Problem Statement - Part II

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?

Answer:

The coding solution for this question is present in the jupyter notebook from line 79 – 92.

◆   Optimal value of lambda for Ridge Regression = 10
◆   Optimal value of lambda for Lasso Regression = 0.001

If we double the values of alpha for both Ridge and Lasso
i.e Ridge(alpha=20) and Lasso(alpha=0.002) we get the following changes

Changes in Ridge Regression metrics:
        1. R2 score of train set decreased from 0.94 to 0.93

        2. R2 score of test set remained same at 0.93

Changes in Lasso Regression metrics:
        1. R2 score of train set decreased from 0.92 to 0.91

        2. R2 score of test set decreased from 0.93 to 0.91

So, the most important predictor variables after we double the alpha values are:-

        `GrLivArea`

        `OverallQual_8`

        `OverallQual_9`

        `Functional_Typ`

        `Neighborhood_Crawfor`

        `Exterior1st_BrkFace`

        `TotalBsmtSF`

        `CentralAir_Y`

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:

        1. The model we will choose to apply will depend on the use case.

        2. If we have too many variables and one of our primary goal is feature selection, then we will use **Lasso Regression**.

        3. If we don't want to get too large coefficients and reduction of coefficient magnitude is one of our prime goals, then we will use **Ridge Regression**.

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:

The coding solution for this question is present in the jupyter notebook from line 94 – 105.

After dropping our top 5 lasso predictors, we get the following new top 5 predictors:-

```
2ndFlrSF

Functional_Typ

1stFlrSF

MSSubClass_70

Neighborhood_Somerst
```

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:

1. A model is robust when any variation in the data does not affect its performance much.

2. A generalizable model is able to adapt properly to new, previously unseen data, drawn from the same distribution as the one used to create the model.

3. To make sure a model is robust and generalizable, we have to take care it doesn't overfit. This is because an overfitting model has very high variance and a smallest change in data affects the model prediction heavily. Such a model will identify all the patterns of a training data, but fail to pick up the patterns in unseen test data.

4. In other words, the model should not be too complex in order to be robust and generalizable.

5. If we look at it from the perspective of Accuracy, a too complex model will have a very high accuracy. So, to make our model more robust and generalizable, we will have to decrease variance which will lead to some bias. Addition of bias means that accuracy will decrease.

6. In general, we have to find some balance between model accuracy and complexity. This can be achieved by Regularization techniques like Ridge and Lasso Regression.