ARUN PERI

PGDML, MLC38 (FEB BATCH)

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 1

The optimal value of alpha for the Lasso regression is obtained as 0.001 and for the Ridge regression it is 0.1.
If we double the value of alpha for lasso i.e., 0.002, we are getting the top important predictor variables and their coefficients are as follows:

	Featuere	Coef
0	MSSubClass	1.228122
1	LotArea	0.448666
7	BsmtUnfSF	0.444011
2	OverallQual	0.369056
28	PoolArea	0.183455
17	BedroomAbvGr	0.164196
30	MoSold	0.096387
29	MiscVal	0.086592

If we double the value of alpha for ridge i.e., 0.2, we will get the top important predictor variables and their coefficients are as follows:

	Feaure	Coef
0	MSSubClass	1.695342
30	MoSold	0.805584
7	BsmtUnfSF	0.472715
1	LotArea	0.420151
29	MiscVal	0.389370
23	WoodDeckSF	0.380451
2	OverallQual	0.362497
28	PoolArea	0.338625
14	BsmtHalfBath	-0.103419
17	BedroomAbvGr	-0.158968

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 2

We are getting the optimal value of alpha for Lasso ss 0.001 and for the Ridge ss 0.1. So, the optimal value of the lambda is 1000 for lasso and 10 for ridge.
For Lasso regression, we are getting the R2 score for train and test data for the optimal value as 0.88 (@ alpha = 0.001) and 0.796 (@ alpha = 0.001) respectively.
For Ridge regression, we are getting the R2 score for train and test data for the optimum value as 0.882 (@ alpha = 0.1) and 0.786 (@ alpha = 0.1) respectively.
We have chosen the Lasso model with optimum alpha value as 0.001 or lambda value as 1000. We are getting better test R2 score with a simpler model as the lambda value is very high. So, the Lasso model is more generalizable.

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 3

The first five most important variable and their constant for lasso model with the optimum value of alpha = 0.001 are as follows:

	Featuere	Coef
0	MSSubClass	1.364407
30	MoSold	0.492701
7	BsmtUnfSF	0.459485
1	LotArea	0.433104
2	OverallQual	0.364987

After removing these variable 'MSSubClass', 'MoSold', 'BsmtUnfSF', 'LotArea', 'OverallQual'. The Lasso model will show the following 5 coefficients as important as shown below:

	Featuere	Coef
0	OverallCond	0.990807
29	YrSold_old	0.664801
6	2ndFlrSF	0.506703
1	MasVnrArea	0.451921
13	BedroomAbvGr	0.450908

- R2 score for training and test data after removing the coefficient is coming out to be 0.8698 for training data and 0.767 for test data.

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 4

1. A model is robust when there is not much variation in the train and the test accuracy.

2. There should be a balance between the bias and regularization term. The Regularization term is controlled by parameter lambda.

3. If lambda if very less, then the regularization term will be very less and all the concentration to reduce the cost function will be on reducing the bias or the error term and the model will become very complex.

4. The accuracy of such model will be very high on the training set as the error will be very less, but the variance will be very high so the accuracy of the test data set will be very less.

5. On the other hand, when the lambda is very high then the regularization term is on more focuses to minimize the cost function. Then the model will be a simple model.

6. The accuracy of the training set will be very less but the variance will be very high. So, we have to make a balance between the error term and the regularization term, so need to have the model which is more generalizable.