PREDICT BLOOD DONATION

Sabrina Gouveia

DATA SCIENCE General Assembly - 2016

“Can we predict whether a donor will return to donate blood given their donation history”?

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.**

DATA

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.

- Combined data into a single file so that Test & Training data could be randomly split in Python

- Files 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

- `tv_don` (Total volume donated) values were replaced with new variables so that the values were not perfectly collinear with the `dons_n` values.

- 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.

PROCESS

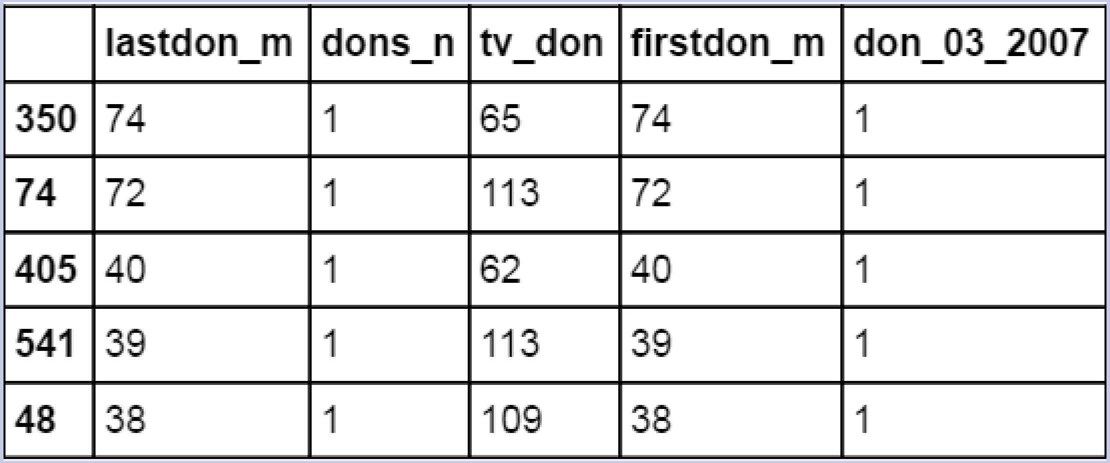
- Clean the data
- Perform exploratory data analysis
- Split the data into training and test subsets
- Fit predictive models with a binary response variable, using the following ML algorithms:
  - k-Nearest-Neighbors (kNN) with k-fold cross validation
  - Logistic regression
  - Random forest classifier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

EXPERIMENTAL PROCEDURE

Load data set

- Set working directory
- Import libraries
- Import ‘bloodonation data set” as a pandas DataFrame (called *bd*)
- Display first (5) rows of Dataframe ‘bd’

bd.head(n = 5)



	lastdon_m	dons_n	tv_don	firstdon_m	don_03_2007
350	74	1	65	74	1
74	72	1	113	72	1
405	40	1	62	40	1
541	39	1	113	39	1
48	38	1	109	38	1

DATA CLEANING

- Check NULL values in the data

bd.isnull().sum()

EXPLORATORY DATA ANALYSIS

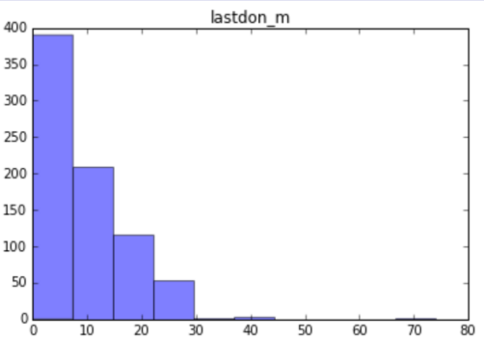
- Count number of rows in the data

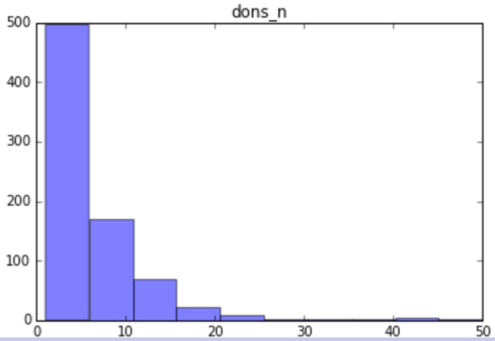
```
bd.count()
```

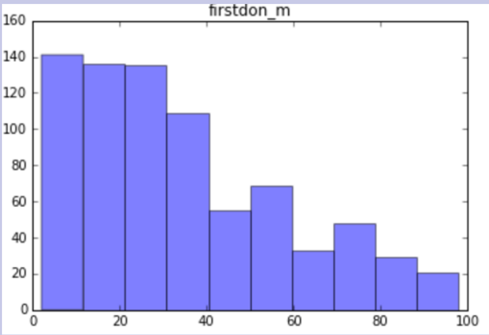
- Count number of observations w/ positive outcome

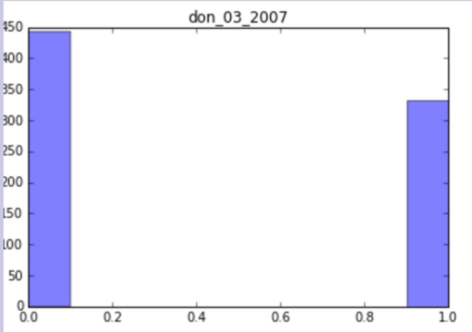
```
len(bd[don_03_2007 == 1]

100*(len(bd[don_03_2007 == 1]) / bd.count()[0]
```

- Plot Histograms to describe the distribution of each variables (Code Here)









There are several observations in *lastdon_m*, *dons_n* and *tv_don* that may be outliers or influential points.