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 to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:

For lasso regression, the optical alpha value is 0.0005 and for ridge regression optimal value of alpha is 5.

If we double the value of alpha in ridge regression, the model will apply more penalty on the curve and try to make model more generalized. Similarly, when we increase the value of alpha for lasso we try to penalize more and coefficient of the variables will reduced to zero.

The variables that generally affect the price are the Living area square feet, Zoning classification, Overall quality and condition of the house, Foundation type of the house, Number of cars that can be accommodated in the garage, Total basement area in square feet and the Basement finished square feet area.

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:

The Mean Squared error in case of Ridge and Lasso are:

- Ridge - 0.01362
- Lasso - 0.01341

The Mean Squared Error of Lasso is slightly lower than that of Ridge

Lasso helps in feature reduction as the coefficient value of one of the features became 0, hence 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:

The next five most important predictor variables are:

TotalBsmtSF - Total square feet of basement area

OverallCond - Rates the overall condition of the house

Foundation_PConc - Poured Concrete type of foundation

GarageCars - Size of garage in car capacity

BsmtFinSF1 - Type 1 finished square feet

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 robust and generalisable as possible so that they are not impacted by outliers in the training data. The test model accuracy should not lesser than the training score. The model should be accurate for datasets other than the ones which were used during training. Too much weightage should not give to the outliers so that the accuracy predicted by the model is high.

The outlier analysis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which it does not make sense to keep must be removed from the dataset. This would help increase the accuracy of the predictions made by the model. Confidence intervals can be used, typically 3-5 standard deviations. This would help standardize the predictions made by the model.