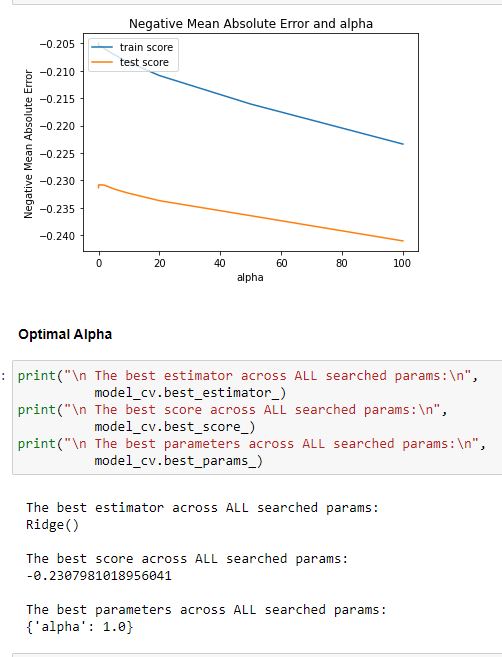
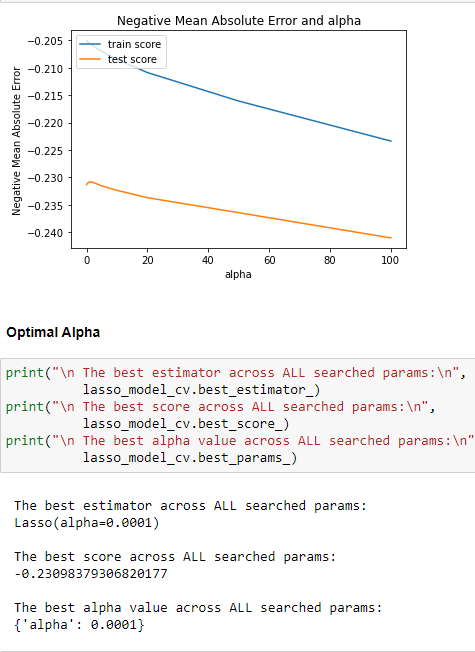
# **Advanced Regression Subjective Assignment**

**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?**





The optimal values for alpha are as follows:

Ridge Regression : 1.0

Lasso Regression : 0.0001

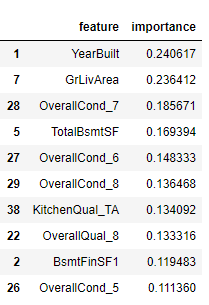
Alpha penalizes the model for using more no of features . In other words ridge and Lasso techniques regularize the coefficients by reducing their magnitude ie **cause shrinkage of the coefficients**. So doubling their values will-

* In case of Ridge - Coefficients of predictors decrease, then their value in the model decreases. That is, their effect decreases. And thus the flexibility of the model should decrease.
* In case of Lasso- More coefficients will become zero , thus removing them entirely from the model.

After Change implemented these are the most important features :

Left -Ridge

Right -Lasso

**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?**

In regularized regression -

Objective Function = RSS + α \* (sum of square/absolute value of coefficients)

= error term + α \* regularization term

Here, α (alpha) is the parameter which balances the amount of emphasis given to minimizing RSS vs minimizing the regularization term .

As the value of alpha increases, the model complexity reduces. Though higher values of alpha reduce overfitting, significantly high values can cause underfitting as well .

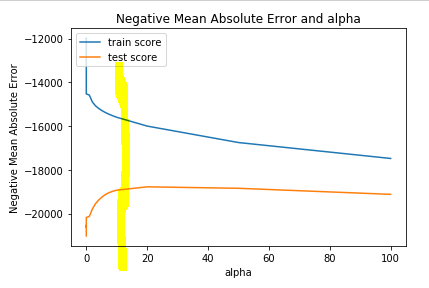
Thus alpha should be chosen wisely. In my assignment I am using cross-validation, i.e. the value of alpha is iterated over a range of values and the one giving higher cross-validation score is chosen.

The optimal values for alpha are as follows:

Ridge Regression : 1.0

Lasso Regression : 0.0001

Another way to choose alpha is by looking at the graph , when the difference between the Test and Training graph is minimum as well as error should not be too much that is the optimal point.



**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?**

So after excluding these 5 features-

* OverallCond\_7
* OverallCond\_6
* YearBuilt
* OverallCond\_5
* OverallQual\_7

The new Lasso model have these dominant features:



**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?**

Models are trained on a set of training data but their efficiency is determined by the ability to perform well on the unseen (test) data. This is the classic case of bias vs variance trade off.

Extremely complex models don’t generalize well since they are prone to change with small changes in the input data. Extremely simple models are likely to fail in predicting hence are prone to make errors and less accurate.

Hence, we should always select a model which is just complex enough to understand the variance in the data without much inaccuracy at the same time not too complex to overfit.

This can be achieved **using regularization**. Regularization is the process of deliberately simplifying models to achieve the correct balance between keeping the model simple and yet not too naïve.

Other ways can be - Weighted Least Square/Generised Least Square