Machine Bias:

quando un algoritmo* ha pregiudizi**

Elena Pesce

pesce@dima.unige.it

Terzo Ciclo

BIG DATA: USI E ABUSI

Prima di iniziare

*Algoritmo: procedimento che risolve un determinato problema attraverso un numero finito di passi elementari, chiari e non ambigui, in un tempo ragionevole. [Def. Wikipedia]

**Pregiudizio: atteggiamento sfavorevole od ostile, in partic. quando esso presenti, oltre che caratteri di superficialità e indebita generalizzazione, anche caratteristiche di rigidità, cioè quando implichi il rifiuto di metterne in dubbio la fondatezza e la resistenza a verificarne la pertinenza e la coerenza. [Def. Treccani]

Immaginate di essere accusati di aver commesso un crimine...

Preferireste che a giudicarvi fosse una persona o un algoritmo?

I FATTI

Valutazione del rischio

In alcuni Stati americani (Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia, Washington e Wisconsin) si usano algoritmi per valutare la *probabilità* di recidiva (*risk assessment of relapse*) per reati minori durante le udienze preliminari.

Questo è spesso fatto insieme a una valutazione delle necessità di riabilitazione.

Sottoporre a giudizio vs sottoporre a processo

Infatti *questi* algoritmi sono stati originariamente progettati come *sistemi di supporto alle decisioni* per dare ai giudici una *stima* degli *effetti* di trattamenti alternativi (trattamento farmacologico, consulenza per la salute mentale, ...) rispetto alla *reclusione carceraria.*

Alcuni dati

Gli USA sottopongono a giudizio moltissime persone, con una percentuale spropositata di afroamericani

Efficacia dei trattamenti alternativi:

- Risparmio fino a 18 milioni all'anno dal 2000 nello stato della California
- Calo della recidiva per chi ha completato un trattamento alternativo: il *tasso* di arresto è diminuito dell'85%, il *tasso* di condanna del 75%, il *tasso* di carcerazione dell'83%

http://www.apa.org/monitor/julaug03/alternatives.aspx

Perché usare questi algoritmi?

Quindi i dati dimostrano che si può ridurre la popolazione carceraria con un conseguente risparmio sui costi, ma si può ridurre anche la *probabilità* che un imputato ricommetta un crimine.

Se si potesse prevedere in modo *accurato* quali imputati hanno una maggiore probabilità di commettere nuovi crimini, il sistema giudiziario penale potrebbe essere più selettivo riguardo a chi incarcerare e a chi proporre un trattamento alternativo.

Stimare oggettivamente queste probabilità è difficile: un algoritmo può aiutare.

Alcuni esempi

- 41 ipotetici casi sono stati presentati a 81 giudici inglesi
- In nessun caso è stata raggiunta l'unanimità di giudizio

- 7 dei 41 casi erano ripetuti 2 volte, cambiando solo il nome dell'accusato
- La maggior parte dei giudici non ha preso la stessa decisione sullo stesso caso

Un algoritmo prenderebbe sempre la stessa decisione

L'ALGORITMO

Northpointe

Northpointe azienda fondata nel 1989 da Tim Brennan (professore di statistica in Colorado) e Dave Wells (dirigente di un programma di correzione in Michigan) per sviluppare l'algoritmo COMPAS (*Correctional Offender Management Profiling for Alternative Solutions*), uno tra i più usati per la valutazione del rischio di recidiva.

Nel 2011 Brennan e Wess hanno venduto Northpointe a un azienda basata a Toronto, e COMPAS è diventato un vero e proprio software.

L'algoritmo COMPAS

«Lo scopo di COMPAS è quello di ridurre il numero di reati e non la pena.»
[Brennan]

L'algoritmo fornisce un punteggio da 1 (basso rischio) a 10 (alto rischio) sulla base di 137 domande (non si chiede l'etnia):

- Quante persone conosci che assumono droghe?
- Se ti fanno innervosire perdi il controllo?
- Quante volte ti sei ritrovato in una rissa a scuola?
- ...

Definire il recidivo

Per recidivo si intende una persona che ha già ricevuto una valutazione di rischio tramite COMPAS e commette un reato dello stesso tipo entro i due anni successivi.

OTTIMA IDEA MA...

- Push the button
- Mancato test di validità
- Uso fuori contesto

Push the button

L'algoritmo non deve essere considerato come uno strumento per prendere decisioni alternativo al giudice, ma come uno strumento di supporto alle decisioni.

Mancato test di validità

Alcune giurisdizioni hanno iniziato a utilizzare algoritmi di valutazione del rischio prima ancora di *testare in modo rigoroso* se funzionassero davvero.

Lo Stato di New York, per esempio, ha iniziato a utilizzare l'algoritmo per valutare le persone in libertà vigilata in un progetto pilota del 2001, per poi estenderlo a tutti i dipartimenti dello Stato, ma non ha mai pubblicato una valutazione statistica completa dello strumento fino al 2012.

Inoltre non sono mai state valutate le differenze raziali, nonostante l'etnia non sia presente tra le domande.

Uso scorretto

In alcune contee gli algoritmi per la valutazione del rischio vengono utilizzati durante i processi.

Il punteggio però dovrebbe essere utilizzato prima del processo per stabilire quali imputati possono beneficiare della libertà vigilata o dei trattamenti alternativi.

Disparità raziale?

Northpointe non ha mai pubblicato i calcoli effettuati per arrivare al punteggio di rischio. L'algoritmo quindi risulta essere *privato*. Inoltre la validità è stata esaminata dalle stesse persone che hanno sviluppato l'algoritmo.

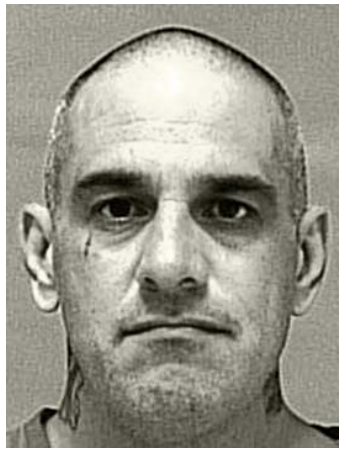
Nel 2014 il procuratore Eric Holder ha denunciato il fatto che questi punteggi potrebbero essere soggetti a pregiudizio raziale (*bias*).

ProPublica ha quindi avviato uno studio analizzando il software COMPAS per capirne l'*accuratezza* e testare se l'algoritmo fosse soggetto a disparità raziali.

Due arresti per piccoli furti

VERNON PRATER
REATI PRECEDENTI • 2 rapine a mano armata • 1 tentativo rapina a mano armata
RISK SCORE: 3

BRISHA BORDEN
REATI PRECEDENTI • 4 infrazioni giovanili
RISK SCORE: 8



Sta scontando 8 anni di carcere per furto aggravato

2 ANNI DOPO...

Nessun reato successivo

L'ALGORITMO SOTTO 'PROCESSO'

Raccolta dati

È stata scelta la contea di Broward (Florida):

- grande giurisdizione in cui si usa COMPAS *correttamente* per decidere se rilasciare un imputato in attesa di giudizio, consigliando eventualmente un trattamento alternativo
- in Florida ci sono leggi che regolano la diffusione dei registri carcerari

ProPublica ha analizzato un periodo di 2 anni, per un totale di 11757 persone che hanno ricevuto un punteggio di rischio nel 2013 e 2014, integrando i dati pubblici di incarcerazione dal 2013 al 2016.

Analisi iniziale

Sono stati analizzati i punteggi relativi al rischio di recidiva per crimini minori.

Pulizia dati

Si considerano solo le persone che hanno uno score COMPAS e che hanno commesso un nuovo reato entro i 2 anni dal primo score, o che non hanno commesso nuovi reati nei due anni successivi => 6172 persone:

- 3175 afroamericani e 2103 bianchi
- 1175 donne e 4997 maschi
- 2809 recidivi

Accuratezza della predizione

Per valutare l'accuratezza del modello si confronta lo score stabilito dall'algoritmo (predetto) con l'effettiva recidiva di una persona. Se questi non sono concordi allora è stato commesso un errore nell'assegnazione dello score.

RECIDIVO/SCORE	BASSO	ALTO
NO	VERO NEGATIVO (VN)	FALSO POSITIVO (FP)
SI	FALSO NEGATIVO (FN)	VERO POSITIVO (VP)

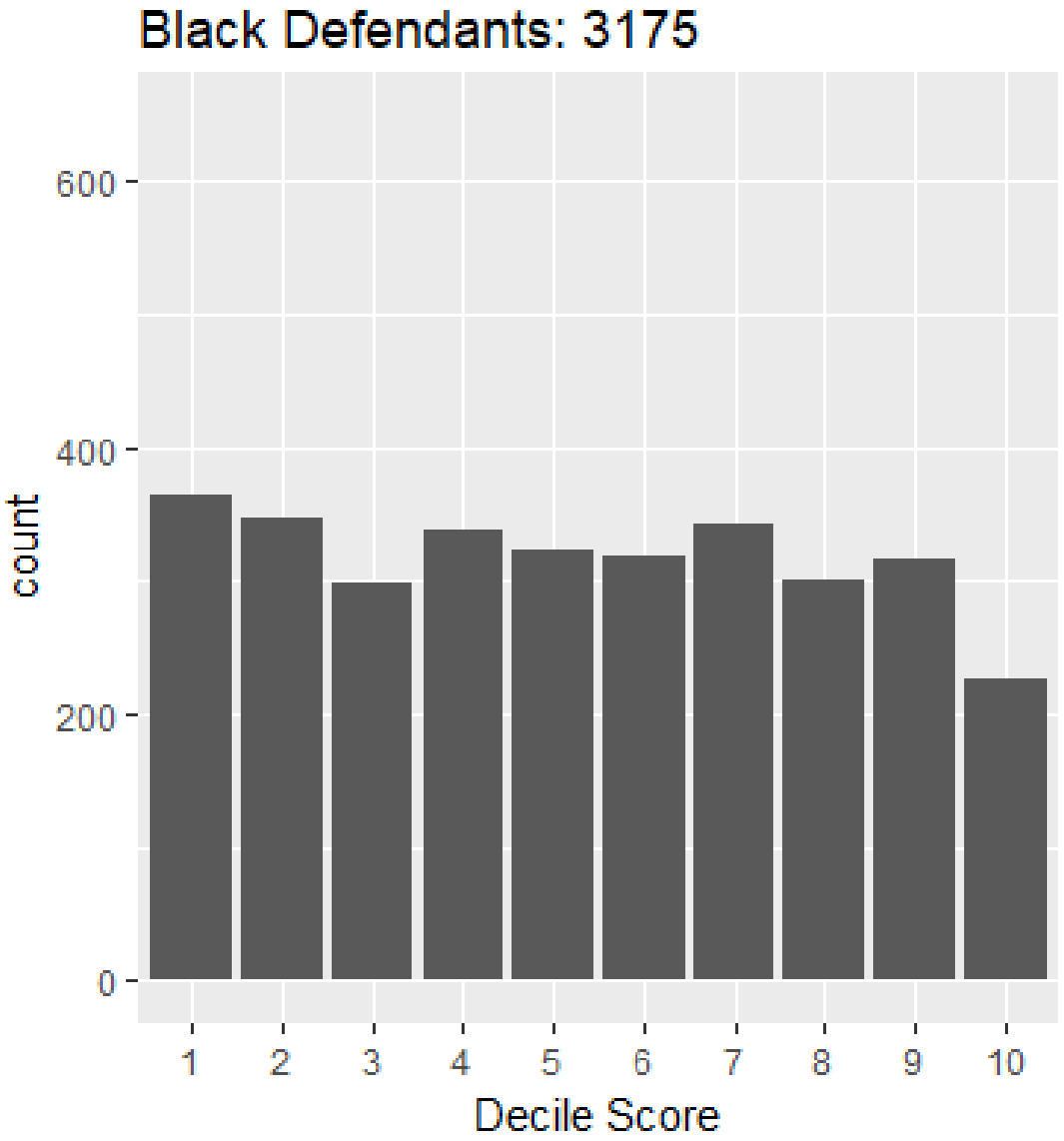
$$\text{ACCURATEZZA} = \frac{VN + VP}{VN + VP + FP + FN} * 100$$

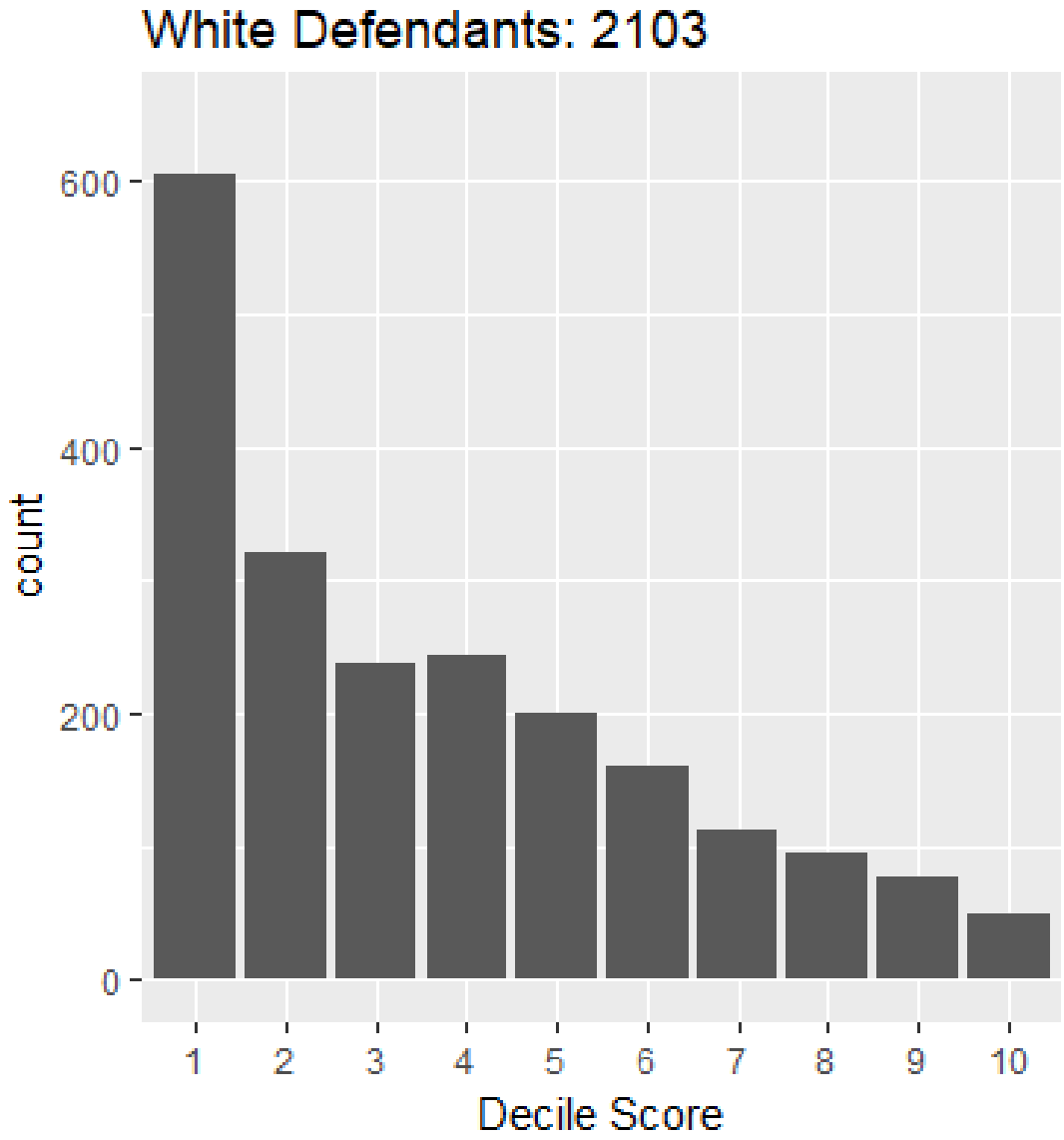
Accuratezza generale: 63,6%

Tipi di errore

ERRORI	BIANCHI	AFROAMERICANI
VALUTATO ALTO RISCHIO, NON HA RICOMMESSO REATO (FP)	23,5%	44,9%
VALUTATO BASSO RISCHIO, HA RICOMMESSO REATO (FN)	47,7%	28,0%

Decile Scores





Testare la disparità raziale

Nonostante sia chiara la differenza tra le distribuzioni dei punteggi COMPAS tra bianchi e afroamericani, guardando solo le precedenti analisi non teniamo conto degli altri fattori demografici e comportamentali.

Un modo per modellare la probabilità

Per testare la disparità nei punteggi è stato utilizzato un *modello di regressione logistica* che considera etnia, età, storia criminale, recidiva futura, grado di accusa, genere.

Raggruppiamo i punteggi COMPAS in due livelli:

- da 1 a 4 --> basso rischio di recidiva
- da 5 a 10 --> alto rischio di recidiva

Risultati (1)

- Il fattore che influenza di più l'esito è l'età: gli imputati con meno di 25 anni hanno una probabilità 2,5 volte maggiore di ottenere un punteggio alto rispetto a chi ha tra i 25 e i 45 anni

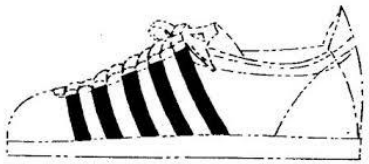
- Anche l'etnia risulta essere predittiva: gli imputati afroamericani hanno il 45% di probabilità in più di ottenere un punteggio alto rispetto ai bianchi

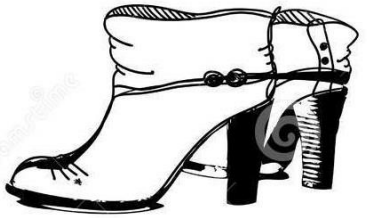
- Le imputate femmine hanno il 19,4% di probabilità in più di ottenere un punteggio alto rispetto agli uomini

Risultati (2)

- Gli imputati bianchi con un punteggio alto hanno una probabilità di recidività 3,61 volte maggiore rispetto a un bianco con un punteggio basso
- Gli imputati afroamericani con un punteggio alto hanno una probabilità di recidività 2,99 volte maggiore rispetto a un afroamericano con un punteggio basso
- Gli afroamericani che hanno ricevuto un punteggio alto hanno ricommesso un reato più spesso rispetto ai bianchi (63% VS 59%)
- COMPAS classifica con un punteggio basso gli imputati bianchi il 70,5% in più delle volte rispetto agli imputati afroamericani

Pensate a una scarpa...





Concetti chiave

- Algoritmi di questo tipo possono essere un valido strumento a supporto delle decisioni, ma non possono e non devono sostituire l'esperienza degli esperti (*innovation VS skepticism*, https://www.psychologytoday.com/blog/the-moment-youth/201206/the-art-positive-skepticism)
- Gli algoritmi dovrebbero essere resi pubblici (*trasparency*, *open data*, *reproducibility*)
- *Trade-off* tra complessità e comprensibilità del modello usato
- *Big Data* ≠ *Good Data*
- *Spurious Correlation* (http://www.tylervigen.com/spurious-correlations)

Bibliografia

- https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

- Hello World: How to be Human in the Age of the Machine, Hannah Fry

- Significance Magazine, Royal Statistical Society

Professione Statistico: materiale conferenze

http://www.dima.unige.it/~riccomag/