Assignment - II 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?

Answer - The optimal value of alpha for ridge and lasso regression is 10 and 0.001. If we double the alpha, the r2 score is observed as below :

Lasso Regression -
Train → 0.91, Test → .90

Ridge Regression -
Train → 0.93, Test → .90

The most important predictor variables after doubling the value of alpha are - BsmtFullBath, OverallCond, Neighborhood_Crawfor, YearBuilt, GarageArea, RoofMatl_CompShg.

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 - Lasso tends to do well if there are a small number of significant parameters and the others are close to zero.

Whereas Ridge works well if there are many large parameters of about the same value.

Hence, we will go ahead with Lasso Regression.

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?

Answer - The five most important predictor variables in the lasso model are - **BsmtFullBath, OverallCond, Neighborhood_Crawfor, YearBuilt and MasVnrArea.** After dropping these columns, the five most important predictor variables are - **2ndFlrSF, LowQualFinSF, Functional_Typ, Neighborhood_MeadowV and Neighborhood_Edwards**

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 - A robust model will continue to make accurate predictions even when faced with challenging situations. In other words, robustness ensures that a model can generalize well to new unseen data. To ensuring that a model is robust and generalizable we can make the following considerations.

1. Cross-Validation: Using k-fold cross-validation helps to evaluate how well the model generalizes to unseen data and reduces the risk of overfitting to a specific dataset.
2. Regularization: Models have a regularization parameter (alpha) that controls the strength of the regularization. Perform hyperparameter tuning to find the optimal value of alpha that balances the trade-off between fitting the training data well and maintaining generalization to new data.
3. Feature Scaling: Ensuring that the features are appropriately scaled.
4. Outlier Handling: Regression model is sensitive to outliers, so we need to handle outliers appropriately.

   Implications on Accuracy -

- If there is a significant difference between the accuracy achieved on the training set and the testing set, it suggests overfitting. The model has likely memorized the training data rather than learning the underlying patterns, leading to poor generalizations.
- Consistent Performance Across Different Data Splits: A robust model should show consistent performance across different random data splits.