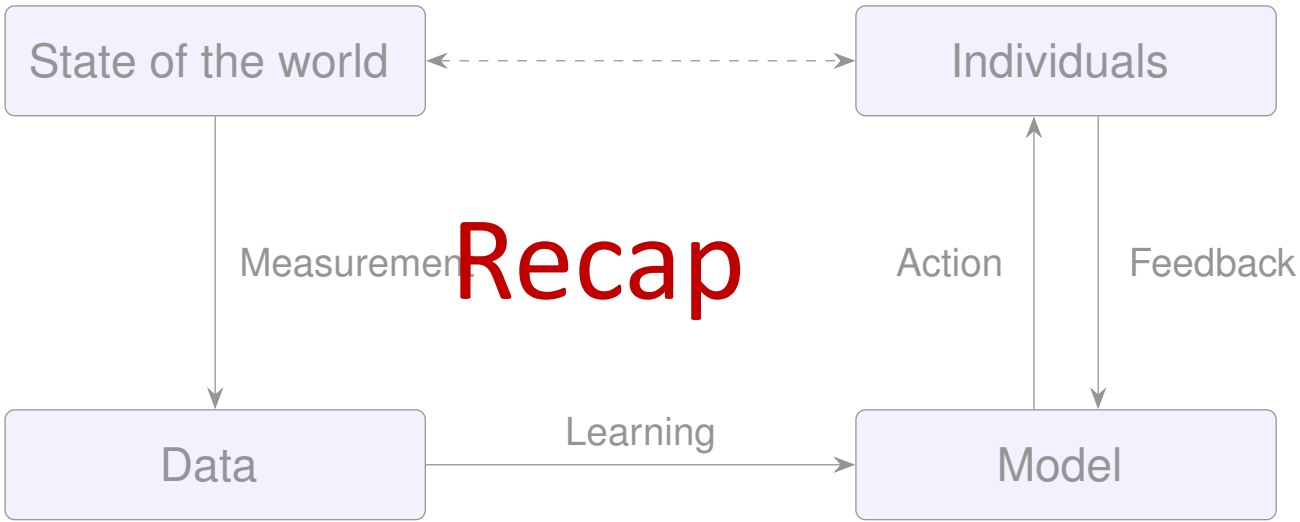
CSCI 618: Trustworthy ML (from an optimization lens)

Vatsal Sharan

Fall 2026

Lecture 10, Sep 28

USC University of Southern California


State of the world
Individuals
Measurement
Recap
Action
Feedback
Data
Learning
Model

Group fairness

Group fairness notions require that the models predictions obey certain properties over protected groups (e.g. by race, gender).

Many different notions have been proposed

- Statistical parity
- Equalized odds
- Calibration across groups

Statistical parity/Demographic parity

Binary classification setup (e.g. admitting a student to a degree program)

- Classifier f
- Datapoint (x, y)
- Sensitive attribute $a \in \{0,1\}$

Statistical parity (also known as demographic parity): $\Pr_x[f(x) = 1 \mid a = 1] = \Pr_x[f(x) = 1 \mid a = 0]$

In words: **Predictions are independent of sensitive attribute**

E.g., admit equal fraction of men or women into program

Can be too strong if labels and sensitive attribute are not independent.

E.g. if one demographic is more likely to be qualified than the other

Equalized odds & Equality of opportunity

Same binary classification setup (e.g. admitting student to degree program)

- Classifier f
- Datapoint (x, y)
- Sensitive attribute $a \in \{0,1\}$

Just having 1st constraint (for $y = 1$) gives **equality of opportunity**

Equalized odds: Following 2 constraints are satisfied

1. $\Pr_{x,y}[f(x) = 1 \mid a = 1, y = 1] = \Pr_{x,y}[f(x) = 1 \mid a = 0, y = 1]$
2. $\Pr_{x,y}[f(x) = 0 \mid a = 1, y = 0] = \Pr_{x,y}[f(x) = 0 \mid a = 0, y = 0]$

Recall for class 1 = $\Pr_{x,y}[f(x) = 1 \mid y = 1]$

FPR = $\Pr_{x,y}[f(x) = 1 \mid y = 0]$

Equivalently:
Recall for $y = 1$ is the same for both groups (1st condition)
False positive rate (**FPR**) is the same for both groups (2nd condition)

Also equivalent to saying: Conditioned on label, prediction is independent of sensitive attribute

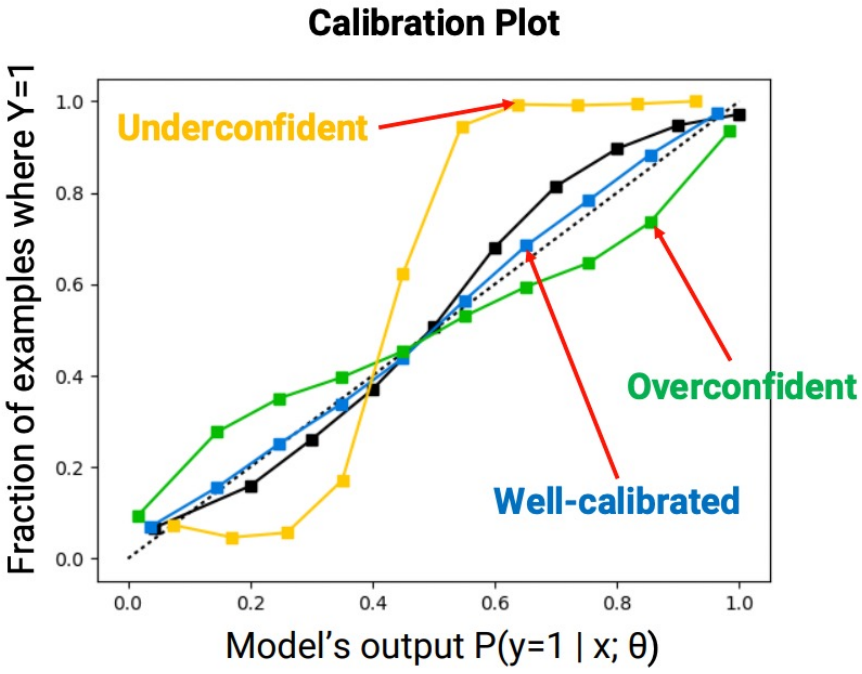
From *Equality of Opportunity in Supervised Learning,* Hardt et al. '16

Calibration

Calibration: A model f for binary classification is calibrated if

$$\Pr_{x,y}[y = 1 \mid f(x) = \alpha] = \alpha$$

Informally, this says that “predictions mean what they should”



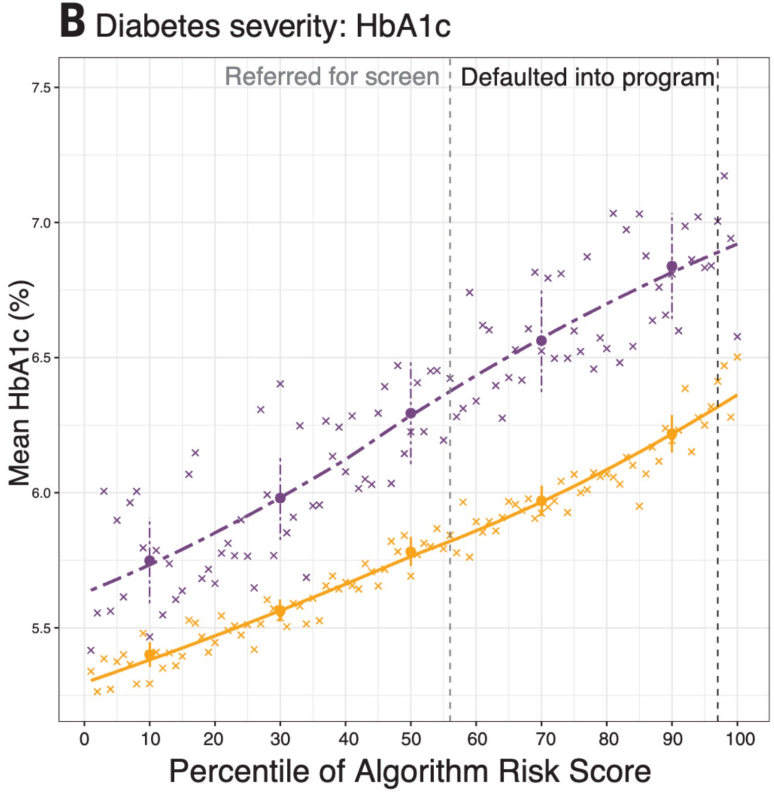
This is known as a ***reliability diagram***

Calibration across groups

A model f for binary classification is calibrated for groups defined by sensitive attribute a if

$$\Pr_{x,y}[y = 1 \mid f(x) = \alpha, a = 1] = \alpha\,,$$
$$\Pr_{x,y}[y = 1 \mid f(x) = \alpha, a = 0] = \alpha.$$

Informally, this says that “predictions mean what they should for each group”



Achieving these notions: Post-processing for statistical parity

Consider binary classification setup (e.g. admitting a student to a degree program)

- Predictor p which predicts a score in $[0,1]$ (higher score => higher probability of label 1)
- Datapoint (x, y)
- Sensitive attribute $a \in \{0,1\}$

We want use p to get classifier f which maximizes accuracy but obeys statistical parity:

$$\Pr_x[f(x) = 1 \mid a = 1] = \Pr_x[f(x) = 1 \mid a = 0]$$

- What would f be if we only wanted to maximize accuracy? (Suppose $p(x) = \Pr[y = 1|x]$)

- How to ensure statistical parity?

Achieving these notions: Post-processing for equality of opportunity

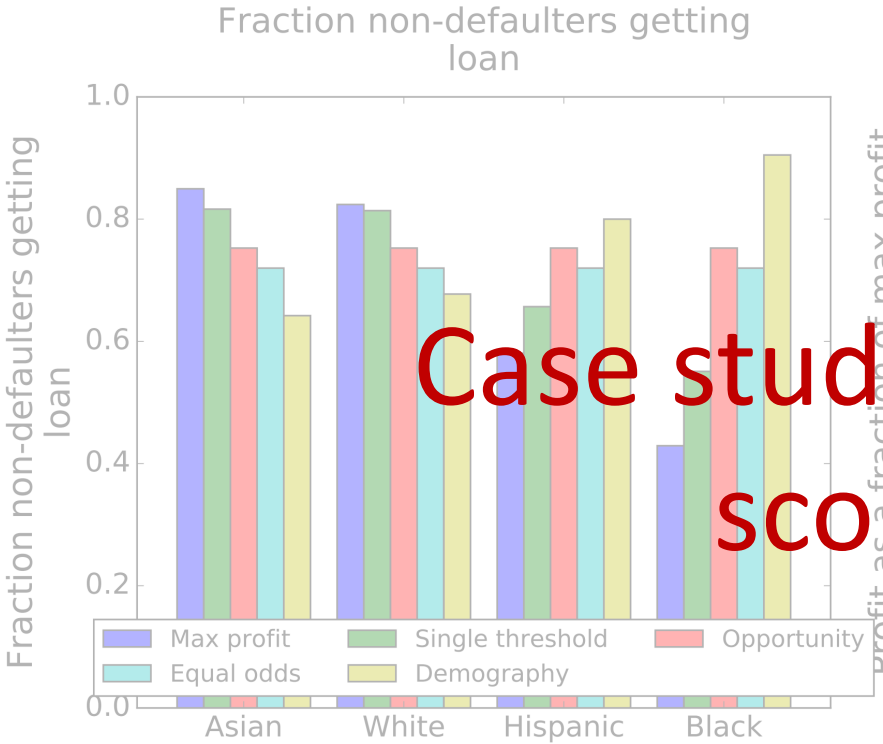
Consider binary classification setup (e.g. admitting a student to a degree program)

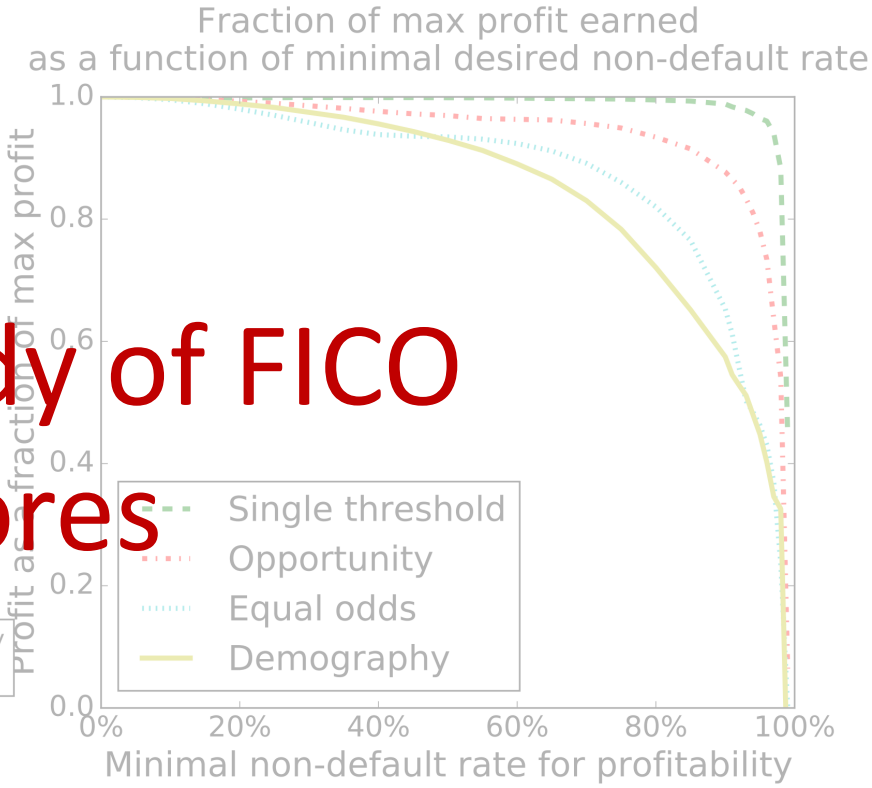
- Predictor p which predicts a score in $[0,1]$ (higher score => higher probability of label 1)
- Datapoint (x, y)
- Sensitive attribute $a \in \{0,1\}$

We want use p to get classifier f which maximizes accuracy but obeys equality of opportunity:

$$\Pr_{x,y}[f(x) = 1 \mid a = 1, y = 1] = \Pr_{x,y}[f(x) = 1 \mid a = 0, y = 1]$$

- How to obtain equality of opportunity?

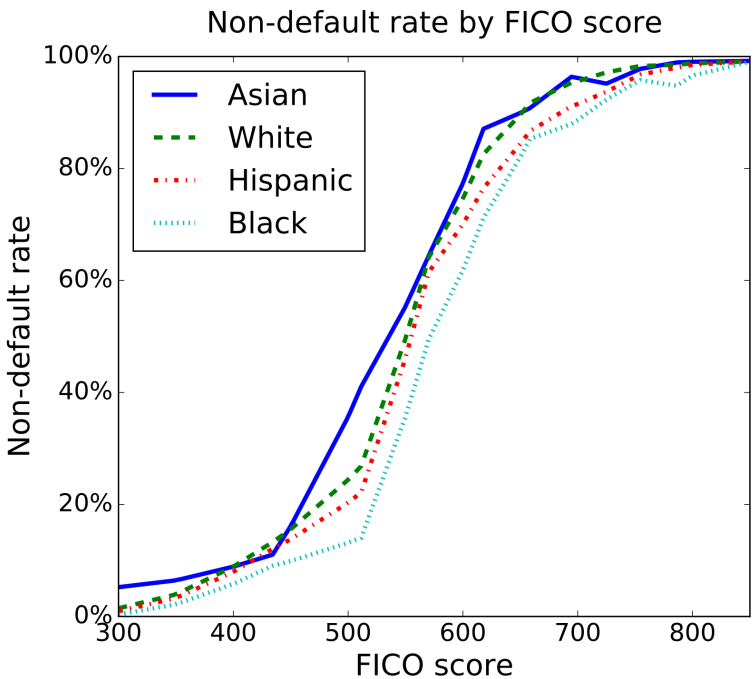




Case study of FICO scores

The input data: FICO scores and their distribution by race

- FICO scores based on 300k TransUnion scores from 2003
- Range from 300-850, aim to predict risk of defaulting on a loan
- 620 is common threshold for good loan rates. Corresponds to 82% non-default rate in the data



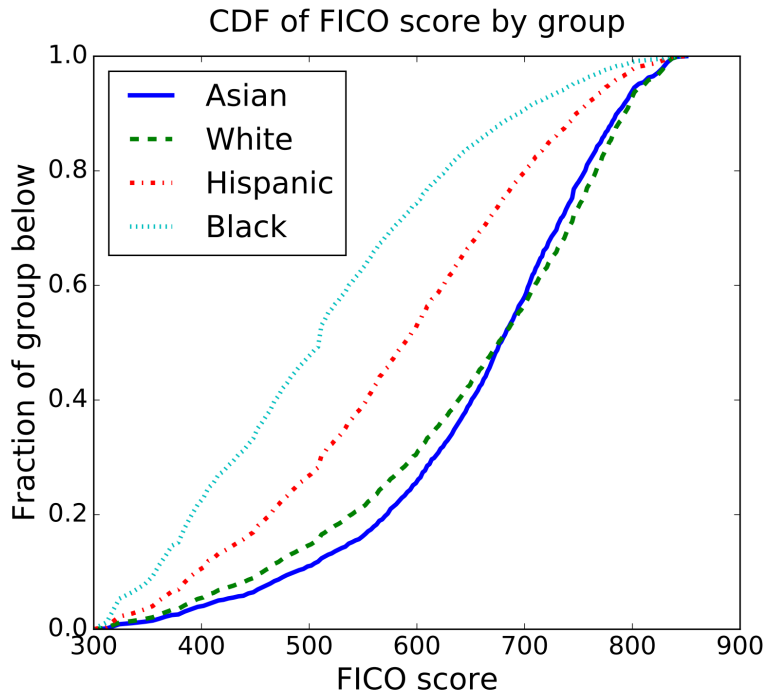
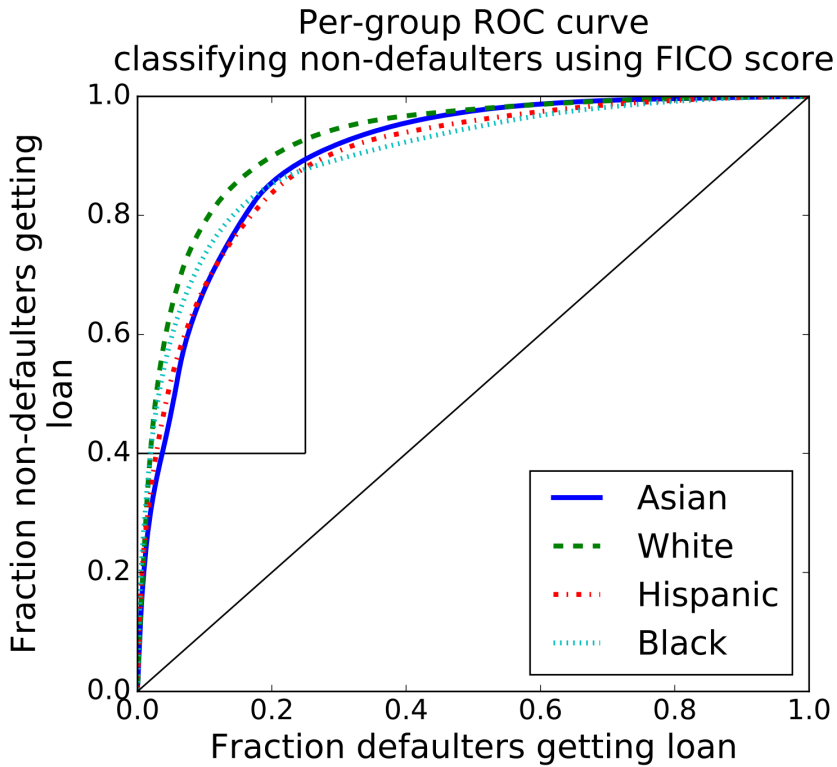


Fig. from *Equality of Opportunity in Supervised Learning,* Hardt et al. '16

The decision-making rules

- **Max profit:** No fairness constraints, will pick for each group the threshold that maximizes profit. This is the score at which 82% of people in that group do not default.
- **Race blind:** Requires threshold to be same for each group. Hence will pick the single threshold at which 82% of people do not default overall.
- **Demographic parity:** Same as statistical parity, ensures fraction of group members that get loan is same across groups.
- **Equal opportunity:** Picks for each group a threshold such that fraction of non-defaulting group members that qualify for loans is the same.
- **Equalized odds:** Requires both fraction of non-defaulters that qualify for loans and fraction of defaulters that qualify for loans to be constant across groups.

How accurate is the base classifier across groups?



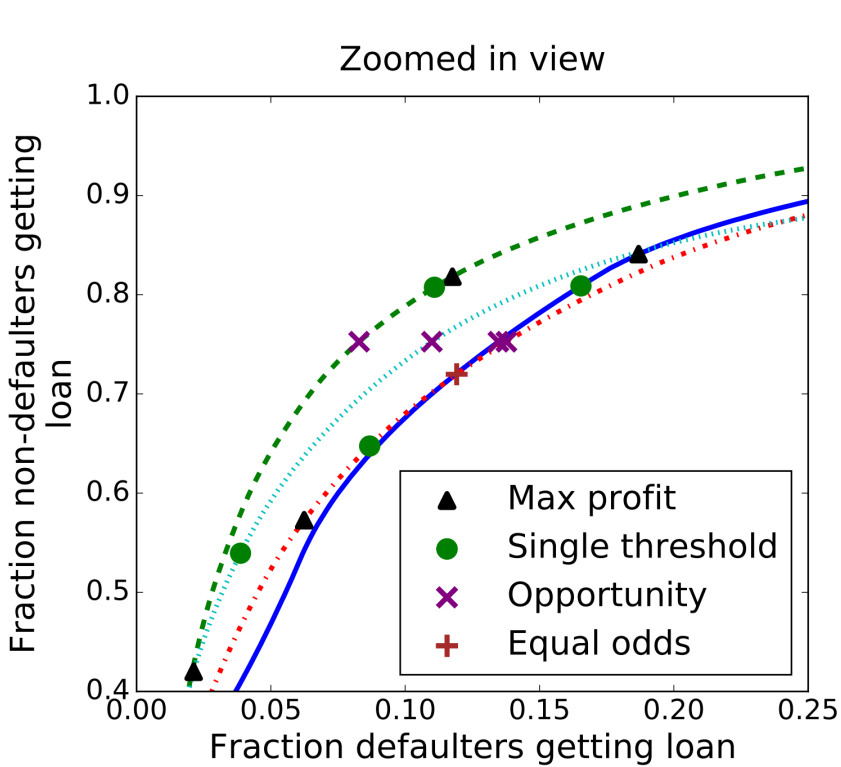
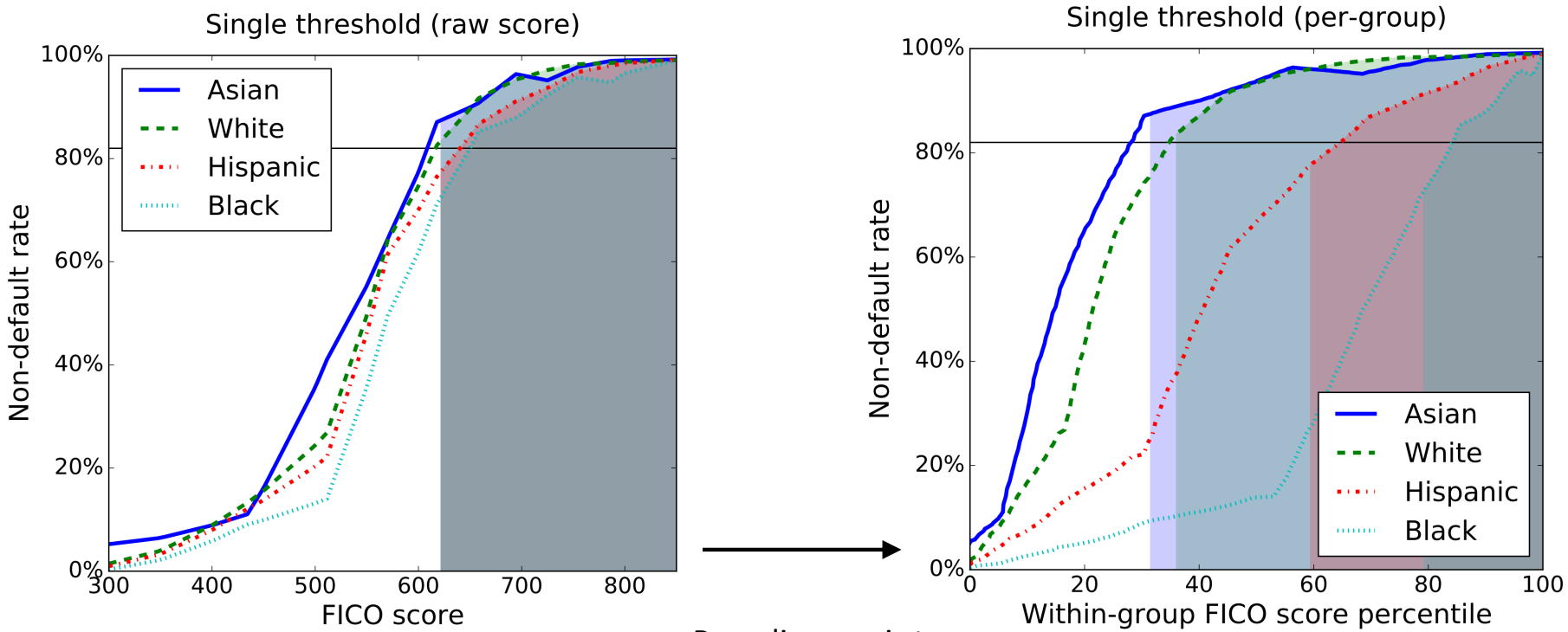


Fig. from *Equality of Opportunity in Supervised Learning,* Hardt et al. '16

What happens when we pick a single threshold?

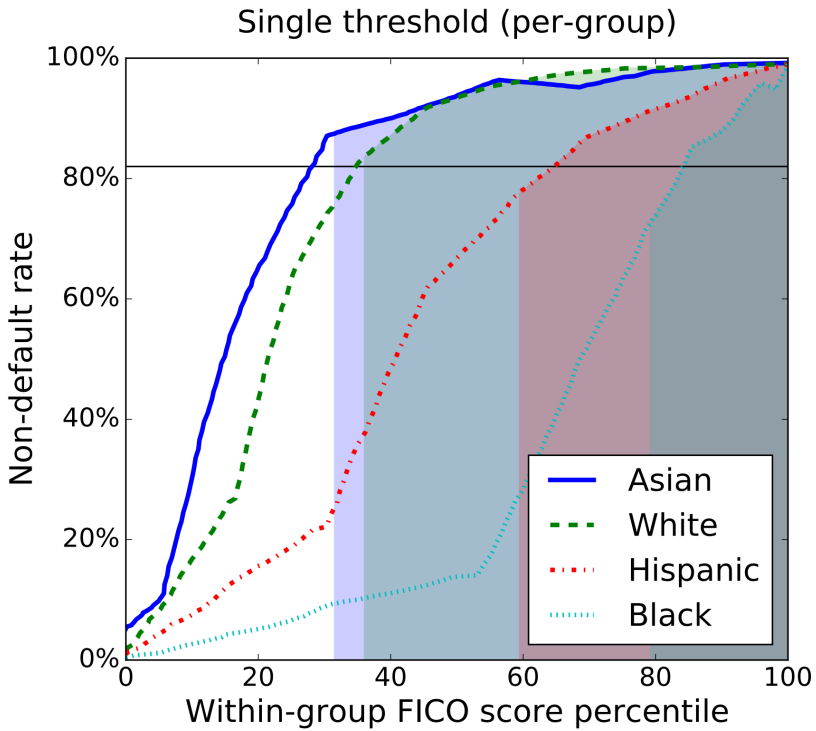


Rescaling x-axis to represent within group thresholds

Claim: $\Pr[f(x) = 1 \mid y = 1, a]$ is fraction of area under the curve that is shaded

Fig. from *Equality of Opportunity in Supervised Learning,* Hardt et al. '16

Equality of opportunity as area under curve



Claim: $\Pr[f(x) = 1 \mid y = 1, a]$ is fraction of area under the curve that is shaded

Reason:

Consider any within-group percentile t.

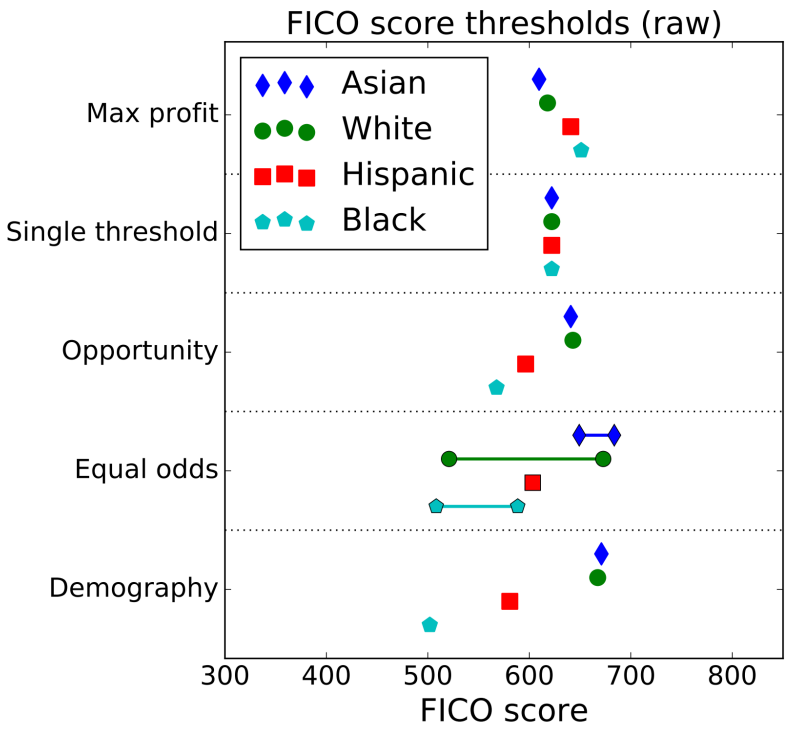
$\Pr[y = 1 \mid a] = \int_t \Pr[y = 1 \mid t, a\,]\ dt$ = area under curve

$\Pr[f(x) = 1, y = 1 \mid a] = \int_t 1\,(\,t > \text{threshold})\ \Pr[y = 1 \mid t, a\,]\ dt$

= area under curve which is shaded

Fig. from *Equality of Opportunity in Supervised Learning,* Hardt et al. '16

What are thresholds for different rules?



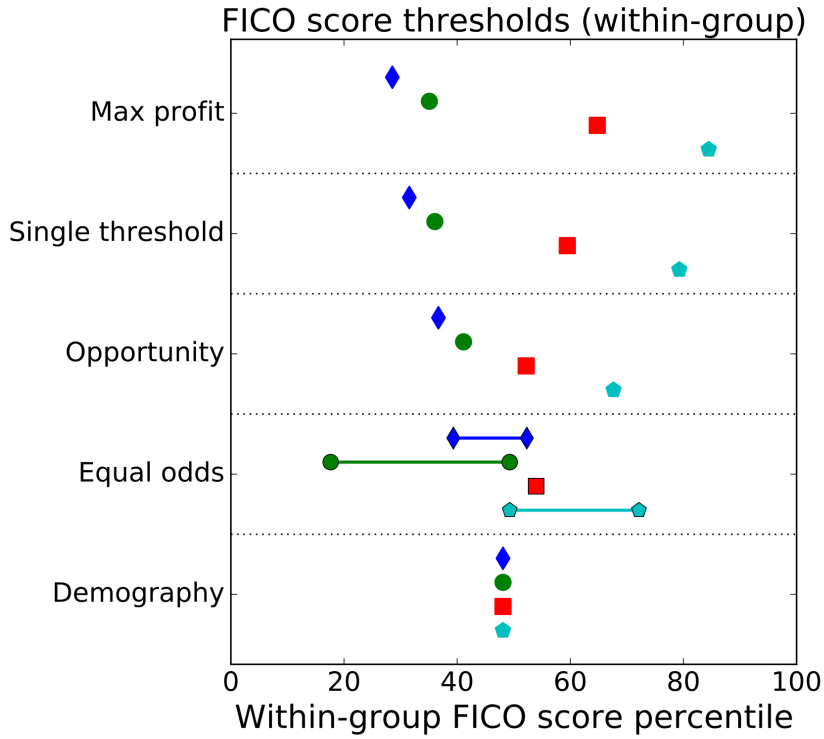
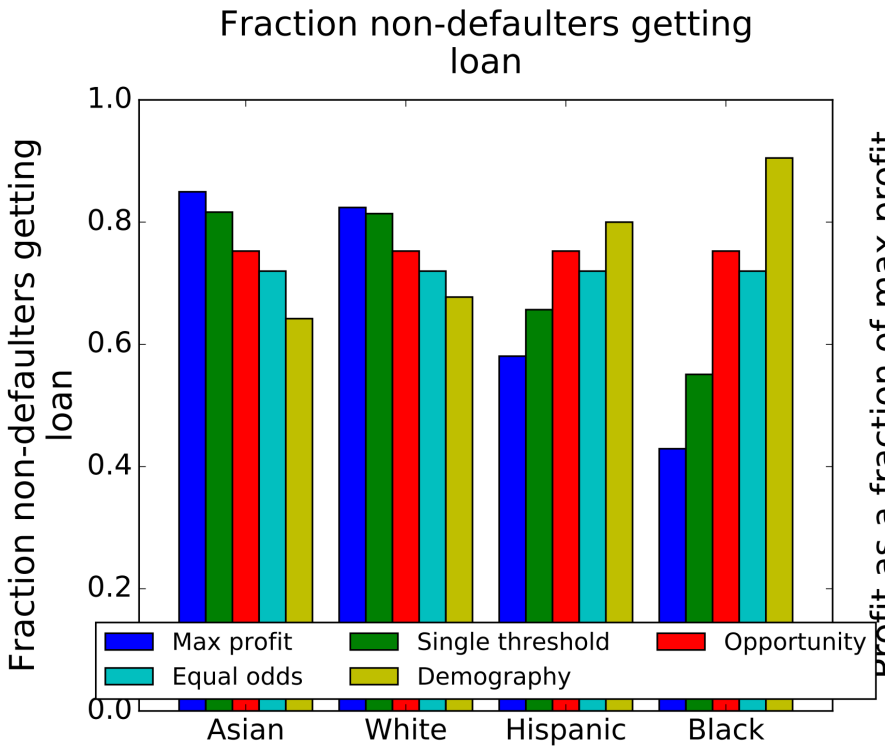


Fig. from *Equality of Opportunity in Supervised Learning,* Hardt et al. '16

What fairness and utility do different rules obtain?



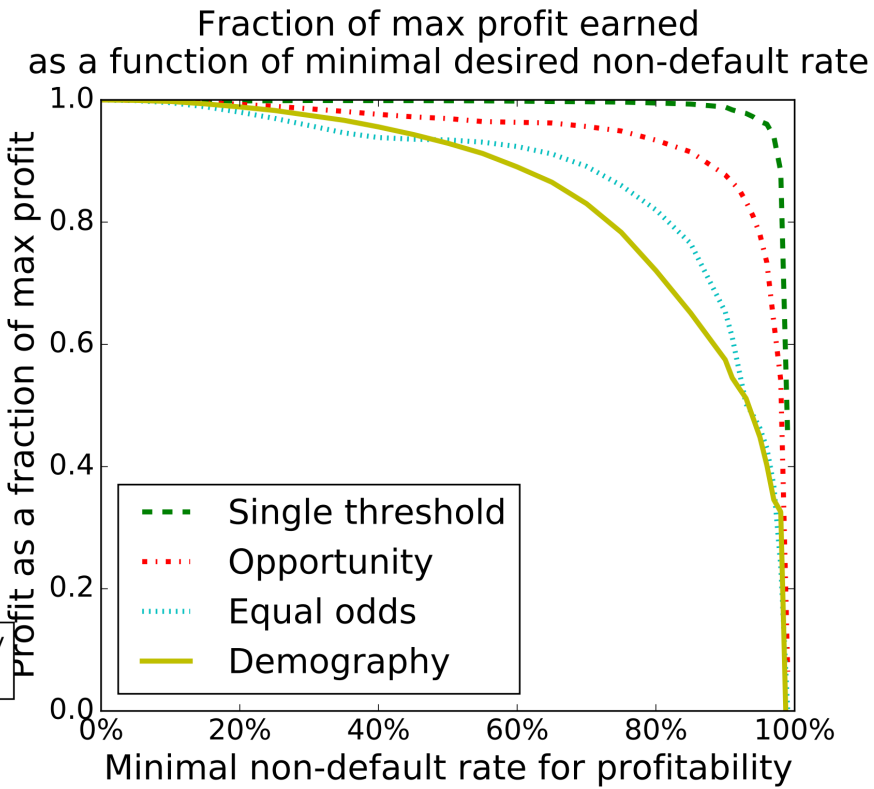
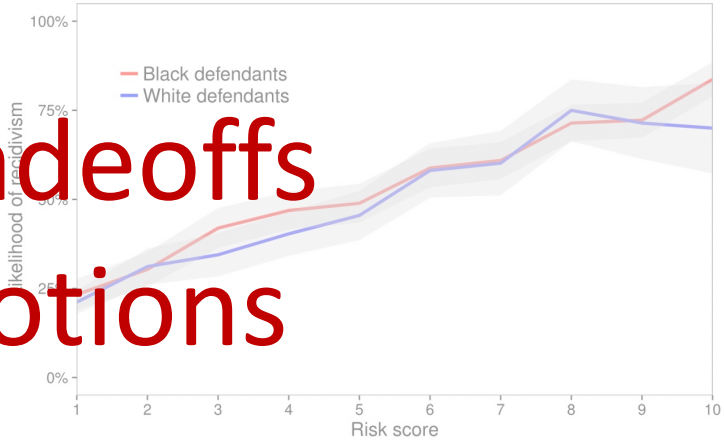


Fig. from *Equality of Opportunity in Supervised Learning,* Hardt et al. '16

Takeaways?



Inherent tradeoffs between notions



Consider three basic fairness notions

- **Calibration for groups:**

$$\Pr_{x,y}[y = 1 \mid f(x) = \alpha, x \in g_0] = \alpha\,,$$
$$\Pr_{x,y}[y = 1 \mid f(x) = \alpha, x \in g_1] = \alpha.$$

- **Balance for positive class:** The average score received by individuals with $y = 1$ is the same for the two groups.

$$\frac{\sum_{(x,y):\, x \in g_0,\, y=1} f(x)}{|(x,y): x \in g_0,\ y = 1|} = \frac{\sum_{(x,y):\, x \in g_1, y=1} f(x)}{|(x,y): x \in g_1,\ y = 1|}$$

- **Balance for negative class:** The average score received by individuals with $y = 0$ is the same for the two groups.

$$\frac{\sum_{(x,y):\, x \in g_0, y=0} f(x)}{|(x,y): x \in g_0,\ y = 0|} = \frac{\sum_{(x,y):\, x \in g_1, y=0} f(x)}{|(x,y): x \in g_1,\ y = 0|}$$

Except in special cases, impossible to achieve all 3!!

Let the base rate for group $g_i = \Pr[y = 1 \mid \ x \in g_i\ ]$.

Theorem (Kleinberg-Mullainathan-Raghavan '16): Consider any predictor f which satisfies (1) calibration across groups, (2) balance for positive class, and (3) balance for negative class. Then,

- either $f(x)$ is a perfect predictor, with $f(x) \in \{0,1\}$,
- or the two groups have equal base rates.

Similar result is also true for classifiers which satisfy approximate versions of the 3 fairness conditions.

Inherent tradeoff in fairness, proof sketch

Theorem (Kleinberg-Mullainathan-Raghavan ‘16): Consider any predictor f which satisfies (1) calibration across groups, (2) balance for positive class, and (3) balance for negative class. Then,

- either $f(x)$ is a perfect predictor, with $f(x) \in \{0,1\}$,
- or the two groups have equal base rates.

Inherent tradeoff in fairness, proof sketch

Step 2 : relate to balance conditions

N_t : # people in group t

M_t : # people in group t in true class

$$\frac{M_t}{N_t} = \alpha_t$$

Let x_t be the avg score given to +ve members in group t

Let y_t " -ve "

$$(N_t - M_t)\, y_t + M_t\, x_t = M_t$$

$$\left(1 - \frac{M_t}{N_t}\right) y_t + \left(\frac{M_t}{N_t}\right) x_t = \left(\frac{M_t}{N_t}\right)$$

$$(1 - \alpha_t)\, y_t + \alpha_t\, x_t = \alpha_t$$

$$y_t / (\alpha_t / (1-\alpha_t)) + x_t = 1$$

By (2), $x_t = x \;\forall t$

(3), $y_t = y \;\forall t$

