PROBLEM STATEMENT - II

Question 1

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

Answer 1

The optimal values for ridge and lasso regression are

- Ridge: .1
- Lasso: .05

Change in the model when the values of alpha for both ridge and lasso were doubled:

When the values were doubled, there was marginal increase in the R 2 score of both test and train data for ridge and lasso regression. However, they were not found to be significant.

The most important predictor variables remained unchanged after the doubling the alpha.

The most important predictor variables are:

Lot Area
GrLivArea
Garage Area
Age of House
Age of remodelling
Overall condition of house
Number of baths
Porch Area

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?

The optimal values for ridge and lasso regression are

- Ridge: .1
- Lasso: .05

I choose to apply ridge regression as it had better R 2 score of 66.51 over 66.47 of lasso.

Question 3

After building the model, you realized that the five most important predictor variables in the lasso model are not available int 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

I took a different approach to the issue. First the missing variables were identified. Thereafter the columns were created with value as zero in the incoming data and the existing model was used to predict the target variable.

The data consistency of test data is always not predictable. Hence, it does not serve to create new models every time there is such a difference

Question 4

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?

Answer

For the model to be generalizable and robust, it should have the following considerations.

1. The outliers in the incoming data should not affect the model drastically. Scaling the outlier values and removing variables with large outliers can enable this.
2. The model should not be overfitting. The test data prediction score should not me higher than the one achieved in train data.