Problem Statement – Part 2

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 to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

- Optimal Value: Lasso = 100, Ridge = 5
- After Doubling the Alpha for Lasso (200), more so ever Coefficients will be moved towards 0 and in some cases exactly 0, resulting in fewer features and most possibly simpler model.
- After Doubling the alpha for Ridge (10), more coefficients are moved towards 0 and never 0. Bias will increase with more sacrifices on Variance
- Most important predictors across both would be OverallQual,GrLivArea,GarageCars,Neighorbood. Quality of the house with living area coupled with garage capacity and neighbood determines the price of House.

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?

Between Lasso and Ridge, my preferred choice is Lasso in this usecase.

Lasso not only does regularisation but also helps in feature selection by reducing the coefficients towards 0 and exactly 0 in some cases, resulting is a more simpler model with lesser features and easy to interpret/understand.

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?

Five most important Predictor variables:

- 1stFlrSF
- 2ndFlrSF
- GarageArea
- OverallCond
- MasVnrArea

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?

Below are the broad-based techniques/approaches that can be followed to have a model robust and Generalisable:

- **Regularization (Lasso & Ridge):** helps in managing the model complexity by reducing the Coefficients of Features towards 0 thereby avoiding overfitting. In this usecase Lasso case helped in indirectly doing **Feature Selection** reducing the features to considerable amount.
- **Cross Validation:** Instead of a single train-test-split, Cross validation using kfolds. This makes the model get trained and validated for k times. Just so that model results are not just “by chance”.
- **Pre-processing and EDA:** This ensures that outliers and other noises are identified and treated well before making sure the model is trained or tested with well vetted features/data.