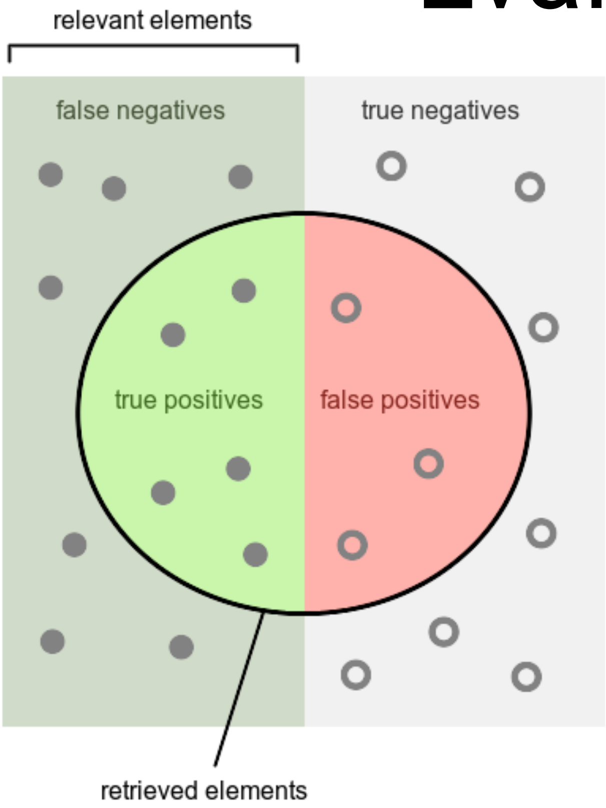
Evaluating Classifiers



	Positive	Negative
Positive	True Positive	False Negative
Negative	False Positive	True Negative

Training a Classifier

- Get training data
- Train some classifier $f : \mathcal{X} \to \mathbb{R}$, e.g. logistic regression

$$p(y \mid x) = \frac{1}{1 + \exp(-yf(x))}$$

- Clearly, larger values $f(x)$ mean that $y = 1$ is more likely, but the numbers need not be properly calibrated.
- Many diagnostics for a classifier …

ROC Curve

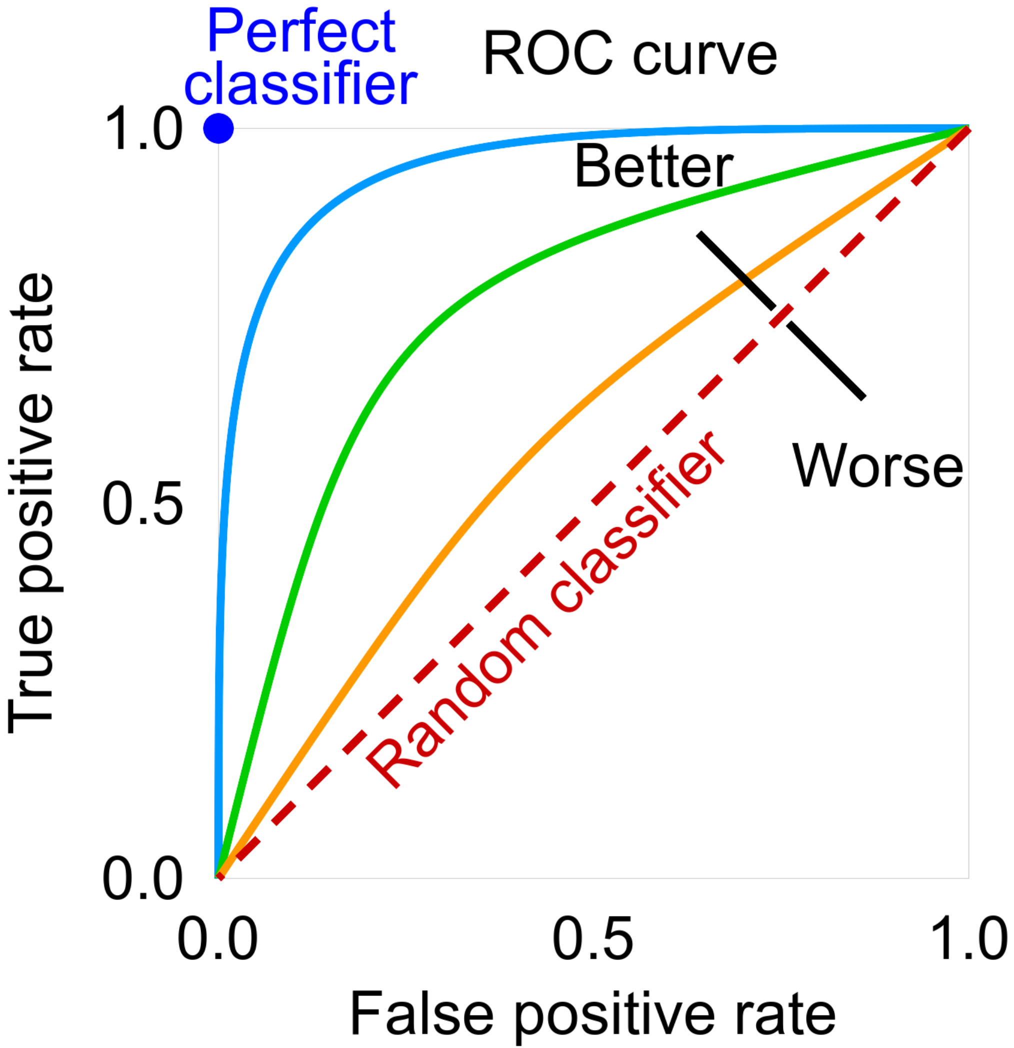
- True positive rate

$$\mathrm{TP}(z) = \frac{\Pr\{f(x) \geq z \text{ AND } y = 1\}}{p(y = 1)}$$

- False positive rate

$$\mathrm{FP}(z) = \frac{\Pr\{f(x) \geq z \text{ AND } y = -1\}}{p(y = -1)}$$

- A good classifier separates both classes well.



Precision Recall Curve

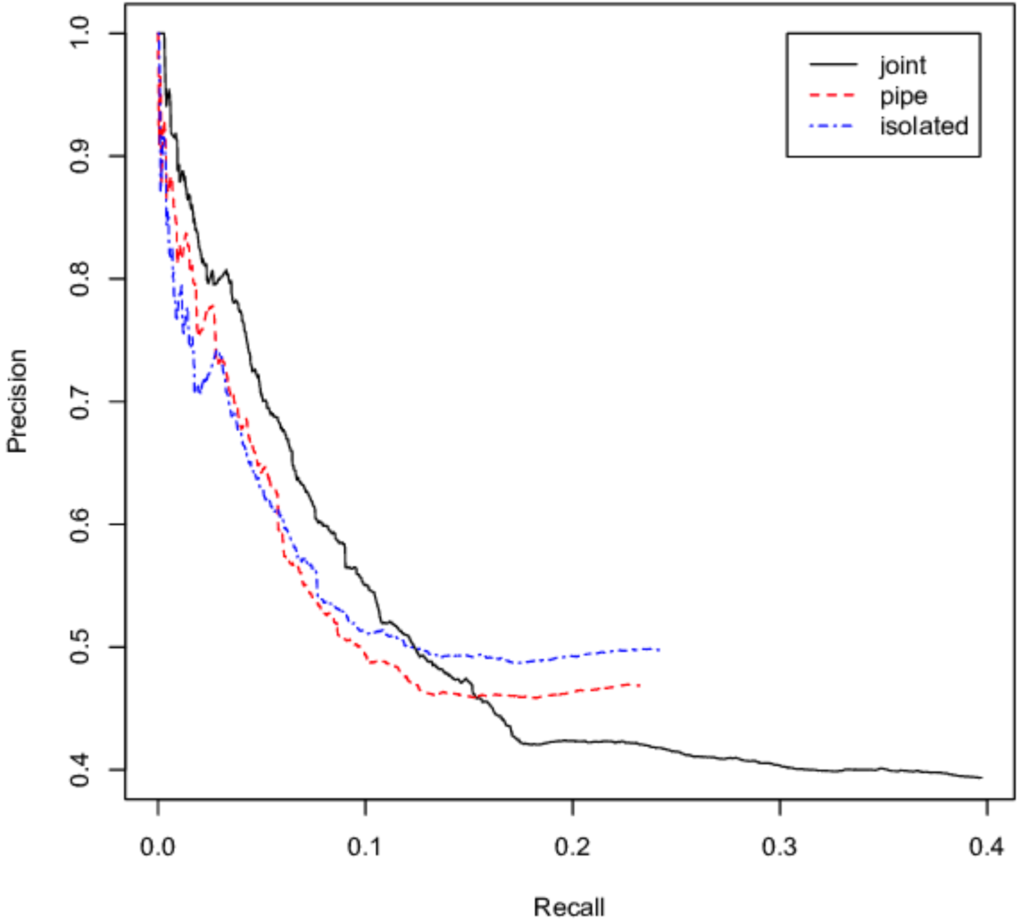
- Precision

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- Recall

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- A good has high precision even at high recall



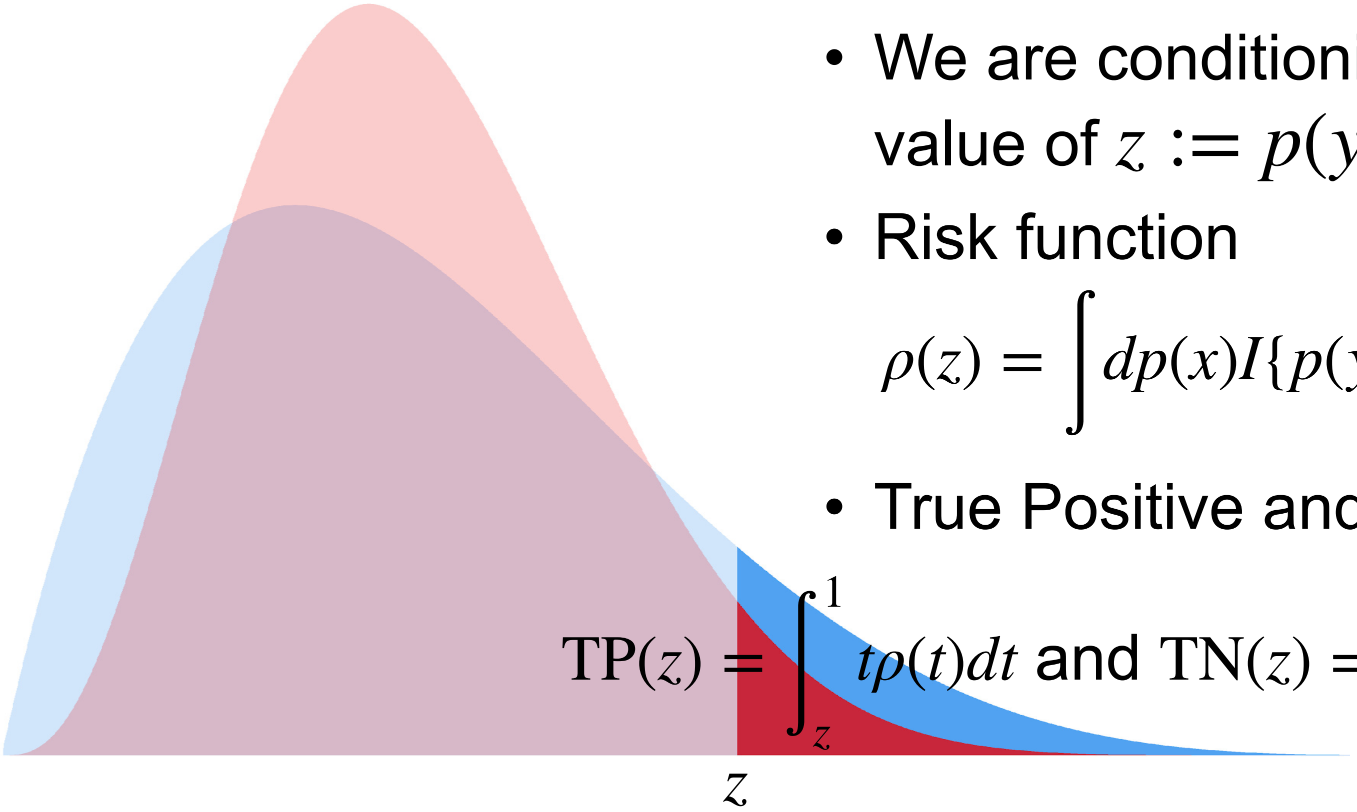
Risk Distribution

- We are conditioning on the value of $z := p(y = 1 \mid x)$
- Risk function

$$\rho(z) = \int dp(x) I\{p(y = 1 \mid x) = z\}$$

- True Positive and Negative

$$\mathrm{TP}(z) = \int_z^1 t\rho(t)dt \text{ and } \mathrm{TN}(z) = \int_0^z (1 - t)\rho(t)dt$$



Even more Metrics

		Actual Class y		
		Positive	Negative	
$h_\theta(x)$ **Test outcome**	Test outcome positive	**True positive** (TP)	**False positive** (FP, Type I error)	Precision = $\frac{\#\mathrm{TP}}{\#\mathrm{TP} + \#\mathrm{FP}}$
	Test outcome negative	**False negative** (FN, Type II error)	**True negative** (TN)	Negative predictive value = $\frac{\#\mathrm{TN}}{\#\mathrm{FN} + \#\mathrm{TN}}$
		Sensitivity = $\frac{\#\mathrm{TP}}{\#\mathrm{TP} + \#\mathrm{FN}}$	Specificity = $\frac{\#\mathrm{TN}}{\#\mathrm{FP} + \#\mathrm{TN}}$	Accuracy = $\frac{\#\mathrm{TP} + \#\mathrm{TN}}{\#\mathrm{TOTAL}}$

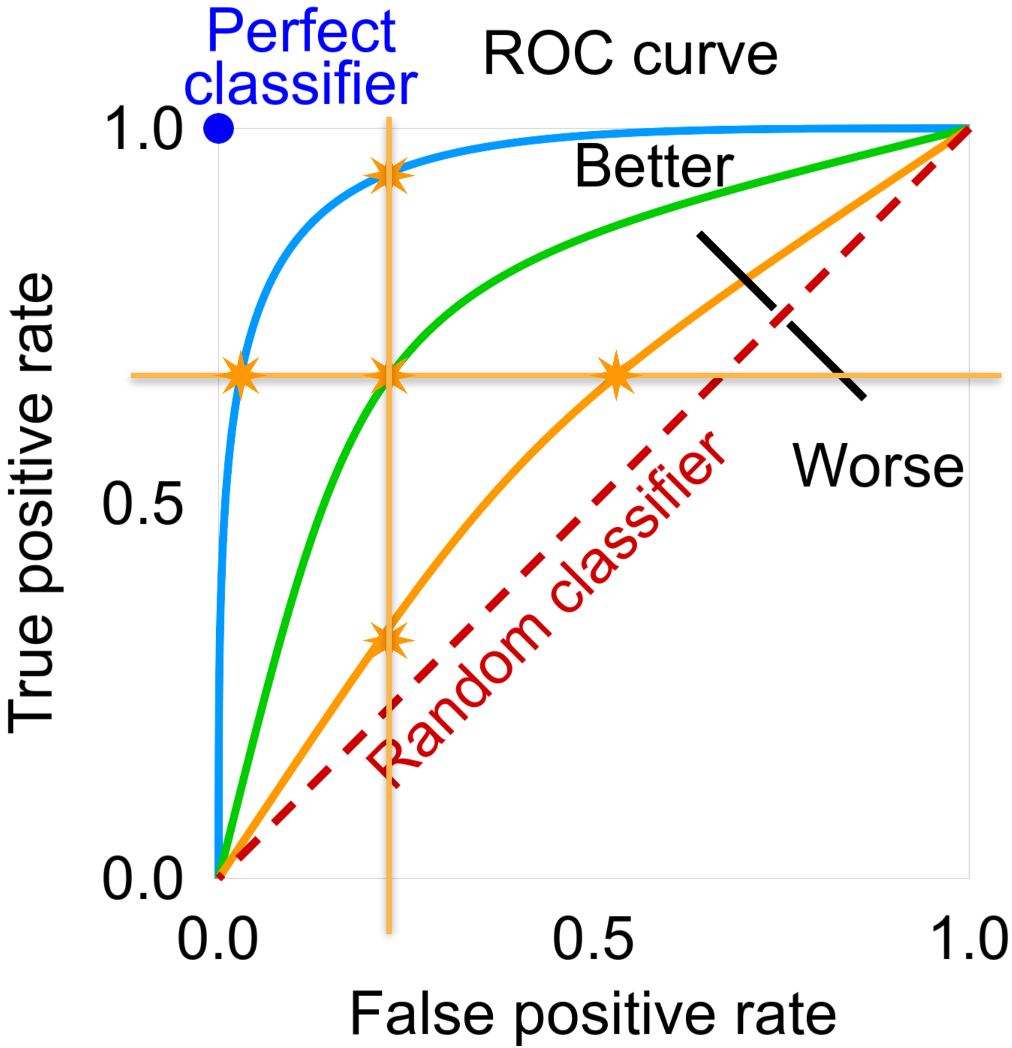
What does Fairness (not) Mean?

- Matching quality scores between diffe
(African American, Asian, Latino, White,
  - E.g. same false positive rate for arrests
  - E.g. same true positive rate for arrests
  - E.g. same precision for arrests
  - E.g. same risk for arrests (all subjects with more t
    50% risk)
  - E.g. same false negative rate for loan applications

Mission Impossible

Example - ROC Curve

- For given TP different curves cut at different FP rates.
- For given FP different curves cut at different TP rates.
- We can assume that the curves look different for different groups.

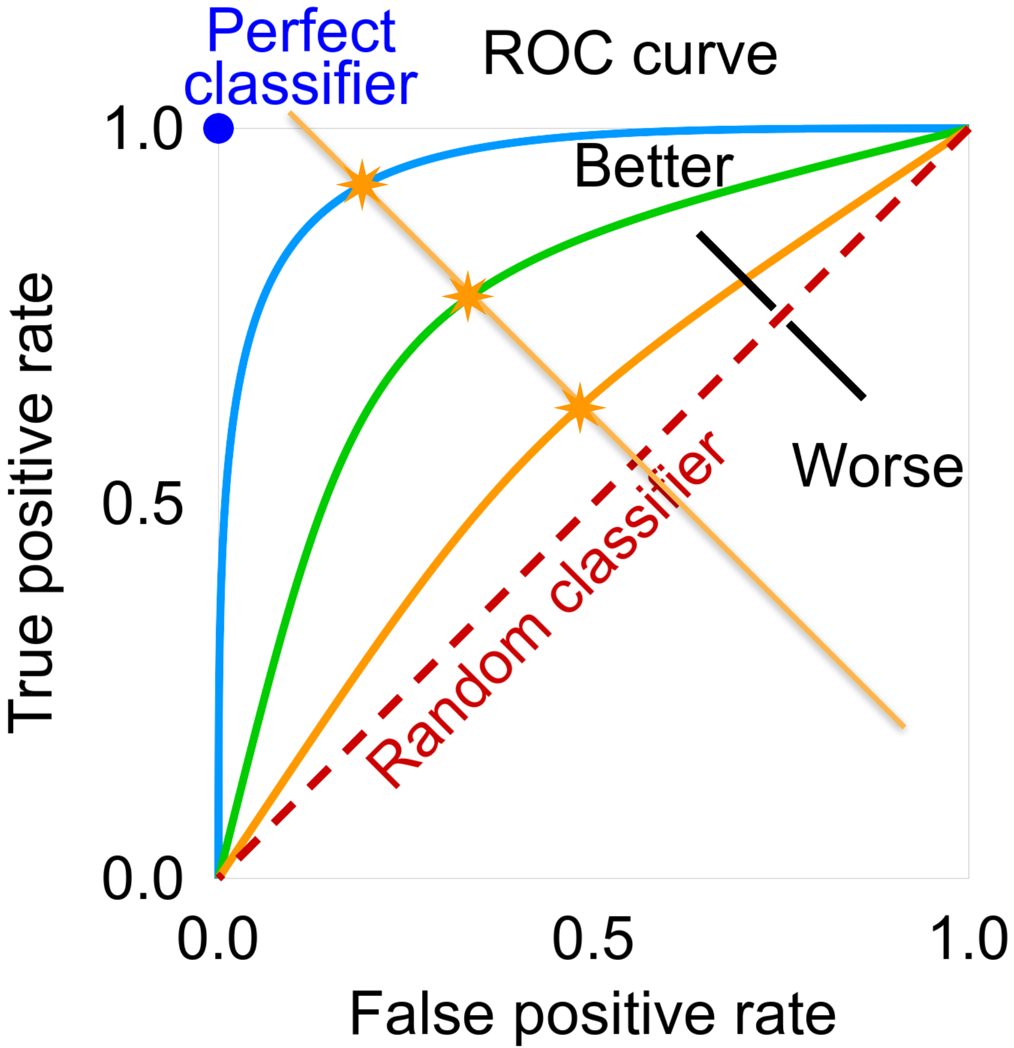


Example - ROC Curve

- We want a given number of positives (regardless of TP or FP) per group.

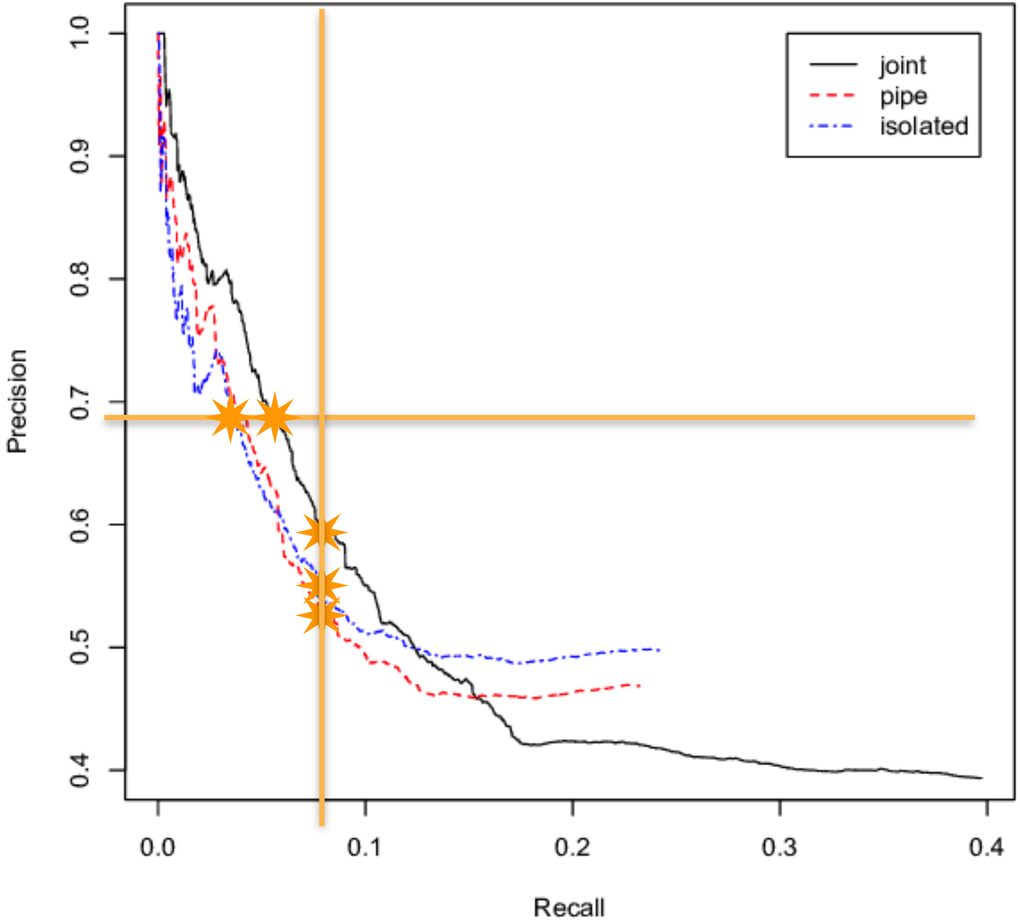
$$\mathrm{TP} + \mathrm{FP} = \mathrm{const}\,.$$

- Cutpoints can vary between groups!



Example - Precision Recall Curve

- For given Precision different curves cut at different Recall rates.
- For given Recall different curves cut at different Precision rates.
- We can assume that the curves look different for different groups.



~~CRITERIA:IMPOSSIBLE~~

Criteria Impossible

- Kleinberg, Mullainathan and Raghavan, 2016
  Inherent Trade-Offs in the Fair Determination of Risk Scores
- Chouldecova, 2017
  Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments

- Impossible to satisfy the following three requirements unless either perfect classification holds or the score distributions are the same across the groups.
  1. Classifier is well calibrated within groups
  2. Balance for positive class (score assigned to members of positive group has same average for all groups)
  3. Balance for negative class (ditto for negative group)

Well Known Problem

Hutchinson & Mitchell, 2018

50 Years of Test (Un)fairness: Lessons for Machine Learning

Different criteria are at odds with each other. Improving one can make the other worse. See diagram from Darlington, 1971.

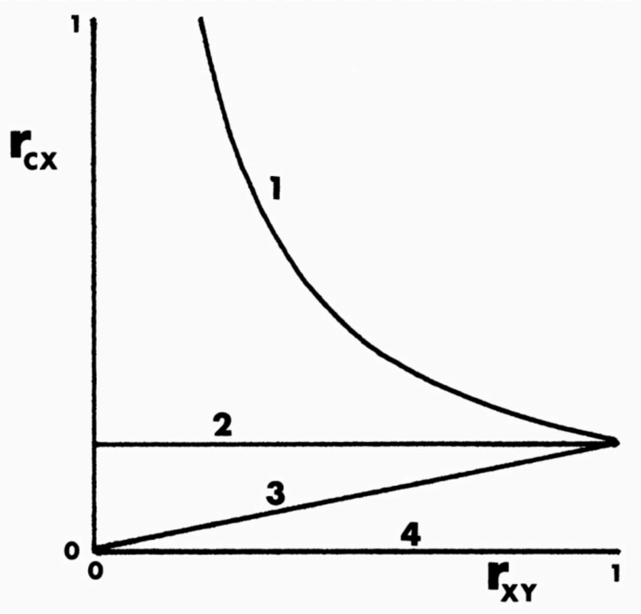
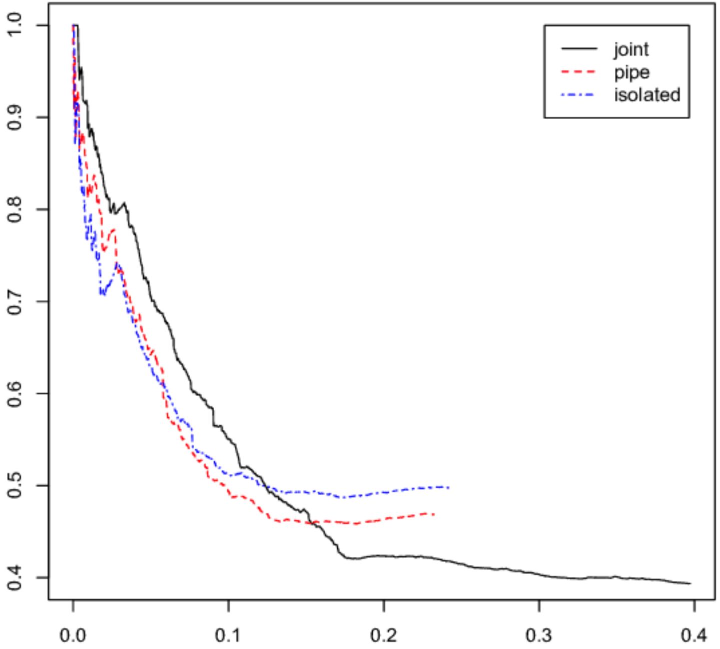
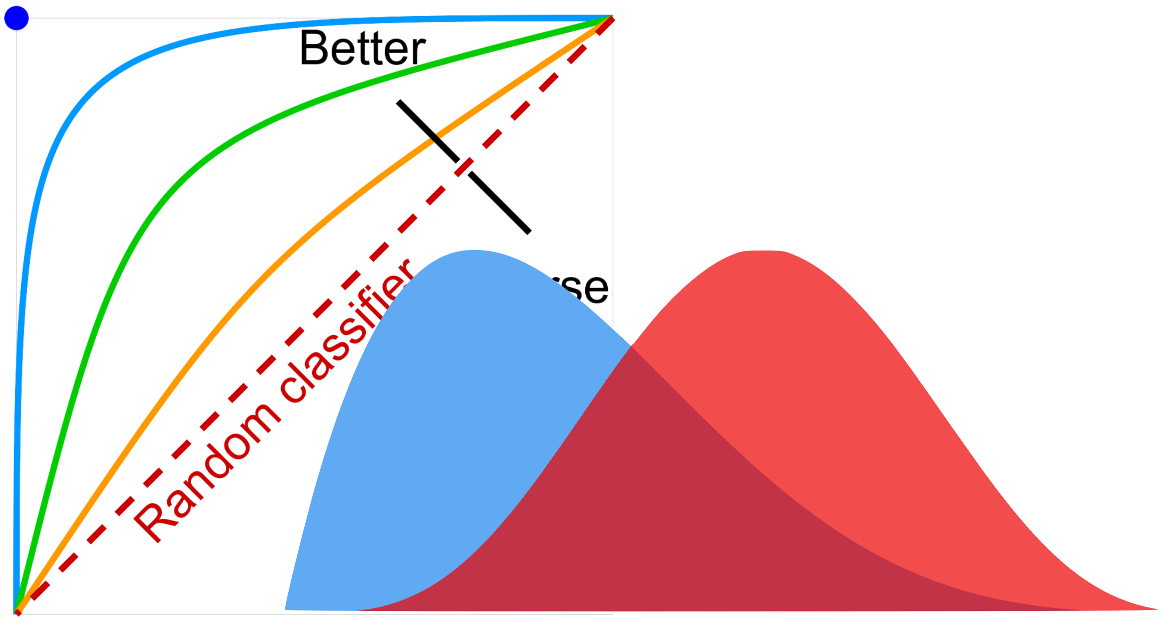


Figure 2: Darlington's original graph of fair values of the correlation between culture and test score (r_{CX} in Darlington's notation), plotted against the correlation between test score and ground truth (r_{XY}), according to his definitions (1–4). (The correlation between the demographic and target variables is assumed here to be fixed at 0.2.)

Deeper Reasons

- Different distributions have different curves
- Cutting them leads to different outcomes in general





s329p

Impossibility Theorem

'Pokemon' Theorem

Denote by p and q distributions on $\mathcal{X} \times \{0,1\}$, e.g. for different protected attriubutes. Let $s_1, \ldots s_n$ be statistics of the distributions for which $s_i[p] = s_i[q]$. If $p \neq q$ there always exists another statistic s' for which $s'[p] \neq s'[q]$.

Interpretation

Regardless of how many fairness criteria we are able to satisfy for different groups, there's always a criterion that we fail.

Related result by Simiou, Corbett-Davies, Goel 2017 (Infra-marginality)

Impossibility Theorem

‘Pokemon’ Theorem

Denote by p and q distributions on $\mathscr{X} \times \{0,1\}$, e.g. for different protected attriubutes. Let $s_1, \ldots s_n$ be statistics of the distributions for which $s_i[p] = s_i[q]$. If $p \neq q$ there always exists another statistic s' for which $s'[p] \neq s'[q]$.

Proof

Recall Maximum Mean Discrepancy (MMD). Two distributions are the same only if all expectations are the same (from an infinite class of test functions). A finite number of statistics is not enough. There always has to be one where things don’t match.

THE CRITERION COLLECTION

Fairness Definition Zoo

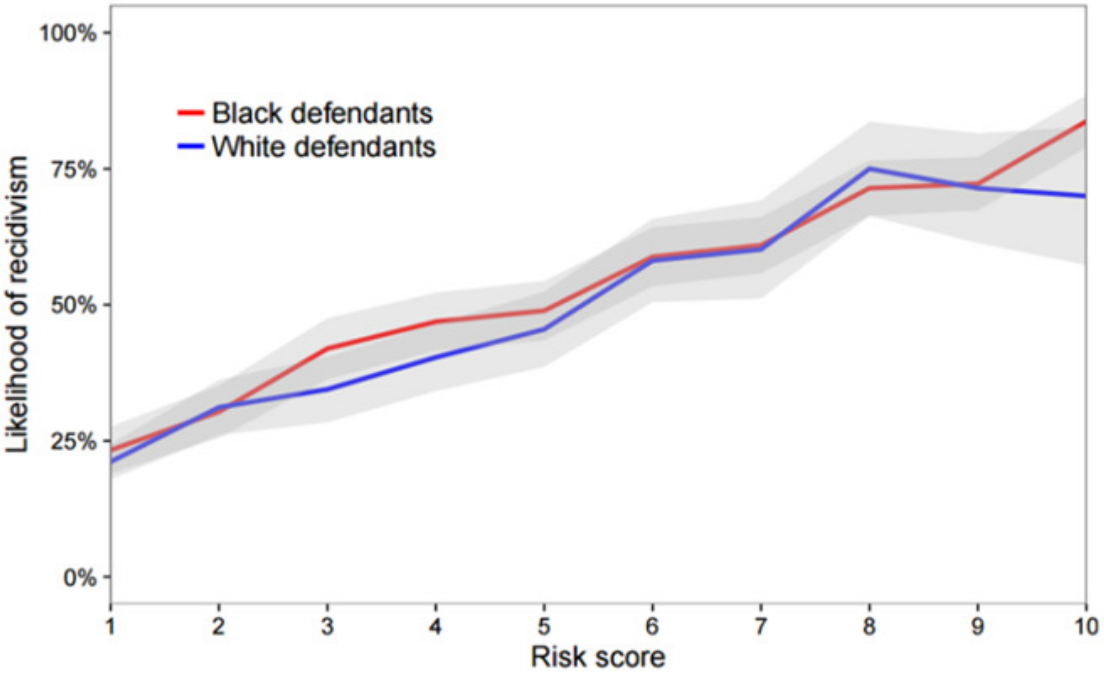
Verma & Rubin, 2018
Fairness Definitions Explained

- **Calibration**
  Outcome is independent of group membership given risk.
- **Classification parity**
  e.g., false positive rates are equal across groups.
- **Anti-classification**
  Protected characteristics are not used by the algorithm.
- **Conditional Demographic Disparity**
  Whether a group has smaller fraction of positive outcomes vs. fraction of negative outcomes relative to demographics.

	Definition	Paper	Citation #	Result
3.1.1	Group fairness or statistical parity	[12]	208	×
3.1.2	Conditional statistical parity	[11]	29	✓
3.2.1	Predictive parity	[10]	57	✓
3.2.2	False positive error rate balance	[10]	57	×
3.2.3	False negative error rate balance	[10]	57	✓
3.2.4	Equalised odds	[14]	106	×
3.2.5	Conditional use accuracy equality	[8]	18	×
3.2.6	Overall accuracy equality	[8]	18	✓
3.2.7	Treatment equality	[8]	18	×
3.3.1	Test-fairness or calibration	[10]	57	✓✗
3.3.2	Well calibration	[16]	81	✓✗
3.3.3	Balance for positive class	[16]	81	✓
3.3.4	Balance for negative class	[16]	81	×
4.1	Causal discrimination	[13]	1	×
4.2	Fairness through unawareness	[17]	14	✓
4.3	Fairness through awareness	[12]	208	×
5.1	Counterfactual fairness	[17]	14	–
5.2	No unresolved discrimination	[15]	14	–
5.3	No proxy discrimination	[15]	14	–
5.4	Fair inference	[19]	6	–

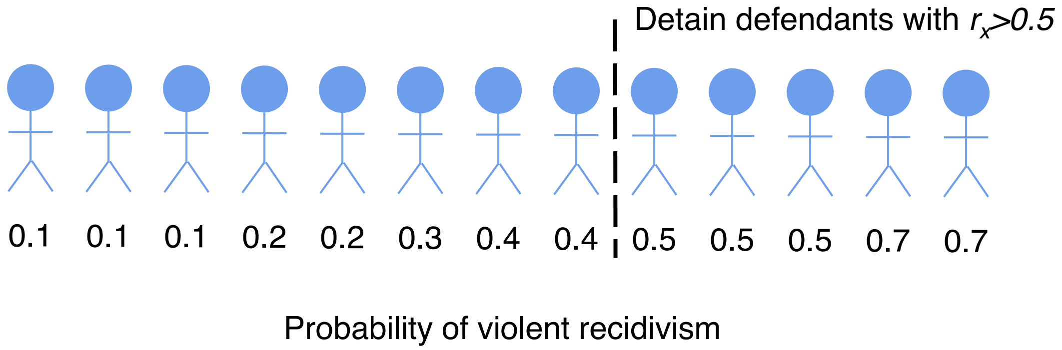
Calibration

Outcome is independent of group membership given risk.



From Corbett Davis & Goel - ICML 2019 tutorial

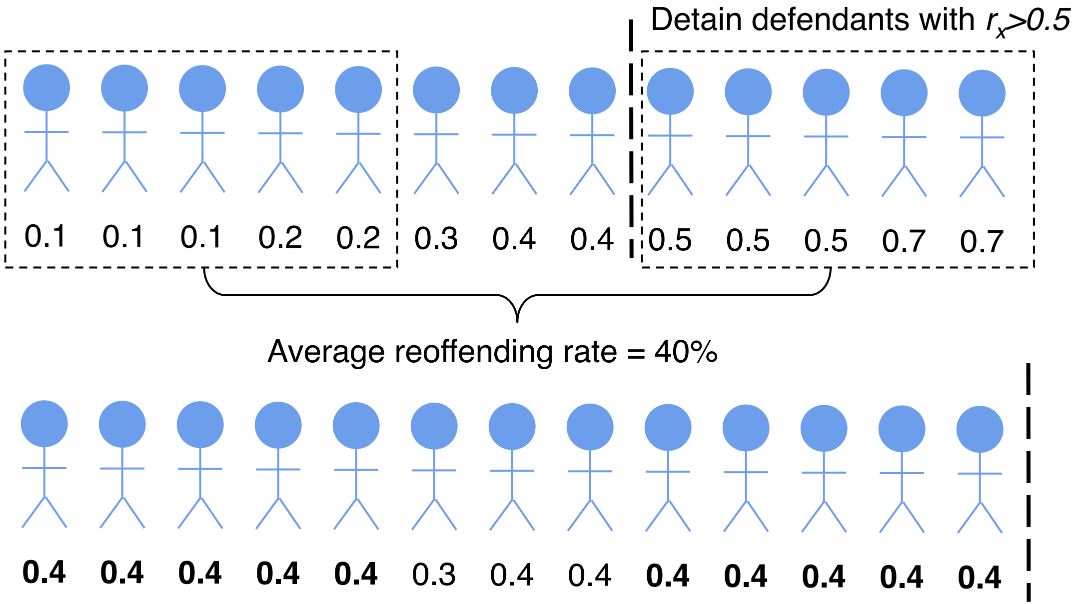
Hacking Calibration



Change decision function by changing features such that one group falls below the threshold

From Corbett-Davis & Goel - ICML 2019 tutorial

Hacking Calibration

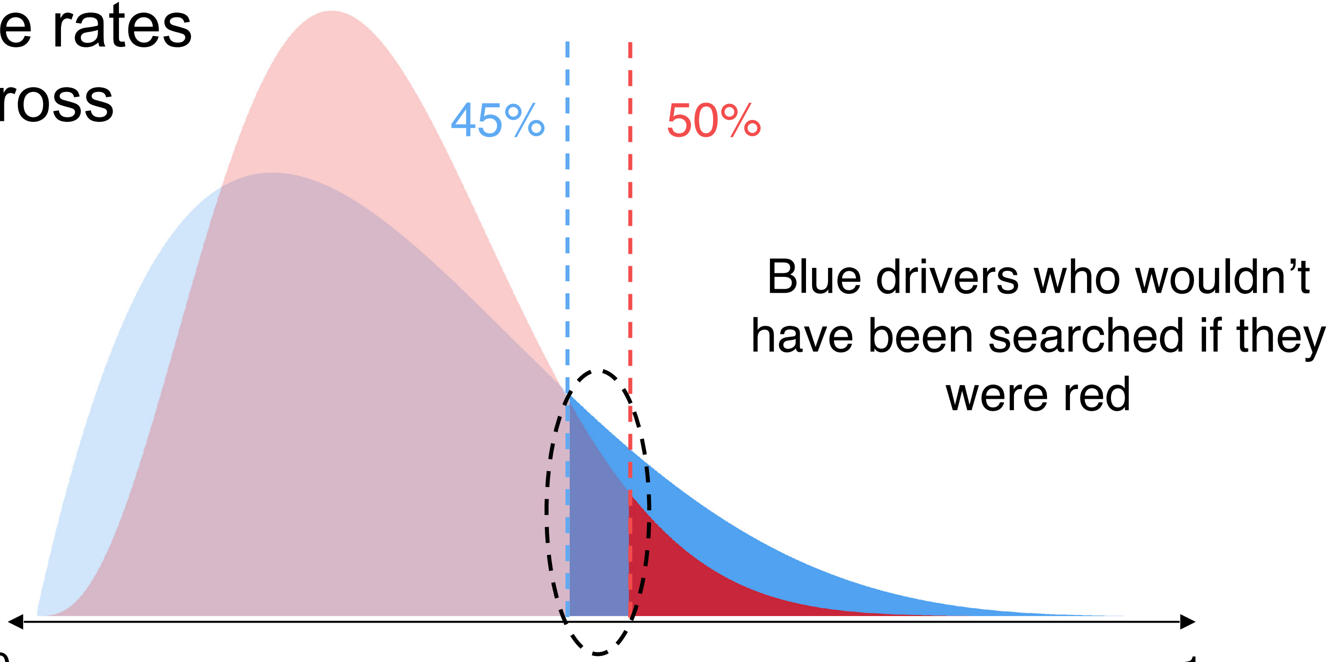


The scores are still calibrated, but no blue defendants are detained. In practice this could be achieved by choosing features that aren't predictive for the blue group.

From Corbett Davis & Goel - ICML 2019 tutorial

Classification Parity

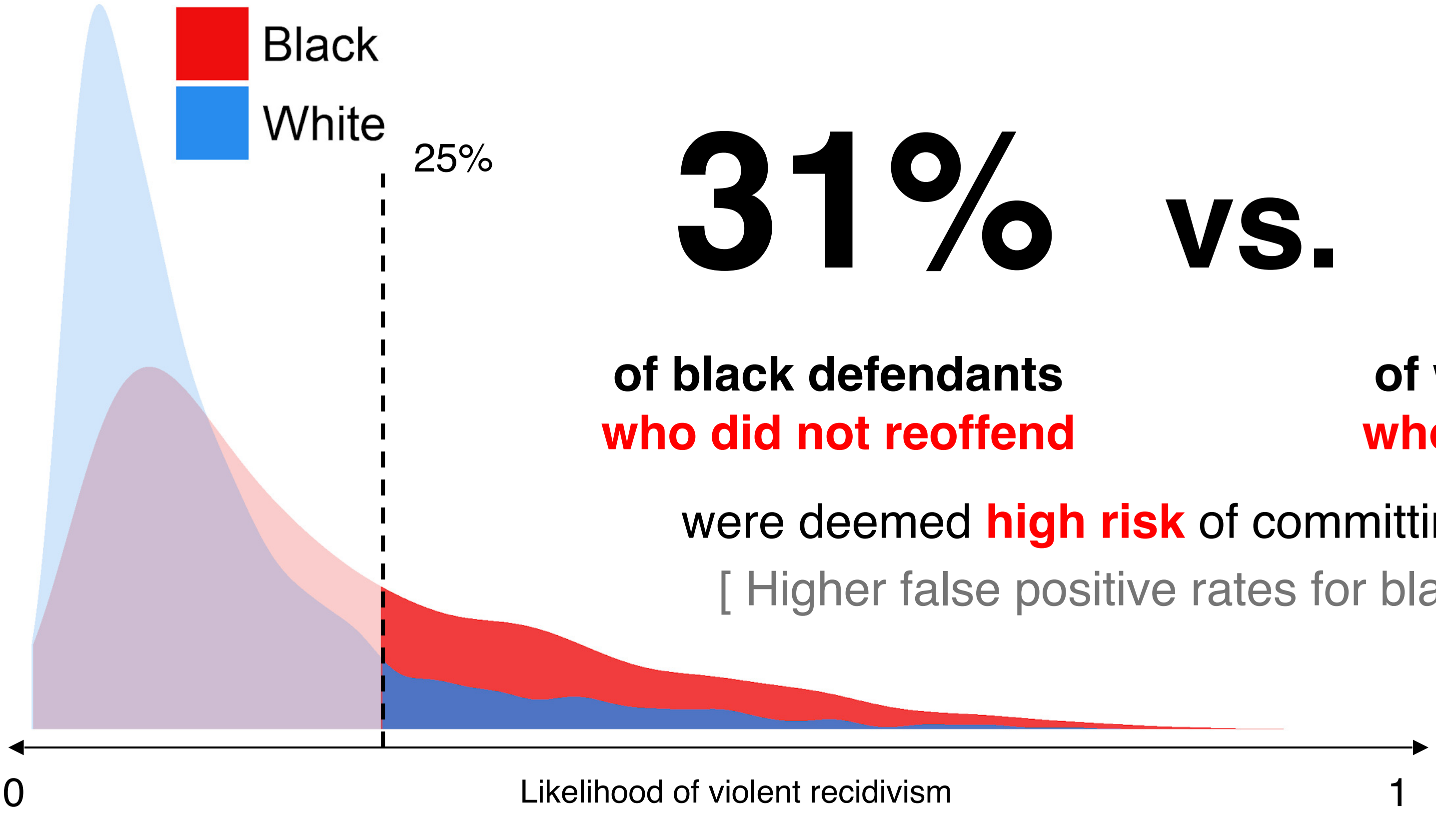
False positive rates are equal across groups.



Police are acting suboptimally in order to discriminate against blue drivers

From Corbett-Davis & Goel - ICML 2019 tutorial

Error rate disparities in Broward County



31% vs. 15%

of black defendants **who did not reoffend** | **of white defendants** **who did not reoffend**

were deemed **high risk** of committing a violent crime

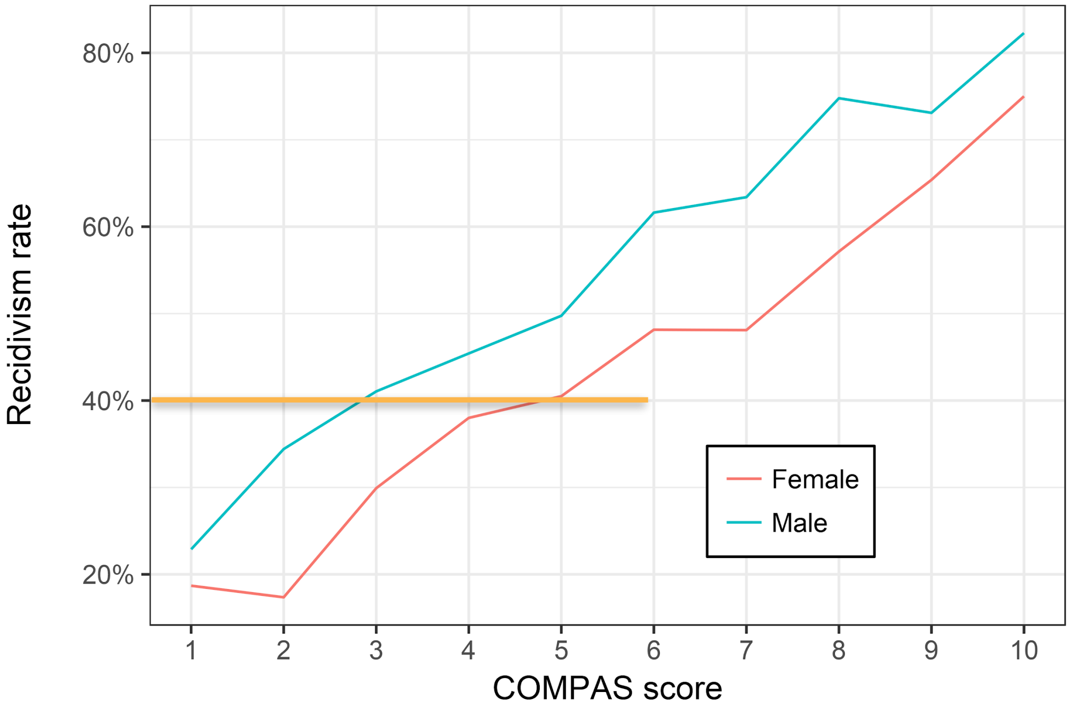
[Higher false positive rates for black defendants]

From Corbett Davis & Goel - ICML 2019 tutorial

Anti-Classification

Protected attributes are not used by the algorithm.

- Possible to discriminate based on secondary attributes (pony vs. motorbike, dreadlocks, Oxford shirt, etc.)
- Can make outcomes worse



Conditional Demographic Parity

Whether a group has smaller fraction of positive outcomes vs. fraction of negative outcomes relative to demographics.

$$\text{CDD} = \sum_c p(c)\left[\frac{\hat{p}(y=1\,|\,c)}{\hat{p}(y=1)} - \frac{\hat{p}(y=-1\,|\,c)}{\hat{p}(y=-1)}\right]$$

If necessary, need to partition by subcategories to avoid Simson's paradox (Wachter, Mittelstadt, Russell, 2020).

(Different risk thresholds between categories break it!)