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?

ANS 1

The optimal value of alpha for Ridge is 3 and for Lasso is 0.0001. The magnitude of the coefficients decreases as we choose to double the value of alpha for both ridge and lasso.

The top ten predictor variables will largely remain unchanged. Some slight shuffling can be seen before and after doubling the value of alpha but the top two variables remain at there original position, i.e., MSZoning_RL and 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?

ANS 2

The Mean Squared error in case of Ridge and Lasso are:

Ridge - 0.01586

Lasso - 0.01579

The Mean Squared Error of Lasso is slightly lower than that of Ridge

Lasso helps in feature reduction (as the coefficient value of one of the feature became 0), Lasso has a better edge over Ridge. Therefore, the preferred option will be to choose Lasso over Ridge.

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?

ANS 3

The five most important predictor variables now will be :-

2ndFlrSF, 1stFlrSF, OverallCond, TotalBsmtSF and BsmtFinSF1

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?

ANS 4

To negate impact of outliers in the training data a model should be robust and generalisable. Also, the model should be generalisable so that the test accuracy is not lesser than the training score. It should be accurate for datasets other than employed during training. To ensure high accuracy is predicted by the model by not providing too much weightage to outliers. This can be done by outlier analysis and only relevant ones should be retained while all others be dropped from the dataset. This will ensure better accuracy of predictions made by the model. Confidence intervals typically three to five standard deviations can be used. It will help standardise the predictions made by the model. If the model is not robust, it will be unreliable for predictive analysis.