Q1. 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?

Ans: To improve prediction accuracy, reduce variance, and enhance model interpretability, regularizing coefficients is essential. Both Ridge and Lasso regression use a tuning parameter called lambda, identified through cross-validation, but they differ in their approaches to regularization.

Regularizing coefficients is essential for improving prediction accuracy, reducing variance, and making the model more interpretable.

Ridge Regression:

- Ridge regression uses a tuning parameter called lambda, identified through cross-validation, as a penalty based on the square of the magnitude of coefficients. This penalty aims to minimize the residual sum of squares by penalizing larger coefficients. As lambda increases, the variance of the model decreases while the bias remains constant. Unlike Lasso regression, Ridge regression retains all variables in the final model.

Lasso Regression:

- Lasso regression also uses a tuning parameter called lambda, identified through cross-validation, but the penalty is based on the absolute value of the magnitude of coefficients. As lambda increases, Lasso shrinks some coefficients towards zero, effectively performing variable selection. When lambda is small, Lasso behaves like simple linear regression. As lambda increases, coefficients of less important variables are reduced to zero, simplifying the model by excluding these variables.

Choosing between the two depends on the specific needs of the model:

- If you need to retain all features and reduce multicollinearity, **Ridge Regression** is preferable.
- If you aim for a more interpretable model with feature selection, **Lasso Regression** is the better choice.

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?

Ans: The five most important predictor variables to be excluded are:

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

The next step is to identify the new five most important predictor variables without these excluded variables.

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: Ensuring a model is robust and generalizable involves keeping it as simple as possible. While this may lead to a decrease in accuracy, it makes the model more robust and generalizable. This concept can be understood through the Bias-Variance trade-off. A simpler model typically has higher bias but lower variance, making it more generalizable.

The implications for accuracy are that a robust and generalizable model will perform consistently well on both training and test data, meaning the accuracy does not significantly change between the two datasets.

Bias: Bias is the error introduced by the model when it fails to learn from the data. High bias indicates the model is too simple and unable to capture the underlying patterns, leading to poor performance on both training and testing data.

Variance: Variance is the error introduced when the model overfits the data, capturing noise instead of the actual pattern. High variance means the model performs exceptionally well on the training data but poorly on the testing data due to its inability to generalize.

Achieving a balance between bias and variance is crucial to avoid overfitting and underfitting, ensuring the model performs well on unseen data.

Q4. 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?**

Ans: For ridge regression, when we plot the curve between negative mean absolute error and alpha, we observe that as the value of alpha increases from 0, the error term decreases, while the train error shows an increasing trend. When the value of alpha is 2, the test error is minimized, so we decided to use an alpha of 2 for our ridge regression.

For lasso regression, I decided to use a very small value of 0.01. Increasing the value of alpha causes the model to apply more penalization, driving more coefficients to zero. Initially, we observed that at alpha 0.4, the negative mean absolute error stabilized, but a lower alpha balances the bias-variance trade-off better.

When we double the value of alpha for ridge regression to 4, the model applies more penalty, making it more generalized and simpler, but it also increases the error for both test and train datasets.

Similarly, doubling the value of alpha for lasso regression increases penalization, leading to more coefficients being reduced to zero and a decrease in the R^2 score.

The most important variables after the changes have been implemented are as follows:

Ridge Regression:
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

Lasso Regression:
1. GrLivArea
2. OverallQual
3. OverallCond
4. TotalBsmtSF
5. BsmtFinSF1
6. GarageArea
7. Fireplaces
8. LotArea
9. LotArea
10. LotFrontage