Equality of Opportunity in Supervised Learning

Speaker : Yupei Du
AntNLP

Introduction

Background & Former Work

Background

- Anti-discrimination Law
- ML helps obtain more accurate predictions in sensitive areas
- However, algorithm may introduce new biases
  - Embeddings: Bolukbasi et al.(2014)
  - Coreference Resolution: Zhao et al.(2016)
  - Structured Prediction: Zhao et al.(2016)

Other Approaches

- Fairness through unawareness
  - Redundant Encoding

- Demographic Parity
  - Doesn’t ensure fairness
  - Cripple utility

This Work

- Goal : Predict Y from feature X while ensuring “non-discriminatory” w.r.t. protected attribute A

- Data : Labeled training data

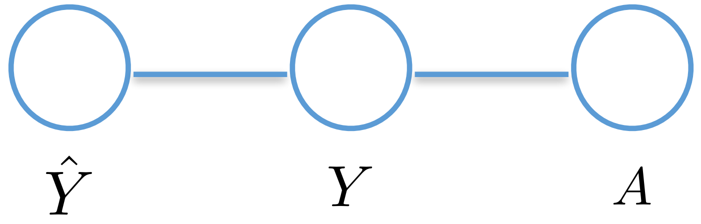
- Notion : “oblivious”, based only on joint distribution $(\hat{Y}, Y, A)$

- Pros:
  - Allow for perfect predictor
  - Possible when functional form aren’t public

Criterion

Equal odds & Equal opportunity

Equal Odds & Equal Opportunity



- Equal odds

$$Pr\{\hat{Y} = 1 | A = 0, Y = y\} = Pr\{\hat{Y} = 1 | A = 1, Y = y\}, y \in \{0, 1\}$$

- Demographic Parity

$$Pr\{\hat{Y} = 1 | A = 0, Y = 1\} = Pr\{\hat{Y} = 1 | A = 1, Y = 1\}$$

Real-valued scores

- Set a threshold t $\hat{Y} = \mathbb{I}\{R > t\}$
- Tradeoff between true positive rate(TPR) and false positive rate(FPR)
- Randomized thresholds to change TPR and FPR

Methods

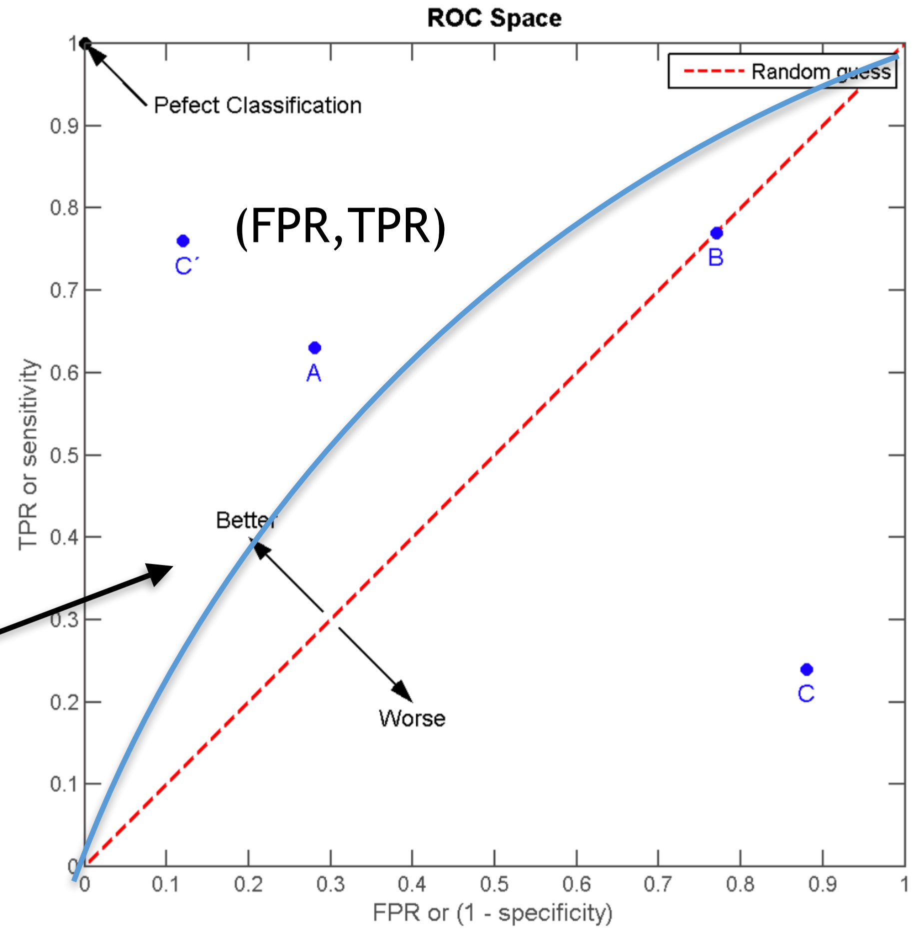
Achieving Equal odds & Equal opportunity

ROC Curve

	Observed positive	Observed negative
Predicted positive	TP	FP
Predicted negative	FN	TN

TPR = TP/(TP+FN)
FPR = FP/(FP+TN)

Convex



Derived Predictor

- Definition

A predictor 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

- Notations

$$\gamma_a(\hat{Y}) = (Pr\{\hat{Y} = 1 | A = a, Y = 0\}, Pr\{\hat{Y} = 1 | A = a, Y = 1\})$$

$$P_a(\hat{Y}) = convhull\{(0,0), \gamma_a(\hat{Y}), \gamma_a(1 - \hat{Y}), (1,1)\}$$

Derived Predictor

A predictor $\tilde{Y}$ is derived if and only if for all $a \in \{0,1\}$,we have $\gamma_a(\tilde{Y}) \in P_a(\hat{Y})$

- Optimization Problem

$$\min_{\tilde{Y}} \mathbb{E}\mathcal{L}(\tilde{Y}, Y)$$

$$s.t.\ \forall a \in \{0,1\} : \gamma_a(\tilde{Y}) \in P_a(\hat{Y}) \qquad \gamma_0(\tilde{Y}) = \gamma_1(\tilde{Y})$$

- Solution is an optimal equalized odds predictor
- Linear program whose coefficients can be computed from joint distribution $(\hat{Y}, A, Y)$

Derived Predictor

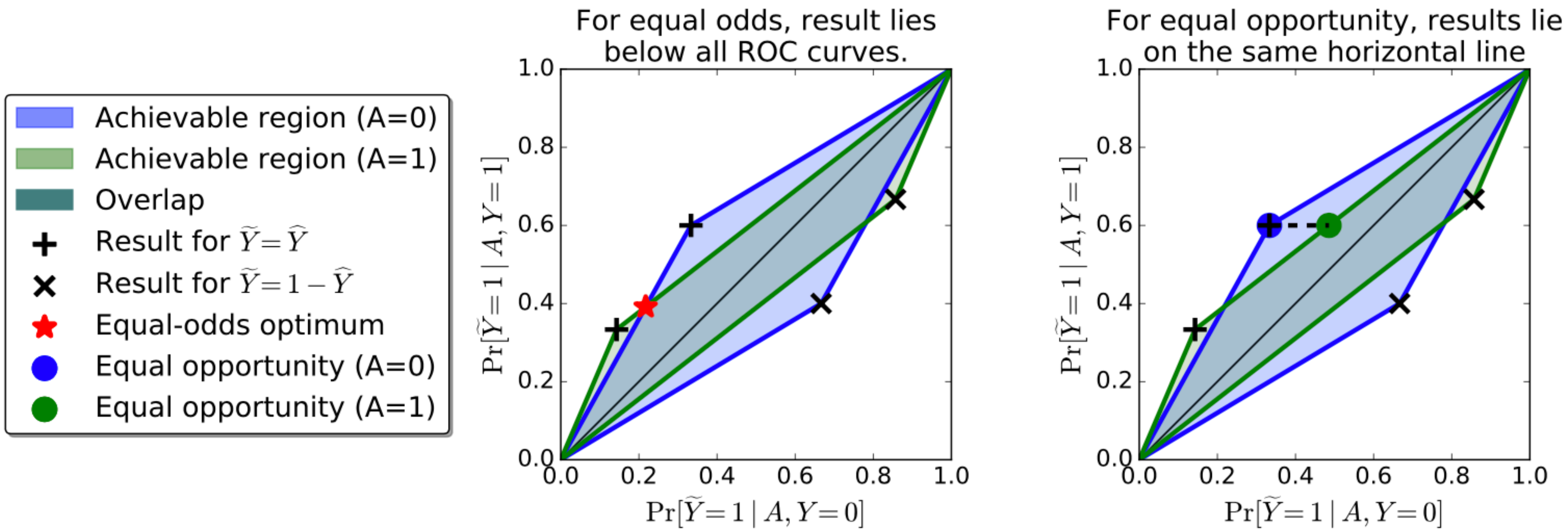


Figure 1: Finding the optimal equalized odds predictor (left), and equal opportunity predictor (right).

Deriving from score function

- When score function satisfies equality
  - Thresholding it!

- When doesn't
  - Choose different thresholds for different As

A-conditional ROC curves:

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

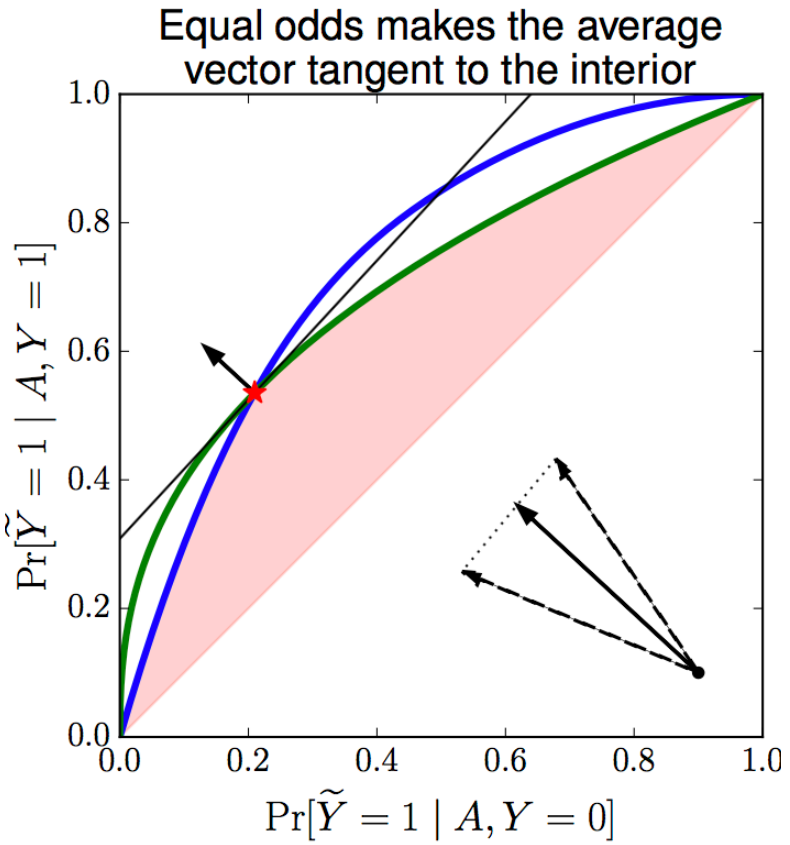
$$D_a \stackrel{\text{def}}{=} \text{convhull}\{C_a(t) \colon t \in [0, 1]\}$$

Threshold Predictor

- Intersect
  - Poor tradeoff between TPR & FPR
- Equal odds
  - Minimum performance
  - Randomized($t_a \& t^a$)

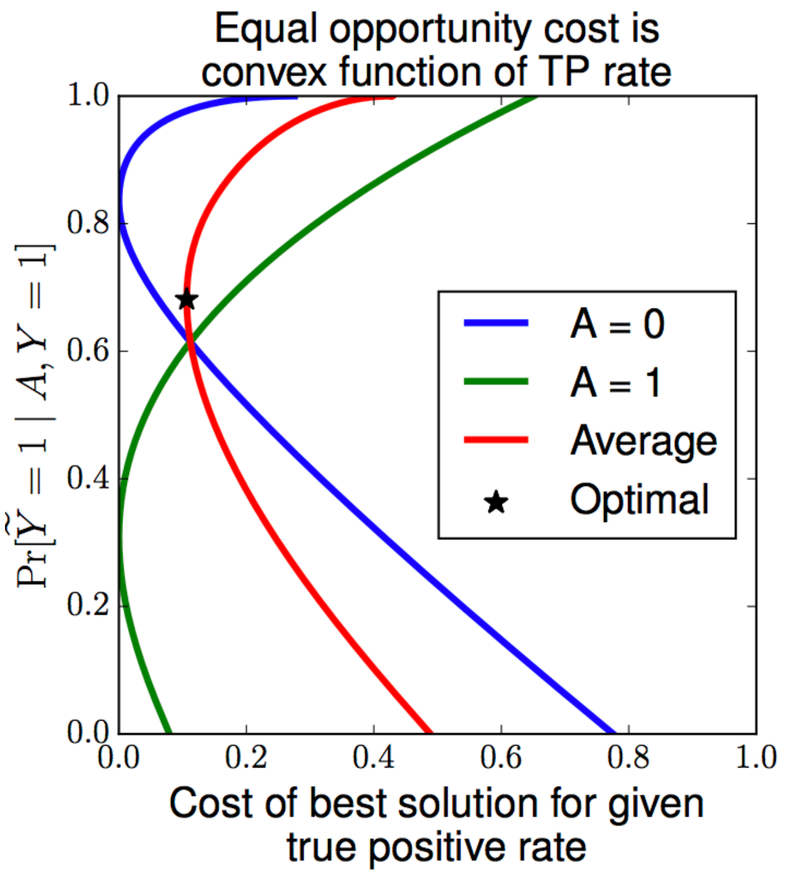
Optimize: $\min_{\forall a:\ \gamma \in D_a} \gamma_0 \mathcal{L}(1,0) + \gamma_1 \mathcal{L}(0,1)$

Ternary Search



Threshold Predictor

- Equal opportunity
  - No randomize
  - Convex function
  - Ternary search



Bayes Optimal Predictor

- Bayes optimal regressor

$$R = \arg\min_{r(x,a)} \mathbb{E}[(Y - r(X, A))^2] = r^*(X, A) \qquad r^*(x, a) = \mathbb{E}[Y|X = x, A = a]$$

- Bayes optimal predictor with constraints

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,* $\mathbb{E}\mathcal{L}(Y^*, Y) \leq \mathbb{E}\mathcal{L}(\hat{Y}, Y)$ *for any predictor* $\hat{Y}(X, A)$ *which satisfies* C.

2. Y^* *is derived from* (R, A).

Nearly Optimality

- Conditional Kolmogorov distance

$$d_K(R,R') \stackrel{\text{def}}{=} \max_{a,y\in\{0,1\}} \sup_{t\in[0,1]} \left|\Pr\{R>t \mid A=a, Y=y\} - \Pr\{R'>t \mid A=a, Y=y\}\right|$$

- Distance between points

Let $R, R' \in [0,1]$ *be random variables in the same probability space as* A *and* Y. *Then, for any point* p *on a restricted ROC curve of* R, *there is a point* q *on the corresponding restricted ROC curve of* R' *such that* $\|p-q\|_2 \sqrt{2 \cdot d_K(R,R')}$.

- Difference between Losses

$$\mathbb{E}\ell(\widehat{Y}, Y) \leqslant \mathbb{E}\ell(Y^*, Y) + 2\sqrt{2} \cdot d_K(\widehat{R}, R^*)$$

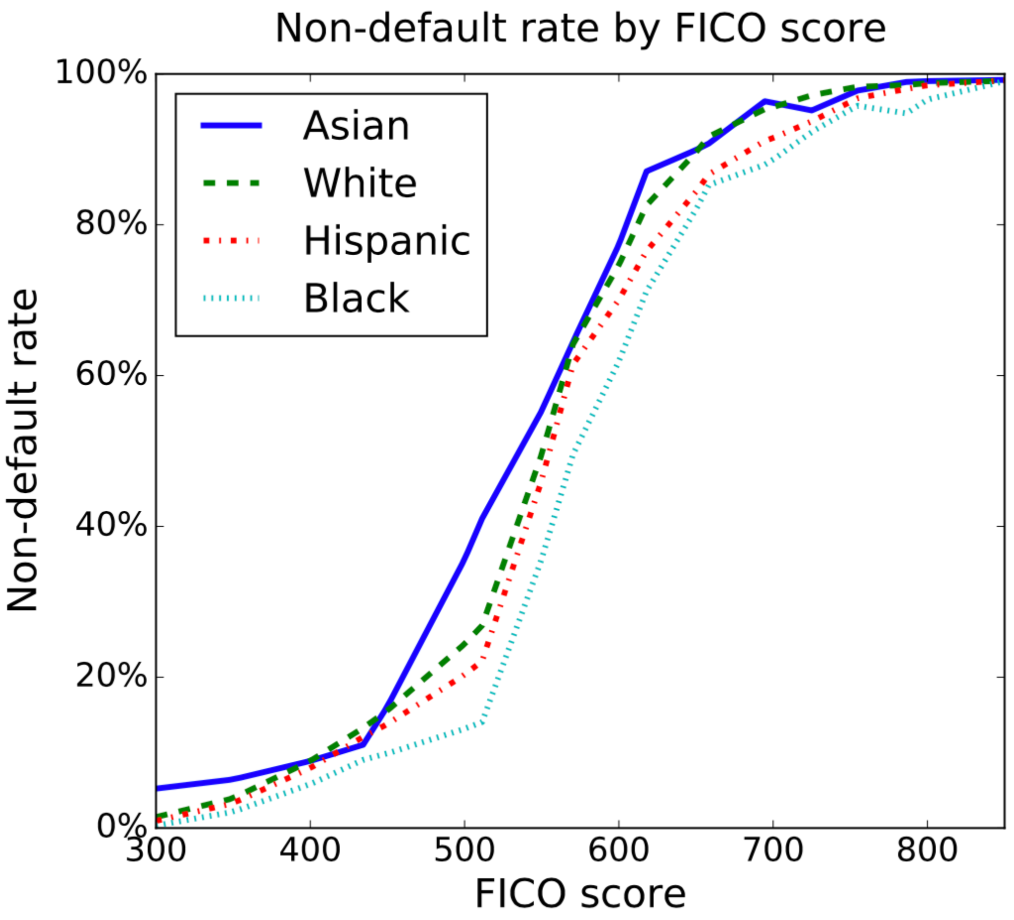
Experiment

Case study: FICO scores

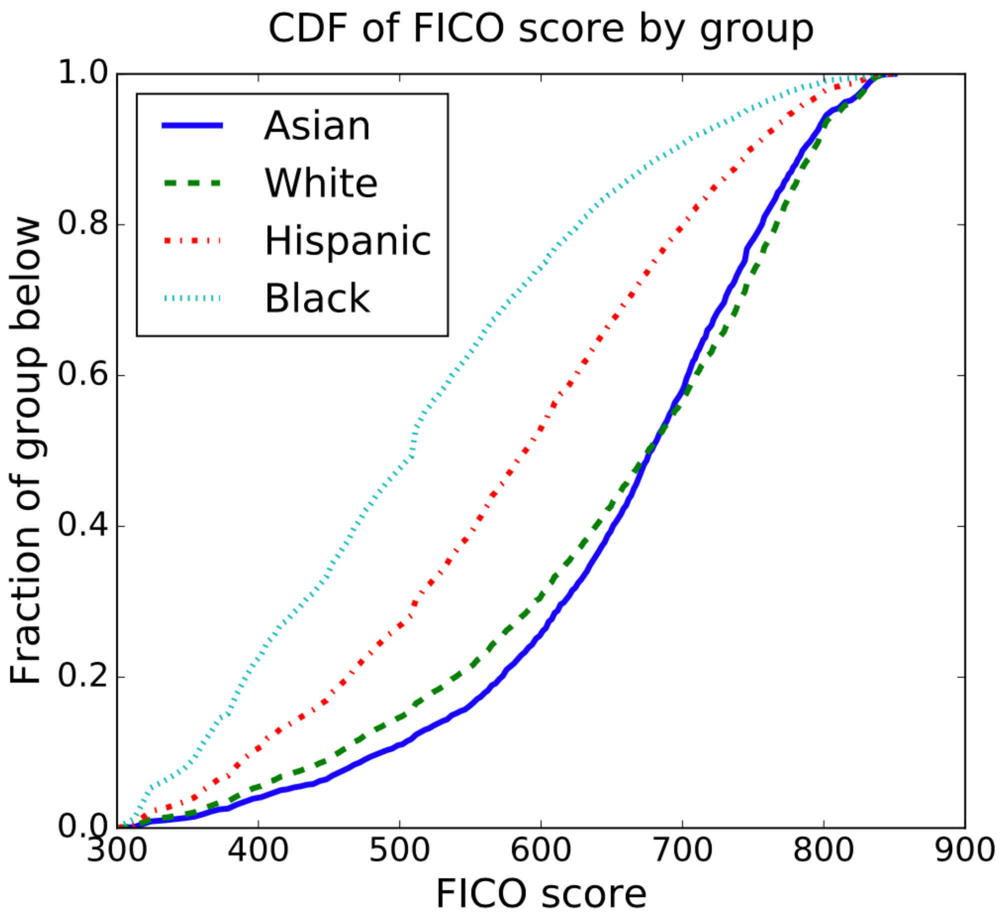
FICO Scores

- Dataset
  - 301536 TransUnion TransRisk scores from 2003
  - Score : 300~800
  - Label : default/payback
  - Protected Attribute: race (white, hispanic, black, asian)
  - Threshold : 620, corresponding to default rate at 18%
  - Cost : FP is 82/18 expensive as FN
- Strategy : max profit / race blind / demographic parity/\
  equal odds / equal opportunity

Input Data

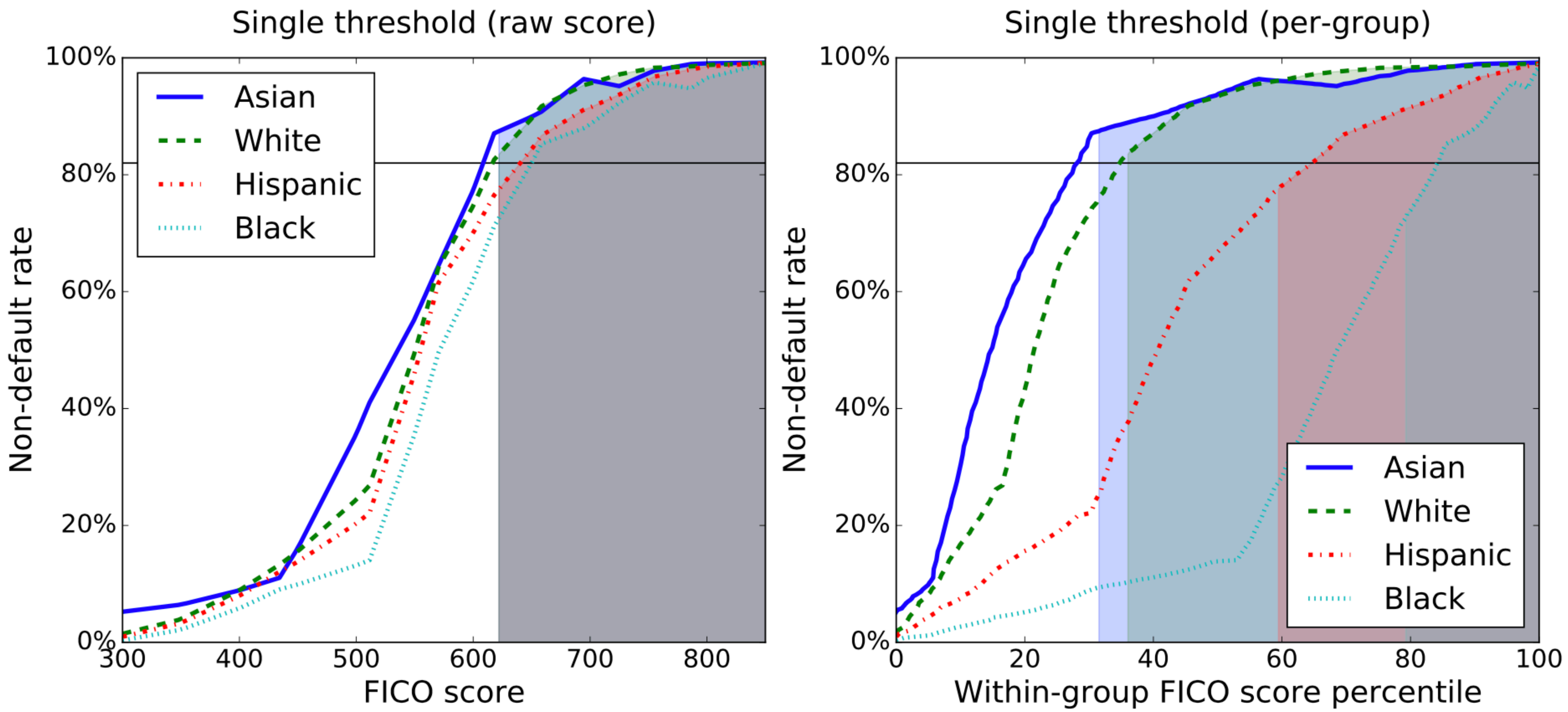


Non-default rate by FICO score
100%
80%
60%
40%
20%
0%
Non-default rate
300
400
500
600
700
800
FICO score
Asian
White
Hispanic
Black



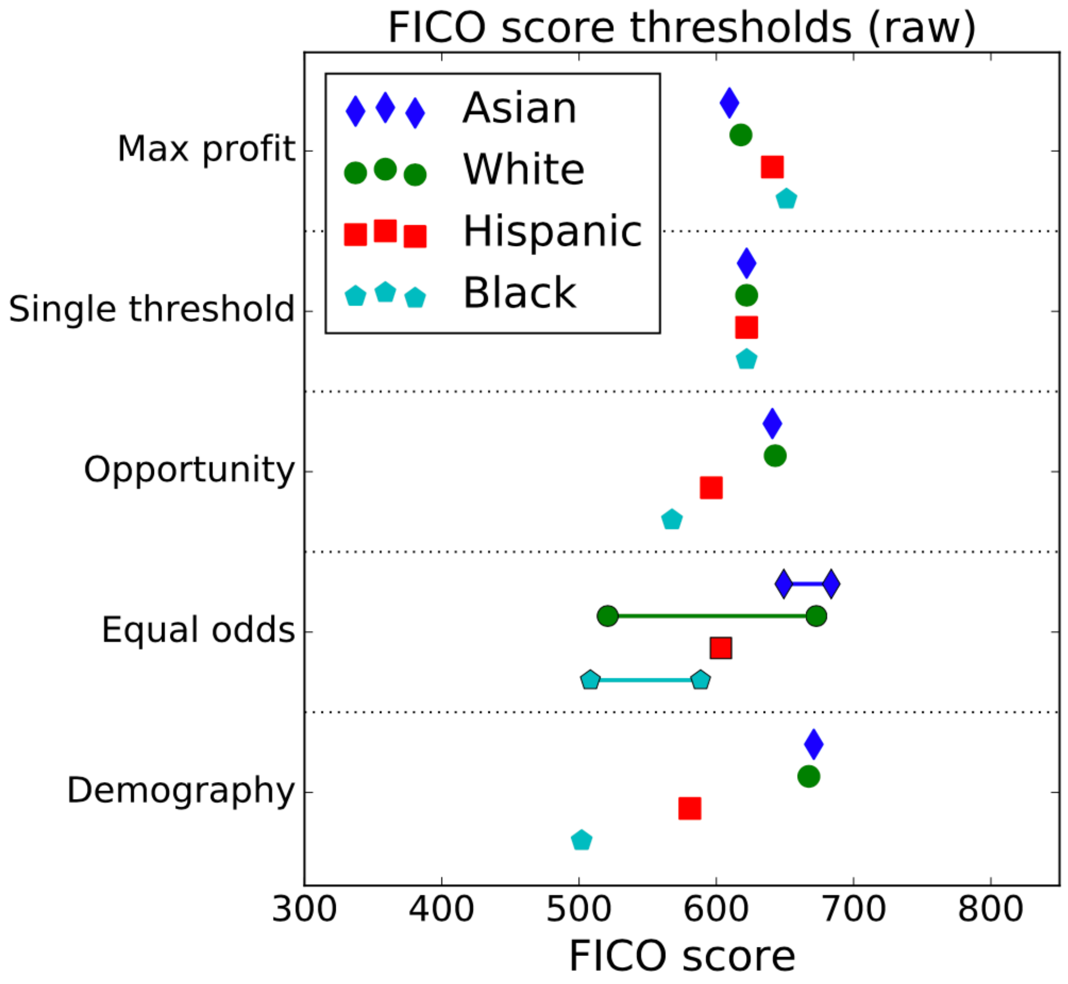
CDF of FICO score by group
1.0
0.8
0.6
0.4
0.2
0.0
Fraction of group below
300
400
500
600
700
800
900
FICO score
Asian
White
Hispanic
Black

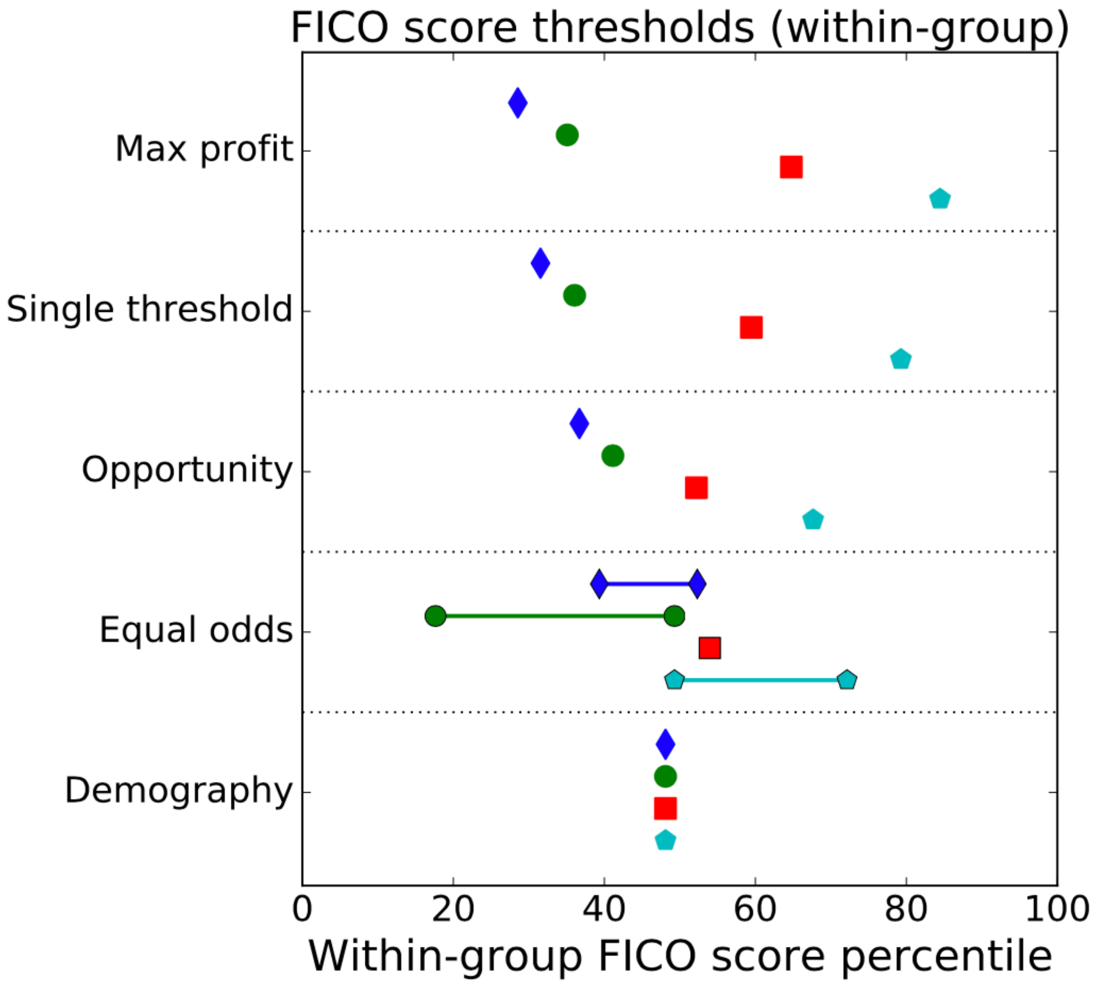
Input Data



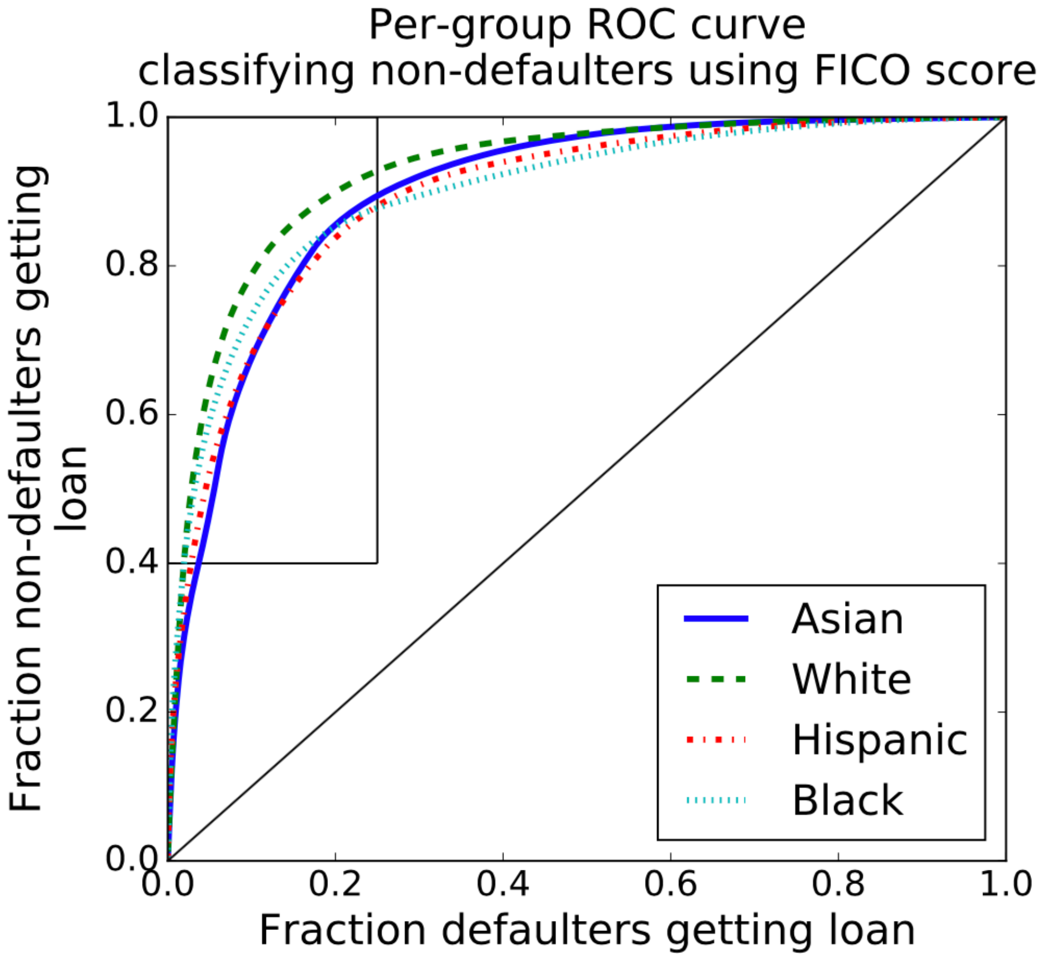
Single threshold (raw score)
100%
80%
60%
40%
20%
0%
Non-default rate
300
400
500
600
700
800
FICO score
Asian
White
Hispanic
Black
Single threshold (per-group)
100%
80%
60%
40%
20%
0%
Non-default rate
0
20
40
60
80
100
Within-group FICO score percentile
Asian
White
Hispanic
Black

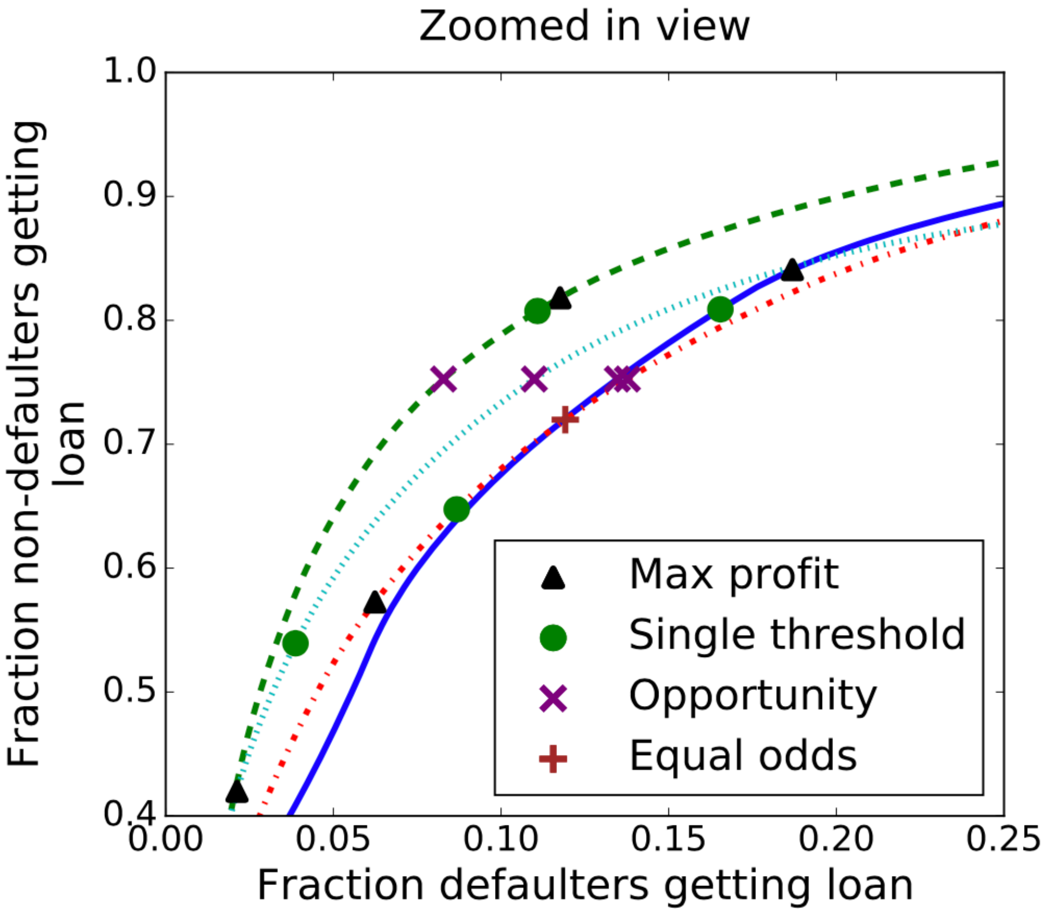
Thresholds


FICO score thresholds (raw)
Asian
White
Hispanic
Black
Max profit
Single threshold
Opportunity
Equal odds
Demography
300
400
500
600
700
800
FICO score

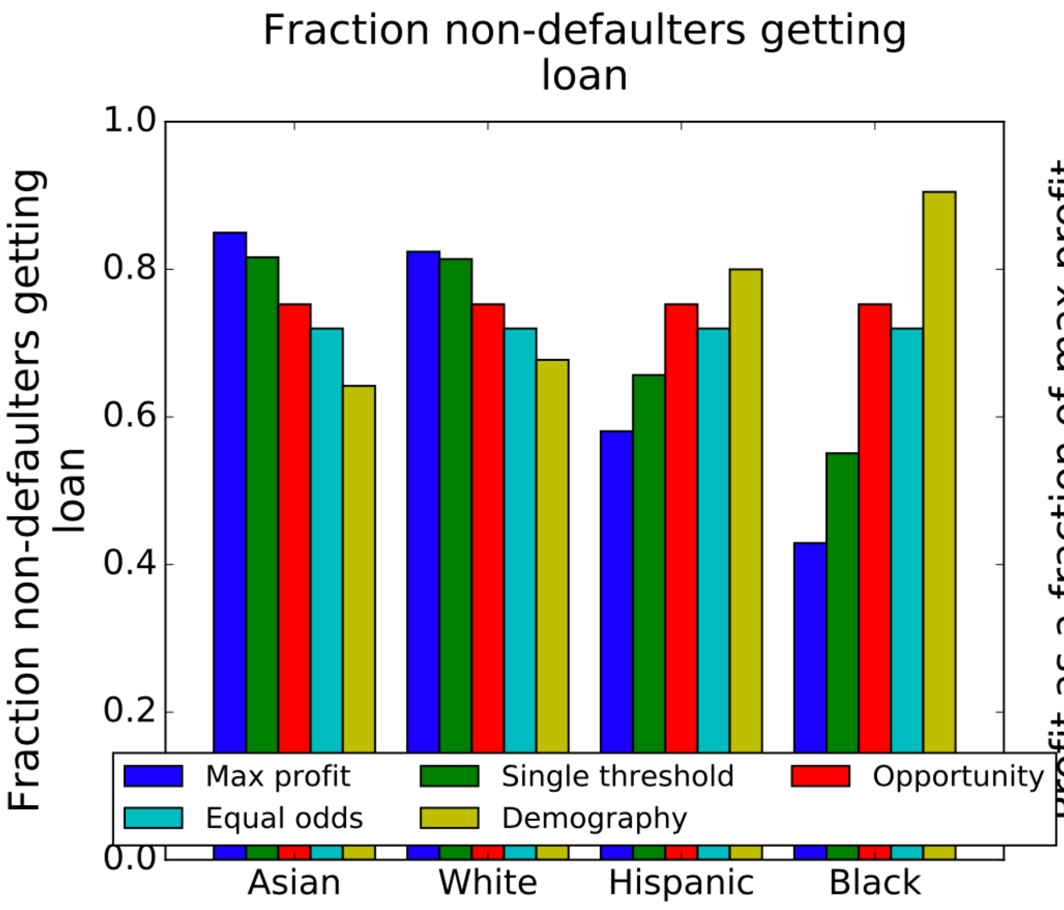

FICO score thresholds (within-group)
Max profit
Single threshold
Opportunity
Equal odds
Demography
0
20
40
60
80
100
Within-group FICO score percentile

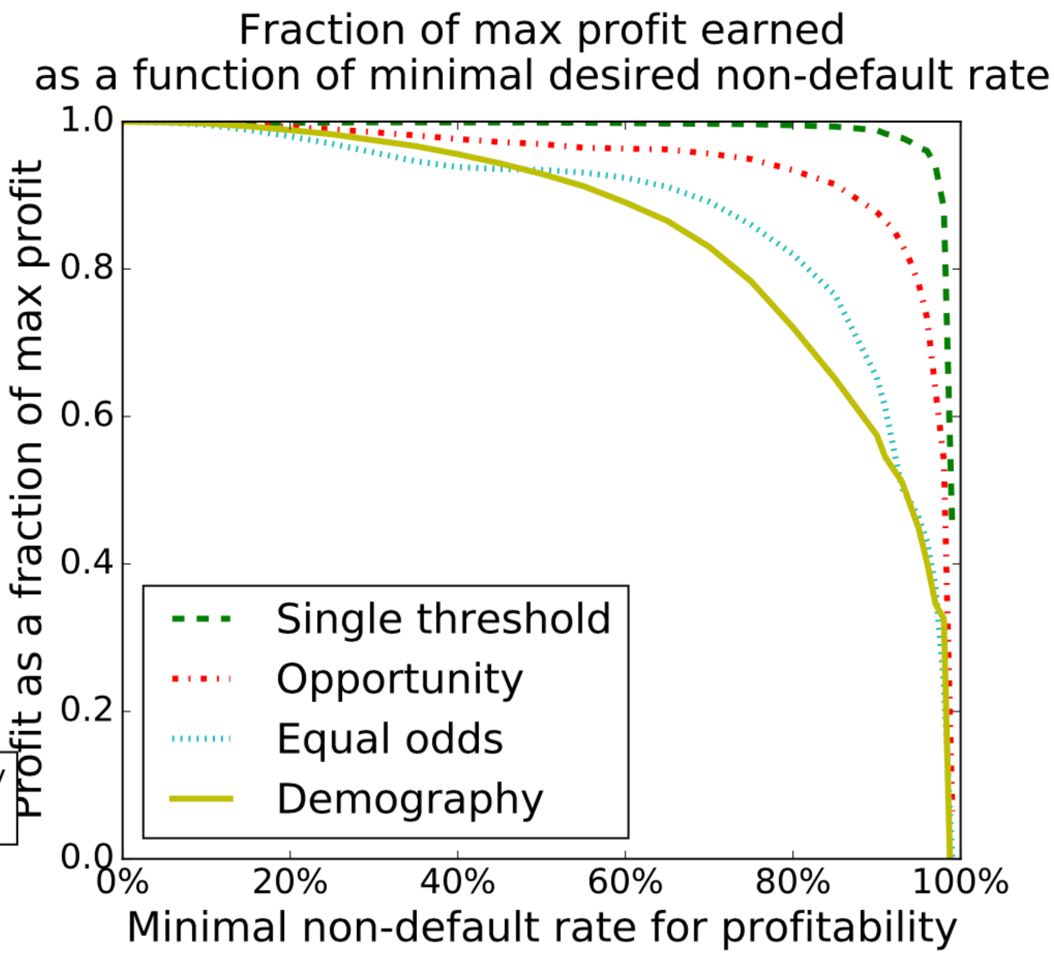
Default Rate





Profit


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


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

Lessones

- Measuring unfairness rather than proving fairness
  - Satisfy notions shouldn’t be considered as a proof of fairness
- Proper Incentives
- When to use post-processing
  - Post-posting process should be processed only when better features and more data are no longer a option
- Burden shifting
  - R is shifted on A compensate for more biases in minor groups on target label caused by uncertainty

THANKS FOR LISTENING