

# 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?

## Answer1:

Optimal value of alpha for ridge: 1.0

Optimal value of alpha for Lasso: 0.001

After make the double alpha for ridge and lasso i.e. 2 and 0.002

**For Ridge:** Coeff values are increasing as alpha will increase.  $r^2$ \_score of train data is also drop from .807 to 0.45

**For Lasso:** As alpha value increased more feautrues removed from model. But  $r^2$ score is also dropped by 1% in both test and train data

**Top Features:** GarageFinish, KitchenQual, TotRmsAbvGrd, Fireplaces, HeatingQC

---

## 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 2:

We will choose **Lasso** as it gives us the option of feature selection.

It has removed unwanted features from model without affecting the model accuracy, Which makes our model robust and generalized.

---

## 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 3:

Top 5 features are GarageFinish, KitchenQual, TotRmsAbvGrd, Fireplaces, HeatingQC.

After dropping them model accuracy reduced from 80 and 81% to 45% and 39%.

Now top most features are:

Next top 5 features after dropping 5 main predictors 1stFlrSF, MSSubClass\_90, MSSubClass\_120, TotalBsmtSF, HouseStyle\_1Story

---

**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:**

Below points need to be kept in mind while building model to make model robust:

1. Model accuracy should be  $> 70-75\%$ : In our case its coming  $80\%$ (Train) and  $81\%$ (Test) which is correct.
2. P-value of all the features is  $< 0.05$
3. VIF of all the features are  $< 5$

In our model, all the three points are kept in mind while building it.

---