# Advanced Regression – Part II Subjective questions

## 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 alpha for Ridge and Lasso regression is determined using Grid- Search Cross validation. From the CV conducted in model, below are the optimal values obtained

Optimal value of alpha for Lasso for the given model : 200
Optimal value of alpha for Ridge for the given model : 30

The value of alpha determines the amount of regularization done to the model. If the value of alpha is doubled , the model will tend to underfit. The training accuracies will be reduced due to the same.

After doubling the alpha value of 'Lasso' , from 200 to 400 the most important predictor variable is determined as '*GrLivArea*' since it has the highest co-efficient value.

After doubling the alpha value of 'Ridge' , from 30 to 60 the most important predictor variable is determined as '*GrLivArea*' since it has the highest co-efficient value.

## 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:

With the analysis I perform, the Test R2 Score for both the regression are below,

Test R2 Score (Ridge) : 0.710256<br>
Test R2 Score (Lasso) : 0.701650

The RMSE of both regression models are below,

Test RMSE (Ridge) : 48705.8<br>
Test RMSE (Lasso) : 49423.8

Since the test R2 score for Ridge is slightly higher than test R2 score for Lasso. So I choose Ridge regression for identifying the significant features of the dataset.

## 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 most important predictor variables can be determined using the coefficients of the model. If the value of the coefficient is higher, the predictor is considered to be more important.

Below is the list of most important 5 predictor variables for **Ridge**,

1. GrLivArea
2. 2ndFlrSF
3. BsmtQual
4. KitchenQual
5. BsmtExposure

Below is the list of most important 5 predictor variables for **Lasso**,

1. GrLivArea
2. ExterQual
3. Neighborhood
4. BsmtExposure
5. BsmtQual

## Question 4

How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?\

## Answer:

A model is considered to be robust and generalizable, if it is as less complex as possible. A model is said to become complex if the data points has large number of features, or the data represents polynomial of higher degree (ex:2,3).
To ensure the model is robust, it should have low bias and low variance.

A model is said to be overfitting if it has low bias and high variance. A model is said to be performing very well with training dataset but fails to perform with new/test dataset.

In order to make the model more generalizable , regularization technique is used. The regularization reduce the coefficients of the features towards zero, ensuring the complexity of the model is as low as possible.

Regularization may cost a small amount of training accuracy , however the model performs better with new data