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:

Ridge regression: When I plot the graph for ‘negative mean absolute error’ vs ‘alpha score’ we see that as the value of alpha increase from 0 and the error term decrease.

The train error is showing increasing trend when value of alpha increases. when the value of alpha is 2 the test error is minimum, hence I decided for further evaluation with value of alpha equal to 2 for our ridge regression.

Lasso regression: I have decided to use very small value alpha 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 produces values as 0.4 in ‘negative mean absolute error’ vs ‘alpha score’. 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 variable will reduced to zero, when we increase the value of our r2 square also decreases.

The most important variables after the changes have been implemented for ridge regression are as follows	**The most important variable after the changes has been implemented for lasso regression are as follows**
MSZoning_FV	GrLivArea
MSZoning_RL	OverallQua
Neighborhood_Crawfor	
MSZoning_RH	
MSZoning_RM	
SaleCondition_Partial	
Neighborhood_StoneBr	
GrLivArea	
SaleCondition_Normal	
Exterior1st_BrkFace	

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 - It 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 does have one obvious disadvantage. It would include all the predictors in the final model. This may not affect the accuracy of the predictions but can make model interpretation challenging when the number of predictors is very large.

Lasso regression - It 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.

Generally, Lasso should perform better in situations where only a few among all the predictors that are used to build our model have a significant influence on the response variable. So, feature selection, which removes the unrelated variables, should help

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 five most important predictor variables that will be excluded are:

1. GrLivArea

2. OverallQual

3. OverallCond

4. TotalBsmtSF

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:

The model should be as simple as possible, though its accuracy will decrease little, 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 (no overfitting/ underfitting).

Bias: 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.

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.