Problem Statements

Q1. You have determined the optimal value of lambda for ridge and lasso regression during theassignment. Now, which one will you choose to apply and why?

Ans: Regularizing coefficients is crucial for enhancing prediction accuracy, reducing variance, and improving model interpretability.

Ridge regression applies a penalty equal to the square of the coefficient magnitudes, controlled by a tuning parameter called lambda, determined through cross-validation. The goal is to minimize the residual sum of squares, adding a penalty that is the product of lambda and the sum of the squared coefficients. Larger coefficients are more heavily penalized, leading to a decrease in model variance as lambda increases, while bias remains relatively constant. Ridge regression includes all variables in the final model, differing from Lasso Regression.

Lasso regression also uses lambda as a tuning parameter, but the penalty is the absolute value of the coefficient magnitudes. With increasing lambda values, Lasso progressively shrinks coefficients towards zero, effectively setting some to exactly zero, thus performing variable selection. At lower lambda values, it resembles simple linear regression, but as lambda increases, more pronounced shrinkage occurs, and variables with zero-value coefficients are excluded from the model.

Q2. 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?

Those 5 most important predictor variables that will be excluded are :-

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

Q3. 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?

Ans: The principle of model simplicity aligns with the Bias-Variance trade-off, advocating for a model that is robust and generalizable, even if it means a slight decrease in accuracy. A simpler model tends to have higher bias but lower variance, making it more generalizable. This is evident in its performance across training and test data, where a robust and generalizable model will exhibit consistent accuracy on both, indicating minimal fluctuation between training and test results.

Bias refers to the error due to overly simplistic models that fail to capture the complexity of the data. High bias indicates that the model is unable to learn key details from the data, leading to poor performance on both training and test datasets.

Variance, on the other hand, is the error due to overly complex models that overfit the training data. High variance models excel on training data, as they are finely tuned to that specific dataset, but they perform poorly on unseen test data.

Striking a balance between bias and variance is crucial to avoid both overfitting and underfitting, ensuring the model is both accurate and generalizable.

Q4. What is the optimal value of alpha for ridge and lasso regression? What will be the changes in themodel 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?

Ans: In Ridge Regression, the relationship between the negative mean absolute error and alpha shows that increasing alpha from 0 initially decreases the error. However, as alpha continues to rise, the training error tends to increase. At an alpha value of 2, the test error reaches its minimum, leading to the decision to set alpha at 2 for optimal Ridge Regression performance.

For Lasso Regression, a small alpha value of 0.01 was chosen. Increasing alpha in Lasso leads to greater penalization, pushing more coefficients towards zero. Initially, the combination of a 0.01 alpha with a negative mean absolute error resulted in 0.4.

Doubling the alpha value in Ridge Regression to 10 increases the penalty, which in turn simplifies the model, moving away from fitting each data point in the dataset. The graph illustrates that an alpha of 10 results in increased error for both the training and test datasets.

In a similar vein, raising the alpha in Lasso Regression leads to more severe penalization. This results in a greater number of coefficients being reduced to zero. Consequently, as the alpha value increases, there is a decrease in the model's R2 score, reflecting a compromise between complexity and fit.

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

1. MSZoning_FV
2. MSZoning_RL
3. Neighborhood_Crawfor
4. MSZoning_RH
5. MSZoning_RM
6. SaleCondition_Partial
7. Neighborhood_StoneBr
8. GrLivArea
9. SaleCondition_Normal
10. Exterior1st_BrkFace

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

1. GrLivArea
2. OverallQual

3. OverallCond
4. TotalBsmtSF
5. BsmtFinSF1
6. GarageArea
7. Fireplaces
8. LotArea
9. LotArea
10. LotFrontage