Big Data's Disparate Impact

Maria Wyrzykowska, Adrian Urbański

November 3, 2020

Źródło

Prezentacja przygotowana została na podstawie pracy pod tytułem „Big Data's Disparate Impact" opublikowanej na łamach California Law Review, której autorami są Solon Barocas i Andrew D. Selbst.

Big Data's Disparate Impact

Solon Barocas* & Andrew D. Selbst**

Advocates of algorithmic techniques like data mining argue that these techniques eliminate human biases from the decision-making process. But an algorithm is only as good as the data it works with. Data is frequently imperfect in ways that allow these algorithms to inherit the prejudices of prior decision makers. In other cases, data may simply reflect the widespread biases that persist in society at large. In still others, data mining can discover surprisingly useful regularities that are really just preexisting patterns of exclusion and inequality. Unthinking reliance on data mining can deny historically disadvantaged and vulnerable groups full participation in society. Worse still, because the resulting discrimination is almost always an unintentional emergent property of the algorithm's use rather than a conscious choice by its programmers, it can be unusually hard to identify the source of the problem or to explain it to a court.

Spis treści

1. Problemy z Machine Learningiem
   - Trudne decyzje
   - Słabe dane
   - Celowa dyskryminacja
2. Prawo vs dyskryminacja
   - Źródła prawa przeciwdziałającego dyskryminacji
   - Disparate treatment
   - Disparate impact
3. Trudności przy reformach
   - Problemy wewnętrzne
   - Problemy zewnętrzne

1. Problemy z Machine Learningiem
   - Trudne decyzje
   - Słabe dane
   - Celowa dyskryminacja
2. Prawo vs dyskryminacja
   - Źródła prawa przeciwdziałającego dyskryminacji
   - Disparate treatment
   - Disparate impact
3. Trudności przy reformach
   - Problemy wewnętrzne
   - Problemy zewnętrzne

Określanie „Target Variable” i „Class Labels”

Co to znaczy że ktoś jest dobrym pracownikiem?

- Efektywność?
- Oceny przełożonych?
- Długość zatrudnienia?

Etykiety

Dane, na których uczy się nasz algorytm, mogą mieć przydzielone etykiety. Czasem już one dyskryminują.

- St George's, University of London
- LinkedIn

Bias danych

Czasami problemem jest polaryzacja danych.

- Niedostateczna reprezentacja
- Nadmierna reprezentacja

Feature selection

Feature selection

Nazwa procesu, podczas którego organizacje decydują których zmiennych używać do podejmowania decyzji.

Dane są drogie. Z tego powodu czasem bardziej opłaca się brać pod uwagę wyłącznie te dane, które łatwo jest zdobyć.

- Prestiżowe szkoły
- „Redlining“

Proxies

Informacja o przynależności do danej grupy, nawet jeśli nie jest podana wprost, najczęściej i tak zakodowana jest w innych zmiennych.

Masking

Każdy z poruszonych wcześniej problemów może zostać wykorzystany przez zdeterminowaną osobę do celowej dyskryminacji jakiejś grupy.

1. Problemy z Machine Learningiem
   - Trudne decyzje
   - Słabe dane
   - Celowa dyskryminacja
2. Prawo vs dyskryminacja
   - Źródła prawa przeciwdziałającego dyskryminacji
   - Disparate treatment
   - Disparate impact
3. Trudności przy reformach
   - Problemy wewnętrzne
   - Problemy zewnętrzne

Title VII

Title VII to akt praw cywilnych USA z 1964 roku, który "prohibits employment discrimination based on race, color, religion, sex and national origin". W tym celu klasyfikuje on sytuacje w których dochodzi do dyskryminacji na dwie kategorie:

- Disparate treatment ("odmienne traktowanie")
- Disparate impact ("odmienny wpływ"?)

Czym jest disparate treatment?

Intuicja: dyskryminacja wynika z intencji twórcy modelu

- formalna dyskryminacja - jedną ze zmiennych w danych wejściowych modelu jest informacja o należności do chronionej klasy
- celowa dyskryminacja - wprawdzie w danych nie mamy zmiennej "rasa", ale model celowo dyskryminuje (maskowanie/racjonalny rasizm)

Formalna dyskryminacja a proxies

W przypadku data miningu taka definicja nie ma sensu z powodu istnienia proxies.

- jeśli ktoś chce dyskryminować po rasie to może ją łatwo wywnioskować na podstawie innych zmiennych
- dodanie rasy do zmiennych pewnie miałoby niewielki wpływ na wyniki

Jak udowadnia się disparate treatment? - McDonell-Douglas

1. Pokrzywdzony pokazuje, że podobna osoba nienależąca do chronionej klasy nie zostałaby pokrzywdzona
2. Pracodawca oferuje "niedyskryminujący" powód swojej decyzji, nie musi udowadniać jego prawdziwości
3. Pokrzywdzony pokazuje, że ten powód jest tylko pretekstem

Jak udowadnia się disparate treatment? - Mixed-motive

Pokrzywdzony nie musi pokazywać że oferowany powód to pretekst, tylko że dyskryminacja była motywującym czynnikiem dla akcji podjętej przez pracodawcę.
Czym jest "motywujący czynnik"?

Disparate treatment - podsumowanie

Co podchodzi pod disparate treatment?

- model celowo stworzony tak, by dyskryminował
- model, który jest używany bo lubimy jak dyskryminuje

Co nie podchodzi pod disparate treatment?

- Street Bump, LinkedIn itp
- pracodawca wie, że dyskryminuje, ale nie jest to jego główny cel

Wniosek: disparate treatment, oprócz najbardziej oczywistych przypadków, ma problemy z regulowaniem politycznej poprawności.

Czym jest disparate impact?

Intuicja: dotyczy niecelowej dyskryminacji - występującej niezależnie od zamiaru twórcy modelu.

Jak udowadnia się disparate impact?

1. Pokrzywdzony pokazuje, że chroniona klasa jest inaczej traktowana
2. Pracodawca może się bronić, że jego akcja była związana z charakterystyką pracy, była koniecznością biznesową
3. Pokrzywdzony pokazuje, że istnieje alternatywne, mniej dyskryminujące rozwiązanie, którego pracodawca nie wykorzystał

Jakie są problemy?

Wszystko jest bardzo nieścisłe:

- O ile gorsza musi być sytuacja chronionej klasy?
- Czym i jak duża ma być "biznesowa konieczność"/bycie "związanym z charakterystyką pracy"?:
  - Czy target variable jest związana z pracą?
  - Czy model ją przewiduje?
  - Czy model ją dobrze przewiduje?
- Co to znaczy, że pracodawca nie wykorzystał mniej dyskryminującego rozwiązania?

1. Problemy z Machine Learningiem
   - Trudne decyzje
   - Słabe dane
   - Celowa dyskryminacja
2. Prawo vs dyskryminacja
   - Źródła prawa przeciwdziałającego dyskryminacji
   - Disparate treatment
   - Disparate impact
3. Trudności przy reformach
   - Problemy wewnętrzne
   - Problemy zewnętrzne

Określanie „Target Variable”

Ponieważ różni pracodawcy mają różne cele, niemożliwe jest jednoznaczne określenie co jest „dobrą“ zmienną. Czasami dyskryminacja jest oczywista, ale jest to rzadkość.

- Hayden v. County of Nassau

Dane - etykiety

Nie można zabronić korzystania z historycznych danych przy tworzeniu modelu, ale historyczne dane prawie zawsze będą dyskryminować. Co więcej, ponieważ część procesu rekrutacji jest celowo subiektywna, niemożliwe jest utworzenie żadnego obiektywnego zestawu zasad.

- Długość stażu pracy

Dane - zbieranie

Trudności z prawem które miałoby kontrolować jakość danych:

- Zauważenie ewentualnego problemu z danymi może być trudne
- Poprawienie go jeszcze trudniejsze
- Czy należy karać pracodawców, którzy tego nie zrobili?
- A co z tymi którzy próbowali, ale im się nie udało?

Feature selection

Bardzo trudno jest określić kiedy pracodawca dyskryminuje poprzez ograniczanie danych. Aby to zrobić, należy stwierdzić czy w ogóle można było uniknąć błędów, ponieważ uzyskanie lepszych, bardziej szczegółowych danych byłoby proste i nie wymagałoby poniesienia kosztów przekraczających zyski.

Proxies

Jak traktować dane, które wskazują na przynależność do jednej z chronionych grup?

- Usunąć je?
- Jak bardzo muszą być powiązane żeby w ogóle się nimi przejmować?

Większy problem: nawet po usunięciu wrażliwych danych, może się okazać, że jakieś grupy nadal są gorzej traktowane - ale tym razem, winne są inne zmienne.

Antyklasyfikacja i antysubordynacja

Istnieją dwie zasady stanowiące źródło inspiracji dla prawa przeciwdziałającego dyskryminacji:

- antyklasyfikacja - przynależność do chronionej grupy nigdy nie powinna być brana przy podejmowaniu decyzji dotyczących danej osoby (vs disparate treatment) - dominująca
- antysubordynacja - celem jest eliminacja wszelkich nierówności spowodowanych przynależnością do chronionej klasy (vs disparate impact)

Które wybrać?

Antyklasyfikacja i antysubordynacja często wzajemnie sobie przeczą.

- Jeśli chcemy zlikwidować nierówności wynikające z przynależności do chronionej grupy, musimy wiedzieć kto należy do tej grupy.
- Co jeśli wyniki dyskryminują, a nie możemy naprawić działania modelu?
- Co jeśli nie zatrudniamy żadnych kobiet, bo statystycznie krócej zostają one w pracy?

Im dalej jesteśmy od antysubordynacji, tym gorzej sobie radzimy z takimi problemami. Im bliżej - tym więcej pojawia się dalszych pytań.

Podsumowanie

Data mining może być zbyt "wrong" albo zbyt "right".
Żeby przygotować się do walki z dyskryminacją z tego wynikającą, musimy sobie najpierw odpowiedzieć na liczne pytania: co chcemy osiągnąć, jaki stopień dyskryminacji jest dopuszczalny, jak wiele strat ma ponieść pracodawca żeby jej zapobiegać?