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:

For ridge regression, when we plot the curve between negative mean absolute error and alpha we can see that the optimum value of alpha is 100 where the test error is minimum.

For lasso regression, the optimum value of alpha should be 0.01.

When we double the value of alpha for ridge, we need to penalize the curve more and try to make the model more generalized

When we double the value of alpha for lasso, we need to penalize the model more coefficients will be reduced to zero.

The most important variables after the changes are:

BsmtQual_Ex
RoofMatl_WdShngl
RoofMatl_Membran
BsmtQual_None
BsmtFinType1_None
RoofMatl_Tar&Grv
RoofMatl_CompShg
RoofMatl_Roll
PavedDrive_P
PavedDrive_Y

For lasso regression:

BsmtQual_Ex
RoofMatl_WdShngl
RoofMatl_Membran
BsmtQual_None
BsmtFinType1_None
RoofMatl_Tar&Grv
RoofMatl_CompShg

RoofMatl_Roll
PavedDrive_P
PavedDrive_Y

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:

Ridge regression uses lambda as the penalty is square of the magnitude of coefficients. Residual sum of squares should be small by the penalty. Penalty= lambda*(sum of squares of coefficients). Hence, coefficients with greater value gets penalized and as we increase the lambda the variance model is dropped, and bias remains constant. Ridge regression also includes all the variables in the final model unlike lasso regression. Lasso regression shrinks the coefficient towards zero and it makes the variables exactly zero which are neglected by the model.

Hence, we will prefer Ridge regression.

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 another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Answer:

1. GrLivArea
2. OverallQual
3. OverallCond
4. TotalBsmntSF
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?

Answer:

We need to make sure that

- The model is as simple as possible
- The simpler the model, them more bias but less variance and more generalizable.

Its implication on accuracy is that a robust and generalized model will perform equally well on both training and test data. The accuracy does not change much for training and test data.