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 1:

Optimal value of alpha for Ridge = 2

Optimal value of alpha for Lasso = 0.01

For lasso regression, when we increase the value of alpha the model tries to penalize more and makes most of the coefficients value equal to zero.

When we double the value of alpha for our ridge regression, the model will apply more penalty on the curve and try to make the model more generalized and simpler to fit every data of the data set. Here when alpha is doubled, error increases for both test and train sets.

Most important predictor variables after the change :

1. GrLivArea
2. OverallQual
3. OverallCond
4. TotalBsmtSF
5. BsmtFinSF1
6. GarageArea
7. Fireplaces
8. LotArea
9. LotFrontage
10. BsmtFullBath

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 2: Though the model performance by Ridge Regression was better in terms of R2 values of Train and Test,

I will choose Lasso, since it brings and assigns a zero value to insignificant features, enabling us to choose better predictive variables.

Lasso model is simple yet robust, in this case.

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 3 :

GarageArea

Fireplaces

LotArea

LotFrontage

BsmtFullBath

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 4 : This can be checked using the Bias-Variance trade-off.

The simpler the model the more will be 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 is error in the model when the model is weak to learn from the data. High bias would mean that the model is unable to learn important details in the data. Model would perform poor on training and testing data.

Variance is error in the model when model tries to over learn from the data. High variance would mean that the model performs exceptionally well on training 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.The model should be as simple as possible, though its accuracy may decrease but it will be more robust and generalisable.