Subjective Assignment – Advanced Regression

Q1. 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?

Ans: After plotting a curve between the negative mean absolute error and alpha we see value of alpha is equal to 2 for ridge regression and for lasso regression the alpha value is 0.01.

With an increase in the value of alpha, the lower the value of the model coefficients, and more is the regularization for ridge regression.

For Lasso, the increase in alpha value, variance reduces with a slight compromise in terms of bias. The most important variables are Neighborhood_StoneBr, GarageArea, Neighborhood_NridgHt, TotalBsmtSF, GrLivArea , KitchenQual, Neighborhood_Names, Neighborhood_Edwards, BldgType_TwnhsE, GarageFinish

Q2. **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?**

Ans: We observe that Both Ridge and Lasso give very similar results in terms of performance. Although Ridge model performs slightly (2%) better than lasso on the test dataset, we still decide to choose the Lasso model to apply finally. Lasso helps with feature elimination and as our dataset has over 130+ columns, so feature elimination can be an advantage in realising the most important predictor variables.
Hence, our final model is Lasso with r2 score of 88 on train and 89 on test datasets respectively.

Q3: 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?

Ans: The five most important predictor variables in our Lasso model are:

OverallQual, OverallCond, Fireplaces, FireplaceQu_Gd, LotFrontage

If we remove these and rebuild the model, the five most important predictor variables now

are – 3SsnPorch, ScreenPorch, 2ndFlrSF, 1stFlrSF , GarageArea

Q4: 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?

Ans: By making sure the model is not over fitting and is as simple as possible, we are ensuring that it is robust and generalizable. The accuracy of the model will go up if we try to over fit the model but that no longer makes it generalizable. When the model is generalized the accuracy should be pretty good on both the training and the testing dataset making the model robust.