.860 to 0.854
2. R2_score on train dataset changes from 0.916 to 0.908
3. 5 Most important predictors now are OverallQual,Neighborhood_Crawfor,Neighborhood_NridgHt,OverallCond, BsmtFullBath

In this case, we observe degradation of r2_score on the test and train set.

Increasing alpha increases bias and reduces variance of the model.

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?

Soln:

Both Ridge and Lasso have similar r2_scores but Lasso is significantly simpler compared to ridge regression since the insignificant coefficients are converted to complete zeroes. So, It is beneficial to select Lasso for its simplicity but where even a 2 percent increase in r2_score matters, there ridge regression can be selected.

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?

Soln:

For Lasso regression, the 5 most important predictors for new model are:
MSZoning_FV,Exterior1st_BrkFace,Functional_Typ,BsmtExposure_Gd,Neighborhood_StoneBr

For Ridge regression, the 5 most important predictors for new model are:
PoolQC_None,RoofMatl_Membran,MSZoning_RH,MSZoning_FV,MSZoning_RM

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?

Soln:

The model is robust when it performs as well on the test data as it performs on the training data .Since, a model can overfit the data from training data, it needs to be checked with an example testing data.This is usually done by partitioning the given data randomly into train and test sets. The model performs on the test set and the robustness can be demonstrated. Usually, the complex model has higher accuracy but lesser variance. The lesser number of features make a model generalisable which means reducing the complexity of the model.Reducing the complexity reduces accuracy but the compromise in accuracy will result in increase in variance. Right balance needs to be found between Accuracy and variance for obtaining a proper model.