7/5/2022

1. Subjective Based

Naveen Kumar Jagadeesan

1. Assignment-based Subjective Questions

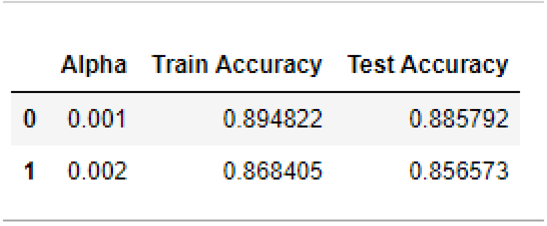
Question 1:

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:

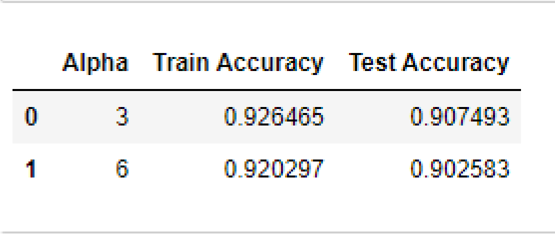
- The Optimal value of alpha for Ridge = 3
- The Optimal value of alpha for Lasso = 0.001
- After Doubling the Value of Alpha for Ridge = 6
- After Doubling the Value of Alpha for Lasso = 0.002
- When doubling the value, R2 Score Getting reduced for both the Ridge and Lasso Regression. But it is not too much different between them.

For Lasso Regression



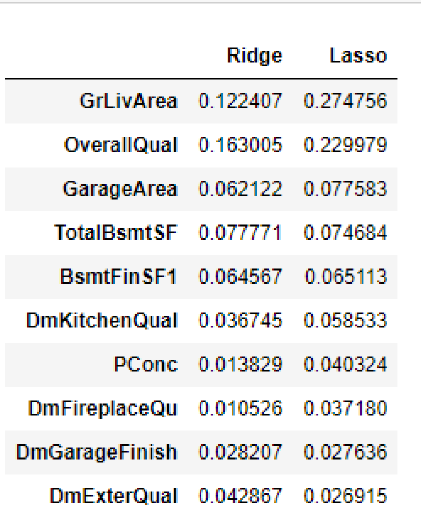
	Alpha	Train Accuracy	Test Accuracy
0	0.001	0.894822	0.885792
1	0.002	0.868405	0.856573

For Ridge Regression



	Alpha	Train Accuracy	Test Accuracy
0	3	0.926465	0.907493
1	6	0.920297	0.902583

Most important predictor variables after the change are implemented



	Ridge	Lasso
GrLivArea	0.122407	0.274756
OverallQual	0.163005	0.229979
GarageArea	0.062122	0.077583
TotalBsmtSF	0.077771	0.074684
BsmtFinSF1	0.064567	0.065113
DmKitchenQual	0.036745	0.058533
PConc	0.013829	0.040324
DmFireplaceQu	0.010526	0.037180
DmGarageFinish	0.028207	0.027636
DmExterQual	0.042867	0.026915

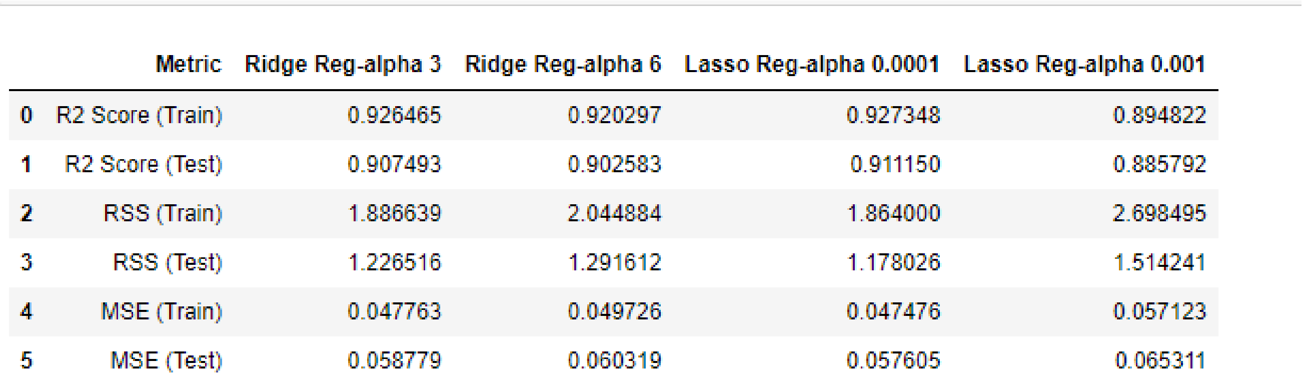
The Working are shown in the python notebook. If need further details, you can refer the notebook.

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 Optimal value of alpha for Ridge = 3
- The Optimal value of alpha for Lasso = 0.001
- After Doubling the Value of Alpha for Ridge = 6
- After Doubling the Value of Alpha for Lasso = 0.002
- When doubling the value, R2 Score Getting reduced for both the Ridge and Lasso Regression. But it is not too much different between them.



	Metric	Ridge Reg-alpha 3	Ridge Reg-alpha 6	Lasso Reg-alpha 0.0001	Lasso Reg-alpha 0.001
0	R2 Score (Train)	0.926465	0.920297	0.927348	0.894822
1	R2 Score (Test)	0.907493	0.902583	0.911150	0.885792
2	RSS (Train)	1.886639	2.044884	1.864000	2.698495
3	RSS (Test)	1.226516	1.291612	1.178026	1.514241
4	MSE (Train)	0.047763	0.049726	0.047476	0.057123
5	MSE (Test)	0.058779	0.060319	0.057605	0.065311

The number of feature parameters selected by lasso is 43 for 0.001 and 107 for 0.0001. but for Ridge it takes all the 180 columns as the feature parameters.

Lasso helps in feature reduction as the coefficient value of some features would be 0 and for ridge it will be close to 0 and not zero. Ridge is better when compared with lasso for r2 score and lasso is better when feature selection comes into play.

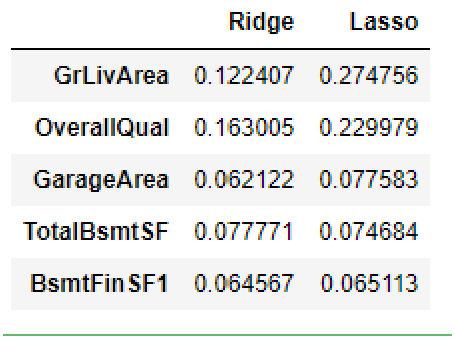
Anyway, got good results in both techniques. But will choose lasso so that manual feature selection is not required.

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:

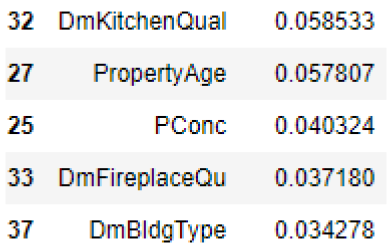
First Five variables



	Ridge	Lasso
GrLivArea	0.122407	0.274756
OverallQual	0.163005	0.229979
GarageArea	0.062122	0.077583
TotalBsmtSF	0.077771	0.074684
BsmtFinSF1	0.064567	0.065113

These variables are worked in the python notebook and found the first 5 predictive variables.

Next Five variables



32	DmKitchenQual	0.058533
27	PropertyAge	0.057807
25	PConc	0.040324
33	DmFireplaceQu	0.037180
37	DmBldgType	0.034278

These are the next 5 important predictor variables after the first 5 variables are excluded

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:

1. Remove Outliers, missing values and imputing the mode values are the best measures to make a model is robust or not
2. Use median as a measure instead of mean so that it performs better when there are more outliers
3. Simple Models would perform well on unseen data, complex models would perform on training data, but it might fail in testing data as it called as overfitting
4. Ensuring correct trade off between bias and variance to make model more robust and generalisable.
5. Model Accuracy can be managed easily by reducing total error.
6. So finally, overfitting / underfitting the dataset to make the model more robust and generalisable