## 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 for Lasso : “alpha”: 0.001

Optimal value for Ridge : “alpha“: 0.9

If we choose the double value of alpha for both ridge and lasso then

**For Ridge:** In the model observed that the Coefficient values are increasing as alpha value increases , r2_score of train data also drops

**For Lasso**: As alpha value increases more features are removed from the model, but r2_score also drops in both train and test data.

**Top Features**: MiscVal, BsmtHalfBath , LowQualFinSF , BsmtFullBath, HalfBath

## 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?

It is important to regularize coefficients and improve the prediction accuracy also with the decrease in variance and making the model interpretable.

I have chosen Lasso as its helps in feature selection option, It also removes unwanted features from the model without effecting the model accuracy. Which helps to keep the model more generalized simple and accurate.

Since Lasso uses a tuning parameter lambda which applies a penalty term which is absolute value of magnitude of coefficients which is identified by cross validation. As lambda increase the coefficient shrinks towards zero. Lasso also does the variable selection.

## Question 3

After building the model, you realized 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?

The top 5 features in the model are : MiscVal, BsmtHalfBath , LowQualFinSF , BsmtFullBath, HalfBath.

After dropping them the accuracy drops, and the next best features based on their coefficients and next best predictors are

BsmtFinType1 (0.324486), Neighborhood_Gilbert (0.282619), LotShape (0.194391), HeatingQC (0.192283), Neighborhood_BrkSide (0.160200).

## 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:

The model should be as simple as possible, though its accuracy might decrease but it will be more robust and generalizable. It can be understood using the Bias- Variance tradeoff.

A model is robust when any variation in the data does not affect its performance much. A generalizable model is able to adapt properly to new, previously unseen data, drawn from the same distribution as the one used to create the model. The Model should not overfit.

The simpler the model more the bias and lower the variance, complex the model lower the bias and higher the variance. Hence the bias and variance should be at their optimal level so that each of these values are considerate and balanced.

Its implication in terms of accuracy is that a robust and generalizable model will perform equally well on both training and test data sets, the accuracy does not tend to change much for train and test data.

1. Model Accuracy should be >70- 75% .
2. P-Value of features is < 0.05
3. VIF of all features are < 5