ADVANCE LINEAR REGRESSION QUESTIONS

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 :

1. Optimal alpha value of ridge is 20.0 which is coming as part of best parameter identified and for lasso it is 0.001.
2. When we double the values most of coefficients try to move towards the Zero
3. Important predictor variables are below
   A) BsmtFinSF2
   B) LotShape
   C) ExterCond
   D) GarageCars
   E) Neighborhood_Gilbert

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 : Based on accuracy score derived for Train vs test will go with lasso , because based on assumptions comparison between train and test set error terms of residuals are normally distributed between test and train data and no skewness in data. Additionally actuals and predicted between train and test is more or less on same lines.

Ridge Accuracy score of Train Data set : 0.909

Ridge Accuracy score of Test Data set : 0.899

Lasso Accuracy score of Test Data set : 0.909

Lasso Accuracy score of Test Data set : 0.9

best lasso alpha Value : {'alpha': 0.001}

best Ridge alpha value : {'alpha': 20.0}

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?

Answer : Below are five top predictors of the model

BsmtFinSF2 0.30

LotShape 0.15

ExterCond 0.12

GarageCars 0.11

Neighborhood_Gilbert 0.09

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 : Based on accuracy derived between two model complexities went into model where accuracy between train and test data is close with no overfitting the requirements . if we see the data is optimal where heteroscedastic and homoscedastic was not found in distribution of error terms We can see the accuracy of test data is not significantly high or less as compared with train data accuracy .