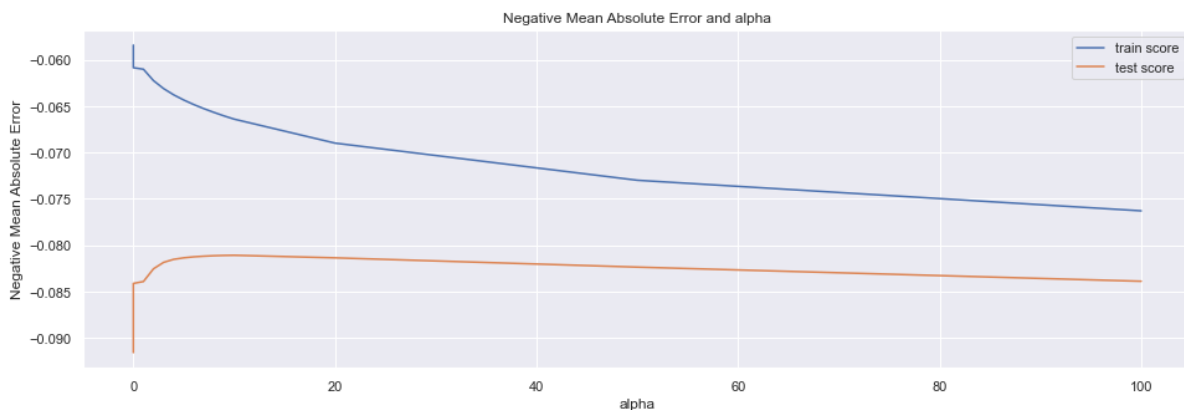


Question 1

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?

Ans:

In Ridge regression, When we plot the curve between Negative Mean Absolute Error and alpha., we observed that value of lambda/alpha increases from 0, error term decreases as shown in below graph.



In final model optimal value of alpha for ridge is "10 "and for Lasso is "0.001"

If we double the value of alpha for our Ridge regression, then model will apply more penalty on the curve and try to make model more generalized and simpler. As alpha increased to 20 we get more error for both train and test.

And incase of Lasso when we try to increase the value of alpha, model will be penalized, and more coefficient of the model will be reduced to zero

After the implementation, below are the predictor variables.

- MSSubClass
- Neighborhood_Crawfor
- SaleCondition_Normal
- OverallCond
- BsmtFullBath
- SaleType_New
- Neighborhood_StoneBr
- MSZoning_FV
- MSZoning_RL
- Exterior1st_BrkFace

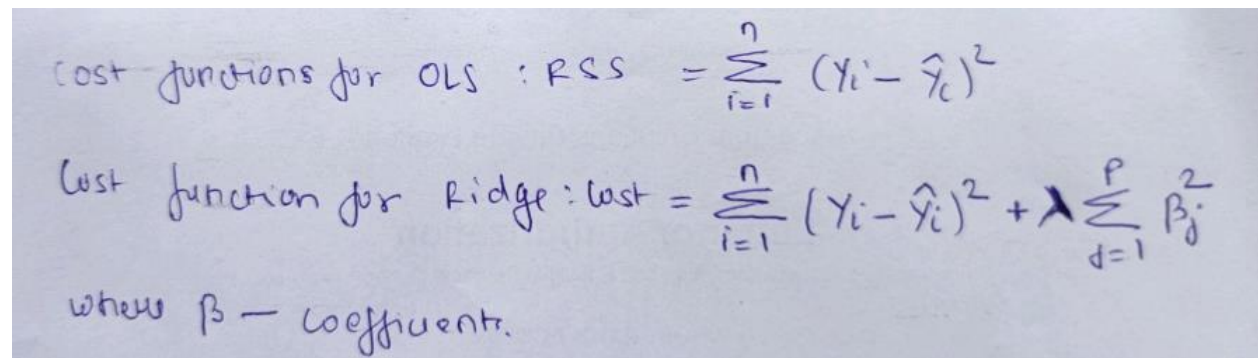
Question 2

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?

Ans:

Regularization helps with managing model complexity by shrinking the model coefficient estimates towards 0 and this discourages the model from becoming too complex, thus avoiding the risk of overfitting.

Ridge regression uses a tuning parameter called lambda and it is square of magnitude of coefficients as shown in below expression.



Handwritten text showing the cost functions for OLS and Ridge regression:

$$\text{Cost function for OLS : } RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$
$$\text{Cost function for Ridge : } \text{Cost} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

where β - coefficients.

The penalty is lambda times sum of square of coefficients, hence the coefficient with higher lambda will be penalized. As we increase the value of lambda the variance in the model is dropped and the bias remains constant. If there are lot of features, then Ridge will not do feature selection.

The primary difference between Lasso and Ridge regression is their penalty term. The penalty term in case of Lasso is the sum of the absolute values of all the coefficients present in the model as shown in below expression.

$$\text{Lasso Regression Cost} = \sum_{i=1}^n (y - \hat{y}_i)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

As with Ridge regression, Lasso regression shrinks the coefficient estimates towards 0 and in Lasso, the penalty pushes some of the coefficient estimates to be exactly 0, provided the tuning parameter, λ , is large enough and performs feature selection.

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?

Ans.

The five most important predictor variables that will be excluded are:

'GrLivArea', 'GarageType_Attchd', 'MSZoning_RM', 'SaleType_New', 'TotalBsmtSF'.

If we remove these and rebuild the model, the five most important predictor variables now are – MSSubClass, Neighborhood_Crawfor, SaleCondition_Normal, OverallCond, BsmtFullBath

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?

Ans:

We are ensuring model is robust and generalizable by making sure that model is not over fitting and is as simple as possible. The simpler the model the more the bias but less variance and more generalizable. The accuracy of the model will go up if we try to over fit but will no longer makes it generalizable and when the model is generalized it will perform equally well on both training and test data sets.