CSCI 618: Trustworthy ML (from an optimization lens)

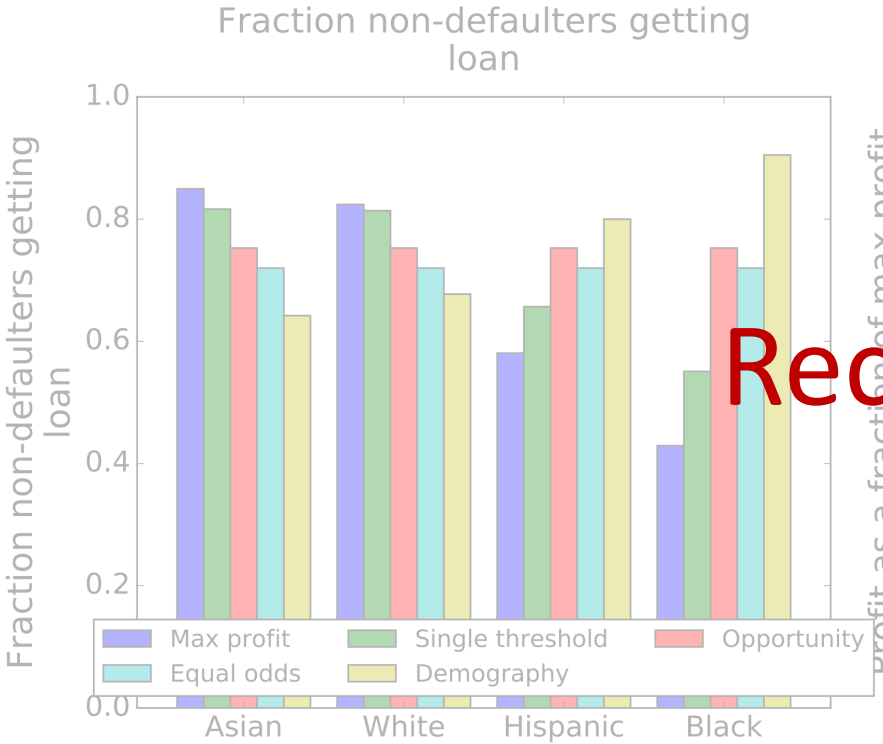
Vatsal Sharan

Fall 2026

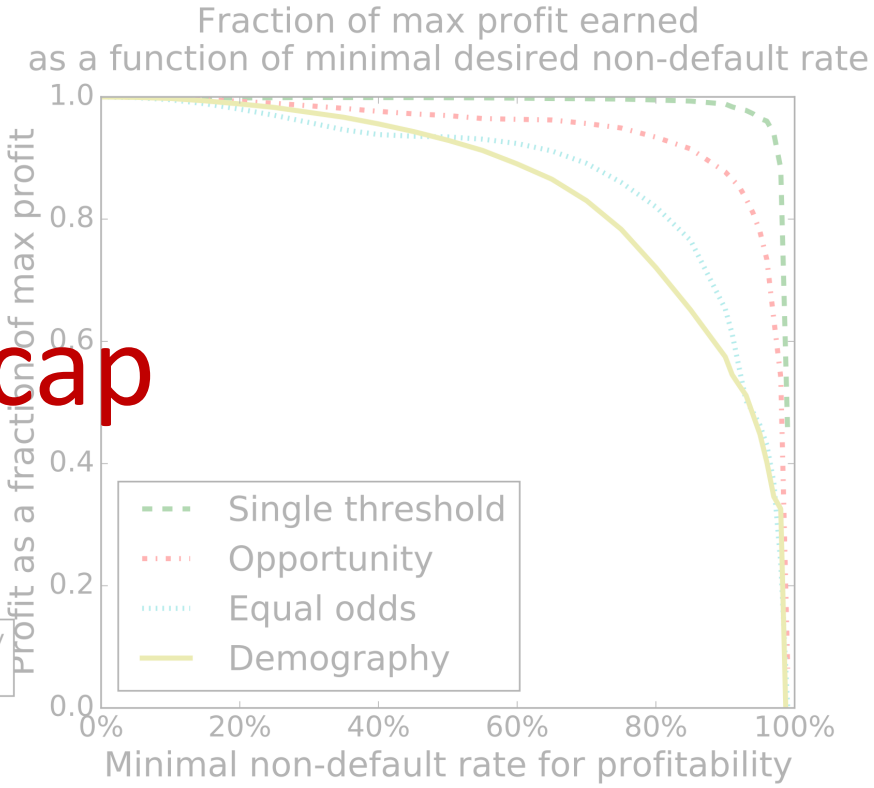
Lecture 11, Sep 30



USC University of Southern California

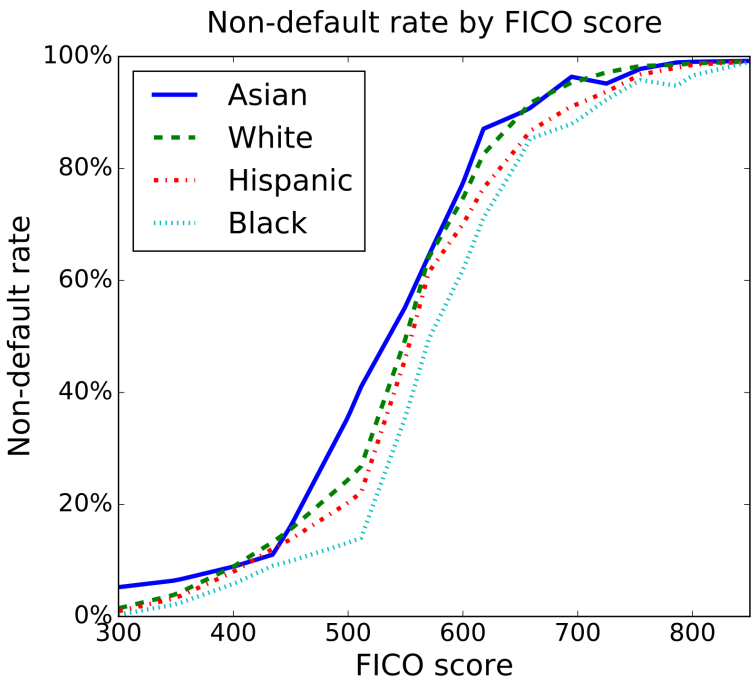

Fraction non-defaulters getting loan
Fraction non-defaulters getting loan
1.0
0.8
0.6
0.4
0.2
0.0
Asian
White
Hispanic
Black
Max profit
Single threshold
Opportunity
Equal odds
Demography

Recap


Fraction of max profit earned as a function of minimal desired non-default rate
Profit as a fraction of max profit
1.0
0.8
0.6
0.4
0.2
0.0
0%
20%
40%
60%
80%
100%
Minimal non-default rate for profitability
Single threshold
Opportunity
Equal odds
Demography

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



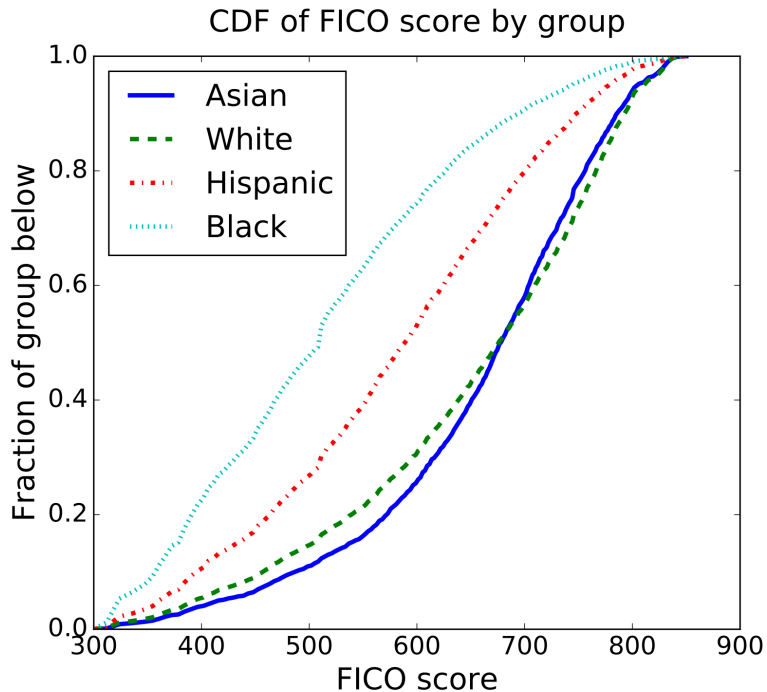
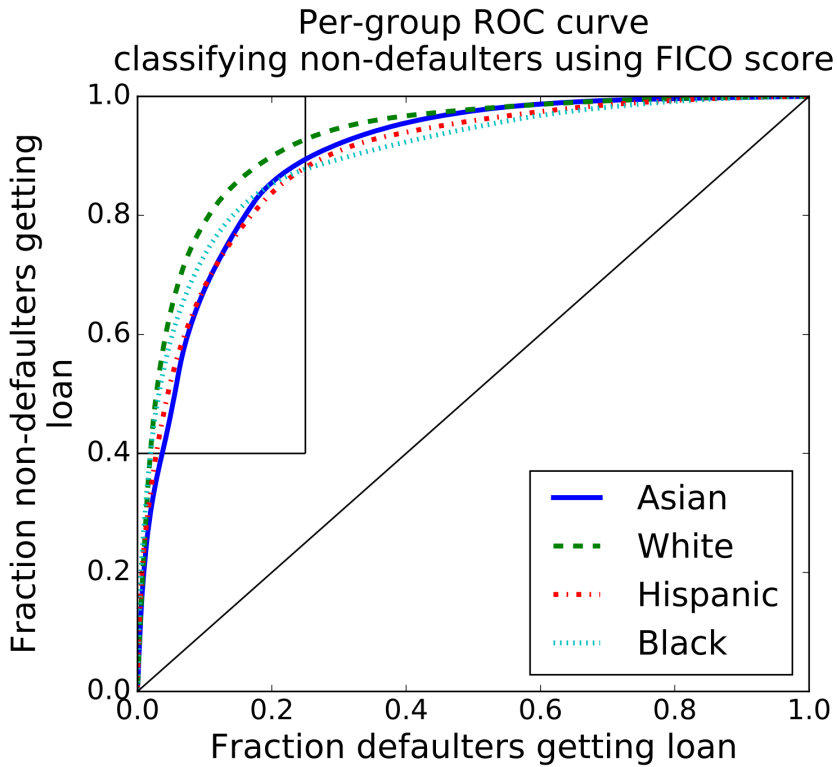


Fig. from *Equality of Opportunity in Supervised Learning,* Hardt et al. '16

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?



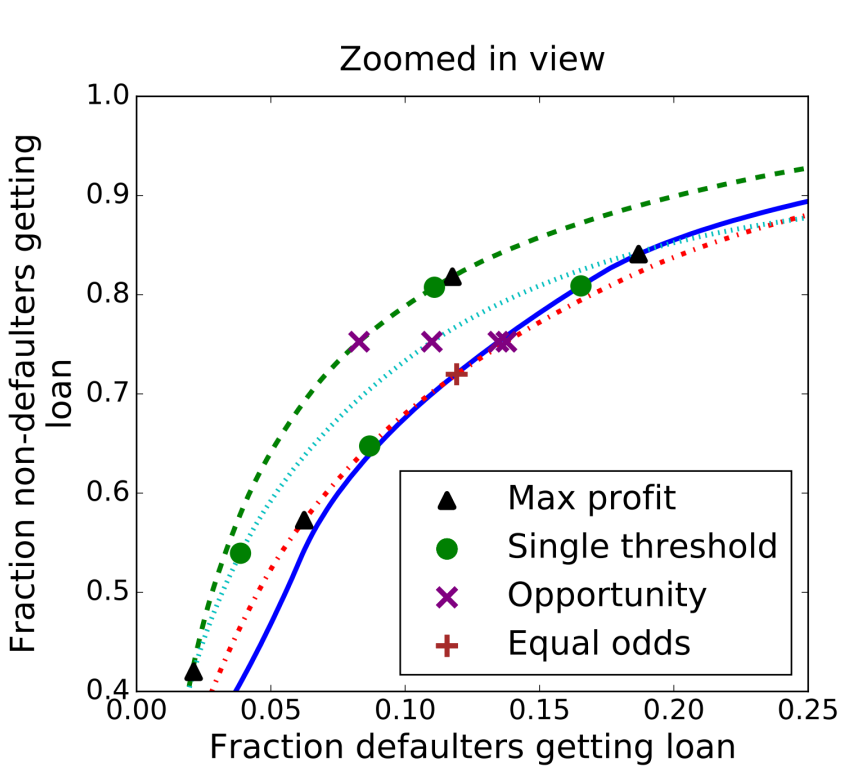
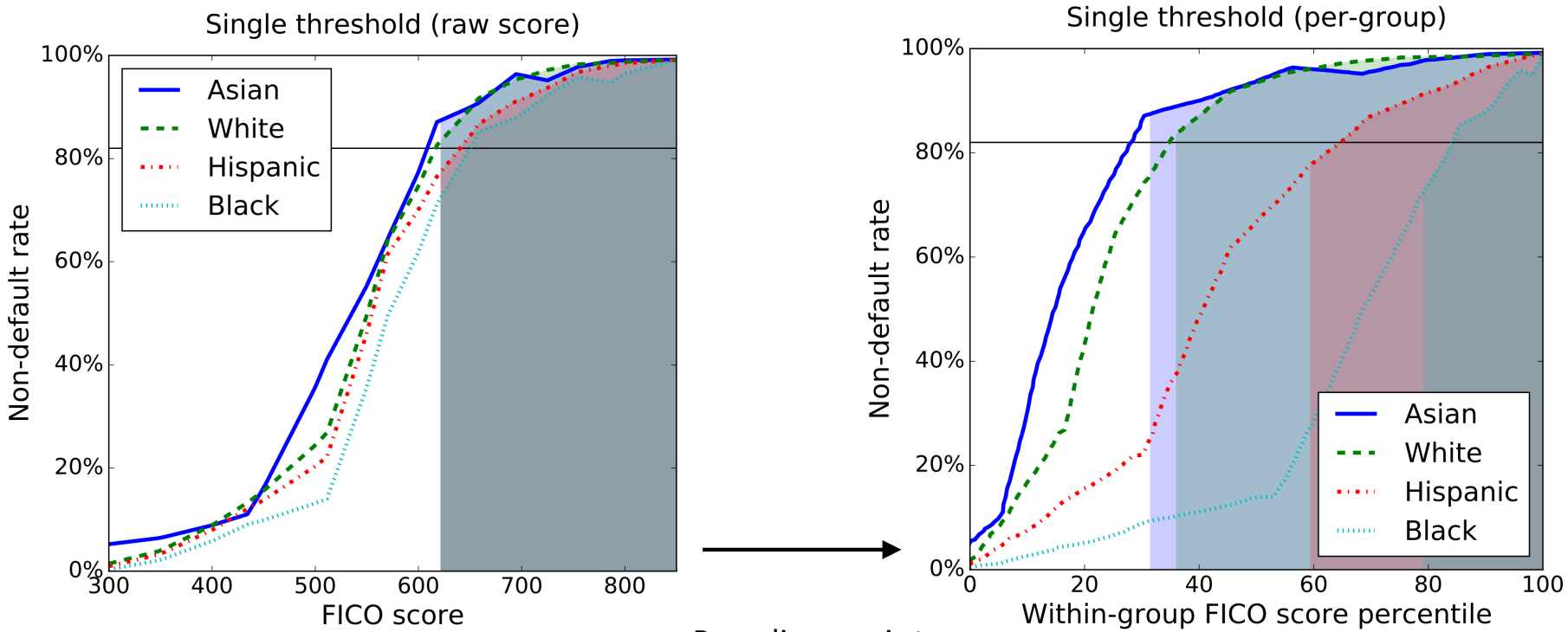


Fig. from *Equality of Opportunity in Supervised Learning,* Hardt et al. '16

What happens when we pick a single threshold?

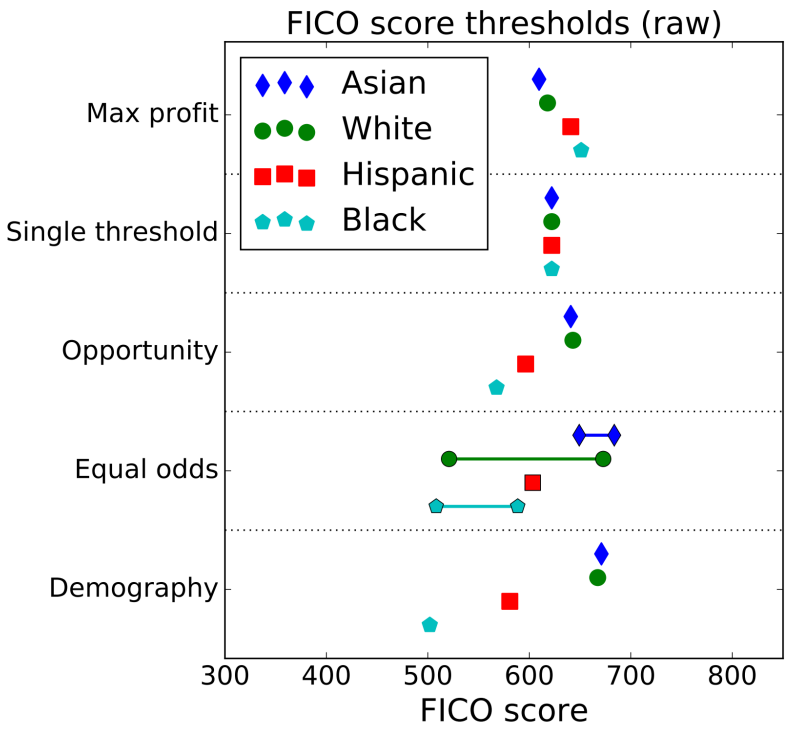


Rescaling x-axis to represent within group thresholds

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

What are thresholds for different rules?



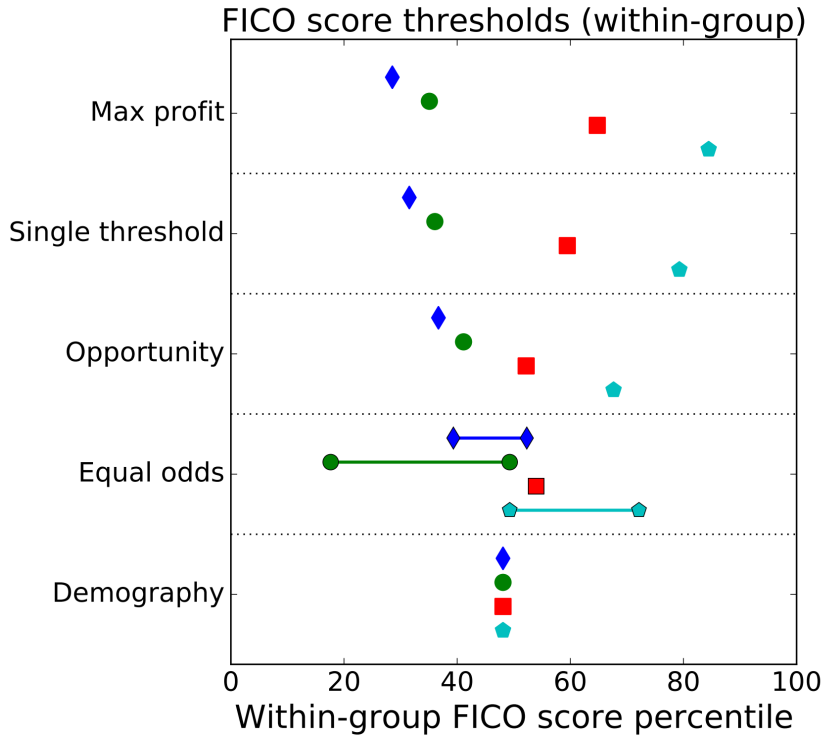
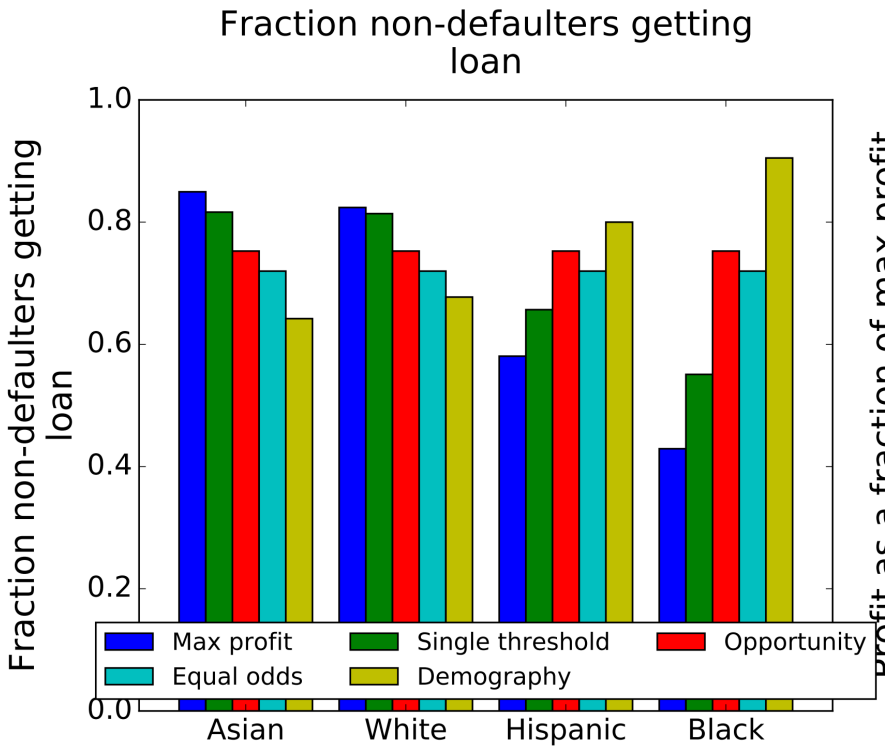


Fig. from *Equality of Opportunity in Supervised Learning,* Hardt et al. '16

What fairness and utility do different rules obtain?



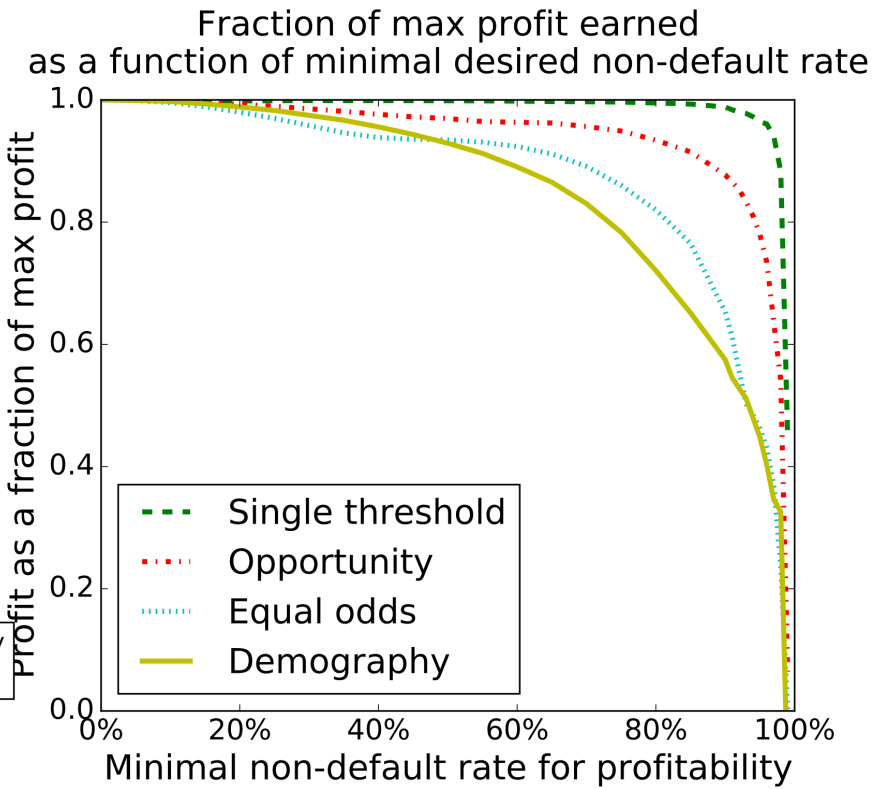


Fig. from *Equality of Opportunity in Supervised Learning,* Hardt et al. '16

Is it possible to get both equalized odds and demographic parity?

Equalized odds : $\Pr[\hat{Y}=1 \mid Y=1, a=1] = \Pr[\hat{Y}=1 \mid Y=1, a=0] = t$

$\Pr[\hat{Y}=1 \mid Y=0, a=1] = \Pr[\hat{Y}=1 \mid Y=0, a=0] = f$

Statistical parity : $\Pr[\hat{Y}=1 \mid a=1] = \Pr[\hat{Y}=1 \mid a=0]$

Let $\pi_1 = \Pr[Y=1 \mid a=1]$, $\pi_0 = \Pr[Y=1 \mid a=0]$

$$\begin{aligned}\Pr[\hat{Y}=1 \mid a=1] &= \Pr[\hat{Y}=1 \mid Y=1, a=1] \cdot \Pr[Y=1 \mid a=1] \\ &\quad + \Pr[\hat{Y}=1 \mid Y=0, a=1] \cdot \Pr[Y=0 \mid a=1] \\ &= t \cdot \pi_1 + f \cdot (1-\pi_1)\end{aligned}$$

For statistical parity : $t\pi_1 + f(1-\pi_1) = t\pi_0 + f(1-\pi_0)$

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,

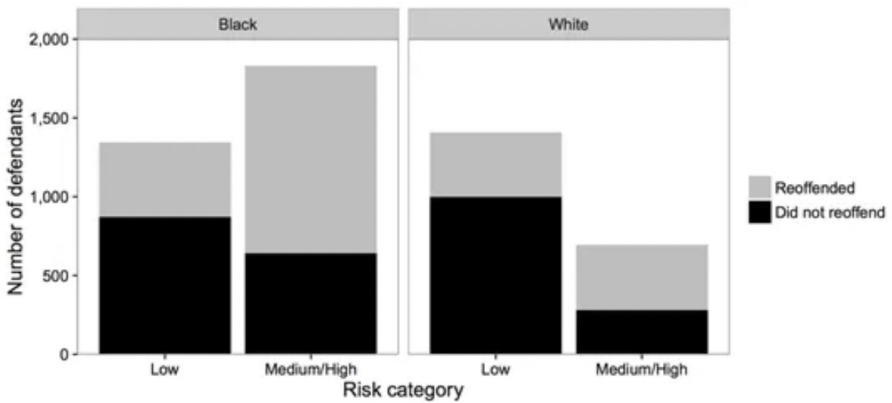
- either $f(x)$ is a perfect predictor, with $f(x) \in \{0,1\}$,
- or the two groups have equal base rates.

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

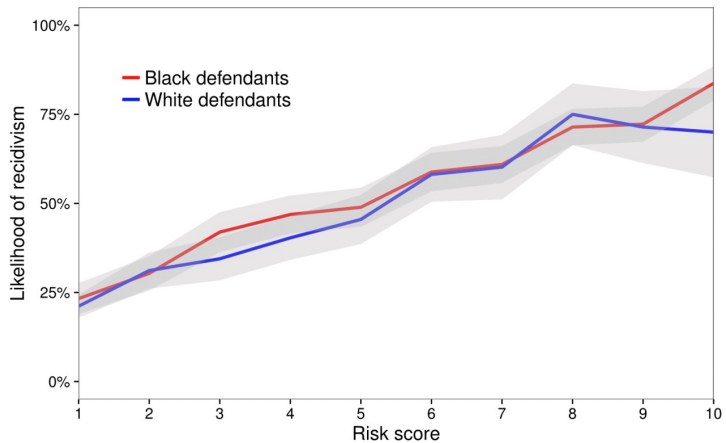
What does this say about COMPAS?

- The recidivism rate among black defendants in the data is 51%, compared to 39% for White defendants.
- Classifier is far from perfect (~60% accuracy)

COMPAS: Unfair because black defendants who did not recommit crime are assigned higher score (i.e. does not obey balance)



COMPAS: Fair because probability of recommitting crime is similar for a given risk score, for both groups (i.e. is calibrated)



https://medium.com/soal-food/what-makes-an-algorithm-fair-6ad64d75dd0c

Also see *Fair prediction with disparate impact: A study of bias in recidivism prediction instruments,* Chouldechova et al. '17

Achieving these notions: Post-processing for calibration

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 $a \in \{0,1\}$

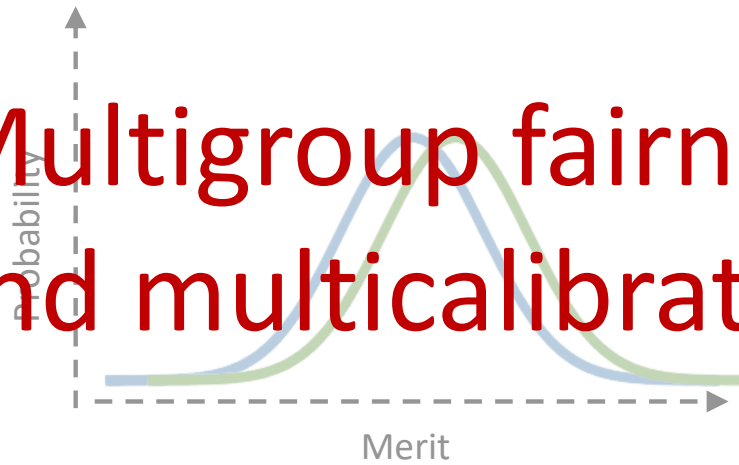
We want use p to get classifier f which is calibrated:

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

Basic idea: Suppose we have datapoints on which p has the same score, say 0.3. Assign all of these datapoints to score $f(x) = \Pr[y = 1 \mid p(x) = 0.3]$

Histogram binning and Isotonic regression implement and generalize this.

Multigroup fairness and multicalibratoin

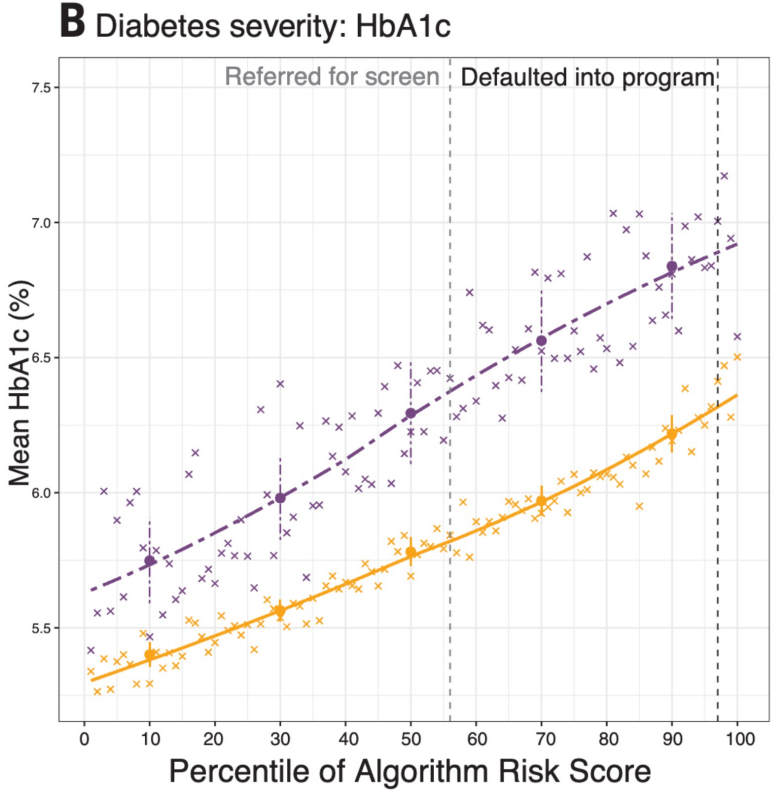


Earlier: Calibration across groups

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.



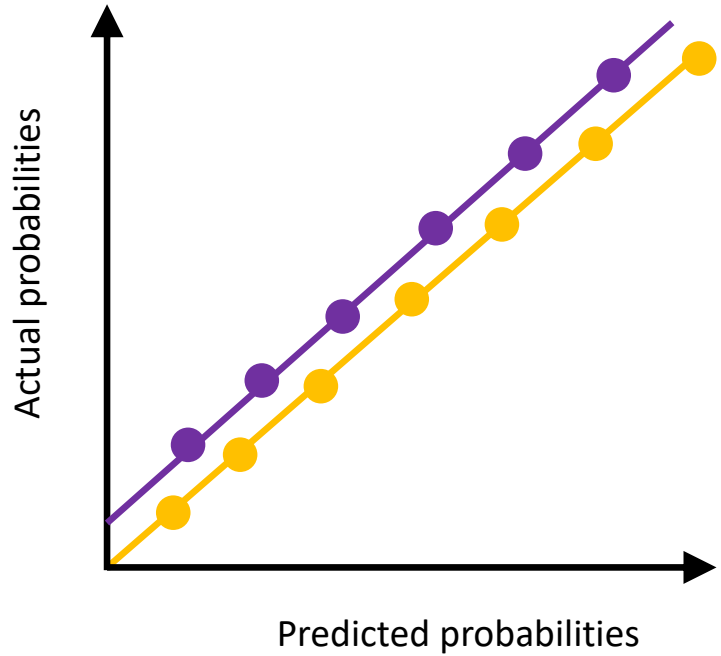
Existing medical risk predictors are miscalibrated across groups

Is calibration across groups enough?

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.



Ideal scenario: Calibrated, meaningful predictions for both yellow and purple groups.

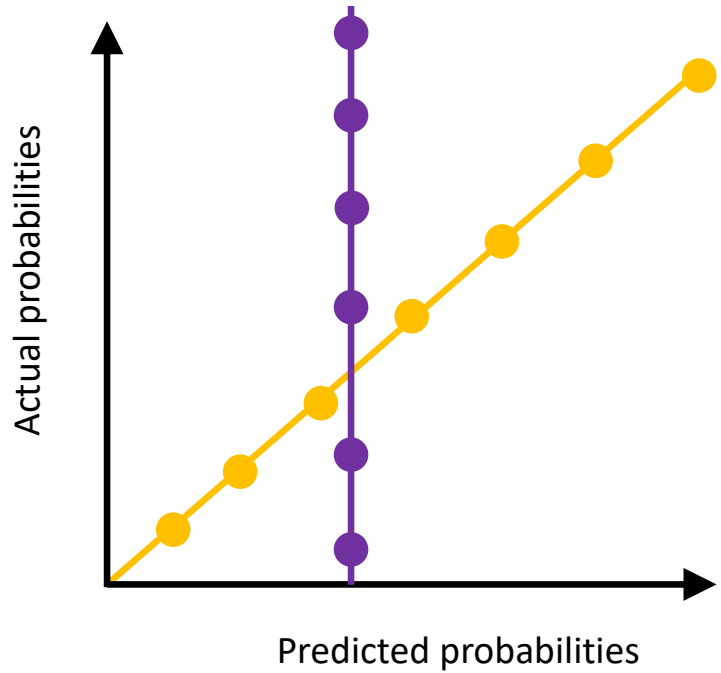
Fig. based on talk by Michael Kim, Multicalibration and Outcome Indistinguishability, Simons Institute

Is calibration across groups enough?

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***, but they may not say much!***



Issue (**Algorithmic stereotyping**): Can achieve calibration on purple group by just predicting average on the group

Fig. based on talk by Michael Kim, Multicalibration and Outcome Indistinguishability, Simons Institute

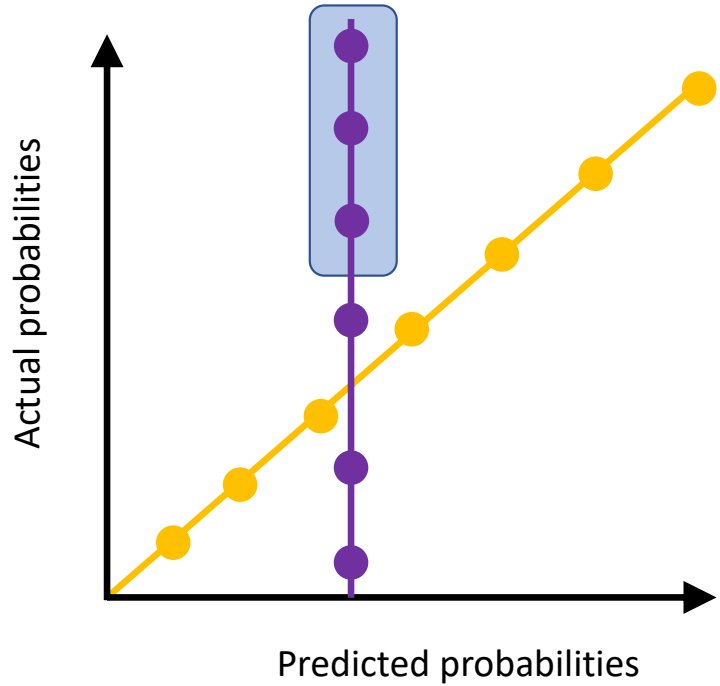
Calibration across rich set of groups

Idea: Let's try to get calibration on ***all*** groups in the data.

Can we get calibration on groups which correspond to single datapoints in the population?

No, since it would require estimating the ground truth probabilities accurately for each datapoint.

Instead: Can we capture some large, meaningful set of groups?



The model here is not calibrated on the individuals in the group in the blue box .

Fig. based on talk by Michael Kim, Multicalibration and Outcome Indistinguishability, Simons Institute

Multicalibration: Calibration for identifiable groups

Definition (Multicalibration, Hebert-Johnson et al. '18): A predictor f is multicalibrated with respect to a collection of groups C, if f is calibrated for every $c \in C$, i.e.

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

Group calibration

- $C =$ small number of groups
- **Easy to achieve**
- **Weak, allows algorithmic stereotyping**

Individual fairness

- $C = \{\{x\} : x \in \text{domain}\}$
- **Cannot be efficiently achieved**
- **Strong individual-level fairness guarantee**

Multicalibration: Calibration for identifiable groups

Definition (Multicalibration, Hebert-Johnson et al. '18): A predictor f is multicalibrated with respect to a collection of groups C, if f is calibrated for every $c \in C$, i.e.

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

Group calibration

- C = small number of groups
- **Easy to achieve**
- **Weak, allows algorithmic stereotyping**

Individual fairness

- $C = \{\{x\} : x \in \text{domain}\}$
- **Cannot be efficiently achieved**
- **Strong individual-level fairness guarantee**

Multi-calibration

- C = potentially infinite
- **Efficiently achievable (under certain conditions)**
- **Much stronger fairness protection than group calibration**

Multicalibration: Some nice properties

Definition (Multicalibration, Hebert-Johnson et al. '18): A predictor f is multicalibrated with respect to a collection of groups C, if f is calibrated for every $c \in C$, i.e.

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

- **(Efficiently achievable)** If C is defined by a collection of groups such that it is easy to check for violation of calibration for $c \in C$, then can efficiently postprocess a predictor to be multicalibrated with respect to C (without hurting its accuracy).

- **(Gives predictors optimal for multiple downstream loss functions)** If f is multicalibrated, then is f is optimal with respect to multiple loss functions at the same time (up to postprocessing). This is called *omniprediction* (Gopalan et al. 2022).

- **(Gives predictors robust to distribution shifts)** If f is multicalibrated with respect to some groups, then can get group robustness for those groups (Kim et al. 2022).