Assignment Based Subjective Answers

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?

Alpha is the penalty added to co-efficient to identify the best bias-variance tradeoff. Lower the value of alpha lower RSS and high variance. As alpha increases towards infinity the betas reduce to zero and hence variance is reduced with maximum RSS.

In our assignment example for Australian housing prediction dataset optimal value of alpha for Ridge

{'alpha': 20}

In [222]:

and for Lasso

{'alpha': 500}

When you double the value of alpha the coefficients will be further reduced. In case of Ridge it includes all variables for model prediction so we will not see much impact but incase of Lasso feature selection happens, so it will pick the predictor variables which has high influence on the response variable.

In our assignment example for Australian housing prediction dataset when Alpha value is doubled the co-efficient reduces

Ridge - Alpha :20

TotalSF	22237.2269
BsmtBath	4599.406601

Ridge-Alpha:40

TotalSF	4696.028157
BsmtBath	4406.406601

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?

Our dataset is having huge number of predictor variables. In the list of predictor variables not all of them have positive correlation with the response variable SalePrice so in this case choosing Lasso technique will be more appropriate as it helps in feature selection especially when you have a large dataset and provides a good accuracy of the model.

Also Ridge regression model did not give zero value for co-efficients whereas Lasso had reduced the co-efficients to few of the variables to zero.

Question 3

After building the model, you realized 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?

The five most important predictor variables which are strongly correlated with the target variable SalePrice is Total SF, OverallQual of the property, SquareFeet of the Bsmt, Garage Area, Fireplace availability,

If we exclude these variables and try to build a new model. Next set of most important predictor variables are Full Bath, TotalRmsAbvGrd, MasVnrArea, TotalPorchArea, HouseStoreyStyle

Question 4

How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

A robust model is something which performs better with unseen real dataset. The predictions are at high level of accuracy able to accommodate variance in the data with medium to low bias.

Unseen real data may have outliers. They can be introduced by human error, system error, natural deviation in measurement, variations in the scale. Models behavior is affected due to outliers. So handling outlier by using standard scaling techniques. retaining only the outlier relevant to the model we will be able to effectively handle outlier which will help us in building a robust mode and the accuracy is not affected which helps the model to be more generalizable.