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:
In the model created, we have used GridSearchCV to find the optimum value of alpha for Ridge and Lasso.

Based on the models generated the optimum values are as follows:

For **Ridge Regression** the value came as **0.2.**
For **Lasso Regression** the value came as **0.0001.**

On doubling the values of alpha, the R2 values for Train and Test slightly lowered for both ridge and lasso regression.

Metric	**Linear Regression**	**Ridge Regression (alpha=0.2)**	**Lasso Regression (alpha=0.0001)**	**Ridge Regression (alpha=0.4)**	**Lasso Regression (alpha=0.0002)**
R2 Score (Train)	8.55E-01	0.85518	0.851177	0.854736	0.845957
R2 Score (Test)	-1.15E+24	0.845036	0.843213	0.844063	0.83606
RSS (Train)	2.45E+00	2.448653	2.516342	2.456158	2.604601
RSS (Test)	9.10E+24	1.229052	1.243513	1.236766	1.300243
MSE (Train)	4.89E-02	0.048972	0.049645	0.049047	0.050508
MSE (Test)	1.44E+11	0.052912	0.053222	0.053078	0.054423

After the values of alpha are doubled, the new top predictor variables are as follows:

Ridge	**Lasso**
BsmtFinSF1	BsmtFinSF1
LotArea	OverallQual_9
OverallQual_9	OverallQual_3
OverallQual_2	2ndFlrSF
OverallQual_3	BsmtUnfSF

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

In the model generated, the ridge and lasso models are having very close R2 values. However Lasso model, generates a much simpler model as few of the features are eliminated by the model. So I would prefer the Lasso model because we would be having a much simpler 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?

We can choose Lasso as the final model based on the reason mentioned in question 2. Based on that, we can remove the columns ['BsmtFinSF1', 'OverallQual_9', 'LotArea', 'OverallQual_3', '2ndFlrSF'] which came as top columns in Lasso. After remove the columns, the new columns are :

1. Fireplaces
2. Neighborhood_StoneBr
3. Neighborhood_NridgHt
4. MSSubClass_30
5. OverallCond_3

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

Inorder to make the model robust and generalisable we can follow the below.

- Select only the relevant features and remove features which won’t help in predictions. This also involve removing any highly correlated features as well.
- Avoid overfitting and underfitting. Overfitted models won’t perform well with unseen data.
- Tuning hyperparameters help avoid overfitting and underfitting

- If the model is having high bias or high variance, then it might not perform well. High bias and low variance is underfit model and low bias and high variance is overfitted. Choose a model which is having an adequate bias and variance.

When we choose a model with low bias, then its overfitted model and may not work predict well for unseen data. If we try to increase the variance, then the bias need to be compromised. If we choose high bias, then the variance would be low and the model is underfit. We have to compromise between these two.