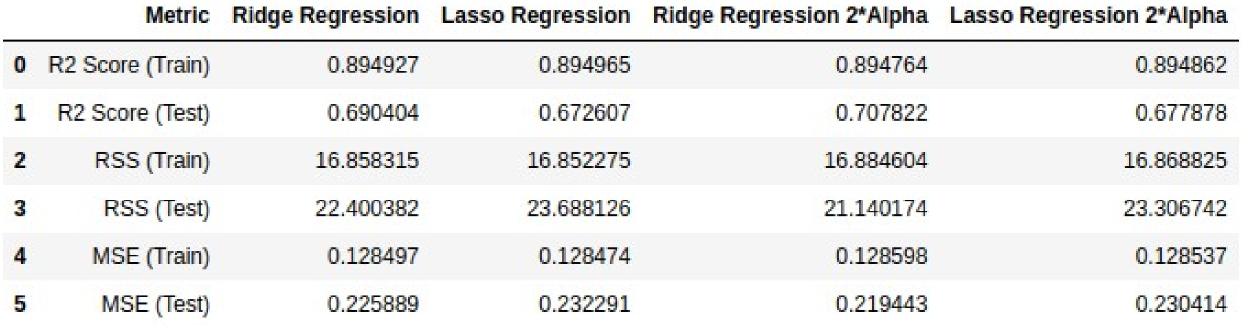
Question 1

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

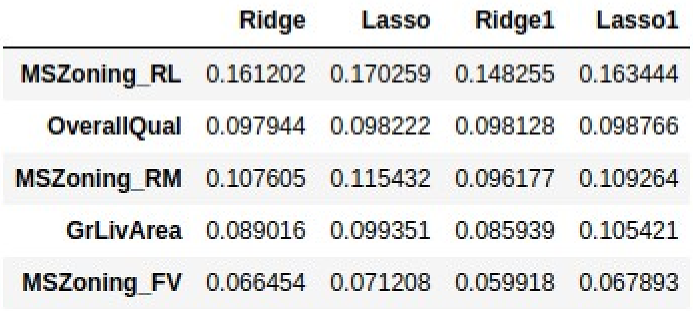
Optimal values of alpha are as follows:
For Ridge:2
For Lasso:0.001
Below is the comparison of Ridge and lasso before and after doubling alpha value:



	Metric	Ridge Regression	Lasso Regression	Ridge Regression 2*Alpha	Lasso Regression 2*Alpha
0	R2 Score (Train)	0.894927	0.894965	0.894764	0.894862
1	R2 Score (Test)	0.690404	0.672607	0.707822	0.677878
2	RSS (Train)	16.858315	16.852275	16.884604	16.868825
3	RSS (Test)	22.400382	23.688126	21.140174	23.306742
4	MSE (Train)	0.128497	0.128474	0.128598	0.128537
5	MSE (Test)	0.225889	0.232291	0.219443	0.230414

From above table it was clear that after doubling value of alpha R2 score (Test as well as train) of both models (Ridge and lasso) goes up. Which tells us that performance of model goes up which was also implied by RSS and MSE values as both these values goes down which is a good indication for a model.
Below are the top 5 features after doubling the alpha value:

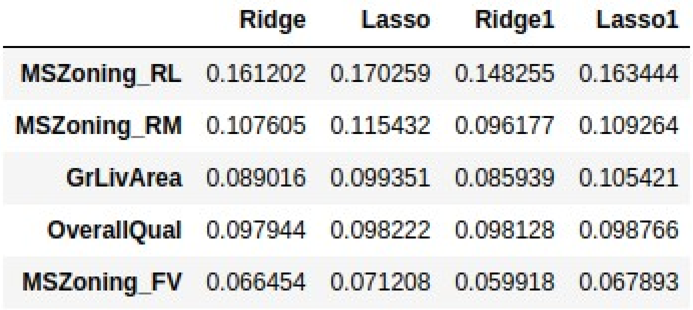
Ridge:



	Ridge	Lasso	Ridge1	Lasso1
MSZoning_RL	0.161202	0.170259	0.148255	0.163444
OverallQual	0.097944	0.098222	0.098128	0.098766
MSZoning_RM	0.107605	0.115432	0.096177	0.109264
GrLivArea	0.089016	0.099351	0.085939	0.105421
MSZoning_FV	0.066454	0.071208	0.059918	0.067893

1. MSZoning_RL : Zoning classification of sale - Residential Low Density
2. OverallQual: Rates the overall material and finish of the house
3. MSZoning_RM : Zoning classification of sale - Residential Medium Density
4. GrLivArea: Above grade (ground) living area square feet
5. MSZoning_FV : Zoning classification of sale - Floating Village Residential

Lasso:



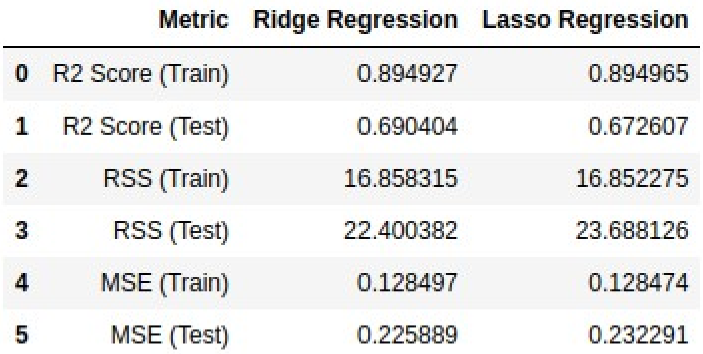
	Ridge	Lasso	Ridge1	Lasso1
MSZoning_RL	0.161202	0.170259	0.148255	0.163444
MSZoning_RM	0.107605	0.115432	0.096177	0.109264
GrLivArea	0.089016	0.099351	0.085939	0.105421
OverallQual	0.097944	0.098222	0.098128	0.098766
MSZoning_FV	0.066454	0.071208	0.059918	0.067893

1. MSZoning_RL : Zoning classification of sale - Residential Low Density
2. MSZoning_RM : Zoning classification of sale - Residential Medium Density
3. GrLivArea: Above grade (ground) living area square feet
4. OverallQual: Rates the overall material and finish of the house
5. MSZoning_FV : Zoning classification of sale - Floating Village Residential

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



	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.894927	0.894965
1	R2 Score (Test)	0.690404	0.672607
2	RSS (Train)	16.858315	16.852275
3	RSS (Test)	22.400382	23.688126
4	MSE (Train)	0.128497	0.128474
5	MSE (Test)	0.225889	0.232291

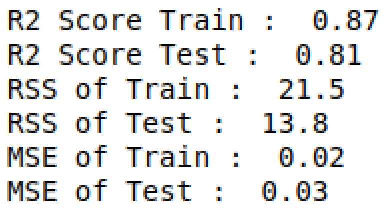
Based on above metrics we will choose Ridge regression as it will be giving better results on test data as compared to Lasso regression model.
R2 score of Ridge regression is better than Lasso regression and in case of Ridge we have lower MSE and lower RSS on Train and Test data.
We can not compare both models with dependent variables as dependent variables count is same for the 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?

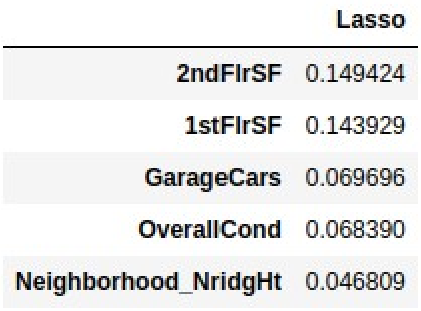
Answer

After removing top 5 features from our model we get below R2 score and RSS, MSE values, which are better from actual model.



```
R2 Score Train :  0.87
R2 Score Test :  0.81
RSS of Train :  21.5
RSS of Test :  13.8
MSE of Train :  0.02
MSE of Test :  0.03
```

After building model we get below 5 features as most important features:



	Lasso
2ndFlrSF	0.149424
1stFlrSF	0.143929
GarageCars	0.069696
OverallCond	0.068390
Neighborhood_NridgHt	0.046809

2ndFlrSF (Second floor square feet) is the most important feature as it has biggest coefficient value. After 2ndFlrSF, 1stFlrSF (First floor square feet) is the most important one. Followed by GarageCars (Size of garage in car capacity), OverallCond (Rates the overall condition of the house) and Neighborhood_NridgHt (Physical locations within Ames city limits- Northridge Heights).

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 4

For making a robust model we should take care below mentioned points:

1. Our model does not contain Outliers.
2. Using robust evaluation metric
3. Transform data to reduce impact of outliers
4. Over fitting on training data must be avoided

All of these will have very much impact on accuracy of model:

1. If our data have outliers then our model tries to fit them, in fitting them it will not bale to predict normal data hence accuracy will be low.
2. Robust evaluation metrics will help in building model more accurate.
3. Transforming data will help in reducing impact of outliers, which will help in building better model.
4. As our model tries to overfit train data or memorising training data will increase chance of model failure on unseen data very much hence accuracy will be impacted very much.