PREDICT BLOOD DONATION

Sabrina Gouveia

DATA SCIENCE General Assembly - 2016

INTRODUCTION

Blood donation has been around for a long time. The first successful recorded transfusion was between two dogs in 1665, and the first medical use of human blood in a transfusion occurred in 1818. Even today, donated blood remains a critical resource during emergencies.

What data are you planning to use to answer that question?

The dataset is from a mobile blood donation vehicle in Taiwan. The Blood Transfusion Service Center drives to different universities and collects blood as part of a blood drive. The goal is to predict whether or not a donor will give blood the next time the vehicle comes to campus.

What do you know about the data so far?

Data is courtesy of Yeh, I-Cheng via the UCI Machine Learning repository: "Knowledge discovery on RFM model using Bernoulli sequence".

The data set *bloodonation* includes five (5) variables, describing the blood donation history of 776 individuals.

In order to prepare the blood donation data for analysis and model building, I combined the original "Test Data – Blood Donation" and "Train Data – Blood Donation" into a single file so that the test and training data could be randomly split into training and testing data in Python. The files were also combined in order to impute values for the response variable `don_03_2007` since the test data did not include any values for this variable: having classification values that we know to be true in the test data is essential in leveraging the test data to assess the predictive power of a model since I will essentially compare how well models correctly predict classification values relative to the true classification. New binary values for `don_03_2007` were inserted into the .csv file using the Excel random number generator, drawing from a Bernoulli distribution (p = 0.4).

Next, the `tv_don` values were replaced with new variables so that the values were not perfectly collinear with the `dons_n` values. Finally, the data set was sorted on descending or ascending values of the different explanatory variables and the `don_03_2007` values were changed to 0 or 1 so that the response variable covaried with the explanatory variables in an intuitive way.

Finally, the headers were removed from the .csv file so that Python could easily interpret the data.

OBJECTIVE

Train several machine learning (ML) models to estimate the probability that an individual donated blood in March 2007 using the *bloodonation* data set, compare these predictive power these models and select the best performing model. In subsequent sections of this analysis, I discuss how to:

- Clean the data
- Perform exploratory data analysis
- Split the data into training and test subsets
- Fit predictive models with a binary response variable, using the following ML algorithms:
    - k-Nearest-Neighbors (kNN) with k-fold cross validation
    - Logistic regression
    - Random forest classifier
- Make predictions with these models
- Test and compare the predictive power of each model

EXPERIMENTAL PROCEDURE

Load data set

I begin by setting the working directory, importing libraries that I will need throughout this analysis and importing the *bloodonation* data set as a pandas DataFrame, which I will call *bd*.