2022

Advanced Regression – Graded Assignment

AUSTRALIAN HOUSE PRICE - ADVANCED REGRESSION - ASSIGNMENT

BHAVESHKUMAR THAKER

UPGRAD – CHANDIGARH UNIVERSITY – M.SC. IN DATA SCIENCE

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 to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer 1a: Following is the optimal value of alpha for ridge and lasso regression retrieved using GridSearchCV.

Regression	**Alpha**
Ridge	6.0
Lasso	0.001

Answer 1b: The alpha value passed as hyper parameter to the Ridge and Lasso regressors needs be double.

```
ridge_reg = Ridge(alpha=(6.0 * 2))

lasso_reg = Lasso(alpha=(0.001 * 2))
```

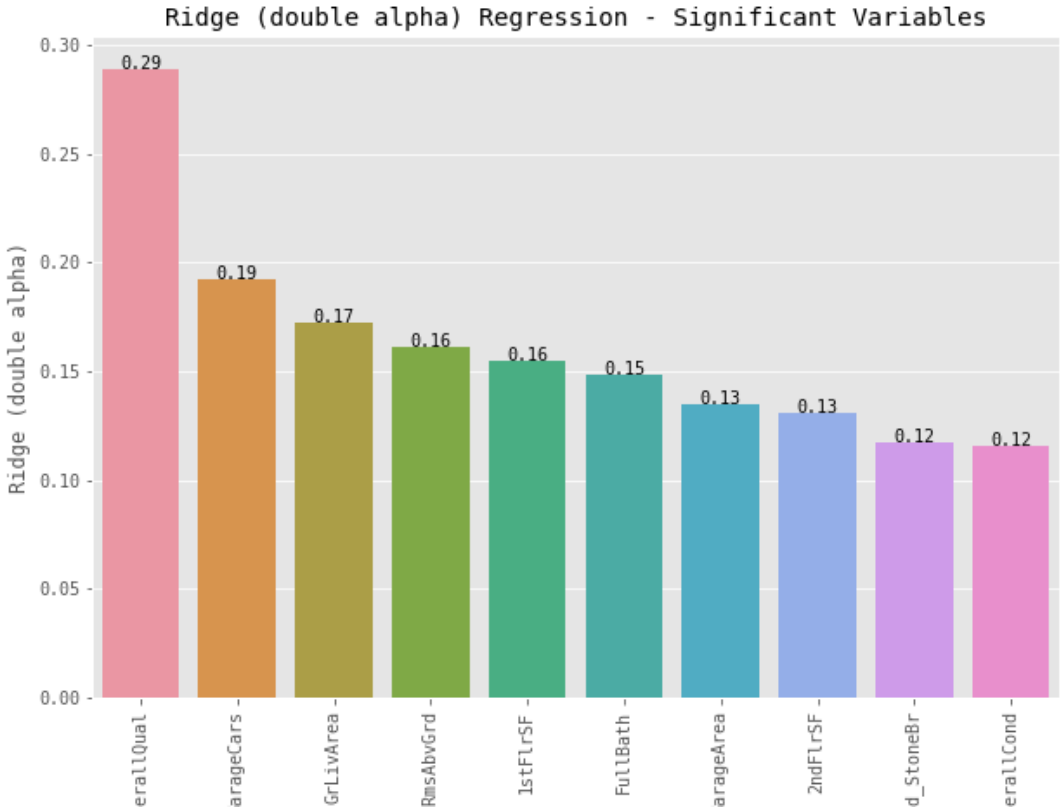
This will increase the value of alpha passed to Ridge and Lasso regressors. Due to increased alpha we can see in the Jupyter Notebook that the r2_score is reduced for both regression.

	Model_Name	MSE_Score	RMSE_Score	R2_Score
1	LinearRegression (RFE)	0.02573606259572235	0.16042463213522526	0.8346595297612565
2	Ridge	0.021128182665372728	0.14535536682686584	0.8642626996965949
3	Ridge (alpha*2)	0.022128609699399258	0.14875688118335656	0.8578354897987999
4	Lasso	0.019392145867684382	0.13925568522571846	0.8754158098281059
5	Lasso (alpha*2)	0.02122058821779823	0.14567288085912983	0.8636690433269043

Answer 1c: Following are the important predictors after the change of alpha to double its value is implemented.

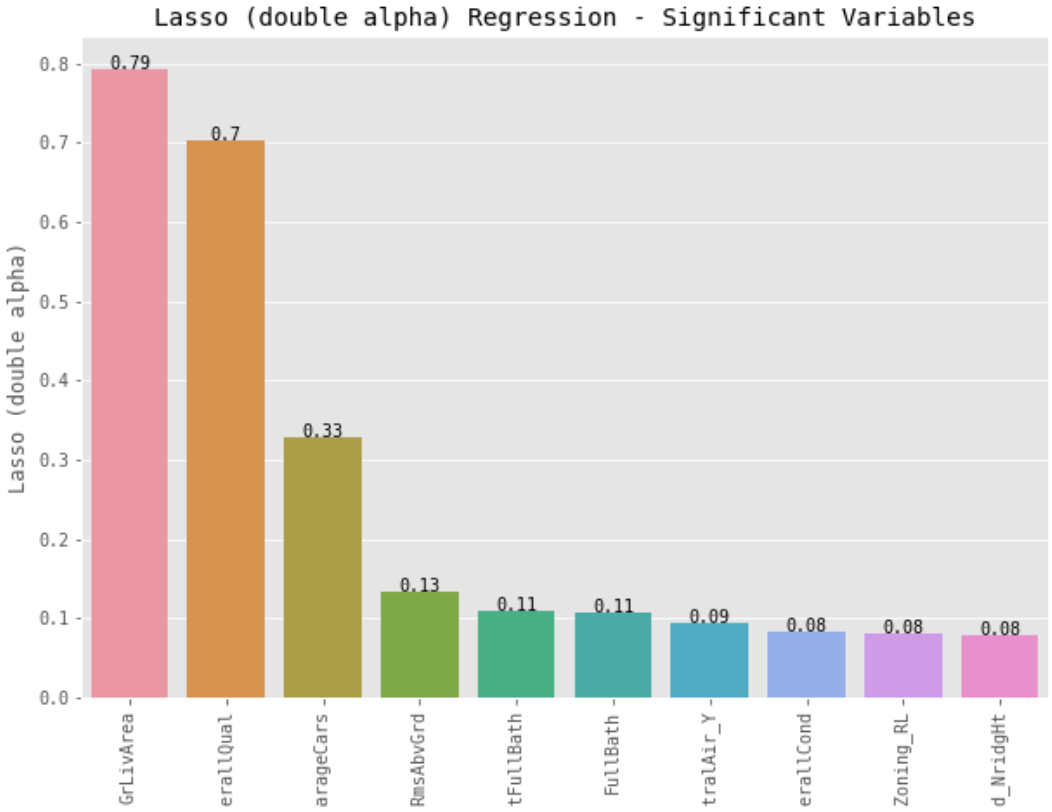
Ridge Regression (double alpha)

OverallQual	0.28917
GarageCars	0.19214
GrLivArea	0.17271
TotRmsAbvGrd	0.16135
1stFlrSF	0.15500
FullBath	0.14821
GarageArea	0.13466
2ndFlrSF	0.13111
Neighborhood_StoneBr	0.11737
OverallCond	0.11555



Lasso Regression (double alpha)

GrLivArea	0.79357
OverallQual	0.70333
GarageCars	0.32740
TotRmsAbvGrd	0.13430
BsmtFullBath	0.10844
FullBath	0.10719
CentralAir_Y	0.09400
OverallCond	0.08391
MSZoning_RL	0.08136
Neighborhood_NridgHt	0.07789


Lasso (double alpha) Regression - Significant Variables
Lasso (double alpha)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.79
0.7
0.33
0.13
0.11
0.11
0.09
0.08
0.08
0.08
GrLivArea
erallQual
arageCars
RmsAbvGrd
tFullBath
FullBath
tralAir_Y
erallCond
Zoning_RL
d_NridgHt

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 2: The Lasso Regression be the best model for solving the problem because

- The r2_score of the Lasso Regression is higher than the Ridge Regression on test dataset
- The RMSE score of the Lasso Regression is lower than the Ridge Regression on test dataset

	Model_Name	MSE_Score	RMSE_Score	R2_Score
1	LinearRegression (RFE)	0.02573606259572235	0.16042463213522526	0.8346595297612565
2	Ridge	0.021128182665372728	0.14535536682686584	0.8642626996965949
3	Ridge (alpha*2)	0.022128609699399258	0.14875688118335656	0.8578354897987999
4	Lasso	0.019392145867684382	0.13925568522571846	0.8754158098281059
5	Lasso (alpha*2)	0.02122058821779823	0.14567288085912983	0.8636690433269043

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?

Answer 3: Following are the important predictors after dropping the top 5 most important predictors in the Lasso Regression Model.

1stFlrSF	0.99654
2ndFlrSF	0.41257
GarageArea	0.35087
Neighborhood_StoneBr	0.17759
FullBath	0.17356
Neighborhood_NridgHt	0.15433
Neighborhood_Crawfor	0.10983
CentralAir_Y	0.10632
Functional_Typ	0.10363
BsmtFullBath	0.10159

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?

Answer 4: A model is robust if its output dependent variable is consistently accurate even if one or more of the input independent variables or assumptions are drastically changed due to unforeseen circumstances. A model should also be generalizable so that the test accuracy is not lesser than the training score. Thus, A model needs to be made robust and generalizable so that they are not impacted by outliers in the training data.

The outlier analysis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which it does not make sense to keep must be removed from the dataset. This would help standardize the predictions made by the model. This would help increase the accuracy of the predictions made by the model.