Advanced Linear Regression

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:

Optimal Value of alpha for ridge and lasso regression are:

- Optimal Value of lambda for Ridge: 7
- Optimal Value of lambda for Lasso: 0.001

If we choose to double the value of alpha for both ridge and lasso:

In case of ridge that will lower the coefficients and makes the coefficients near the zero but in case of Lasso there would be features selection by making less important features coefficients to 0 and remaining as if ridge near to 0

. The most important predictor variable after the change is implemented are those which are significant features that having high coefficients values

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:

Optimal Value of alpha for ridge and lasso regression are:

- Optimal Value of lambda for ridge: 7
- Optimal Value of lambda for Lasso: 0.001

As if we already got good r2_score for both the models Its better to choose lasso because it selects the features which are significant and eliminate features by make coefficients of the feature to zero were as Ridge will make the coefficients value near zero but not exactly zero so all the features will be there and becomes complex compared to lasso

As we know simpler models are better than complex model so, lasso is better than ridge

Ridge: Train: 91, Test: 88

Lasso: Train: 90, Test: 87

Question 3

After building the model, you realized 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:

On running the same notebook and removing the top 5 significant variables:

We found below variables as next 5 significant.

```
MSZoning_FV            0.096870
Exterior1st_BrkFace    0.072128
BsmtExposure_Gd        0.064040
MSZoning_RL            0.056037
GarageCars             0.047854
```

These are just next top 6-10 variables (excluding Mszoning features) before removing top 5 after removing they become top 5 (Mszoning added and became bit significant)

Last cells I ran these in Jupiter notebook

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 considered to be robust if the model is stable, i.e. does not change drastically upon changing the training set. The model is considered generalizable if it does not overfits the training data, and works well with new data. Its implication in terms of accuracy is that a robust and generalizable model will perform equally well on both training and test data i.e. the accuracy does not change much for training and test data.