Subjective Assignment – 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:

• In my final model, the optimal value of alpha for Ridge regression is 50, and for Lasso regression, it is 10.

• When we double these values, the model's performance remains the same in both cases.

• After these changes, the significant predictor variables for Ridge regression include Neighborhood_Stone Br, GarageArea, Neighborhood_NridgHt, TotalBsmtSF, GrLivArea, KitchenQual, Neighborhood_Names, Neighborhood_Edwards, BldgType_TwnhsE, and GarageFinish.

• In Lasso regression, the significant predictor variables are TotalBsmtSF, SaleType_New, MSZoning_RM, GarageType_Attchd, GrLivArea, Neighborhood_NAmes, Neighborhood_OldTown, KitchenQual, SaleCondition_Partial, and RoofStyle_Gable.

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:

We notice that both Ridge and Lasso yield comparable results in terms of performance. Despite the Ridge model exhibiting a slightly better performance (1%) than Lasso on the test dataset, we opt for the Lasso model for the final application. Lasso facilitates feature elimination, which is advantageous considering our dataset comprises over 130+ columns.

Therefore, feature elimination can aid in identifying the most significant predictor variables. Consequently, our final model is Lasso with an R-squared score of 88 on the training set and 84 on the test set, respectively.

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:

The top five significant predictor variables in our Lasso model are: 'GrLivArea', 'GarageType_Attchd', 'MSZoning_RM', 'SaleType_New', and 'TotalBsmtSF'. Upon removal of these variables and subsequent model rebuilding, the new top five significant predictor variables are: 'MasVnrArea', 'Neighborhood_StoneBr', 'Neighborhood_NridgHt', 'Fireplaces', and 'GarageArea'.

Question 4

How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

Answer:

Ensuring our model strikes the right balance between complexity and simplicity is key to its resilience and adaptability across various contexts. While pushing for higher accuracy by introducing more complexity might seem advantageous, it often compromises the model's ability to be applied broadly. A robust model, one that performs consistently well across both training and testing datasets, reflects its capacity to effectively generalize to new scenarios and data inputs.