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, the best alpha value is 10.

• For Lasso Regression, the ideal alpha value is 0.001.

After doubling the alpha value, these are the observed alterations in the metrics for Ridge Regression:

The R2 score of the training set fell slightly from 0.94 to 0.93, while the R2 score of the testing set remained stable at 0.93.

Similarly, in the case of Lasso regression, after doubling the alpha value, the following changes were observed:

The R2 score of the training set dropped from 0.92 to 0.91, while the R2 score of the testing set fell from 0.93 to 0.91.

The most influential predictor variables post the adjustment in alpha value for Lasso were:

GrLivArea 1.11, OverallQual_8 1.09, OverallQual_9 1.08, Functional_Typ 1.07, Neighborhood_Crawfor 1.07, TotalBsmtSF 1.05, Exterior1st_BrkFace 1.05, CentralAir_Y 1.04, YearRemodAdd 1.04, Condition1_Norm 1.03.

For Ridge Regression, the most prominent predictor variables after doubling the alpha value were:

GrLivArea 1.08, OverallQual_8 1.07, OverallQual_9 1.07, Neighborhood_Crawfor 1.07, Functional_Typ 1.06, Exterior1st_BrkFace 1.06, OverallCond_9 1.06, TotalBsmtSF 1.05, CentralAir_Y 1.05, OverallCond_7 1.04.

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 decision to employ either Ridge or Lasso regression largely depends on the specific requirements of the task at hand.

• If the scenario involves numerous variables and we are mainly focused on feature selection, Lasso regression would be the preferred choice.

• On the other hand, if the task primarily involves reducing the magnitude of the coefficients to avoid overfitting, Ridge regression would be a better option.

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?

In this scenario, we would exclude the top five features identified in the initial Lasso model and proceed to construct a new model. Following this, the five most significant predictors that emerged are as follows:

• **2ndFlrSF**

• **Functional_Typ**

• **1stFlrSF**

• **MSSubClass_70**

• **Neighborhood_Somers**

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:

• Robustness of a model refers to its ability to maintain performance despite variations in the input data.

• A generalizable model is one that performs well not just on the training data, but also on new, unseen data, drawn from the same underlying distribution.

• Ensuring robustness and generalizability primarily involves preventing overfitting. An overfit model captures the noise or fluctuations in the training data too well, which negatively impacts its performance on new data. It tends to have high variance, meaning small changes in the data can cause large changes in the predictions.

• If we consider accuracy, an overly complex model might show high accuracy on the training data but perform poorly on unseen data. To make a model more robust and generalizable, we often have to accept some increase in bias, which can lead to a decrease in accuracy. This bias-variance tradeoff is an integral part of model design.

• Finding the right balance between model complexity and accuracy is key. Regularization techniques such as Ridge Regression and Lasso are effective tools for achieving this balance, as they can prevent overfitting by adding a penalty term to the loss function that discourages complex models.