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?

Ans:

Optimal value of alpha for Ridge regression =5

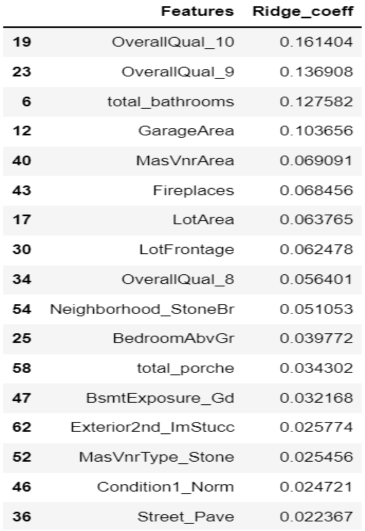
Optimal value of lasso for Ridge regression =0.0001

On doubling value of alpha , r2 score remains the same.

But the coefficients are changed , lowered a little on increasing alpha . This is because on increasing alpha , penalty is more due to which coefficients tend to 0.

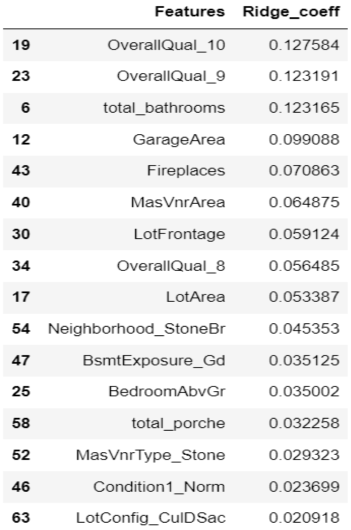
Coefficients before and after doubling alpha in ridge regression:

Before doubling alpha



	Features	Ridge_coeff
19	OverallQual_10	0.161404
23	OverallQual_9	0.136908
6	total_bathrooms	0.127582
12	GarageArea	0.103656
40	MasVnrArea	0.069091
43	Fireplaces	0.068456
17	LotArea	0.063765
30	LotFrontage	0.062478
34	OverallQual_8	0.056401
54	Neighborhood_StoneBr	0.051053
25	BedroomAbvGr	0.039772
58	total_porche	0.034302
47	BsmtExposure_Gd	0.032168
62	Exterior2nd_ImStucc	0.025774
52	MasVnrType_Stone	0.025456
46	Condition1_Norm	0.024721
36	Street_Pave	0.022367

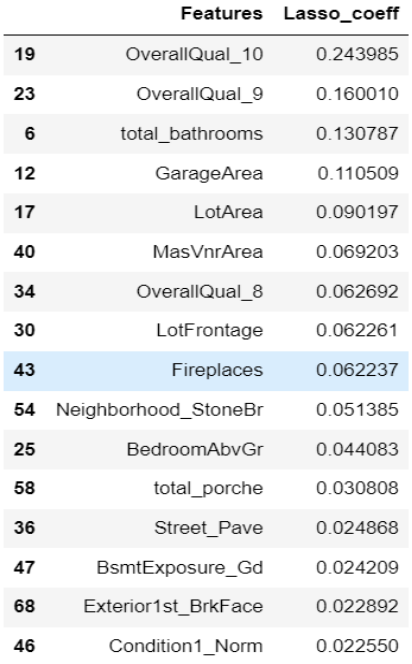
After doubling alpha



	Features	Ridge_coeff
19	OverallQual_10	0.127584
23	OverallQual_9	0.123191
6	total_bathrooms	0.123165
12	GarageArea	0.099088
43	Fireplaces	0.070863
40	MasVnrArea	0.064875
30	LotFrontage	0.059124
34	OverallQual_8	0.056485
17	LotArea	0.053387
54	Neighborhood_StoneBr	0.045353
47	BsmtExposure_Gd	0.035125
25	BedroomAbvGr	0.035002
58	total_porche	0.032258
52	MasVnrType_Stone	0.029323
46	Condition1_Norm	0.023699
63	LotConfig_CulDSac	0.020918

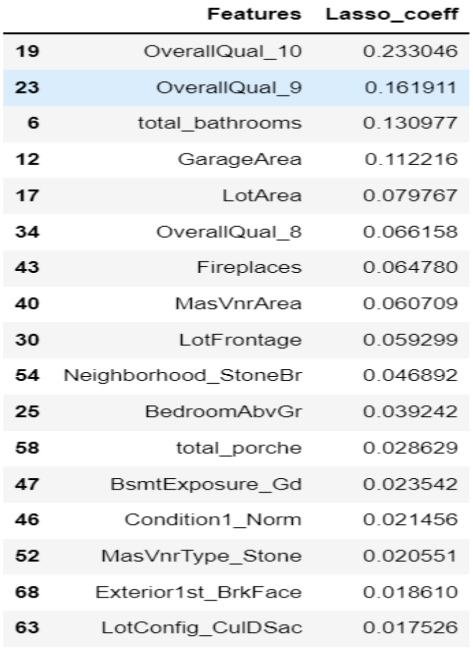
Coefficients before and after doubling alpha in lasso regression:

Before doubling alpha



	Features	Lasso_coeff
19	OverallQual_10	0.243985
23	OverallQual_9	0.160010
6	total_bathrooms	0.130787
12	GarageArea	0.110509
17	LotArea	0.090197
40	MasVnrArea	0.069203
34	OverallQual_8	0.062692
30	LotFrontage	0.062261
43	Fireplaces	0.062237
54	Neighborhood_StoneBr	0.051385
25	BedroomAbvGr	0.044083
58	total_porche	0.030808
36	Street_Pave	0.024868
47	BsmtExposure_Gd	0.024209
68	Exterior1st_BrkFace	0.022892
46	Condition1_Norm	0.022550

After doubling alpha



	Features	Lasso_coeff
19	OverallQual_10	0.233046
23	OverallQual_9	0.161911
6	total_bathrooms	0.130977
12	GarageArea	0.112216
17	LotArea	0.079767
34	OverallQual_8	0.066158
43	Fireplaces	0.064780
40	MasVnrArea	0.060709
30	LotFrontage	0.059299
54	Neighborhood_StoneBr	0.046892
25	BedroomAbvGr	0.039242
58	total_porche	0.028629
47	BsmtExposure_Gd	0.023542
46	Condition1_Norm	0.021456
52	MasVnrType_Stone	0.020551
68	Exterior1st_BrkFace	0.018610
63	LotConfig_CulDSac	0.017526

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?

Ans:

R2 value for both is same i.e 84% in train set and 79% on test set.

But Lasso regression helps in feature selection , i.e by pushing some coeffecients to exactly 0 value unlike ridge regression. Therefore I will apply lasso regression.

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?

Top 5 predictors in current lasso model:

1. OverallQual_9
2. total_bathrooms
3. OverallQual_10
4. GarageArea
5. LotArea

After excluding these predictors, The r2 value dropped to 69%. 5 most predictors now are:

1. MasVnrArea
2. LotFrontage
3. Fireplaces
4. Neighborhood_StoneBr
5. BedroomAbvGr

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?

Ans:

To make the model robust and generalizable, we need to make the model simple.

1. Simpler models are more generic and perform better for unseen data.
2. Simpler models have low variance. The model coefficients don’t change much when there is change in training data.
3. Simpler models have high training errors, but complex models tend to overfit.
4. We use regularization to make models simple. Regularisation will add penalty term to cost function whichg will tend the coefficients towards 0.

   In terms of accuracy a robust and generalized model means the accuracy will be the same on train and test sets. There won’t be much deviation which means model will not overfit.