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 for alpha (Working in the python file)

Ridge=0.2

Lasso=50

If we double the value of alpha for ridge to 0.4 and Lasso model to 100

r2 of training data has decreased but increased for the testing data

Predictors are same but the coefficient of these predictor has changed

Five important predictor

GrLivArea - Above grade (ground) living area square feet

OverallQual_10 - Very Excellent

OverallQual_9 - Excellent

TotalBsmtSF - Total square feet of basement area

BsmtFinSF1 - finished square feet

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:Lasso as the R2 value is better for the testing data.

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?

Answer.

Five most important predicted variable after dropping the previous top 5 variable found earlier in the model.(Calculation in the python file)

TotRmsAbvGrd : Total rooms above grade (does not include bathrooms

Neighborhood_Crawfor : Neighborhood location Crawfor

BsmtFullBath_3, -- Basement full bathrooms is 3

SaleCondition_Partial, : Home was not completed when last assessed (associated with New Homes)

GarageCars_3: Size of garage in car capacity is 3

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?

Model can be made robust and generalizable when the test accuracy is not lesser than the training accuracy. R2 value of the test data should remain similar and same for different data set.

the case, the outlier analysis needs to be done and only those which are relevant to the

datasets need to be retained. Those outliers which it does not make sense to keep must be

removed from the dataset. This would help increase the accuracy of the predictions made by the model