Trustworthy AI

Spring 2024

Yuan Tian

#15: Machine Learning Fairness

Reminders

- Final course project presentation will be on June 5
  - 5 minutes presentation + 2 minutes Q&A
  - 50% peer eval, 50% Yuan's eval
  - Group assignment on Gradescope
- Course evaluation (1% bonus) due on June 7
- Final report due on June 10
  - Group assignment on Gradescope
- Final project presentation evaluation and teammate evaluation (course project) due on June 11

What is Fairness?

- Sameness
  - Everybody is equal.
- Deservedness
  - You get what you deserve, e.g. If you work hard, you succeed.
- Treating Same Individuals Similarly

Uses of Fairness in ML

Candidate evaluations for job positions

Lending trustworthiness assessments

Personalized product recommendations

Goal: Prevent discrimination against individuals based on their membership in some group, while maintaining utility for the classifier

Sources of Unfairness

- Bias in data
  - Data collection: temporal, behavioural and geographical biases
  - Imbalance data or imbalance labels (more labels for one race)
  - Historical biases: gender roles in texts and images, racial stereotypes in languages
  - Inappropriate data handling
- Model
  - Inappropriate model selection
  - Incorrect algorithm design or application

Fairness in Supervised Learning

Formal Setup:

- Available features X (e.g. credit history)
- Protected attribute A or S (e.g. race, gender)
- Prediction target Y (e.g. load defaulting)
- Learn predictor $\hat{\mathrm{Y}}(\mathrm{X})$ or $\hat{\mathrm{Y}}(\mathrm{X}, \mathrm{A})$ for Y

Definitions of Fairness

There are many ways to describe fairness.

- Fairness Through Unawareness
- Individual Fairness: Each two similar individuals should be classified similarly
- Group Fairness: Model's outcome should be same across different subgroups
  - Statistical (Demographic) Parity
  - Equality of Odds (Paper 3)
  - Equality of Opportunity (Paper 3)

Fairness Through Unawareness

- It has been the default fairness method in machine learning
- Refers to leaving out protected attributes such as gender, race, and other characteristics deemed sensitive.
- Ineffective: protected variables could be correlated with other variables in the data ⇒ *redundant encodings*

~~Race~~ ← Postal code

Recap: Common Metrics

True Class

Predicted Class	**Positive**	**Negative**
Positive	TP	FP
Negative	FN	TN

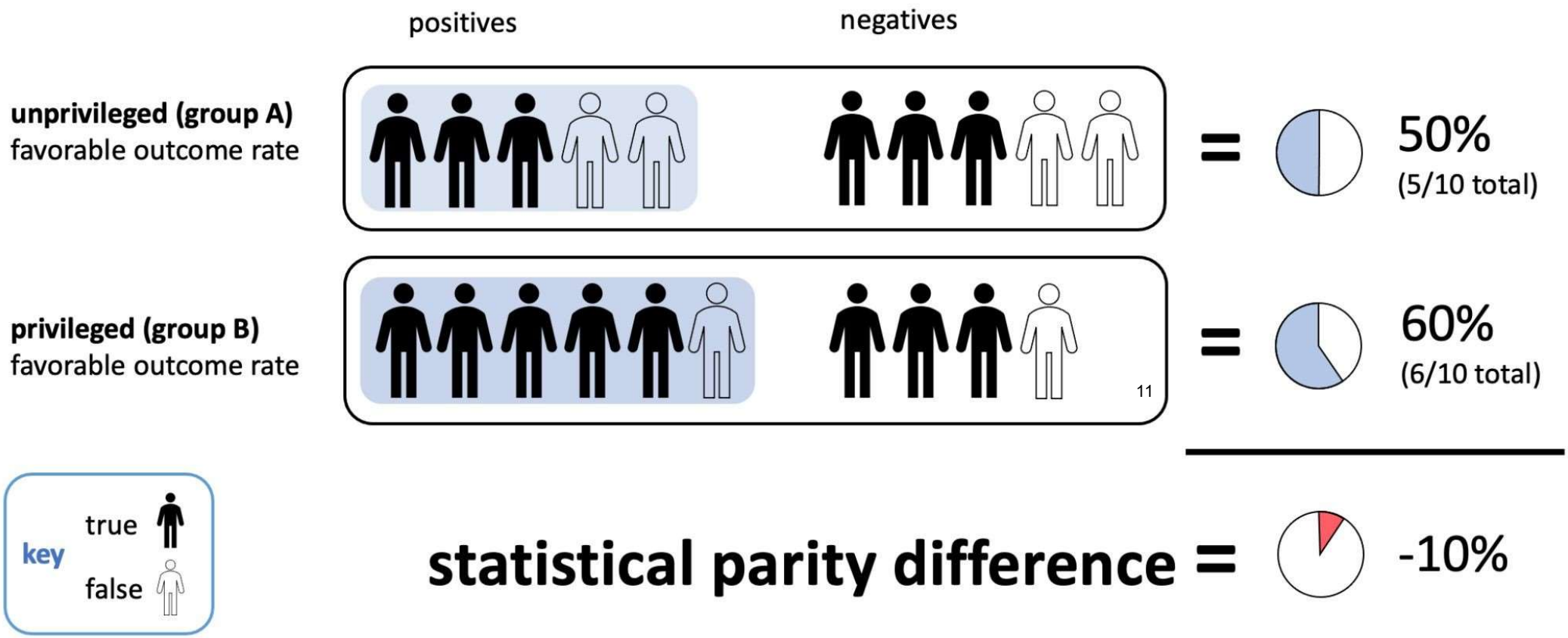
Statistical Parity (Demographic Parity)

- Statistical parity states that the proportion of each subclass of a protected class (e.g. gender or race) should receive outcomes at equal rates

$$P(\hat{Y} = 1 \mid A = a) = P(\hat{Y} = 1 \mid A = b)$$

- When to use this notion of fairness?
  - Neutralizes redundant encodings
  - Does not prevent all unfairness, especially regarding subsets of each subclass

Statistical Parity (Demographic Parity)



Equality of Odds

Equality of odds is satisfied if the prediction $\hat{Y}$ is conditionally independent to the protected attribute A, given the true value Y:

$$\mathrm{P}(\hat{Y}|Y, \mathrm{P}) = \mathrm{P}(\hat{Y}|\ Y)$$

This means that the true positive rate and false positive rate will be the same for each population

Equality of Opportunity

It is similar to the definition of equality of odds, except it is focused on the particular label of Y = 1:

$$\mathrm{P}(\hat{Y}|Y = 1, \mathrm{P}) = \mathrm{P}(\hat{Y}|\ Y = 1)$$

It states that each group should get the positive outcome at equal rates, assuming that people in this group qualify for it.

Timeline of Papers

2011 - *Fairness Through Awareness*

2013 - *Learning Fair Representations*

2016 - *Equality of Opportunity in Supervised Learning*

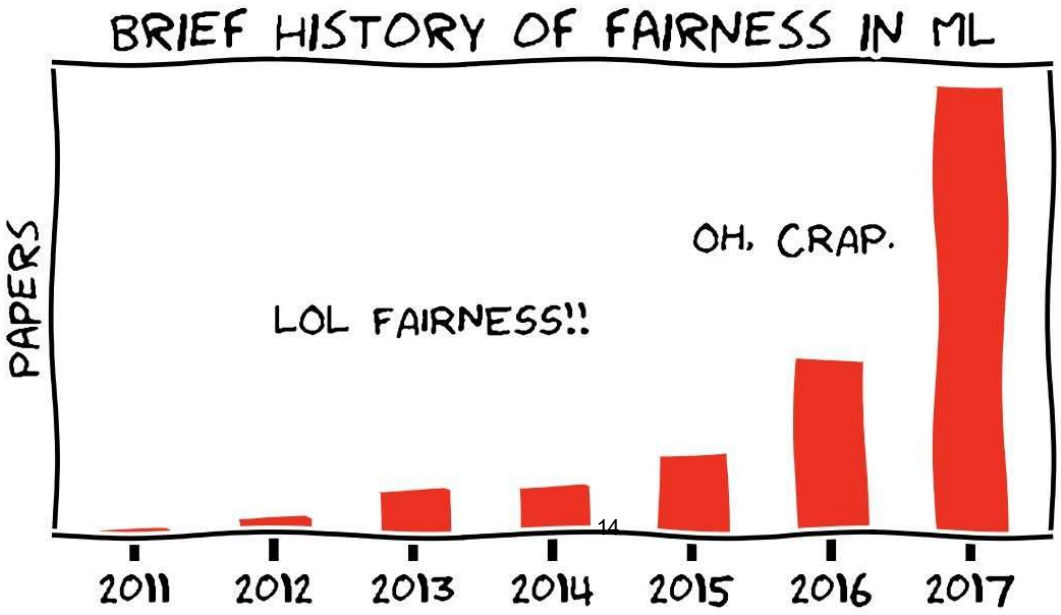


Image Credit: https://fairmlclass.github.io/

Some interesting resources for further reading

- https://fairmlbook.org
- https://fairmlclass.github.io/

01

Fairness Through Awareness

Cynthia Dwork, Moritz Hardt, Toniann Pitassi,
Omer Reingold,
Richard Zemel

November 2011

Fairness through awareness

General idea:

- A framework for **Individual Fairness**
- Formulated as a linear optimization problem
- Evaluates the “alikeness” between members requiring classification

Prevents against:

- Explicit discrimination

Assumptions

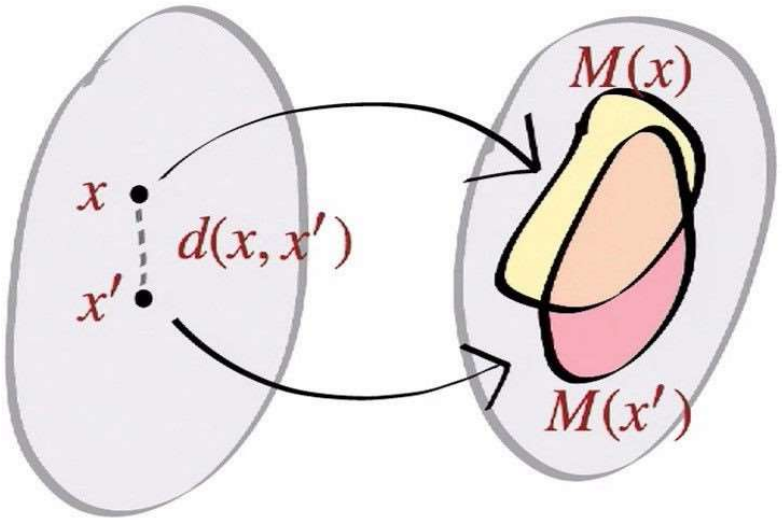
Basic formulation

- Individuals V ⟶ Outcomes A
- Distributions mappings M: $V \to \Delta(A)$

$$M = \{\mu_x\}_{x \in V} \text{ where } \mu_x = M(x) \in \Delta(A).$$

- "Similarity of individuals distance" *d:* $V \times V \to \mathbb{R}$
- "Similarity of distributions distance" D
- **Lipschitz mapping:**
  - $D(Mx, My) \leq d(x, y)$

Idea: "Map similar people similarly"



Basic formulation

There are many possible classifiers that satisfy the Lipschitz condition $D(Mx, My) \leq d(x, y)$

How to pick which classifier to use?

Choose by a loss function L, reducing the decision to an optimization problem:

$$\begin{aligned} \mathrm{opt}(\mathcal{I}) \stackrel{\mathrm{def}}{=} & \min_{\{\mu_x\}_{x \in V}} \mathop{\mathbb{E}}_{x \sim V} \mathop{\mathbb{E}}_{a \sim \mu_x} L(x, a) \\ \text{subject to} \quad & \forall x, y \in V, : \quad D(\mu_x, \mu_y) \leq d(x, y) \\ & \forall x \in V : \quad \mu_x \in \Delta(A) \end{aligned}$$

Choosing d and D

d represents the distance between individuals

- Different for each task
- Challenging to quantify, especially when there are many variables
- If chosen poorly, has the potential to introduce bias

D represents the distance in the output space

- More quantifiable as D only depends on the output distribution
- Challenge lies in ensuring d and D are comparable

Potential D metrics: statistical distance

Let P, Q denote probability measures on a finite domain A;

D as total variation norm (statistical distance):

$$D_{\text{tv}}(P, Q) = \frac{1}{2} \sum_{a \in A} |P(a) - Q(a)|$$

Limitations of D_{tv}:

Requires that d is scaled between {0, 1}.

Potential D metrics: relative l_∞ metric

Let P, Q denote probability measures on a finite domain A

D as the relative l_∞ metric:

$$D_\infty(P, Q) = \sup_{a \in A} \log\left(\max\left\{\frac{P(a)}{Q(a)}, \frac{Q(a)}{P(a)}\right\}\right)$$

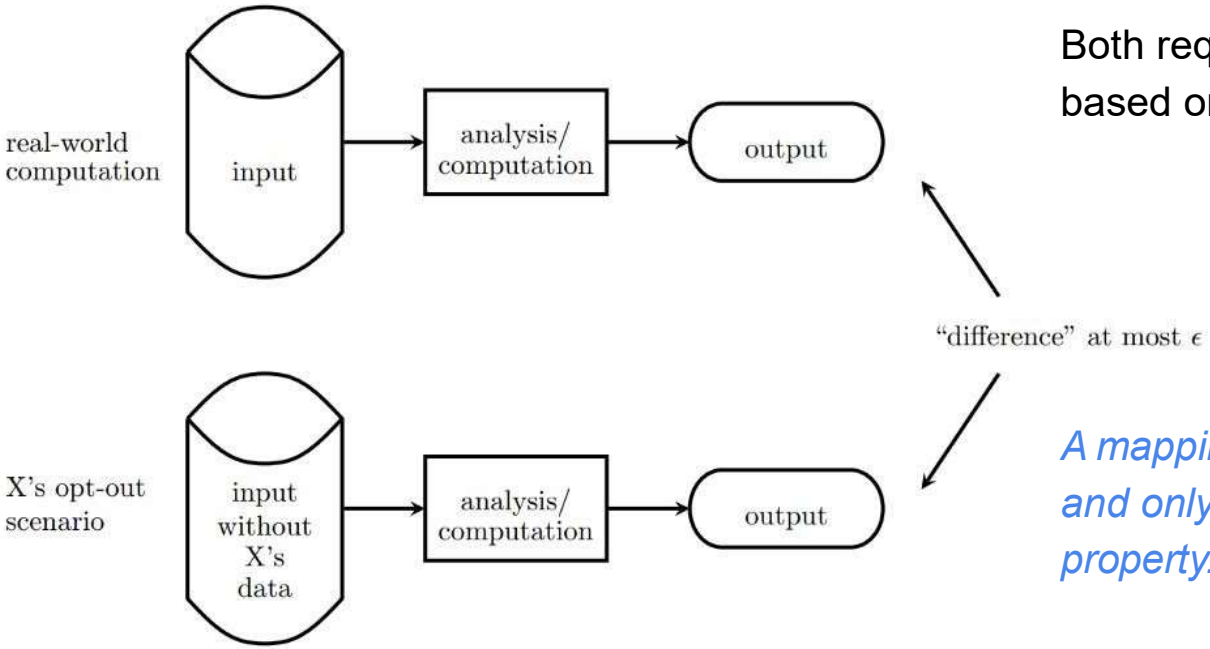
Better metric as D_∞ imposes a strong constraint for $d(x, y)$ << 1 and a weak constraint for $d(x, y)$ >> 1.

Both metrics can be computed with a program of size poly($|V|$, $|A|$).

Relation to differential privacy

Recall: Differential privacy is a system for describing the patterns of groups within the dataset while withholding information about individuals in the dataset.

ε-differential privacy:



Similarity with (D, d)-Lipschitz principle:

Both require that outputs differ minimally based on the difference between inputs

A mapping M satisfies ε-differential privacy if and only if M satisfies the (D_∞, d)-Lipschitz property.

Summary / Contributions

- Introduced a framework for characterizing individual fairness
- Outlined an optimization approach to maximize functionality while maintaining a strict level of fairness
- Determined when the approach implies statistical parity
- Provided alternative formation that enforces statistical parity

Limitations and Future Work

Limitations

- Assume the data owner is trustworthy
- Provides a local solution; does not solve the global problem
- Does not provide a clear definition of the distance metric d
- Only explores effects on disjoint subsets
- Only true for the set of real individuals and does not generalize to all possible individuals
    - It is not a learning problem

Discussion questions:

- Potential usage of *metric labeling* to build a d metric
- Does fairness hide information?

02

Learning Fair Representations

Richard Zemel, Yu Wu, Kevin Swerksy, Toniann Pitassi, Cynthia Dwork

Motivation

In the previous work [Fairness through Awareness]:

- The distance metric that defines the similarity between the individuals is assumed to be given which is unrealistic.
  - The problem of establishing fairness will be reduced to finding a fairness distance metric
- It is not formulated as a learning problem, and cannot generalize.
  - it forms a mapping for a given set of individuals without any procedure for generalizing to novel unseen data.

General Idea

Find a representation Z from data to remove sensitive information

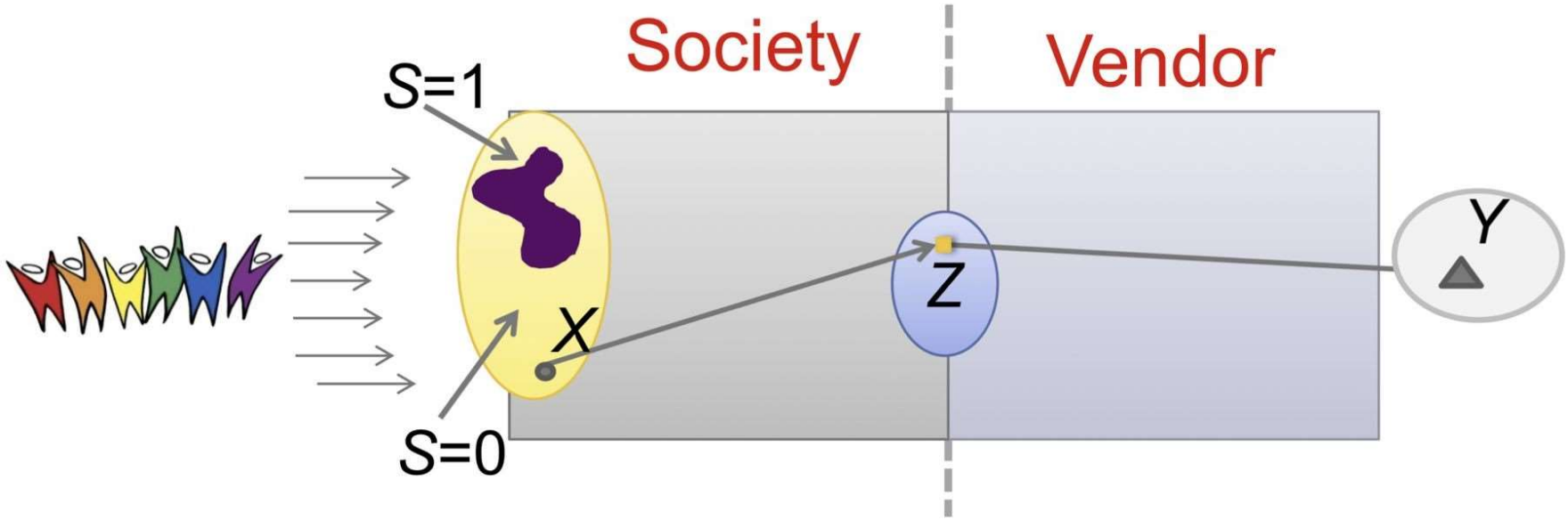


Image Credit: Richard Zemel

Naive Solutions

- Removing the sensitive attributes (fairness through unawareness)
  - But the other attributes combined together can reveal some information
- Using only a small subset of attributes that we are sure they don't reveal any information
  - It may hurt the accuracy of the model
- Relabeling training data or changing sensitive attributes
  - It may hurt the accuracy of the model
  - Still may leak some information

“Learning Fair Representation (LFR)” in a nutshell

Formulating fairness as an **optimization problem** of finding a good representation of the data with two competing goals:

- **Encode** the data as well as possible ⟹ **Preserve utility**
- **Obfuscate** any sensitive information ⟹ **Protect sensitive groups**

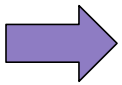
The main idea is to map each individual, represented as a data point in a given input space, to a probability distribution in a new representation space.

Notation

- X is the dataset where each $x \in X$ is a D-dimensional vector
  - X_o: training dataset
- $S = \{0, 1\}$ is the sensitive attribute.
  - $X+$: data points with $S = 1$
  - $X-$: data points with $S = 0$
- $Y = \{0,1\}$ is the target labels.
- Z is a multinomial random variable, where each of the K values represents one of the intermediate set of ”prototypes”. Associated with each prototype is a vector v_k in the same space as the individuals x.
- $d(x, x')$ is the distance measure on X (e.g. Euclidean distance).

Recall: Goals of the Representation

Goals	Meaning
Obfuscate S (Fairness)	Low mutual information between Z and S
Preserve information in X (Reconstruction)	High mutual information between X and Z
High classification accuracy (Accuracy)	High mutual information between Z and Y



Probabilistic Mapping

Given the definitions of the prototypes as points in the input space, a set of prototypes induces a natural probabilistic mapping from X to Z via the softmax:

$$P(Z = k|\mathbf{x}) = \exp(-d(\mathbf{x}, \mathbf{v}_k))/\sum_{j=1}^{K} \exp(-d(\mathbf{x}, \mathbf{v}_j))$$

Similar to (soft) K-Means Clustering. Prototypes acts as clusters.

Recall: K-Means vs. Soft K-Means

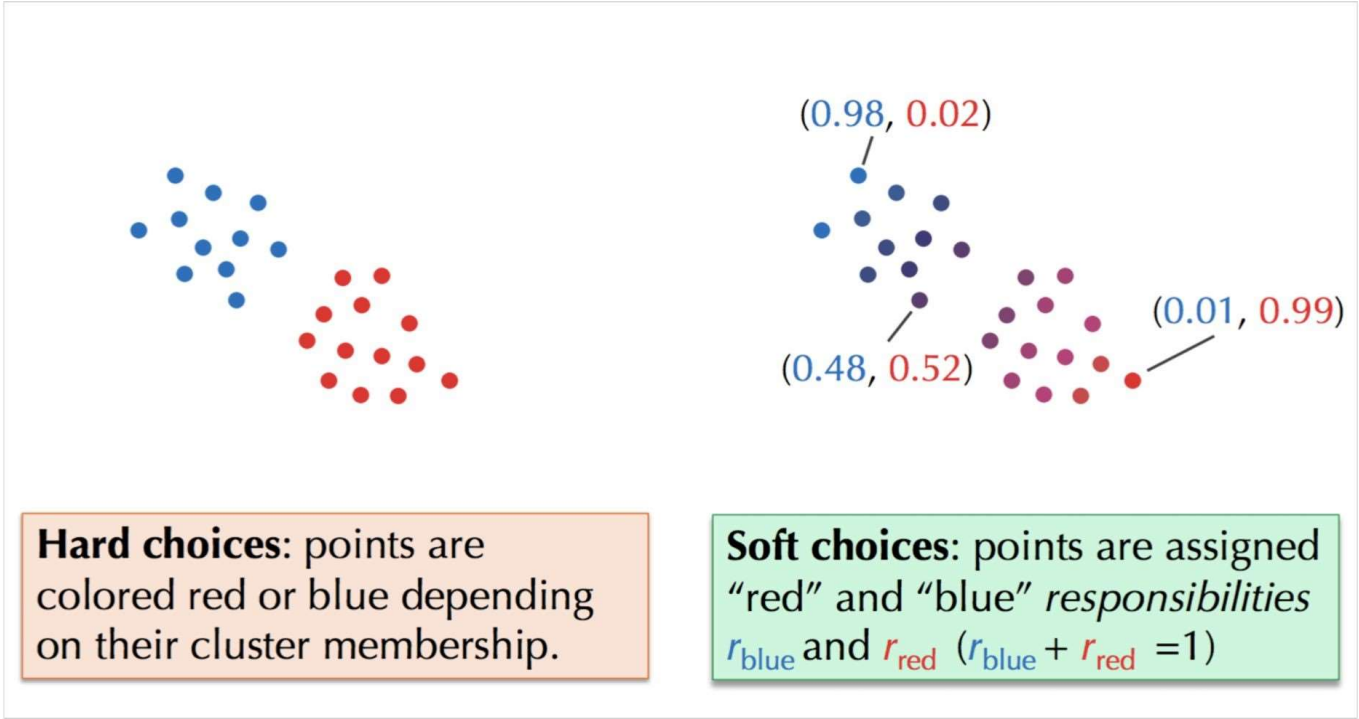


Image Credit: https://www.cs.cmu.edu/~02251/recitations/recitation_soft_clustering.pdf

LFR Objective Function

The LFR model aims to minimize the following objective function:

$$L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y$$

Fairness Term

Reconstruction Term

Accuracy Term

Where A_x, A_y and A_z are hyperparameters governing the trade-offs.

LFR Objective Function: Fairness

Loss function to ensure the group fairness:

$$L_z = \sum_{k=1}^{K} \left| M_k^+ - M_k^- \right|$$

Where:

$$M_k^+ = \mathbb{E}_{\mathbf{x} \in X^+} P(Z = k|\mathbf{x}) = \frac{1}{|X_0^+|} \sum_{n \in X_0^+} M_{n,k}$$

$$M_{n,k} = P(Z = k|\mathbf{x}_n) \quad \forall n, k$$

Meaning: Each class should contain roughly a same ratio from the protected and unprotected group

LFR Objective Function: Reconstruction

Loss function for the reconstruction term:

$$L_x = \sum_{n=1}^{N} (\mathbf{x}_n - \hat{\mathbf{x}}_n)^2$$

Where:

$$\hat{\mathbf{x}}_n = \sum_{k=1}^{K} M_{n,k} \mathbf{v}_k \qquad M_{n,k} = P(Z = k|\mathbf{x}_n) \quad \forall n, k$$

Meaning: The learned representation of the data should resemble the actual data and contain as much information as possible.

LFR Objective Function: Accuracy

Loss function to ensure accuracy:

$$L_y = \sum_{n=1}^{N} -y_n \log \hat{y}_n - (1 - y_n) \log(1 - \hat{y}_n)$$

Where:

$$\hat{y}_n = \sum_{k=1}^{K} M_{n,k} w_k \qquad M_{n,k} = P(Z = k | \mathbf{x}_n) \quad \forall n, k$$

Meaning: The learned representation should still predict labels with high accuracy

LFR Distance Metric

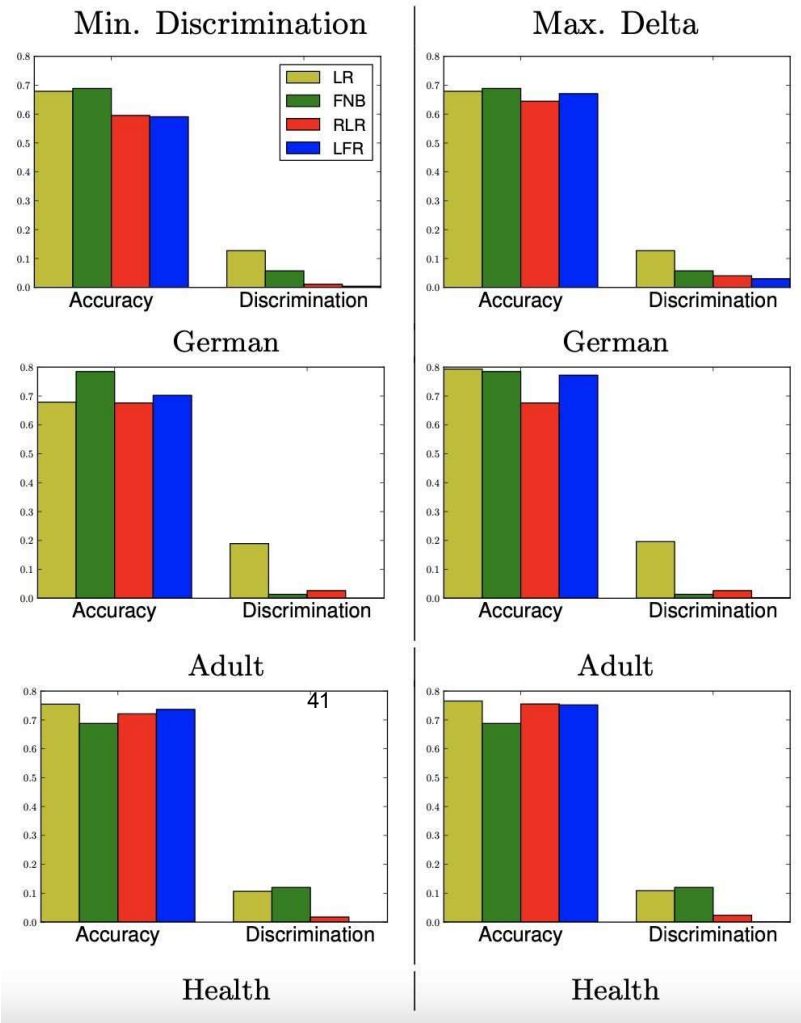
To allow different input features to have different levels of impact, they introduce individual weight parameters for each feature dimension, α_i, which act as inverse precision values in the distance function:

$$d(\mathbf{x}_n, \mathbf{v}_k, \alpha) = \sum_{i=1}^{D} \alpha_i (x_{ni} - v_{ki})^2$$

More flexible than Euclidean distance

Experiments

Results on test sets for the three datasets (German, Adult, and Health), for two different model selection criteria: minimizing discrimination and maximizing the difference between accuracy and discrimination.

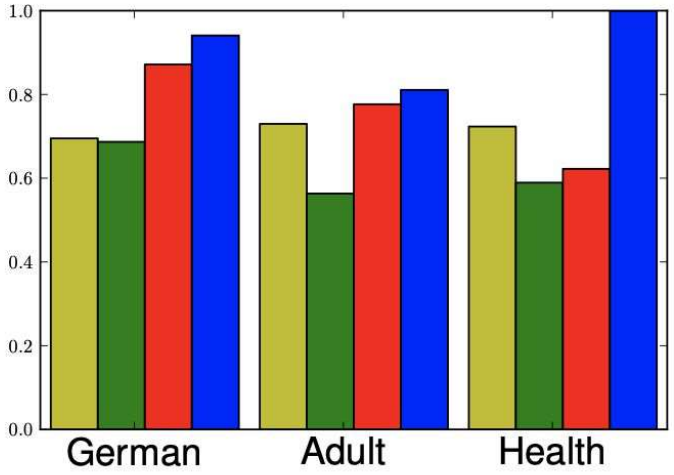


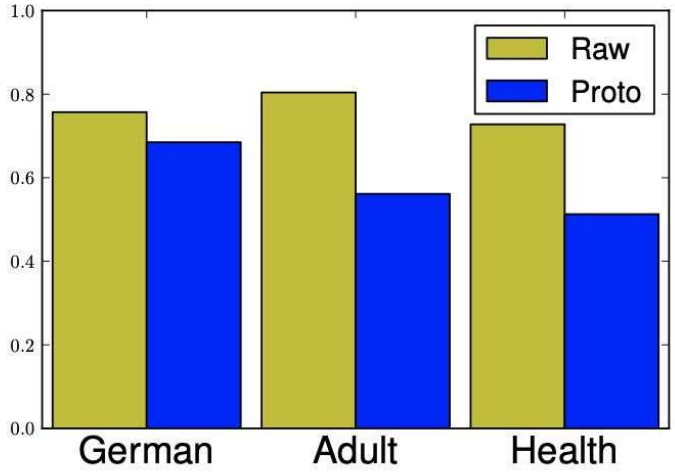
Experiments (cont.)

- Individual fairness: Comparing consistency of each model's classification decisions, based on the yNN measure.

$$yNN = 1 - \frac{1}{Nk} \sum_{n} |\hat{y}_n - \sum_{j \in kNN(\mathbf{x}_n)} \hat{y}_j|$$

- Accuracy of predicting the sensitive variable for the different datasets.
  - Raw: predictions using the input by removing s
  - Proto: predictions using the LFR





Limitations and Future Work

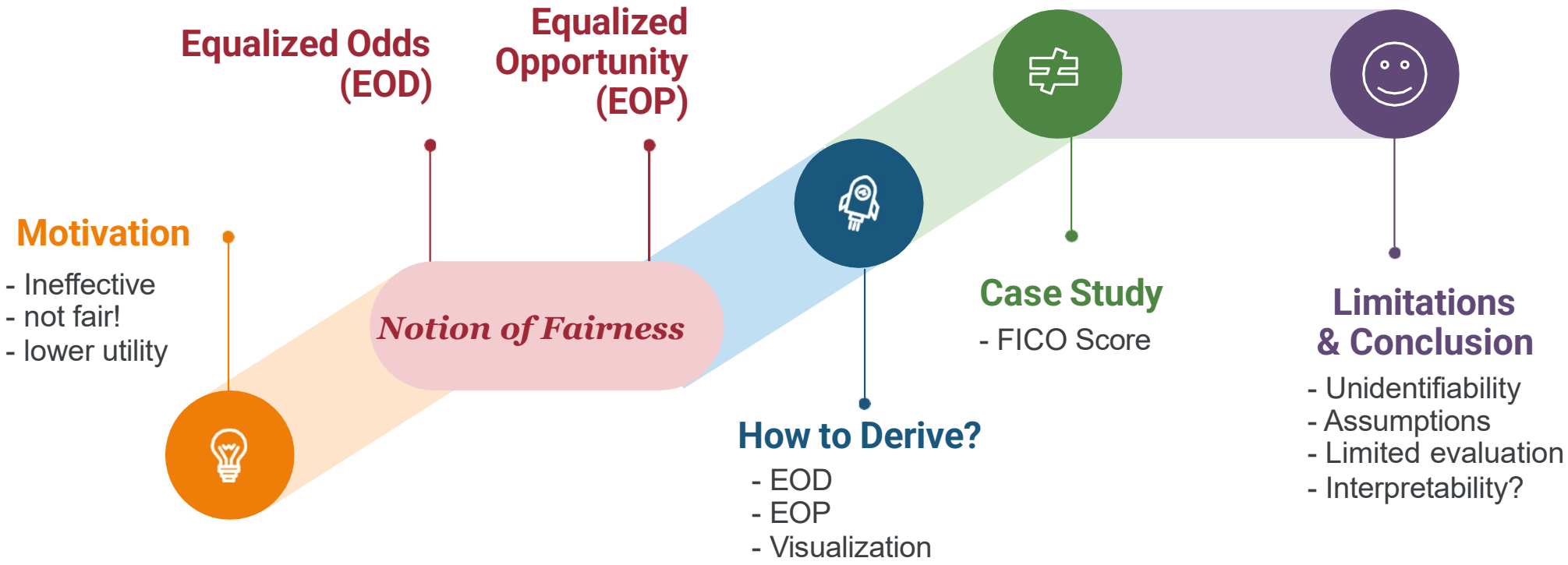
- They didn't evaluate their approach in high-dimensional datasets (e.g. images) and for more than two protected subgroups.
- This approach mostly considers group fairness or statistical parity. It is worth trying to apply other notions of fairness.
- It would be interesting to investigate the tradeoffs between fairness and accuracy more thoroughly.
- It would be interesting to investigate the relation of fairness and privacy. Do we achieve some levels of privacy by using these fair representations?

03

Equality of Opportunity in Supervised Learning

Moritz Hardt,
Eric Price,
Nathan Srebro

Outline


Equalized Odds (EOD)
Equalized Opportunity (EOP)
Motivation
- Ineffective
- not fair!
- lower utility
Notion of Fairness
How to Derive?
- EOD
- EOP
- Visualization
Case Study
- FICO Score
Limitations & Conclusion
- Unidentifiability
- Assumptions
- Limited evaluation
- Interpretability?

Motivation

Question: What does it mean for Y to be fair?

Existing notions:

- Fairness through unawareness:
  - Ineffective due to redundant encodings
- Demographic parity
  - Doesn't ensure fairness
  - Decrease utility ⇒ cannot achieve perfect accuracy!

~ **Solution:** establish a new notion of fairness in supervised learning

New Notion of Fairness: Equality

Goal & Requirement ⇒ align accuracy & fairness!

- Measure of discrimination
- High utility ⇒ allows perfect accuracy of $\hat{Y} = Y$
- Better incentive

Proposed Notion

- Predict a true outcome Y from features X based on labeled training data
- Does not discriminate with respect to a specified protected attribute A
- Oblivious: based only on the joint distribution of $(Y, \hat{Y}, A)$

~ **Equalized Odds** and **Equalized Opportunity**

Equalized Odds (EOD)

Predictor $\hat{Y}$ satisfies EOD with respect to protected attribute A and outcome Y, if $\hat{Y}$ and A are independent conditional on Y

$$\Pr\left\{\widehat{Y}=1 \mid A=0, Y=y\right\}=\Pr\left\{\widehat{Y}=1 \mid A=1, Y=y\right\}, \quad y \in\{0,1\}$$

Features:

- Equalize TP and FP rates
- Align fairness with accuracy
  - Allow $\hat{Y} = Y$ as a solution
  - Enforce accuracy in all classes (not only the majority)

Equalized Odds (EOD)

Predictor $\hat{Y}$ satisfies EOD with respect to protected attribute A and outcome Y, if $\hat{Y}$ and A are independent conditional on Y

$$\Pr\left\{\widehat{Y}=1 \mid A=0, Y=y\right\} = \Pr\left\{\widehat{Y}=1 \mid A=1, Y=y\right\}, \quad y \in \{0,1\}$$

Compared to **Demographic Parity**:

- EOD allows $\hat{Y}$ to depend on A but only through the target variable Y
- Encourage the use of features that allow to directly predict Y
- Prohibit abusing A as a proxy for Y

Equalized Odds (EOD)

Predictor $\hat{Y}$ satisfies EOD with respect to protected attribute A and outcome Y, if $\hat{Y}$ and A are independent conditional on Y

$$\Pr\left\{\widehat{Y}=1 \mid A=0, Y=y\right\}=\Pr\left\{\widehat{Y}=1 \mid A=1, Y=y\right\}, \quad y \in\{0,1\}$$

Example: job hiring, A = gender; 2 female (1 qualified) + 3 male (1 qualified)

- Demographic Parity: *female hiring rate % = male hiring rate %*
- EOD: *qualified female hiring rate % = qualified male hiring rate %*
  unqualified female hiring rate % = unqualified male hiring rate %

~ EOD can be fair while being perfectly accurate!

Equalized Opportunity (EOP)

Predictor $\hat{Y}$ satisfies EOP with respect to protected attribute A and outcome Y when $Y=1$, if $\hat{Y}$ and A are independent conditional on Y

$$\Pr\{\widehat{Y} = 1 \mid A = 0, Y = 1\} = \Pr\{\widehat{Y} = 1 \mid A = 1, Y = 1\}$$

Y = y

Y = y

EOP vs. EOD:

- weaker constraint
- but allows stronger utility

Equalized Opportunity (EOP)

Predictor $\hat{Y}$ satisfies EOP with respect to protected attribute A and outcome Y when $Y=1$, if $\hat{Y}$ and A are independent conditional on Y

$$\Pr\left\{\widehat{Y}=1 \mid A=0, Y=1\right\}=\Pr\left\{\widehat{Y}=1 \mid A=1, Y=1\right\}$$

Hiring Example:

~ hired people regardless their gender should have been offered with the same opportunity!

- Demographic Parity: *female hiring rate % = male hiring rate %*
- EOD: *qualified female hiring rate % = qualified male hiring rate %* ⟵ **EOP**
  unqualified female hiring rate % = unqualified male hiring rate %

Finding EOD/EOP Predictor $\hat{Y}$

- Goal:
  - Find a non-discriminating predictor $\tilde{Y}$ derived from a (possibly discriminatory) learned model
- Based on the existing training pipeline of the problem, models can be:
  - Binary predictor $\hat{Y}$
  - Score R
- Post-learning process:
  - Do not require changes in training process
- Oblivious

How?

~ 1) minimizing the loss function 2) given some constraints

Predictor $\hat{Y}$ Constraints (Binary)

A predictor $\hat{Y}$ satisfies:

1. EOD, if and only if $\gamma_0(\hat{Y}) = \gamma_1(\hat{Y})$
2. EOP, if and only if $\gamma_0(\hat{Y})_2 = \gamma_1(\hat{Y})_2$

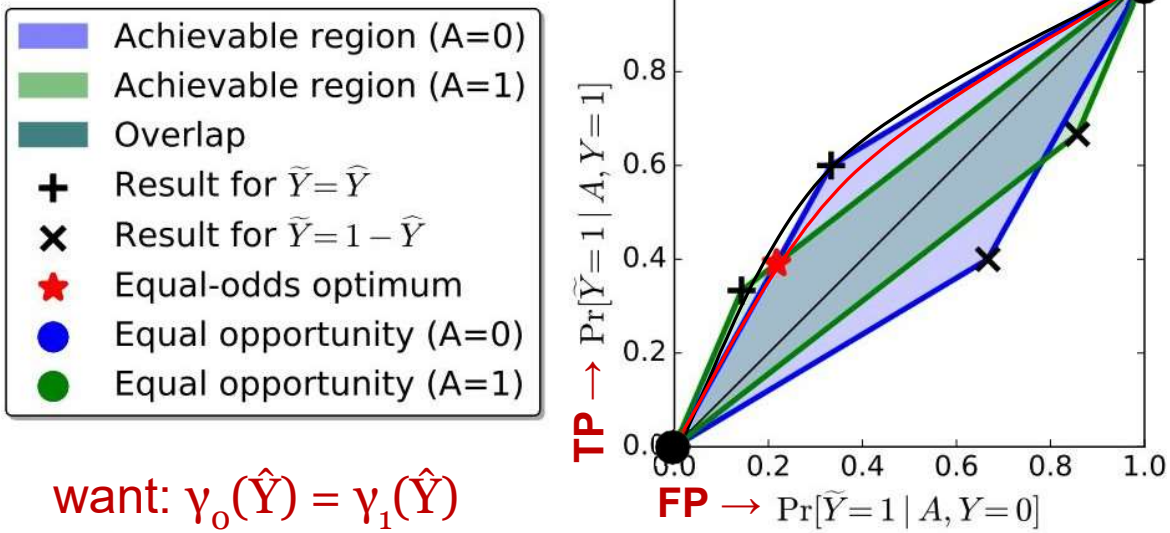
~ But! Trivial if without Loss Minimization

$$\min_{\widetilde{Y}} \quad \mathbb{E}\ell(\widetilde{Y}, Y)$$

Find the best (fair) predictor with a minimal cost **~ accuracy and fairness!**

Visual Representation: EOD

$$\Pr\left\{\widehat{Y}=1 \mid A=0, Y=y\right\} = \Pr\left\{\widehat{Y}=1 \mid A=1, Y=y\right\}, \quad y \in \{0,1\}$$



want: $\gamma_0(\hat{Y}) = \gamma_1(\hat{Y})$

Figure: Finding the optimal equalized odds predictor

x and +: min loss results

Intersections: ●

- To satisfy the EOD constraint: intersect

Non-trivial Intersection: ★

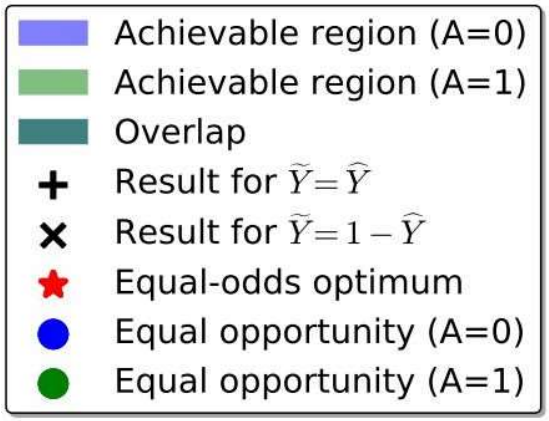
- Also minimize loss

~ EOD: result lies below all ROC curves

Visual Representation: EOP

want: $\gamma_0(\hat{Y})_2 = \gamma_1(\hat{Y})_2$

$$\Pr\left\{\widehat{Y} = 1 \mid A = 0, Y = 1\right\}$$
$$= \Pr\left\{\widehat{Y} = 1 \mid A = 1, Y = 1\right\}$$



EOP Constraint:

- Won't show intersections because no restriction on the FP-axis
- Only focus on the TP-axis ($\gamma_0(\hat{Y})_2 = \gamma_1(\hat{Y})_2$)
- care only about Y=1, so shifting along the FP-axis while maintaining the same TP rate

Loss Minimization:

- Consider both TP and FP
- Non + points are all worse

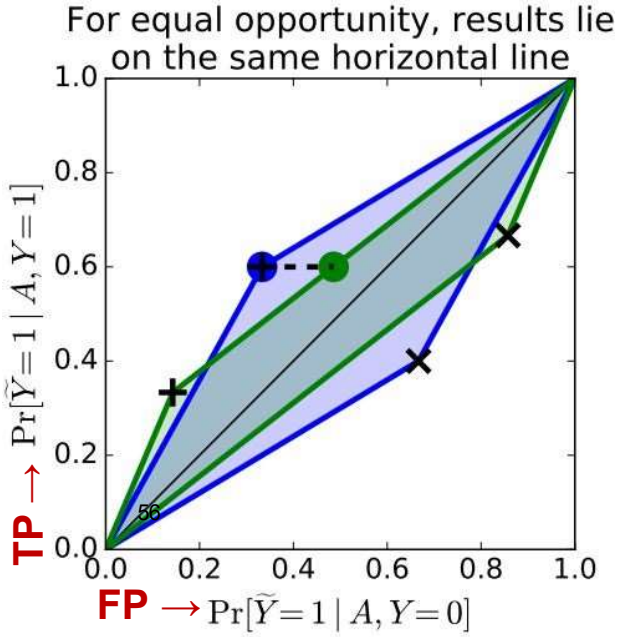
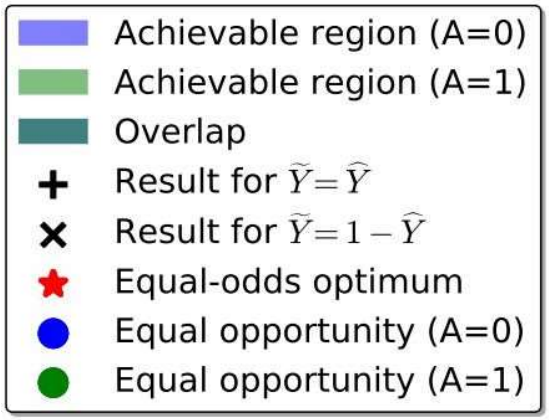


Figure: Finding the optimal equal opportunity predictor (right).

Visual Representation: EOP

want: $\gamma_0(\hat{Y})_2 = \gamma_1(\hat{Y})_2$

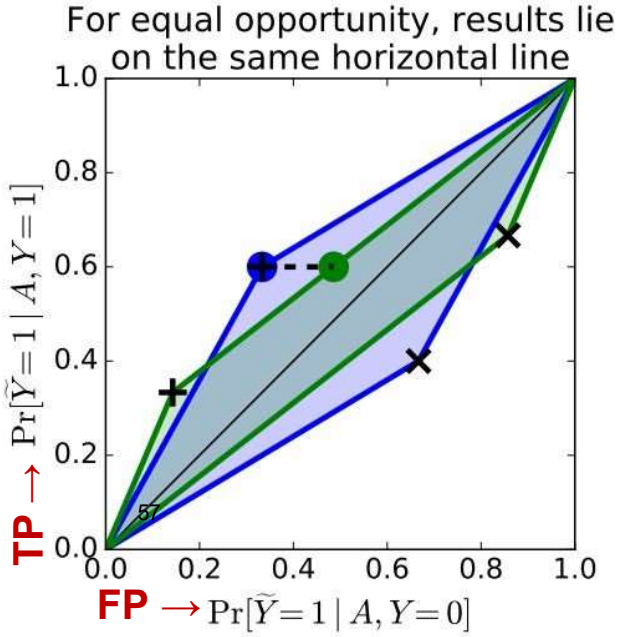
$$\Pr\{\widehat{Y} = 1 \mid A = 0, Y = 1\}$$
$$= \Pr\{\widehat{Y} = 1 \mid A = 1, Y = 1\}$$



Example:

Female vs male job hiring
Suppose A=0 is race1, A=1 is race2;
From the figure we can see that the final results of EOD should be:

- Rate of high-scored race1 people getting hired = 0.6
- Rate of high-scored race2 people getting hired = 0.6
- Rate of low-scored race1 people getting hired = 0.3
- Rate of low-scored race2 people getting hired = 0.5

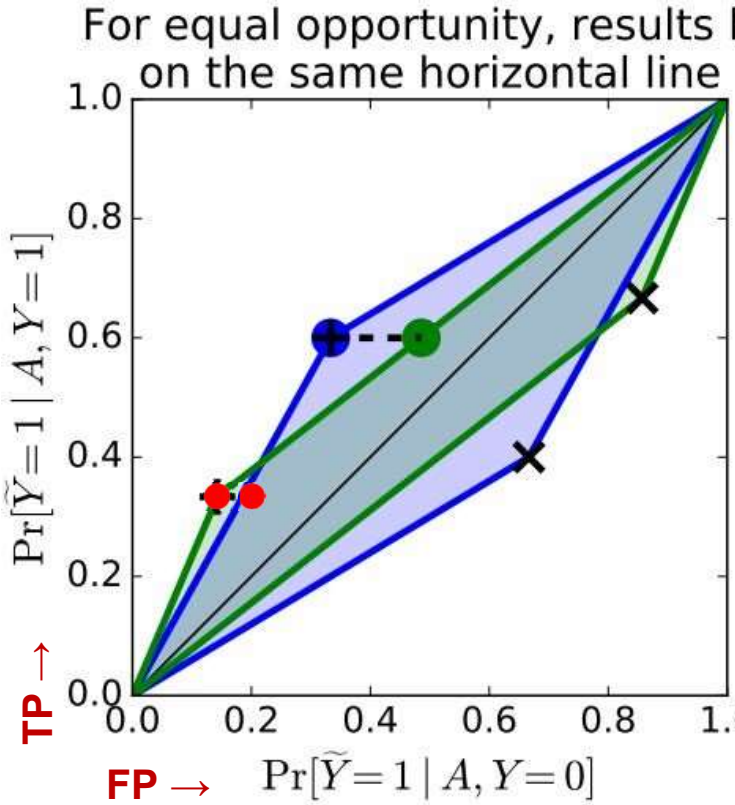


If this was EOD, then the low-scored hiring rate should also equals!

Question:

Is this (indicated in red) also a set of answers?

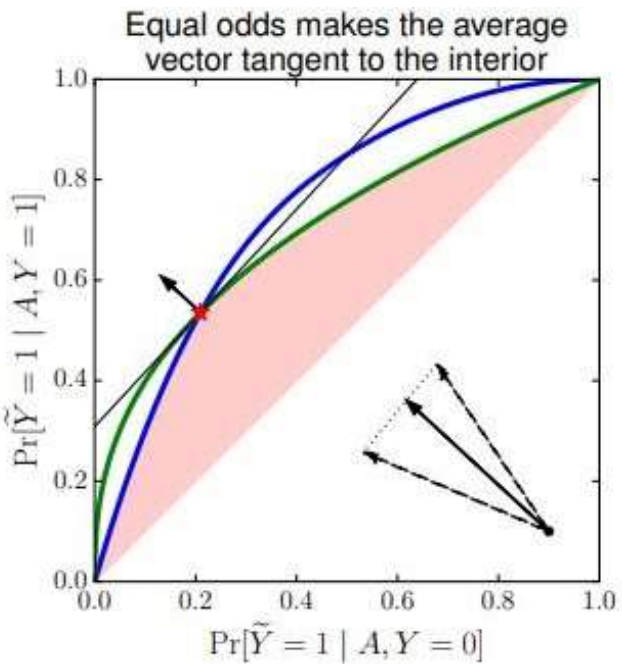
- Can still be considered as EOP
- However, not really an optimal EOP because of the Lower TP rate

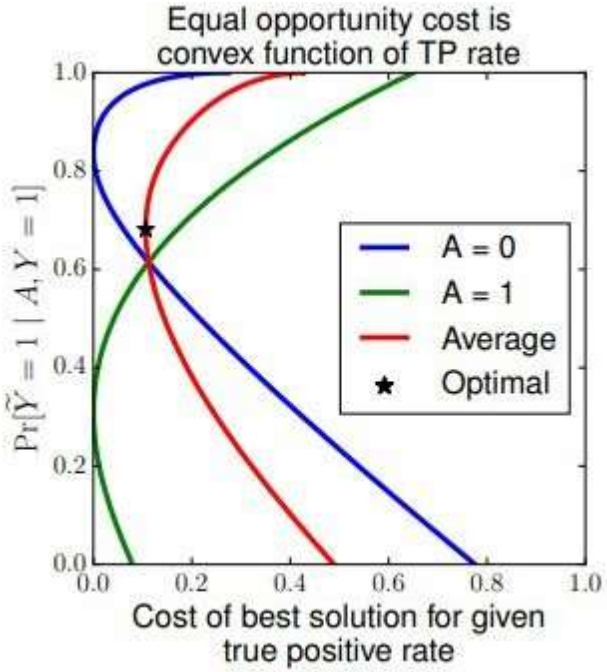


How to Derive (Score Function)

A-conditional ROC Curve: (Continuous & smooth)

$$C_a(t) \stackrel{\text{def}}{=} \left(\Pr\left\{\widehat{R} > t \mid A = a, Y = 0\right\}, \Pr\left\{\widehat{R} > t \mid A = a, Y = 1\right\}\right)$$ *, where t is the threshold value*

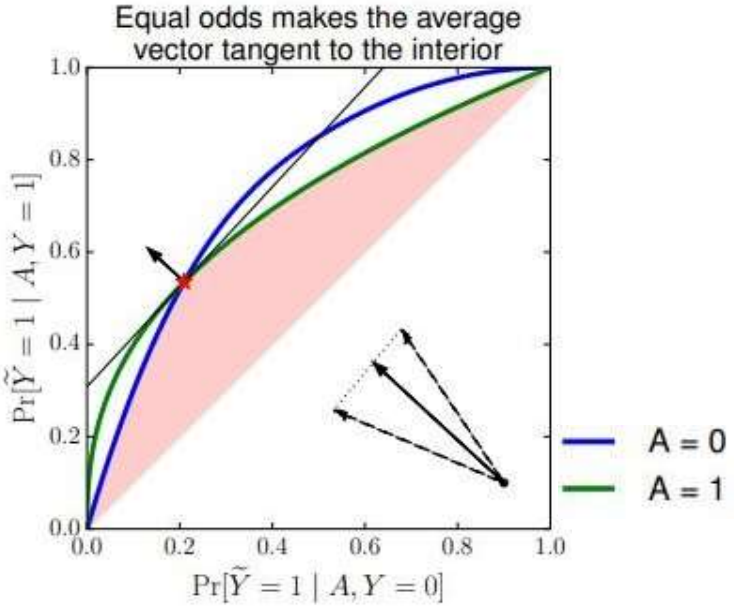




How to Derive (Score Function)

A-conditional ROC Curve: (Continuous & smooth)

$$C_a(t) \stackrel{\text{def}}{=} \left(\Pr\left\{\widehat{R} > t \mid A = a, Y = 0\right\}, \Pr\left\{\widehat{R} > t \mid A = a, Y = 1\right\}\right)$$



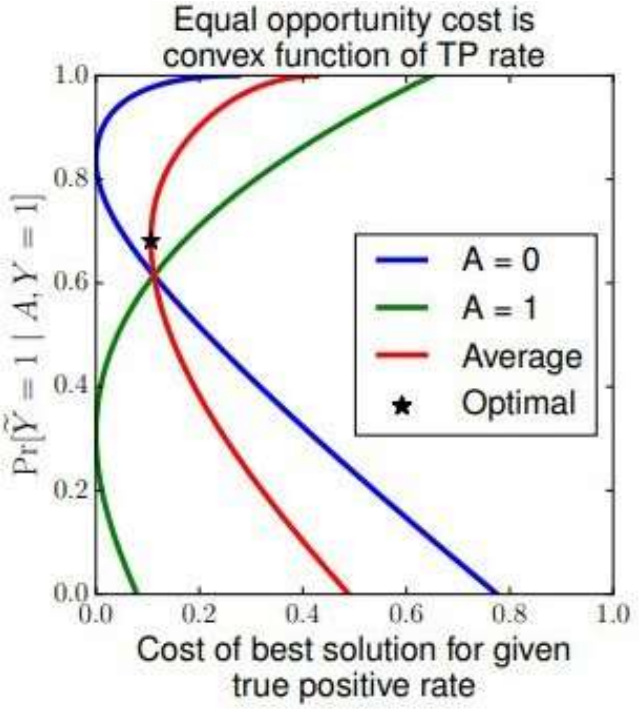
Equalized Odds:

- if the ROC curves for all values of A agree → Intersection of ROC curves
- May choose different t for different a
- Feasible set of F/TP rates of possible EOD predictors: intersected areas
- Pointwise min of all A-conditional ROC curves
  incentivize good utility in *all* classes

How to Derive (Score Function)

Equalized Opportunity:

- Points on the curves with only same TP in both group
- No randomization
- Optimal solution: 2 deterministic thresholds one for each group



Solving the optimization problems:

Both EOD and EOP can be efficiently optimized numerically using ternary search

Bayes Optimal Predictors

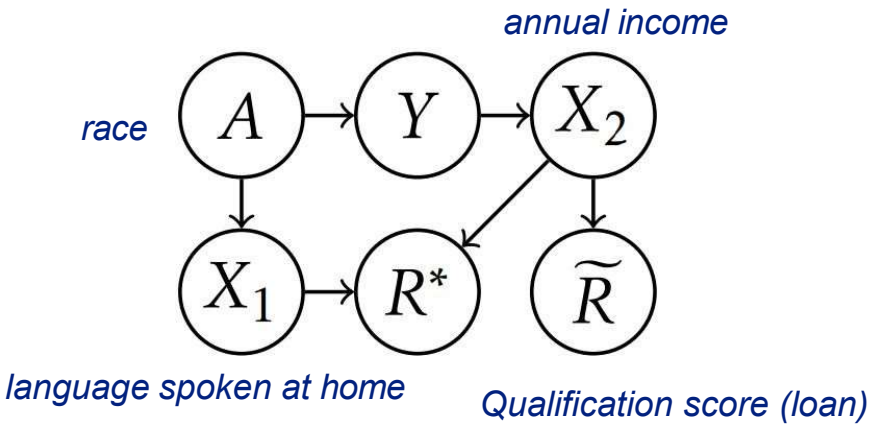
- **Goal**: construct a nearly optimal non-discriminating classifier
  - A Bayes optimal regressor -(derived threshold)> Bayes optimal EOD predictor
  - Quantify the loss of:

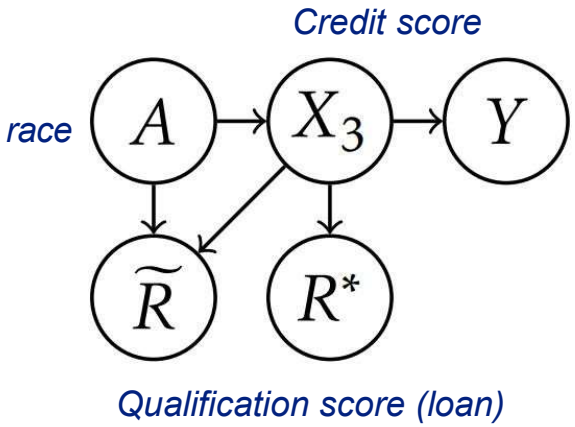
an EOD predictor derived based on a regressor (nearly Bayes-optimal) In terms of the conditional Kolmogorov distance

$\mathbf{Y}$ is nearly optimal if $\mathbf{R}$ is nearly optimal

Same for EOP

Indistinguishable Scenarios





R: Optimal score (accuracy)*
$\tilde{R}$: EOD score (fairness)

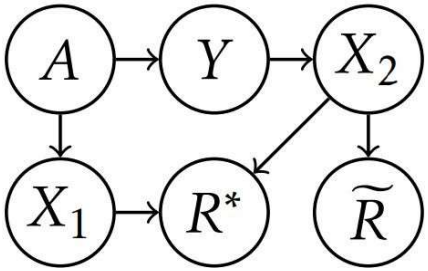
$$\mathbf{R}^* = X1 + X2 = A + X2 \qquad \mathbf{R}^* = X3$$

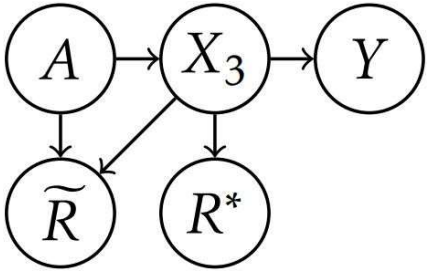
$$\tilde{\mathbf{R}} = X2 \qquad \tilde{\mathbf{R}} = X3 - A$$

$$\Rightarrow \mathbf{R}^* = \tilde{\mathbf{R}} + A \qquad \Rightarrow \mathbf{R}^* = \tilde{\mathbf{R}} + A$$

Indistinguishable!

Indistinguishable Scenarios





R: Optimal score (accuracy)*
$\tilde{R}$: EOD score (fairness)

$\mathbf{R^*} = X1 + X2 = A + X2$

$\tilde{\mathbf{R}} = X2$

$\Rightarrow \mathbf{R^*} = \tilde{\mathbf{R}} + A$

$\mathbf{R^*} = X3$

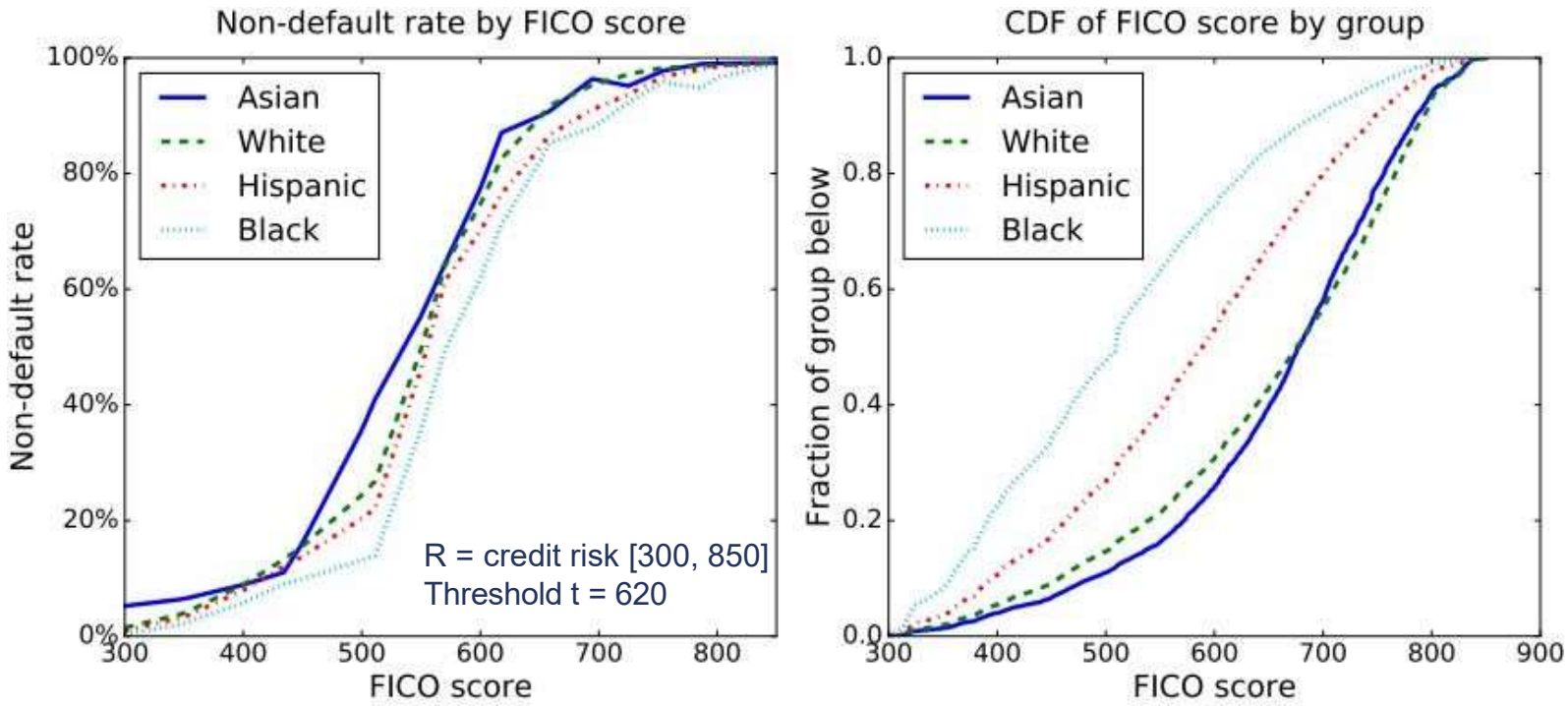
$\tilde{\mathbf{R}} = X3 - A$

$\Rightarrow \mathbf{R^*} = \tilde{\mathbf{R}} + A$

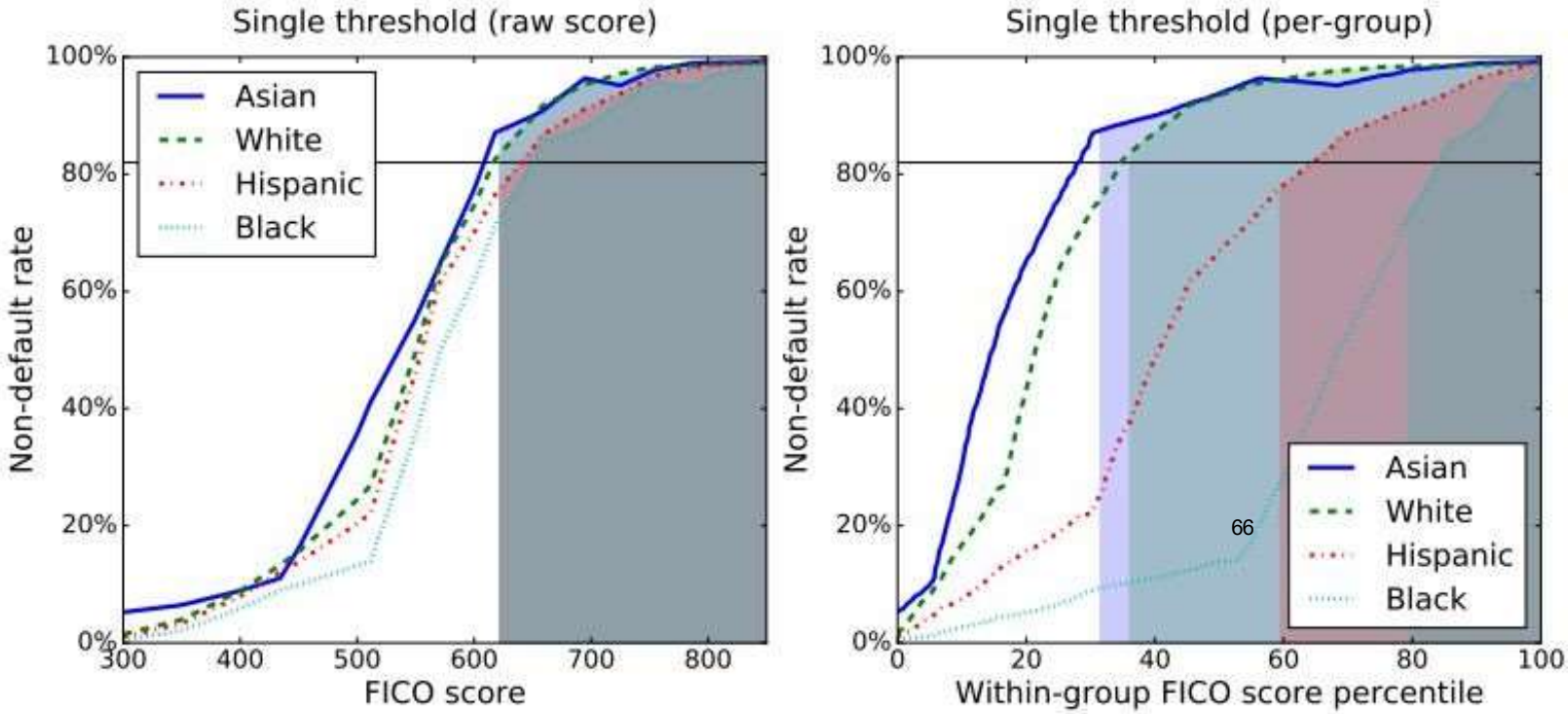
Shift the burden of uncertainty from the protected class to the decision maker

Case Study: FICO Scores

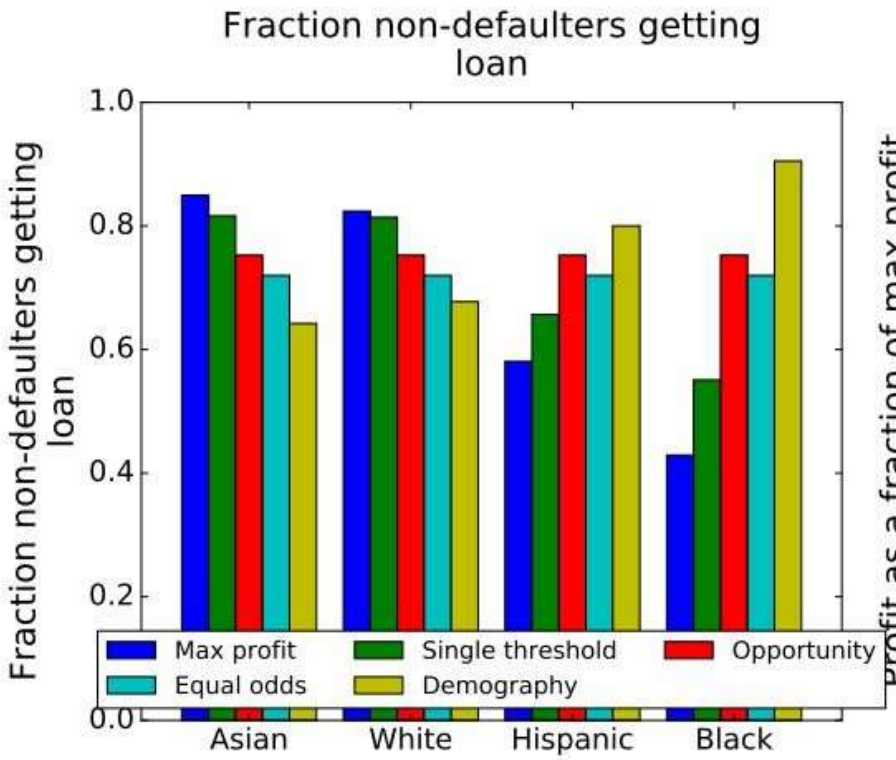
Scenario: a lender wants to provide loans for people who are able to pay back

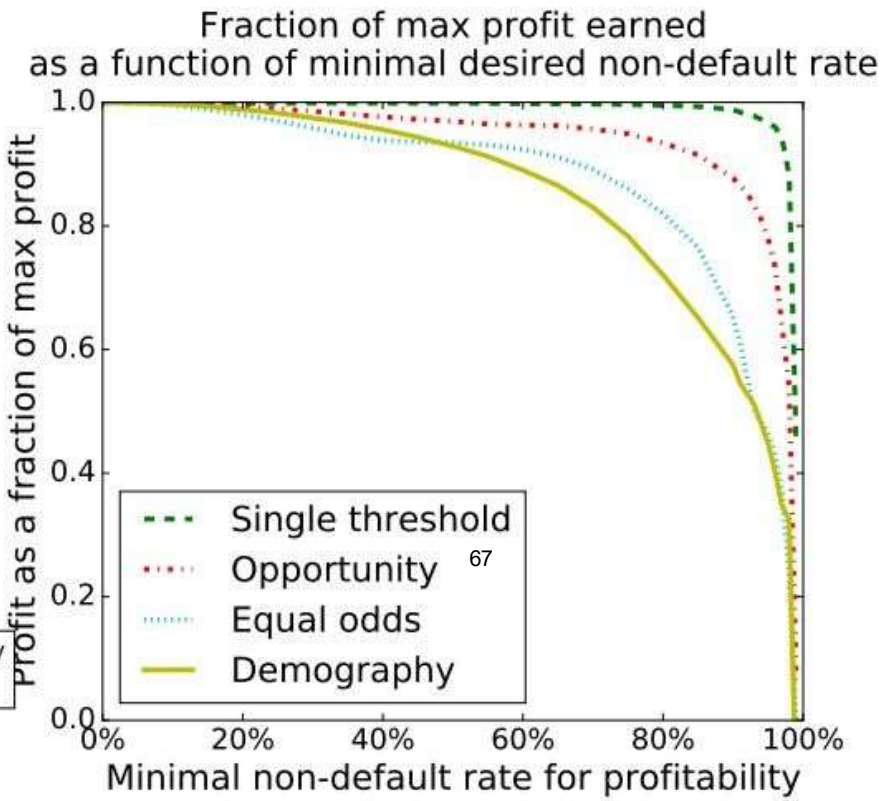


Case Study: FICO Scores



Case Study: FICO Scores





Limitations and Future Work

- Assumption on the data
  - Reliable “labeled data”
  - A and Y reasonably well balanced
- Evaluation only on binary classifiers, what about more complex models?
- Might not be a “good predictor” anymore after the post-processing
- Further support needed to verify the proposed notion of fairness: Interpretability?

  → to what extent biases may be learned by the model?

  → is the model interpretable enough to identify bias?
- Is it easy to find the unbiased true Y?

Revisiting the question: ***What does it mean to be “fair”?***