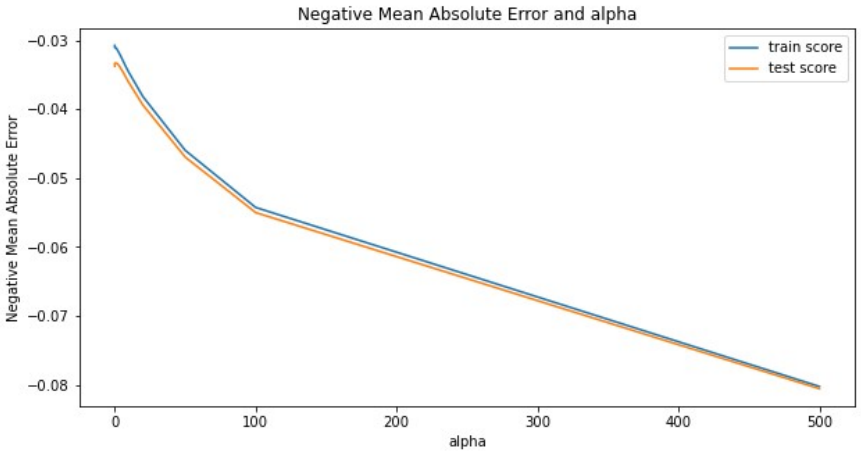
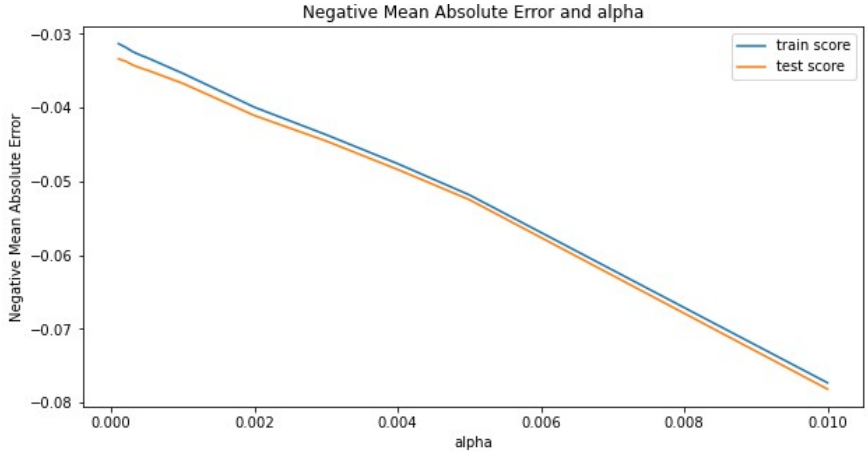
**Question 1: What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?**

**Answer:**

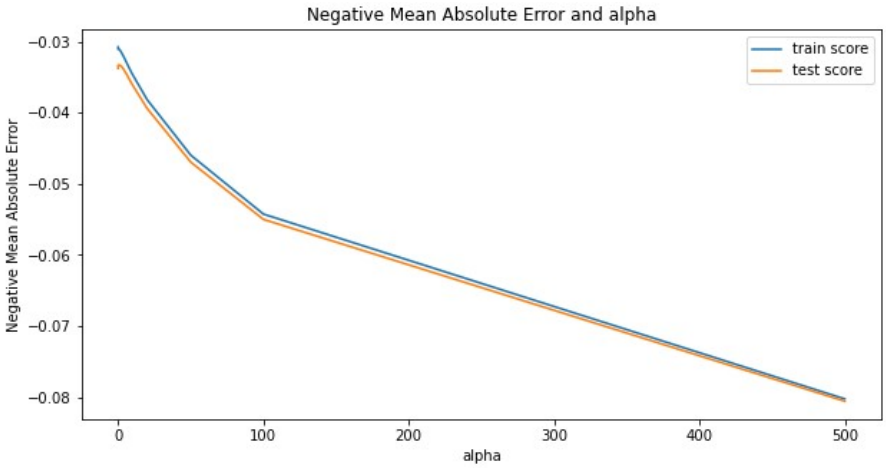
- Optimal Value of alpha:
  - Ridge : 1.0
  - Lasso : 0.0001
- Impact after choose double the value of alpha:
  - If we double the alpha it will penalize the curve more and model will be more complex.
  - In case of Lasso more coefficients will be 0.
  - It reduced little r2 value so that means error increased in training and test data.
  - Before:
    - Ridge :



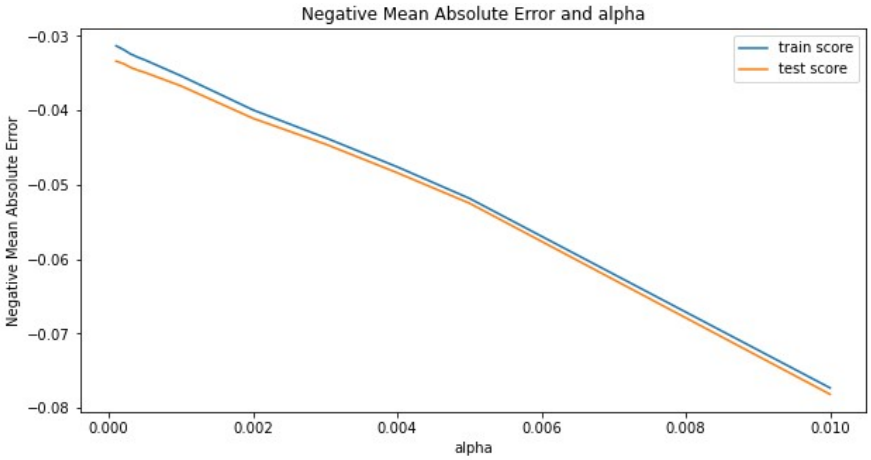

    - Lasso :




  - After:
    - Ridge :




- Lasso




-
  - Ridge : OverallQual, 1stFlrSF, OverallCond, TotalBsmtSF, 2ndFlrSF, GarageArea, BsmtQual, MSZonoing_FV(Floating Village Residential), Foundation Slab, HeatingQC.
  - Lasso : OverallQual, 1stFlrSF, TotalBsmtSF, OverallCond, 2ndFlrSF, GarageArea, BsmtQual, LotArea, Foundation Slab, HeatingQC.

- Important predictor variables after double the value of alpha:
  - We could see that for Ridge MSZonoing_RL(Residential Low Density) is not in the top 10 predictors and HeatingQC came into top 10 predictors.
  - For Ridge KitchenQual replaced by HeatingQC
  - Before:
    - Ridge : OverallQual, 1stFlrSF, OverallCond, TotalBsmtSF, 2ndFlrSF, GarageArea, BsmtQual, MSZonoing_FV(Floating Village Residential), MSZonoing_RL(Residential Low Density), Foundation Slab.
    - Lasso : OverallQual, 1stFlrSF, TotalBsmtSF, OverallCond, 2ndFlrSF, GarageArea, BsmtQual, LotArea, Foundation Slab, KitchenQual.
  - After
    - Ridge : OverallQual, 1stFlrSF, OverallCond, TotalBsmtSF, 2ndFlrSF, GarageArea, BsmtQual, MSZonoing_FV(Floating Village Residential), Foundation Slab, HeatingQC.
    - Lasso : OverallQual, 1stFlrSF, TotalBsmtSF, OverallCond, 2ndFlrSF, GarageArea, BsmtQual, LotArea, Foundation Slab, 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:**

- Hyperparameter Lambda:
  - Ridge : 1.0
  - Lasso : 0.0001
- R2 value on train data & test data for Ridge & Lasso
  - Train : Ridge(0.92), Lasso(0.91)
  - Test : Ridge(0.87), Lasso(0.87)
- MSE
  - Ridge : 0.002728
  - Lasso : 0.002730

We could see from the above parameters, R2 value on test and train data are same and comparable.

Also we could observe that MSE value in both the case also comparable.

But Lasso helps to reduce few coefficients to 0.

So Lasso will be better compare to Ridge

**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:**

Following factors are the good predictor and impact on the sales price of House in case of Ridge & Lasso:

- Ridge : OverallQual, 1stFlrSF, OverallCond, TotalBsmtSF, 2ndFlrSF, GarageArea, BsmtQual, MSZonoing_FV(Floating Village Residential), MSZonoing_RL(Residential Low Density), Foundation Slab.
- Lasso : OverallQual, 1stFlrSF, TotalBsmtSF, OverallCond, 2ndFlrSF, GarageArea, BsmtQual, LotArea, Foundation Slab, KitchenQual.

If the most 5 important variables are not present in the model then next 5 important variables will be considered as most important variables.

In that case following will be the scenario for each case:

- Ridge : GarageArea, BsmtQual, MSZonoing_FV(Floating Village Residential), MSZonoing_RL(Residential Low Density), Foundation Slab.
- Lasso : GarageArea, BsmtQual, LotArea, Foundation Slab, KitchenQual.

**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:**

- Need to follow Occam Rajor's principle and according to that Model should be as simple as it will be required.
- Model should not be impacted by outliers. To much focus on outliers while building model can impact while using test data and model can fail on that.
- It need to be ensured that accuracy of training data of the model and test data of the model should be comparable. Otherwise model may not perform well in unseen data and high training r2 and low test r2 is sign of over fitting and that is not generalised.
- Regularization method should be used to keep the model optimum simpler. It penalizes the model if it becomes more complex.
- Regularization method helps to achieve the Bias-Variance trade off. It compromise by increasing bias to a optimum position where Total Error is minimum. This point also known as Optimum Model Complexity where Model is sufficient simpler to be generalisable and also complex enough to be robust.
-