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 to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer1:

Optimal Value of alpha in Ridge: 4

Optimal Value of alpha in Lasso: .001

On double the value of alpha, In case of ridge that will lower the coefficients and in case of Lasso there would be more less important features coefficients turning 0.

The most important predictor variable after the change is implemented are those which are significant.

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:

Optimal Value of alpha for ridge and lasso regression are:

• Optimal Value of lambda for ridge: 4

• Optimal Value of lambda for Lasso: 0.001

As we got good score for both the models so we can go with Lasso regression as it results in model parameters such that lesser important features coefficients become zero.

Ridge: Train :88.79 Test :86.8 and

Lasso : Train :88.62 Test :86.7

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 :

On running the same notebook and removing the top 5 significant variables :

We found below variables as next 5 significant for Ridge:

- Neighborhood_StoneBr : 0.21
- BsmtExposure_Gd : 0.22
- Neighborhood_Somerst : 0.25
- Neighborhood_Crawfor : 0.26
- OverallQual_Very Good : 0.32

We found below variables as next 5 significant for Lasso:

- BsmtExposure_Gd : 0.2
- GrLivArea : 0.26
- Neighborhood_Somerst : 0.28
- constant : 0.29
- Neighborhood_Crawfor : 0.31

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 :

Here are some changes you can make to your model:

• Use a model that's resistant to outliers. Tree-based models are generally not as affected by outliers, while regression-based models are. If you're performing a statistical test, try a non-parametric test instead of a parametric one.

• Use a more robust error metric. switching from mean squared error to mean absolute difference (or something like Huber Loss) reduces the influence of outliers. I explain a bit about why this is the case at Why is the median a measure of central tendency? It doesn't have anything to do with any other values of the data set, so how does it "describe" the data set?

Here are some changes you can make to your data:

• Winsorize your data. Artificially cap your data at some threshold. See What are some applications of winsorization?

• Transform your data. If your data has a very pronounced right tail, try a log transformation.

• Remove the outliers. This works if there are very few of them and you're fairly certain they're anomalies and not worth predicting