Summer Data Scientist Data Assessment

Crime and Education Lab New York

This exercise is intended to help us see how you approach a standard data processing and statistical analysis task, and for you to demonstrate your thought process and workflow. The scenario described in this task is fictional, but it is indicative of the type of work we do at Crime and Education Lab New York.

Logistically, you should spend no more than 3 hours on this task. The task must be completed in either R or python, but feel free to submit your work in whatever format you prefer (i.e., scripts + separate write-up, or a mixed format like R markdown or Jupyter notebook). This task should be entirely your own work. This task may be used as the basis for discussion during a subsequent phone interview, where we'll ask about the approach that you used and about related analytics concepts.

Please email completed tasks to Sibella Matthews (sibellamatthews@uchicago.edu). If you have any questions about the task, please email Lucie Parker (laparker@uchicago.edu) and Melissa McNeill (mmcneill@uchicago.edu).

Background

Back in January 2010, the NYC District Attorney's (DA's) Office implemented a program designed to reduce felony re-arrest rates city wide. That is, individuals arrested post-implementation were offered an intervention on the spot, with the hopes of reducing their chance of getting re-arrested for a felony at some point in the future. As their trusted data partner, the DA's Office has asked you to help them study the program and its rollout.

Data

There are two main datasets, *arrests.csv* and *demo.csv*. The *arrests.csv* file contains information on each arrest made from 2008 through 2011, and includes an arrest ID, a person ID, the date of arrest, and whether the arrest was for a misdemeanor or felony crime. The *demo.csv* file contains demographic data on each person found in the arrest file, including birthdate, gender, and home precinct.

Analysis

Part 1: Variable Creation

In this section, you'll create an analysis dataset allowing you to look more closely at arrestees impacted by the program.

1. For the arrests that occurred post-implementation, create the following covariates:
    - Age
    - Gender
    - Home precinct
    - Law code (misdemeanor or felony)
    - Number of prior misdemeanor arrests (in the last 2 years)
    - Number of prior felony arrests (in the last 2 years)
    - Number of prior misdemeanor arrests (in the last 6 months)
    - Number of prior felony arrests (in the last 6 months)
2. Generate a binary outcome that measures any felony re-arrest in a 1-year period following the arrest.

Part 2: Statistical Analysis

Please select only **one** of the following tasks to complete - program evaluation or predictive model.

Program Evaluation

In order to study the impact of the program, the DA's Office had randomly selected some precincts to receive the treatment (i.e. intervention). The other precincts serve as control precincts. All individuals who live in a treatment precinct that were arrested post-implementation received the treatment. The treatment assignments can be found in *treatment_assignment.csv*.

1. We're only interested in measuring the effect of the program for the first time an individual receives treatment. Limit your analysis dataset from Part 1 to the first post-implementation arrest for each individual.
2. Did the program significantly reduce felony re-arrest in a 12-month follow-up? Is your result robust to covariate inclusion?

OR

Predictive Model

The DA's Office would like to continue this program, but due to budget constraints, they'd like to offer it only to arrested individuals *most at risk* of felony re-arrest. They've asked you to determine, using this historical data, whether a predictive model could be used to reliably identify at-risk individuals.

1. Since individuals in the treatment precincts have already received the program, which might affect the felony re-arrest rate observed in the data, limit your analysis dataset from Part 1 to exclude the treatment precincts.
2. Build and evaluate a model* predicting felony re-arrest in a 12-month follow-up period.
3. Which features seem to be driving your model? What seems to matter more, recent history (last 6 months) or fuller history (last 2 years)?

*Note: This isn't a model performance competition – we're just trying to see how you think about modeling. We understand that you're working with a very limited feature set here.