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

The optimal value of alpha for lasso is taken as .01 and for ridge it is taken as 2.

When the alpha value for lasso is doubled, it increases the penalty on the curve and more coefficient are set to zero. As far as ridge is concerned doubling the value of alpha will penalize the model reducing the selected feature variable.

Ridge top 10 features

- MSSubClass
- LotFrontage
- LotArea
- OverallQual
- OverallCond
- MasVnrArea
- BsmtFinSF1
- BsmtFinSF2
- TotalBsmtSF
- 1stFlrSF

Lasso top 10 features

- OverallQual
- GrLiveArea
- GarageArea
- OverallCond
- Fireplaces
- BsmtFullBath
- TotalBsmtSF
- FullBath
- WorkDeckSF
- LotArea

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?

In my analysis both ridge and lasso have given similar performance in terms of r2 and rmse score. I have opted to apply lasso as it provides the top features out of box, lasso model is easier to comprehend. On the other ridge does not provide the list of top features out of the box and it requires some extra analysis to zero down the top features.

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?

- GarageCars
- GarageType_Attchd
- BsmtFinType1_GLQ
- Foundation_PConc
- CentralAir_Y

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:

The simple models are prone to have more bias and less ability to explain the variance in test dataset. The complex model will have low bias but high variance, the complex models are difficult to comprehend. To identify the robust and generalisable model, once has to find a difficult between analysing the variance-bias trade off. The robust and generalisable model may not give extremely high value of accuracy but there won't be high difference between train and test set performances.