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 case of ridge regression, when we plot curve between negative mean absolute error and alpha we see that as the value of alpha increases from 0 the error tem decreases and train error is showing increasing trend when value of alpha increases.

For lasso regression I decided to keep very small value that is 0.001, when we increase the value of alpha the model try to penalize more and try to make most of the coefficient value zero.

The most important variables for ridge regression are as follows:-

01 MiscVal
02 BsmtHalfBath
03 HalfBath
04 LowQualFinSF
05 BsmtFullBath
06 Neighborhood_Gilbert
07 EnclosedPorch
08 TotRmsAbvGrd
09 GrLivArea
10 Neighborhood_IDOTRR

The most important variables for lasso regression are as follows:-

01 MiscVal
02 BsmtHalfBath
03 LowQualFinSF
04 BsmtFullBath
05 HalfBath
06 BsmtFinType
07 Neighborhood_Gilbert
08 LotShape
09 HeatingQC
10 Neighborhood_BrkSide

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

It is important to regularize coefficients and improve the prediction accuracy also with the decrease in variance, and making the model interpretably.

Ridge regression uses a tuning parameter called lambda as the penalty is square of magnitude of coefficients which is identified by cross validation. Residual sum or squares should be small by using the penalty. The penalty is lambda times sum of squares of the coefficient, hence the coefficients have greater values get penalized. As we increase the value of lambda the variance in model is dropped and bias remains constant.

Lasso regression, uses a tuning parameter called lambda as the penalty is absolute value of magnitude of coefficient which is identified by cross validation. As the lambda value increases Lasso shrinks the coefficient towards zero and it make the variable exactly equal to 0. Lasso also does variable selection. When lambda value is small it performs simple linear regression and as lambda value increases.

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?

01 Neighborhood_Gilbert
02 EnclosedPorch
03 TotRmsAbvGrd
04 GrLivArea
05 Neighborhood_IDOTRR

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:
The model must be as simple as possible, though its accuracy will decrease but it will be more robust and regenerlisable. It can also be Using the Bias variance trade off. The simpler the model the more the bias, less variance and more generalisable. It’s an implications 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 the difference between the average prediction of our model and the correct value which we are trying to predict. Model with high bias pays very little attention to the training data and oversimplifies the model. It always leads to high error on training and test data.

Variance: Variance is the variability of model prediction for a given data point or a value which tells us spread of our data. Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn't seen before. As a result, such models perform very well on training data but has high error rates on test data.