Subjective Questions – Advanced Regression

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:

- Optimal Value of alpha for Ridge and Lasso Regression:

Optimal value of alpha for Ridge Regression : "6.0"

Optimal value of alpha for Lasso Regression : "0.001"

- Changes observed when the alpha for both ridge and lasso are doubled:

	Metric	Linear Regression	Ridge Regression	Lasso Regression	Ridge 2x	Lasso 2x
0	R2 Score (Train)	0.974452	0.967836	0.959609	0.966083	0.952749
1	R2 Score (Test)	0.943502	0.947092	0.944523	0.947003	0.938574
2	RSS (Train)	4.099018	5.160547	6.480546	5.441751	7.581104
3	RSS (Test)	4.087863	3.828103	4.013962	3.834506	4.444371
4	MSE (Train)	0.004015	0.005054	0.006347	0.005330	0.007425
5	MSE (Test)	0.009312	0.008720	0.009143	0.008735	0.010124

Note: Ridge 2x and Lasso 2x represents the values when the alpha values are doubled.

- We observe that overall accuracy of a model in both Ridge and lasso regression decrease slightly.
- In our case, it is “MSZoning_FV” is the most important predictor variable.

  The five important predictors are:
  "MSZoning_RL",
  "MSZoning_FV",
  "SaleCondition_Normal",
  "MSSubClass_70",
  "OverallCond”

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?

Lasso has feature elimination feature by doing coefficient zero. Therefore the model would be more robust in nature. So Lasso Regression is applied.

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?

The five most important variables obtained are

1. "Neighborhood_Crawfor"
2. "Neighborhood_ClearCr"
3. "FireplaceQu_Gd"
4. "FireplaceQu_TA"
5. "Exterior1st_BrkFace"

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 generalized so that the test accuracy is not lesser than the training score.
- The model should be accurate for datasets other than the ones which were used during training. Too much importance should not be given to the outliers so that the accuracy predicted by the model is high.
- To ensure that this is not the case, the outliers analysis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which it does not make sense to keep must be removed from the dataset. If the model is not robust, It cannot be trusted for predictive analysis.