

# Advanced LR - Assignment

## 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 after passing the data points through GridSearchCV for ridge came out to be 1 and for Lasso optimal value came out to be 0.01.

### Ridge with Alpha = 1

```
ridge_model_cv_alpha Ridge(alpha=1)
R2 Score of the training data: 0.8770463231488708
R2 Score of the testing data: 0.8511381569087493
Residual Mean Error is: 0.0
```

### Ridge with Alpha = 2

```
ridge_model_cv_alpha 2
R2 Score of the training data: 0.8737273514905264
R2 Score of the testing data: 0.8491180383763313
Residual Mean Error is: 0.0
```

### Lasso with Alpha = 100 (Y variable is not transformed here)

```
R2 Score of the training data: 0.8737273514905264
R2 Score of the testing data: 0.8491180383763313
Residual Mean Error is: 0.0
```

### Lasso with Alpha = 200 (Y variable is Transformed here)

```
R2 Score of the training data: 0.865545965643375
R2 Score of the testing data: 0.8325114160314759
Residual Mean Error is: -0.0
```

### Lasso with Y variable Transformed (log transformation) Alpha = 0.01

```
R2 Score of the training data: 0.8614470111879391
R2 Score of the testing data: 0.8468385398855675
Residual Mean Error is: 0.0
```

## 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 Apply Lasso as the number of variables is also less. Lasso does its own feature selection which is not available in ridge. This will make the model simpler also Lasso penalizes the model more than Ridge when a greater number of variables are used.

## 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

**The top 5 predictor variables of lasso:**

```
R2 Score of the training data: 0.8701526017200429
R2 Score of the testing data: 0.8231812328590055
Residual Mean Error is: -0.0
```

TOP 5 Features

	Features	Coefficient
13	OverallQual_10	116199.2112
3	2ndFlrSF	94100.9222
19	OverallQual_9	87757.3208
2	1stFlrSF	78733.7584
9	Neighborhood_NoRidge	47386.7439

**Results after dropping the top 5 predictors:**

```
R2 Score of the training data: 0.8067393295826615
R2 Score of the testing data: 0.7626234010404764
Residual Mean Error is: 0.0
```

TOP 5 Features

	Features	Coefficient
0	LotArea	58804.0270
34	GarageCars_3	45059.5392
1	BsmtFinSF1	39678.0683
9	Neighborhood_StoneBr	38186.0986
8	Neighborhood_NridgHt	28539.4531

#### **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. We can keep the variance and bias the least.
2. To keep the variance low, we need to make sure that our model does not overfit and this can be done by using transformations or regularization techniques.
3. If the bias is high and variance is low, we should try introducing new training data.
4. Keeping the model simple and robust and avoiding multicollinearity in the data.