Advanced linear regression assignment

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 regression is 50, while for Lasso it is 100.

When we double the value of alpha:

For Ridge: R2 score decreases, RSS increases while MSE increases.

For Lasso: R2 score increases, RSS decreases while MSE decreases.

Below are the most important predictor variables after the change is implemented:

Ridge:

OverallQual

KitchenQual_TA

Neighborhood_NoRidge

LotArea

KitchenQual_Gd

Lasso:

OverallQual

KitchenQual_TA

KitchenQual_Gd

Neighborhood_NoRidge

LotArea

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:

Though there is not much difference between the metrics of obtained model after Ridge and Lasso regression, I will choose to apply Lasso because it has further eliminated the dependent features from the model, by bringing the lesser dependent variables coefficient to zero.

Note: We have already taken care of multicollinearity before, by using VIF.

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:

We need to choose the next 5 predictor variables, after the initially obtained top 5 features. These variables will be:

Ridge:

LotArea

ExterQual_TA

BsmtQual_TA

KitchenQual_Fa

BsmtFinSF1

Lasso:

KitchenQual_Fa

LotArea

BsmtQual_TA

ExterQual_TA

BsmtFinSF1

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:

We can follow below steps to ensure model is robust and generalisable:

- We can handle missing data,
- Remove non relevant features from model building
- We can remove outliers.
- We can take important features using feature selection from RFE.
- W can standardize the data, do log transform on dependent variable
- We can do feature Selection using Lasso – dropping 0 coefficient features from Lasso regression.
- We can do hyperparameter tuning- this reduces overfitting without making the model too simple that it underfits.

Above actions will increase the accuracy of the model, as we are reducing improper data taken into consideration for model and also taking steps to reduce overfitting whilst not compromising on model complexity.