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:

Optimum Value for alpha - Ridge Regression:

5.0 Optimum Value for alpha - Lasso

Regression: 0.0004

After doubling the optimal value for alpha:

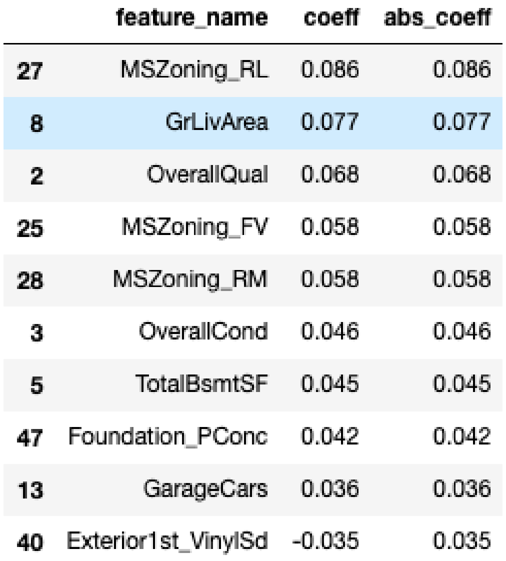
Ridge Regression: 10.0

Lasso Regression: 0.0008

For Ridge regression we saw that the coefficient values for a few features have reduced a little. The change in alpha does not seems to have significantly impacted the accuracy between the models.

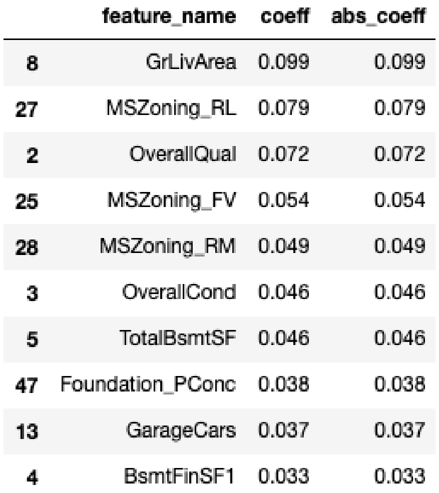
10 most important predictor variables after the change shown below (Calculated in the Notebook):

Ridge:



	feature_name	coeff	abs_coeff
27	MSZoning_RL	0.086	0.086
8	GrLivArea	0.077	0.077
2	OverallQual	0.068	0.068
25	MSZoning_FV	0.058	0.058
28	MSZoning_RM	0.058	0.058
3	OverallCond	0.046	0.046
5	TotalBsmtSF	0.045	0.045
47	Foundation_PConc	0.042	0.042
13	GarageCars	0.036	0.036
40	Exterior1st_VinylSd	-0.035	0.035

Lasso:



	feature_name	coeff	abs_coeff
8	GrLivArea	0.099	0.099
27	MSZoning_RL	0.079	0.079
2	OverallQual	0.072	0.072
25	MSZoning_FV	0.054	0.054
28	MSZoning_RM	0.049	0.049
3	OverallCond	0.046	0.046
5	TotalBsmtSF	0.046	0.046
47	Foundation_PConc	0.038	0.038
13	GarageCars	0.037	0.037
4	BsmtFinSF1	0.033	0.033

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:

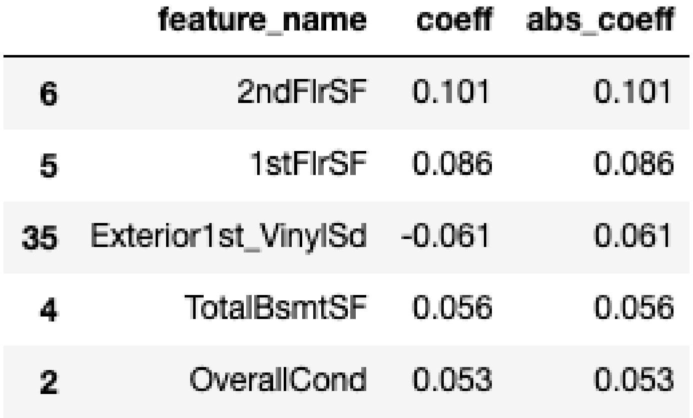
After finalizing the model and the R2 scores on both train and test data for both the models do not vary much, but as studied will apply Lasso regression model in place of ridge regression.

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:

After creating the new model below are the new top five important predictor variables based on the absolute value of their coefficients (Calculated as per the Jupyter Notebook).



	feature_name	coeff	abs_coeff
6	2ndFlrSF	0.101	0.101
5	1stFlrSF	0.086	0.086
35	Exterior1st_VinylSd	-0.061	0.061
4	TotalBsmtSF	0.056	0.056
2	OverallCond	0.053	0.053

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:

A model can be considered as generalizable when it does not overfits the training data set and performs same on the test data set as well. A model can be considered robust if it works for broad range of input data set i.e. is does not drastically change its behavior on changing of input data. Ideally accuracy should not vary much for training and test datasets.