Assignment 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:

The optimal value of alpha for ridge is 2 and for Lasso it is 0.0001 on the given dataset.

Metric	Ridge (alpha = 2)	Ridge (alpha=4)	Lasso(alpha=0.0001)	Lasso(alpha=0.0002)
R2	0.817	0.813	0.820	0.820
RSS	266.508	272.302	261.508	262.109
MSE	0.427	0.431	0.423	0.423

- Ridge: Error terms are slightly higher when the value of alpha is doubled.
- Lasso: Not much effect on R2 and error terms when the value of alpha is doubled.

For both Ridge and Lasso **GrLivArea , 1stFlrSF , OverallQual_9, YearBuilt, OverallQual_10** are the top 5 important predictor variables after the change is implemented.

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 will choose Lasso with double alpha value that is alpha=0.0002 because Lasso will make few variables completely zero and the final model will be simple and more robust. Eg. In my Lasso model when alpha is 0.0002, the columns related to KitechAbvGr_2, KitechAbvGr_3, BedroomAbvGr_2, BedroomAbvGr_3 are zero. So, I can completely remove them from my predictor variables list.

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:

After excluding the five important predictor variables: **GrLivArea , 1stFlrSF , OverallQual_9, YearBuilt, OverallQual_10,** the new model will the following important predictor variables: **TotalBsmtSF, BsmtFinSF1, GarageArea, TotRmsAbvGrd, LotArea.**

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 a model more robust and generalisable, we should try to make it simpler. A simpler model will penalise the model for involving more predictor variables, and thus can remove the problem of overfitting. Such model will may not have a higher accuracy but can work better on unseen data. This can be well understood by the concept of bias-variance trade-off. The simpler the model the more the bias and less variance. Bias is the error made in learning the model and variance is the error when the model perform good on training data but poor on test data.