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 optimal value of LAMBDA we got in case of Ridge and Lasso is :

- **Ridge - 4.0 (lowest MSE value)**
- **Lasso - 0.0002**

Ridge

The output when alpha is 8:
The mean squared error value is 0.0037900269015508827
The r2 value of train data is 0.9452041481276376
The r2 value of test data is 0.866734002656996

Top predictor features when alpha is 8 are :

	Coefficient
OverallQual	0.087402
GrLivArea	0.084575
1stFlrSF	0.066656
TotalBsmtSF	0.060994
BsmtFinSF1	0.054475
2ndFlrSF	0.053555
Neighborhood_StoneBr	0.047511
Neighborhood_Crawfor	0.046191
LotArea	0.045719
GarageArea	0.044744

Lasso

The output when alpha is 0.0004:
The mean squared error value is 0.003963020587913445
The r2 value of train data is 0.9384520359553216
The r2 value of test data is 0.8653962719556855

Top predictor features when alpha is 0.0004 are:

	Coefficient
GrLivArea	0.239656
OverallQual	0.180856
TotalBsmtSF	0.065726
Neighborhood_Crawfor	0.064841
Neighborhood_StoneBr	0.061488
BsmtFinSF1	0.057596
GarageArea	0.050611
BsmtQual_Ex	0.049914
OverallCond	0.043671
Neighborhood_NoRidge	0.042211

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 optimal value of LAMBDA we got in case of Ridge and Lasso is :

- **Ridge - 4.0 (lowest MSE value)**
- **Lasso - 0.0002**

The r2 value for Ridge and Lasso is:

Ridge - Train = 0.947 , Test = 0.869

Lasso - Train = 0.945 , Test = 0.866

The Mean Squared error for Ridge and Lasso is:

Ridge - 0.00379

Lasso - 0.00392

Proceeding with lasso as coefficients are closer to 0 even though MSE is lower for ridge

LASSO Regularization the TOP 5 Predicted variables

- **GrLivArea**
- **OverallQual**
- **YearBuilt**
- **BsmtQual_Ex**
- **BsmtExposure_Gd**

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:

	Coefficient
2ndFlrSF	0.214363
1stFlrSF	0.187527
Neighborhood_StoneBr	0.100623
TotalBsmtSF	0.085750
OverallCond	0.081471

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.Use a training Data set that covers diverse scenarios of the problem being solved

2.Use feature selection appropriately. Select relevant features that capture the key aspects of the problem.

3.Use regularization techniches such as

1. Lasso(adds a penalty term proportional to the absolute value of the weights in the model.)
2. Ridge (regularization adds a penalty term proportional to the square of the weights in the model)

to shrink high values coefficients towards zero.

Regularization helps to achieve Bias-Variance trade off.It increase bias to an optimum position where total error is minimum.