Capstone

Overview and first tasks

As a near-graduate of the Academy, you have been hired by the consultancy *Awkward Problem Solutions*™, that solves the tough data science problems that no one else will touch.

The consultancy has accepted a contract by the police department of your city. The police department has received lots of complaints about its stop and search policy. Every time a car is stopped, the police officers have to decide whether or not to search the car for contraband. According to critics, these searches have a bias against people of certain backgrounds.

Your company has been hired to (1) determine whether these criticisms seem to be substantiated, and (2) create a service to fairly decide whether or not to search a car, based on objective data. This service will be used by police officers to request authorization to search, and your service will return a Yes or No answer.

The police department has asked for the following requirements:

1. A minimum 50% success rate for searches (when a car is searched, it should be at least 50% likely that contraband is found)

2. No police sub-department should have a discrepancy bigger than 5% between the search success rate between protected classes (race, ethnicity, gender)

3. The largest possible amount of contraband found, given the constraints above.

The police department has collected a few years of data about the car stops, including whether the car was searched, and if whether any contraband was found.

- The training set is here
- The data dictionary is here

Your first objective is to produce a report about the historical data, including what seems to indicate contraband, potential sources of descrimination in searches, and proposing potential actions. Your report will include two parts:

1. A report for the Police department
2. A technical report for your boss at the consultancy (with your technical analysis)

Note: this will get updated with the grading criteria to include the model report.

A few notes from your instructors

This problem has a number of characteristics that problems in the real world will have.

For instance:

1. The dataset may contain things you do not need

2. The requirements from the client may contain some level of ambiguity

3. The requirements from the client may not be completely achievable, and you may need to explain trade-offs and recommend solutions

4. The modelling problem may contain difficult yet unavoidable ethical questions

5. The dataset may contain historical biases, which can affect your models and produce unexpected results.

During the capstone, assume the posture of a professional instead of a student. This means that your professional opinion may be to tell the client why certain aspects of the problem should be posed differently, or drawing their attention to unexpected consequences of the requirements.

In some cases, there will be no right answers, and your job will be to make the best you can with what you have.

During the next few weeks (see full calendar here) you will get more detailed instructions about grading criteria and what is expected at each section. For now it's best to start getting accustomed to the dataset, start exploring in depth and making a few toy models.

Good luck!