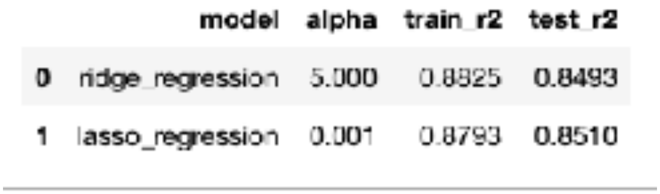
Advanced Regression Assignment

Problem Statement Part-II

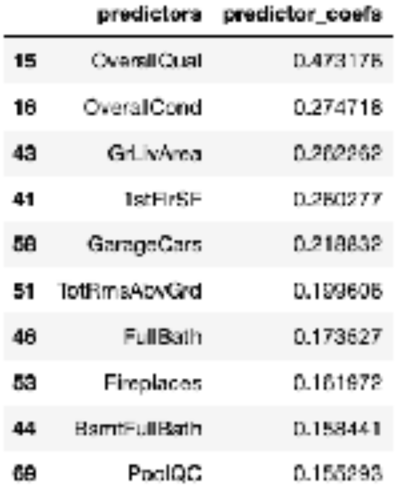
Question 1
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:
The model parameters in terms of alpha and training and test r2 scores are articulated in table below



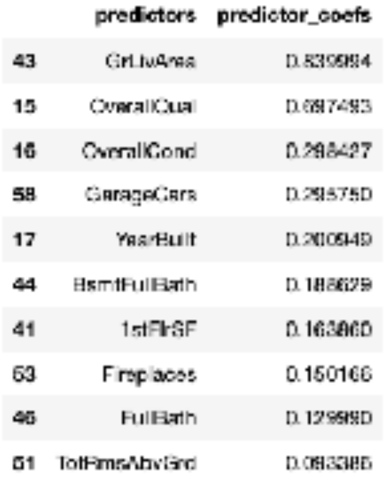
	model	alpha	train_r2	test_r2
0	ridge_regression	5.000	0.8825	0.8493
1	lasso_regression	0.001	0.8793	0.8510

Ridge Regression Important Features alpha=5.0.



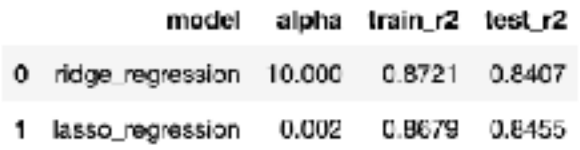
	predictors	predictor_coefs
15	OverallQual	0.473176
16	OverallCond	0.274716
43	GrLivArea	0.262262
41	1stFlrSF	0.260277
58	GarageCars	0.218832
51	TotRmsAbvGrd	0.199606
46	FullBath	0.173627
53	Fireplaces	0.161672
44	BsmtFullBath	0.158441
69	PoolQC	0.155293

Lasso Regression Important Features alpha = 0.01



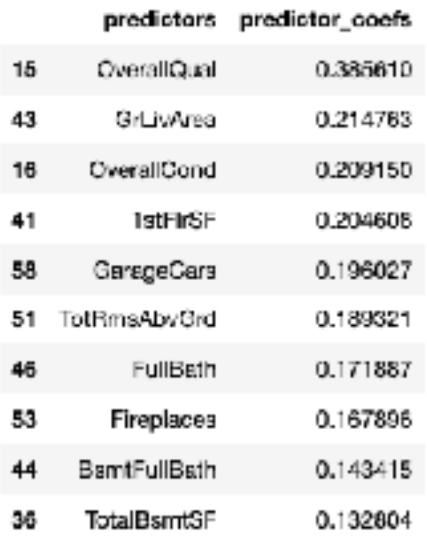
	predictors	predictor_coefs
43	GrLivArea	0.839984
15	OverallQual	0.697493
16	OverallCond	0.298427
58	GarageCars	0.295750
17	YearBuilt	0.200949
44	BsmtFullBath	0.188629
41	1stFlrSF	0.163860
53	Fireplaces	0.150166
46	FullBath	0.129990
51	TotRmsAbvGrd	0.093385

Model Params with double alpha value



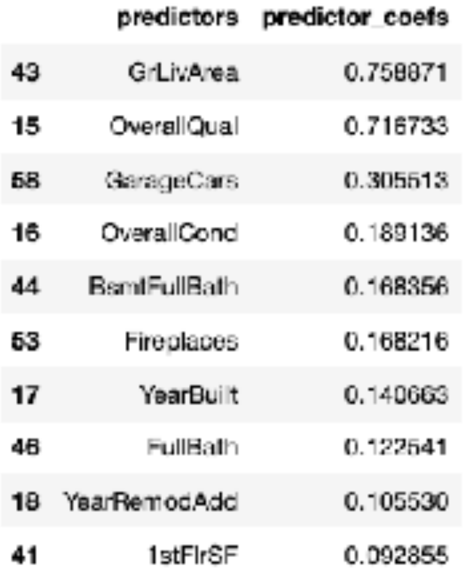
	model	alpha	train_r2	test_r2
0	ridge_regression	10.000	0.8721	0.8407
1	lasso_regression	0.002	0.8679	0.8455

Ridge Regression Important Features alpha = 10.0



	predictors	predictor_coefs
15	OverallQual	0.385610
43	GrLivArea	0.214763
16	OverallCond	0.209150
41	1stFlrSF	0.204606
58	GarageCars	0.196027
51	TotRmsAbvGrd	0.189321
46	FullBath	0.171887
53	Fireplaces	0.167896
44	BsmtFullBath	0.143415
36	TotalBsmtSF	0.132804

Lasso Regression important features alpha=0.02



	predictors	predictor_coefs
43	GrLivArea	0.758871
15	OverallQual	0.716733
58	GarageCars	0.305513
16	OverallCond	0.189136
44	BsmtFullBath	0.168356
53	Fireplaces	0.168216
17	YearBuilt	0.140663
46	FullBath	0.122541
18	YearRemodAdd	0.105530
41	1stFlrSF	0.092855

Conclusions
- with double alpha we see the r2 scores changing (decreasing) , and the model being able to explain less variation in data (in terms of r2 score=

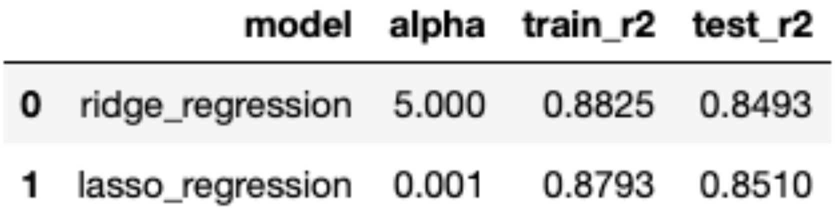
- The important features with alpha and double the alpha are articulated in the tables above
  - In general we see different features being projected as prominent ones

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:

- The choice of Lasso or Ridge regression depends on wether we want all the predictor parameters in the final model (Ridge Regression) or we want Lasso model which would help us in forcing some coefficients to 0 and thus helping with feature/predictor elimination.

- In our case if we take r2 scores during training and testing as a yardstick

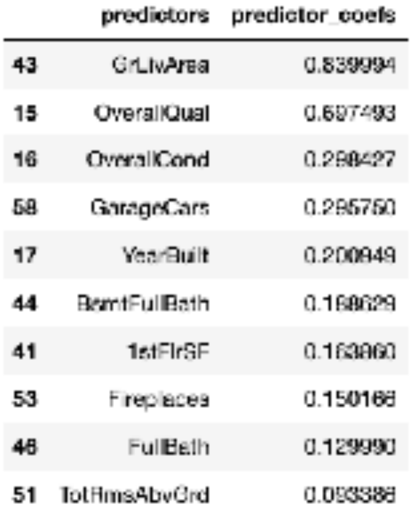


	model	alpha	train_r2	test_r2
0	ridge_regression	5.000	0.8825	0.8493
1	lasso_regression	0.001	0.8793	0.8510

We would go with lasso as the model can better explain variance in given data for both training and testing (With the training and testing dataset fed being same to both ridge and lasso models).

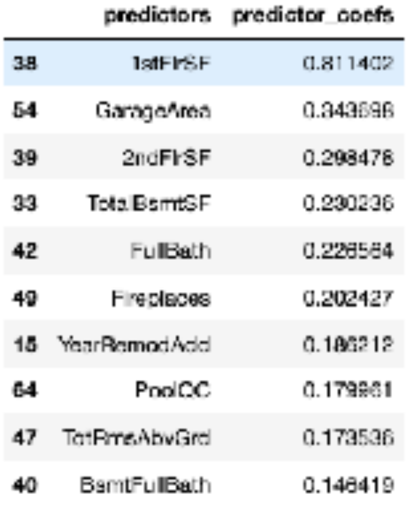
Question 3
After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Answer:
With alpha = 0.001 the important predictor variables in lasso are



	predictors	predictor_coefs
43	GrLivArea	0.839994
15	OverallQual	0.697493
16	OverallCond	0.298427
58	GarageCars	0.295760
17	YearBuilt	0.200849
44	BsmtFullBath	0.168629
41	1stFlrSF	0.163860
53	Fireplaces	0.150168
46	FullBath	0.129990
51	TotRmsAbvGrd	0.093386

Now with top 5 features namely ["GrLivArea", "OverallQual", "OverallCond", "GarageCars", “YearBuilt"] removed the important features predicted by lasso with alpha = 0.001 are



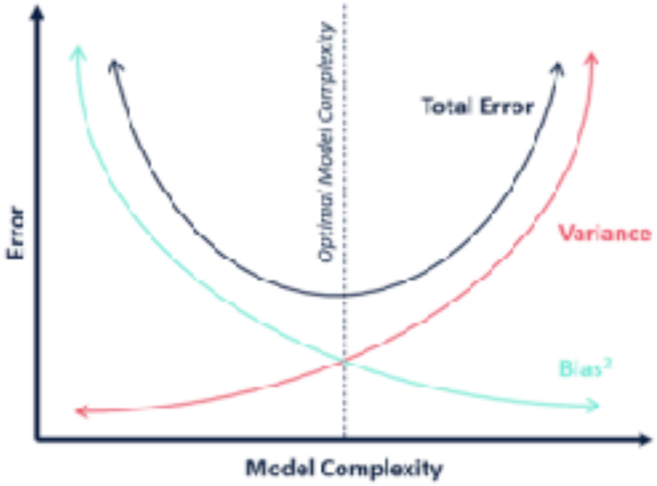
	predictors	predictor_coefs
38	1stFlrSF	0.811402
54	GarageArea	0.343698
39	2ndFlrSF	0.298476
33	TotalBsmtSF	0.230236
42	FullBath	0.228564
49	Fireplaces	0.202427
15	YearRemodAdd	0.186212
64	PoolQC	0.179861
47	TotRmsAbvGrd	0.173536
40	BsmtFullBath	0.148419

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:

This can be explained with the help of Model Bias vs Variance tradeoff graph

The below figure is taken from the internet.



To make sure the model is robust and generalisable the idea is to give up a bit of bias (increase bias) in tradeoff to decrease variance.

In the above graph if the choice of hyper-parameter makes the model go to extreme right i.e low bias and high variance we would have a highly complex model which generalises well on training data but fails on testing data(overfitting). This leads to high accuracy during training and lower during testing

While if hyper-parameter choice gives a model to extreme left we have high variance and low bias. In this case the model may perform better on unseen testing data but fails to generalise on training data (underfitting)

The optimal model would be one in which the hyper parameter tuning results in a model in the intersection of bias and variance curve such a model will be able to learn patterns on training data and also generalise well on testing data thereby producing a robust model with good training and testing accuracies.