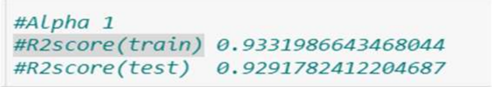
Advanced Regression : Answers for Subjective Questions

By
Renuka Mokka

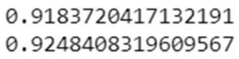
Q1. What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

- The optimal value of lambda for Ridge regression – 4, R2 score for test and train are as below.



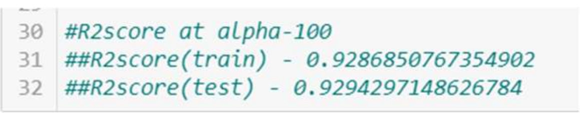
```
#Alpha 1
#R2score(train) 0.9331986643468044
#R2score(test)  0.9291782412204687
```

When it doubled, the R2 score decreases slightly as below.



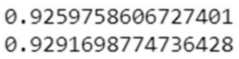
```
0.9183720417132191
0.9248408319609567
```

- The optimal value of lambda for Lasso Regression – 100 - R2 score for test and train are as below



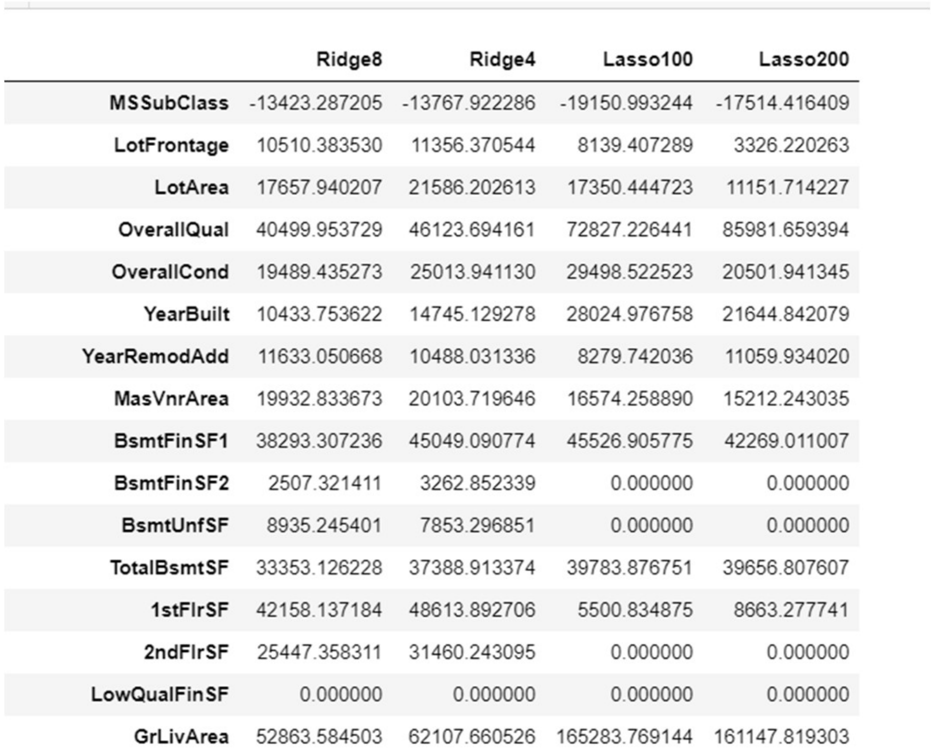
```
#R2score at alpha-100
##R2score(train) - 0.9286850767354902
##R2score(test) - 0.9294297148626784
```

When it doubled to 200, the R2 score for train and test data slightly decreased as below.



```
0.9259758606727401
0.9291698774736428
```

The coefficients changed as below when we double the lambdas in Ridge and Lasso:



	Ridge8	Ridge4	Lasso100	Lasso200
MSSubClass	-13423.287205	-13767.922286	-19150.993244	-17514.416409
LotFrontage	10510.383530	11356.370544	8139.407289	3326.220263
LotArea	17657.940207	21586.202613	17350.444723	11151.714227
OverallQual	40499.953729	46123.694161	72827.226441	85981.659394
OverallCond	19489.435273	25013.941130	29498.522523	20501.941345
YearBuilt	10433.753622	14745.129278	28024.976758	21644.842079
YearRemodAdd	11633.050668	10488.031336	8279.742036	11059.934020
MasVnrArea	19932.833673	20103.719646	16574.258890	15212.243035
BsmtFinSF1	38293.307236	45049.090774	45526.905775	42269.011007
BsmtFinSF2	2507.321411	3262.852339	0.000000	0.000000
BsmtUnfSF	8935.245401	7853.296851	0.000000	0.000000
TotalBsmtSF	33353.126228	37388.913374	39783.876751	39656.807607
1stFlrSF	42158.137184	48613.892706	5500.834875	8663.277741
2ndFlrSF	25447.358311	31460.243095	0.000000	0.000000
LowQualFinSF	0.000000	0.000000	0.000000	0.000000
GrLivArea	52863.584503	62107.660526	165283.769144	161147.819303

Impact on predictors for ridge and Lasso:

The top 10 features remained similar when doubled the lambda in ridge and Lasso, but just the coefficients changes slightly, and order shifted for few.

1. GrLivArea - Above grade (ground) living area square feet

2. 1stFlrSF - First Floor square feet

3. OverallQual - Rates the overall material and finish of the house

4. BsmtFinSF1 - Type 1 finished square feet

5. TotalBsmtSF - Total square feet of basement area

6.TotRmsAbvGrd - Total rooms above grade (does not include bathrooms)

7. 2ndFlrSF - Second floor square feet

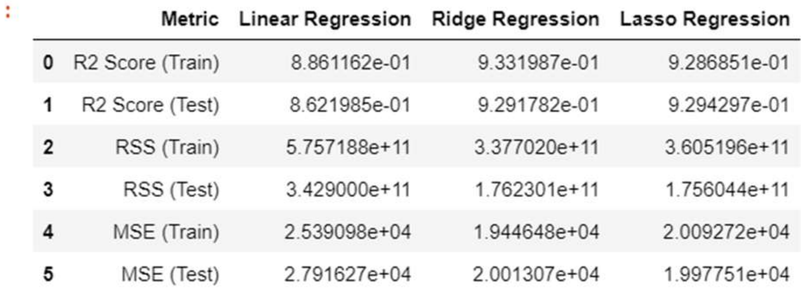
8. MasVnrArea - Masonry veneer area in square feet

9. OverallCond - Rates the overall condition of the house

10. LotArea - Lot size in square feet

Q2 - 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?

- The optimal value of lambda for Ridge regression – 4
- The optimal value of lambda for Lasso Regression – 100
- The results are very much improved in Ridge and Lasso compared to Linear Regression.
- Ridge and Lasso results are very similar, I would prefer to choose Lasso regression as we are getting better test results, also we can avoid highly correlated features.



	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	8.861162e-01	9.331987e-01	9.286851e-01
1	R2 Score (Test)	8.621985e-01	9.291782e-01	9.294297e-01
2	RSS (Train)	5.757188e+11	3.377020e+11	3.605196e+11
3	RSS (Test)	3.429000e+11	1.762301e+11	1.756044e+11
4	MSE (Train)	2.539098e+04	1.944648e+04	2.009272e+04
5	MSE (Test)	2.791627e+04	2.001307e+04	1.997751e+04

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?

The tope 5 predictors which we got in Lasso are:

1. GrLivArea - Above grade (ground) living area square feet
2. OverallQual - Rates the overall material and finish of the house
3. BsmtFinSF1 - Type 1 finished square feet
4. TotalBsmtSF - Total square feet of basement area
5. Neighborhood_NridgHt - Physical locations within Ames city limits - Northridge Heights

After dropping top 5 predictors from lasso, and rebuilding model we get the below as top 5 predictors.

1. 1stFlrSF - First Floor square feet
2. 2ndFlrSF - Second floor square feet
3. Neighborhood_StoneBr - Physical locations within Ames city limits - Stone Brook
4. OverallCond - Rates the overall condition of the house
5. Street_Pave - Type of road access to property -Paved

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?

To make a model robust and generalizable , we should try to solve the problem of overfitting, should not try to learn data too much that it performs well on train data but performs poor on unseen real data.

To avoid the overfitting problem, we should use regularization methods like Ridge and lasso.

Data cleaning is also very important to make the model robust. Too much weightage should not be given to outlier data to improve the train accuracy, but it fails on test data. We should first check for outliers in data and remove the outliers which do not make sense and build the model to increase the test accuracy.

So, for a model to be robust and generalizable, we should take care of important steps like data quality and diversity, data extraction and feature selection, model selection and evaluation, model optimization and tuning.

If the model is not robust, It cannot be trusted for predictive analysis.