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:

In the case of ridge regression, when we plot the curve between negative mean absolute error and alpha we see that as the value of alpha increase from 0 the error term decreases; the train error is showing increasing trend when value of alpha increases. When the value of alpha is 2 the test error is minimum, so we decided to go with value of alpha equal to 2 for our ridge regression.

In Lasso regression, a very small value 0.01 is chosen. As we increase the alpha value, the model intensifies its efforts to penalize coefficients more aggressively, pushing many coefficients towards zero. Initially, the negative mean absolute error stood at 0.4 with the corresponding alpha value.

When we double the value of alpha for our ridge regression we will take the value of alpha equal to 4 the model will apply more penalty on the curve and try to make the model more generalized that is making model more simpler and no thinking to fit every data of the data set. From the graph we can see that when alpha is 10 we get more error for both test and train. Similarly, when we increase the value of alpha for lasso we try to penalize more our model and more coefficient of the variable will reduced to zero, when we increase the value of our r2 square also decreases.

The most important variable after the changes has been implemented for ridge regression are as follows:

1. Neighborhood_Crawfor
2. MSZoning_FV
3. SaleType_New
4. MSZoning_RL
5. Neighborhood_StoneBr
6. SaleCondition_Normal
7. MSZoning_RH
8. SaleCondition_Alloca
9. MSZoning_RM
10. Exterior1st_Stucco

The most important variable after the changes has been implemented for lasso regression are as follows: -

1. GrLivArea
2. OverallQual
3. TotalBsmtSF
4. OverallCond

5. GarageArea
6. BsmtFinSF1
7. YearRemodAdd
8. Fireplaces
9. LotArea
10. MSZoning_RL

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:

Regularizing coefficients becomes pivotal to improve prediction accuracy, decrease variance, and ensure model interpretability. In Ridge regression, a tuning parameter named lambda is employed, penalizing coefficients based on their squared magnitude, which is determined through cross-validation. This penalty aims to minimize the residual sum of squares by scaling coefficients relative to lambda. Consequently, higher lambda values penalize larger coefficient values more significantly. As lambda increases, Ridge regression lowers model variance while preserving bias, and it retains all variables in the final model.

Conversely, Lasso regression also uses lambda as a penalty, targeting coefficients' absolute magnitudes via cross-validation. With increasing lambda values, Lasso progressively reduces coefficients toward zero, allowing variable selection by nullifying specific coefficients. When lambda is small, Lasso behaves similarly to simple linear regression. However, as lambda grows, shrinkage occurs, causing the model to exclude variables with zero coefficients.

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 5 most important variables now are:

1. BsmtFinSF1
2. YearRemodAdd
3. Fireplaces
4. LotArea
5. MSZoning_RL

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:

Simplicity in the model is key, even at the cost of reduced accuracy, as it leads to greater robustness and generalizability. This concept aligns with the Bias-Variance trade-off, where a simpler model tends to have higher bias but lower variance, resulting in enhanced generalizability. For accuracy, a robust and generalizable model maintains consistent performance across both training and test data, displaying minimal accuracy fluctuations between the two.

Bias denotes the model's error when it struggles to learn from data. High bias signifies the model's inability to capture intricate data details, leading to poor performance on both training and testing data sets.

Variance represents the model's error when it overly learns from the data. High variance indicates exceptional performance on training data but dismal performance on testing data, as the model wasn't exposed to this specific data during training.

Maintaining a balance between Bias and Variance is crucial to prevent both overfitting and underfitting of data, ensuring the model's optimal performance.