Problem Statement – 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:

Optimal Values:

- Optimal value of alpha for Ridge Regression = **10**
- Optimal value of alpha for Lasso Regression = **0.001**

Changes in Ridge Regression metrics:

- R2 score of train set remained same at 0.94
- R2 score of test set remained same at 0.94

Changes in Lasso Regression metrics:

- R2 score of train set decreased from 0.92 to 0.91
- R2 score of test set decreased from 0.91 to 0.89

So, the most important predictor variables after alpha values are doubles:

- GrLivArea
- OverallQual_8
- OverallQual_9
- Neighborhood_Crawfor
- Functional_Typ
- TotalBsmtSF
- Exterior1st_BrkFace
- YearBuilt

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:

- I will choose **Lasso Regression**.
- Lasso regression has feature section which has removed unwanted columns without affecting model accuracy.
- The R2 score for Lasso Regression model also did not varied significantly.

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 dropping our top 5 lasso predictors, we get new top 5 predictors:

- 2ndFlrSF
- 1stFlrSF
- TotalBsmtSF
- Exterior1st_BrkFace
- 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

- To make sure about model being **Robust** is the performance of the model should not be affected by any variation in input data.
- The model is able to classify or forecast new unseen data is called **generalizable**. After training dataset, when model is provided with new data set should perform well.
- We can make sure is the model is robust and generalized by checking is if it is not **overfitting**.
- When model is overfitted it has very high variance and smallest change in data affects the model prediction heavily. Such model will be able to identify all the patterns of training data, but will fail to classify patterns on unseen test data.
- The model should not be too simple so that it would not identify required patterns in any data set.
- The model should not be too complex and too simple in order to be robust and generalizable.
- **Accuracy** is a measure which shows how correctly the model has predicted True positives out of total actual positive, complex model will have 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.
- In general, we have to find some balance between model accuracy and complexity. This can be achieved by **Regularization** techniques like Ridge Regression and Lasso.