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:

The optimum alpha for Ridge is 5, and the optimum alpha for Lasso is 0.0001.

After doubling the value of alpha for Ridge and Lasso the R2 score is almost similar and MSE is 0.

Below are the important features from Ridge regression

[('OverallQual', 0.113), ('GrLivArea', 0.095), ('Neighborhood_NoRidge', 0.076), ('2ndFlrSF', 0.071), ('GarageArea', 0.058), ('TotalBsmtSF', 0.053), ('FullBath', 0.052), ('BsmtFullBath', 0.046), ('Neighborhood_NridgHt', 0.044), ('Neighborhood_Crawfor', 0.041)]

Below are the important features from Lasso regression

[('GrLivArea', 0.299), ('OverallQual', 0.146), ('Neighborhood_NoRidge', 0.082), ('YearBuilt_Age', -0.076), ('KitchenAbvGr', -0.065), ('OverallCond', 0.055), ('LotArea', 0.054), ('KitchenQual_TA', -0.051), ('BsmtFullBath', 0.048), ('Neighborhood_NridgHt', 0.047)]

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 at Optimal ALPHA: 5.0

R2 SCORE: 0.8440% MSE: 0.0

Lasso Regression at Optimal ALPHA: 0.0001

R2 SCORE: 0.8568% MSE: 0.0

Lasso Provides a better R2 score and Mean square error. So I will choose Lasso Regression model for predicting the Housing prices.

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:

After removing the top five predictor variables and with the new Lasso Regression model below are the list of important predictor variables and their corresponding coefficient values.

[('TotalBsmtSF', 0.364), ('2ndFlrSF', 0.174), ('KitchenQual_Fa', -0.084), ('MasVnrArea', 0.083), ('GarageArea', 0.081), ('KitchenQual_TA', -0.081), ('BsmtQual_Fa', -0.076), ('BsmtQual_TA', -0.069), ('LotArea', 0.056), ('KitchenQual_Gd', -0.054)]

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 simpler the model the more the bias but less variance and more generalizable. The 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: High bias model is unable to learn details in the data. The model performs poor on training and testing data.
- Variance: High variance means the model performs exceptionally well on training data.
  It is important to have a balance between Bias and Variance to avoid overfitting and under-fitting of data.