Sunrise housing Assignment: Abhinav Joshi

abhinavjoshi7891@gmail.com

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?

As per our model:
The optimal value of alpha for ridge came out to be 50
The optimal value of alpha for lasso came out to be .001

When we doubled the value of alpha for

both ridge and lasso we observed some changes in the model metrics and Coefficients

For, the most Important predictor variables after the change was implemented, some came out to be the same as before but With slightly different coefficient values ,some new predictors were added, for some the importance reduced and for some the importance increased.

Top 10 Features for Ridge(alpha=50)

	Features	Coefficient
14	OverallQual	0.0749
7	GrLivArea	0.0571
5	TotalBsmtSF	0.0421
26	MSZoning_RL	0.0416
8	GarageArea	0.0316
40	Neighborhood_Crawfor	0.0282
27	MSZoning_RM	0.0270
154	SaleCondition_Partial	0.0262
153	SaleCondition_Normal	0.0233
73	Exterior1st_BrkFace	0.0224

Metrics of Ridge model with alpha = 50

r2_train: 0.9194710981783326
r2_test: 0.8848017375950764
RSS_train: 12.706289918804082
RSS_test: 8.21261424771026
MSE_train: 0.012568041462714225

MSE_test: 0.01892307430347986

Top 10 Features for Ridge(alpha = 100)

	Features	Coefficient
14	OverallQual	0.0700
7	GrLivArea	0.0524
5	TotalBsmtSF	0.0403
8	GarageArea	0.0289
26	MSZoning_RL	0.0267
40	Neighborhood_Crawfor	0.0263
154	SaleCondition_Partial	0.0237
3	BsmtFinSF1	0.0234
153	SaleCondition_Normal	0.0203
50	Neighborhood_NridgHt	0.0190

Metrics of Ridge model with alpha =

100
r2_train100: 0.9167835987876092
r2_test100: 0.8837851284272086
RSS_train100: 12.706289918804082
RSS_test100: 8.21261424771026
MSE_train100:
0.012568041462714225
MSE_test100: 0.01892307430347986

**Top 10 Features for Lasso(alpha =.
001)**

	Features	Coefficient
14	OverallQual	0.0872
26	MSZoning_RL	0.0854
7	GrLivArea	0.0758
27	MSZoning_RM	0.0592
5	TotalBsmtSF	0.0412
24	MSZoning_FV	0.0392
8	GarageArea	0.0361
40	Neighborhood_Crawfor	0.0267
154	SaleCondition_Partial	0.0265
153	SaleCondition_Normal	0.0222

Metrics of Lasso model with alpha =.001

r2_train: 0.9198393878567181
r2_test: 0.886297338325777
RSS_train: 12.648179162022217
RSS_test: 8.105991182280132
MSE_train: 0.012510562969359264

MSE_test: 0.018677399037511824

Top 10 Features for Lasso(alpha =. 002)

	Features	Coefficient
14	OverallQual	0.0950
7	GrLivArea	0.0848
5	TotalBsmtSF	0.0399
8	GarageArea	0.0350
154	SaleCondition_Partial	0.0260
40	Neighborhood_Crawfor	0.0241
3	BsmtFinSF1	0.0219
153	SaleCondition_Normal	0.0203
26	MSZoning_RL	0.0183
55	Neighborhood_Somerst	0.0173

Metrics of Lasso Model with alpha =. 002

r2_train2: 0.9130180993368373

r2_test2: 0.8823198602610391
RSS_train2: 13.72447931752856
RSS_test2: 8.389550086229576
MSE_train2: 0.013575152638505006
MSE_test2: 0.019330760567349253

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?
Both the models had almost the same performance in terms of the metrics that we chose to evaluate the models as shown below:

Metrics of Ridge model with alpha = 50
r2_train: 0.9194710981783326
r2_test: 0.8848017375950764
RSS_train: 12.706289918804082
RSS_test: 8.21261424771026

MSE_train: 0.012568041462714225
MSE_test: 0.01892307430347986

Metrics of Lasso model with alpha =. 001
r2_train: 0.9198393878567181
r2_test: 0.886297338325777
RSS_train: 12.648179162022217
RSS_test: 8.105991182280132
MSE_train: 0.012510562969359264
MSE_test: 0.018677399037511824

We will apply lasso regression for the following reasons:
A) Occam’s Razor:
When in doubt choose the simpler model.
Lasso produced a model which was simpler as lasso performed variable selection and removed unnecessary noise. We could afford this as the number of predictor variables was high.

B) Only a few among the predictors had

a significant impact on the response variable as we saw during the EDA by making the heatmaps

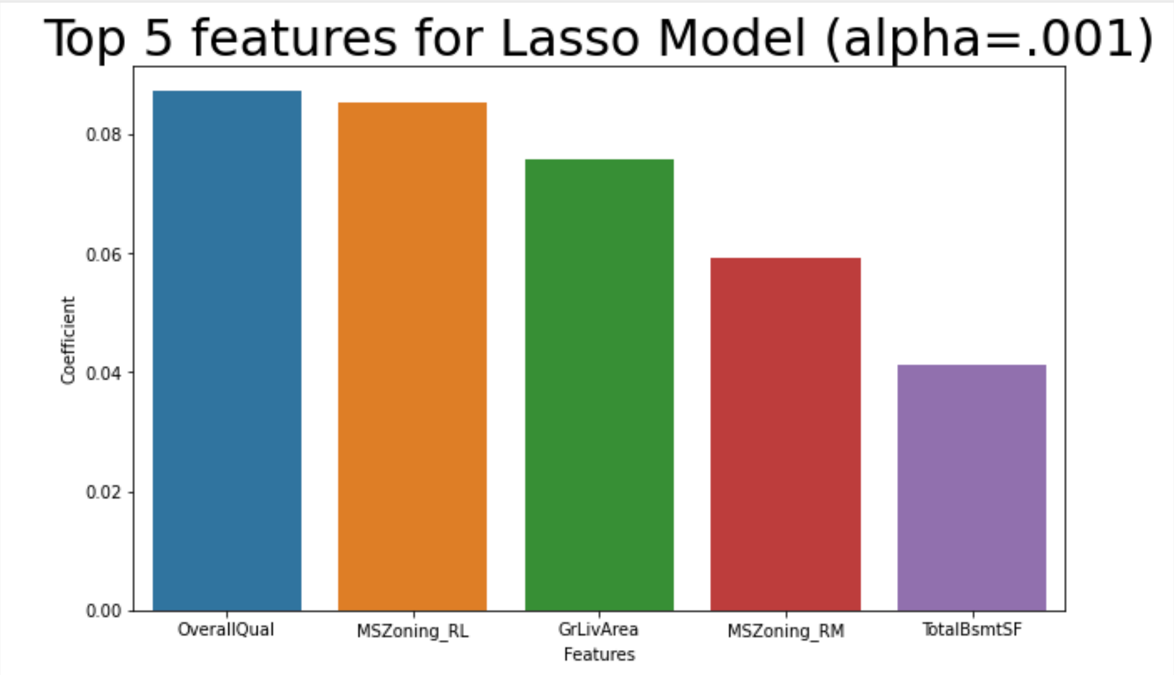
C) The numbers of variables was high and even though we eliminated some columns, the chance of noisy variables being still being in the dataset was high.

D) Because of the feature selection it is easier to interpret the model generated by lasso as compared to that generated by ridge

Q3. 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?

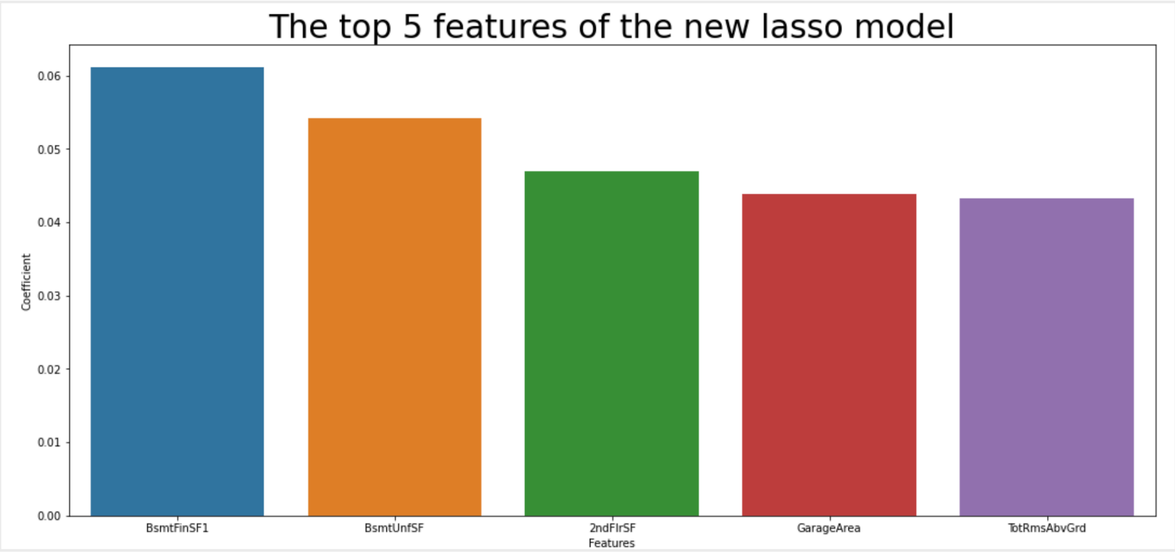
In the original lasso model, the five most important predictor variables were:

	Features	Coefficient
14	OverallQual	0.0872
26	MSZoning_RL	0.0854
7	GrLivArea	0.0758
27	MSZoning_RM	0.0592
5	TotalBsmtSF	0.0412



After removing these five most important predictor variables ,we created a new model. And for this new model, the five most important predictor variables are:

	Features	Coefficient
3	BsmtFinSF1	0.0612
4	BsmtUnfSF	0.0542
5	2ndFlrSF	0.0470
6	GarageArea	0.0438
16	TotRmsAbvGrd	0.0433



Q4. 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?
To make sure that a model is robust and generalisable, we simply follow the principle of Occam’s Razor which states that a model should be as simple as possible but not too simple and the model should be robust.

A model being robust means that it is relatively less sensitive to the changes in the input data I.e. when we make changes to the input dataset, the model’s algorithm is less likely to get affected.
In terms of accuracy, a robust model will have lower accuracy on the training dataset when compared to a more complex model. But for the test set and

the real world ‘unseen’ data, the robust model will have higher accuracy I n the long term over a complex model.

A Model is more generalisable when it is simpler and applicable to a large variety of data that it can possibly encounter.
In terms of accuracy, more generalisable model is bound to have a higher accuracy on ‘unseen’ data as against the a more complex model.
A complex model makes too many assumptions about the ‘unseen’ data and assumptions over unseen things are very-very likely to be wrong.