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. Optimal Value of alpha and lasso regression are 6 and 100

2. If we double the value of alpha for Ridge to 12 and Lasso to 200 than there is not much difference in the r2 values of both the models

- Ridge Regression train r2: 0.8126
- Ridge Regression test r2: 0.8033
- Lasso Regression train r2: 0.8056
- Lasso Regression test r2: 0.803

3. The most important predictor variables also remain the same

Before and after doubling alpha for Ridge

- MSSubClass
- Fireplaces
- LandContour_Lvl

Before and after doubling alpha for Lasso

- OverallQual
- Neighborhood_NridgHt
- Exterior1st_CemntBd

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: We can use Lasso in this case as LASSO gives feature selections option by setting some feature coefficients exactly to 0, i.e., removing them from the model.

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: R2 value of the new model without top 5 earlier predictors now reduced to

Lasso Regression train r2: 0.789

Lasso Regression test r2: 0.7564

and top 5 predictors now are

- ExterQual
- SaleType_ConLI
- Exterior1st_HdBoard
- Neighborhood_Veenker
- Neighborhood_OldTown

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: A model is said to be generalisable if it has not been overfitting the training data, that is it doesn't memorize the training data. When it come across the unseen data that is test data, it should give the reasonable response with acceptable error. Model is considered as robust if the result is consistently accurate even if some of the variables changed

A model can be made robust and generalisable by having the balance between overfitting, underfitting and accuracy.

Regularization is one such technique that reduces the overfitting by penalizing the coefficients that are large.

The implications of keeping the model robust and generalisable brings down the accuracy score of the model on the training data set whereas will be more consistent on the test data set. This is because we compromise on the complexity of the model to make it more generalisable.

Below diagram shows that there is a trade-off between bias and variance with respect to model complexity. A simple model would usually have high bias and low variance, whereas a complex model would have low bias and high variance. In either case, the total error would be high.