Assignment

Problem Statement – Part II

Subjective Questions

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: 10
Optimal value of alpha for ridge: 100

After make the double alpha for ridge and lasso i.e. 20 and 200

For Ridge: ***Coeff values are increasing as alpha will increase. r2_score of train data is also drop from .807 to 0.45***

For Lasso: ***As alpha value increased more features removed from model. But r2score is also dropped by 1% in both test and train data***

Important Predictor Variables are: LotArea, OverallQual, OverallCond, YearBuilt, BsmtFinSF1, TotalBsmtSF, GrLivArea

Predictors are same but the coefficient of these predictor has changed.

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 lasso for the test dataset so we will choose lasso regression to solve this problem

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:

Five most important predictor variables are:

1stFlrSF
GrLivArea
Street_Pave
RoofMatl_Metal
RoofStyle_Shed

Predictor variables will be same but the coefficient of these predictor will slightly get changed.

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:

The model should be generalized so that the test accuracy is not lesser than the training score. The model should be accurate for datasets other than the ones which were used during training. Too much importance should not given to the outliers so that the accuracy predicted by the model is high. To ensure that this is not the case, the outliers analysis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which it does not make sense to keep must be removed from the dataset. If the model is not robust, It cannot be trusted for predictive analysis.