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 :

Optimal Value :

Ridge(alpha=5.0)
Lasso(alpha=0.0004)

When we double the value of alpha for our ridge regression no 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 .from the graph we can see
that when alpha is 10 we get more error for both test and train.
Similarly when we increase the value of alpha for lasso we try to penalize more our model and more coefficient of

The most important variable after the changes has been implemented for ridge regression are as follows:

MSZoning_RL
GrLivArea
OverallQual
MSZoning_RM
MSZoning_FV
OverallCond
TotalBsmtSF
Foundation_PConc
GarageCars
BsmtFinSF1

The most important variable after the changes has been implemented for lasso regression are as follows:-

GrLivArea
MSZoning_RL
OverallQual
MSZoning_FV
MSZoning_RM
TotalBsmtSF
OverallCond
Foundation_PConc
GarageCars
BsmtFinSF1

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 make the model interpretable.

Ridge regression, uses a tuning parameter called lambda as the penalty is square of magnitude of coefficients which are 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 coefficients, hence the coefficients that have
greater values get penalized. As we increase the value of lambda the variance in model is dropped and
bias remains constant. Ridge regression includes all variables in the final model unlike Lasso Regression.

Lasso regression, uses a tuning parameter called lambda as the penalty is absolute value of magnitude
of coefficients which are identified by cross validation. As the lambda value increases Lasso shrinks the
coefficient towards zero and it makes the variables exactly equal to 0. Lasso 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 0 value are neglected by the model

The Mean Squared Error of Lasso is slightly lower than that of Ridge
Also, since Lasso helps in feature reduction, Lasso has a better edge over Ridge.

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 :

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

1) MSZoning_RL
2) GrLivArea
3) MSZoning_RM
4) OverallQual
5) MSZoning_FV

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 more 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 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 an error in a 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 poorly 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.