Housing Problem Statement 2

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 : 10

Optimal value of alpha for lasso : 100

When we double the value of alpha for both ridge and lasso:

Alpha for ridge : 20 - Coeff values are decreasing as alpha will increase. r2_score of train data is drop from .818 to 0.815 and test data is drop from 0.81 to 0.808

Alpha for lasso : 200 - As alpha value increased more feature coeff are decreasing in value. r2score is train data is also drop from .82 to 0.818 and test data is drop from 0.811 to 0.807

Top 5 features are `Neighborhood_NoRidge`, `Neighborhood_NridgHt`, `1stFlrSF`, `OverallQual`, `2ndFlrSF`.

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

I have chosen Lasso as it helps in feature selection. It has removed the unwanted features without affecting the model accuracy. Hence, making the model more generalized, simple and accurate.

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

Top 5 features are `Neighborhood_NoRidge`, `Neighborhood_NridgHt`, `1stFlrSF`, `OverallQual`, `2ndFlrSF`. After dropping the top 5 features model r2 score is low i.e. 60.9% for Train and 61.8% for Test.
Five most important predictor variables now are `Neighborhood_Somerst`, `Neighborhood_StoneBr`, `Exterior1st_BrkFace`, `GarageFinish_NA`, `KitchenQual`.

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

A model is said to be robust and generalisable when :

1) Model accuracy should be > 70%. In our case it is 82% for train and 81% for test.
2) P-value of all features is < 0.05
3) VIF of all features is < 5

Thus, we are sure that the module is robust and generalisable.