House Price Prediction – Subjective Questions

- Dhandapani Subramanian

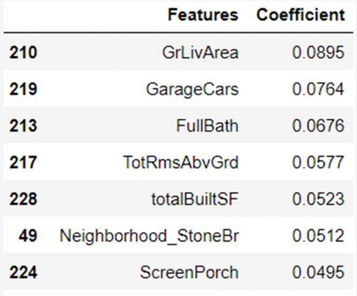
Question 1

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?

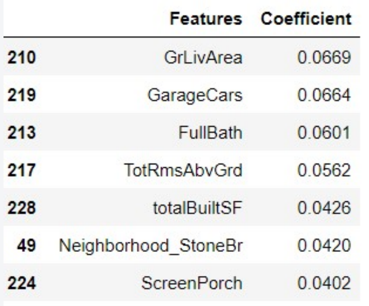
Answer)

The optimal value of alpha for ridge was 4.0 and that of lasso regression was 0.0002. When we doubled the value of alpha for ridge and lasso i.e 8.0 and 0.0004 respectively, then the predictor coefficients got reduced. Thus forming more regularized solution. Few variables changed their positions as well.

Here is the change in the model for ridge at 4 and 8 alpha values.

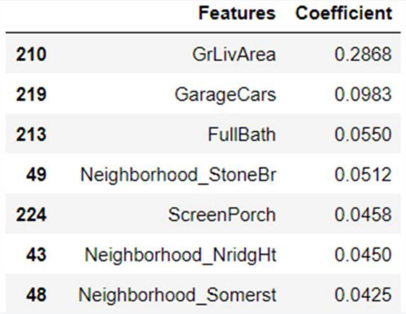


	Features	Coefficient
210	GrLivArea	0.0895
219	GarageCars	0.0764
213	FullBath	0.0676
217	TotRmsAbvGrd	0.0577
228	totalBuiltSF	0.0523
49	Neighborhood_StoneBr	0.0512
224	ScreenPorch	0.0495

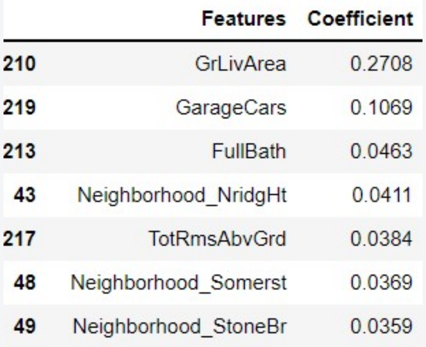


	Features	Coefficient
210	GrLivArea	0.0669
219	GarageCars	0.0664
213	FullBath	0.0601
217	TotRmsAbvGrd	0.0562
228	totalBuiltSF	0.0426
49	Neighborhood_StoneBr	0.0420
224	ScreenPorch	0.0402

Lasso regression changes are shown here for alpha 0.0002 and 0.0004



	Features	Coefficient
210	GrLivArea	0.2868
219	GarageCars	0.0983
213	FullBath	0.0550
49	Neighborhood_StoneBr	0.0512
224	ScreenPorch	0.0458
43	Neighborhood_NridgHt	0.0450
48	Neighborhood_Somerst	0.0425



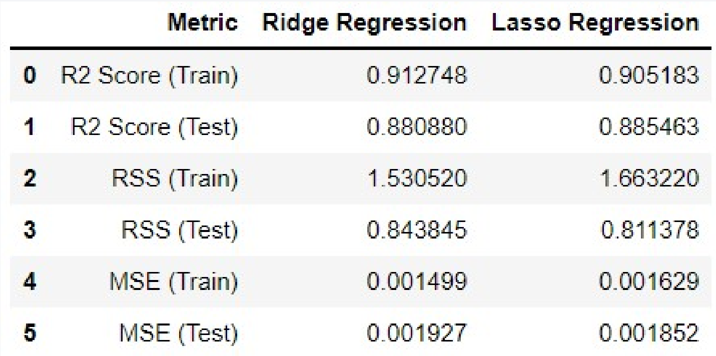
	Features	Coefficient
210	GrLivArea	0.2708
219	GarageCars	0.1069
213	FullBath	0.0463
43	Neighborhood_NridgHt	0.0411
217	TotRmsAbvGrd	0.0384
48	Neighborhood_Somerst	0.0369
49	Neighborhood_StoneBr	0.0359

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)

Optimal value for the hyperparameter lambda for ridge was 4.0 and lasso was 0.0002 for the assignment. Here is the metrics for both the models.



	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.912748	0.905183
1	R2 Score (Test)	0.880880	0.885463
2	RSS (Train)	1.530520	1.663220
3	RSS (Test)	0.843845	0.811378
4	MSE (Train)	0.001499	0.001629
5	MSE (Test)	0.001927	0.001852

From the above results we can conclude that Ridge is the better model for finding the house price. Because R square of ridge for both test and train value is high compare with Lasso and linear model.

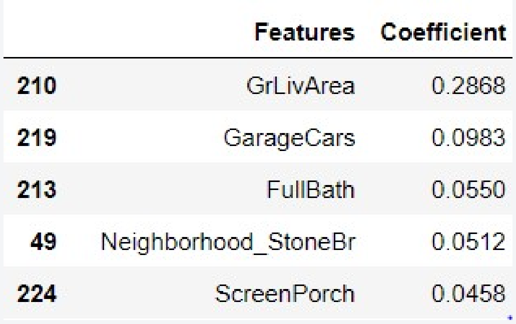
RSS and MSE (Train and test) values are also low for Ridge model when comparing with Lasso.

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?

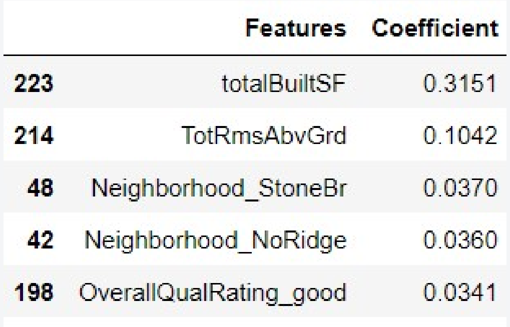
Answer)

Here were the initial top 5 predictor variables in lasso model for house price prediction.



	Features	Coefficient
210	GrLivArea	0.2868
219	GarageCars	0.0983
213	FullBath	0.0550
49	Neighborhood_StoneBr	0.0512
224	ScreenPorch	0.0458

When we delete these top 5 variables from the input data and recreate lasso model. We get the next top 5 predictors



	Features	Coefficient
223	totalBuiltSF	0.3151
214	TotRmsAbvGrd	0.1042
48	Neighborhood_StoneBr	0.0370
42	Neighborhood_NoRidge	0.0360
198	OverallQualRating_good	0.0341

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)

The model must be able to work with outliers and missing data. They should have better R square score for their test data than the training data. We can use proper EDA techniques like standardization and normalization. This world help in good extend for the algorithms to work up on. We can also make use of confidence intervals.