Advanced Regression Assignment

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?

Solution:
Optimal value of Alpha for Ridge Regression: 2.0
Optimal value of Alpha for Lasso Regression: 0.001

If we double the value of alpha for Ridge and Lasso, ie.

- Alpha for Ridge: 4.0
- Alpha for Lasso: 0.002

The most important features after change in alpha,

1. R-square :
    a. Ridge
        i. Train Data: 0.8868666302897418
        ii. Test Data: 0.8472134504321858
    b. Lasso
        i. Train Data: 0.01777804360115409
        ii. Test Data: 0.025181144315137335
2. Mean Squared Error
    a. Ridge
        i. Train Data: 0.8489659879231168
        ii. Test Data: 0.8338141483962711
    b. Lasso
        i. Train Data: 0.023733839616346137
        ii. Test Data: 0.027389517756666766

	Ridge	**Lasso**
1	OverallQual_9	OverallQual_9
2	TotRmsAbvGrd_12	GrLivArea
3	OverallCond_9	Neighborhood_NridgHt
4	GrLivArea	ConstructionAge
5	TotRmsAbvGrd_11	OverallQual_8

6	MSSubClass_160	KitchenQual_Fa
7	OverallQual_2	MSSubClass_160
8	Fireplaces_3	BsmtExposure_No Basement
9	MSZoning_RL	Neighborhood_Somerst
10	Neighborhood_NridgHt	BsmtExposure_Gd

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?

Solution:

Lasso regression provides a more simple model as it reduces the number of features. The number of features with 0 coefficients is more in Lasso than in Ridge. Also the difference in Train and Test r-square is less in Lasso. This means the model is better fitted than Ridge Regression.

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?

Solution

The new 5 most important predictor variables are:

1. OverallCond_3
2. OverallQual_4
3. KitchenQual_Fa
4. BsmtQual_No Basement
5. OverallQual_2

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?

Solution:

To verify if the model developed by any method is robust and generalizable, following things needs to be checked:

1. It is not impacted by the outliers in the data
2. Test and the training metrics are not much different.

Too much weightage should not be given to the outlier data. This ensures high accuracy of the model. This can either be done by handling outliers prior to the model building. Only if the outliers are the requirement of the problem concerned, then they shall be kept unhandled.

In case the model is robust and generalisable, then it leads to a decrease in standard accuracy of the model. There is always a trade-off between accuracy and robustness. If the model is more generalised, it tends to have a low standard accuracy.