Problem Statement - Part II

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 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:

Optimal value of alpha for Ridge: 4

Optimal value of alpha for Lasso: 100

After doubling the value if alpha for ridge and lasso i.e. 8 and 200, the results are:

Ridge Alpha 4

#R2score(train) = 0.892762756977153

#R2score(test) = 0.8285049824629558

Ridge Alpha 8

#R2score(train) = 0.8773897973861292
#R2score(test) = 0.8241224673254787

R2score on training data has decreased but it is almost same on test data.

Lasso Alpha 100

#R2score(train) = 0.9007107296092232

#R2score(test) = 0.8363565465614051

Lasso Alpha 200

#R2score(train) = 0.88811921460748

#R2score(test) = 0.8331858718448046

R2score of training data has decreased and is almost same on testing data.

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:

The r2_score of Lasso is slightly higher than ridge for the test dataset. Also, it gives feature selection option. It has removed unwanted features from model without affecting the model accuracy.

Hence, we will choose **Lasso Regression.**

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:

Top 5 features are:

1. GrLivArea
2. OverallCond
3. MasVnrArea
4. 1stFlrSF
5. 2ndFlrSF

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:

To make model robust and generalisable 3 features are required:

1. **Model accuracy should be > 70-75%:** We have got 90%(Train) and 83%(Test), through Lasso Regression.

2. **P-value of all the features is < 0.05**

3. **VIF of all the features are < 5**

4. The test accuracy is not lesser than the training score

5. The model should be accurate for datasets other than the ones which were used during training.

Our model satisfies all these conditions. Thus, the model is robust and generalisable.