Q&A

Problem Statement - Part II

Assignment Part-II

The following questions are the second part of the graded assignment. Please submit the answers in one PDF file. For writing normal text, please use MS Word (or similar software that can convert documents to PDF). For equations and figures, you can write/draw them on a blank sheet of paper using a pen, click images and upload them in the same Word document.

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?

The optimal lambda value in case of Ridge and Lasso is as below:

- Ridge - 4
- Lasso - 0.0001

Most important variables after value of alpha for both Ridge and Lasso is doubled:

Metric	Ridge Regression	Lasso Regression	Ridge2 Regression	Lasso2 Regression
R2 Score (Train)	0.953467	0.956628	0.949605	0.952638
R2 Score (Test)	0.888607	0.885036	0.886147	0.885712
SSE (Train)	5.792917	5.3993968	6.273654	5.896091
SSE (Test)	7.099186	7.326771	7.255946	7.283685
MSE (Train)	0.006605	0.006157	0.007154	0.006723
MSE (Test)	0.018881	0.019486	0.019298	0.019372

Performance of model diminishes when the optimum value of alpha is doubled. R2 score is reduced whereas SSE and MSE have increased.

Ridge alpha = 8 - Top Five Predictors

index	Ridge2
Neighborhood_Crawfor	0.081568
OverallCond_9	0.079698
OverallQual_9	0.077819
OverallCond_8	0.073737
Exterior1st_BrkFace	0.068517

Lasso , alpha = 0.0002 – Top Five Predictors.

index	Lasso2
OverallQual_9	0.140863
OverallQual_10	0.127948
OverallCond_9	0.123125
Neighborhood_Crawfor	0.117366
Neighborhood_StoneBr	0.104996

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?

The optimal lambda value in case of Ridge and Lasso is as below:

- Ridge - 4
- Lasso - 0.0001

Most important variables after value of alpha for both Ridge and Lasso

Metric	Ridge Regression	Lasso Regression	Ridge2 Regression	Lasso2 Regression
R2 Score (Train)	0.953467	0.956628	0.949605	0.952638
R2 Score (Test)	0.888607	0.885036	0.886147	0.885712
SSE (Train)	5.792917	5.3993968	6.273654	5.896091
SSE (Test)	7.099186	7.326771	7.255946	7.283685
MSE (Train)	0.006605	0.006157	0.007154	0.006723
MSE (Test)	0.018881	0.019486	0.019298	0.019372

Ridge shows less variability between test and train. Variation in Ridge is less compared to lasso. So Ridge will be applied to build the model.

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?

After removing the best five predictor variables of Lasso are as follows:

Top Five Predictors	Lasso5
SaleType_Oth	0.159851
Neighborhood_StoneBr	0.122142
Neighborhood_Crawfor	0.110341
RoofMatl_Tar&Grv	0.098209
MSZoning_FV	0.094939

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?

As Per, Occam's Razor— given two models that show similar 'performance' in the finite training or test data, we should pick the one that makes fewer on the test data due to following reasons:-

- Simpler models are usually more 'generic' and are more widely applicable
- Simpler models require fewer training samples for effective training than the more complex ones and hence are easier to train.
- Simpler models are more robust.
    - Complex models tend to change wildly with changes in the training data set
    - Simple models have low variance, high bias and complex models have low bias, high variance
- Simpler models make more errors in the training set. Complex models lead to over fitting — they work very well for the training samples, fail miserably when applied to other test samples

Therefore to make the model more robust and generalizable, make the model simple but not too simple which will not be of any use.
Regularization can be used to make the model simpler. Regularization helps to strike the delicate balance between keeping the model simple and not making it too naive to be of any use. For regression, regularization involves adding a regularization term to the cost that adds up the absolute values or the squares of the parameters of the model.
Also, Making a model simple leads 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 quantifies how accurate is the model likely to be on test data. Models that are too naïve, for e.g., one that gives same answer to all test inputs and makes no discrimination whatsoever has a very large bias as its expected error across all test inputs are very high.
A complex model can do an accurate job prediction provided there is enough training data. Variance refers to the degree of changes in the model itself with respect to changes in the training data.