Fairness in Machine Learning

SML201: Introduction to Data Science, Spring 2019

Michael Guerzhoy

Content from Moritz Hardt, Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel

Terminology note

- "Supervised machine learning" is the same as "predictive modelling"
  - Machine Learning is usually applied to larger datasets

Glossary

Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = $1,000,000	large grant= $50,000
nice place to have a meeting: Snowbird, Utah, French Alps	nice place to have a meeting: Las Vegas in August

Robert Tibshirani

The dangers of math snobbery according to Moritz Hardt

- “Technical work without understanding social context”
- “Thinking we’re more rigorous than social scientists”
- “Justifying an approach by the math it entails”

Running examples

- A model that estimates the probability that a person will recidivate
  - Used in deciding whether to grant bail
- A model that estimates the probability that a person will default
  - Used in deciding whether to offer a loan

COMPAS

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

https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html

https://www.propublica.org/article/technical-response-to-northpointe

https://www.liebertpub.com/doi/pdf/10.1089/big.2016.0047

COMPAS

- "Correctional Offender Management Profiling for Alternative Sanctions"
  - Developed by Northpointe (currently Equivant)
  - Used by *a lot* of probation departments to assess the likelihood of a defendant becoming a recidivist
  - Defendants who are defined as medium or high risk are more likely to be detained before trial
    - (N.B., this is only suggestive of importance)
  - Race is not an input to the algorithm

COMPAS Probation Risk and Needs Assessment Questionnaire

OFFENDER NAME:	NYSID:	STATUS:
RACE:	SEX:	DOB:
DATE OF ASSESSMENT:	MARITAL STATUS:	
SCALE SET: Full COMPAS Assessment v2	AGENCY/COUNTY NAME:	

PART ONE: CRIMINAL HISTORY / RISK ASSESSMENT

CURRENT CHARGES

What offenses are covered by the current charges (check all that apply)?

Homicide	Arson	Property/Larceny
Assault	Weapons	Fraud
Robbery	Drug Sales	DWI / DWAI
Sex Offense (with force)	Drug Possession	AUO
Sex Offense (without force)	Burglary	Other

1 Do any of the current offenses involve domestic violence?

Yes No

2 What offense category represents the most serious current charge?

Misdemeanor Non-Assault Felony Assaultive Felony

3 Was there any degree of physical injury to a victim in the current offense?

Yes No

4 Based on your judgment, after reviewing the history of the offender from all known sources of information (PSI, police reports, prior supervision, victim, etc.) does the defendant demonstrate a pattern of violent behavior against people resulting in physical injury?

Yes No

http://www.northpointeinc.com/downloads/research/DCJS_OPCA_COMPAS_Probation_Validity.pdf

PART TWO: NEEDS ASSESSMENT

A. ASSOCIATES / PEERS

17 The offender has peers and associates who *(check all that apply)* :

- Use illegal drugs
- Have been arrested
- Have been incarcerated
- None
- Lead law-abiding lifestyles
- Are gainfully employed
- Are involved in pro-social activities

18 What is the gang affiliation status of the offender :

- Current gang membership
- Previous gang membership
- Not a member but associates with gang members
- None

19 Does the offender have a criminal alias, a gang-related or street name?

Yes No

20 Does unstructured idle time contribute to the opportunity for the offender to commit criminal offenses?

Yes Unsure No

21 Does offender report boredom as a contributing factor to his or her criminal behavior?

Yes Unsure No

B. FAMILY

22 Are the offender 's family or household members able and willing to support a law abiding lifestyle?

Yes Unsure No

23 Is the offender's current household characterized by *(check all that apply)* :

COMPAS Probation Risk and Needs Assessment Questionnaire – *Continued*

PART THREE: OFFENDER QUESTIONNAIRE

NYSID : Name : DOB :

Please look at the following areas and let us know which of them you think will present the greatest problems for you. *Please check one response for each question in the column provided* .

	Please answer questions as either No, Yes or Don't Know	No	Yes	Don't Know
48	Do you feel you need assistance with finding or maintaining a steady job?			
49	Do you feel you need assistance with finding or maintaining a place to live?			
50	Will money be a problem for you over the next several months?			

	How difficult will it be for you to...	Not Difficult	Somewhat Difficult	Very Difficult
51	manage your money?			
52	keep a job once you have found one or if you currently have one?			
53	find or keep a steady place to live?			
54	have enough money to get by?			
55	find or keep people that you can trust?			
56	find or keep friends who will be a good influence on you?			
57	avoid risky situations?			
58	learn to control your temper?			
59	find things that interest you?			
60	learn better skills to get or keep a job?			
61	find a safe place to live where you won't be hassled or threatened?			
62	get along with people?			

COMPAS Probation Risk Assessment

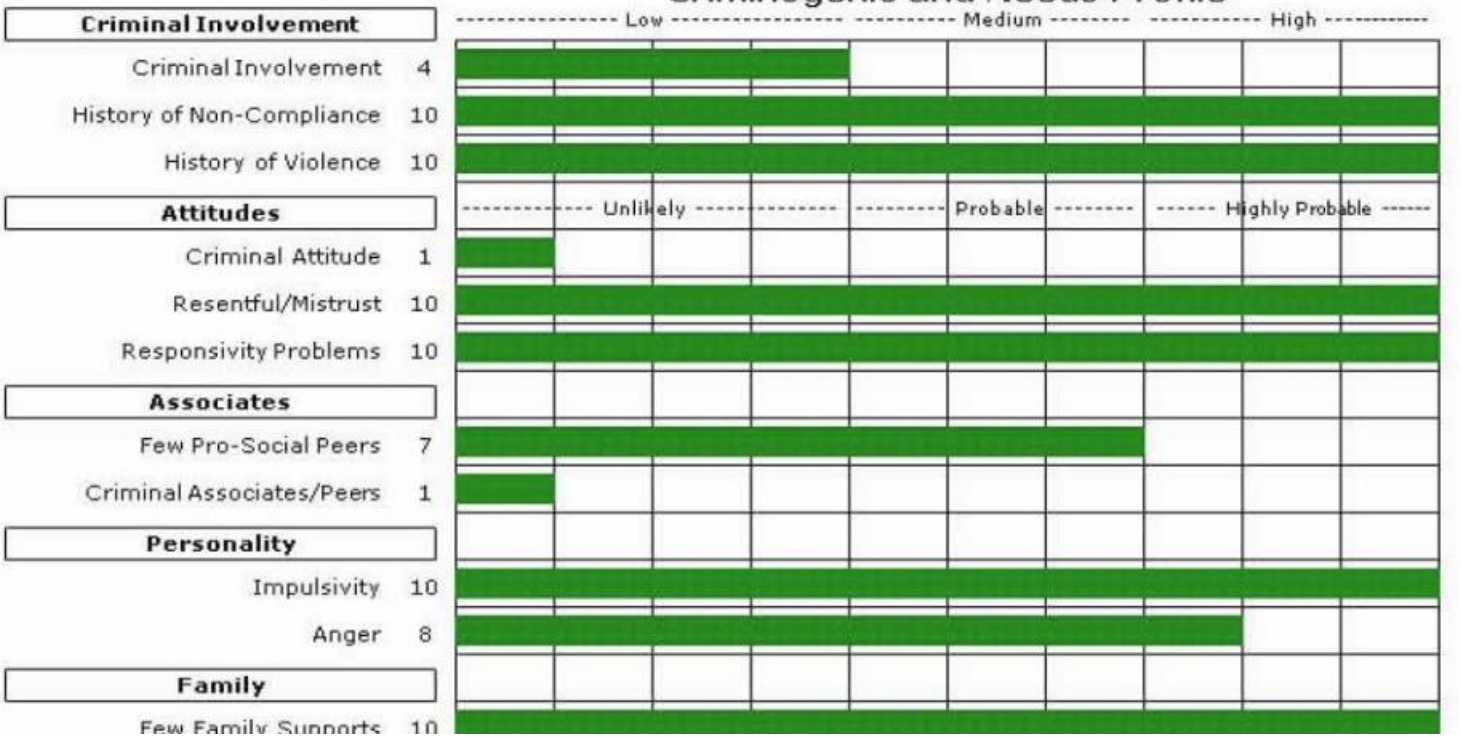
Offender: **Joe Sample** DOB: **2/2/1950** Gender: **Male**

Screening Date: **9/13/2007** Screener: **Hellem, Dan** Ethnicity: **Native A**

Scale Set: **DMB-PSI** Case: **009943** Marital Status: **Single**

Overall Risk Potential

Criminogenic and Needs Profile



Compas

- Basically, a logistic regression model, with the answers to the questionnaire as the predictors and the probability of another arrest as the outcome

COMPAS

- Correct predictions about ~65% of the time, for both white and black defendents
  - Black defendants who did not recidivate were incorrectly predicted to reoffend at a rate of 44.9%
  - White defendants who did not recidicate were predicted to reoffend at a rate of 23.5%
  - White defendants who did recidivate were incorrectly predicted to not reoffend at a rate of 47.7%
  - Black defendants who did recidivate were incorrectly predicted to not reoffend at a rate of 28.0%

What is a fair model, mathematically?

What is a fair model, legally?

- Various legal rules in the US
  - https://en.wikipedia.org/wiki/Disparate_treatment
  - https://en.wikipedia.org/wiki/Disparate_impact
  - 80% rule: if a group is hired at less than 80% the rate of another group, that is (sometimes) evidence of adverse impact
  - Various other tests used

Demographic parity

- Assume C and A are binary
  - E.g. $C = 1$ means "likely to reoffend" and A=1 indicates a protected group
  - The classifier *C satisfies* demographic parity if
    - The probability of saying "yes" and "no" is the same regardless of the value of *A*
    - $\frac{\#(C=1,A=0)}{\#(A=0)} = \frac{\#(C=1,A=1)}{\#(A=1)}$

Accuracy parity

- The Classifier *C satisfies* accuracy parity if
  - The prediction accuracy is the same for different demographics

True positive parity

- The Classifier *C satisfies* true positive parity if
  - The probability of getting a correct “yes” is the same regardless of demographics
  - “Equal opportunity”

False positive parity

- The Classifier *C satisfies* true positive parity if, for correct answer Y,
  - The probability of getting an incorrect “yes” is the same regardless of demographics

Predictive value parity

- The Classifier *C satisfies* predictive value parity
  - The probability of getting an incorrect “yes” is the same regardless of demographics
  - The probability of getting an incorrect “no” is the same regardless of demographics

Back to COMPAS

- Likelihood of a nonrecidivating black defendant being assessed as high risk is nearly twice that of white nonrecidivating defendants
  - No false positive parity
- But accuracy parity is satisfied
  - The probability that a defendant assessed as high risk will recidivate is roughly the same regardless of race
- Mathematically, it is not in general possible to satisfy both accuracy parity and false positive parity at the same time
  - Only possible if the “base rates” – the proportions of people recidivating – are the same

Accuracy Parity vs. False Positive Parity

Low-risk: 10% chance of re-arrest

High-risk: 80% chance of re-arrest

Group A	Group B
Low-risk: 40, High-risk: 60	Low-risk: 50, High-risk: 50

- Assume the system perfectly identifies low vs. high-risk
- Group A: Predict 60 will be arrested. 12/60 won't be.
- Group B: Predict 50 will be arrested. 10/50 won't be.
- Group A: error rate is $\frac{12+4}{100} = 16\%$
- Group B: error rate is $\frac{10+5}{100} = 15\%$
- Larger differences in error rates for larger discrepancies
- Equalizing the error rates (perhaps by randomly erring when deciding about group B, if the user is acting in bad faith) will mess up the false-positive parity

Accuracy Parity vs. False Positive Parity

- What if arrests are more likely in one neighborhood than another?

Demographic Parity

- The Classifier *C satisfies* accuracy parity if
  - The probability of saying “yes” and “no” is the same regardless of *A*
- Does not rule out accepting random people in group *a* but only qualified people in group *b*
  - Can happen if there is not enough data about group *a*
- If the *base rates* – proportions of people for whom Y=1 – are different across different groups, the perfect classifier *C=Y* is ruled out
  - Y=1 is the correct answer

Accuracy parity

- The Classifier *C satisfies* accuracy parity if
- The prediction accuracy is the same for different demographics
- Allows for the perfect predictor *C=Y*
- Discourages laziness by equalizing error rates in all groups
- False positive and negative rates will not in general be equal
  - Might “make up” for rejecting qualified women by accepting unqualified men, making the accuracy the same across demographics

True positive parity

- The Classifier *C satisfies* true positive parity if, for correct answer Y,
  - The probability of getting a correct "yes" is the same regardless of demographics
  - "Equal opportunity"
- Suitable when a positive outcome is desirable, and we want everyone who is qualified to have an equal shot at it
- E.g., a system that decides if to grant loans to people

Fairness through unawareness

- The model is not allowed to use/see demographic information
- But that information can often be inferred
- Will not lead to any of the other notions of fairness
  - COMPAS does not measure demographic information
  - Demographic information can often be fairly easily inferred given enough information
  - Proxies for demographic information (e.g., the address) might be used by the model instead

The classifier can only be as good as the training set

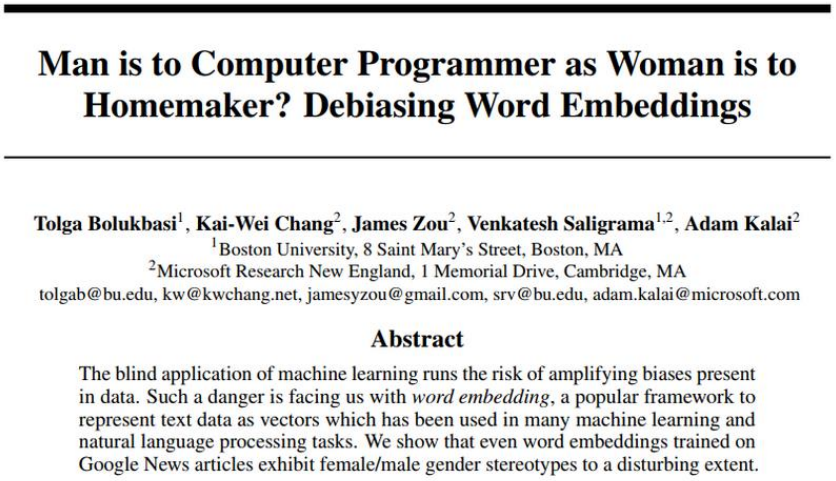
- Re-arrests may be a biased measure of public safety. Predominantly black neighborhoods are policed more heavily.

Sources of unfairness (without explicit wrongdoing)

- Sample size disparity
  - More training data generally leads to smaller errors
  - More likely that there will be less training data for minority populations
- Argument: checking algorithms for the different notions of fairness will encourage companies to collect more data to improve their classifiers

Sources of unfairness: biases in data

- Data collection procedures may be biased
- Decisions about how to measure data may be biased
- Extant text and image data may be biased



Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi[1], **Kai-Wei Chang**[2], **James Zou**[2], **Venkatesh Saligrama**[1,2], **Adam Kalai**[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

Abstract

The blind application of machine learning runs the risk of amplifying biases present in data. Such a danger is facing us with *word embedding*, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. We show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent.

> “Machine learning is like money laundering for bias. It's a clean, mathematical apparatus that gives the status quo the aura of logical inevitability. The numbers don't lie.”

Maciej Cegłowski

Model Cards for Model Reporting

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru
{mmitchellai,simonewu,andrewzaldivar,parkerbarnes,lucyvasserman,benhutch,espitzer,tgebru}@google.com

ABSTRACT

Trained machine learning models are increasingly used to perform high-impact tasks in areas such as law enforcement, medicine, education, and employment. In order to clarify the intended use cases of machine learning models and minimize their usage in contexts for which they are not well suited, we recommend that released models be accompanied by documentation detailing their performance characteristics. In this paper, we propose a framework that we call model cards, to encourage such transparent model reporting. Model cards are short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions, such as across different cultural, demographic, or phenotypic groups (e.g., race, geographic location, sex, Fitzpatrick skin type [17]) and intersectional groups (e.g., age and race, or sex and Fitzpatrick skin type) that are relevant to the intended application domains. Model cards also disclose the context in which models are intended to be used, details of the performance evaluation procedures, and other relevant information. While we focus primarily on human-centered machine learning models in the application fields of computer vision and natural language processing, this framework can be used to document any trained machine learning model. To solidify the concept, we provide cards for two supervised models: One trained to detect smiling faces in images, and one trained to detect toxic comments in text. We propose model cards as a step towards the responsible democratization of machine learning and related AI technology, increasing transparency into how well AI technology works. We hope this work encourages those releasing trained machine learning models to accompany model releases with similar detailed evaluation numbers and other relevant documentation.

problematic when models are used in applications that have serious impacts on people's lives, such as in health care [16, 39, 41], employment [3, 15, 27], education [23, 42] and law enforcement [4, 9, 20, 31].

Researchers have discovered systematic biases in commercial machine learning models used for face detection and tracking [6, 11, 43], attribute detection [7], criminal justice [12], toxic comment detection [13], and other applications. However, these systematic errors were only exposed after models were put into use, and negatively affected users reported their experiences. For example, after MIT Media Lab graduate student Joy Buolamwini found that commercial face recognition systems failed to detect her face [6], she collaborated with other researchers to demonstrate the disproportionate errors of computer vision systems on historically marginalized groups in the United States, such as darker-skinned women [7, 38]. In spite of the potential negative effects of such reported biases, documentations accompanying publicly available trained machine learning models (if supplied) provide very little information regarding model performance characteristics, intended use cases, potential pitfalls, or other information to help users evaluate the suitability of these systems to their context. This highlights the need to have detailed documentation accompanying trained machine learning models, including metrics that capture bias, fairness and inclusion considerations.

As a step towards this goal, we propose that released machine learning models be accompanied by short (one to two page) records we call model cards. Model cards (for model reporting) are complements to "Datasheets for Datasets" [21] and similar recently proposed documentation paradigms [5, 26] that report details of the datasets used to train and test machine learning models. We

Conclusions

- Most fairness measures are not compatible
- Enforcing algorithmic fairness can reduce the classification accuracy of the algorithm
  - But algorithms are not static: e.g., more data can be gathered to improve the accuracy and the fairness
- The way people use algorithms is probably more of an issue than formal fairness criteria
- Should always consider various fairness criteria when designing/deploying opaque systems