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?

Answer:

Optimal value for alpha for
Ridge : 10
Lasso : 100

If I choose double the value of alpha for both Ridge and Lasso, i.e., 20 for Ridge and 200 for Lasso, the following changes might occur in the model:

1. For Ridge regression, increasing the alpha from 10 to 20 will result in a stronger penalty on the coefficients, causing them to shrink towards zero. The model complexity will be reduced, potentially preventing overfitting but possibly increasing bias.

2. For Lasso regression, increasing the alpha from 100 to 200 will also increase the penalty, causing the coefficients to shrink and possibly become exactly zero for some features(in our case **Exterior2nd_ImStucc** coefficients is zero). This can lead to feature selection, providing a simpler and more interpretable model but potentially losing some predictive power.

The most important predictor variables after the change is implemented are mentioned below:
Neighborhood_NridgHt
Neighborhood_NoRidge
OverallQual
GrLivArea
Neighborhood_Somerst

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:

We will lasso regression for model interpretability and feature selection. By choosing ridge regression for stability in multicollinearity. But still we need to compare both models using cross-validation and select the one with better performance based on metrics like MSE, R-squared, or MAE.

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:

The Top 5 feature that we will drop are **'Neighborhood_NoRidge', 'Neighborhood_NridgHt', 'GrLivArea', 'OverallQual', 'Neighborhood_Somerst'**. After dropping the feature we have observed that the R2 score is drop to 56% for Train and 58% for test from 81.5% for Train and 82% for Test
Next top 5 features after drooping 5 main predictors are
GarageCars ,Neighborhood_Crawfor, MSZoning_RL, TotalBsmtSF, Exterior1st_CemntBd

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:

To ensure a model is robust and generalisable, use a diverse and representative dataset, split data into training, validation, and test sets, apply cross-validation, use regularization techniques, perform feature selection, choose an appropriate model, and optimize hyperparameters. Ensuring robustness and generalisability improves the model's ability to make accurate predictions on unseen data, as it reduces overfitting and increases adaptability to various conditions.