Advanced Regression 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 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 and lasso regression are below:

Alpha for Ridge Regression is 1
Alpha for Lasso Regression is 10

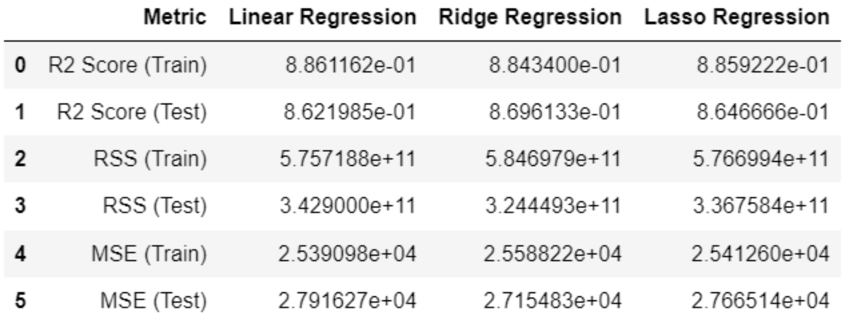
If we double the value of alpha for both Lasso and Ridge, R2score of training data will decrease while it will increase on testing data.

Important predictors after this change are implemented will be same as that of before but the coefficients of these predictors will change.

LotArea--------------Lot size in square feet
OverallQual--------Rates the overall material and finish of the house
OverallCond-------Rates the overall condition of the house
YearBuilt------------Original construction date
BsmtFinSF1-------Type 1 finished square feet.
TotalBsmtSF------- Total square feet of basement area
GrLivArea----------Above grade (ground) living area square feet.
TotRmsAbvGrd----Total rooms above grade (does not include bathrooms)
Street_Pave--------Pave Road access to property
RoofMatl_Metal----Roof material_Metal

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 go ahead with lasso regression to solve this problem.



	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	8.861162e-01	8.843400e-01	8.859222e-01
1	R2 Score (Test)	8.621985e-01	8.696133e-01	8.646666e-01
2	RSS (Train)	5.757188e+11	5.846979e+11	5.766994e+11
3	RSS (Test)	3.429000e+11	3.244493e+11	3.367584e+11
4	MSE (Train)	2.539098e+04	2.558822e+04	2.541260e+04
5	MSE (Test)	2.791627e+04	2.715483e+04	2.766514e+04

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:
LotArea, OverallQual, YearBuilt, BsmtFinSF1, TotalBsmtSF are the top 5 important predictors.

We will drop them and check for the next most important predictors. Below are the next most important predictors:

11stFlrSF-----------First Floor square feet
GrLivArea-----------Above grade (ground) living area square feet.
Street_Pave---------Pave Road access to property
RoofMatl_Metal------Roof material_Metal
RoofStyle_Shed------Type of roof (Shed)

***Note:** R2 score will decrease on dropping the top 5 important predictors.

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:
Model generalization is important so that the test accuracy is not lesser than the training score. Model should not only work for the data on which it was trained but also work perfectly on other datasets as well. We should make sure that the model doesn’t overfit. In other words, the model should not be too complex to be robust and generalizable.

If we look at the accuracy, a too complex model will have a very high accuracy. So, to make our model more robust and generalizable, we will have to decrease variance which will lead to some bias. Addition of bias means that accuracy will decrease.