

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:

Below are the optimal value of alpha results achieved for both Ridge and Lasso.

The best alpha value for Lasso : {'alpha': 0.001}

The best alpha value for Ridge : {'alpha': 0.1}

When we double the value of alpha for ridge and lasso regression, we tend to apply more penalty on the curve and the model may become generalized. Also, the coefficients of the variables might reduce to zero.

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: I would personally go with Lasso regression will be an ideal option, the reason being that the model will be more robust and can also help in feature elimination.

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?

Answer: As per my understanding, I would go with the below five most important variables which can be excluded.

- TotalBSmtSF
- OverallQual
- GrLivArea
- GarageArea
- OverallCond

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: The model must be robust at the same time generalisable, hence it is our responsibility to check for any outliers and make sure they are not impacting the training data set. The Outliers detection activity needs to be carried out in a very diligent way to make sure we are not giving too much weightage to it, which may cause the accuracy to spike up for a given model. Also, it is generally a good practice to remove any outliers from the dataset if they don't make any sense and

if we are very much sure that they are not useful for our analysis. The test accuracy also should be taken into consideration such that it is not less than the training scores.