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:
For Ridge Regression (Doubled alpha model, alpha=8*2=16):

The most important top 10 predictor variables after the change is implemented are as follows:
- 'GrLivArea'
- 'AgeofProperty'
- 'OverallQual'
- 'MSZoning_FV'
- 'MSSubClass_160'
- 'MSZoning_RL'
- 'Neighborhood_Crawfor'
- 'OverallCond'
- 'MSSubClass_70'
- 'TotalBsmtSF'

For Lasso Regression (Doubled alpha model, alpha=0.001*2=0.002):

The most important top 10 predictor variables after the change is implemented are as follows:
- 'GrLivArea'
- 'AgeofProperty'
- 'MSZoning_FV'
- 'MSSubClass_160'
- 'OverallQual'
- 'Neighborhood_Crawfor'
- 'MSZoning_RL'
- 'MSSubClass_70'
- 'OverallCond'
- 'MSSubClass_90'

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:

The R2 test score for the Lasso Regression Model is slightly better compared to the Ridge Regression Model. Additionally, there's a slight reduction in training accuracy, indicating that the Lasso model performs better on unseen data, making it an optimal choice.

The Mean Squared Error (MSE) for the test set in Lasso Regression is slightly lower than that of the Ridge Regression Model, suggesting that Lasso Regression performs better on unseen test data. Moreover, Lasso Regression facilitates feature selection by setting the coefficients of some insignificant predictor variables to zero, giving it an advantage over Ridge Regression.

In real-world regression analysis, analysts often encounter challenges such as outliers, non-normality of errors, and overfitting, particularly in sparse datasets. Using L2 norm regularization (Ridge) may expose analysts to these risks. Hence, employing L1 norm regularization (Lasso) could be beneficial as it offers robustness against such risks, resulting in better and more robust regression models.

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 5 most important features in the Original Lasso Model are:

'GrLivArea', ‘MSZoning_FV’, ‘MSSubClass_160’,’Exterior1st_BrkComm’, and ‘AgeofProperty’

The 5 most important features in the New Lasso Model are:

'Neighborhood_IDOTRR', 'Neighborhood_MeadowV', 'Neighborhood_OldTown', 'Neighborhood_BrDale', and 'Exterior2nd_Brk Cmn'.

Question 4
How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

Answer:
The robustness of a model refers to its ability to maintain consistent performance on both training and testing data, even when subjected to noise or variations in the dataset. Achieving robustness is essential for ensuring that the model can generalize well to unseen data beyond the training set.

Regularization techniques play a crucial role in enhancing the robustness of a model by controlling the trade-off between complexity and bias. By penalizing overly complex models, regularization helps maintain an optimal level of complexity, thereby improving the model's generalizability.

To enhance robustness and generalizability, it's vital to strike a balance between model simplicity and effectiveness. A model should be sufficiently simple to avoid overfitting to the training data while still capturing essential patterns in the data.

The Bias-Variance Trade-off is a key concept in achieving this balance. A highly complex model may fit the training data well but can be unstable and overly sensitive to small changes in the dataset, leading to high variance. On the other hand, overly simple models may exhibit high bias, resulting in inaccurate predictions across different datasets.

Maintaining an optimal balance between bias and variance helps minimize the total error of the model, ensuring accurate predictions on unseen data.