**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:

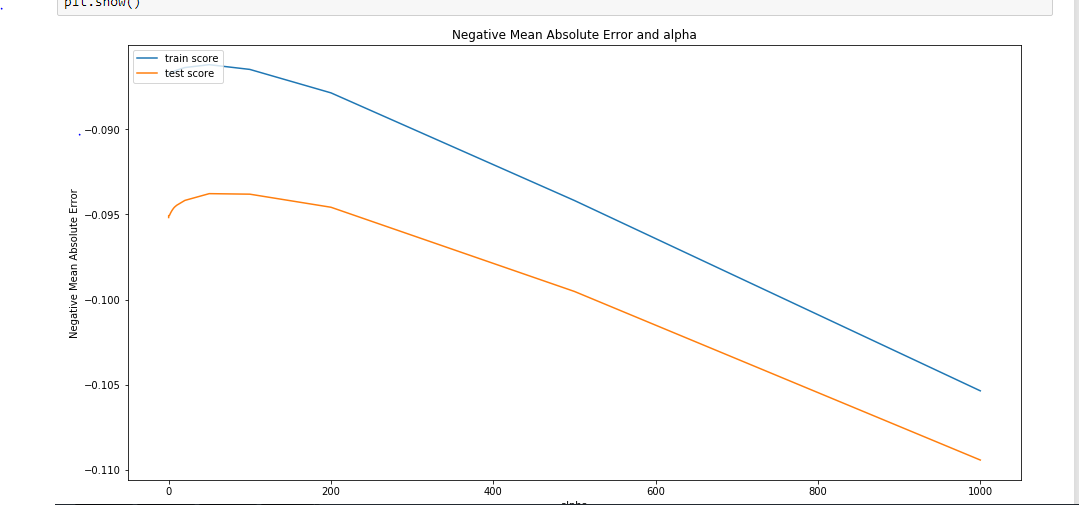
The optimal value of alpha for ridge and lasso regression we can get by using different types of cross validation techniques such as GridsearchCV etc. in our model. For our housing price prediction problem statement I have used gridsearchcv and for lasso the optimal alpha value is 0.001 where as for ridge I got 50 as the optimal alpha value.

If we increase the alpha value by 2 times then the flexibility of the model will decrease, leading to the decrease in variance and increase in bias.

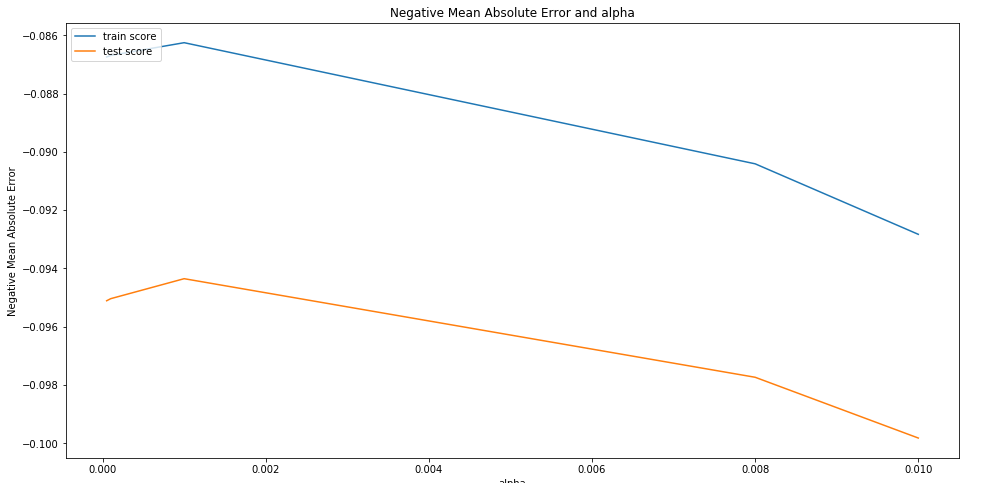
If the bias will increase then the model accuracy in both training and test data will decrease.

If we will consider our model then if we choose the double the value of alpha then it will not significantly impact our model as we can see from the graph but the error term for both training and test set will increase as the bias will increase.

**Ridge**

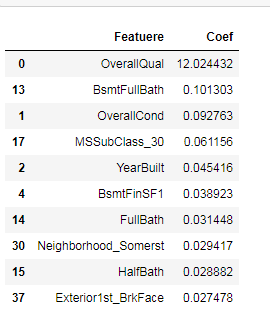


**Lasso**



After the change is implemented in our model then the most important predictor variable **OverallQual** will remain same as the most important predictor as the coefficient of that variable is very large compare to the other variables.

Generally if we increase the alpha value for our model then the penalties for higher coefficient variable will increase and alpha will shrinks the coefficient of the variable to zero.



**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:

The optimal value of alpha for ridge is 50 whereas for lasso it is 0.001.

Below is the output of lasso model with alpha value 0.001



Below is the o/p of Ridge model with alpha value 50:



If we compare two models then we will know that the accuracy of lasso model is slightly higher than the ridge model.

And also our business requirement is to understand how exactly the prices vary with the variables. So here the feature selection will play a crucial role as we have to select variables from 81 variables that are given in the dataset.

I would apply lambda=0.001 with Lasso regression since Lasso regression helps in feature elimination by making the coefficients of insignificant/duplicate variables 0 thus giving only those variables which are significant in determining the house price.

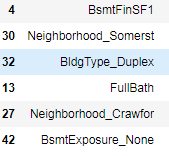
**Question 3**

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:

If predictor variables are missing and there is not any multicolinearity in our model then this will affect the accuracy of the model.

We have shown top 10 important variables as per their coefficients which are affecting the price of the house. Higher the coefficient value of the predictors higher it will affect the target variables. So if we remove the top 5 variables then the next 5 variables as per the coefficients are now will become most important predictor variables as shown below.



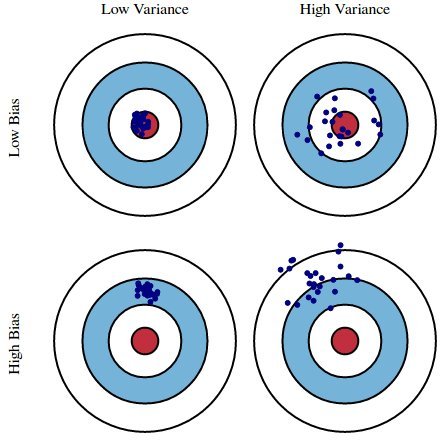
**Question 4**

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:

Models are trained on the training dataset but their efficiency is determined by the ability of the model to perform in test or unseen dataset.

Complex models are not generalized well since any small change in the data will affect the model. Extremely simple models are also likely to fail as they don’t have enough information to learn and predict hence are prone to make errors and less accurate. This is the classic case of **bias vs variance trade off.**

****

Hence we should always select a model which is just complex enough to understand the variance in the model and able to predict without much inaccuracy.

If the model is too complex or over fitting then we can use regularization. Regularization is the process of deliberately simplifying models to achieve the correct balance between keeping the model simple yet not to naïve.