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?

Answer:

The optimal values are as below:

Lasso:0.0004

Ridge:10

On doubling the Ridge alpha I'm get very similar values as per the first time. Slight difference in MSE(Earlier- 0.0132948,Now- 0.0136 204).

On doubling the Lasso alpha I'm get very similar values as per the first time. Slight difference in MSE(Earlier- 0.0132060, Now- 0.0139 213).

Important variables post this change would be:

MSZoning_FV
MSZoning_RL
GrLivArea
Foundation_PConc
SaleCondition_Normal
GarageCars

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:

The optimal values are as below:

Lasso:0.0004

Ridge:10

MSE for Lasso and Ridge are as below:

Lasso: 0.0132060

Ridge: 0.0132948

The MSE of Lasso is slightly lower than that of Ridge, also Lasso is useful in feature reduction. Therefore, variables predicted by Lasso can be applied to choose significant variables for predicting the price of a house.

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 excluding the important predictor variables attained in the prior model, the important predictor variables now will be Lot area, Lot shape, Remodeled, Overall condition.

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:

Test and Training accuracy should be close

Model significance should be determined using P-values,R2 and adjusted R2.

Outliers should not be impacting the model. Therefore, outliers treatment is must using Boxplot, z scores.

Implications for Accuracy:

If outliers are not treated the data given by the model can't be trusted. They can affect the mean and median values which are used to replace null and NA values.

Abundance of training data set is important instead of relying on correlations

Selecting significant variables only

The model shouldn't over pr under fit.