Problem Statement- Part II

Submitted by: Aarushi Gupta

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 1

The optimal value of alpha for Ridge regression is 4.0 and the optimal value of alpha for Lasso regression is 100.

If we double the alpha value to 8.0 and 200 for Ridge and Lasso regression respectively, then For Lasso regression, the R-squared value decreases from 90.3% to 89.7% and for Ridge Regression, The R-squared value decreases from 90% to 89.9% on test data.

The important predictor variables after the change are implemented become:

	Linear	Ridge	Lasso
GrLivArea	1.273635e+18	65103.657055	233296.735635
OverallQual	6.715052e+04	58388.817247	80023.979172
BsmtFinSF1	-2.996471e+17	48877.901936	52318.089723
RoofMatl_WdShngl	7.519592e+04	31583.426143	30753.185003
ExterQual	1.164802e+04	29673.566771	30123.659416
Neighborhood_NridgHt	2.382080e+04	23676.825341	28876.990094
SaleType_New	2.762793e+04	16690.400145	27011.528607
Neighborhood_NoRidge	7.030516e+03	28558.018904	26483.757315
TotalBsmtSF	4.390624e+17	38699.020337	25843.276714
KitchenQual	1.607651e+04	27084.064601	24469.027287
GarageCars	2.377816e+04	24184.923822	23625.679267
BsmtExposure	1.825290e+04	23720.276633	21533.298047
LotArea	4.442043e+04	27688.430487	20720.553847

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 2

The r2_score of Lasso regression model is slightly higher than Ridge regression model for the test dataset so we will choose lasso regression to solve this problem.

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.280320e-01	9.173636e-01	9.124645e-01
1	R2 Score (Test)	-1.032643e+22	9.008204e-01	9.036439e-01
2	RSS (Train)	4.515484e+11	5.184850e+11	5.492233e+11
3	RSS (Test)	2.707771e+34	2.600663e+11	2.526626e+11
4	MSE (Train)	2.113375e+04	2.264605e+04	2.330767e+04
5	MSE (Test)	7.898799e+15	2.447920e+04	2.412824e+04

In case of large number of features present in the dataset, both ridge regression and Lasso find their respective advantages. Ridge regression does not completely eliminate (bring to zero) the coefficients in the model whereas lasso does this along with automatic variable selection for the model. Here also, there are more than 200 features, thus lasso regression is preferred since it gives slightly better results than ridge 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?

Answer 3

The five most important variables in the lasso model including all the important predictor variables are:

	Lasso
GrLivArea	233296.735635
OverallQual	80023.979172
BsmtFinSF1	52318.089723
RoofMatl_WdShngl	30753.185003
ExterQual	30123.659416

After excluding the important predictor variables, the five most important predictor now becomes:

	Lasso
1stFlrSF	191023.141619
2ndFlrSF	139291.334475
TotalBsmtSF	108906.786518
OverallCond	40208.977145
PoolArea	38708.585185

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?

Answer 4

The model should be as simple as possible, though the accuracy will be compromised but it will be more robust and generalisable. The good model understands Bias-Variance trade-off. The simpler model has more bias but less variance and is more generalizable. Also, it will be more generalisable. However, complex model has high variance and low bias but such model leads to overfitting of data and cannot be considered as robust and generalisable model.

The model should be generalized so that the test accuracy is not very much lesser than the training score. In this dataset too much importance should not be given to the outliers since it might result in loss of useful information. Robust model implied that it should perform well on extrapolation or in more simpler words it should perform well on unseen data also so that the model can be trusted for predictive analysis.