Improving on Fairness/Bias

Hari Bandi and Dimitris Bertsimas

MIT

Based on "The price of Diversity"
Discharge to post acute care

Percentage of admissions to PAC:

- American Indian: [bar height]
- Asian: [bar height] 12.2%
- Black: [bar height] 12.2%
- Native Hawaiian: [bar height]
- Other race: [bar height]
- White: [bar height] 21.3%
The Problem: Systemic Bias

• Systemic *bias* with respect to *gender, race and ethnicity*, often *unconscious*, but prevalent in datasets involving *choices* made by people.

• Some examples include datasets related to human choices in *college admissions, hiring, lending, or parole* decisions that discriminate against *African-Americans* or *women*.
Summary

• We propose a novel optimization approach to train classification models on large datasets to *alleviate bias* and *enhance diversity* without significantly compromising on meritocracy.

• Key takeaway: The price of diversity is *low* and sometimes *negative*, that is we can modify our selection processes in a way that *enhances diversity without affecting meritocracy* significantly, and sometimes improving it.
Background: Massachusetts General Hospital

• Discharge planning is the development of an individualized discharge plan for a patient prior to leaving hospital for home or to a post acute care (PAC).

• Early prediction of PAC needs prior to discharge leads to

  • reducing hospital length of stay,
  • unplanned readmissions, and
  • improves patient outcomes.
The problem

• The task is to determine discharge disposition for trauma patients within 48-hours after admission.

• Patients are either sent to a post acute care rehab center or home directly after discharge.

• A successful admission into a PAC depends on
  • Patients’ needs,
  • Rehab center agreeing to admit the patient,
  • Patient agreeing to get admitted into a rehab center.
Dataset

- The American College of Surgeons Trauma Quality Improvement Program (ACS-TQIP) database.

- Dataset is sourced from hospitals around the country.

- Features include:
  - patient demographics (age, gender),
  - comorbidities,
  - Emergency Department (ED) vital signs, and
  - injury characteristics (e.g., severity, mechanism).
ML model

• Determine discharge disposition for trauma patients within 48-hours after admission.

• Patients are either sent to a post acute care rehab center or home directly after discharge.

• Build a Logistic Regression model to predict disposition with AUC = 0.79
Notation

- Each of the patients is assigned an outcome $Y = \{-1,+1\}$ representing either *Entering PAC* (+1) or *not* (−1).
  - $W$: set of white patients.
  - $B$: set of black patients.
  - $n_w$: total number of white patients.
  - $n_b$: total number of black patients.
  - $p_w$: total number of white patients who enter PAC.
  - $p_b$: total number of black patients who enter PAC.
We call a dataset $\alpha$-biased if the difference between the rates of positive observations among a pair of subgroups $W$ and $B$ based on a protected variable (in this case, race) is at least $\alpha$.

**Definition 1 ($\alpha$-biased dataset)** A dataset $\mathcal{X} = \{(x_i, y_i) | y_i \in \{-1, 1\}\}$ is said to be $\alpha$-biased with respect to a pair of subgroups $W, B \subseteq \mathcal{X}$ if

$$\left| \frac{\sum_{i \in W} \mathbb{I}(y_i = +1)}{n_w} - \frac{\sum_{i \in B} \mathbb{I}(y_i = +1)}{n_b} \right| \geq \alpha.$$
Demographic parity

- Demographic parity imposes the condition that a classifier $H$ should *predict a positive outcome* for individuals across groups with *almost equal frequency*.

**Definition 2** (Demographic parity) A classifier $\mathcal{H}: X \rightarrow \{-1, 1\}$ achieves demographic parity with bias $\epsilon$ with respect to groups $W, B \subseteq X$ if and only if

$$\left| \frac{\sum_{i \in W} \mathbb{1}(\mathcal{H}(x_i) = +1)}{n_w} - \frac{\sum_{i \in B} \mathbb{1}(\mathcal{H}(x_i) = +1)}{n_b} \right| \leq \epsilon.$$
Example of Demographic parity

Subpopulation-A

Subpopulation-B

\[ \alpha = \frac{1}{3} \]

Demographic parity (\( \alpha = 0 \))

Y = +1

Y = -1

✓ qualified candidate

The Analytics Edge
Proposed solution

• *Flip outcome labels* \((Y)\) while training your model to achieve demographic parity.

• We propose a Mixed-integer Optimization (MIO) problem that achieves this by introducing *binary variables* \(z_i \in \{0, 1\}, i \in [n]\) to decide which outcome labels to flip.
Proposed solution

• If we decide to *flip the outcome label* of the $i^{th}$ observation: $y_i \in \{-1, 1\}$, the resulting outcome label would be $\tilde{y}_i = y_i(1 - 2z_i)$.

• We define a set of $n$ binary variables ($z$) that flip at most $\tau_w$ proportion of labels in $W$ and $\tau_b$ proportion of labels in $B$ given by,

$$
\mathcal{Z}_{\tau_w, \tau_b} = \left\{ z \in \{0, 1\}^n : \frac{\sum_{i \in W} z_i}{n_w} = \tau_w, \frac{\sum_{i \in B} z_i}{n_b} = \tau_b \right\}.
$$
Proposed solution

- The parameters $\tau_w$ and $\tau_b$ are estimated from the data so that the resulting classifier ensures $\varepsilon$-demographic parity.

\[
\tau_w \leq \frac{n_b \cdot p_w}{n_w(n_w + n_b)} - \frac{p_b}{n_w + n_b} + \frac{n_b \cdot \varepsilon}{n_w(n_w + n_b)},
\]

\[
\tau_b \leq \frac{p_w}{n_w + n_b} - \frac{n_w \cdot p_b}{n_b(n_w + n_b)} - \frac{n_w \cdot \varepsilon}{n_b(n_w + n_b)}.
\]
Logistic Regression

• The dependent variable (Y) is a Bernoulli random variable
  • Y = +1 – “Entering PAC”
  • Y = -1 – “Not entering PAC”

• We seek to predict the probability of a success outcome of the dependent variable Y as a function of independent variables \( x_1, x_2 \ldots x_k \)

• We predict the *likelihood* that \( Y = +1 \) as follows:
  
  \[
  \Pr(Y = +1) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}}
  \]

  *This is guaranteed to be between 0 and 1*
Logistic regression

Logistic regression model

\[
\min_{\beta_0, \beta} \sum_{i=1}^{n} \log \left( 1 + e^{-y_i (\beta^T x_i + \beta_0)} \right).
\]

Parameters of a Logistic regression model

MIO model

\[
\min_{\mathbf{z} \in \mathcal{Z}_{\tau_w, \tau_b}} \min_{\beta_0, \beta} \sum_{i=1}^{n} \log \left( 1 + e^{-y_i (1-2z_i) (\beta^T x_i + \beta_0)} \right).
\]

Updated outcome label

Binary variables

The Analytics Edge
MIO model (linearizing product terms)

\[
\begin{align*}
\min_{\mathbf{z}} \min_{\beta, \gamma} & \quad f(\beta, \gamma) := \sum_{i=1}^{n} \log \left( 1 + e^{-y_i(\beta^T x_i + \beta_0) + 2y_i(\gamma_i^T x_i + \gamma_i,0)} \right) \\
\text{s.t} & \quad \sum_{i \in W} z_i = \tau_w \cdot n_w, \\
& \quad \sum_{i \in B} z_i = \tau_b \cdot n_b, \\
& \quad -z_i M_j \leq \gamma_{i,j} \leq z_i M_j, \quad i \in [n], j \in [p], \\
& \quad -(1-z_i) M_j \leq \gamma_{i,j} - \beta_j \leq (1-z_i) M_j, \quad i \in [n], j \in [p], \\
& \quad \sum_{i \in W} \gamma_{i,j} = \tau_w \cdot n_w \cdot \beta_j, \quad j \in [p], \\
& \quad \sum_{i \in B} \gamma_{i,j} = \tau_b \cdot n_w \cdot \beta_j, \quad j \in [p], \\
& \quad z_i \in \{0, 1\}, i \in [n].
\end{align*}
\]
Additional constraints

- Maximize likelihood
- Demographic parity
- Severity of injuries unchanged
- Age and gender distribution unchanged
Predictive performance

AUC on original outcome labels

AUC: 0.78

AUC on modified outcome labels

AUC: 0.87

• Alleviating bias improves out-of-sample AUC of OCTs by 8-10%.
Implementation tool

• Train Optimal Classification Trees (OCTs) to provide insights on which attributes of individuals lead to flipping of their labels.

• Construct a dataset based on output of the MIO model. Each defendant is labeled as one of the following:
  • negative (patient discharged to home),
  • high (patient discharged to PAC), or
  • no change (outcome label unchanged)
Implementation tool

Left part of the OCT after splitting on Head severity \leq 1.0
Implementation tool

Right part of the OCT after splitting on Head severity ≥ 2.0.
Other applications

• Admissions

• Parole

• Bar exam
Key takeaways

• Demonstrate how alleviating bias can improve selection processes in practice.

• Develop a highly interpretable implementation tool to make changes to the current selection processes to improve diversity.

• Alleviating bias improves predictive performance.