Assignment Part-II

Assignment Question Execution is performed on attached ipython notebook. Only results are mentioned in this document.

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?

Optimal value of alpha for ridge and lasso regression are:

ridge - 5.72

lasso - 123.28

Changes in the model if you choose double the value of alpha for both ridge and lasso shown below:

Ridge (alpha=11.44)

```
Number of non-zero Coefficients 428
MSE Train 341799658.645763
MAE Score Train 12487.671574347201
R2 Score Train 0.9413642588855341

MSE Test 604053693.7884015
MAE Score Test 16091.718095949287
R2 Score Test 0.8909195846471027

Lasso(alpha=246.56)

Number of non-zero Coefficients 73
MSE Train 403684180.9477941
MAE Score Train 13962.86500265628
R2 Score Train 0.9307479673331338

MSE Test 608605235.0792032
MAE Score Test 16192.663047127962
R2 Score Test 0.8900976643118711
```

- R2score of training data has decrease and it has increase on testing data
- Predictors are same but the coefficient of these predictor has changed

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 r2_score of lasso is slightly higher than lasso for the test dataset so we will choose lasso regression to solve this problem

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:

1stFlrSF, 2ndFlrSF, Neighborhood_Neighborhood_StoneBrook GarageArea, Neighborhood_Neighborhood_NorthridgeHeights are next five important predictor variables

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:

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 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.