Advanced Regression

Assignment Part - 2

Submitted

by

Vicky Bachu

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:

Alpha for Ridge : 100

Alpha for Lasso: 0.01

In the case of Ridge Regression: When we plot the curve between negative mean absolute error and alpha we observed that as the value of alpha increase from 0 the error term decreased. When the value of alpha is 100 the test error is minimum so we decided to go with value of alpha equal to 100 for our ridge regression.

For lasso regression we decided to keep very small value that is 0.01, when we increase the value of alpha the model try to penalize more and try to make most of the coefficient value zero.

When we double the value of alpha for our ridge regression, the model will apply more penalty on the curve and try to make the model more generalized that is making model more simpler and not fitting every data of the data set from the graph we can see that when alpha is 200 we get more error for both test and train.

Similarly when we increase the value of alpha for lasso we try to penalize more our model and more coefficient of the variable will reduced to zero, the value of our r2 square also decreases

The most important variable after the changes has been implemented for ridge regression are as follows:-

1. MSZoning_FV
2. MSZoning_RL
3. Neighborhood_Crawfor
4. MSZoning_RH
5. MSZoning_RM
6. SaleCondition_Partial
7. Neighborhood_StoneBr
8. GrLivArea
9. SaleCondition_Normal
10. Exterior1st_BrkFace

The most important variable after the changes has been implemented for lasso regression are as follows:-

1. GrLivArea
2. OverallQual
3. OverallCond
4. TotalBsmtSF
5. BsmtFinSF1
6. GarageArea
7. Fireplaces
8. LotArea
9. LotArea
10. LotFrontage

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 regression would be a better option it would help in feature elimination and the model will be more robust. The r2_score of lasso is slightly higher than lasso for the test dataset so we will choose lasso regression to solve this problem.

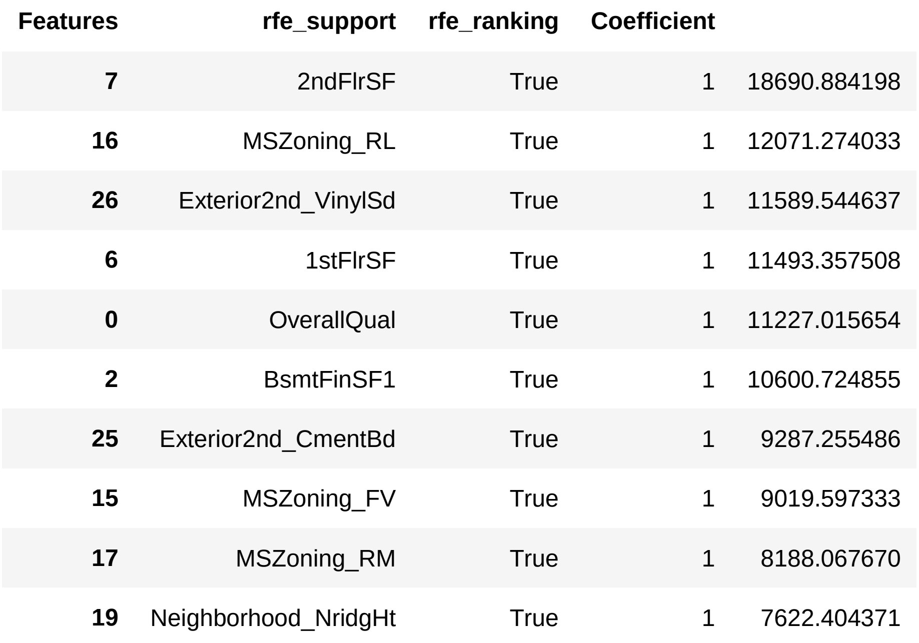
It is important to regularize coefficients and improve the prediction accuracy also with the decrease in variance, and making the model interpretably. Ridge regression, uses a tuning parameter called lambda as the penalty is square of magnitude of coefficients which is identified by cross validation. Residual sum or squares should be small by using the penalty. The penalty is lambda times sum of squares of the coefficients, hence the coefficients that have greater values gets penalized. As we increase the value of lambda the variance in model is dropped and bias remains constant. Ridge regression includes all variables in final model unlike Lasso Regression. Lasso regression, uses a tuning parameter called lambda as the penalty is absolute value of magnitude of coefficients which is identified by cross validation. As the lambda value increases Lasso shrinks the coefficient towards zero and it make the variables exactly equal to 0. Lasso also does variable selection. When lambda value is small it performs simple linear regression and as lambda value increases, shrinkage takes place and variables with 0 value are neglected by the model.

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?

Ans : new top five predictor variables will be :-

1. BsmtFinSF1
2. Exterior2nd_CmentBd
3. MSZoning_FV
4. MSZoning_RM
5. Neighborhood_NridgHt



Features	rfe_support	rfe_ranking	Coefficient	
7	2ndFlrSF	True	1	18690.884198
16	MSZoning_RL	True	1	12071.274033
26	Exterior2nd_VinylSd	True	1	11589.544637
6	1stFlrSF	True	1	11493.357508
0	OverallQual	True	1	11227.015654
2	BsmtFinSF1	True	1	10600.724855
25	Exterior2nd_CmentBd	True	1	9287.255486
15	MSZoning_FV	True	1	9019.597333
17	MSZoning_RM	True	1	8188.067670
19	Neighborhood_NridgHt	True	1	7622.404371

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 be given to the outliers so that the accuracy predicted by the model is high. To ensure that this is not the case, the outlier’s 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.