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 to 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 are:

- Optimal Value of lambda for Ridge: 10
- Optimal Value of lambda for Lasso: 0.001

After double the value of alpha:

- Ridge Regression
  - R2 score for train data decreased from 0.95 to 0.94
  - R2 score for test data remained the same 0.90
- Lasso Regression
  - R2 score for train data decreased from 0.93 to 0.92
  - R2 score for train data decreased from 0.89 to 0.80

In case of ridge that will lower the coefficients and in case of Lasso there would be more less important features coefficients turning 0.

The most important predictor variable after the change are remained significant.

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 model we will choose to apply will depend on the use case.
- If we have too many variables and one of our primary goal is feature selection, then we will use **Lasso**.
- 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:

After dropping top 5 lasso predictors, we got these 5 predictors

- 2ndFlrSF
- 1stFlrSF
- TotalBsmtSF
- 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:

- A model is **robust** when any variation in the data does not affect its performance much.
- A **generalizable** model can adapt appropriately to new, previously unseen data drawn from the same distribution as the one used to create the model.
- To ensure a robust and generalizable model, we must **ensure it is balanced**. An overfitting model has a very high variance, and the slightest change in data affects the model prediction heavily. Such a model will identify all the training data patterns but needs to pick up the ways in unseen test data.
- In other words, the model should be simple to be robust and generalizable.
- If we look at it from the accuracy perspective, a too-complex model will have a very high accuracy. To make our model more robust and generalizable, we will have to decrease variance, leading to some bias. The addition of discrimination means that accuracy will decrease.
- In general, strike some balance between model accuracy and complexity. Regularization techniques like Ridge Regression and Lasso can achieve this.