Advanced Regression 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?

Ans:

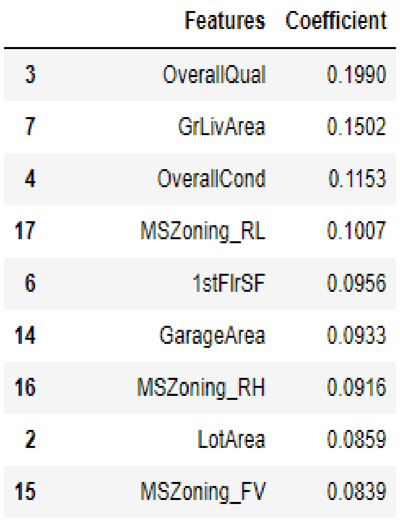
- The optimal value of alpha for **ridge is 0.9** and for **lasso is 0.0001**.

- Doubling the value of alpha in both Ridge and Lasso regression will lead to stronger regularization, resulting in smaller coefficient values for Ridge and in larger coefficient values Lasso. The specific impact depends on the dataset and the original values of alpha.

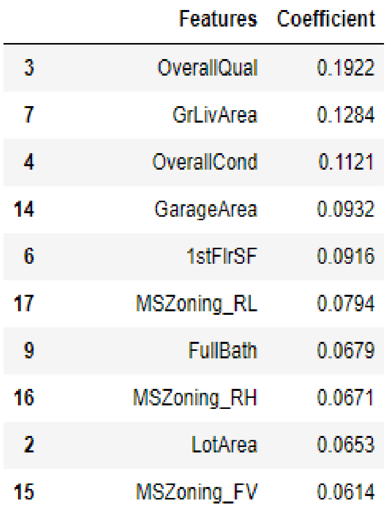
Table for the Lasso and Ridge after doubling the alpha values.

SNO	**Metric**	**Lasso**	**Ridge**
1	R2 Score (Train)	0.8804266832695572	0.8891622007208132
2	R2 Score (Test)	0.8727793661186324	0.8726707201295381
3	RSS (Train)	2.0299624796739018	1.8816620632309768
4	RSS (Test)	0.9739735499111919	0.9748053200136774
5	MSE (Train)	0.001988210068240844	0.001842959905221329
6	MSE (Test)	0.002218618564717977	0.0022205132574343446

- The most important 10 predictor variables after us double the value of alpha for Ridge is
- Optimal value of alpha = 0.9 and double of the alpha = 1.8.

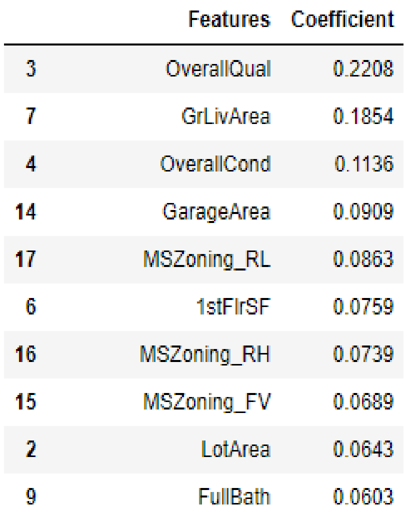


	Features	Coefficient
3	OverallQual	0.1990
7	GrLivArea	0.1502
4	OverallCond	0.1153
17	MSZoning_RL	0.1007
6	1stFlrSF	0.0956
14	GarageArea	0.0933
16	MSZoning_RH	0.0916
2	LotArea	0.0859
15	MSZoning_FV	0.0839

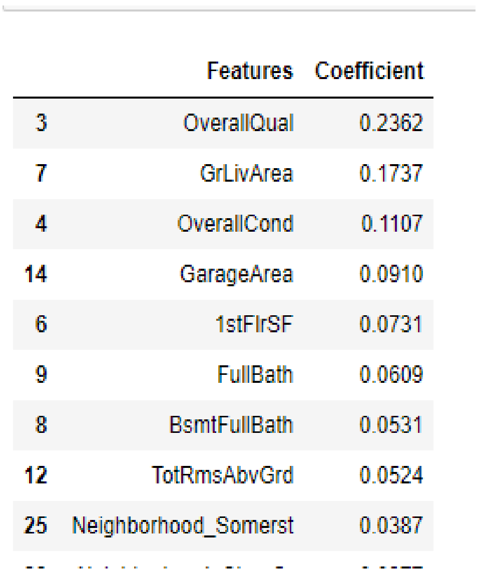


	Features	Coefficient
3	OverallQual	0.1922
7	GrLivArea	0.1284
4	OverallCond	0.1121
14	GarageArea	0.0932
6	1stFlrSF	0.0916
17	MSZoning_RL	0.0794
9	FullBath	0.0679
16	MSZoning_RH	0.0671
2	LotArea	0.0653
15	MSZoning_FV	0.0614

- The most important 10 predictor variables after us double the value of alpha for Lasso is
- Optimal value of alpha = 0.0001 and double of the alpha = 0.0002.



	Features	Coefficient
3	OverallQual	0.2208
7	GrLivArea	0.1854
4	OverallCond	0.1136
14	GarageArea	0.0909
17	MSZoning_RL	0.0863
6	1stFlrSF	0.0759
16	MSZoning_RH	0.0739
15	MSZoning_FV	0.0689
2	LotArea	0.0643
9	FullBath	0.0603



	Features	Coefficient
3	OverallQual	0.2362
7	GrLivArea	0.1737
4	OverallCond	0.1107
14	GarageArea	0.0910
6	1stFlrSF	0.0731
9	FullBath	0.0609
8	BsmtFullBath	0.0531
12	TotRmsAbvGrd	0.0524
25	Neighborhood_Somerst	0.0387

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?

Ans:

From the below points we can consider **Lasso Regression** is better for the Housing data set.

- We can observe the Lasso Regression performance well on the test data as compared to Ridge and Linear.
- The Difference between train and test data for Lasso is approx 1% where as for Ridge and Linear is approx 2%.
- Lasso Regression performs well on unseen data as compared to Ridge and Linear Regression.
- MSE (Test) of Lasso is less than Ridge and Linear Regression.

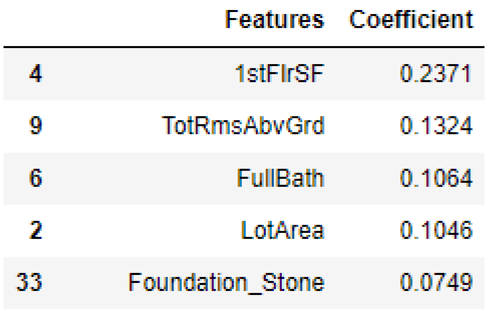
Question 3

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?

Ans:

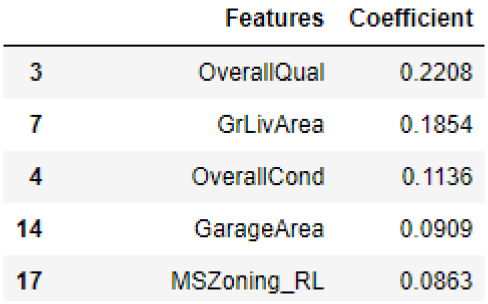
These are the top 5 predictors which we got after dropping the top 5 from the lasso.

- After dropping.



	Features	Coefficient
4	1stFlrSF	0.2371
9	TotRmsAbvGrd	0.1324
6	FullBath	0.1064
2	LotArea	0.1046
33	Foundation_Stone	0.0749

- Before dropping.



	Features	Coefficient
3	OverallQual	0.2208
7	GrLivArea	0.1854
4	OverallCond	0.1136
14	GarageArea	0.0909
17	MSZoning_RL	0.0863

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?

Ans:

A robust and generalizable model is one that performs well on both training and test data, it will not show high sensitive to noise on training data and predict accurately on unseen data.

The above can be achieved from the following,

Regularization:

- Regularization helps to prevent over fitting with the help of Lasso and Ridge Regression.
- Which are L1 (Lasso) and L2 (Ridge) regularization techniques, which will penalizes the complex models and helps avoid fitting the noise in the training data.

Cross-Validation:

- Use techniques like CV-fold, K-fold to know how the model performs on different data.
- It helps to detect over fitting and help model to perform well on unseen data.

Train-Test Split:

- Split the data into train and test, train the model with the training set and evaluate the model on the testing set.
- Observe how well the model performs on the test data set.

Feature Scaling:

- Feature scaling is a technique to standardize the independent features present in the data in a fixed range.
- It is to handle highly varying magnitudes or values which are present in the data.