Advanced Regression Assignment

Part II

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:

Optimal Value of Alpha found is as below:

1. Ridge -> 500
2. Lasso -> 0.01

If we double the value of Alpha in both models, as expected, r2 score (cross-val) will decrease. There are changes in the top-5 predictor variables are well.

For Lasso, with alpha as 0.01 top-5 predictor variables in ranking order were: GrLivArea, OverallQual, GarageCars, YearBuilt, OverallCond

On taking alpha=0.02 top-5 variables are: GrLivArea, OverallQual, GarageCars, YearBuilt, KitchenQual.

Hence KitchenQual has replaced OverallCond.

Similarly, for Rasso, with alpha=500, top-5 predictor variables in ranking order were: OverallQual, GrLivArea, 1stFlrSF, TotRmsAbvGrd, OverallCond

On taking alpha=1000, we get: OverallQual, GrLivArea, 1stFlrSF, TotRmsAbvGrd, FullBath

Therefore, FullBath replaced OverallCond in the list.

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:

After finding the optimal values of lambda for Ridge and lasso regression, Ridge performs well with better r2 as well as RMSE score. Therefore, I will choose Ridge after comparing performances on optimal lambda.

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 building new Lasso Regression model after removing top-5 predictors we get below columns as new top-5:

TotalBsmtSF, 2ndFlrSF, GarageCars, YearBuilt, KitchenQual

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:

To make sure model is generalizable we look at the difference between train and test accuracy/r2-score. If this difference is large then it means model is overfitting and hence is not generalizable. Here we will need to drop some variables. On the other hand if this difference is not too big, then the model performs fairly as expected on unseen dataset and is not overfitting.

Note: Code for finding out top-5 variables on required changes is present at the end of Solution Jupyter Notebook (part I).