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 and lasso are as mentioned bellow:

- Ridge: 0.2
- Lasso: 0.001

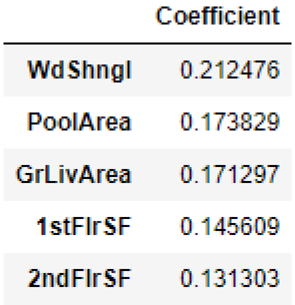
Changes in model when alpha is doubled:

Ridge: In the scenario of Ridge when the model was trained with two times the optimal alpha below changes were observed.

1. Change is coefficient

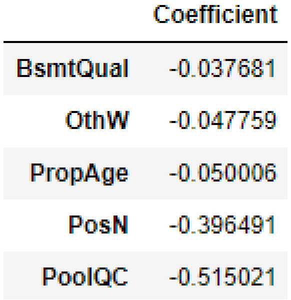
Optimal alpha

Top 5 +ve



	Coefficient
WdShngl	0.212476
PoolArea	0.173829
GrLivArea	0.171297
1stFlrSF	0.145609
2ndFlrSF	0.131303

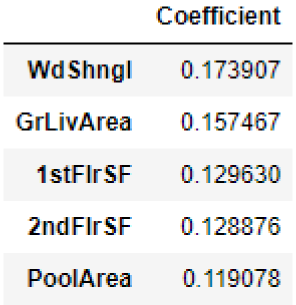
Top 5 -ve



	Coefficient
BsmtQual	-0.037681
OthW	-0.047759
PropAge	-0.050006
PosN	-0.396491
PoolQC	-0.515021

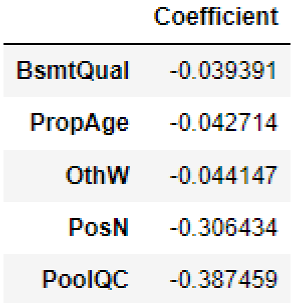
Double of Optimal Alpha

Top 5 +ve



	Coefficient
WdShngl	0.173907
GrLivArea	0.157467
1stFlrSF	0.129630
2ndFlrSF	0.128876
PoolArea	0.119078

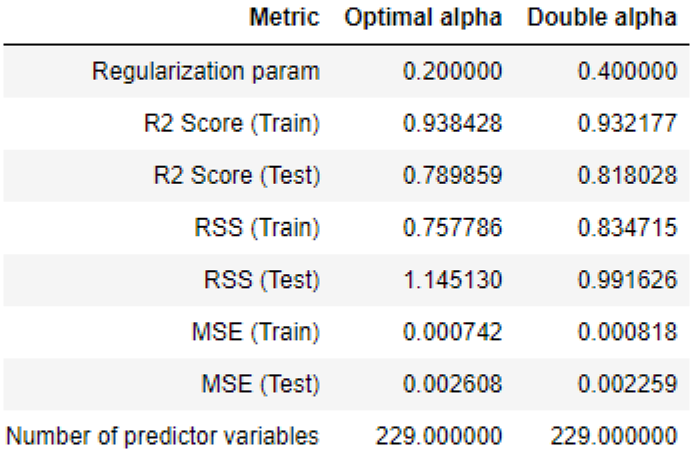
Top 5 -ve



	Coefficient
BsmtQual	-0.039391
PropAge	-0.042714
OthW	-0.044147
PosN	-0.306434
PoolQC	-0.387459

The value of coefficients has decreased, indicating that the alpha has increased the penalty due to which the coefficient value decreased. This also will make the model more generalized.

2. Change in other measures



Metric	Optimal alpha	Double alpha
Regularization param	0.200000	0.400000
R2 Score (Train)	0.938428	0.932177
R2 Score (Test)	0.789859	0.818028
RSS (Train)	0.757786	0.834715
RSS (Test)	1.145130	0.991626
MSE (Train)	0.000742	0.000818
MSE (Test)	0.002608	0.002259
Number of predictor variables	229.000000	229.000000

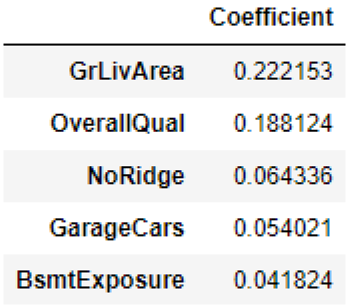
In this particular scenario all the metrics have improved.

Lasso: In the scenario of Lasso when the model was trained with two times the optimal alpha below changes were observed.

1. Change is coefficient

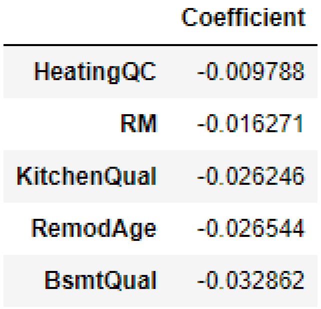
Optimal alpha

Top 5 +ve



	Coefficient
GrLivArea	0.222153
OverallQual	0.188124
NoRidge	0.064336
GarageCars	0.054021
BsmtExposure	0.041824

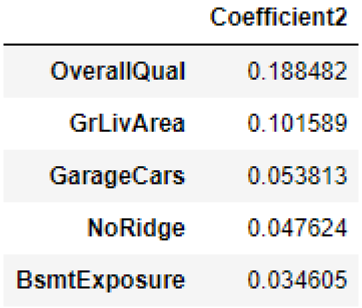
Top 5 -ve



	Coefficient
HeatingQC	-0.009788
RM	-0.016271
KitchenQual	-0.026246
RemodAge	-0.026544
BsmtQual	-0.032862

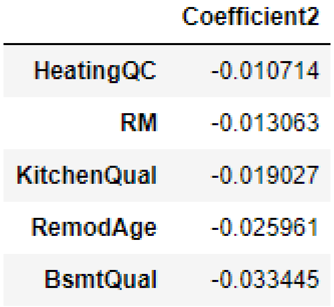
Double of Optimal Alpha

Top 5 +ve



	Coefficient2
OverallQual	0.188482
GrLivArea	0.101589
GarageCars	0.053813
NoRidge	0.047624
BsmtExposure	0.034605

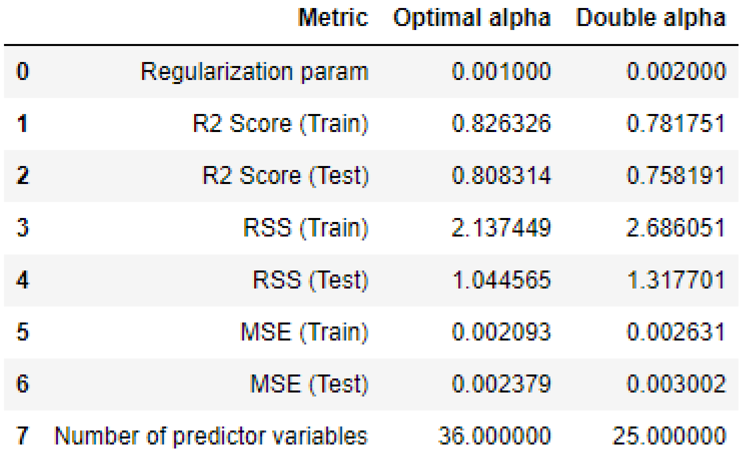
Top 5 -ve



	Coefficient2
HeatingQC	-0.010714
RM	-0.013063
KitchenQual	-0.019027
RemodAge	-0.025961
BsmtQual	-0.033445

The value of coefficients has decreased, indicating that the alpha has increased the penalty due to which the coefficient value decreased. This also will make the model more generalized.

1. Change in other measures



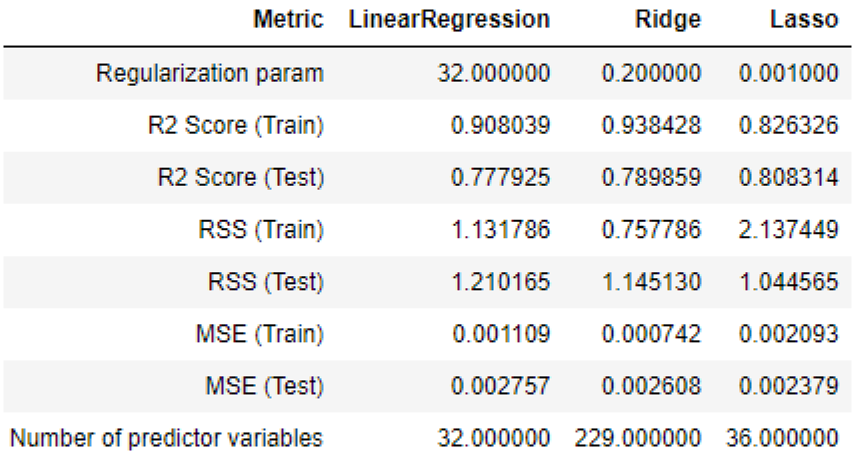
	Metric	Optimal alpha	Double alpha
0	Regularization param	0.001000	0.002000
1	R2 Score (Train)	0.826326	0.781751
2	R2 Score (Test)	0.808314	0.758191
3	RSS (Train)	2.137449	2.686051
4	RSS (Test)	1.044565	1.317701
5	MSE (Train)	0.002093	0.002631
6	MSE (Test)	0.002379	0.003002
7	Number of predictor variables	36.000000	25.000000

In the second scenario due to high value of alpha phew more variables got filtered but this also resulting the underfitting of the model which is indicated by lower value of R2 score, and higher value of RSS and MSE.

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: Below id the statistic used for finalizing the model



Metric	LinearRegression	Ridge	Lasso
Regularization param	32.000000	0.200000	0.001000
R2 Score (Train)	0.908039	0.938428	0.826326
R2 Score (Test)	0.777925	0.789859	0.808314
RSS (Train)	1.131786	0.757786	2.137449
RSS (Test)	1.210165	1.145130	1.044565
MSE (Train)	0.001109	0.000742	0.002093
MSE (Test)	0.002757	0.002608	0.002379
Number of predictor variables	32.000000	229.000000	36.000000

The R2 score for Lasso regression is consistent across train and test data and the test R2 score is better. In the similar way the test RSS and MSE are also better for Lasso model. From the point of view of simplifying the model, Lasso uses only 36 fields where as Ridge required more number of fields.

Due to above mentioned reasons Lasso will be chosen for implementation

Question-3

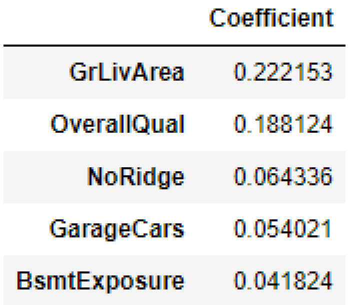
After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Answer:
Since the top 5 variables were missing from incoming data we trained a new model by excluding the 5 variables. Bellow changes were observed.

2. Change is coefficient

Original

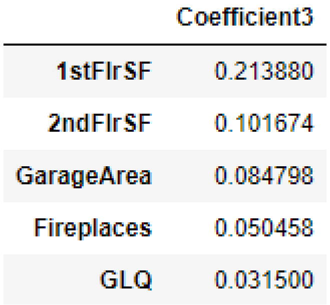
Top 5 +ve



	Coefficient
GrLivArea	0.222153
OverallQual	0.188124
NoRidge	0.064336
GarageCars	0.054021
BsmtExposure	0.041824

After removing top 5

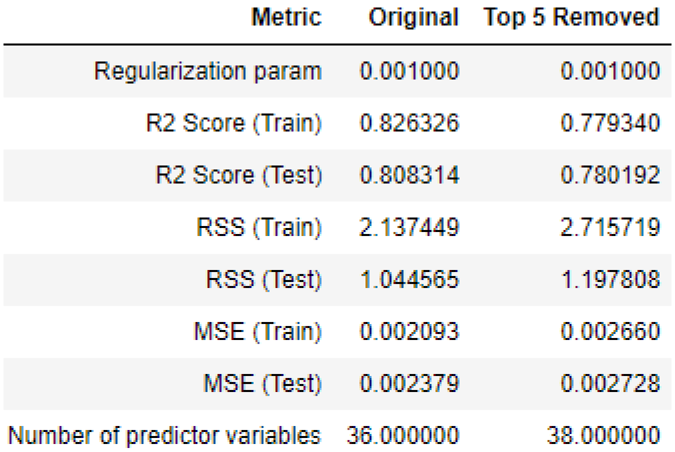
Top 5 +ve



	Coefficient3
1stFlrSF	0.213880
2ndFlrSF	0.101674
GarageArea	0.084798
Fireplaces	0.050458
GLQ	0.031500

New top 5 predictors.

2. Change in other measures



Metric	Original	Top 5 Removed
Regularization param	0.001000	0.001000
R2 Score (Train)	0.826326	0.779340
R2 Score (Test)	0.808314	0.780192
RSS (Train)	2.137449	2.715719
RSS (Test)	1.044565	1.197808
MSE (Train)	0.002093	0.002660
MSE (Test)	0.002379	0.002728
Number of predictor variables	36.000000	38.000000

The accuracy of the model decreased substantially, and the complexity also increased.

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: Below parameters can be considered for generalizing the model

1. Number of predictor variables required. More the number of variables more complex
2. The accuracy in case of train and test data should be similar, the model should neither underfit not overfit
    a. Underfitting: low R2 for Test and Train
    b. Overfitting: High training R2 low Test R2
    c. Good Model: High training and Test R2