Faith Meyer
Lab 6

Logistic Regression vs Naive Bayes

I picked the data sets Hitters from the ISLR package from the options of the data sets preloaded into R Studio. I chose this data set because I am a sports fan and when I looked through the data set it looked like something fun to play around with. While I was looking through the data I thought it would be fun to predict the division each player played in. I predicted that based on the number of years they played the sport. Since these two variables don't have anything to do with each other, the best classification algorithm for me to use would be Naive Bayes. Here I will tell you the difference between Naive Bayes and Logistic Regression, and why I chose to use one over the other.

The data set I used would benefit from using Naive Bayes instead of Logistic Regression. This is because when predicting the division the player is in has nothing to do with the amount of years they have been playing. They are conditionally independent, and that is what Naive Bayes is useful for when classifying data. This means that each data point is treated independently and point A is targeted towards point B so it can explain that if A happens whenever B occurs. This is different from Logistic Regression since that is more of a linear classification. The points are split up in a linear way and they are mapped directly to one another showing if one event happened at a certain time because the other event happened. Also, the data set is not huge, so that is another reason to use Naive Bayes. Another key difference between the two classifications is the size of the data set that it will be successful with. Naive Bayes is better for smaller data sets so you can focus on features as individuals. Logistic Regression is more successful with larger data sets because it tends to overfit the data in the plot making it more generalized.

When trying to figure out what type of classification algorithm you want to use, you need to look at multiple important factors. These factors include things such as the size of the data set, what you are trying to do with the data set, and if the variables you are working with are independent of each other or do they have some sort of correlation to each other. With the Hitters data set, Naive Bayes is the better choice for the prediction of Division based on how many years a player played baseball. This is because each aspect of the data is independent of each other. The data set is also a smaller size, which Naive Bayes is better at dealing with.

Sources:

https://mgimond.github.io/Stats-in-R/Logistic.html to make the R graph for Logistic Regression
https://www.andreaperlato.com/mlpost/naive-bayes-classification/ To make R graph for Naive Bayes
Difference between logistic regression and Naive Bayes:
https://www.educba.com/naive-bayes-vs-logistic-regression/
https://dataespresso.com/en/2017/10/24/comparison-between-naive-bayes-and-logistic-regression/