Homework 1 CSCE 421

Arya Rahmanian
Department of Computer Science
Texas A&M University
College Station
aryarahmanian@tamu.edu

February 19, 2023

1

The data is just a really long csv with two columns, x and y, and hundreds of rows of values.

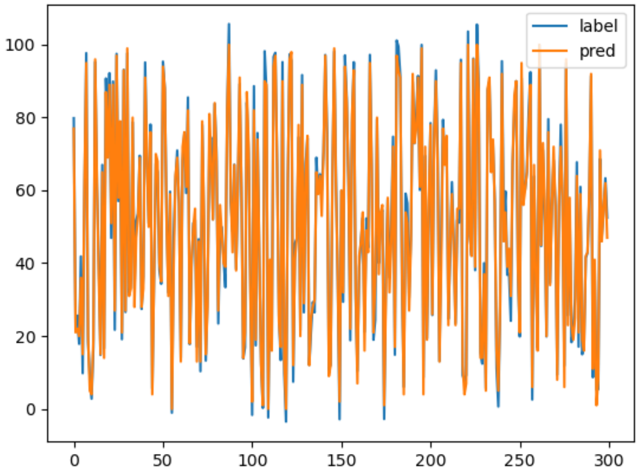

Figure 1: The figure above is an overlay of the original data with my prediction plot over it.

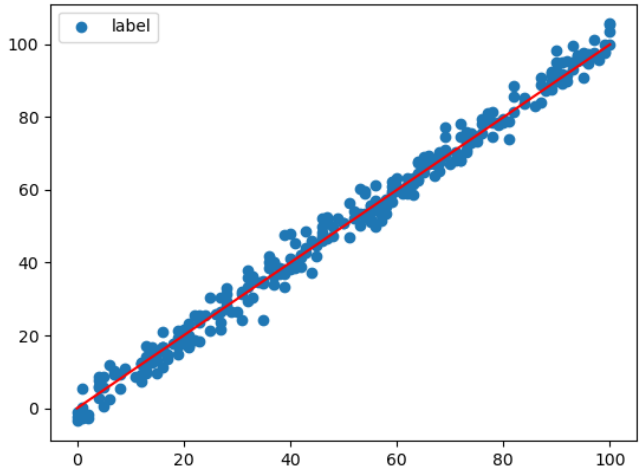

Figure 2: My line of best fit for the training x and y data points

2

The data for this part of the assignment has several columns (features) with several different stats for each baseball player. We are using NewLeague as the label for this data set. For this question we are just prepossessing the data and will make models in question 3.

3

The coefficients are different for the linear and logistic model. This is because linear and logistic are different models and SHOULD produce different coefficients. They have different objectives when training.
When I made my models I got the following measurements

	Linear	Log
AUC	0.9560	0.9692
Optimal Thresholds	0.11195	0.71061

To find the "optimal threshold" we want a high tpr and low fpr, this corresponds to the top left of the ROC curve. So I found the threshold that yields the highest tpr and lowest fpr.

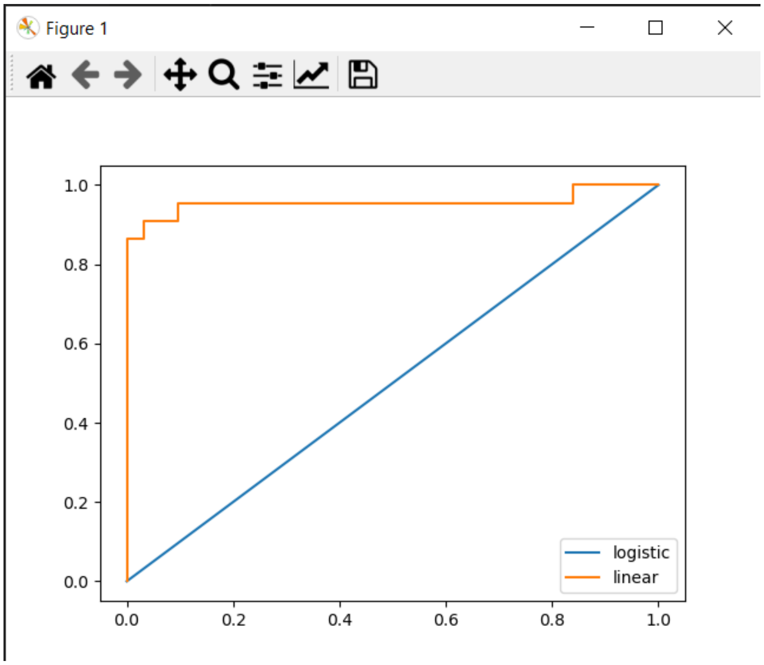

Figure 3: The figure above is the ROC curve for both linear and log models.

Then I did 5 fold cross validation, and the number of features don't change with each fold because it's divided up evenly into n segments.

	Mean-h	Mean	Mean + h
Linear AUC	0.86215	0.926978	0.9918
Log AUC	0.8580	0.9278	0.9976
Linear F1	0.8591	0.9269	0.9947
Log F1	0.8351	0.9106	0.986176