Problem Statement – II

Name: Saaigoutham A

Email: saaigoutham@gmail.com

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 to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Solution:

Optimal Value of Alpha:

Ridge Regression: 0.8

Lasso Regression: 0.001

i) Ridge Regression:

Changes in Model after doubled alpha rate:

Ridge : 0.8

- Regression Model Ridge(alpha=0.8) :

For Train Set:

- R2 score: 0.9174107805492033
- MSE score: 0.08258921945079668
- RMSE score: 0.2873834014879716

For Test Set:

- R2 score: 0.8687753882586958
- MSE score: 0.12714453067688283
- RMSE score: 0.3565733173933277

```
Ridge Regression Model with doubled alpha rate (alpha=1.6):

For Train Set:
R2 score: 0.9165677266249391
MSE score: 0.08343227337506087
RMSE score: 0.2888464529383404

For Test Set:
R2 score: 0.869602529301643
MSE score: 0.12634310738963833
RMSE score: 0.3554477562028467
```

Most Important Variables after change is implemented:

	Features	Coefficient	Absolute value
0	Exterior2nd_Brk Cmn	-0.4457	0.4457
1	Exterior1st_BrkComm	-0.4457	0.4457
2	Neighborhood_MeadowV	-0.4039	0.4039
3	2ndFlrSF	0.3849	0.3849
4	Neighborhood_Crawfor	0.3527	0.3527
5	BldgType_Duplex	-0.3393	0.3393
6	SaleCondition_Partial	0.3191	0.3191
7	CentralAir_Y	0.3143	0.3143
8	1stFlrSF	0.3040	0.3040
9	KitchenQual_Fa	-0.2840	0.2840

Observations after doubling alpha rate

- The test accuracy of the new model increases
- Overall, the old model seems to perform better, since it has a good training and test score

(ii) Lasso Regression:

Lasso : 0.001

- Regression Model Lasso(alpha=0.001) :

For Train Set:

- R2 score: 0.9343218725091461
- MSE score: 0.06567812749085392
- RMSE score: 0.2562774424151566

For Test Set:

- R2 score: 0.8903239653895467
- MSE score: 0.10626594936731987
- RMSE score: 0.32598458455472995

```
Lasso Regression Model with doubled alpha rate (alpha=0.002):

For Train Set:
R2 score: 0.9280398141262559
MSE score: 0.07196018587374407
RMSE score: 0.26825395779698025

For Test Set:
R2 score: 0.8911927383138581
MSE score: 0.1054241886315814
RMSE score: 0.32469091245611015
```

Most Important Variables after change is implemented:

	Features	Coefficient	Absolute value
0	GrLivArea	0.3162	0.3162
1	Neighborhood_Crawfor	0.2935	0.2935
2	SaleCondition_Partial	0.2552	0.2552
3	Neighborhood_StoneBr	0.2260	0.2260
4	OverallQual	0.2008	0.2008
5	CentralAir_Y	0.1802	0.1802
6	Exterior1st_BrkFace	0.1751	0.1751
7	selling_age	-0.1725	0.1725
8	OverallCond	0.1433	0.1433
9	SaleCondition_Normal	0.1428	0.1428

Observations after doubling alpha rate:

- Both train and test accuracy decrease after doubling the alpha rate
- The initial model is the better model here also

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?

Solution:
Optimal Value of Alpha:

Ridge Regression: 0.8

Lasso Regression: 0.001

The R2 test score on the Lasso Regression Model is slightly better than that of Ridge Regression Model. Moreover, the training accuracy is slightly reduced; hence, making the model an optimal choice as it seems to perform better on the unseen data. (Image is attached)

Ridge : 0.8

- Regression Model Ridge(alpha=0.8) :

For Train Set:

- R2 score: 0.9174107805492033
- MSE score: 0.08258921945079668
- RMSE score: 0.2873834014879716

For Test Set:

- R2 score: 0.8687753882586958
- MSE score: 0.12714453067688283
- RMSE score: 0.3565733173933277

Lasso : 0.001

- Regression Model Lasso(alpha=0.001) :

For Train Set:

- R2 score: 0.9343218725091461
- MSE score: 0.06567812749085392
- RMSE score: 0.2562774424151566

For Test Set:

- R2 score: 0.8903239653895467
- MSE score: 0.10626594936731987
- RMSE score: 0.32598458455472995

The MSE for Test set (Lasso Regression) is slightly lower than that of the Ridge Regression Model; implies Lasso Regression performs better on the unseen test data. Also,since Lasso helpsin feature selection (the coefficient values of some of the insignificant predictor variables became 0), implies Lasso Regression has a better edge over Ridge Regression. Therefore, the variables predicted by Lasso can be applied to choose significant variables for predicting the price of a house in this analysis.

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?

Solution: After Building the Initial lasso model the top 5 features are shown in the image below, Top 5 Features Dropped:

	Features	Coefficient	Absolute value
0	Exterior1st_BrkComm	-0.4459	0.4459
1	Neighborhood_Crawfor	0.3277	0.3277
2	Neighborhood_StoneBr	0.3267	0.3267
3	GrLivArea	0.3098	0.3098
4	SaleCondition_Partial	0.2637	0.2637

The above features were dropped, and new model was built, now the new top 5 potential features and their coefficients are,

	Features	Coefficient	Absolute value
0	Exterior2nd_Brk Cmn	-0.4565	0.4565
1	2ndFlrSF	0.2944	0.2944
2	Neighborhood_Edwards	-0.2466	0.2466
3	SaleType_New	0.2396	0.2396
4	1stFlrSF	0.2376	0.2376

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?
Solution:

Robustness of a model implies, the testing error of the model is consistent with the training error, the model performs well with enough stability even after adding some noise to the dataset. Thus, the robustness (or generalizability) of a model is a measure of its successful application to data sets other than the one used for training and testing.

By the implementing regularization techniques, we can control the trade-off between model complexity and bias which is directly connected the robustness of the model. Regularization helpsin penalizing the coefficientsfor making the model too complex; thereby allowing only the optimal amount of complexity to the model. It helps in controlling the robustness of the model by making the model optimal simpler. Therefore, to make the model more robust and generalizable, one need to make sure that there is a delicate balance between keeping the model simple and not making it too naive to be of any use. Also, making a model simple lead to Bias Variance Trade-off:

A complex model will need to change for every little change in the dataset and hence is very unstable and extremely sensitive to any changes in the training data. • A simpler model that abstracts out some pattern followed by the data points given is unlikely to change wildly even if more points are added or removed.

Bias helps you quantify, how accurate is the model likely to be on test data. A complex model can do an accurate job prediction provided there has to be enough training data. Models that are too naïve, for e.g., one that gives same results for all test inputs and makes no discrimination whatsoever has a very large bias as its expected error across all test inputs are very high. Variance is the degree of changes in the model itself with respect to changes in the training data. Thus, accuracy of the model can be maintained by keeping the balance between Bias and Variance as it minimizes the total error as shown in the below graph.

Thus, accuracy and robustness may be at the odds to each other as too much accurate model can be prey to over fitting hence it can be too much accurate on train data but fails when it faces the actual data or vice versa.