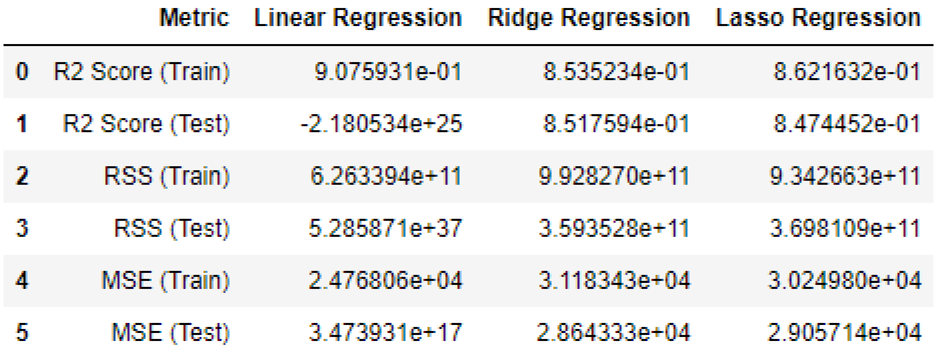
Question 1

Answer

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?

Optimal value of alpha for ridge regression: 500

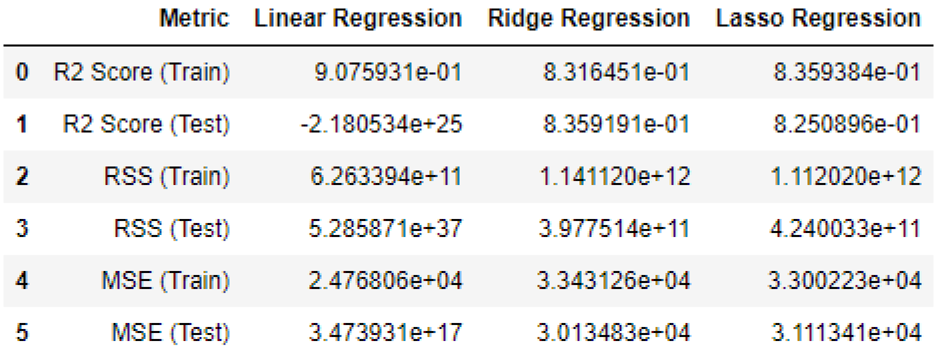
Optimal value of alpha for lasso regression: 1000



	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.075931e-01	8.535234e-01	8.621632e-01
1	R2 Score (Test)	-2.180534e+25	8.517594e-01	8.474452e-01
2	RSS (Train)	6.263394e+11	9.928270e+11	9.342663e+11
3	RSS (Test)	5.285871e+37	3.593528e+11	3.698109e+11
4	MSE (Train)	2.476806e+04	3.118343e+04	3.024980e+04
5	MSE (Test)	3.473931e+17	2.864333e+04	2.905714e+04

Doubled alpha for ridge regression: 1000

Doubled alpha for lasso regression: 2000



	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.075931e-01	8.316451e-01	8.359384e-01
1	R2 Score (Test)	-2.180534e+25	8.359191e-01	8.250896e-01
2	RSS (Train)	6.263394e+11	1.141120e+12	1.112020e+12
3	RSS (Test)	5.285871e+37	3.977514e+11	4.240033e+11
4	MSE (Train)	2.476806e+04	3.343126e+04	3.300223e+04
5	MSE (Test)	3.473931e+17	3.013483e+04	3.111341e+04

For ridge regression the most important predictor variables after the changes is:

zoning_FV	9.6
PavedDrive_Y	9.3
Neighborhood_ClearCr	8.7
GarageType_Attchd	8.8

BsmtCond_Gd	7.82
HouseStyle_2Story	7.76
LandContour_Lvl	7.5
RoofMatl_Membran	7.4
GrLivArea	7.3
Electrical_SBrkr	7.2

For lasso regression the most important predictor variables after the changes is:

GrLivArea	26380
Neighborhood_NridgHt	12643
GarageCars	11379
Neighborhood_NoRidge	8878
TotalBsmtSF	7648
BsmtExposure_Gd	6762
BsmtFinType1_GLQ	5380
RoofMatl_WdShngl	5379
SaleType_New	4401
Foundation_PConc	4112

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

I would choose ridge over lasso because of several factors:

1. ridge has lesser alpha therefore lesser penalty and regularization as more regularization occurs the more the model is closer to underfitting

2. r2 score of ridge is lesser.

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

Before:

GrLivArea	26974.716924
Neighborhood_NridgHt	12751.967853
GarageCars	9729.528831
Neighborhood_NoRidge	9702.718035
BsmtExposure_Gd	7153.316067

After:

2ndFlrSF	20358.139141
1stFlrSF	16156.906414
TotalBsmtSF	9392.504423
GarageArea	8719.892347
RoofMatl_WdShngl	7308.885343

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

1. By eliminating extreme outliers the dataset
2. By having a threshold in the dataset
3. By making sure you are using the right model for the dataset

These optional steps will bridge the gap closer between train and test data metrics such as accuracy, rss, rmse, etc.

A model is more robust and generalizable when train and test metrics are closer in value.