Problem Statement - Part II

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?

For given dataset and Model,
Ridge : Optimal Value of alpha is 2.0, As we increase the value of lambda the variance in model is dropped
Lasso : Optimal Value of alpha is 0.01,when we increase alpha the model try to penalize and most of the coefficient value zero.

When we double the alpha value for ridge, the model Predictors are same but the coefficient of these predictor has changed,

the model will apply more penalty on the curve. Similarly when we increase alpha for lasso more coefficient of the variable will reduced to zero. R2 also decreases.
The most important variable for ridge: MSZoning_FV, MSZoning_RL, Neighborhood_Crawfor, MSZoning_RH, MSZoning_RM, SaleCondition_Partial, Neighborhood_StoneBr, GrLivArea, OverallQual
The most important variable for lasso: GrLivArea, OverallQual, OverallCond, TotalBsmtSF, BsmtFinSF1

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?

Will be using Lasso due to below reasons,

a) Lasso gives a better model with close R2 score between train and test data

meaning varience is low and also bias is also low.

b) Lasso also does variable selection. So better to go with Lasso Regression 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?

Those 5 most important predictor variables that will be excluded are :- 1. GrLivArea 2. OverallQual 3. OverallCond 4. TotalBsmtSF 5. 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?

The model should be simple and also accurate meaning low variance and low bias,

The generalizable model will perform equally well on both training and test data.

Better to have model which is highly stable meaning low variance and little bit acceptable compromise in accuracy meaning Bias.