Question 1

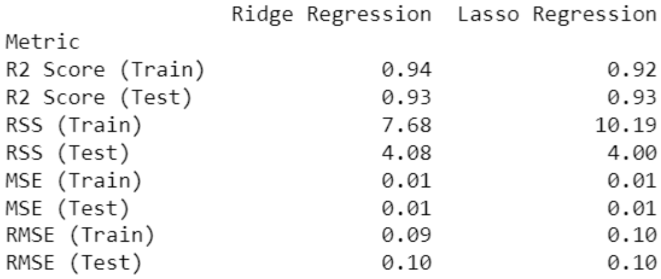
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer

- The optimal value of alpha for Ridge Regression is **10**
- the optimal value of alpha for Lasso Regression is **0.001**

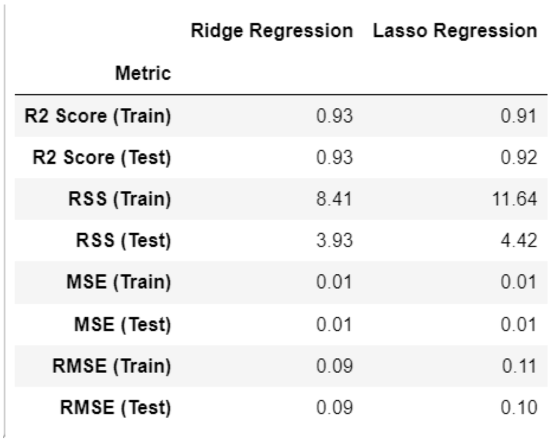
When we doubled the value of alpha for ridge and lasso regression models, we observed changes in the model metrics.

Earlier the model metrics were as follows

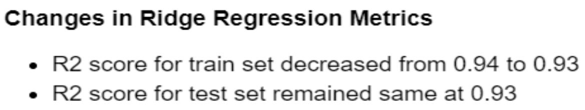


```
                  Ridge Regression  Lasso Regression
Metric
R2 Score (Train)              0.94              0.92
R2 Score (Test)               0.93              0.93
RSS (Train)                   7.68             10.19
RSS (Test)                    4.08              4.00
MSE (Train)                   0.01              0.01
MSE (Test)                    0.01              0.01
RMSE (Train)                  0.09              0.10
RMSE (Test)                   0.10              0.10
```

After doubling the value of alpha, the model metrics were as follows

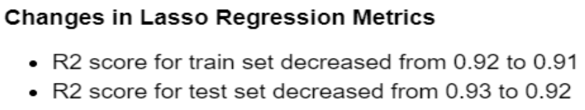


Metric	Ridge Regression	Lasso Regression
R2 Score (Train)	0.93	0.91
R2 Score (Test)	0.93	0.92
RSS (Train)	8.41	11.64
RSS (Test)	3.93	4.42
MSE (Train)	0.01	0.01
MSE (Test)	0.01	0.01
RMSE (Train)	0.09	0.11
RMSE (Test)	0.09	0.10



Changes in Ridge Regression Metrics

- R2 score for train set decreased from 0.94 to 0.93
- R2 score for test set remained same at 0.93

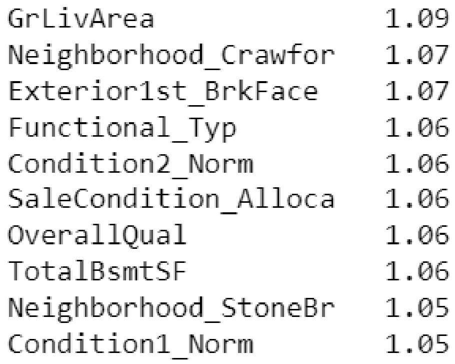


Changes in Lasso Regression Metrics

- R2 score for train set decreased from 0.92 to 0.91
- R2 score for test set decreased from 0.93 to 0.92

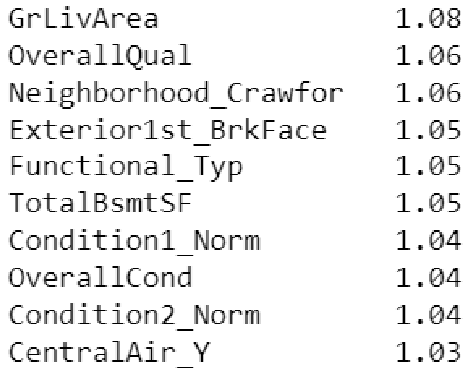
We also observed changes in predictor variables along with their coefficient values.

OLD PREDICTOR VARIABLES (RIDGE)



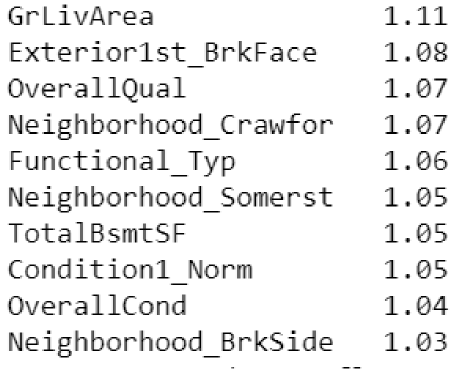
```
GrLivArea               1.09
Neighborhood_Crawfor    1.07
Exterior1st_BrkFace     1.07
Functional_Typ          1.06
Condition2_Norm         1.06
SaleCondition_Alloca    1.06
OverallQual             1.06
TotalBsmtSF             1.06
Neighborhood_StoneBr    1.05
Condition1_Norm         1.05
```

NEW PREDICTOR VARIABLES (RIDGE) WITH 2*ALPHA



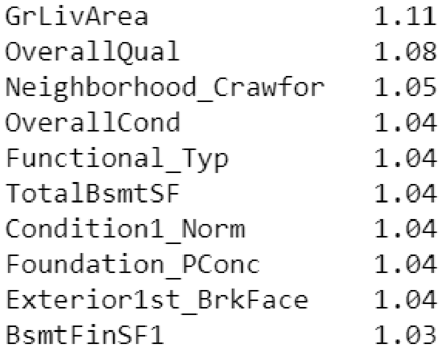
```
GrLivArea               1.08
OverallQual             1.06
Neighborhood_Crawfor    1.06
Exterior1st_BrkFace     1.05
Functional_Typ          1.05
TotalBsmtSF             1.05
Condition1_Norm         1.04
OverallCond             1.04
Condition2_Norm         1.04
CentralAir_Y            1.03
```

OLD PREDICTOR VARIABLES (LASSO)



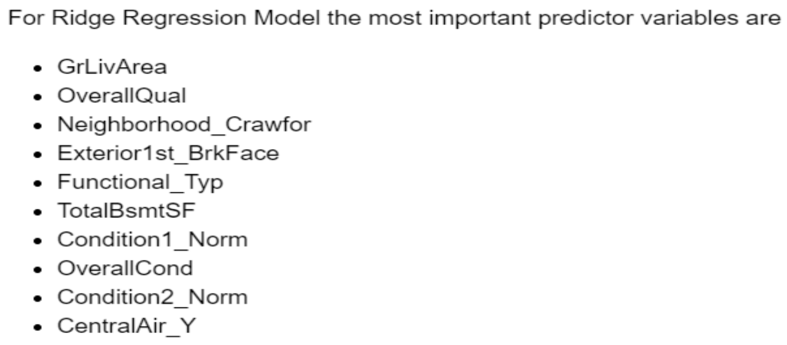
```
GrLivArea               1.11
Exterior1st_BrkFace     1.08
OverallQual             1.07
Neighborhood_Crawfor    1.07
Functional_Typ          1.06
Neighborhood_Somerst    1.05
TotalBsmtSF             1.05
Condition1_Norm         1.05
OverallCond             1.04
Neighborhood_BrkSide    1.03
```

NEW PREDICTOR VARIABLES (LASSO) WITH 2*ALPHA



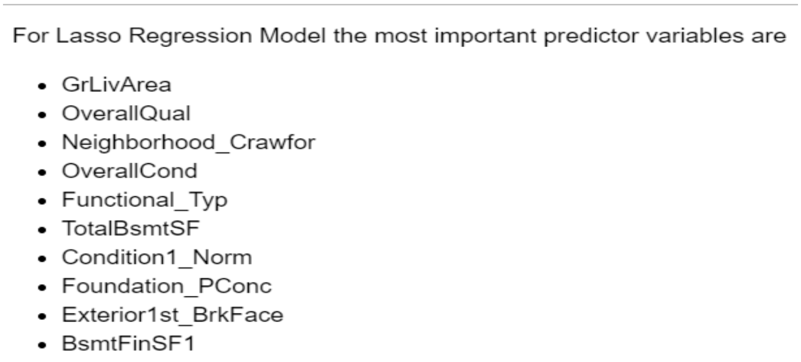
```
GrLivArea               1.11
OverallQual             1.08
Neighborhood_Crawfor    1.05
OverallCond             1.04
Functional_Typ          1.04
TotalBsmtSF             1.04
Condition1_Norm         1.04
Foundation_PConc        1.04
Exterior1st_BrkFace     1.04
BsmtFinSF1              1.03
```

Hence after the changes are implemented (alpha doubled)



For Ridge Regression Model the most important predictor variables are

- GrLivArea
- OverallQual
- Neighborhood_Crawfor
- Exterior1st_BrkFace
- Functional_Typ
- TotalBsmtSF
- Condition1_Norm
- OverallCond
- Condition2_Norm
- CentralAir_Y



For Lasso Regression Model the most important predictor variables are

- GrLivArea
- OverallQual
- Neighborhood_Crawfor
- OverallCond
- Functional_Typ
- TotalBsmtSF
- Condition1_Norm
- Foundation_PConc
- Exterior1st_BrkFace
- BsmtFinSF1

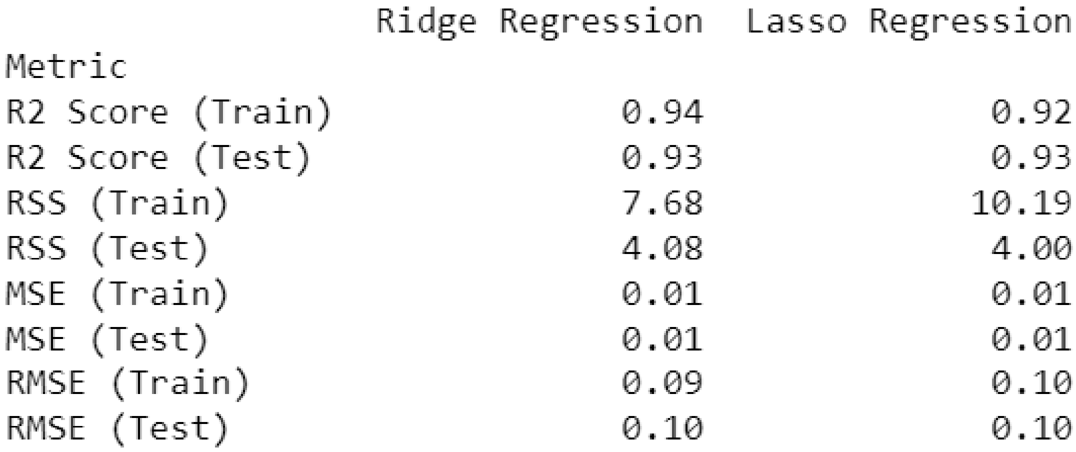
Question 2.

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 determined optimal value of alpha for Ridge Regression is **10**
- the determined optimal value of alpha for Lasso Regression is **0.001**

When we compared the model metrics of Ridge and Lasso Regression models, we observed the following



```
                  Ridge Regression  Lasso Regression
Metric
R2 Score (Train)              0.94              0.92
R2 Score (Test)               0.93              0.93
RSS (Train)                   7.68             10.19
RSS (Test)                    4.08              4.00
MSE (Train)                   0.01              0.01
MSE (Test)                    0.01              0.01
RMSE (Train)                  0.09              0.10
RMSE (Test)                   0.10              0.10
```

- The R2 Score on test data was same (**0.93**) for both Ridge and Lasso Regression models

However, since Lasso helps in **feature reduction**, Lasso had better edge over Ridge Regression Model. **Hence, we chose Lasso Regression Model.**

Question 3.

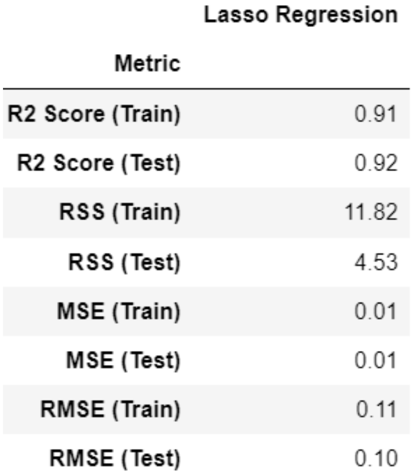
After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Answer

For the current Lasso Regression Model, the top 5 predictor variables were

- GrLivArea
- Exterior1st_BrkFace
- OverallQual
- Neighborhood_Crawfor
- Functional_Typ

We dropped the top 5 predictor variables and recreated the Lasso Regression model in Jupiter Notebook. For this new model, the model metrics were



	Lasso Regression
Metric	
R2 Score (Train)	0.91
R2 Score (Test)	0.92
RSS (Train)	11.82
RSS (Test)	4.53
MSE (Train)	0.01
MSE (Test)	0.01
RMSE (Train)	0.11
RMSE (Test)	0.10

And the 5 most important predictor variables now are

- 2ndFlrSF
- MSSubClass_70
- 1stFlrSF
- Neighborhood_Somerst
- Neighborhood_StoneBr

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?

As per Occam's Razor

- Simpler models are usually more generic and are more widely applicable
- Simpler models require fewer training samples for effective training than the more complex ones and hence are easier to train
- Simpler models are more robust
  - Complex models have tendency to change wildly with changes in the training data set
  - Simple models have low variance, high bias and complex models have low bias, high Variance
  - Simpler models make more errors in the training set. Complex models lead to overfitting- they work very well for the training samples, fail miserably when applied to other test data

Therefore, to make the model more robust and generalizable, we have to make the model simple but not simpler which will not be of any use.

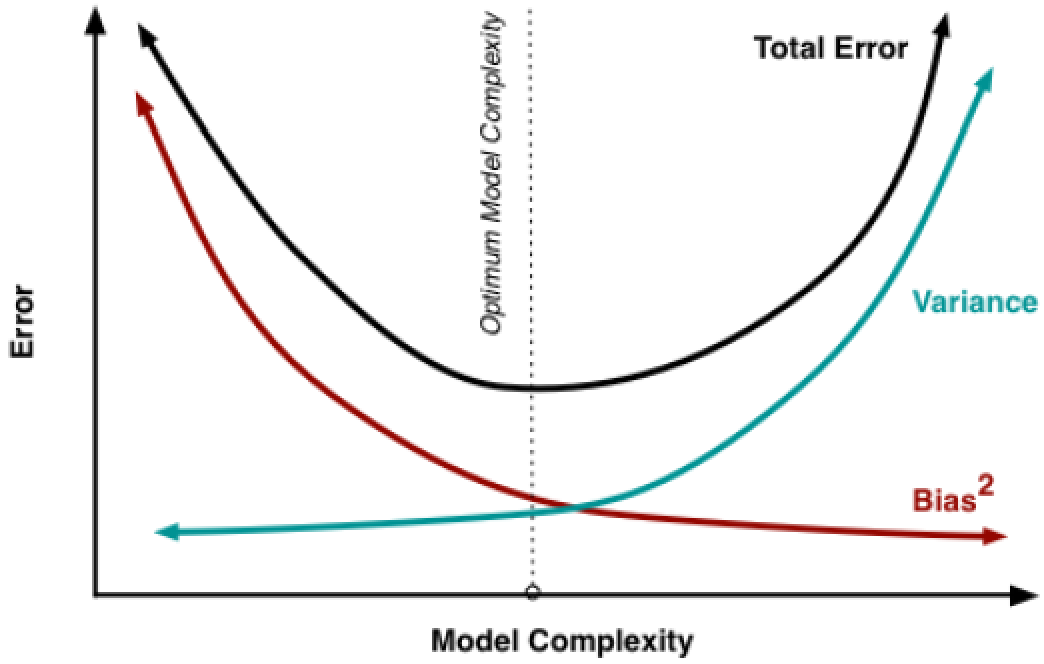
Regularization can be used to make the model simpler. Regularization helps to strike the delicate balance between keeping the model simple and not making it too naive to be of any use.

Making a model simple lead to Bias-Variance Trade-off

- A complex model will need to change for every little change in the dataset and hence is very unstable and extremely sensitive to any changes in the training data
- A simpler model that abstracts out some pattern followed by the data points given is unlikely to change wildly even if more points are added or removed

Bias quantifies how accurate is the model likely to be on test data. A complex model can do an accurate job prediction provided there is enough training data. Variance refers to the degree of changes in the model itself with respect to changes in the training data.

Thus accuracy of the model can be maintained by keeping the balance between Bias and Variance as it minimizes the total error as shown in the below graph.


Optimum Model Complexity
Total Error
Error
Variance
Bias^2
Model Complexity