

# 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?

Ans:-

Optimal Value of alpha for Ridge and Lasso regression

Ridge Alpha: 2

Lasso Alpha: 0.01

If we double the value of alpha in Ridge and lasso regression the R2 score reduced

Ridge Alpha. 4

# ridge regression

```
lm = Ridge(alpha=4)
```

```
lm.fit(X_train, y_train)
```

# predict

```
y_train_pred = lm.predict(X_train)
```

```
print(metrics.r2_score(y_true=y_train, y_pred=y_train_pred))
```

```
y_test_pred = lm.predict(X_test)
```

```
print(metrics.r2_score(y_true=y_test, y_pred=y_test_pred))
```

0.9293153913681106

0.8673123125036433

Lasso. Alpha 0.02

# lasso regression

```
la = Lasso(alpha=0.02)
```

```
la.fit(X_train, y_train)
```

# predic=on on the test set(Using R2)

```
y_train_pred = la.predict(X_train)
```

```
print(metrics.r2_score(y_true=y_train, y_pred=y_train_pred))
```

```
y_test_pred = la.predict(X_test)
```

```
print(metrics.r2_score(y_true=y_test, y_pred=y_test_pred))
```

0.8274703555203093

0.8321493248301828

## 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?

**Ans:-**

We will choose Lasso as its giving feature selection option also. It has removed unwanted features from model without affecting the model accuracy. Which makes are model generalized and simple and accurate.

The  $r^2$ \_score of ridge is slightly higher than lasso for the test dataset so we will choose ridge regression to solve this problem

**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?

**Ans:-**

'OverallQual', 'GarageCars', 'GrLivArea', 'FullBath', 'TotalBsmtSF' are the top five important predictor variables . After removing we get 'Foundation\_PConc', '1stFlrSF', 'GarageArea', '2ndFlrSF', 'BsmtFinType1\_GLQ'.

**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 4:**

We have to get a generalised model so that we don't get overfitting and underfitting . The model should be accurate for datasets other than the ones which were used during training. Too much importance should not given to the outliers so that the accuracy predicted by the model is high. Root mean square error should be low as possible and the model should be build according to the business requirements.