Assignment 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?

- ➔ The optimal value for both Ridge and Lasso is 0.0001.
- ➔ After doubling the alpha value there is no changes for both Ridge and Lasso algorithms as the R2-score and MSE are almost for both the cases.
- ➔ Only minor change that can be highlighted is that the difference between the R2-score of train and the test set obtained after doubling alpha is less compared to the one without doubling.
- ➔ Important Predictors (Top 5- Considering absolute value of the coefficients):
    1. Ridge
        1. MSZoning_RL
        2. MSZoning_RH
        3. MSZoning_FV
        4. MSZoning_RM
        5. ExterCond_Fa
    2. Lasso
        1. MSZoning_RL
        2. MSZoning_FV
        3. MSZoning_RH
        4. MSZoning_RM
        5. BsmtQual_No Basement

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?

- ➔ Doubling the lambda value for Ridge and Lasso doesn't make much difference and the predictor variables obtained in both cases are same so I will be considering the optimal lambda value without doubling i.e., 0.0001 itself.

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?

➔ After excluding the five most predictor variables and building the Lasso model with optimal alpha value = 0.0001, the five most predictor variables now are:

1. KitchenQual_Fa
2. ExterCond_Fa
3. KitchenQual_TA
4. Neighborhood_OldTown
5. BsmtExposure_Gd

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?

➔ To make evaluate the model, metrics such as R2-score and Mean Squared Error can be used for both train and the test set.

➔ Ridge R2-Score:

1. Training - 0.8480382278040002
2. Testing - 0.839244564120565
3. Difference - 0.008793663683435149

➔ Lasso R2-Score:

1. Training – 0.8474381713379648
2. Testing – 0.8418410647769412
3. Difference - 0.005597106561023635

➔ From the R2-Scores above it can be said that the model is working pretty well as there is no overfitting of the data (seeing the Difference between the train and test is minimum).

➔ So the model obtained is robust and generalizable.