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 Alpha Values for Ridge and Lasso Regression:

- ➔ Ridge Regression: Optimal Alpha = 10
- ➔ Lasso Regression: Optimal Alpha = 0.001

When doubling the alpha values for both Ridge and Lasso regression:

- ➔ Ridge Regression: Increasing alpha leads to a reduction in coefficients, potentially diminishing the influence of predictors.
- ➔ Lasso Regression: Higher alpha values may prompt more coefficients to converge to zero, indicating the insignificance of certain features.

Following this adjustment, the most crucial predictor variables are those that retain significance despite the alterations in alpha values.

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 :
Optimal values of alpha for Ridge and Lasso regression:

- ➔ Optimal Value of alpha for Ridge: 10
- ➔ Optimal Value of alpha for Lasso: 0.001

Considering the good scores achieved by both models, Lasso regression is favored as it results in model parameters where less important features have coefficients reduced to zero.

Model Performance:

- ➔ Ridge Regression: Train - 90.9, Test - 87.4
- ➔ Lasso Regression: Train - 89.8, Test - 86.4

These results suggest that while Lasso slightly underperforms compared to Ridge on the training set, it demonstrates similar performance on the test set. Therefore, Lasso's ability to discard less relevant features while maintaining comparable performance makes it a favorable choice.

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 :
After discovering that the five most crucial predictor variables identified by the Lasso model are absent in the incoming data, we need to create a new model excluding these variables.

Initially, the top 5 predictors identified by Lasso were: OverallQual_9, GrLivArea, OverallQual_8, Neighborhood_Crawfor, and Exterior1st_BrkFace. With an optimal alpha value of 0.001, we proceed to build a new Lasso regression model.

Subsequently, we examine the top 5 features significant in predicting house values according to the updated Lasso model:

```
2ndFlrSF              0.10
Functional_Typ        0.07
1stFlrSF              0.07
MSSubClass_70         0.06
Neighborhood_Somerst  0.06
Name: Lasso, dtype: float64
```

After omitting the original top 5 predictors, the new top 5 predictors are:

- ➔ 2ndFlrSF
- ➔ Functional_Typ
- ➔ 1stFlrSF
- ➔ MSSubClass_70
- ➔ Neighborhood_Somerst

This adjustment ensures that the model remains robust and relevant despite the absence of the previously identified key predictors.

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 :
Here are some recommended changes to improve your model and data handling:

Model Improvements:

- ➔ Consider using a model that is less sensitive to outliers. Tree-based models tend to be more robust to outliers compared to regression-based models.
- ➔ When conducting statistical tests, opt for non-parametric tests over parametric ones, as they are less affected by outliers.
- ➔ Switch to a more robust error metric such as mean absolute difference or Huber Loss, which reduces the influence of outliers. This is particularly useful in situations where outliers significantly impact model performance.

Data Handling Recommendations:

- ➔ Apply Winsorization to your data, which involves artificially capping extreme values at a predetermined threshold. This helps mitigate the impact of outliers on your analysis.
- ➔ Explore data transformations such as log transformations, especially if your data exhibits a pronounced right tail. This can help normalize the distribution and make it more suitable for modeling.
- ➔ Consider removing outliers from your dataset if they are few in number and are determined to be anomalies rather than representative of the overall pattern.

Implementing these changes can help enhance the robustness and accuracy of your model while mitigating the impact of outliers on your analysis.