ANSWERS TO DUBJECTIVE QUESTIONS
SUBMISSION BY SIDDHARTHA GHOSH

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

I have observed from my model that with increasing levels of Alpha, the Negative Mean absolute Error has decreased. For both Ridge and Lasso Regression.

In Lasso Regression, a higher Alpha equates to stronger regularisation and therefore a simpler model. Conversely, Lower Alpha = more parameters = weaker regularisation.

It is common knowledge that there is usually no optimum value of Alpha, as its choice is largely judgmental based on model objectives and the type of data.

In my particular model I have introduced different levels of Alpha in the Ridge Regressions.

I did not notice much change in the R2 value between Alpha values 10 and 20.

The R2 Value was very similar in the Lasso Regression with Alpha 20 (0.94)

The Important Predictor Variables have been found to be: Neighbourhood, ExterQuality, Basement Quality and KitchenQuality.

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 have seen the R2s in both the Ridge and Lasso models be in around 0.94 and 0.95.

I would prefer choose Lasso because of the small number of significant parameters.

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:

In such case, I would remove them and choose the standardized coefficients with the next 5 largest absolute values.

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:

I would use three methods to do that:

1. R2: The Closer the R2 score is to 1, the better.
2. Cross Validation: Especially if the data set is small. Each fold generates its set of indicators, that are then cross-checked.
3. Sensitivity Test: To understand model tolerance to noise.