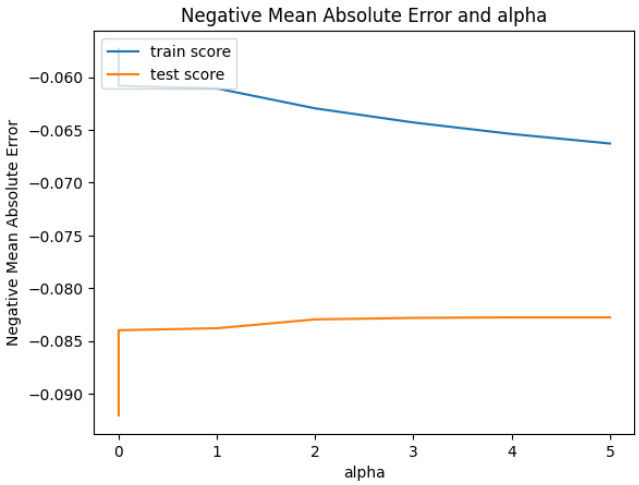
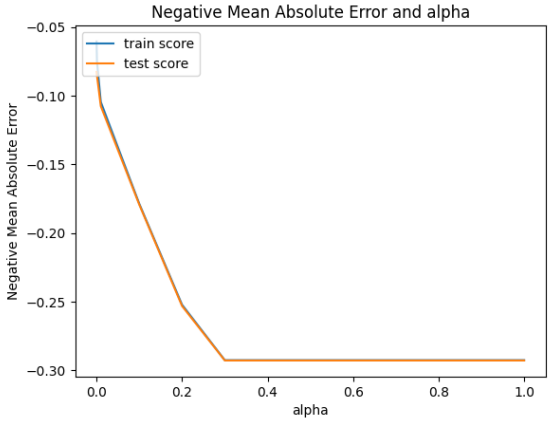
Q1 .What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Ans: When we plot the curve between negative mean absolute error and alpha in Ridge Regression, we see that as the value of alpha increase from 0 the error term decrease and the train score is increasing 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 I have decided to keep very 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 0.3 in negative mean absolute error and alpha.



When value of alpha is doubled, more penalty is applied. Hence increase in error on both train and test set. But as penalty increases, model will become more simpler and general.

Similarly when we increase the value of alpha for lasso regression, more coefficients of the variables will reduced to zero, when we increase the value of our r2 square also decreases.

Q2. 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?

Ans. There are two method of regularisation – Ridge and lasso. Ridge regression introduces bias into the parameter estimates in order to reduce variance. This is a great solution to the problem of colinearity whereas Lasso eliminates some variables from the model altogether. This is a great solution to overfitting.

During the assignment, we had performed both ridge and lasso regression. Both have their own advantages and limitations. Various values of lambda has been used for both regression. Afterwards, one out of them has been chosen called optimal value of lambda. Model has been fitted on optimal value of lambda and compared both models.

I will choose Ridge regression over lasso regression due to following reasons:-

(a) Ridge is a bit easier to implement and faster to compute as compare.
(b) It does not tends coefficients to zero, which may lead to chance of elimination of nay important feature.
(c) Ridge gives biased parameter estimates in order to lower the variance of those estimates. This is ideal for colinearity problems.

Q3. 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?

Ans. The five most important predictor variables now are as follow:

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

Q4. 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?

Ans. Model can be more robust and generalizable if it has fulfil following conditions:-

(a) The model should be as simple as possible, though its accuracy will decrease but it will be more robust and generalizable.
(b) 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. Its implication in terms of accuracy is that a robust and generalizable model will perform equally well on both training and test data i.e. the accuracy does not change much for training and test data.

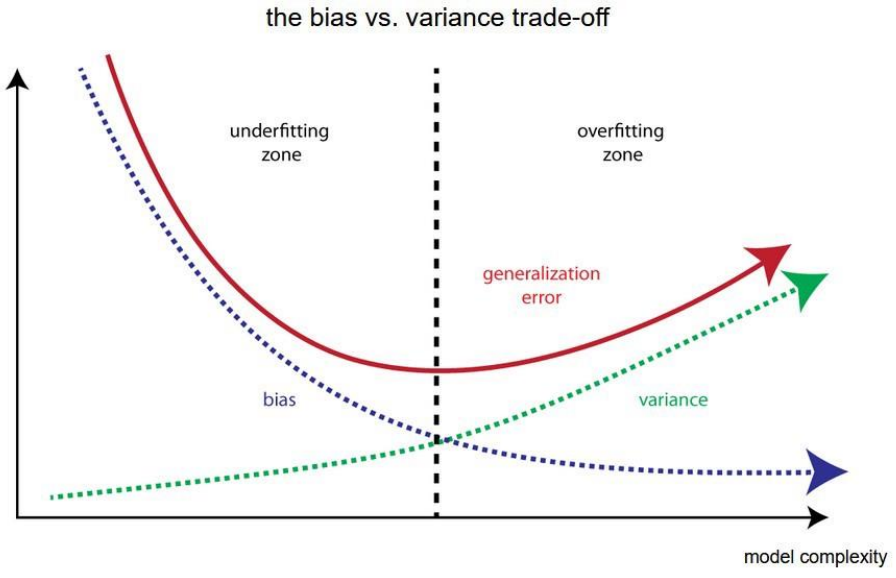


Image Source [https://towardsdatascience.com/bias-and-variance-but-what-are-they-really-ac539817e171]

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 as it was unseen data for the model. It is important to have balance in Bias and Variance to avoid overfitting and under-fitting of data.