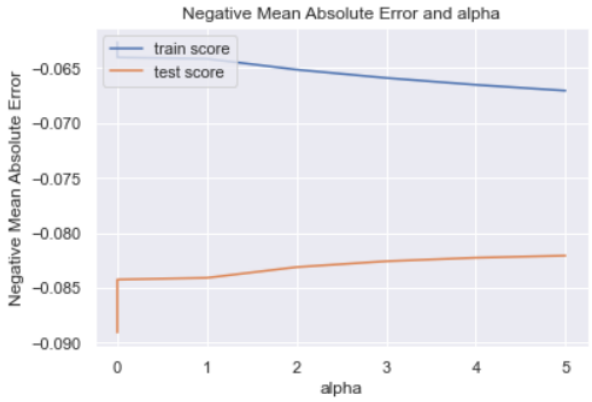
Assignment Part-II – Subjective questions

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 to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer: In the case of ridge regression: - As per below graph:



When we plot the curve between negative mean absolute error and alpha, we see that as the value of alpha increase from 0, the error term decrease and the train error is showing increasing trend when value of alpha increases. when the value of alpha is 2 the test error is minimum, so we decided to go with value of alpha equal to 2 for our ridge regression.

For lasso regression it is decided to keep small value that is 0.01, when we increase the value of alpha the model try to penalize more and try to make most of the coefficient value zero. Initially it came as 0.4 in negative mean absolute error and alpha.

The most important variable after the changes has been implemented for ridge regression are as follows: -

1. MSZoning_FV
2. MSZoning_RL
3. Neighborhood_Crawfor
4. MSZoning_RH
5. MSZoning_RM
6. SaleCondition_Partial
7. Neighborhood_StoneBr
8. GrLivArea
9. SaleCondition_Normal
10. Exterior1st_BrkFace

The most important variable after the changes has been implemented for lasso regression are as follows: -

1. GrLivArea
2. OverallQual
3. OverallCond
4. TotalBsmtSF

5. BsmtFinSF1
6. GarageArea
7. Fireplaces
8. LotArea
9. LotArea
10. LotFrontage

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 regularization of the model coefficients are needed to increase accuracy.

Ridge regression uses a tuning parameter called lambda as the penalty is **square of magnitude** of coefficients which is identified by cross validation. Residual sum or squares should be small by using the penalty.

Lasso regression uses a tuning parameter called lambda as the penalty is **absolute value of magnitude** of coefficients which is identified by cross validation. As the lambda value increases Lasso shrinks the coefficient towards zero and it make the variables exactly equal to 0. Lasso also does variable selection.

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 will decrease but it will be more robust and generalisable. It can be also understood using the Bias-Variance trade-off. The simpler the model the more the bias but less variance and more generalizable.

Bias: Bias is error in model, when the model is weak to learn from the data. High bias means model is unable to learn details in the data. Model performs poor on training and testing data.

Variance: Variance is error in model when model tries to over learn from the data. High variance means model performs exceptionally well on training data as it has very well trained on this of data but performs very poor on testing data.