Assignment 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?

The optimal value of alpha for ridge and lasso regression

Ridge Alpha 9

Lasso Alpha 0.001

After we double

Ridge Alpha 18

Lasso Alpha 0.002

Then we get these values

Ridge Regression-R2 value 0.9474634277506608

after we duble the value of apha

Ridge Regression-R2 value 0.9442547413095183

Lasso Regression-R2 value 0.9357785918945452

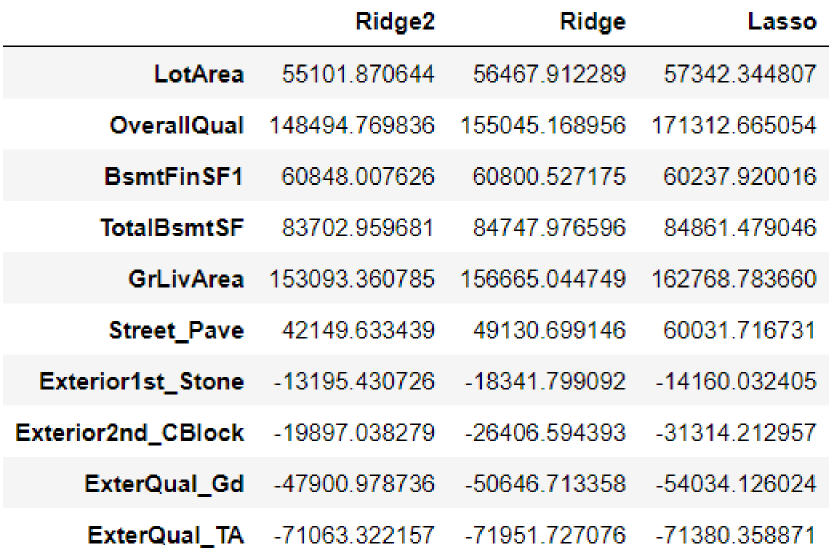
after we duble the value of apha

Lasso Regression-R2 value 0.9266265916896469

Here we see the R2 train vales has decrease after we double the alpha value.

And we can see here that R2 test value also decrease after we double the value of alpha.

most important predictor variables after the change is implemented



	Ridge2	Ridge	Lasso
LotArea	55101.870644	56467.912289	57342.344807
OverallQual	148494.769836	155045.168956	171312.665054
BsmtFinSF1	60848.007626	60800.527175	60237.920016
TotalBsmtSF	83702.959681	84747.976596	84861.479046
GrLivArea	153093.360785	156665.044749	162768.783660
Street_Pave	42149.633439	49130.699146	60031.716731
Exterior1st_Stone	-13195.430726	-18341.799092	-14160.032405
Exterior2nd_CBlock	-19897.038279	-26406.594393	-31314.212957
ExterQual_Gd	-47900.978736	-50646.713358	-54034.126024
ExterQual_TA	-71063.322157	-71951.727076	-71380.358871

LotArea---------------Lot size in square feet

OverallQual-----------Rates the overall material and finish of the house

BsmtFinSF1------------Type 1 finished square feet

TotalBsmtSF-----------Total square feet of basement area

GrLivArea-------------Above grade (ground) living area square feet.

Street_Pave-----------Pave Road access to property

Predictors are same but the coefficent of this predictor has changed.

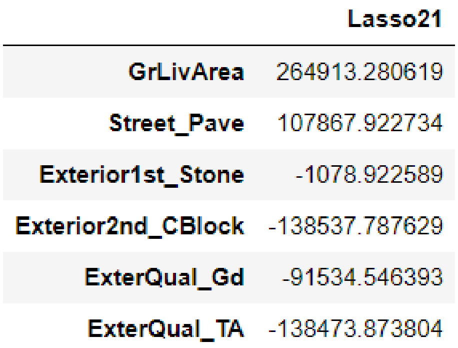
Question 2

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 ridge is slightly higher then lasso for the rest dataset so we will chose ridge 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?



	Lasso21
GrLivArea	264913.280619
Street_Pave	107867.922734
Exterior1st_Stone	-1078.922589
Exterior2nd_CBlock	-138537.787629
ExterQual_Gd	-91534.546393
ExterQual_TA	-138473.873804

these five most important variables

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 give 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.