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 lambda for Ridge Regression = **4**
- Optimal value of lambda for Lasso Regression= **0.001**

Effect on train set:

In Ridge regression the R squared value will decrease from 0.954 to 0.950 once the lambda will be doubled and in the case of Lasso R squared value will decrease from 0.957 to 0.953.

Effect on test set:

In Ridge regression the R squared value will decrease from 0.886 to 0.884 once the lambda will be doubled and in the case of Lasso R squared value will be constant.

So, the most important predictor variables after we double the alpha values are as below :

1) GrLivArea
2) OverallQual_8
3) OverallQual_9
4) Functional_Typ
5) Neighborhood_Crawfor
6) Exterior1st_BrkFace
7) TotalBsmtSF
8) CentralAir_Y

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: Ridge and Lasso regularize the coefficients and improve the prediction accuracy with decrease in variance and bias.

Ridge Regression: Ridge regression uses the hyperparameter called lambda as a penalty to the square of the magnitude of coefficients. Penalty is the lambda times the square of the coefficients so if the lambda will penalise the sum of squares of the coefficients. Ridge regression keeps all the variables unlike Lasso Regression.

Lasso Regression: Lasso doesn't keep all the variables and as lambda increases the coefficients starts shrinking and becomes zero for few of coefficients which doesn't impact the response variable.

Based on analysis R squared value for train and test is better in the case of Ridge rather than Lasso. But Lasso regression brings the insufficient variables coefficients to zero. Lasso is more robust and simpler.

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?

Answer: After dropping our top 5 lasso predictors, we get the following new top 5 predictors: -

1) 2ndFlrSF
2) Functional_Typ
3) 1stFlrSF
4) MSSubClass_70
5) Neighborhood_Somerst

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?

Answer: The simpler the model the more bias but less variance and more ssgeneralizable though accuracy will decrease. It can be understood through the Bias-Variance trade-off as well. The model should perform equally good on train and test data in term of robust and generalizable to achieve better accuracy. To make sure a model is robust and generalizable, we have to take care it doesn't overfit. This is because an overfitting model has very high variance and a smallest change in data affects the model prediction heavily. Such a model will identify all the patterns of a training data but fail to pick up the patterns in unseen test data.

If we from the viewpoint of **Accuracy**, a too complex model will have a very high accuracy. So, to make our model more robust and generalizable, we will have to decrease variance which will lead to some bias. Addition of bias means that accuracy will decrease. We have to find strike some balance between model accuracy and complexity. This can be achieved by Regularization techniques like Ridge and Lasso regression. Ridge and Lasso regression are some of the simple techniques to reduce model complexity and prevent over-fitting which may result from simple linear regression.