Subjective Questions and Solution- Advanced Regression

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:-

The Optimal Value of Ridge Regression is 2 and that of Lasso Regression is 0.001.
The R2 Score when value of alpha = 2 is 0.87405 and that of Lasso for alpha = 0.001 is 0.8663. When we double the value of alpha for ridge and lasso respectively, we d on't find any significant changes in the value of R2 Score.

The most important variables after changes are made in case of Ridge Regression ar e as follows:-

1) OverallQual
2) BsmtQual
3) TotalBsmtSF
4) HeatingQC
5) GrLivArea

The most important variables after changes are made in case of Lasso Regression ar e as follows:-

1) OverallQual
2) BsmtQual
3) HeatingQC
4) GrLivArea
5) KitchenQual

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:-

The optimal value that I have achieved in the assignment is as follows:-
Ridge- 0.6
Lasso- 0.0001

The Mean Squared error in case of Ridge and Lasso are:

-Ridge - 0.0216

-Lasso - 0.0215

The R2 Score in case of Ridge and Lasso are:

-Ridge - 0.8839

-Lasso - 0.8847

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. 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.

Models generated in Lasso Regression are more easier to interpret as compared to the Ridge Regression . Here, a good and optimal value of Lambda is crucial.

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-

After removing the top 5 predictor variables, the new important predictors are as follows:-

1) Neighborhood_ClearCr 2) Neighborhood_NoRidge 3) Neighborhood_Crawfor

4) KitchenQual 5) Fireplaces

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-

Bias: Bias is error in model, when the model is weak to learn from the data. High bias means model isunable to learn details in the data. Model performs poor on training and testing data.

Variance: Variance is error in model, when model tries to over learn from the data. High variance meansmodel performs exceptionally well on training data as it has very well trained on this of data but performs very poor on testing data as it was unseen data for the model.

It is important to have balance in Bias and Variance to avoid overfitting and under-fitting of data.

In case of Regularization techniques such as Ridge and Lasso , we need to find the optimal value , so as to reduce the variance with the small amount of compromise in the Bias.

Both the techniques shrinks the coefficients estimates towards 0