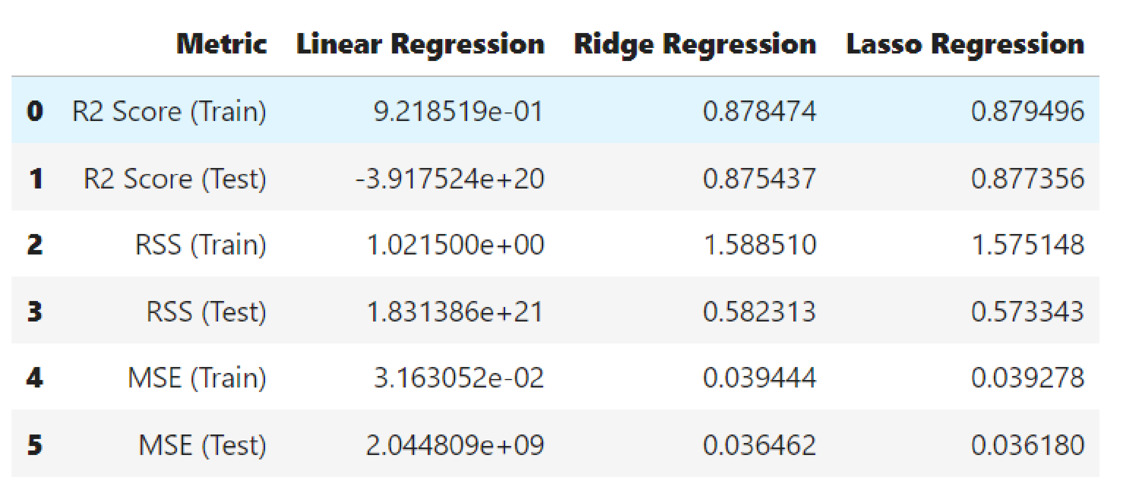
Question 1

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?

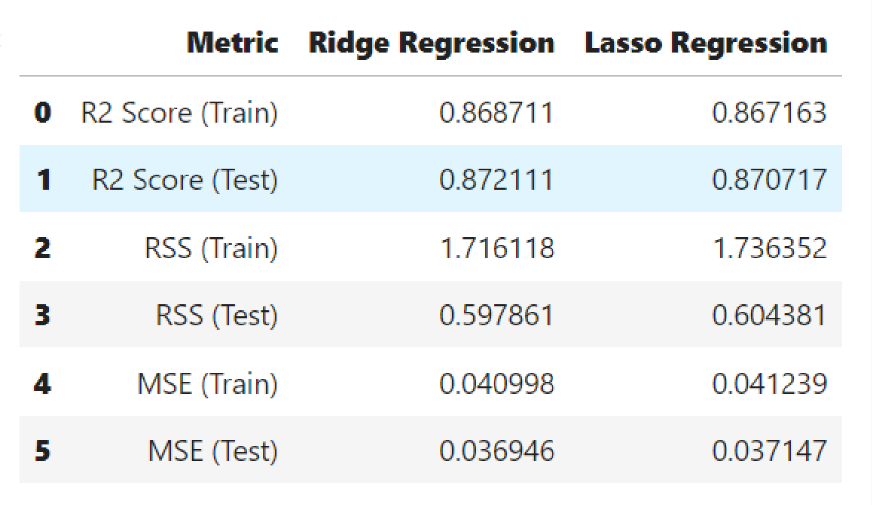
Answer :

The optimal value of alpha for Ridge regression came out to be 4 while the same for Lasso regression is 0.0001. The metrics model using these alpha models are as below



	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.218519e-01	0.878474	0.879496
1	R2 Score (Test)	-3.917524e+20	0.875437	0.877356
2	RSS (Train)	1.021500e+00	1.588510	1.575148
3	RSS (Test)	1.831386e+21	0.582313	0.573343
4	MSE (Train)	3.163052e-02	0.039444	0.039278
5	MSE (Test)	2.044809e+09	0.036462	0.036180

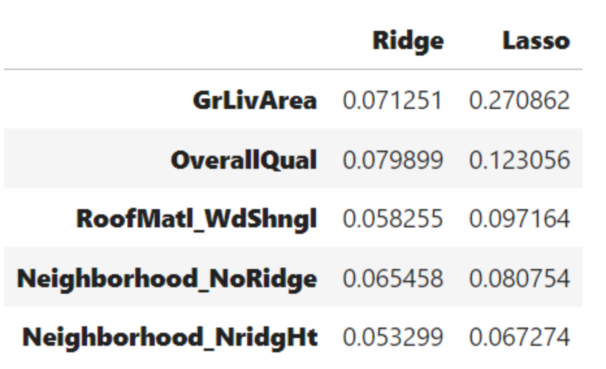
If the alpha values are doubled i.e. 8 for Ridge and 0.0002 for Lasso regression, the model metrics become



	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.868711	0.867163
1	R2 Score (Test)	0.872111	0.870717
2	RSS (Train)	1.716118	1.736352
3	RSS (Test)	0.597861	0.604381
4	MSE (Train)	0.040998	0.041239
5	MSE (Test)	0.036946	0.037147

It is observed that, after doubling the value the R2 score decreased for both Training and Testing dataset. While the RSS and MSE values increased.

The top 5 predictor variables after doubling the alpha values are



	Ridge	Lasso
GrLivArea	0.071251	0.270862
OverallQual	0.079899	0.123056
RoofMatl_WdShngl	0.058255	0.097164
Neighborhood_NoRidge	0.065458	0.080754
Neighborhood_NridgHt	0.053299	0.067274

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 :

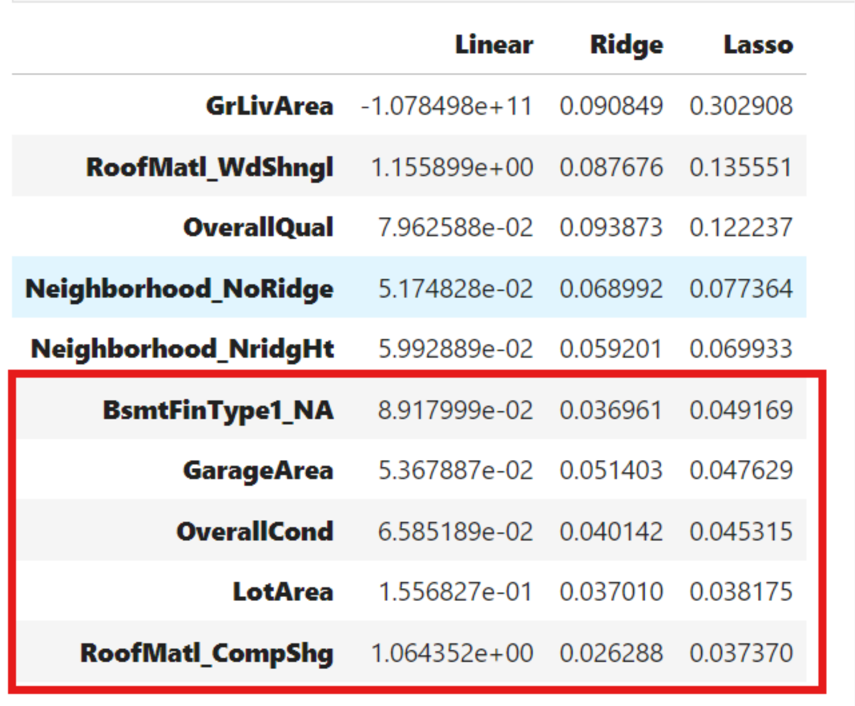
I will choose the Lasso regression as it helps in selecting variables by making the least important variables' co-efficient as 0.

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 top 10 predictor variables in Lasso regression are as follows



	Linear	Ridge	Lasso
GrLivArea	-1.078498e+11	0.090849	0.302908
RoofMatl_WdShngl	1.155899e+00	0.087676	0.135551
OverallQual	7.962588e-02	0.093873	0.122237
Neighborhood_NoRidge	5.174828e-02	0.068992	0.077364
Neighborhood_NridgHt	5.992889e-02	0.059201	0.069933
BsmtFinType1_NA	8.917999e-02	0.036961	0.049169
GarageArea	5.367887e-02	0.051403	0.047629
OverallCond	6.585189e-02	0.040142	0.045315
LotArea	1.556827e-01	0.037010	0.038175
RoofMatl_CompShg	1.064352e+00	0.026288	0.037370

After excluding the top 5, next five most important predictor variables are - BsmtFinType1_NA i.e. No Basement, GarageArea, OverallCond, LotArea, RoofMatl_CompShg i.e. Standard (Composite) Shingle.

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 :

A machine learning model will be robust and generalised when it can perform well on the unseen data. We can ensure the same by performing train-test split, regularization, cross validations.

The accuracy of the model will be impacted if performs well in the training data but not on the testing data.