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 Ridge regression, when we plot the curve, alpha vs negative mean absolute error, the value of alpha increase from 0 the error term decrease and the train error is showing increasing trend when value of alpha increases. When the value of alpha: 2 the test error is min, so I decided to go with value of alpha equal to 2 in ridge regression.

In Lasso regression, I kept a 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 zero. Initially, it came as 0.4 in negative mean absolute error and alpha.

When I 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 I increase the value of alpha for lasso, it try to penalize more our model and most coefficient of the variable was reduced to 0, when we increase the value of our r2 square also decreases.

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

- MSZoning_FV
- MSZoning_RL
- Exterior1st_BrkFace
- SaleCondition_Normal
- Neighborhood_Crawfor
- SaleCondition_Partial
- Neighborhood_StoneBr
- GrLivArea
- MSZoning_RH
- MSZoning_RM

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

- GrLivArea
- OverallQual
- OverallCond
- TotalBsmtSF
- BsmtFinSF1
- GarageArea
- Fireplaces
- LotArea
- LotFrontage

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’s 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 coefficients, hence the coefficients that have greater values gets penalized. As we increase the value of lambda the variance in model is dropped and bias remains constant. Ridge regression includes all variables in final model unlike Lasso Regression.

Lasso regression, uses a tuning parameter called lambda as the penalty is absolute value of magnitude of coefficients which is identified by cross validation. As the lambda value increases Lasso shrinks the coefficient towards zero and it make 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.

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:

The most important predictor variables that will be excluded are :-
1. OverallQual
2. GrLivArea
3. TotalBsmtSF
4. GarageArea
5. OverallCond

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 point to be noted while designing the model is to keep the model simple as possible. By doing this, the model itself become robust and generalisable. There may be a chance that the accuracy would be compromised but the model will be more robust and generalisable.

Its implications in terms of model accuracy would be such that, it will perform equally well both training & test data sets. This means that there is not much difference between the accuracy of both test and train set.

Variance is error in the model, when the model tries to memorizes the data(training dataset). Higher the variance, model performance exceptionally well on training dataset but poorly performs on test dataset, as the data is unseen.

Bias is error in 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 poor on training and testing data.