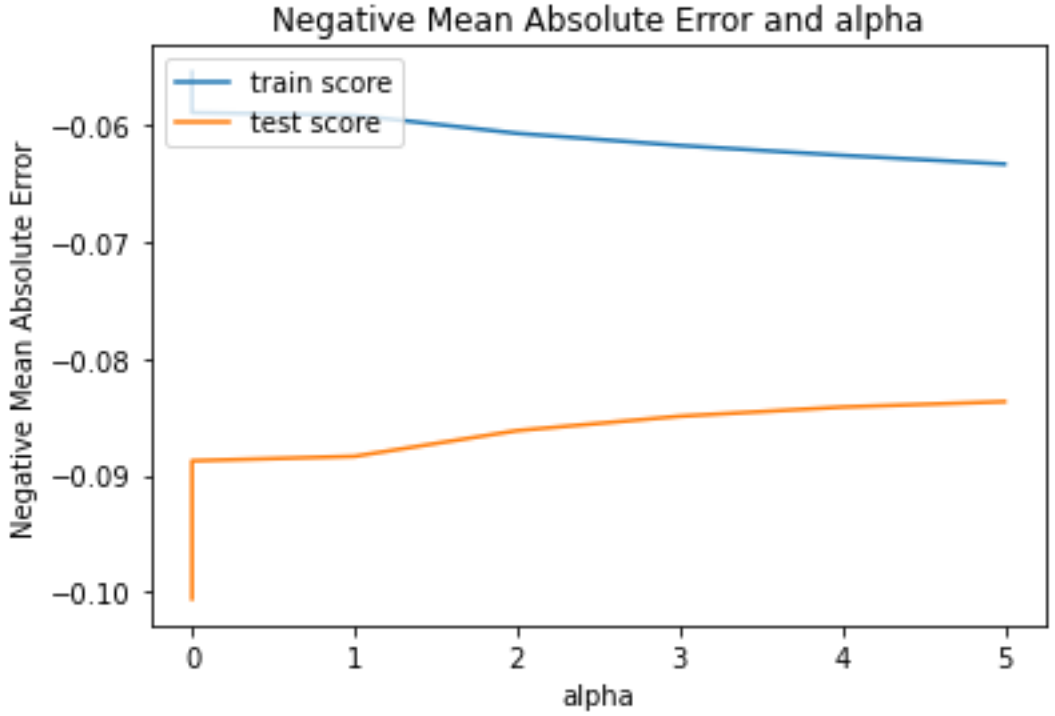
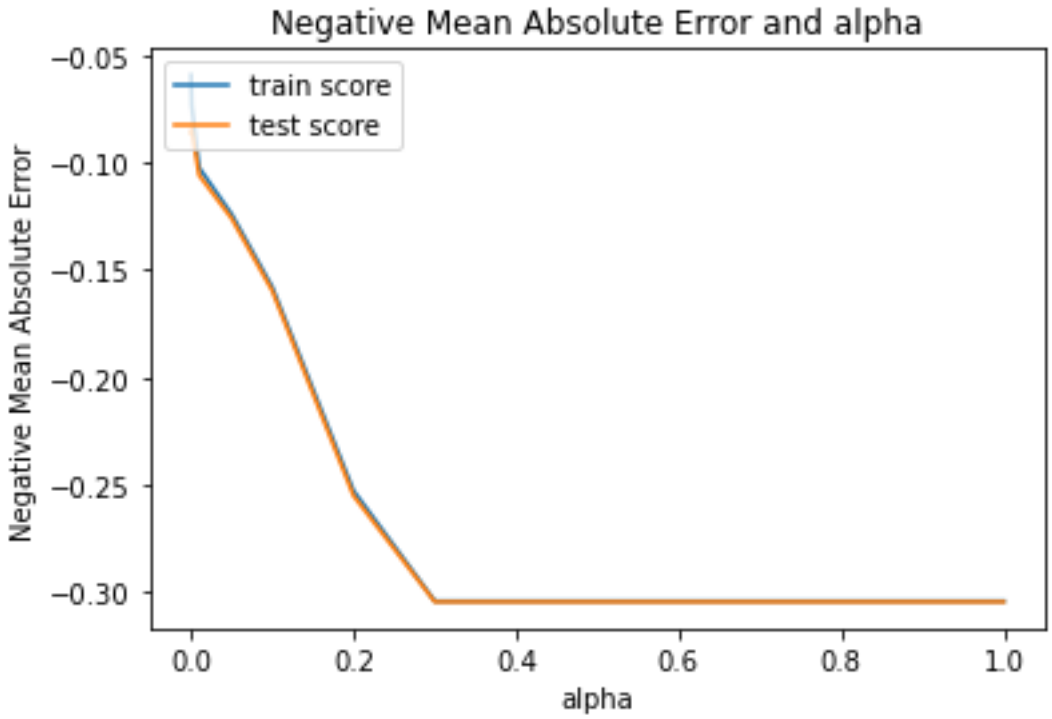
Problem Statement - Part II

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: In the case of ridge regression: - 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 5 the test error is minimum so we decided to go with value of alpha equal to 5 for our ridge regression.



For lasso regression I decided to choose 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 0.4 in negative mean absolute error and alpha.



When we double the value of alpha for our ridge regression, we will take the value of alpha equal to 10 the model will apply more penalty on the curve and try to make the model more generalized that is making model simpler and no thinking to fit every data of the data set.
Similarly, when we increase the value of alpha for lasso, we try to penalize more our model and more coefficient of the variable will reduced to zero, when we increase the value of alpha, our r2 square also decreases.
The most important variable after the changes has been implemented for ridge regression are as follows: -

1. Neighborhood_Crawfor
2. MSZoning_FV
3. MSZoning_RL
4. Neighborhood_StoneBr
5. OverallCond_9

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

1. GrLivArea
2. TotalBsmtSF
3. GarageArea
4. BsmtFinSF1
5. CentralAir

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:

Ridge regression is a method of estimating the coefficients of multiple-regression model in scenarios where linearly independent variables are highly correlated. It 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. The penalty is lambda times sum of squares of the model coefficients, hence the coefficients that have greater values gets penalized. As we increase the value of lambda the variance in model is dropped and bias remains constant. Ridge regression includes all variables in final model unlike Lasso Regression.

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e., models with fewer parameters). It 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. When lambda value is small it performs simple linear regression and as lambda value increases, shrinkage takes place and variables with 0 value are neglected by the model.

We will choose Lasso Regression over Ridge regression as it does variable selection which further results in models that are easier to interpret.

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?

Next 5 most important predictor variables that will be included are: -

1. FireplaceQu_Gd
2. LotArea
3. ScreenPorch
4. OpenPorchSF
5. MasVnrArea

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: 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. Its implication in terms of accuracy is that a robust and generalisable model will perform equally well on both training and test data i.e. the accuracy does not change much for training and test data.

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.