Advanced Regression

HOUSING PRICE PREDICTION

Rohini Kannapiran
MLC61

Q1 :

a. What is the optimal value of alpha for ridge and lasso regression?
Optimal values are as follows:
Ridge =10.0
Lasso =100

b. What will be the changes in the model if you choose double the value of alpha for both ridge and lasso?
After doubling the values of alpha, changes are as follows:

Optimal values:

Model	**R2 train**	**R2 test**
Ridge	83.8	85.3
Lasso	84.1	84.7

Doubled alpha values:

Model	**R2 train**	**R2 test**
Ridge	82.8	85.4
Lasso	83.3	85

Observation - For the given dataset, R2 scores of training data are reduced slightly and test data shows a slight increase in both Ridge and Lasso.

c. What will be the most important predictor variables after the change is implemented?
The top 5 important predictor variables remain the same, even after changing the hyperparameter, but the coefficients differ.
OverallQual, MasVnrArea, BsmtFullBath, BsmtHalfBath, FullBath

Q2: 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?

Optimal value for Ridge found as alpha =10 and the same for Lasso is found to be alpha =100.

Metrics for the above alpha values are as shown below

	Metric	Linear Regression	RFE	Ridge Regression	Lasso Regression
0	R2 Score (Train)	8.787247e-01	8.526555e-01	8.385029e-01	8.417587e-01
1	R2 Score (Test)	8.335035e-01	8.085899e-01	8.531680e-01	8.477299e-01
2	RSS (Train)	8.225097e+11	9.993148e+11	1.095301e+12	1.073219e+12
3	RSS (Test)	4.036249e+11	4.640211e+11	3.559537e+11	3.691369e+11
4	MSE (Train)	2.838296e+04	3.128515e+04	3.275320e+04	3.242137e+04
5	MSE (Test)	3.035651e+04	3.254856e+04	2.850754e+04	2.903065e+04

With reference to above metrics, I prefer Ridge as it performs better compared to the Lasso Regression, which shows almost equal performance in both train and test datasets.

Q3: 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?

After removal of top 5 features, the R2 score value shows a significant drop in accuracy. Also, R2 of train is more than that of test data set.

HalfBath, BedroomAbvGr, KitchenAbvGr, TotRmsAbvGrd, Fireplaces

	Lasso
HalfBath	6318.568743
BedroomAbvGr	-6768.282209
KitchenAbvGr	-42687.324302
TotRmsAbvGrd	17693.720448
Fireplaces	16087.355180

Q4: 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?

The model should be as simple as possible, while attention is given to bring in too much error. Implication in terms of accuracy is defined by the values remain closer for both train and test data.