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 value of alpha for Ridge Regression is alpha =4.0

Top 5 predictor variables are 1. GrLivArea, 2. Neighborhood_NoRidge,   ,3.1stFlrSF, 4. GarageCars,5. FullBath

```
GrLivArea: Above grade (ground) living area square feet
Neighborhood_NoRidge : Northridge
1stFlrSF: First Floor square feet
GarageCars: Size of garage in car capacity
FullBath: Full bathrooms above grade
```

Optimal value of alpha for Lasso Regression is alpha =0.0001

Top 5 predictor Variables are ### 1. GrLivArea, 2. Neighborhood_NoRidge, 3. LotArea   , 4. GarageCars ,5.   Neighborhood_NridgHt

```
GrLivArea: Above grade (ground) living area square feet
Neighborhood_NoRidge : Northridge
LotArea: Lot size in square feet
GarageCars: Size of garage in car capacity
Neighbourhood_NridgHt  Northridge Heights
```

When the alpha values are doubled for Ridge and Lasso Regression

When the alpha values are doubled the value of coefficients are reduced in both Rigde and lasso Regression. In Lasso more coefficients are reduced to zero. R2 score for both Train and test data is slightly reduced in both Lasso and Ridge Regression.

Ridge Regression alpha=8.0

Important predictor variables are 1.GrLivArea , 2. Neighborhood_NoRidge ,3. FullBath , 4. GarageCars 5.1stFlrSF

```
GrLivArea: Above grade (ground) living area square feet
Neighborhood_NoRidge : Northridge
FullBath: Full bathrooms above grade
GarageCars: Size of garage in car capacity
Neighbourhood_NridgHt  Northridge Heights
1stFlrSF: First Floor square feet
```

Lasso Regression alpha=0.0002

Important predictor variables are 1. GrLivArea, 2. Neighborhood_NoRidge, 3. GarageCars , 4. Neighborhood_NridgHt ,5. Neighborhood_StoneBr

GrLivArea: Above grade (ground) living area square feet
Neighborhood_NoRidge : Northridge
GarageCars: Size of garage in car capacity
Neighbourhood_NridgHt Northridge Heights
Neighbourhood_StoneBr Stone Brook

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 would choose Lasso Regression, because the the value of all parameters is almost same in lasso and Ridge Regression

		Ridge	Lasso
0	R2 Score (Train)	0.859417	0.865776
1	R2 Score (Test)	0.833147	0.824399
2	RSS (Train)	1.729873	1.651634
3	RSS (Test)	1.428656	1.503566
4	MSE (Train)	0.041162	0.040220
5	MSE (Test)	0.057112	0.058590

Though r2 score of Lasso is slightly less in Test data but the number of features are also less as many coefficient are equal to zero in Lasso 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:
Top 5 important predictor variables after building model with Lasso Regression were

1. GrLivArea, 2. Neighborhood_NoRidge, 3. LotArea , 4. GarageCars ,5. Neighborhood_NridgHt

After dropping these 5 features and building the model again,
We get for Ridge regression top 5 features now are
1stFlrSF ,FullBath , MasVnrArea, TotalBsmtSF ,Foundation_Stone
1stFlrSF: First Floor square feet
FullBath: Full bathrooms above grade
MasVnrArea: Masonry veneer area in square feet
TotalBsmtSF: Total square feet of basement area
Foundation_stone: Type of foundation Stone

And for Lasso Regression Model top 5 features now are
1stFlrSF,MasVnrArea ,FullBath, Exterior2nd_ImStucc,Fireplaces
1stFlrSF: First Floor square feet
MasVnrArea: Masonry veneer area in square feet
FullBath: Full bathrooms above grade
Exterior2nd: Exterior covering on house (if more than one material)
Imitation Stucco
Fireplaces: Number of fireplaces

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: The model is robust and generalisable some of the important points are ..

1. The R2 score should be optimum and there should be not much difference between train and test R2 score.
2. The model is not overfitted.

To build a Robust and generalized model some of the strategies are
1.Data collection and data cleaning
The train data should well representation of the population.The data should be cleaned and null values,duplicate values and outliers should be removed.
2.EDA and Feature Engineering
Irrelevant data and highly correlated data should be removed.
3.Regularization
We should use Regularization technique to prevent overfitting of data.
4.Cross validation
Use kfold cross validation to increase performance of the model.
5.Hyperparameter Tuning
Use gridsearch to tune hyperparameter to get the optimum value of lamda.