Fairness-aware AI
A data science perspective

Indrė Žliobaitė
Dept. of Computer Science, University of Helsinki

October 15, 2018
Machine intelligence?

Strong AI – machine consciousness and mind
Weak AI – focused on one narrow task

no genuine intelligence, no self-awareness, no life

https://www.translinguoglobal.com/10-reasons-why-google-translate-is-not-better-than-learning-language/
AI in everyday life: data driven decision support

Banking

Image source: https://www.thebalance.com/how-credit-scores-work-315541
http://tcgstudy.com/ranking_to_universities.html
https://www.pinterest.com/pin/443041682071895474/

Sports analytics

Hiring

University admittance
Learning coarse representations

We don't want to know how decisions are made in a doctor's head. We trust the judgement. How about AI?

But to Amazon, and likely others in the tech world, the decision had nothing to do with racism and everything to do with the facts. Amazon argued that it wasn’t acting on prejudice when it excluded those neighborhoods. Instead, algorithms and the underlying data on which those algorithms were based made it clear that Amazon couldn’t make a profit in them. And, given that Amazon is profit-driven, the company excluded them. Race, Amazon said, had nothing to do with it.
How models are built

Data science: model is a summary of data

<table>
<thead>
<tr>
<th>Tid</th>
<th>Attr1</th>
<th>Attr2</th>
<th>Attr3</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Large</td>
<td>125K</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Medium</td>
<td>100K</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>Small</td>
<td>70K</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>Medium</td>
<td>120K</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Large</td>
<td>95K</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>Medium</td>
<td>60K</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>Large</td>
<td>220K</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>No</td>
<td>Small</td>
<td>85K</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>No</td>
<td>Medium</td>
<td>75K</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>No</td>
<td>Small</td>
<td>90K</td>
<td>Yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tid</th>
<th>Attr1</th>
<th>Attr2</th>
<th>Attr3</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>No</td>
<td>Small</td>
<td>55K</td>
<td>?</td>
</tr>
<tr>
<td>12</td>
<td>Yes</td>
<td>Medium</td>
<td>80K</td>
<td>?</td>
</tr>
<tr>
<td>13</td>
<td>Yes</td>
<td>Large</td>
<td>110K</td>
<td>?</td>
</tr>
<tr>
<td>14</td>
<td>No</td>
<td>Small</td>
<td>95K</td>
<td>?</td>
</tr>
<tr>
<td>15</td>
<td>No</td>
<td>Large</td>
<td>67K</td>
<td>?</td>
</tr>
</tbody>
</table>
Rule-based models: decision tree

Training Data

<table>
<thead>
<tr>
<th>ID</th>
<th>Home Owner</th>
<th>Marital Status</th>
<th>Annual Income</th>
<th>Defaulted Borrower</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Single</td>
<td>125K</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Married</td>
<td>100K</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>Single</td>
<td>70K</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>Married</td>
<td>120K</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Divorced</td>
<td>95K</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>Married</td>
<td>60K</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>Divorced</td>
<td>220K</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>No</td>
<td>Single</td>
<td>85K</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>No</td>
<td>Married</td>
<td>75K</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>No</td>
<td>Single</td>
<td>90K</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Model: Decision Tree

- **Home Owner**
  - Yes: NO
  - No: Single, Divorced
    - Single, Divorced: Income
      - < 80K: NO
      - > 80K: YES
    - Married: NO
How white engineers built racist code - and why it's dangerous for black people

As facial recognition tools play a bigger role in fighting crime, inbuilt racial biases raise troubling questions about the systems that create them

Ali Breland
Mon 4 Dec 2017 09.00 GMT
If there is discrimination in data on which a model is trained unless instructed otherwise, a learned model will carry discrimination forward.
If there is discrimination in data on which a model is trained unless instructed otherwise, a learned model will carry discrimination forward.

**Model**

**Data**

**Prediction**

Image source: Introduction to Data Mining, 2nd Edition by Tan, Steinbach, Karpate, Kumar.
Fairness-aware AI
Brief history of fairness-aware ML

2010 First Bayesian solutions by Calders and Verwer
2011 Conditional discrimination paper by Žliobaitė et al
2011 First PhD thesis: F.Kamiran
2012 First dedicated workshop at IEEE ICDM
2013 Edited book
2013 Second PhD thesis: S.Hajian
2014 Special issue in AI and Law
2015 First causality perspective paper by Bonchi at al
2015 First dedicated course, Aalto and UH
2016 Many research papers are coming out, conferences start having dedicated tracks
2016 O’Neil’s book
2017-2018 A new paper or a few every week, many on optimization criteria/ measurement
Machine learning and discrimination

- Discrimination – inferior treatment based on ascribed group rather than individual merits
- Discriminate = *distinguish* (lat.)

- Machine learning
  - uses proxies to distinguish an individual from the mean
  - without judging what is morally right or wrong
  - can enforce constraints *defined by legislation and/or social norms* (external)
There are no “right” or “wrong” variables

Morality is a social convention?

Immanuel Kant introduced the categorical imperative: "Act only according to that maxim whereby you can, at the same time, will that it should become a universal law"
Hidden Bias

When Algorithms Discriminate

By Claire Cain Miller

July 9, 2015

The online world is shaped by forces beyond our control, determining the stories we read on Facebook, the people we meet on OkCupid and the search results we see on Google. Big data is used to make decisions about health care, employment, housing, education and policing.

But can computer programs be discriminatory?
Why unbiased computational processes can lead to discriminative decision procedures

- data is incorrect
  - due to biased decisions in the past
  - population is changing over time

- data is incomplete (omitted variable bias)
- sampling procedure skews the data

Calders and Žliobaitė 2013, book chapter
Desired outcomes?
No direct discrimination

“Twin test”

- $p(+|X, s_1) \neq p(+|X, s_2)$
“Blind” auditioning

Source: https://www.astridbaumgardner.com/blog-and-resources/blog/ysm-mock-auditions/
No indirect discrimination

• Challenges:
  – what is the “right” outcome (qualitatively)?
  – what level of positive decision we are aiming at (quantitatively)

- \[ p(+|X,s1) = p(+|X,s2) \ AND \]
- \[ p(+|X) = p(+|X)^\text{right} \]
London private hire drivers 'don't need written test', mayor told

13 March 2017

London Assembly members voted 15-0 for a motion to replace the written test with a cheaper verbal language test.

Essay writing should not form part of the licensing requirements for private hire drivers, politicians have said.

In a motion signed earlier, 15 London Assembly members unanimously agreed that passing a written English test was an unnecessary requirement for drivers.

Of greater importance, they said, was the ability to communicate verbally, the London Evening Standard reported.
Redlining

Machine learning and discrimination

- Discrimination – inferior treatment based on ascribed group rather than individual merits

Predictive models $y \rightarrow f(X)$

Machine learning and discrimination

- Discrimination – inferior treatment based on ascribed group rather than individual merits

Predictive models $y \rightarrow f(X)$

Source: https://ilmatieteenlaitos.fi/
Machine learning and discrimination

- Discrimination – inferior treatment based on ascribed group rather than individual merits

Predictive models $y \rightarrow f(X)$

Source: https://ilmatieteenlaitos.fi/
• Removing protected characteristic
  - $y \rightarrow f(X,s)$  \[ x \rightarrow f(X) \]
"...until neither the Plain nor the Star-Bellies knew

whether this one was that one... or that one was this one...

or which one was what one... or what one was who."
- Removing protected characteristic does not solve the problem if \( s \) is correlated with \( X \)
  - desired: \( y \rightarrow f(X) \)
• Removing protected characteristic does not solve the problem if \( s \) is correlated with \( X \)
  
  - desired: \( y \rightarrow f(X) \)
- Removing protected characteristic does not solve the problem if s is correlated with X
  - desired: \( y \rightarrow f(X) \)
  - what happens: \( y \rightarrow f(X,s^*), s^* \rightarrow f(X) \)
Naive baseline
Naive baseline: removing sensitive variable

- Suppose salary is decided (in decision maker's head) as

\[ salary = 1000 + 100 \times \text{education} - 500 \times \text{ethnicity} \]
Naive baseline: removing sensitive variable

- Suppose salary is decided (in decision maker's head) as
  \[ salary = 1000 + 100 \times education - 500 \times ethnicity \]

- Data scientist assumes
  \[ salary = b_0 + b_1 \times education \]
Naive baseline: removing sensitive variable

- Suppose salary is decided (in decision maker's head) as
  \[
  \text{salary} = 1000 + 100 \times \text{education} - 500 \times \text{ethnicity}
  \]
- Data scientists assumes
  \[
  \text{salary} = b_0 + b_1 \times \text{education}
  \]
- Observes data

<table>
<thead>
<tr>
<th>education</th>
<th>ethnicity</th>
<th>salary</th>
<th>education</th>
<th>ethnicity</th>
<th>salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>600</td>
<td>1</td>
<td>0</td>
<td>1100</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>700</td>
<td>6</td>
<td>0</td>
<td>1600</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>800</td>
<td>7</td>
<td>0</td>
<td>1700</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>900</td>
<td>9</td>
<td>0</td>
<td>1900</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1500</td>
<td>10</td>
<td>0</td>
<td>2000</td>
</tr>
</tbody>
</table>
Naive baseline: removing sensitive variable

- Suppose salary is decided (in decision maker's head) as

  \[ salary = 1000 + 100 \times education - 500 \times ethnicity \]

- Data scientist assumes

  \[ salary = b_0 + b_1 \times education \]

- Observes data

<table>
<thead>
<tr>
<th>education</th>
<th>ethnicity</th>
<th>salary</th>
<th>education</th>
<th>ethnicity</th>
<th>salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>600</td>
<td>1</td>
<td>0</td>
<td>1100</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>700</td>
<td>6</td>
<td>0</td>
<td>1600</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>800</td>
<td>7</td>
<td>0</td>
<td>1700</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>900</td>
<td>9</td>
<td>0</td>
<td>1900</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1500</td>
<td>10</td>
<td>0</td>
<td>2000</td>
</tr>
</tbody>
</table>

- Learned model

  \[ salary = 602 + 128 \times education \]
Removing sensitive variable makes it worse

- Suppose salary is decided (in decision maker's head) as
  \[ salary = 1000 + 100 \times education - 500 \times ethnicity \]

- Data scientist assumes
  \[ salary = b_0 + b_1 \times education \]

- Observes data

<table>
<thead>
<tr>
<th>education</th>
<th>ethnicity</th>
<th>salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>600</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>700</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>800</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>900</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1500</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>education</th>
<th>ethnicity</th>
<th>salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1100</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>1600</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1700</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>1900</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>2000</td>
</tr>
</tbody>
</table>

- Learned model
  \[ salary = 602 + 128 \times education \]

  Lower base salary, higher reward for education, the model punishes ethnical minorities
How large is the bias?

- Suppose salary is decided (in decision maker's head) as

$$\text{salary} = 1000 + 100 \times \text{education} - 500 \times \text{ethnicity}$$

- Learned model

$$\text{salary} = 602 + 128 \times \text{education}$$

- Salary = \((1000 - 398) + (100 + 28) \times \text{education}\)

Punishment on base salary

$$\gamma = \alpha \times \text{mean(Ethnicity)} - \beta \times \text{mean(education)}$$

$$= (-500) \times 0.5 - 28 \times 5.3 = 398$$

Award for education

$$\beta = \alpha \times \frac{\text{Cov(Education,Ethnicity)}}{\text{Var(Education)}}$$

$$= (-500) \times (-0.72) / 12.9 = 28$$

Žliobaitė and Custers 2016
A simple solution (special case)

- Learn a model on the full dataset

\[ salary = 1000 + 100 \times education - 500 \times ethnicity \]

- Remove the sensitive component of the model

\[ salary = 1000 + 100 \times education - 500 \times ethnicity \]
Measuring discrimination
Measuring

\[
p(+|M) = 50% \\
p(+|F) = 25%
\]
Measuring

\[ p(\text{+}|\text{M}) = 50\% \]
\[ p(\text{+}|\text{F}) = 25\% \]

\[ p(\text{+}|\text{M}) = 38\% \]
\[ p(\text{+}|\text{F}) = 38\% \]
Measuring Reverse discrimination

experience in years

(a) accept

(b) accept
Measuring

\[ p(+|1..3,M) = \text{n.a.} \quad p(+|4..6,M) = 0\% \quad p(+|7..9,M) = 67\% \quad p(+|10..12,M) = 100\% \]

\[ p(+|1..3,F) = 0\% \quad p(+|4..6,F) = 0\% \quad p(+|7..9,F) = 67\% \quad p(+|10..12,F) = \text{n.a.} \]
Measuring

- "Twin test" / counterfactual fairness
  - $p(+|X,M) = p(+|X,F)$

- No "Redlining"
  - $p(+|X,M) = p(+|X,F) = p(+|X) = "right"$
Which is the “right” level?

<table>
<thead>
<tr>
<th></th>
<th>hours per week</th>
<th>annual income (K$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>female</td>
<td>36.4</td>
<td>5.8$/h</td>
</tr>
<tr>
<td>male</td>
<td>42.4</td>
<td>13.8$/h</td>
</tr>
<tr>
<td>all data</td>
<td>40.4</td>
<td>11.4$/h</td>
</tr>
</tbody>
</table>

Redistribution of the same total resources?

Like male?

Like all?
Indirect discrimination
Is there discrimination?

- **Males**
  - To medicine: 25%
  - To computer science: 45%

- **Females**
  - To medicine: 15%
  - To computer science: 35%
Is there discrimination?

<table>
<thead>
<tr>
<th></th>
<th>medicine</th>
<th></th>
<th>computer</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>female</td>
<td>male</td>
<td>female</td>
<td>male</td>
</tr>
<tr>
<td>Number of applicants</td>
<td>800</td>
<td>200</td>
<td>200</td>
<td>800</td>
</tr>
<tr>
<td>Acceptance rate</td>
<td>15%</td>
<td>25%</td>
<td>35%</td>
<td>45%</td>
</tr>
<tr>
<td>Accepted (+)</td>
<td>120</td>
<td>50</td>
<td>70</td>
<td>360</td>
</tr>
</tbody>
</table>

How to correct?
What should be the acceptance rate?

Kamiran et al 2013, KAIS
## Is there discrimination?

<table>
<thead>
<tr>
<th></th>
<th>medicine</th>
<th>computer</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>female</td>
<td>male</td>
<td>female</td>
</tr>
<tr>
<td><strong>No discrimination</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of applicants</td>
<td>800</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>Acceptance rate</td>
<td>20%</td>
<td>20%</td>
<td>40%</td>
</tr>
<tr>
<td>Accepted (+)</td>
<td>160</td>
<td>40</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Discrimination is present</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of applicants</td>
<td>800</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>Acceptance rate</td>
<td>15%</td>
<td>25%</td>
<td>35%</td>
</tr>
<tr>
<td>Accepted (+)</td>
<td>120</td>
<td>50</td>
<td>70</td>
</tr>
</tbody>
</table>

Kamiran et al 2013, KAIS
Adversary action

to medicine

males

5%

55%

to computer science

females

5%

55%
### Redlining / indirect discrimination

Is there discrimination, or not?

<table>
<thead>
<tr>
<th></th>
<th>medicine</th>
<th></th>
<th>computer</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>female</td>
<td>male</td>
<td>female</td>
<td>male</td>
</tr>
<tr>
<td>Number of applicants</td>
<td>800</td>
<td>200</td>
<td>200</td>
<td>800</td>
</tr>
<tr>
<td>Acceptance rate</td>
<td>5%</td>
<td>5%</td>
<td>55%</td>
<td>55%</td>
</tr>
<tr>
<td>Accepted (+)</td>
<td>40</td>
<td>10</td>
<td>110</td>
<td>440</td>
</tr>
</tbody>
</table>

2000
30%
600
## Redlining / indirect discrimination

<table>
<thead>
<tr>
<th></th>
<th>medicine</th>
<th></th>
<th>computer</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>female</td>
<td>male</td>
<td>female</td>
<td>male</td>
</tr>
<tr>
<td>Number of applicants</td>
<td>800</td>
<td>200</td>
<td>200</td>
<td>800</td>
</tr>
<tr>
<td>Acceptance rate</td>
<td>20%</td>
<td>20%</td>
<td>40%</td>
<td>40%</td>
</tr>
<tr>
<td>Accepted (+)</td>
<td>160</td>
<td>40</td>
<td>80</td>
<td>320</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>400</td>
<td>200</td>
<td>600</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>medicine</th>
<th></th>
<th>computer</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>female</td>
<td>male</td>
<td>female</td>
<td>male</td>
</tr>
<tr>
<td>Number of applicants</td>
<td>800</td>
<td>200</td>
<td>200</td>
<td>800</td>
</tr>
<tr>
<td>Acceptance rate</td>
<td><strong>5%</strong></td>
<td><strong>5%</strong></td>
<td><strong>55%</strong></td>
<td><strong>55%</strong></td>
</tr>
<tr>
<td>Accepted (+)</td>
<td>40</td>
<td>10</td>
<td>110</td>
<td>440</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>550</td>
<td>200</td>
<td>600</td>
</tr>
</tbody>
</table>
See literature for more on measuring

- Starting points:

- We can measure on model outputs the same way we measure on data
Algorithmic solutions
Algorithmic solutions

• Preprocessing
  – Modify historical data (inputs, protected or outcomes)
  – Resample input data

• Post-processing
  – Modify models
  – Modify predictions

• Optimization with non-discrimination constraints
Modify input data

- Modify $y$ - massaging

Kamiran and Calders 2009
Modify input data

- Modify X – massaging
  - Any attributes in X that could be used to predict $s$ are changed such that a fairness constraint is satisfied
  - Approach is similar to sanitizing datasets for privacy preservation

Feldman et al. 2014
Resample

- Preferential sampling

Kamiran and Calders 2010
Post-processing

- Modifying the resulting model

Relabel tree leaves to remove the most discrimination with the least damage to the accuracy

Kamiran et al. 2010
Optimization with constraints

- Decision tree

Regular tree induction

\[ IGC := H_{\text{Class}}(D) - \sum_{i=1}^{k} \frac{|D_i|}{|D|} H_{\text{Class}}(D_i) \]

Entropy wrt class label

Discrimination-aware tree induction

\[ IGS := H_{B}(D) - \sum_{i=1}^{k} \frac{|D_i|}{|D|} H_{B}(D_i) \]

Entropy wrt protected characteristic

Tree splits are decided on: IGC - IGS

Kamiran et al. 2010
Prevention solutions

• Preprocessing
  – Modify input data X, s or y
  – Resample input data

• Post-processing
  – Modify models
  – Modify outputs

• Learning with constraints
  From the legal perspective ??
  – Decision manipulation – very bad
  – Data manipulation – quite bad
  – Learning with constraints - ok
    • Protected characteristic should not be used in decision making
Baselines
Problem

- Baseline accuracy and baseline discrimination varies with varying overall acceptance rate.
- Classifiers with different overall acceptance rates are not comparable.

Acc = 50%
Acc = 90%
Experiment

Adult dataset from UCI
### Discrimination-accuracy tradeoffs

<table>
<thead>
<tr>
<th></th>
<th>p(+)</th>
<th>Acc.</th>
<th>Disc.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\pi$</td>
<td>$A$</td>
<td>$d$</td>
</tr>
<tr>
<td>Data/oracle</td>
<td>24.7</td>
<td>100</td>
<td>19.9</td>
</tr>
<tr>
<td>Logistic with s</td>
<td>20.2</td>
<td>84.9</td>
<td>18.3</td>
</tr>
<tr>
<td>Logistic no s</td>
<td>20.1</td>
<td>84.9</td>
<td>17.6</td>
</tr>
<tr>
<td>Logistic massage</td>
<td>22.1</td>
<td>83.5</td>
<td>6.9</td>
</tr>
<tr>
<td>NB with s</td>
<td>15.4</td>
<td>81.9</td>
<td>13.5</td>
</tr>
<tr>
<td>NB no s</td>
<td>14.4</td>
<td>81.4</td>
<td>10.9</td>
</tr>
<tr>
<td>NB massaged</td>
<td>15.4</td>
<td>81.5</td>
<td>6.8</td>
</tr>
<tr>
<td>Tree J48 with s</td>
<td>19.6</td>
<td>85.1</td>
<td>17.9</td>
</tr>
<tr>
<td>Tree J48 no s</td>
<td>19.6</td>
<td>85.0</td>
<td>17.9</td>
</tr>
<tr>
<td>Tree massage</td>
<td>22.9</td>
<td>83.5</td>
<td>6.1</td>
</tr>
</tbody>
</table>

Decreasing acceptance rates may show lower nominal discrimination!
Preferential treatment is a ranking problem

- No discrimination baseline: random order
- Maximum discrimination: all members of the favored community go before all the members of the protected community

Žliobaitė 2015, discrimination-accuracy trade-offs
Normalized measures

- We propose to normalize discrimination by \( d_{\text{max}} \)
  - \( d = D/d_{\text{max}} \), where
    - 1 max, 0 no discrimination, <0 reverse discrimination
  
- We recommend normalizing accuracy – Cohen's kappa
  - \( k = (\text{Acc} - \text{RAcc}) / (1 - \text{Racc}) \)
    - 1 max, 0 like random, <0 very bad
Experiment cont.
Tradeoffs
Baselines and tradeoffs

Census income dataset

Kamiran et al 2010
Testing on discriminatory data?

- **accuracy**
  - 100%

- **discrimination**
  - 0

- **data**
  - random/constant

- **predictive models**
Testing on discriminatory data?

Shouldn't removing discrimination improve accuracy?
Testing on discriminatory data?

Shouldn't removing discrimination improve accuracy?
Algorithm auditing?
Deon - Data Science Ethics Checklist

- Use with caution!

An ethics checklist for data scientists

deon is a command line tool that allows you to easily add an ethics checklist to your data science projects. We support creating a new, standalone checklist file or appending a checklist to an existing analysis in many common formats.

déon • (déon) [n.] (Ancient Greek) wiktionary

Duty; that which is binding, needful, right, proper.
A. Data Collection

☐ **A.1 Informed consent**: If there are human subjects, have those subjects have given informed consent, where users clearly understand what they are consenting to and there was a mechanism in place for gathering consent?

☐ **A.2 Collection bias**: Have we considered sources of bias that could be introduced during data collection and survey design and taken steps to mitigate those?

☐ **A.3 Limit PII exposure**: Have we considered ways to to minimize exposure of personally identifiable information (PII) for example through anonymization or not collecting information that isn't relevant for analysis?

B. Data Storage

☐ **B.1 Data security**: Do we have a plan to protect and secure data (e.g., encryption at rest and in transit, access controls on internal users and third parties, access logs, and up-to-date software)?

☐ **B.2 Right to be forgotten**: Do we have a mechanism through which an individual can request their personal information be removed?

☐ **B.3 Data retention plan**: Is there a schedule or plan to delete the data after it is no longer needed?
C. Analysis

☐ C.1 Missing perspectives: Have we sought to address blindspots in the analysis through engagement with relevant stakeholders (e.g., checking assumptions and discussing implications with affected communities and subject matter experts)?

☐ C.2 Dataset bias: Have we examined the data for possible sources of bias and taken steps to mitigate or address these biases (e.g., stereotype perpetuation, confirmation bias, imbalanced classes, or omitted confounding variables)?

☐ C.3 Honest representation: Are our visualizations, summary statistics, and reports designed to honestly represent the underlying data?

☐ C.4 Privacy in analysis: Have we ensured that data with PII are not used or displayed unless necessary for the analysis?

☐ C.5 Auditability: Is the process of generating the analysis well documented and reproducible if we discover issues in the future?

D. Modeling

☐ D.1 Proxy discrimination: Have we ensured that the model does not rely on variables or proxies for variables that are unfairly discriminatory?

☐ D.2 Fairness across groups: Have we tested model results for fairness with respect to different affected groups (e.g., tested for disparate error rates)?

☐ D.3 Metric selection: Have we considered the effects of optimizing for our defined metrics and considered additional metrics?

☐ D.4 Explainability: Can we explain in understandable terms a decision the model made in cases where a justification is needed?

☐ D.5 Communicate bias: Have we communicated the shortcomings, limitations, and biases of the model to relevant stakeholders in ways that can be generally understood?
E. Deployment

☐ **E.1 Redress**: Have we discussed with our organization a plan for response if users are harmed by the results (e.g., how does the data science team evaluate these cases and update analysis and models to prevent future harm)?

☐ **E.2 Roll back**: Is there a way to turn off or roll back the model in production if necessary?

☐ **E.3 Concept drift**: Do we test and monitor for concept drift to ensure the model remains fair over time?

☐ **E.4 Unintended use**: Have we taken steps to identify and prevent unintended uses and abuse of the model and do we have a plan to monitor these once the model is deployed?
Algorithm auditing?

• Assessing a particular model
  - in the context of application on a particular dataset/population
  - with respect to a particular set of sensitive variables and
  - a particular legislation

• Assessing the modeling process
  - Assessing the procedure of how models are made
  - Authorities? Methodology?
  - Checklist everything and there will be no information left to learn upon
AI is not there yet to replace human morality
Fairness-awareness in AI needs to be explicit

There are no “right” or “wrong” variables