ADVANCED REGRESSION ASSIGNMENT-SUBJECTIVE QUESTIONS
ANJANAVA DAS PURKAYASTHA

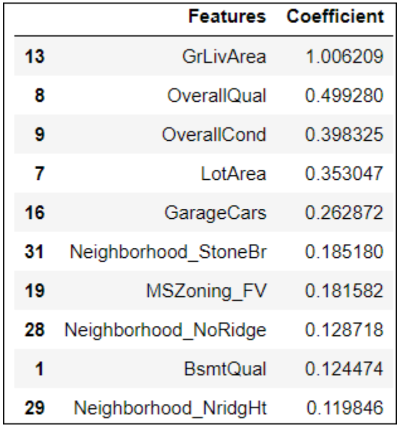
Q1) What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?
Answer: Optimal values of alpha for ridge and lasso regression are 0.5 and 0.0001 respectively.

On doubling the alpha for both ridge and lasso model, we have following R2's as shown in the table:

MODEL	ALPHA	R2_Train	R2_Test	Adjusted R2
Ridge	0.5	0.904	0.91	0.904
	1	0.902	0.907	0.902
Lasso	0.0001	0.904	0.91	0.904
	0.0002	0.902	0.907	0.902

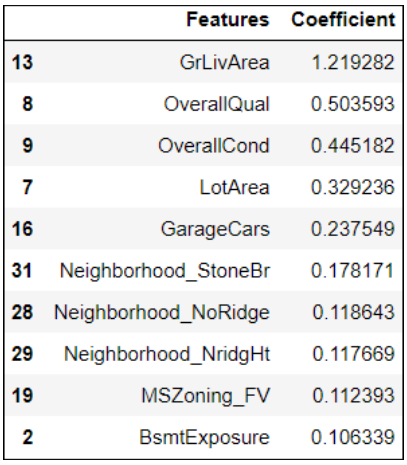
Doubling the alpha for both Ridge and Lasso model caused reduction in the magnitude of the coefficients. Increasing alpha would mean increase in regularization. This means, increase in alpha value increases the bias and reduces the variance. It is seen that there is a very trivial drop in R2 value after doubling.

The 10 most important variables, after the change is implemented, for ridge regression model are:



	Features	Coefficient
13	GrLivArea	1.006209
8	OverallQual	0.499280
9	OverallCond	0.398325
7	LotArea	0.353047
16	GarageCars	0.262872
31	Neighborhood_StoneBr	0.185180
19	MSZoning_FV	0.181582
28	Neighborhood_NoRidge	0.128718
1	BsmtQual	0.124474
29	Neighborhood_NridgHt	0.119846

And for lasso regression model are:



	Features	Coefficient
13	GrLivArea	1.219282
8	OverallQual	0.503593
9	OverallCond	0.445182
7	LotArea	0.329236
16	GarageCars	0.237549
31	Neighborhood_StoneBr	0.178171
28	Neighborhood_NoRidge	0.118643
29	Neighborhood_NridgHt	0.117669
19	MSZoning_FV	0.112393
2	BsmtExposure	0.106339

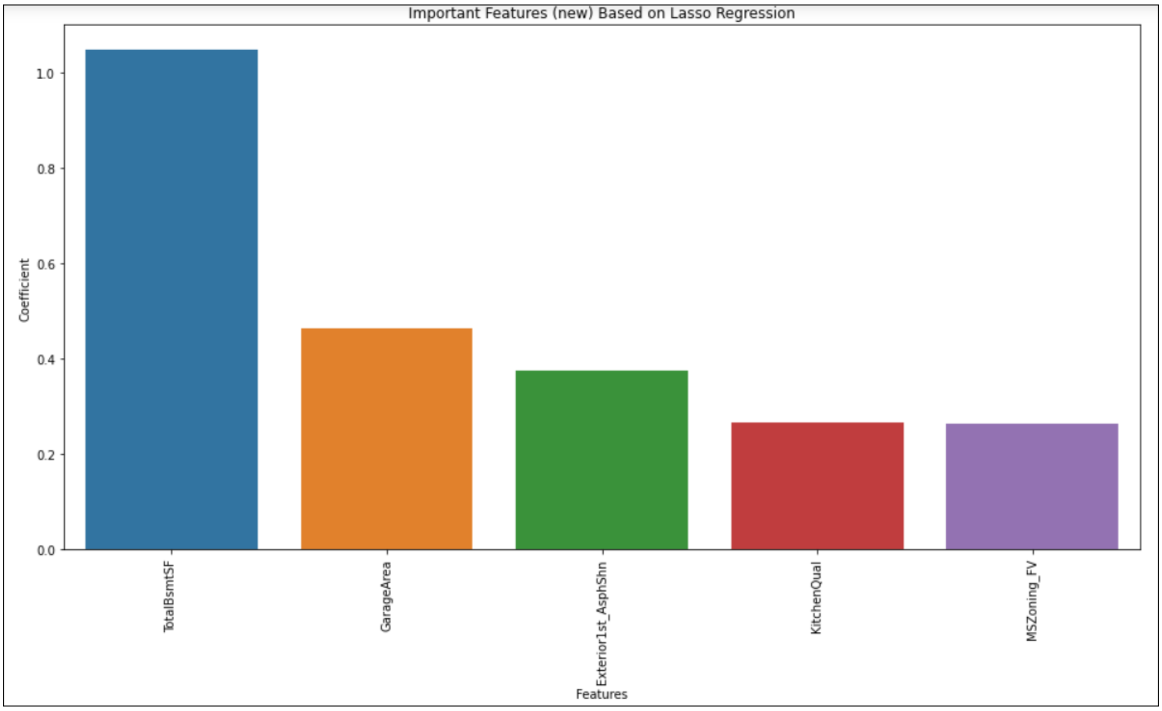
Q2) You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Answer: In terms of model performance, the R2 values of both ridge and lasso model are more-or-less same in this case study. But, following exclusive abilities of lasso gives an edge over ridge at times:

1) An additional ability of feature selection, on top of the usual feature elimination (e.g. RFE) methods.
2) Less computation time and memory
3) Simpler model.

Q3) 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: The new set of top 5 features after excluding the original top 5 features from Lasso regression model, are 'TotalBsmtSF', 'GarageArea', 'Exterior1st_AsphShn', 'KitchenQual' and 'MSZoning_FV'.



Q4) 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: A model is said to be robust and efficient, when it learns over the train data instead of memorizing it. An ideal model will have an optimal balance between bias and variance. The model should neither be overfitting nor underfit.

Optimal model results can be achieved by:

1) Data Cleaning and imputation;
2) Outlier analysis and handling;
3) Feature selection and elimination;
4) Hyper parameter tuning;
5) Dividing the data into 3 sets (train, validation and test) instead of 2 (train and test)
6) Usage of correct performance metric. E.g. Adjusted-R2 instead of R2 for linear regression models and Area under curve (AUC) for classification problems.
7) Usage of Regularization techniques.

The implication of having a robust model is primarily seen in the accuracy of the model. The difference between train and test data becomes trivial. This means the model can perform similarly on the unseen data, as it was performing on the train data. Robust and generalizable models are simple, but not that naive enough to ignore the basic trends in the data.