Problem Statement - Part II

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?

- Optimal value of alpha for Ridge and Lasso regression is determined through process like
    - cross-validation : Perform K-fold validation, by splitting the dataset into K subsets on which we iteratively train our model on k-1 of these folds. We repeat this validation process K times with each fold serving as a validation dataset.
    - Grid Search : Define the range of Alpha from smallest to highest and validate the model to identify the best optimal value of alpha.
- If we choose to double the value of alpha for both Ridge and Lasso
    - In Ridge regression, this makes the regularization term more improved, causing the model to penalize large coefficients even more. This will lead to smaller and more evenly distributed coefficients and reducing risk of overfitting.
    - In Lasso regression, this makes more coefficients to exactly zeros and eliminate less important features from the model.
- In ridge, higher will retain all variables but downscale their importance whereas in case of lasso regression only a subset of predictor variables are considered important and other feature coefficients will be made exactly zero.

Below are the feature considered as high important

- GrLivArea
- GarageCars
- YearRemodAdd
- YearBuilt
- 1stFlrSF
- TotalBsmtSF

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?

Ridge regression is used when we want to reduce multicollinearity and overfitting in datasets with many features, preserving all variables' importance. In contrast, opt for lasso regression when you suspect irrelevant or redundant features in high-dimensional data, as it performs automatic feature selection by setting some coefficients to zero. Lasso is preferable when interpretability and a simpler model with fewer predictors are essential.

Question 3

After building the model, you realized 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?

-Reevaluate the model and identify the new five most important predictor variables among available ones.

Following steps for re evaluating the model

- Remove the top 5 predictor
- Create new lasso regression model
- Examine the coefficient.
- After examine below are the most important predictor.

    - 2ndFlrSF
    - 1stFlrSF
    - Neighborhood_NoRidge
    - KitchenAbvGr
    - Neighborhood_Crawfor

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 model which performs well on unseen data, handles variations and avoids over & underfitting.
- We can use below different strategies to make model more robust and generalisable
    - Regularization
    - Cross-validation
    - Hyperparameter tuning
    - Ensemble Method

- There would be implication of enhancing the model robustness and generalizability like
    - Decrease in Training accuracy, as those strategies try to fit training data too closely which leads to reduced models ability to reduce noise and minor fluctuations in training data.