Sprawozdanie 3

Laboratorium Statystyczne 2

15 stycznia 2017

W Stanach Zjednoczonych jest ponad 2 mln więźniów. Koszty utrzymania systemu są bardzo wysokie i siegają 74 miliardów dolarów. Jednym z pomysłów na zmniejszenie kosztów jest wypuszczenie większej liczby więźniów, którzy nie stanowią bezpośredniego zagrożenia dla innych, na zwolnienie warunkowe (ang. *parole*).

Są jednak przesłanki aby uważać, że system, w którym decyzję o zwolnieniu warunkowym podejmują ludzie, jest wysoce nieefektywny i stronniczy. Opisuje to w swojej książce **Pułapki myślenia** Daniel Kahnemann. Odpowiendni fragment wraz z komentarzem znajduje się tutaj.

Twoim zadaniem jest zbudowanie modelu statystycznego, który na podstawie informacji dostępnych o osobie ubiegającej się o zwolenienie (ang. *parolee*), zdecyduje czy należy warunkowo go zwolnić czy nie. Dane znajdziesz w pliku parole.csv. Poniżej znajduje się opis danych:

- male = 1 if the parolee is male, 0 if female
- race = 1 if the parolee is white, 2 otherwise
- age = the parolee's age in years at the time of release from prison
- state = a code for the parolee's state. 2 is Kentucky, 3 is Louisiana, 4 is Virginia, and 1 is any other state. These three states were selected due to having a high representation in the dataset.
- time.served = the number of months the parolee served in prison (limited by the inclusion criteria to not exceed 6 months).
- max.sentence = the maximum sentence length for all charges, in months (limited by the inclusion criteria to not exceed 18 months).
- multiple.offenses = 1 if the parolee was incarcerated for multiple offenses, 0 otherwise.
- crime = a code for the parolee's main crime leading to incarceration. 2 is larceny, 3 is drug-related crime, 4 is driving-related crime, and 1 is any other crime.
- violator = 1 if the parolee violated the parole, and 0 if the parolee completed the parole without violation.

1. Dokonaj wstępnej eksploracji danych (boxploty, scatterploty).
2. Zbuduj prosty model regresji logistycznej, który przewidywać będzie p-stwo naruszenie zwolnienia warunkowego (ang. *parole violation*).
3. Uprość model poprzez pominięcie zmiennych, które nie są istotne ze względu na predykcję.
4. Podaj interpretację modelu (konkretnych współczynników).
5. Sposród wszystkich zbudowanych przez Ciebie modeli wybierz jeden i oceń jego dokładność.
6. Spróbuj porównać się do danych z książki Kahnemana. Czy domyślne próg $p = 0.5$ daje model bardziej konserwatywny czy bardziej liberalny niż decyzje sędziów? Jaki procent zwolnionych naruszy zwolnienie warunkowe jeśli procent zwolnionych będzie taki jak opisał Kahneman?
7. Dodatkowe: Jakie zalety ma stworzony przez Ciebie model w porównaniu do decyzji podejmowanych przez sędziów? Jakie są jego wady? W jaki sposób można by go wprowadzić w życie?

Termin wykonania zadania to 1 lutego (środa!). Należy przesłać pliki Rmd, pdf i prezentację na adres Piotr.Sobczyk@pwr.edu.pl.

Analizę należy wykonać samodzielnie. Próby oszustwa, w szczególności wszelkiej maści plagiaty, będą skutkować niezaliczeniem całego kursu.

Życzę powodzenia,

Piotr Sobczyk