ADVANCED REGRESSION ASSIGNMENT – SUBJECTIVE QUESTION

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:

The Optimal Value of α (alpha) for ridge regression is 30 and lasso regression is 60
In Case of ridge regression: Doubling the alpha value reduces the score on train and test set but no changes on the variable
In Case of lasso regression: Doubling the alpha value
1. Reduces the score on train and test set
2. Total predictor variable reduces from 136 to 107
3. Co-efficient values increases.S

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:

1. For Ridge, regression, I would choose the optimal value $\alpha=30$
   Because for $\alpha=30$, The negative absolute mean error reduces drastically.
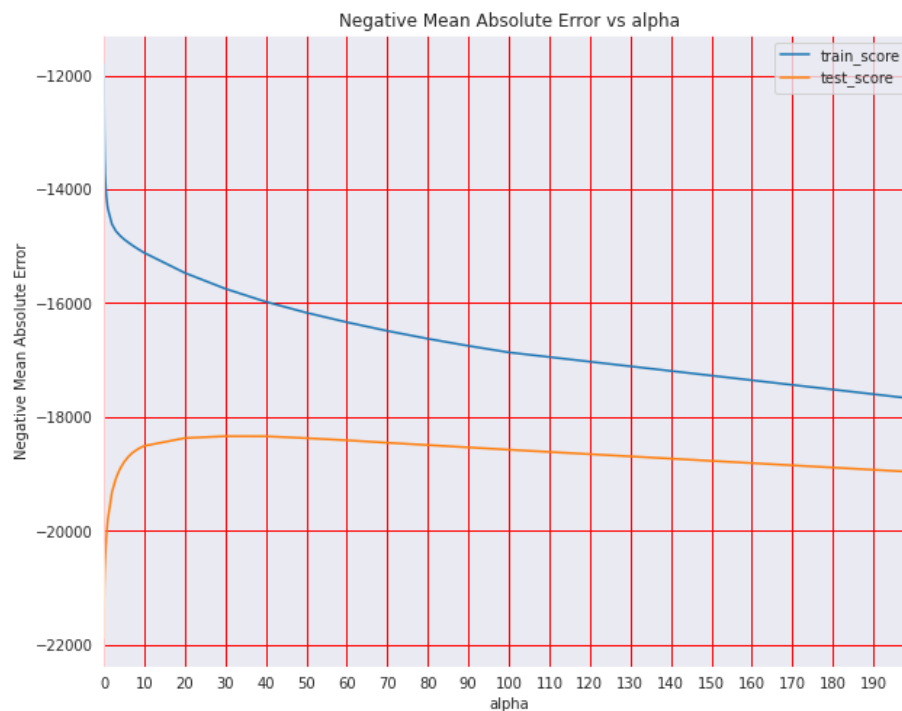
2. For Lasso regression, I would choose the optimal value α=60
   Because for α=60, The negative absolute mean error reduces drastically
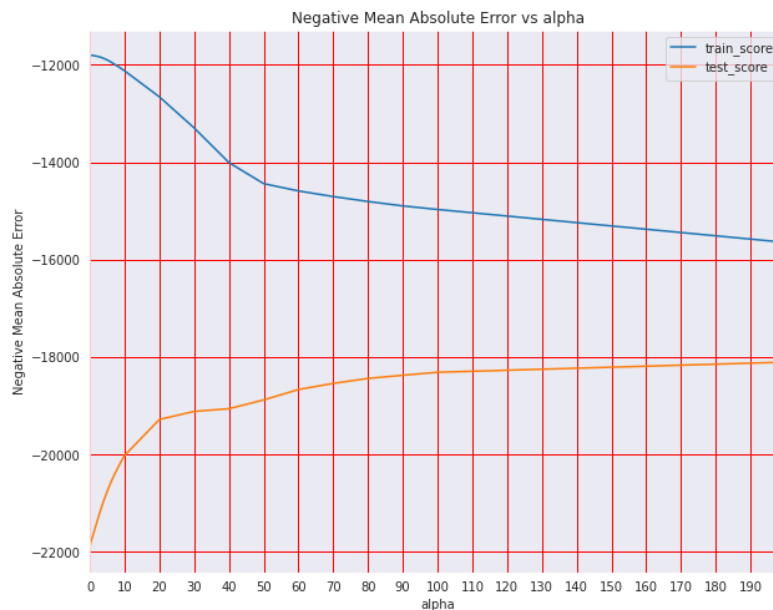
Question 3

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?

The five most important predictors are:

1. 2ndFlrSF - Second Floor square feet
2. Electrical Fuse A- Electric fuse system
3. Foundation PConc- Foundation Poured Concrete
4. Electrical FuseP 60 AMP Fuse Box and mostly knob & tube wiring (poor)
5. Neighbourhood Crawfor - Crawford

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:

A model is robust and can be generalized when it works well on unseen or new data. Hence, the model should be regularized and kept simpler. The accuracy tends to be moderate because of high bias but works well for unseen data as it has low variance. This makes it robust and generalizable.