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?

Ans: For ridge regression-While plotting the curve between negative mean absolute error and alpha we see that as the value of alpha increases from 0 , the error term decreases 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.

For 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 0.Initially it came as 0.4 in negative mean absolute error and alpha.

When we double the value of alpha for our ridge regression ,we will take the value of alpha equal to 10 ,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, the graph we can see that when alpha is 10 ,we get more error for both test and train.

Important predictor variables after change being implemented for ridge regression are as follows:

1.MSZoning_FV

2.MSZONING_RL

3.Neighbourhood_Crawfor

4.MSZoning_RH

5.MSZoning_RM

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

Q2: 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?

Lasso regression would be a better option it would help in feature elimination and the model will be more robust

>Regularization is an important concept that is used to avoid overfitting of the data,

especially when the trained and the test data are much varying

>Ridge regression uses lambda as the tuning parameter

As the penalty is the square of magnitude of coefficients which is identified by cross validation.

Residual sum or residual sum of squares should be small by using the penalty.The penalty is lambda times sum of squares of the coefficients hence the coefficients that have greater value gets penalized.

As the lambda value increases the variance in the model gets dropped and bias remains constant.

Lasso regression ,uses lambda as the tuning parameter

As the penalty is absolute value of magnitude of coefficients which is identified by cross validation.

By adding the L1 regularization term, LASSO regression can shrink the coefficients towards zero. When lambda is sufficiently large, some coefficients are driven to exactly zero. Conversely ,a smaller value of lambda reduced the regularization effect, allowing more variables to have non-zero coefficients.

Question 3: 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 excluding the five most important predictor variables. Which are the five most important predictor variables now?

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

1.GrLivArea

2.OverallQual

3.OverallCond

4.TotalBsmtSF

5.GarageArea

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?

Ans:To ensure that the machine learning model is robust and generalisable so that they are not impacted by outliers in the training data. It should also be generalisable so that the test accuracy is not lesser that the training score. The model should be accurate for datasets other than the ones which were used during the training. Too much weightage should not be given to the outliers so that the accuracy predicted by the model is high.

The implications of having a robust and generalizable model are that it can perform well on new and unseen data. It can also help to reduce the risk of overfitting and underfitting problems. However ,it may come at the cost of reduced accuracy, as the model may not be able to capture everything of the data.