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:

1. Optimal alpha value with respect to the ridge regression model built is 9.0
2. In case of lasso the optimal alpha value is 0.0001.
3. Doubling the alpha value brought down the r2 score in both the models.
4. With the optimal alpha value "OverallQual" acted as an important predictor variable in case of ridge regression model, it remained the important predictor even after doubling the alpha value. However, the value of coefficient corresponding to feature is reduced.
5. With the optimal alpha value "GrLivArea" acted as an important predictor variable in case of lasso regression model, it remained the important predictor even after doubling the alpha value. However, the value of coefficient corresponding to feature is reduced.

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:

Ridge Regression:

- Training R2 Score: 0.88

- Test R2 Score: 0.84
- Number of Features with Coefficients < 0.001: 54 out of 350

Lasso Regression:

- Training R2 Score: 0.90
- Test R2 Score: 0.84
- Number of Features with Coefficients < 0.001: 225 out of 350

Even though there is minimal difference in the R2 scores, between the two models Lasso model seems to be a good option with its higher R2 score and most of the feature's coefficient value zeroed out. Lasso would be simpler than Ridge model.

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?

1. GarageCars
2. 2ndFlrSF
3. Neighborhood_NridgHt
4. OverallCond
5. KitchenQual_TA

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 build a robust model that works well in the real world, it starts with training the model with clean and well-prepared data. Using tricks like k-fold cross-validation ensures that we don't expose the test data while building the model, avoiding the trap of overfitting. Picking the right features and using techniques like L1 and L2 regularizations, plus a bit of hyperparameter tuning, will help fine-tune the model.

With a robust and generalizable model, we will be able to predict any unseen data in the future accurately.