Advanced Regression – Problem Statement – Part II (Subjective Questions)

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?

For both ridge and lasso regression, I found the optimum value to be **100**. This was by running the best_params_ method. When I test with double the lambda (200), here's what I found for each model:

Ridge:

Ridge		
Lambda	Train	Test
100	0.8009	0.7816
200	0.7954	0.7774

With double the lambda in **ridge regression**, the r2 scores for both the train and test datasets **are slightly lower**. Therefore, increasing the lambda from the optimum value suggested by the method is undesirable.

Lasso		
Lambda	Train	Test
100	0.8058	0.787
200	0.8054	0.7859

With double the lambda in **lasso regression**, the r2 scores for both the train and test datasets **are slightly lower**. Therefore, increasing the lambda from the optimum value suggested by the method is undesirable.

Coming to the predictors, first of all, as shown in the answer to Question 3, I will choose lasso regression over ridge.

Going with that, in lasso, with **lambda of 100**, the most important predictors were:

KitchenAbvGr: -20075

OverallQual: 16525

GarageCars: 15277

TotRmsAbvGrd: 14069

Fireplaces: 13606

When we **increased the lambda to 200**, the most important predictors were:

KitchenAbvGr: -17018

OverallQual: 16855

GarageCars: 15056

TotRmsAbvGrd: 13830

Fireplaces: 13564

So, basically, the most important predictors have **remained the same**, although the coefficient values have changed by changing the lambda.

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?

As we saw earlier the optimum value for lambda suggested by the method for both ridge and lasso regression was **100**.

For the train dataset, the r2 value in lasso regression is slightly better than that for ridge, but very slightly. For the test dataset too, the r2 value for lasso is very slightly higher than that for ridge. So, although the difference is very minimal, if we go purely by numbers, **lasso would be the better choice**.

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?

The next 5 most important predictors would be:

BsmtCond: -13435

ExterQual: 11746

KitchenQual: 11698

BsmtQual: 10908

HalfBath: 9712

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?

The most practical approach to ensure that the model is robust and generalisable is to check if the r2 values for the train and test datasets are close enough to each other or not. The closer they are to each other, the better. If the test r2 is significantly lower that the train r2, it is a most probable sign of overfitting, meaning that that model might mostly have memorized the training data, and any slight difference in a new set of data presented to the model could cause chaos.

Complex model = High Accuracy = High variance (not robust enough to handle changes in input data).

We must aim at a model that has the lowest possible total error and that might mean having to decrease the model complexity by compromising on the bias and thereby reducing the variance. As we have learnt and tried out in this session, simple Linear Regression may give us results that seem pretty close, r2 value wise, for both the train and test datasets. However upon performing residual analysis, we might find that the assumptions of LR are not met, which in turn could be a symptom of high variance. Therefore, we can use other regression methods like Ridge or Lasso to give us better model in such scenarios, as they penalize the model equation if the error term is too high in value.