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:

The optimal value of alpha for **Ridge and Lasso are 5 and 0.0004 respectively**. After changing the optimal value of alpha for Ridge from 5 to 10, we can see that there is a slight change in sequence of the important predictor variables as coefficients are changing.

After changing the optimal value of alpha for Lasso from 0.0004 to 0.0008 ,we can see that there is a change in sequence of the important predictor variables as coefficients are changing.

MSZoning_RL is the most important predictor when value of alpha = 10 in Ridge. GrLivArea is the most important predictor when value of alpha = 0.0008 in Lasso.

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:

I will choose **Lasso Regression** with optimal value lambda of 0.0004. Lasso Regression is providing better R2 score. The Mean Squared of Lasso is less than Ridge comparatively.

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 of squares should be small by using the penalty. The penalty is lambda times sum of squares of the coefficients, so the coefficients that have greater values gets penalized. As we increase the value of Lambda the variance in the model dropped and bias remains constant. It includes all variables in final model unlike lasso regression.

Lasso Regression uses a tuning parameter called lambda as penalty is absolute value of magnitude of coefficients which is identified by cross validation. As the lambda increases lasso shrinks the coefficients towards zero and it make the variables exactly equal to zero. It also does variable selection. When lambda value is small it performs simple linear regression and as lambda value increases, shrinkage takes place and variables with zero value are neglected by the model.

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?

Answer:

As per Analysis on the new model, the five most important predictor variables are

1. “2ndFlrSF”
2. “1stFlrSF”
3. “TotalBsmtSF”
4. “OverallCond”
5. “Foundation_PConc”

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 should be as simple as possible, though its accuracy will decrease but it will be most Robust and generalisable. It can be also understood using the Bias-Variance trade-off. The simpler the model the more the bias but less variance and more generalisable. Its implication in terms of accuracy is that a robust and generalizable model will perform equally well on both training and test data set i.e., the accuracy will be similar for training and test data sets.

Other determining factors are Bias and Variance. Bias is error in model, if it is high bias means the model is unable to learn details in the data and performs poor on training and testing data. Variance is error in the model, if it is high variance means model performs exceptionally well on training data as it very well trained on the data but performs very poor on the testing data, means model tries to over learn from the test data.

We should always try to have balance in Bias and Variance to avoid overfitting and underfitting of the data set.