Assignment – 2

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:

The optimal alpha values for ridge and lasso regression are 10 and 100, respectively. Doubling these values to 20 and 200 leads to noticeable shifts in the models. For ridge regression, coefficient values rise, but the R2 score decreases notably from 0.807 to 0.45. In lasso regression, higher alpha values result in the elimination of more features, causing a slight 1% drop in the R2 score for both training and testing data. Following the adjustments, the significant predictor variables in Ridge regression now include 'Neighborhood_StoneBr', 'GarageArea', 'Neighborhood_NridgHt', 'TotalBsmtSF', 'GrLivArea', 'KitchenQual', 'Neighborhood_Names', 'Neighborhood_Edwards', 'BldgType_TwnhsE', and 'GarageFinish'. Meanwhile, for Lasso regression, the important 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:

I will opt for Lasso since it has feature selection. We note that Ridge and Lasso exhibit comparable performance outcomes. Despite Ridge outperforming Lasso marginally (by 1%) on the test dataset, we opt for Lasso for our final application. Lasso's feature elimination capability is particularly beneficial for our dataset, which comprises over 130 columns. This feature elimination offers an advantage in identifying the most crucial predictor variables.

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 predictor variables in our Lasso model are now 'GrLivArea','GarageType_Attchd', 'MSZoning_RM', 'SaleType_New', and 'TotalBsmtSF'. After removing these and reconstructing the model, the new five most 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 generalisable? What are the implications of the same for the accuracy of the model and why?

Answer:

To ensure the model's robustness and generalizability, it's crucial to address three key criteria. Firstly, the model's accuracy should surpass 70-75%. In our scenario, it achieves 80% accuracy on the training set and 81% on the test set, effectively meeting this criterion. Secondly, all feature p-values should be below 0.05, signifying their statistical significance. Lastly, the VIF (Variance Inflation Factor) for all features should be under 5, indicating minimal multicollinearity concerns. Fulfilling these criteria instils confidence in the model's robustness and generalizability.