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?

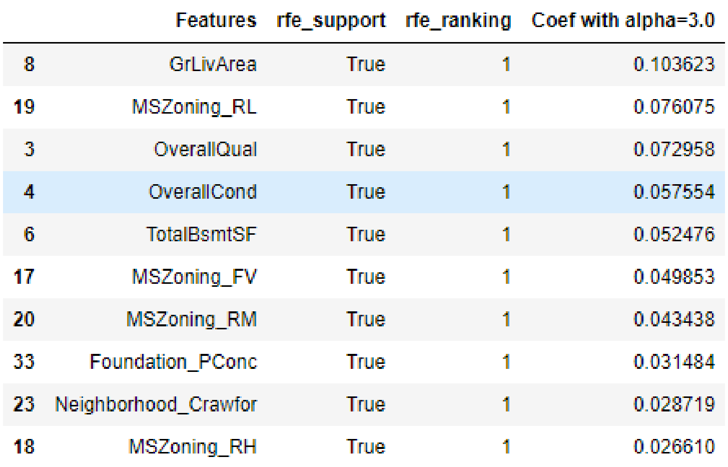
Answer :

The optimal value of alpha for Ridge and Lasso Regression are 1 and 0.0001

In both the Regression model the effects of doubling the value of alpha is that the value of coefficients has moved towards zero for both the coefficients i.e. positive and negative. Some predictor variables have also changed the order of importance . But the first 10 variables have not shown much changes in both the regression models (Ridge and Lasso)

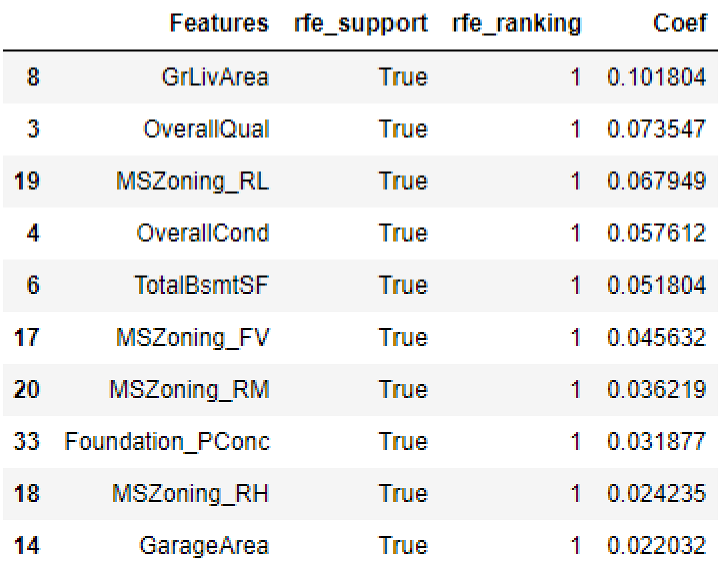
Ridge Regression:

When Alpha= 3.0



	Features	rfe_support	rfe_ranking	Coef with alpha=3.0
8	GrLivArea	True	1	0.103623
19	MSZoning_RL	True	1	0.076075
3	OverallQual	True	1	0.072958
4	OverallCond	True	1	0.057554
6	TotalBsmtSF	True	1	0.052476
17	MSZoning_FV	True	1	0.049853
20	MSZoning_RM	True	1	0.043438
33	Foundation_PConc	True	1	0.031484
23	Neighborhood_Crawfor	True	1	0.028719
18	MSZoning_RH	True	1	0.026610

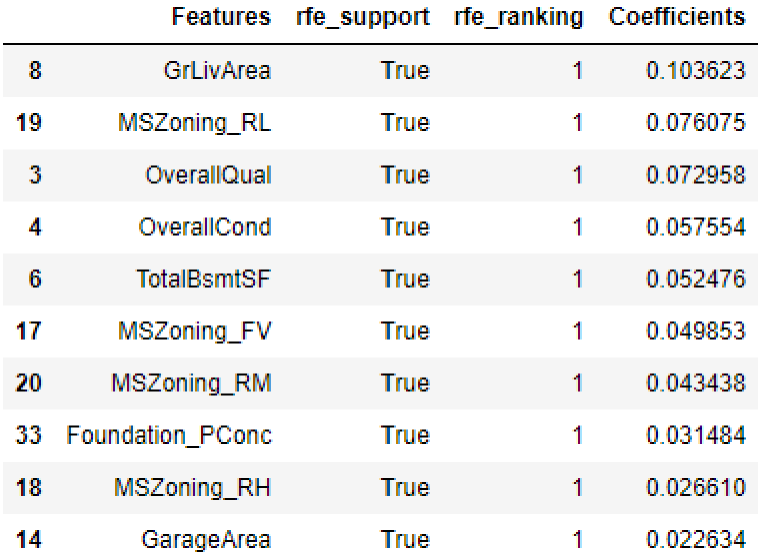
When Alpha=6.0



	Features	rfe_support	rfe_ranking	Coef
8	GrLivArea	True	1	0.101804
3	OverallQual	True	1	0.073547
19	MSZoning_RL	True	1	0.067949
4	OverallCond	True	1	0.057612
6	TotalBsmtSF	True	1	0.051804
17	MSZoning_FV	True	1	0.045632
20	MSZoning_RM	True	1	0.036219
33	Foundation_PConc	True	1	0.031877
18	MSZoning_RH	True	1	0.024235
14	GarageArea	True	1	0.022032

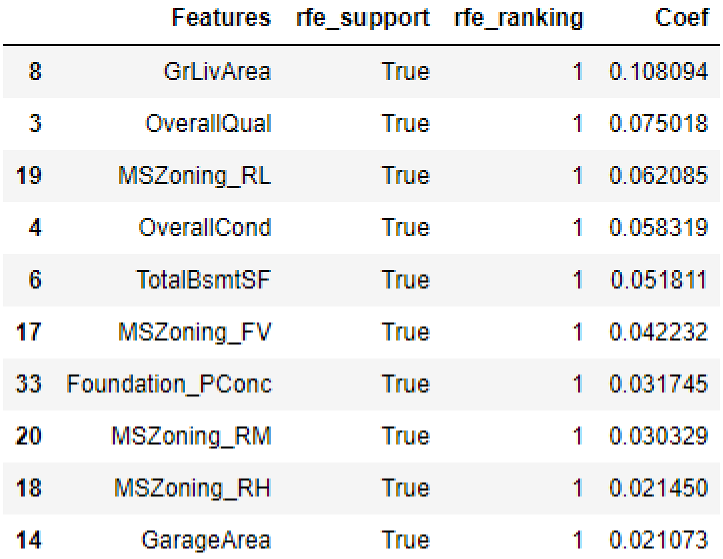
Lasso Regression:

When Alpha =0.0002



	Features	rfe_support	rfe_ranking	Coefficients
8	GrLivArea	True	1	0.103623
19	MSZoning_RL	True	1	0.076075
3	OverallQual	True	1	0.072958
4	OverallCond	True	1	0.057554
6	TotalBsmtSF	True	1	0.052476
17	MSZoning_FV	True	1	0.049853
20	MSZoning_RM	True	1	0.043438
33	Foundation_PConc	True	1	0.031484
18	MSZoning_RH	True	1	0.026610
14	GarageArea	True	1	0.022634

When Alpha = 0.0004



	Features	rfe_support	rfe_ranking	Coef
8	GrLivArea	True	1	0.108094
3	OverallQual	True	1	0.075018
19	MSZoning_RL	True	1	0.062085
4	OverallCond	True	1	0.058319
6	TotalBsmtSF	True	1	0.051811
17	MSZoning_FV	True	1	0.042232
33	Foundation_PConc	True	1	0.031745
20	MSZoning_RM	True	1	0.030329
18	MSZoning_RH	True	1	0.021450
14	GarageArea	True	1	0.021073

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?

Ridge Regression :

```
Alpha = 3.0

Mean Squared Error = 0.01575

Training Accuracy = 93.72%

Test Accuracy = 90.94%
```

Lasso Regression :

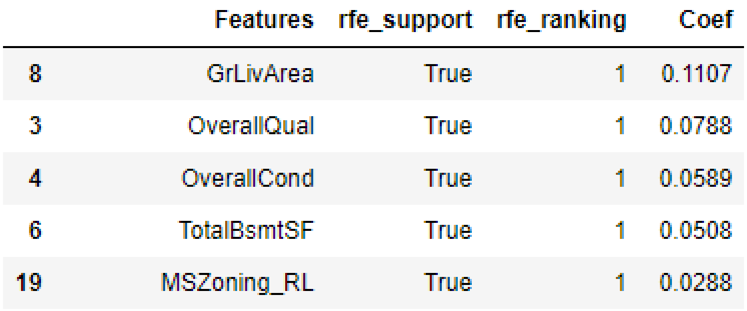
```
Alpha = 0.0002

Mean Squared Error = 0.01564

Training Accuracy = 93.04%

Test Accuracy = 89.22%
```

We can clearly observe from the above values that Mean Squared Error of Lasso is less than the Ridge and Lasso also helps in feature reduction i.e. the coefficient value of one or more the feature can become zero thus , Lasso has edge over the Ridge Regression . So we clearly choose Lasso over Ridge.

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?

After excluding the top five most important predictor variables namely 'MSZoning_RL', 'GrLivArea', 'OverallQual', 'OverallCond' , 'MSZoning_FV' we get alpha = 0.001 . Thus , the top five predictor models after dropping these variables with their respective coefficient values are :



	Features	rfe_support	rfe_ranking	Coef
8	GrLivArea	True	1	0.1107
3	OverallQual	True	1	0.0788
4	OverallCond	True	1	0.0589
6	TotalBsmtSF	True	1	0.0508
19	MSZoning_RL	True	1	0.0288

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?

A model is called generalizable if it doesn't memorize the training data too much. This means when we test the model with new data, it should either give the expected accuracy or make some acceptable errors. A model is considered robust if it gives consistent results even when some values change. Simple models are usually both robust and generalizable. So, it's better to keep the model simple, but not too simple, because that can cause problems.

To make a model generalizable, we need to find a balance between overfitting, underfitting, and accuracy. Regularization is one way to do this—it penalizes large coefficients compared to others. The model should be just complex enough to understand the patterns but not too complex. This means we have to compromise on how complex the model is to improve the accuracy of the test data. The best situation is when the model has enough bias to understand things broadly and enough variance to make few errors.