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?

Ridge and Lasso regression are the techniques used to reduce the model complexity and also prevent model overfitting by adding penalty term(lambda) to RSS (Residual Sum of Squares) to minimize the entire cost towards zero.

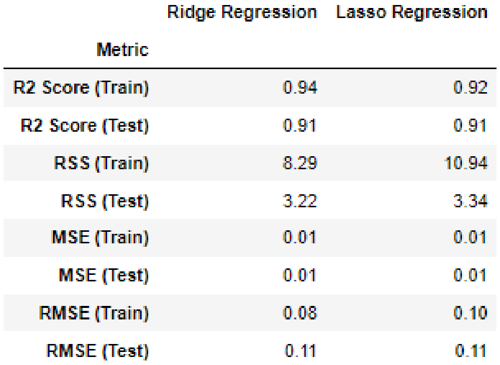
Ridge and lasso both adds penalty term to RSS but the difference is:

Ridge uses sum of squared coefficients and Lasso uses sum of absolute value of coefficients

The optimal value of alpha for ridge and lasso regression for the model created is:

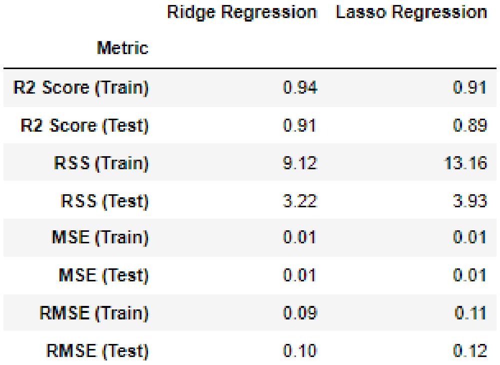
1) Alpha for Ridge regression: 10
2) Alpha for Lasso Regression: 0.001

When we double the value of alpha for Ridge and Lasso:



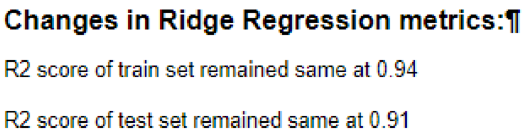
	Ridge Regression	Lasso Regression
Metric		
R2 Score (Train)	0.94	0.92
R2 Score (Test)	0.91	0.91
RSS (Train)	8.29	10.94
RSS (Test)	3.22	3.34
MSE (Train)	0.01	0.01
MSE (Test)	0.01	0.01
RMSE (Train)	0.08	0.10
RMSE (Test)	0.11	0.11

With Alpha 10 and 0.001



	Ridge Regression	Lasso Regression
Metric		
R2 Score (Train)	0.94	0.91
R2 Score (Test)	0.91	0.89
RSS (Train)	9.12	13.16
RSS (Test)	3.22	3.93
MSE (Train)	0.01	0.01
MSE (Test)	0.01	0.01
RMSE (Train)	0.09	0.11
RMSE (Test)	0.10	0.12

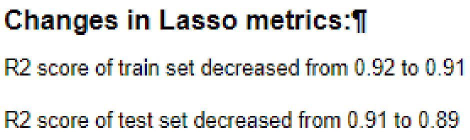
With Alpha 20 and 0.002



Changes in Ridge Regression metrics:¶

R2 score of train set remained same at 0.94

R2 score of test set remained same at 0.91

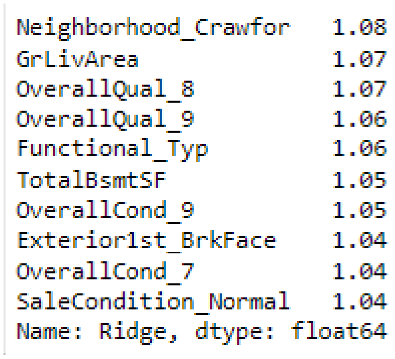


Changes in Lasso metrics:¶

R2 score of train set decreased from 0.92 to 0.91

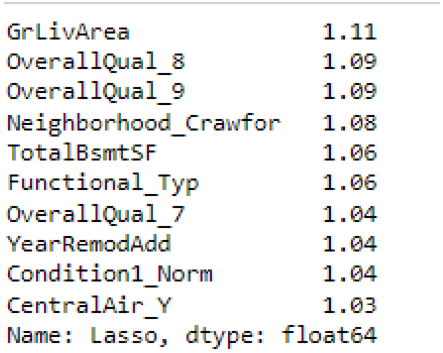
R2 score of test set decreased from 0.91 to 0.89

Most Important Predictors after doubling the values are:



```
Neighborhood_Crawfor   1.08
GrLivArea              1.07
OverallQual_8          1.07
OverallQual_9          1.06
Functional_Typ         1.06
TotalBsmtSF            1.05
OverallCond_9          1.05
Exterior1st_BrkFace    1.04
OverallCond_7          1.04
SaleCondition_Normal   1.04
Name: Ridge, dtype: float64
```

Top Predictors using Ridge after doubling alpha



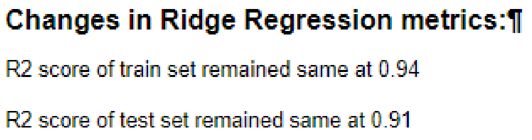
```
GrLivArea              1.11
OverallQual_8          1.09
OverallQual_9          1.09
Neighborhood_Crawfor   1.08
TotalBsmtSF            1.06
Functional_Typ         1.06
OverallQual_7          1.04
YearRemodAdd           1.04
Condition1_Norm        1.04
CentralAir_Y           1.03
Name: Lasso, dtype: float64
```

Top Predictors using Lasso after doubling alpha

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?

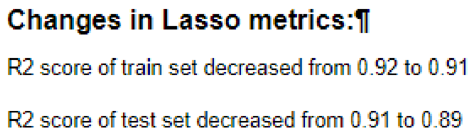
R2 score for ridge (0.94) is greater than Lasso (0.92) for training dataset however for test dataset its same (0.91) so we'll chose Lasso as it does feature selection and tends to make most of the predictor variables as 0 which will make the model simple.



Changes in Ridge Regression metrics:¶

R2 score of train set remained same at 0.94

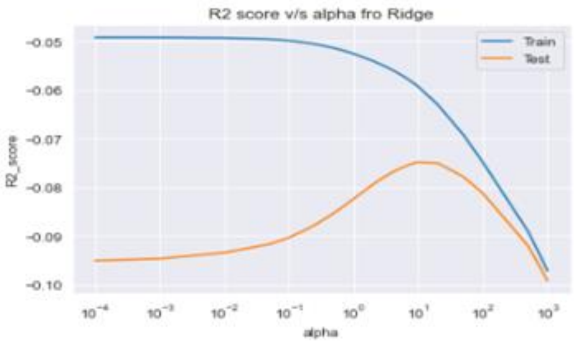
R2 score of test set remained same at 0.91

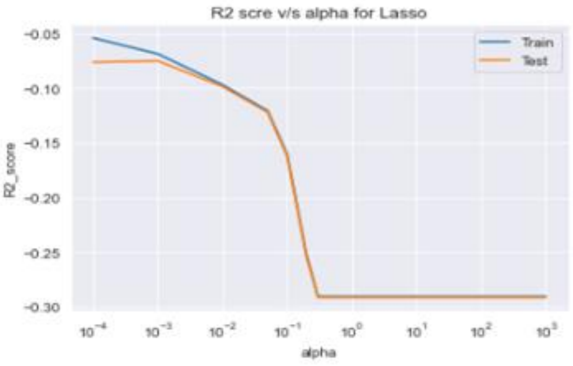


Changes in Lasso metrics:¶

R2 score of train set decreased from 0.92 to 0.91

R2 score of test set decreased from 0.91 to 0.89



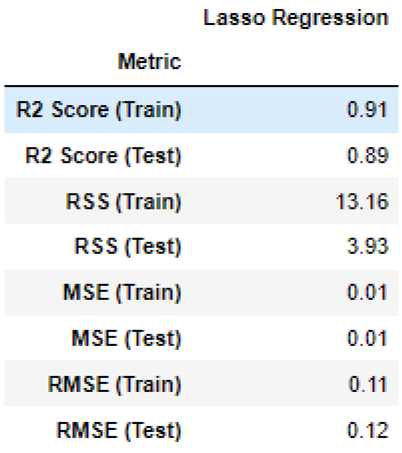


Question 3: After building the model, you realized that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables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

Since we are dropping top 5 predictor variables which are:

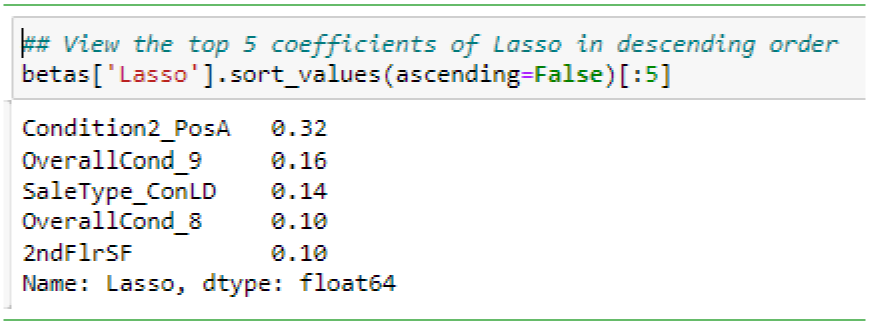
'OverallQual_9', 'GrLivArea', 'OverallQual_8', 'Neighborhood_Crawfor' and 'Functional_Typ'

Alpha and R2_score decreased to 0.0001 from 0.001 and 0.89 and 0.91 for Lasso regression



Metric	Lasso Regression
R2 Score (Train)	0.91
R2 Score (Test)	0.89
RSS (Train)	13.16
RSS (Test)	3.93
MSE (Train)	0.01
MSE (Test)	0.01
RMSE (Train)	0.11
RMSE (Test)	0.12

And the Top 5 Predictive variables will be:



```
## View the top 5 coefficients of Lasso in descending order
betas['Lasso'].sort_values(ascending=False)[:5]
```

```
Condition2_PosA    0.32
OverallCond_9      0.16
SaleType_ConLD     0.14
OverallCond_8      0.10
2ndFlrSF           0.10
Name: Lasso, dtype: float64
```

Question 4: How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

- A model is robust when any variation in the data does not affect its performance much.
- A generalizable model is able to adapt properly to new, previously unseen data, drawn from the same distribution as the one used to create the model.
- To make sure a model is robust and generalizable, we have to take care it doesn't overfit. This is because an overfitting model has very high variance and a smallest change in data affects the model prediction heavily. Such a model will identify all the patterns of a training data, but fail to pick up the patterns in unseen test data.
- In other words, the model should not be too complex in order to be robust and generalizable.
- If we look at it from the perspective of Accuracy, a too complex model will have a very high accuracy. So, to make our model more robust and generalizable, we will have to decrease variance which will lead to some bias. Addition of bias means that accuracy will decrease.

- In general, we have to find strike some balance between model accuracy and complexity. This can be achieved by Regularization techniques like Ridge Regression and Lasso.