Advanced Regression Assignment part 2

Subjective Questions - Lloyd Dsouza

(Note: All code which are required for answering the questions are included in assignment notebook)

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?

Ridge Model optimal value **Alpha = 3.0**

R2 score Train- 0.9454851305736752

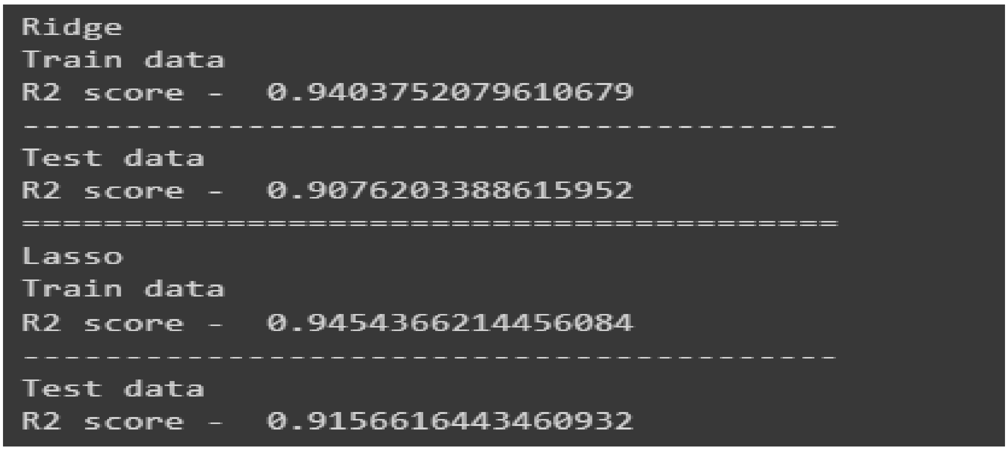
R2 score Test- 0.910460874446027

Lasso Model optimal value **Alpha = 0.0001**

R2 score Train- 0.9500612681333156

R2 score Test- 0.9133209848926895

If we double the values of Alpha I.e., Ridge (Alpha = 6.0) and (Lasso = 0.0002), Then there is a decrease in R2 score for the training and test data. And also, reduction in coefficient values.



```
Ridge
Train data
R2 score -  0.9403752079610679
----------------------------------------
Test data
R2 score -  0.9076203388615952
========================================
Lasso
Train data
R2 score -  0.9454366214456084
----------------------------------------
Test data
R2 score -  0.9156616443460932
```

The top features after the change are:

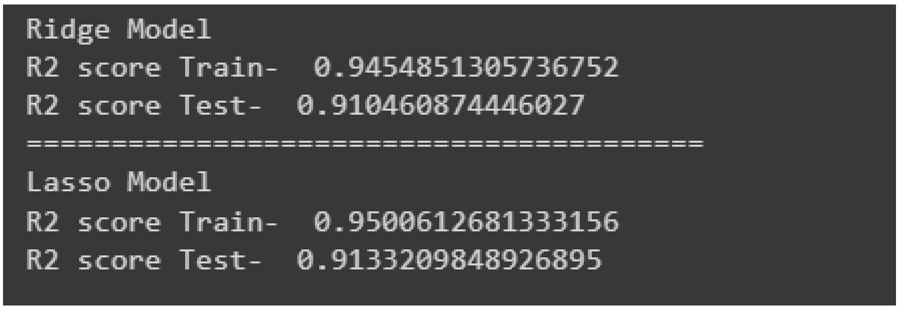
1.	GrLivArea
2.	OverallQual
3.	TotalBsmtSF
4.	YearBuilt
5.	OverallCond
6.	Heating_Grav
7.	Functional_Maj2
8.	MSZoning_FV
9.	LotArea
10.	BsmtFinSF1

(Code included in notebook)

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?

If we see the R2 scores of both ridge and lasso, Lasso scores are slightly better.



```
Ridge Model
R2 score Train-  0.9454851305736752
R2 score Test-  0.910460874446027
========================================
Lasso Model
R2 score Train-  0.9500612681333156
R2 score Test-  0.9133209848926895
```

Lasso has removed unwanted features from model which makes the model generalized, simple and accurate. So, Lasso can be selected.

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?

The top five predictor variables are

GrLivArea, OverallQual,, TotalBsmtSF, MSZoning, TotalBsmtSF, YearBuilt

After removing these variables and building the model. With R2 score - 0.9156616443460932 and mean squared error - 0.014132805539101852

The Top five features obtained are

MSSubClass, BsmtExposure_Mn, Foundation_Slab, Foundation_Stone, Foundation_Wood

(Code included in notebook)

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?

Model can be considered as robust if the model is stable does not change drastically upon changing the training set.

If the model does not overfit the training data, and works well with new data then then that model can be said as generalisable.

As defined in Bias Variance trade off, we can select an optimum point in bias and variance to get a robust and generalisable model.

Its implications are that performance on training and test data or new data would be same.