Equality of Opportunity in Supervised Learning

Moritz Hardt, Eric Price, Nathan Srebro

Presented by: Ronghao Zhang, Wenting Song

Introduction

- proposed a methodology for measuring and preventing discrimination based on a set of sensitive attributes(race, gender)
- individuals who qualify for a desirable outcome should have an equal chance of being correctly classified for this outcome
- implemented it in the AIF360 package as **post-processing** technique

Equalized odds/ Equal opportunity

Goal: predict a true outcome **Y** from features **X** based on labeled training data, while ensuring the prediction is "non-discriminatory" with respect to a specified protected attribute **A**

Definition 2.1 (Equalized odds)**.** We say that a predictor $\widehat{Y}$ satisfies *equalized odds* with respect to protected attribute A and outcome Y, if $\widehat{Y}$ and A are independent conditional on Y.

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

Definition 2.2 (Equal opportunity)**.** We say that a binary predictor $\widehat{Y}$ satisfies *equal opportunity* with respect to A and Y if $\Pr\{\widehat{Y}=1 \mid A=0, Y=1\} = \Pr\{\widehat{Y}=1 \mid A=1, Y=1\}$.

Oblivious measures

Real-valued scores Even if the target is binary, a real-valued predictive score $R = f(X, A)$ is often used (e.g. FICO scores for predicting loan default), with the interpretation that higher values of R correspond to greater likelihood of $Y = 1$ and thus a bias toward predicting $\widehat{Y} = 1$. A binary classifier $\widehat{Y}$ can be obtained by thresholding the score, i.e. setting $\widehat{Y} = \mathbb{I}\{R > t\}$ for some threshold t. Varying this threshold changes the trade-off between sensitivity and specificity.

Definition 2.3. A property of a predictor $\widehat{Y}$ or score R is said to be *oblivious* if it only depends on the joint distribution of $(Y, A, \widehat{Y})$ or (Y, A, R), respectively.

Simulating loan thresholds

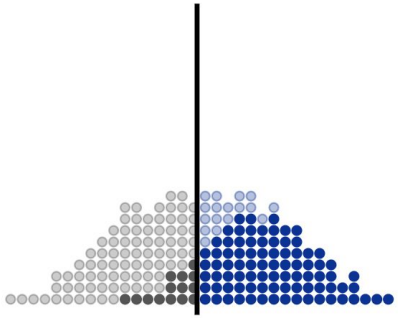
Drag the black threshold bars left or right to change the cut-offs for loans.

Threshold Decision

Credit Score
higher scores represent higher likelihood of payback

0 10 20 30 40 50 60 70 80 90 100

loan threshold: 50



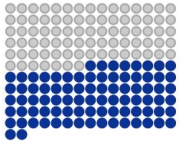
each circle represents a person, with dark circles showing people who pay back their loans and light circles showing people who default

Color

denied loan / would default
denied loan / would pay back
granted loan / defaults
granted loan / pays back

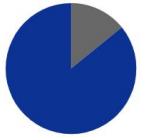
Outcome

Correct 84%
loans granted to paying applicants and denied to defaulters



Incorrect 16%
loans denied to paying applicants and granted to defaulters

True Positive Rate 86%
percentage of paying applications getting loans



Positive Rate 52%
percentage of all applications getting loans

Profit: **13600**

[Link for Simulation]

Achieving non-discrimination

Derived predictor

Definition 3.1 (Derived predictor)**.** A predictor $\widetilde{Y}$ is *derived from a random variable* R *and the protected attribute* A if it is a possibly randomized function of the random variables (R, A) alone. In particular, $\widetilde{Y}$ is independent of X conditional on (R, A).

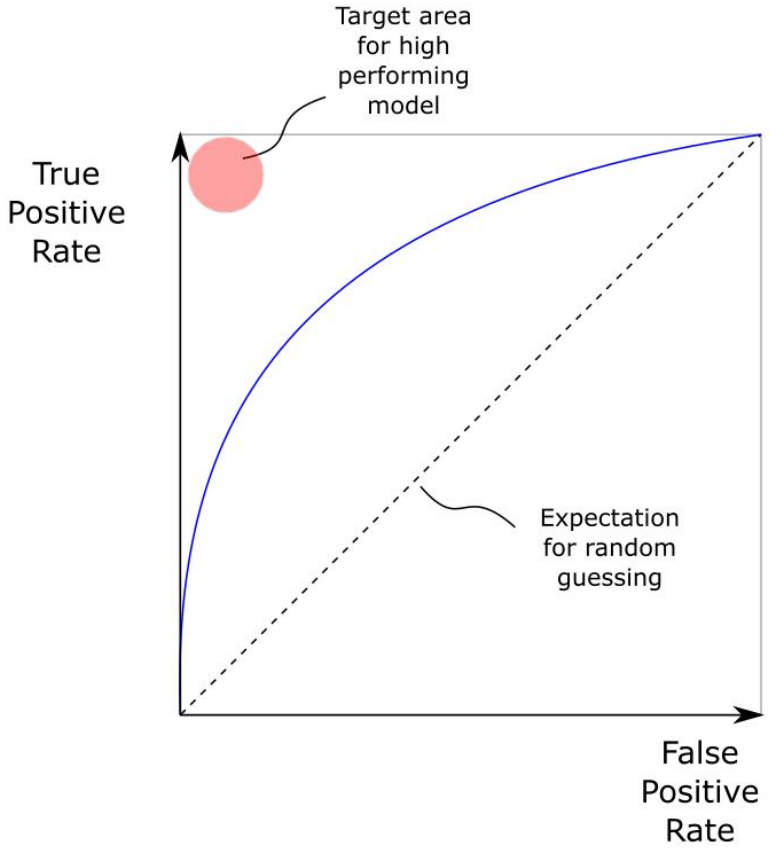
- only depend on **R** and the protected attribute **A**
- only a post-learning step

Deriving from a score function

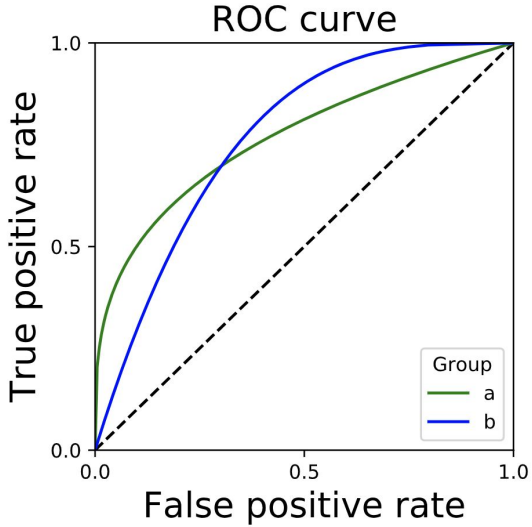
ROC curve of the score

(Receiver Operator Characteristic)

- captures **false positive** (1- specificity) and **true positive** (sensitivity) at different thresholds



Look at ROC curve for each group



Deriving from a score function

- **When the ROC curves can agree**

- A-conditional ROC curves:

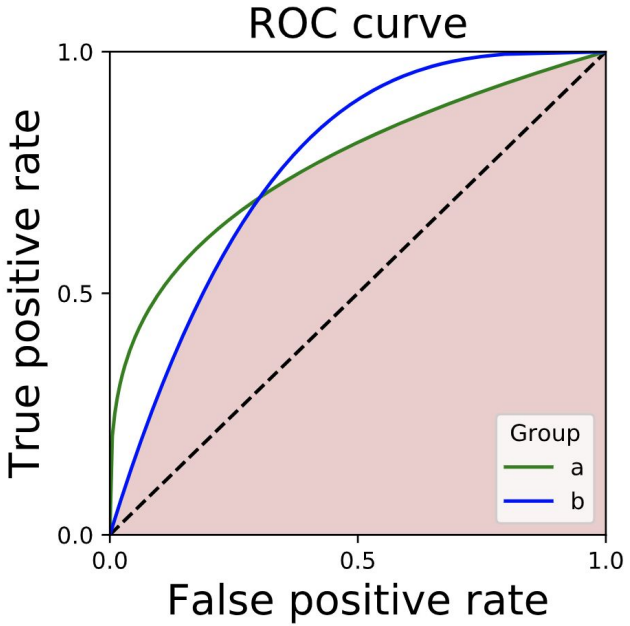
$$C_a(t) \stackrel{\text{def}}{=} \left(\Pr\{\widehat{R} > t \mid A = a, Y = 0\}, \Pr\{\widehat{R} > t \mid A = a, Y = 1\}\right)$$

- **a score function** obeys **equalized odds** if and only if the ROC curves for all values of the protected attribute **agree**.

$$C_a(t) = C_{a'}(t) \text{ for all values of } a \text{ and } t.$$

Deriving from a score function

❑ **When the ROC curves do not agree**



- use **randomization** to fill the span of possible derived predictors.
- for every protected group **a**, consider the **convex hull** of the image of the conditional ROC curve:

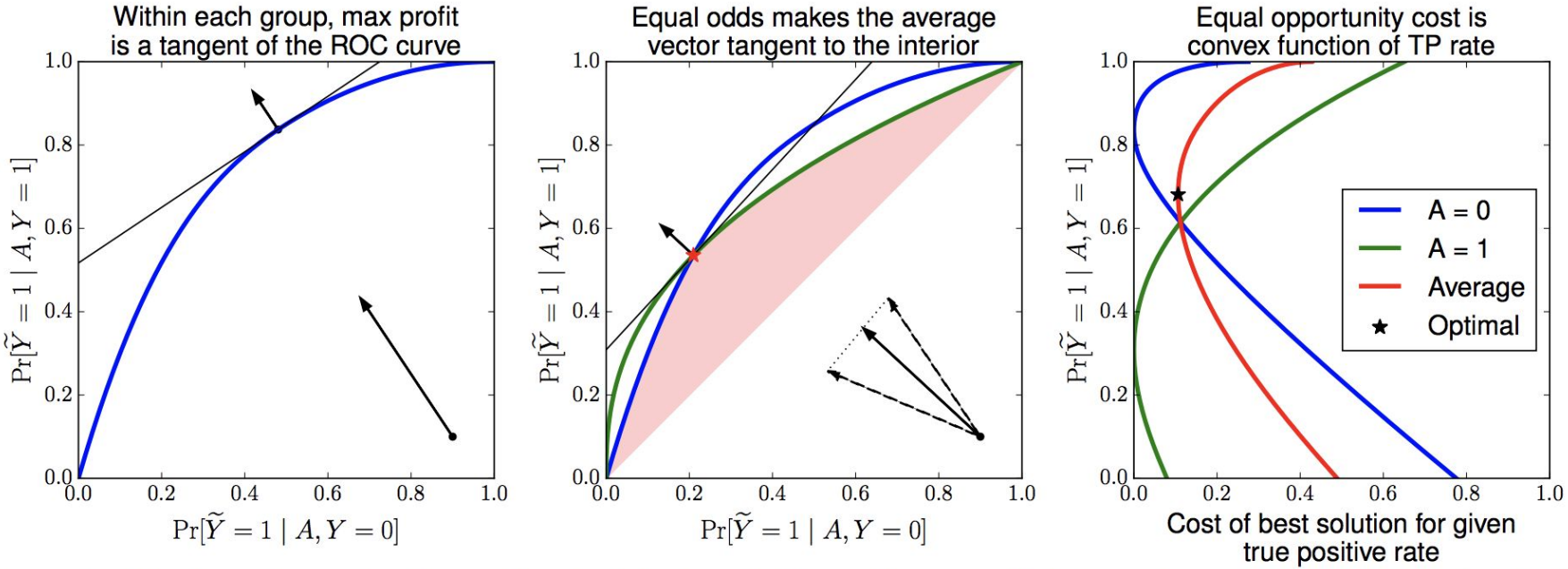


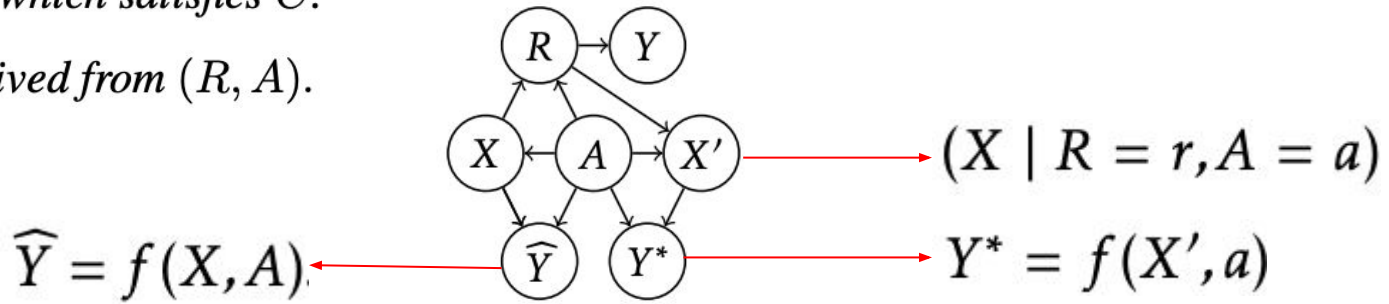
Figure 2: Finding the optimal equalized odds threshold predictor (middle), and equal opportunity threshold predictor (right). For the equal opportunity predictor, within each group the cost for a given true positive rate is proportional to the horizontal gap between the ROC curve and the profit-maximizing tangent line (i.e., the two curves on the left plot), so it is a convex function of the true positive rate (right). This lets us optimize it efficiently with ternary search.

Bayes Optimal Predictors

Bayes optimal equalized odds predictor can be obtained as an derived threshold predictor of the Bayes optimal regressor.

Proposition 4.2. *For any source distribution over* (Y, X, A) *with Bayes optimal regressor* $R(X, A)$*, any loss function, and any oblivious property* C*, there exists a predictor* $Y^*(R, A)$ *such that:*

1. Y^* *is an optimal predictor satisfying* C*. That is,* $\mathbb{E}\ell(Y^*, Y) \leqslant \mathbb{E}\ell(\widehat{Y}, Y)$ *for any predictor* $\widehat{Y}(X, A)$ *which satisfies* C*.*
2. Y^* *is derived from* (R, A)*.*



Case Study
FICO Score Threshold Choice

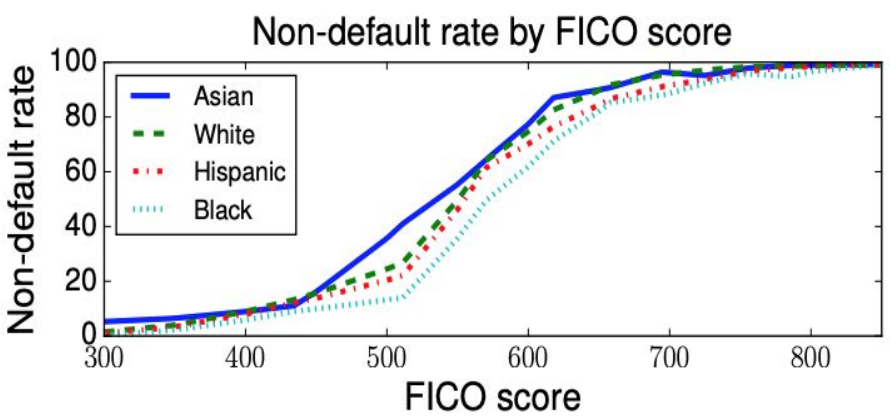
Case study: FICO scores - Score (R) Outcome (Y) Protected Attribute (A)

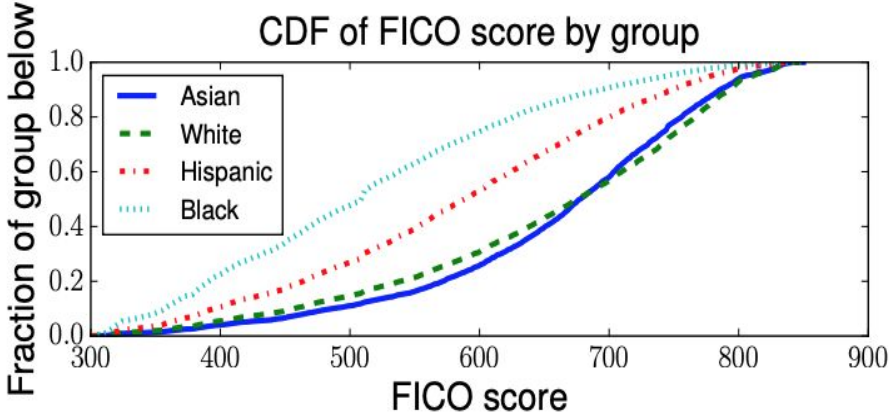
CATEGORY	SCORE
Excellent (30% of People)	750 - 850
Good (13% of People)	700 - 749
Fair (18% of People)	650 - 699
Poor (34% of People)	550 - 649
BAD (16% of People)	350 - 549

Race = { “Asian”, “White”, “Hispanic”, “Black” }

People were labeled as in default if they failed to pay a debt for at least 90 days on at least one account in the ensuing 18-24 month period

Case study: FICO scores - Input Data





Input parameters:

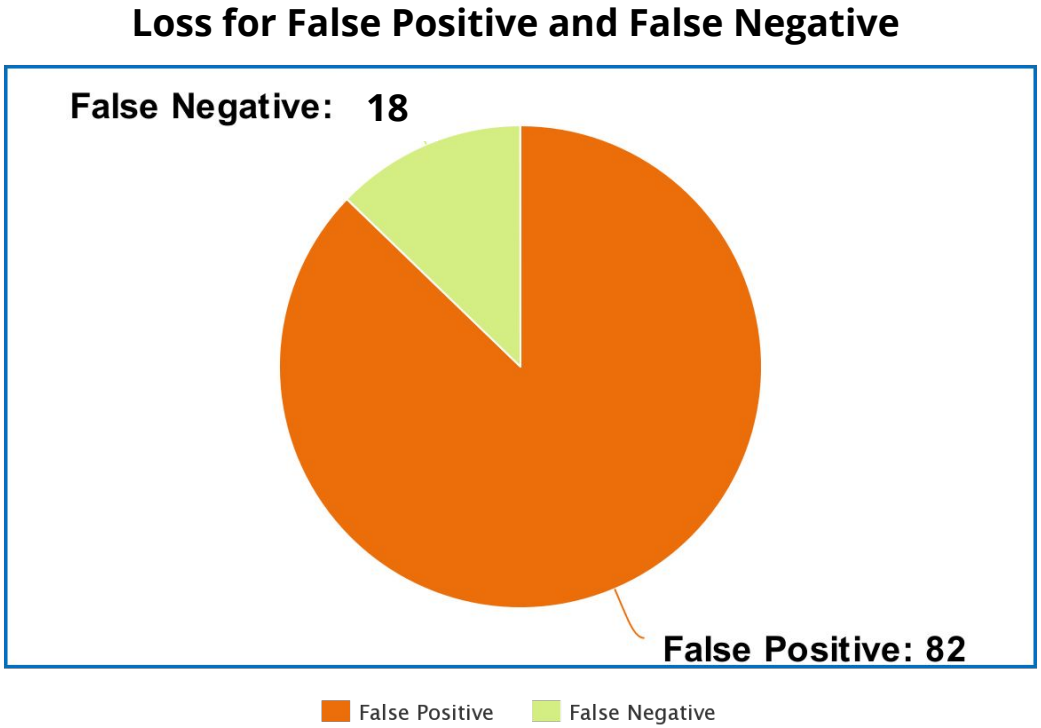
- Non-default rate
- Cumulative Distribution Function (CDF) of FICO
- Number of people per group

Case study: FICO scores - Loss function

- **What's the profit?**

false positives - giving loans to people that default on any account

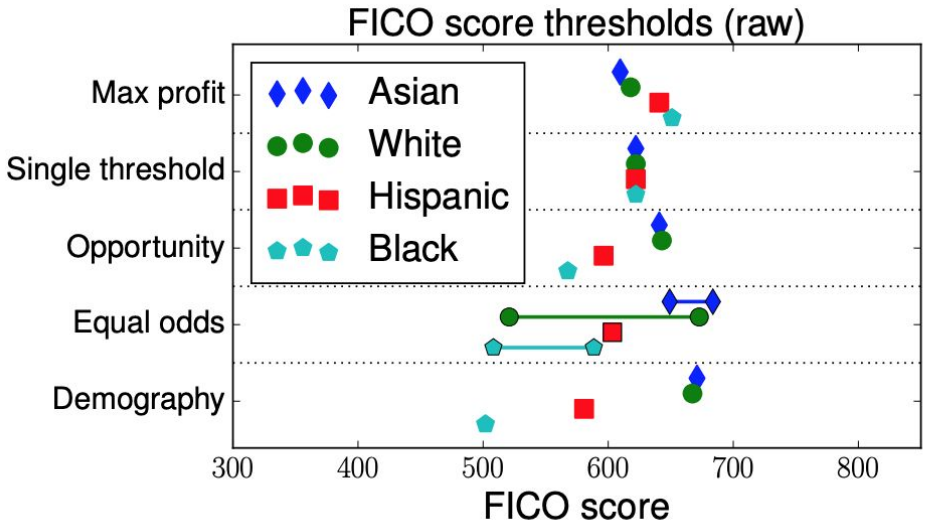
false negatives - not giving a loan to people that don't default

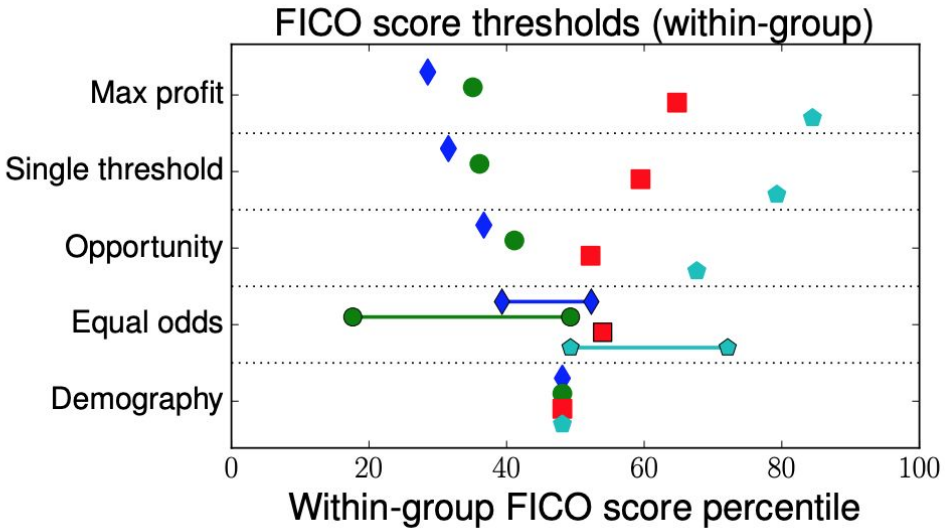


Case study: FICO scores - Different Threshold

Max Profit	Pick threshold at which 82% of people approved are NON-DEFAULTERS for each group	+ Maximize the profit - No fairness constraint
Race Blind	Pick threshold at which 82% of NON-DEFAULTERS get approved in TOTAL	+ Non-discriminating - Scalability issue
Demographic Parity	Fraction of people that get approved is the same for each group	+ Intuitive, widely discussed - Bad profit
Equal Opportunity	Fraction of NON-DEFAULTER that get approved is the same for each group	+ Great balance - Punish defaulters
Equalized Odds	Fraction of NON-DEFAULTER and DEFAULTER that get approved stays the same across groups	+ More fairness - Less profit - Randomization

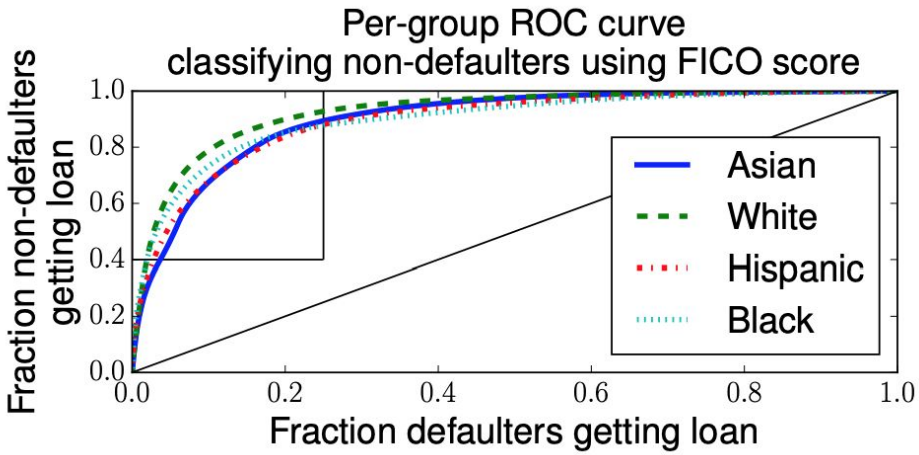
Case study: FICO scores - Results (How fair?)

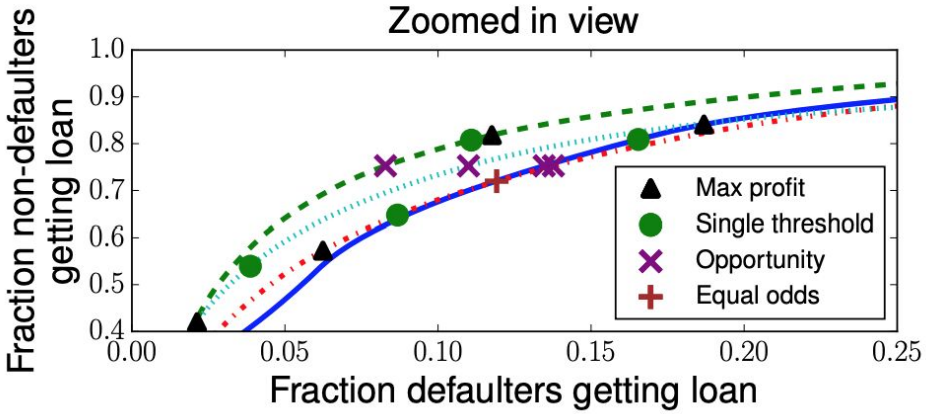




This figure shows the thresholds used by each predictor

Case study: FICO scores - Results (ROC Curve)





Differences in the ROC curve indicate differences in predictive accuracy between groups

Profit achieved by each method, as a fraction of the max profit achievable.

- Race blind threshold gets 99.3% of the maximal profit
- Equal opportunity gets 92.8%
- Equalized odds gets 80.2%
- Demographic parity only 69.8%.

Conclusion

Accomplishments:

- A more accurate and appropriate fairness notion (unlike demo parity)
- Better aligned with the goal of supervised learning
- Simple, efficient and privacy-preserving

Cautions:

- Domain-specific scrutiny is required in defining and collecting a reliable target variable (the variable that is or should be the output)
- Collecting features that directly capture the target (avoid redundant encodings)

Thanks