Question1- 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?

Ridge regression and Lasso regression are regularization techniques aimed at improving prediction accuracy while also reducing variance and ensuring model interpretability.

Answer = Ridge regression introduces a penalty term, lambda, which is determined through cross-validation. This penalty is applied as the square of the magnitude of coefficients, effectively penalizing coefficients with large values. By increasing lambda, the model's variance decreases while bias remains constant. Unlike Lasso regression, Ridge regression includes all variables in the final model.

On the other hand, Lasso regression also employs a tuning parameter, lambda, determined via cross-validation. The penalty in Lasso regression is the absolute value of the magnitude of coefficients. Increasing lambda in Lasso regression leads to coefficients being shrunk towards zero, eventually resulting in some coefficients being exactly zero. This feature of Lasso regression enables variable selection, as variables with zero coefficients are neglected by the model.

In summary, both Ridge and Lasso regression techniques offer ways to improve model performance, reduce overfitting, and enhance interpretability by introducing penalties on the coefficients based on lambda, with Lasso regression additionally providing a mechanism for variable selection

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?

The simpler the model the more the bias but less variance and more generalizable. Its implication in terms of accuracy is that a robust and generalisable model will perform equally well on both training and test data i.e. the accuracy does not change much for training and test data.

Bias: Bias is error in model, when the model is weak to learn from the data. High bias means model is unable to learn details in the data. Model performs poor on training and testing data.

Variance: Variance is error in model, when model tries to over learn from the data. High variance means model performs exceptionally well on training data as it has very well trained on this of data but performs very poor on testing data as it was unseen data for the model. It is important to have balance in Bias and Variance to avoid overfitting and under-fitting of 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?

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 decrease and 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 I have decided to keep very small value that is 0.01, when we increase the value of alpha the model try to penalize more and try to make most of the coefficient value zero. Initially it came as 0.4 in negative mean absolute error and alpha.

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