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?

Ans: - In the case of ridge regression: - We plot the curve between negative mean square absolute error and alpha we see that as the value of alpha increase from 0 the error term decreased & train error increasing trend as alpha value increases. As when the value of alpha reaches 5 test error is minimize, hence decided to take alpha=5 in ridge regression.

In lasso regression, decided to take verry small value that is 0.0001, when we increase the alpha value, when we increase the value of alpha the model try to penalize more.

When we make the alpha value double for Ridge regression now alpha =10 the model will apply more penalty and try to make the model more generalized i.e making more simple model, in the graph we can see that error is more at alpha=10 for both train & test.

Similarly, when we increase value of alpha for lasso try to more penalized & more coefficients of the variables will reduce to zero & r2score also decreases.

The most impotent variables after changing have been implemented for Ridge regression as below: -

1stFlrSF
GrLivArea
OverallQual
TotalBsmtSF
BsmtFinSF1
TotRmsAbvGrd
Neighborhood_StoneBr
GarageArea
2ndFlrSF
Neighborhood_NoRidge

The most impotent variables after changing have been implemented for Lasso regression as below: -

GrLivArea
OverallQual
TotalBsmtSF
BsmtFinSF1
Neighborhood_StoneBr
1stFlrSF
GarageArea
OverallCond
Neighborhood_NridgHt
Neighborhood_NoRidge

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?

Ans:-

The Ridge regression is better in terms of R2 score of Train & Test, but lasso brings & makes zero value to insignificant features, enabling us to choose the predictive variable. Hence choosing Lasso.

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?

Ans:- The 5 most important predictor variables to be remove are below: -

GrLivArea
TotalBsmtSF
OverallQual
1stFlrSF
Neighborhood_StoneBr

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?

Ans:- The model should be as simple as possible, though its accuracy will decrease but it will be more robust & generalisable. it also understands that using the Bias-Variance trade off.

As Simple the model the more the bias but less the variance and more generalisable. its implicates in terms of accuracy is that a robust & generalisable model will perform equality well on both train & test Data

Bias: - Bias is error in model when the model is weaker to learn from the data. High Bias means model is unable to learn detain in the data. Model performs poor on training and testing data.

Variance: - Variance is error in model when the model tries to over learn from the data. High variance means model performs exceptionally well on training data as it has very well trained on this of data but performs very poor on testing data as it was unseen data for the model.

It is important to have balance in bias and variance to avoid overfitting and under-fitting od data.