Fairness in vulnerable attribute prediction on social media

Mariano Beiro & Kyriaki Kalimeri

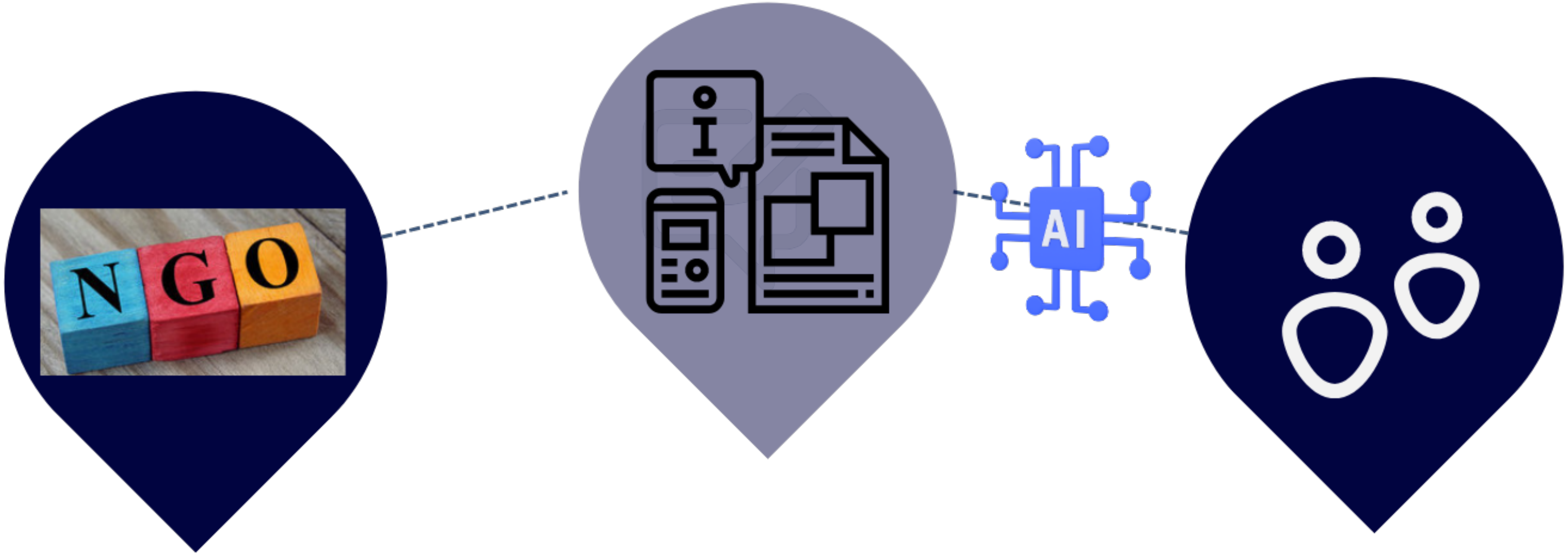
.UBA fiuba FACULTAD DE INGENIERÍA

CONICET

Machine Learning models are increasingly more employed to in the **humanitarian** sector to **inform decision-making**

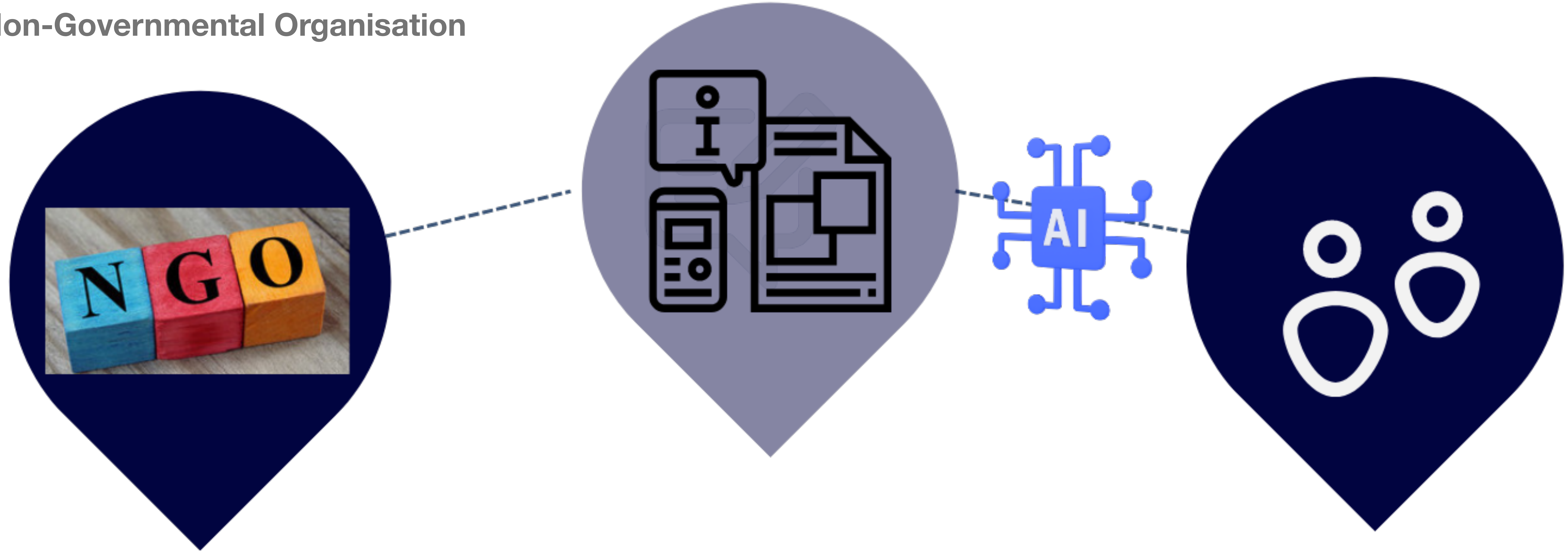
Ensure fairness, transparency, and accountability in model development and deployment

Our Case Study



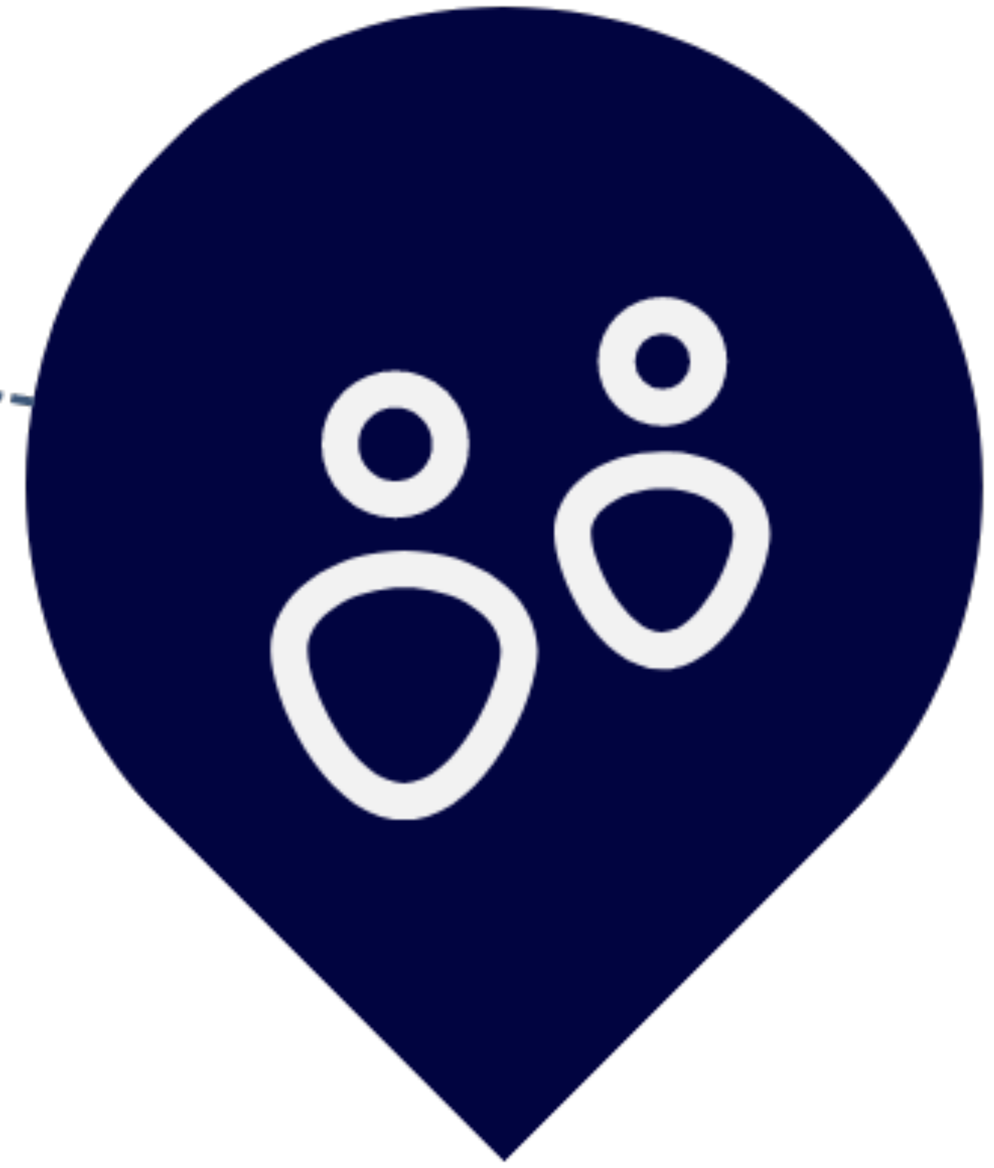
Our Case Study

A Non-Governmental Organisation



Our Case Study

A Non-Governmental Organisation



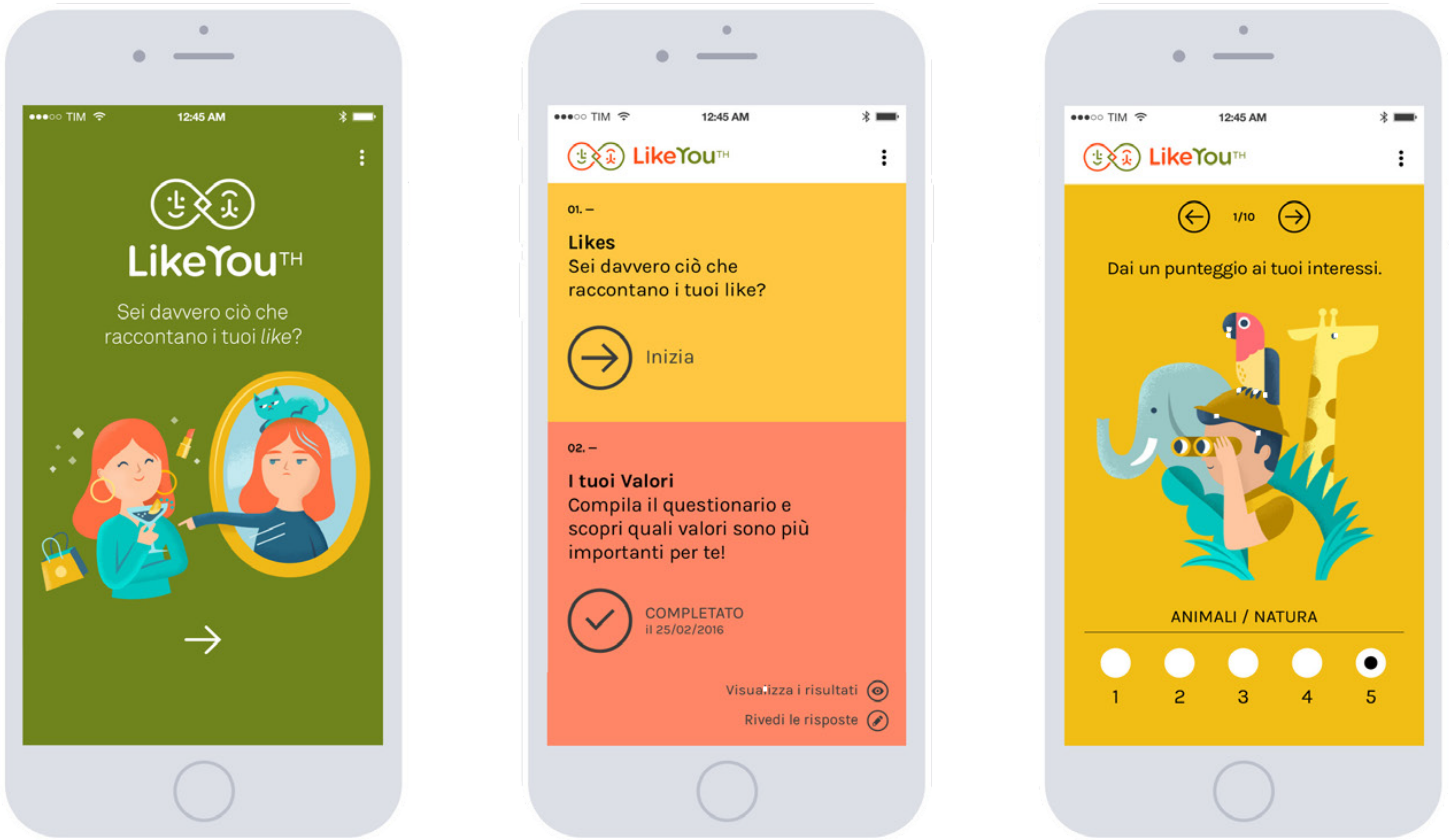
aimed at providing educational/training opportunities

Our Case Study

A Non-Governmental Organisation

aimed at providing educational/training opportunities

to young unemployed via AI

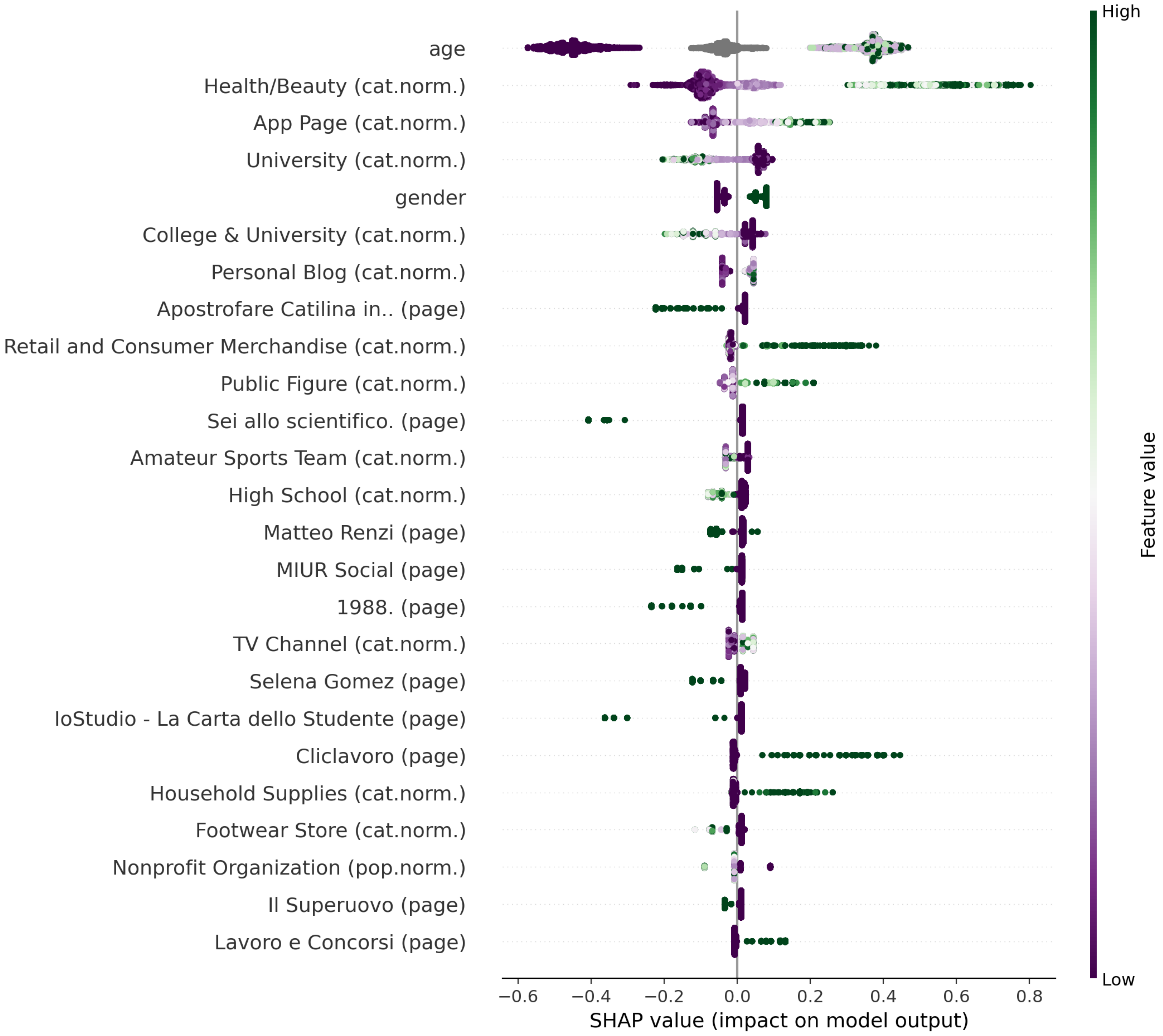

●●●○○ TIM
12:45 AM
LikeYou™
Sei davvero ciò che raccontano i tuoi *like*?
●●●○○ TIM
12:45 AM
LikeYou™
01. –
Likes
Sei davvero ciò che raccontano i tuoi like?
Inizia
02. –
I tuoi Valori
Compila il questionario e scopri quali valori sono più importanti per te!
COMPLETATO
il 25/02/2016
Visualizza i risultati
Rivedi le risposte
●●●○○ TIM
12:45 AM
LikeYou™
1/10
Dai un punteggio ai tuoi interessi.
ANIMALI / NATURA
1
2
3
4
5

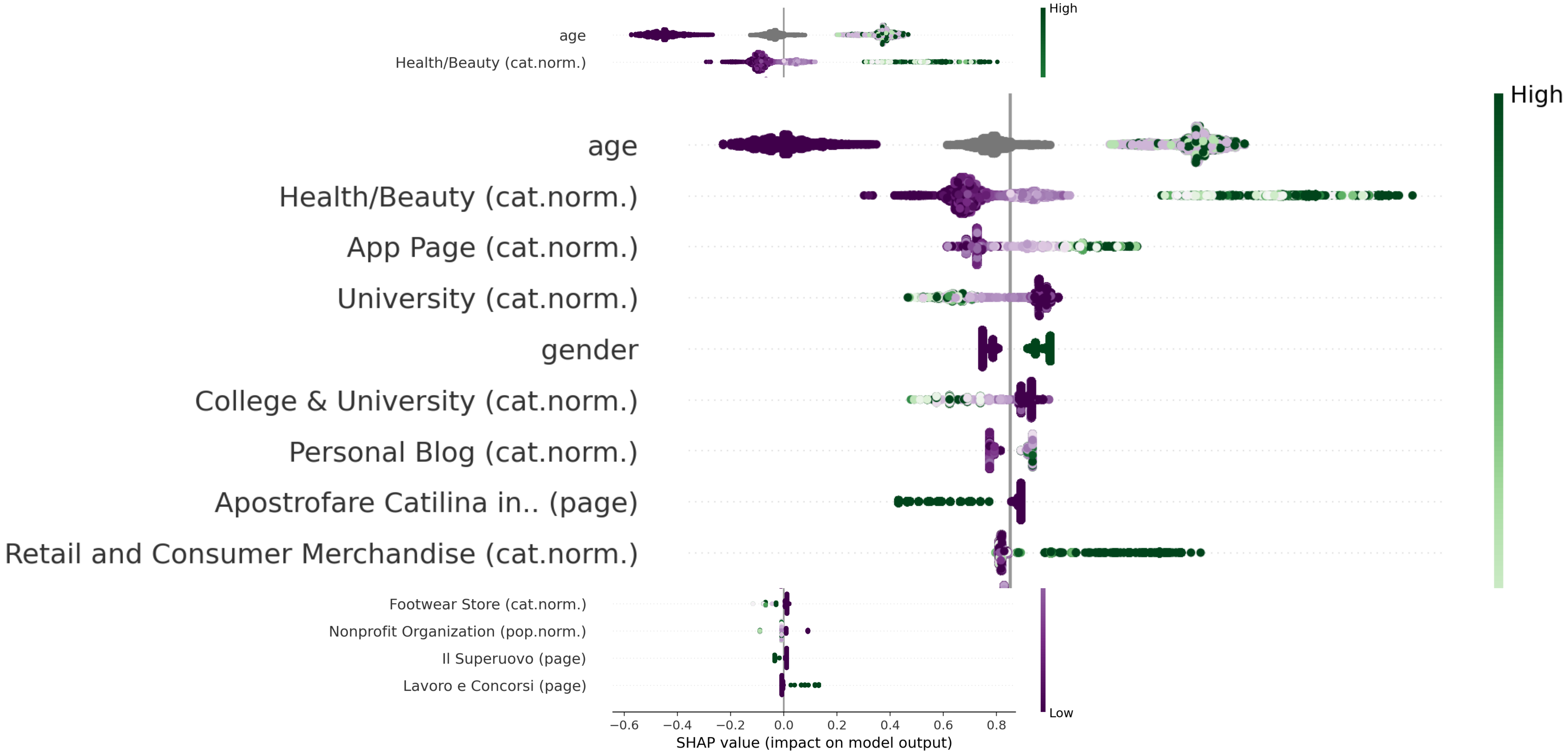
	Census	Dataset *n = 11,393*
Gender		
Female	51.1%	38.1%
Male	48.4%	61.8%
Age		
17–24	7.9%	43.1%
25–34	11.0%	31.2%
35–44	13.8%	13.6%
45–54	16.1%	7.1%
55–64	13.3%	4.5%
65+	24.5%	0.3%
Occupation		
Employed	77%	43.9%
Unemployed	8.7%	7.4%
Student	14.2%	48.5%

	Census	Dataset
		n = 11,393
Gender		
Female	51.1%	38.1%
Male	48.4%	61.8%
Age		
17–24	7.9%	43.1%
25–34	11.0%	31.2%
35–44	13.8%	13.6%
45–54	16.1%	7.1%
55–64	13.3%	4.5%
65+	24.5%	0.3%
Occupation		
Employed	77%	43.9%
Unemployed	8.7%	7.4%
Student	14.2%	48.5%

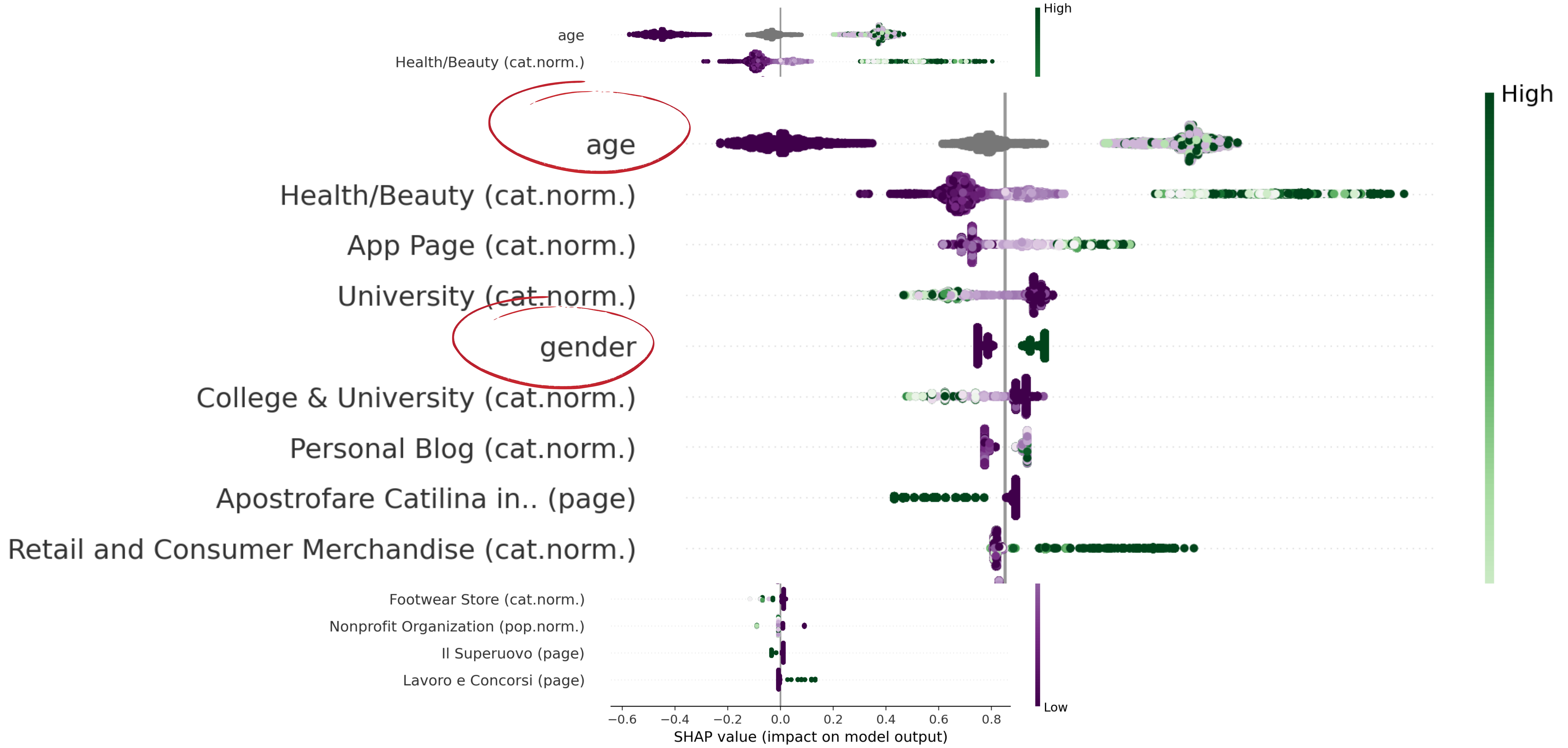
	Census	Dataset
		n = 11,393
Gender		
Female	51.1%	38.1%
Male	48.4%	61.8%
Age		
17–24	[illegible]	43.1%
25–34	[illegible]	31.2%
35–44	[illegible]	13.6%
45–54	[illegible]	7.1%
55–64	13.3%	4.5%
65+	24.5%	0.3%
Occupation		
Employed	77%	43.9%
Unemployed	8.7%	7.4%
Student	14.2%	48.5%

Unemployment rate per Gender:
Male: 5.5%
Female: 9%

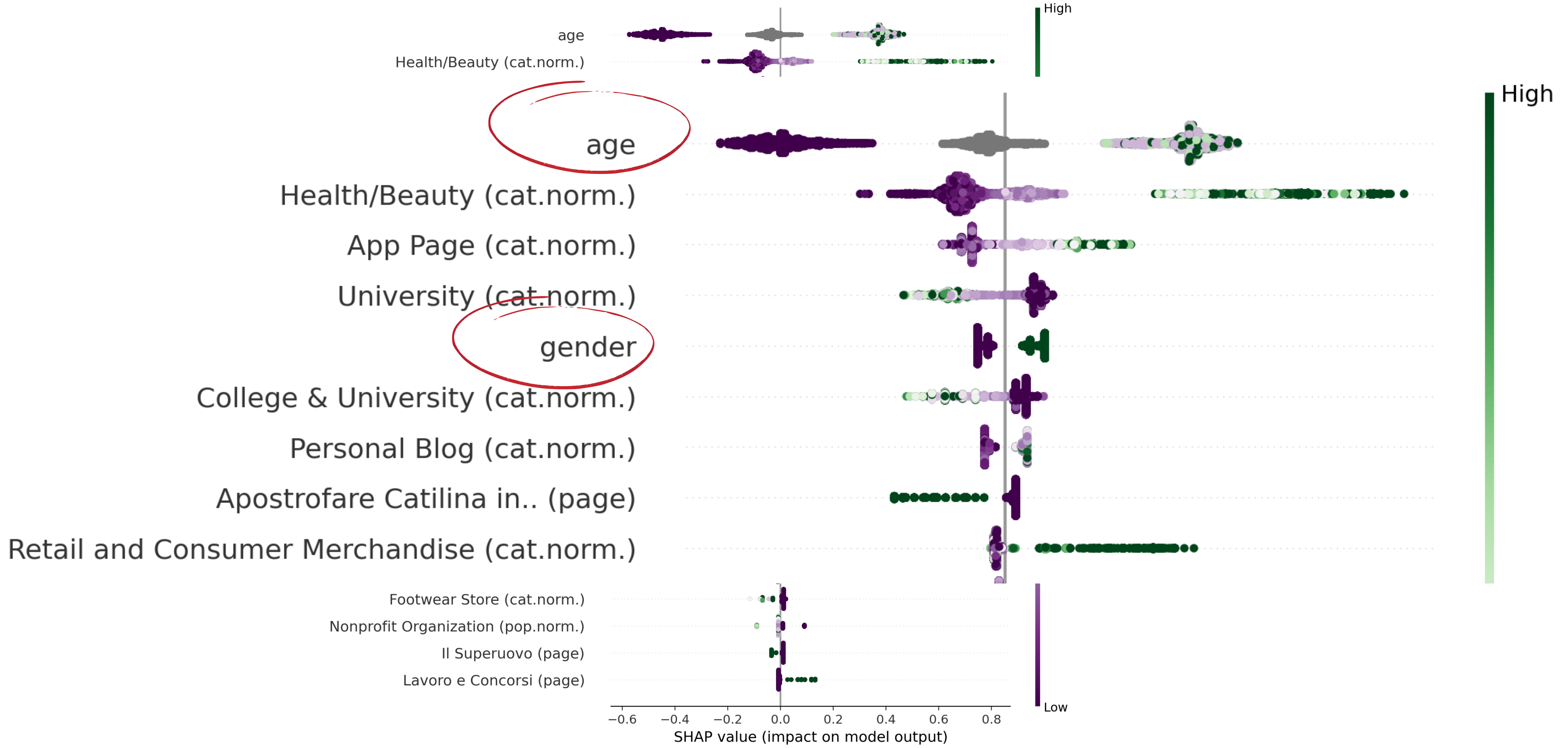

age
Health/Beauty (cat.norm.)
App Page (cat.norm.)
University (cat.norm.)
gender
College & University (cat.norm.)
Personal Blog (cat.norm.)
Apostrofare Catilina in.. (page)
Retail and Consumer Merchandise (cat.norm.)
Public Figure (cat.norm.)
Sei allo scientifico. (page)
Amateur Sports Team (cat.norm.)
High School (cat.norm.)
Matteo Renzi (page)
MIUR Social (page)
1988. (page)
TV Channel (cat.norm.)
Selena Gomez (page)
IoStudio - La Carta dello Studente (page)
Cliclavoro (page)
Household Supplies (cat.norm.)
Footwear Store (cat.norm.)
Nonprofit Organization (pop.norm.)
Il Superuovo (page)
Lavoro e Concorsi (page)
−0.6 −0.4 −0.2 0.0 0.2 0.4 0.6 0.8
SHAP value (impact on model output)
High
Low
Feature value



age
Health/Beauty (cat.norm.)
App Page (cat.norm.)
University (cat.norm.)
gender
College & University (cat.norm.)
Personal Blog (cat.norm.)
Apostrofare Catilina in.. (page)
Retail and Consumer Merchandise (cat.norm.)
Footwear Store (cat.norm.)
Nonprofit Organization (pop.norm.)
Il Superuovo (page)
Lavoro e Concorsi (page)
High
Low
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
SHAP value (impact on model output)


age
Health/Beauty (cat.norm.)
age
Health/Beauty (cat.norm.)
App Page (cat.norm.)
University (cat.norm.)
gender
College & University (cat.norm.)
Personal Blog (cat.norm.)
Apostrofare Catilina in.. (page)
Retail and Consumer Merchandise (cat.norm.)
Footwear Store (cat.norm.)
Nonprofit Organization (pop.norm.)
Il Superuovo (page)
Lavoro e Concorsi (page)
High
High
Low
−0.6 −0.4 −0.2 0.0 0.2 0.4 0.6 0.8
SHAP value (impact on model output)

Well we can predict unemployment with 74% AUROC. Cool!



Our Case Study

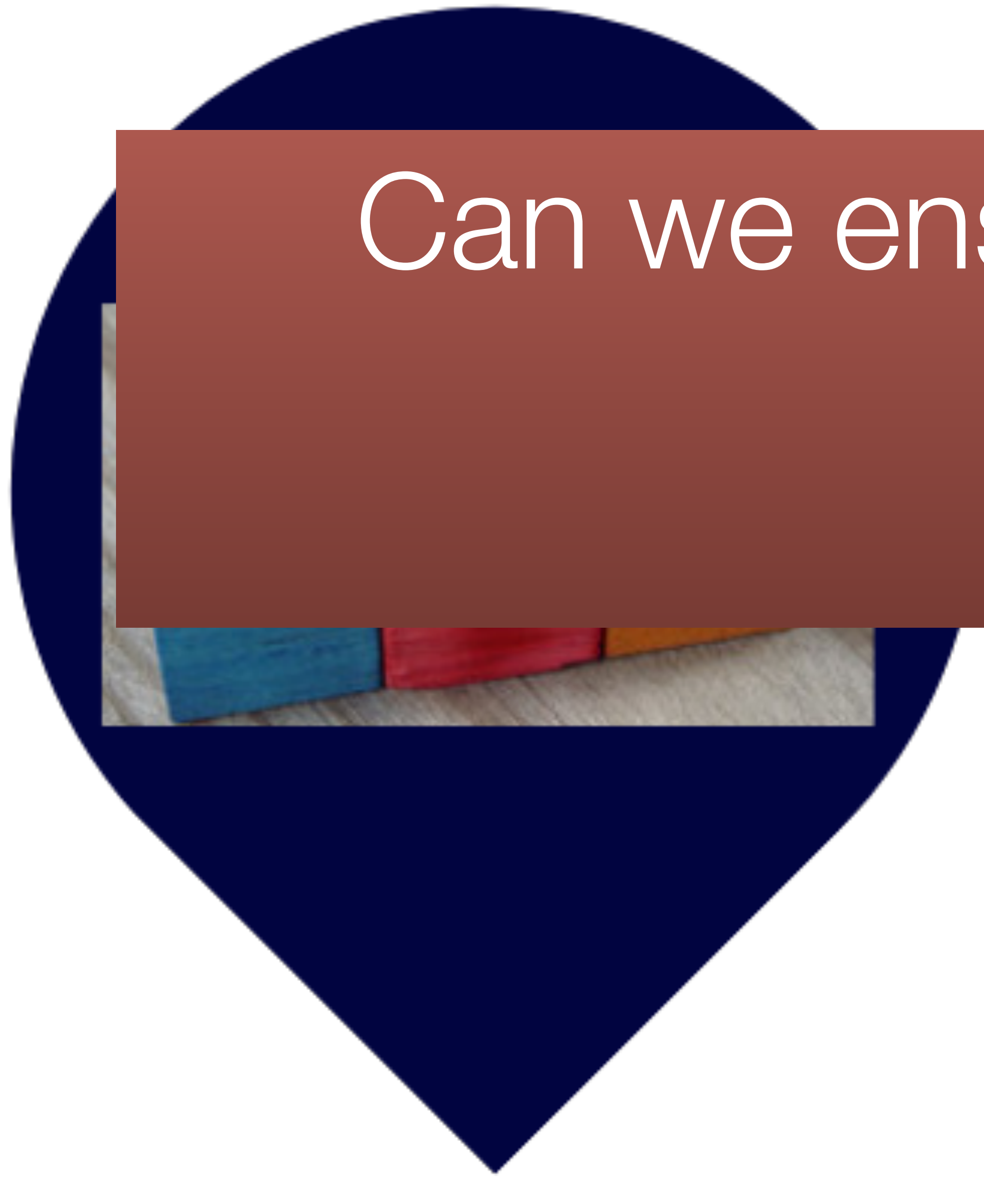
A Non-Governmental Organisation

aimed at providing educational/training opportunities

to young unemployed via AI

Our Case Study

A Non-Governmental Organisation

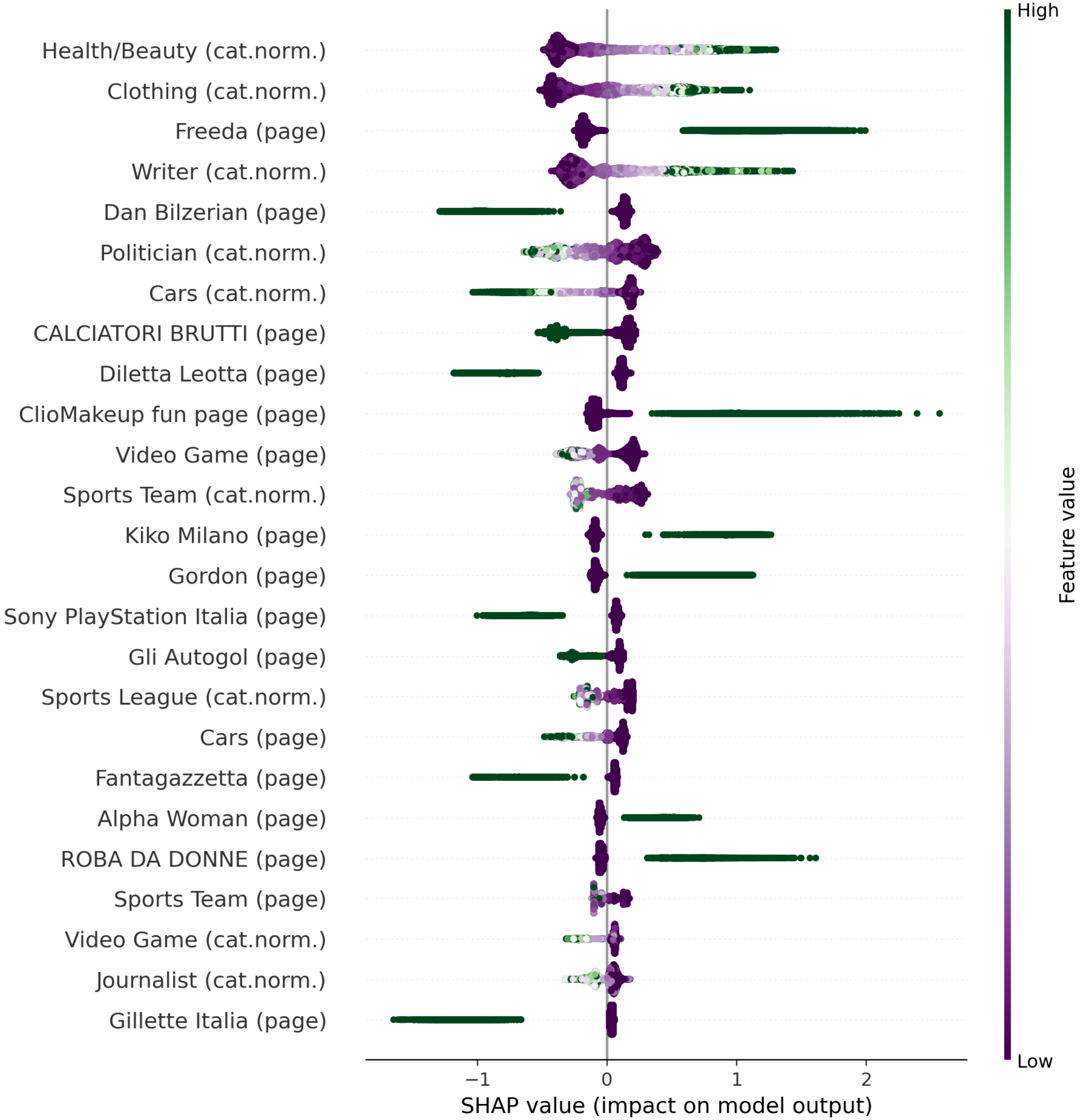


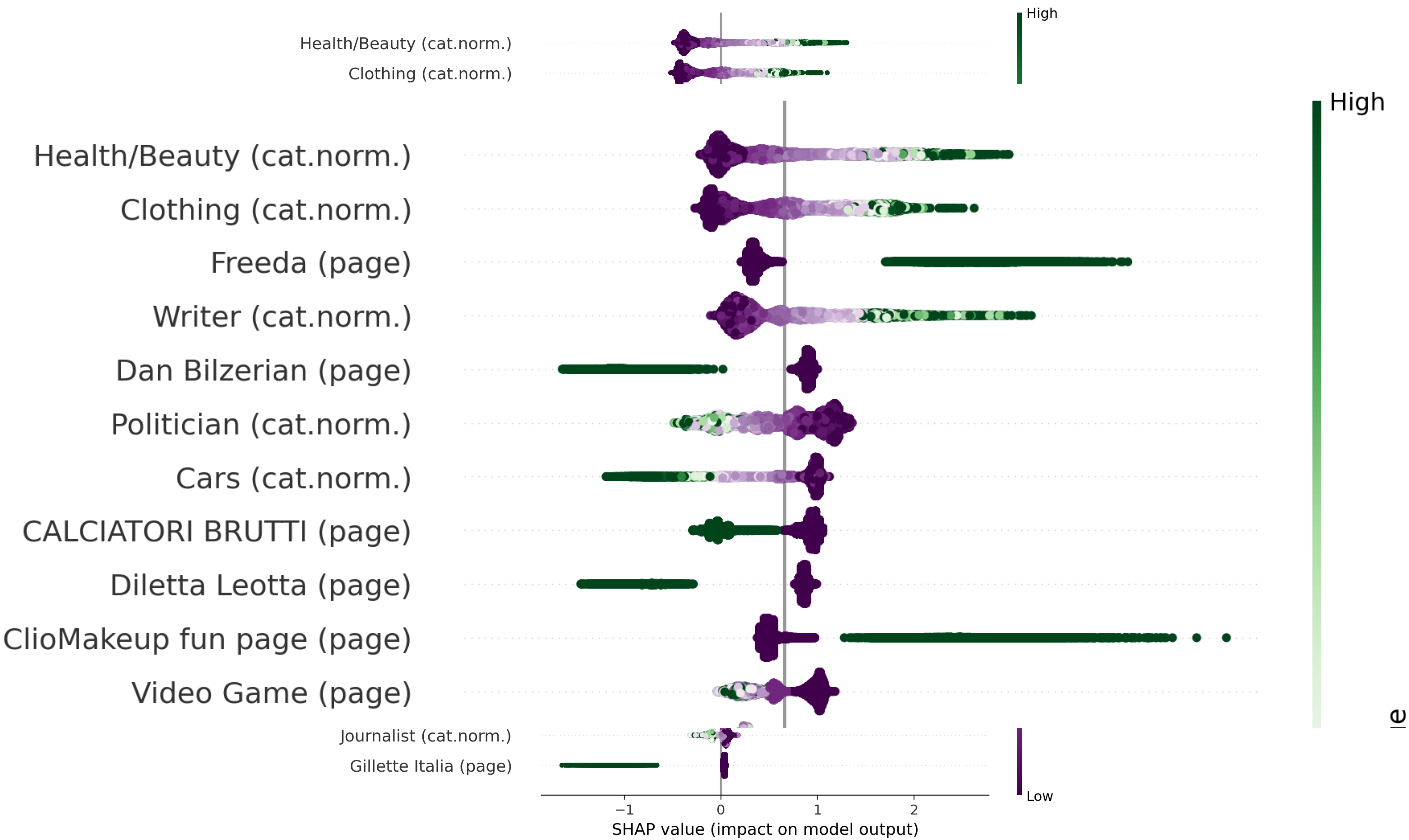
aimed at providing educational/training opportunities

to young unemployed via AI

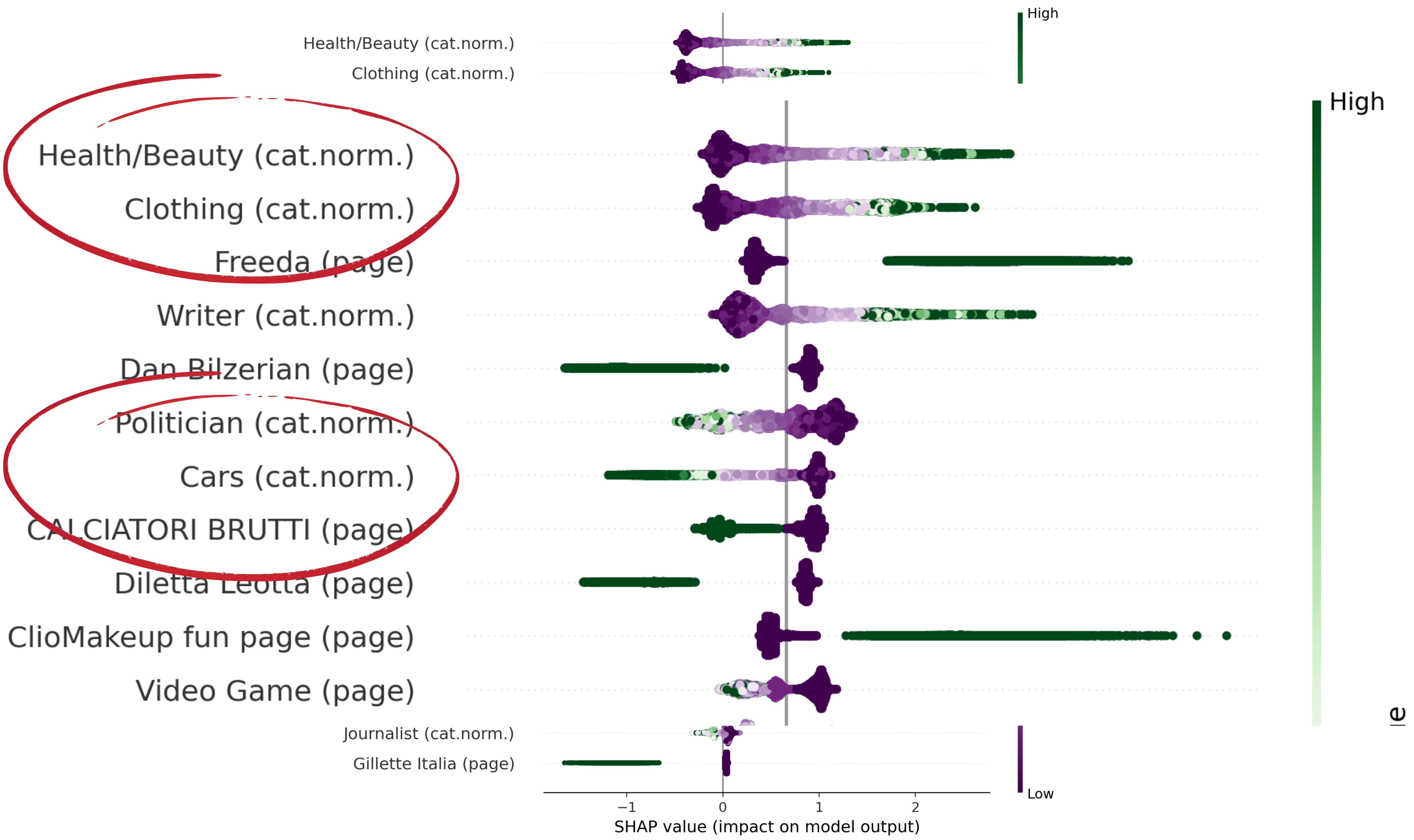
Can we ensure that the machine learning model is fair?

Can we avoid discrimination?


Health/Beauty (cat.norm.)
Clothing (cat.norm.)
Freeda (page)
Writer (cat.norm.)
Dan Bilzerian (page)
Politician (cat.norm.)
Cars (cat.norm.)
CALCIATORI BRUTTI (page)
Diletta Leotta (page)
ClioMakeup fun page (page)
Video Game (page)
Sports Team (cat.norm.)
Kiko Milano (page)
Gordon (page)
Sony PlayStation Italia (page)
Gli Autogol (page)
Sports League (cat.norm.)
Cars (page)
Fantagazzetta (page)
Alpha Woman (page)
ROBA DA DONNE (page)
Sports Team (page)
Video Game (cat.norm.)
Journalist (cat.norm.)
Gillette Italia (page)
−1
0
1
2
SHAP value (impact on model output)
High
Low
Feature value

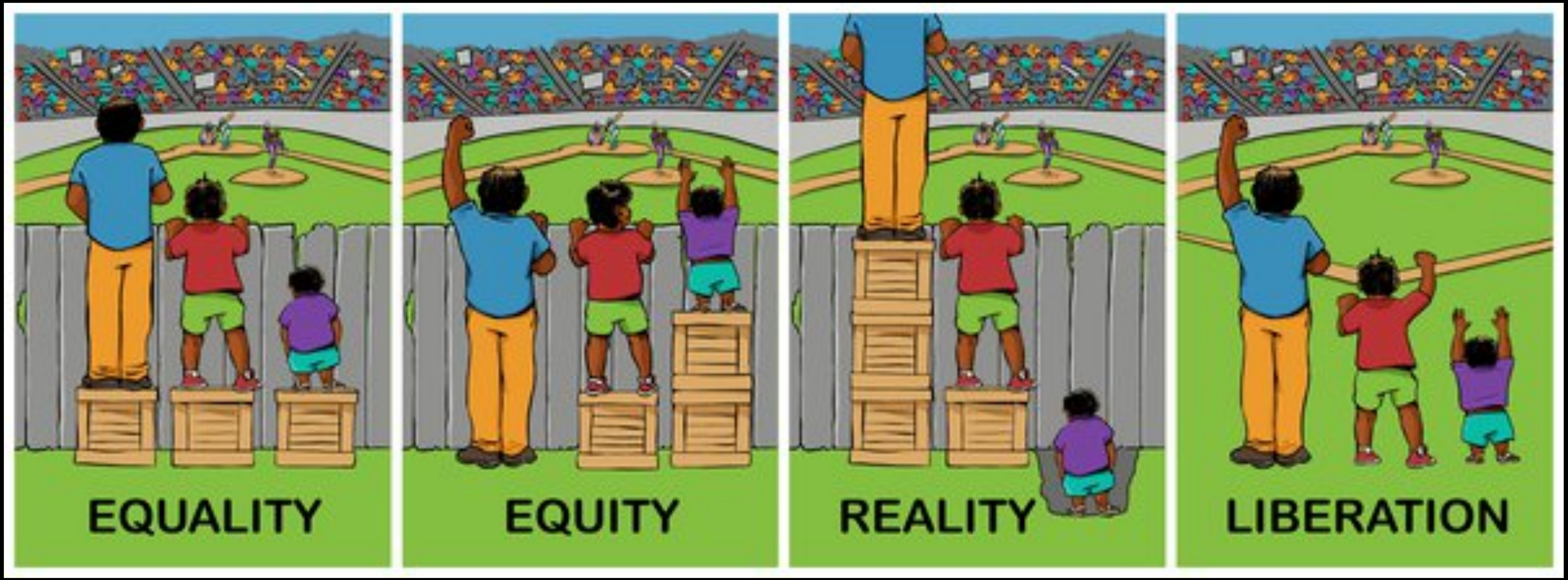


Health/Beauty (cat.norm.)
Clothing (cat.norm.)
Health/Beauty (cat.norm.)
Clothing (cat.norm.)
Freeda (page)
Writer (cat.norm.)
Dan Bilzerian (page)
Politician (cat.norm.)
Cars (cat.norm.)
CALCIATORI BRUTTI (page)
Diletta Leotta (page)
ClioMakeup fun page (page)
Video Game (page)
Journalist (cat.norm.)
Gillette Italia (page)
High
Low
High
−1
0
1
2
SHAP value (impact on model output)

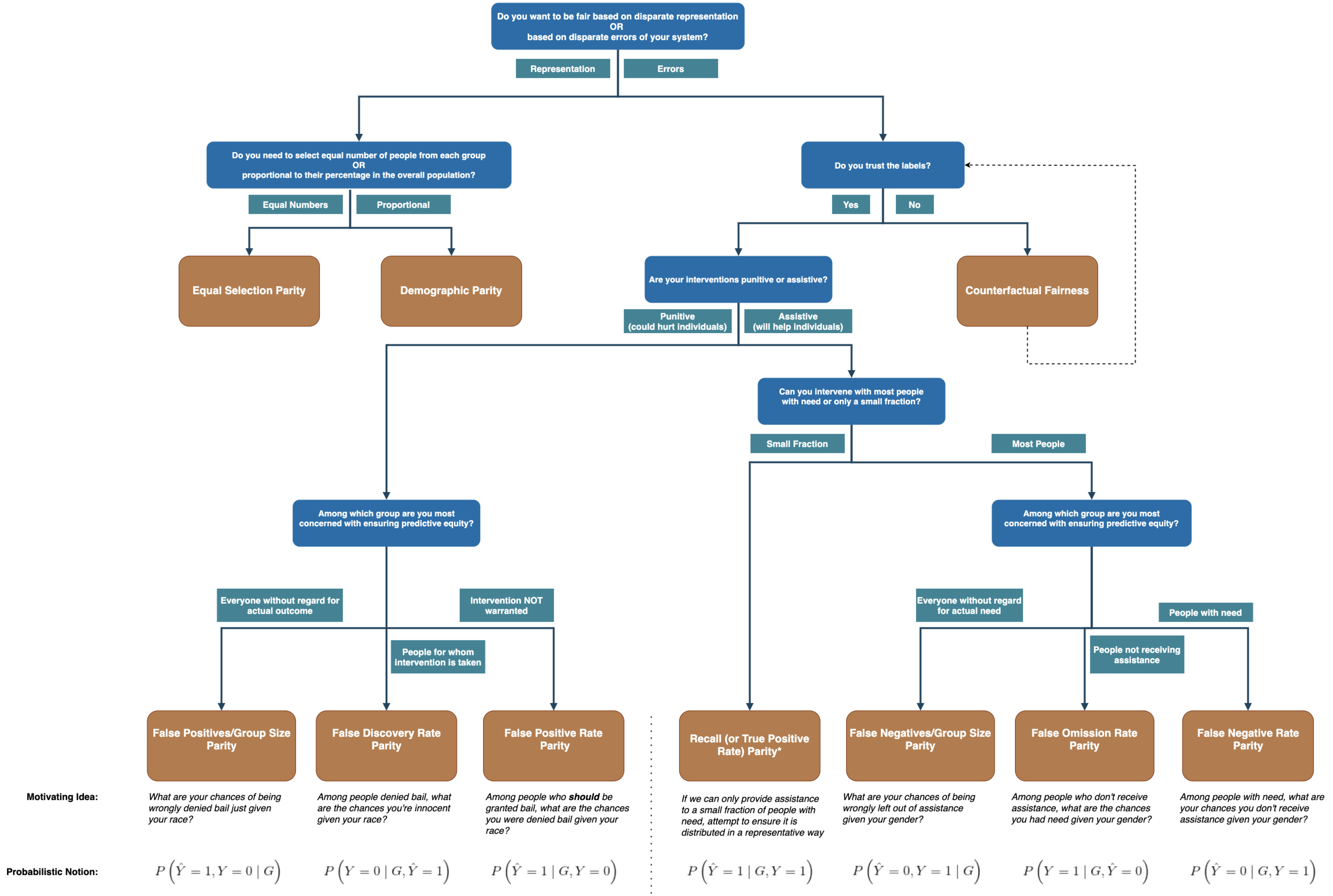

Health/Beauty (cat.norm.)
Clothing (cat.norm.)
Freeda (page)
Writer (cat.norm.)
Dan Bilzerian (page)
Politician (cat.norm.)
Cars (cat.norm.)
CALCIATORI BRUTTI (page)
Diletta Leotta (page)
ClioMakeup fun page (page)
Video Game (page)
Journalist (cat.norm.)
Gillette Italia (page)
High
Low
−1
0
1
2
SHAP value (impact on model output)

With Fairness Through Unawareness we do not avoid discrimination

What is "Fairness"?

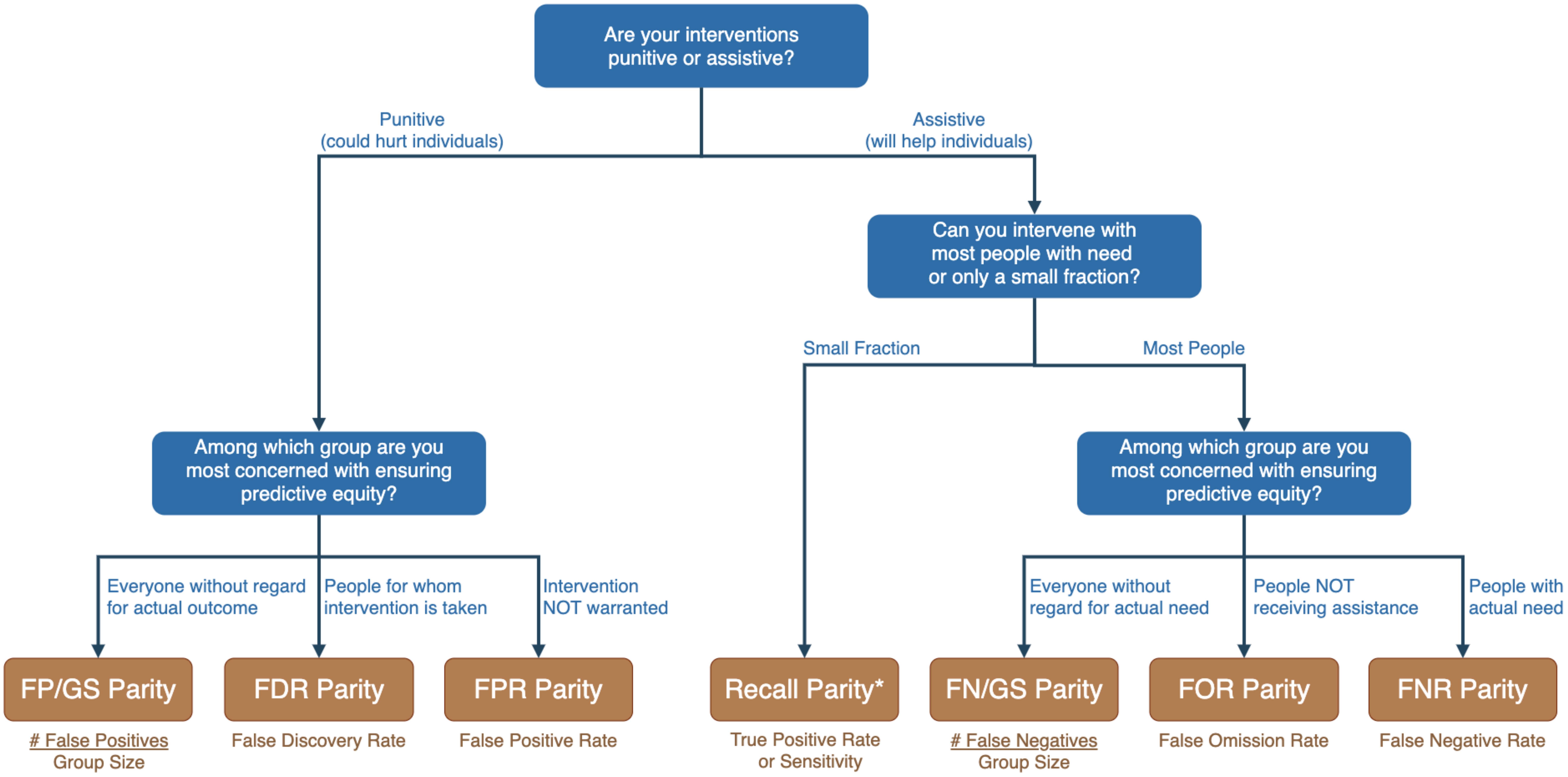


FAIRNESS TREE

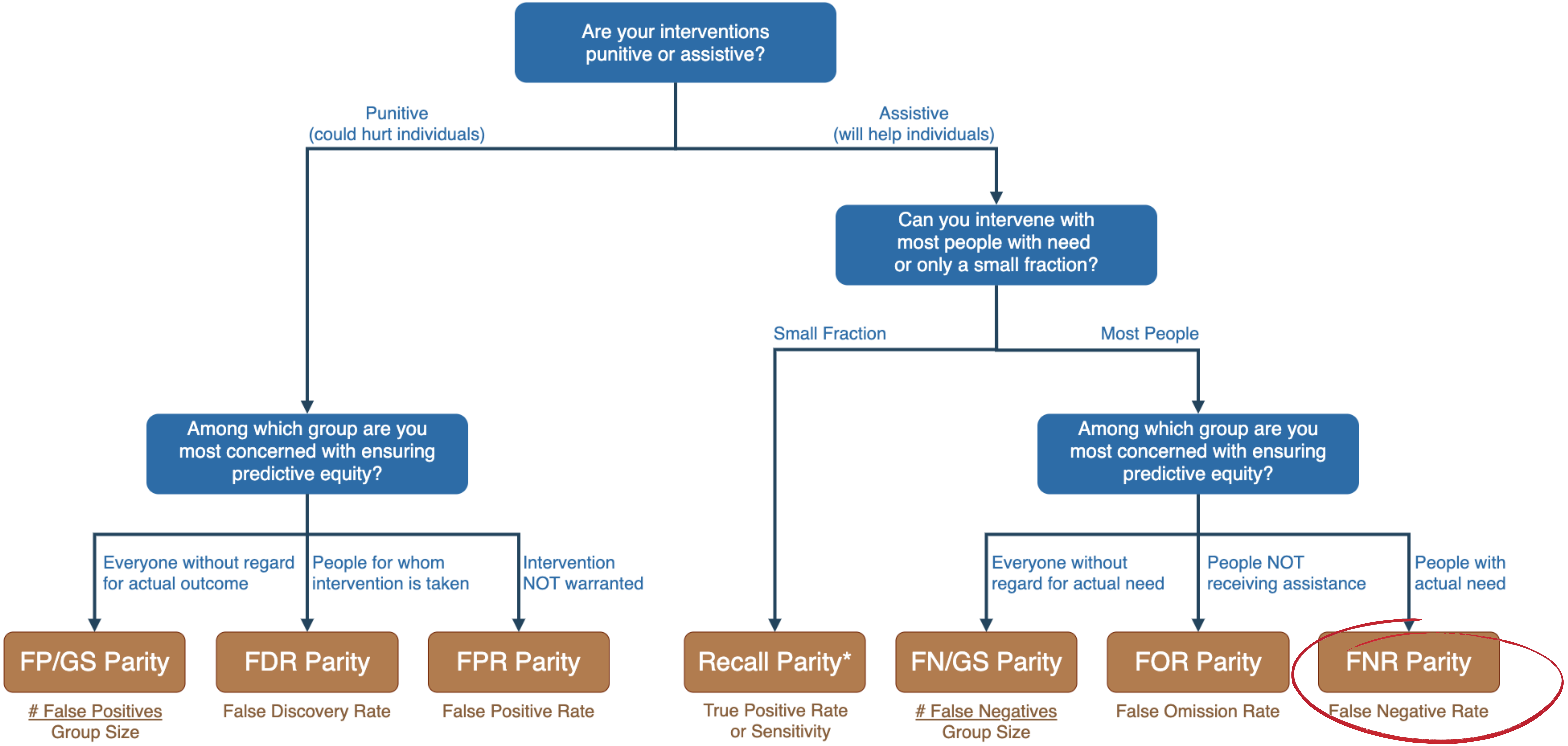


Pedro Saleiro, Benedict Kuester, Abby Stevens, Ari Anisfeld, Loren Hinkson, Jesse London, Rayid Ghani, Aequitas: A Bias and Fairness Audit Toolkit, arXiv preprint arXiv:1811.05577 (2018)

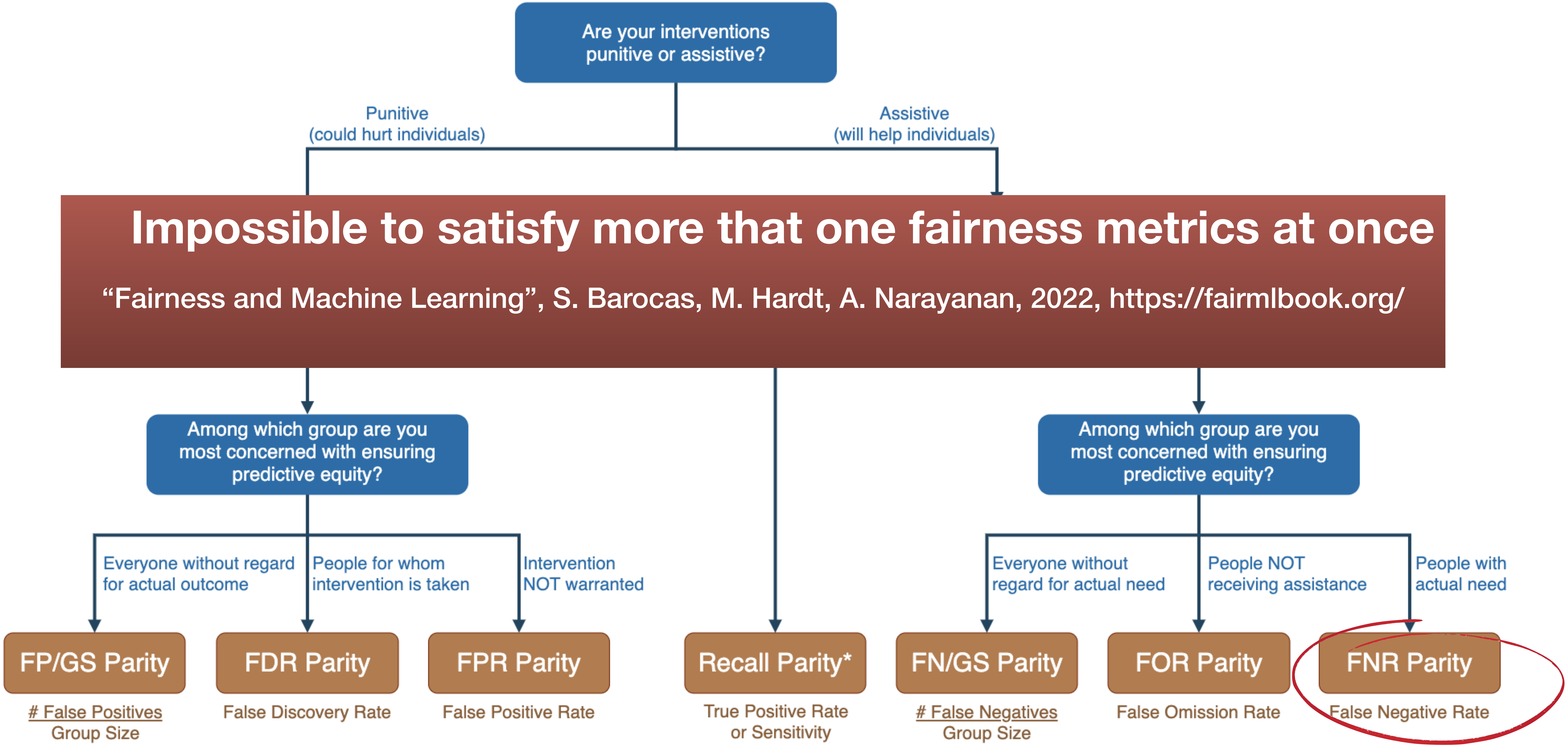
FAIRNESS TREE (Zoomed in)



FAIRNESS TREE
(Zoomed in)



FAIRNESS TREE
(Zoomed in)



Impossible to satisfy more that one fairness metrics at once

"Fairness and Machine Learning", S. Barocas, M. Hardt, A. Narayanan, 2022, https://fairmlbook.org/

Parity of Opportunity

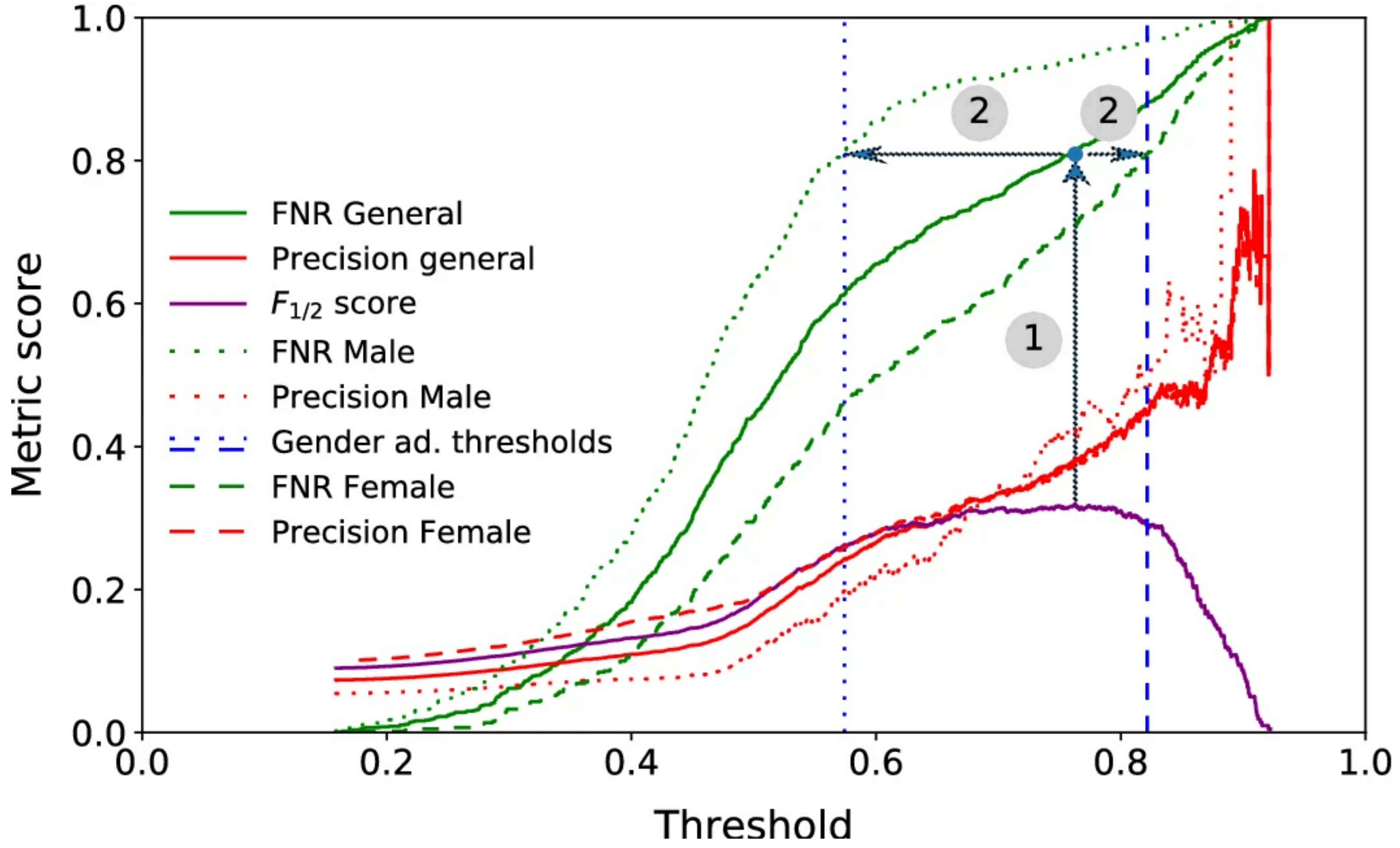
$$\text{FNR}_g \text{ disp.} = \frac{\text{FNR}_g}{\text{FNR}_{ref.group}} = \frac{Pr\left[\hat{Y}=0|Y=1 \wedge G=g\right]}{Pr\left[\hat{Y}=0|Y=1 \wedge G=ref.group\right]}$$

where Y and $\hat{Y}$ represent the real and predicted target values respectively (1 represents the 'unemployed', 0 the employed)

Disparity threshold 80% with respect to a reference group

Adaptive Threshold

$$F_\beta = (1+\beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$ with β at 0.5 to favour precision



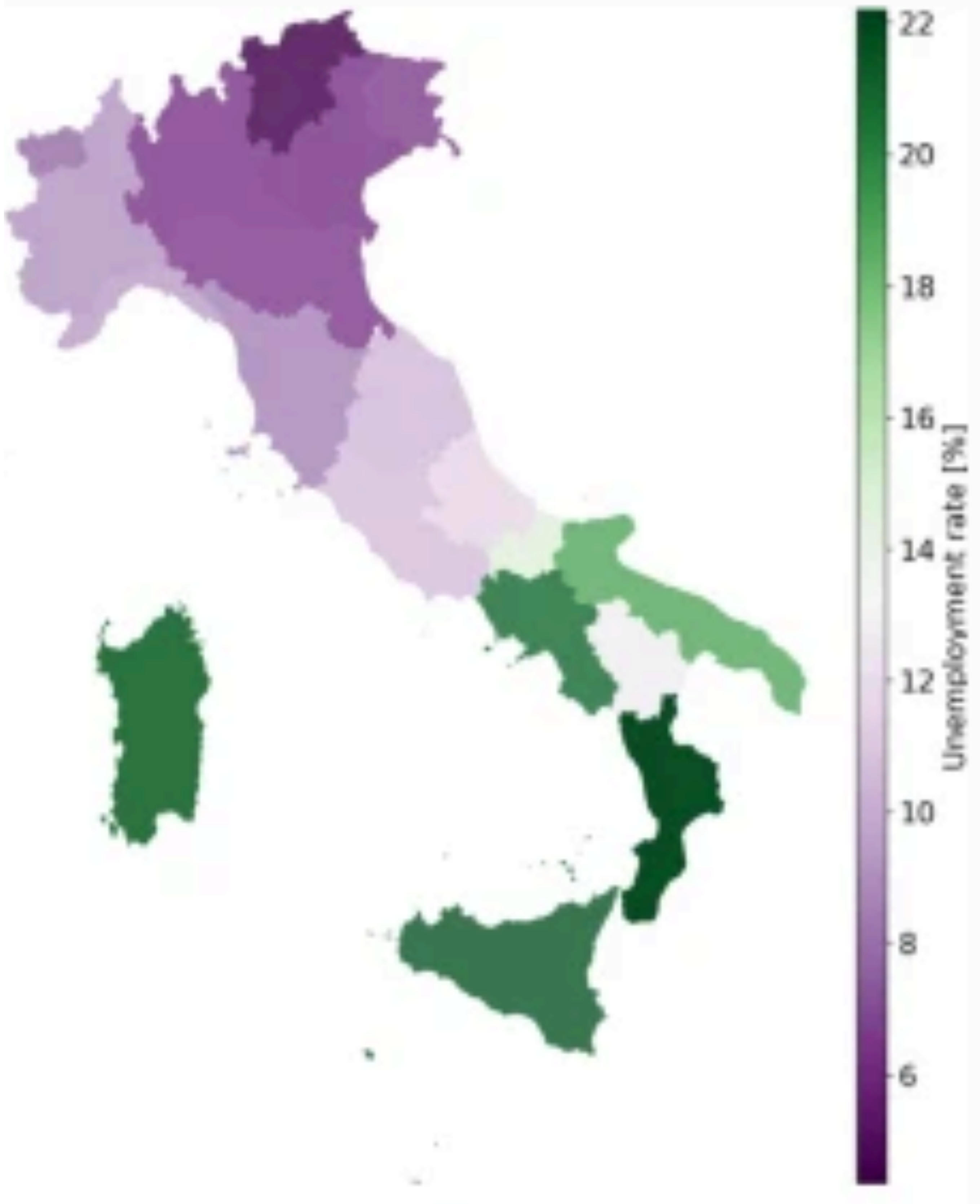
	Demo	NoDemo	Demo+AT.	NoDemo+AT.
Global Accuracy (Metric: AUC(std))				
Baseline	.50	.50	.50	.50
State of the Art	–	.61(.01) (*)	–	–
Our Approach	.74(.02)	.71(.02)	.74(.02)	.71(.02)
Precision and Recall				
Precision	.16(.02)	.18(.01)	**.26(.05)**	**.25(.03)**
Recall	**.56(.05)**	.48(.02)	.21(.05)	.22(.04)
Demographic accuracy (Metric: AUC(std))				
Gender (M)	.66(.05)	.64(.04)	.66(.05)	.64(.04)
Gender (F)	.78(.02)	.76(.02)	.78(.02)	.76(.02)
Age (17-24)	.70(.08)	.69(.08)	.70(.08)	.69(.08)
Age (25-34)	.66(.05)	.65(.05)	.66(.05)	.65(.05)
Age (35-44)	.74(.09)	.73(.08)	.74(.09)	.73(.08)
Age (45-54)	.61(.17)	.54(.16)	.61(.17)	.54(.16)
Age (55+)	.46(.31)	.46(.29)	.46(.31)	.46(.29)
Fairness (Metric: $\frac{FNR}{FNR_{ref}}$)				
Gender (ref.class: Male)				
Female	.47(.11)	.58(.14)	**1.0(.07)**	**1.02(.14)**
Age (ref.class: 17–24)				
25-34	.35(.08)	.62(.12)	.75(.09)	**.80(.08)**
35-44	.26(.12)	.49(.2)	.71(.09)	.73(.1)
45-54	.41(.24)	**.82(.35)**	**.82(.17)**	**.84(.19)**
55+	.59(.36)	**.99(.42)**	**.82(.18)**	**.91(.19)**

	Demo	NoDemo	Demo+AT.	NoDemo+AT.
Global Accuracy (Metric: AUC(std))				
Baseline	.50	.50	.50	.50
State of the Art	–	.61(.01) (*)	–	–
Our Approach	.74(.02)	.71(.02)	.74(.02)	.71(.02)
Precision and Recall				
Precision	.16(.02)	.18(.01)	**.26(.05)**	**.25(.03)**
Recall	**.56(.05)**	.48(.02)	.21(.05)	.22(.04)
Demographic accuracy (Metric: AUC(std))				
Gender (M)	.66(.05)	.64(.04)	.66(.05)	.64(.04)
Gender (F)	.78(.02)	.76(.02)	.78(.02)	.76(.02)
Age (45-54)	[illegible]	[illegible]	[illegible]	[illegible]
Age (55+)	.46(.31)	.46(.29)	.46(.31)	.46(.29)
Fairness (Metric: $\frac{FNR}{FNR_{ref}}$)				
Gender (ref.class: Male)				
Female	.47(.11)	.58(.14)	**1.0(.07)**	**1.02(.14)**
Age (ref.class: 17–24)				
25-34	.35(.08)	.62(.12)	.75(.09)	**.80(.08)**
35-44	.26(.12)	.49(.2)	.71(.09)	.73(.1)
45-54	.41(.24)	**.82(.35)**	**.82(.17)**	**.84(.19)**
55+	.59(.36)	**.99(.42)**	**.82(.18)**	**.91(.19)**

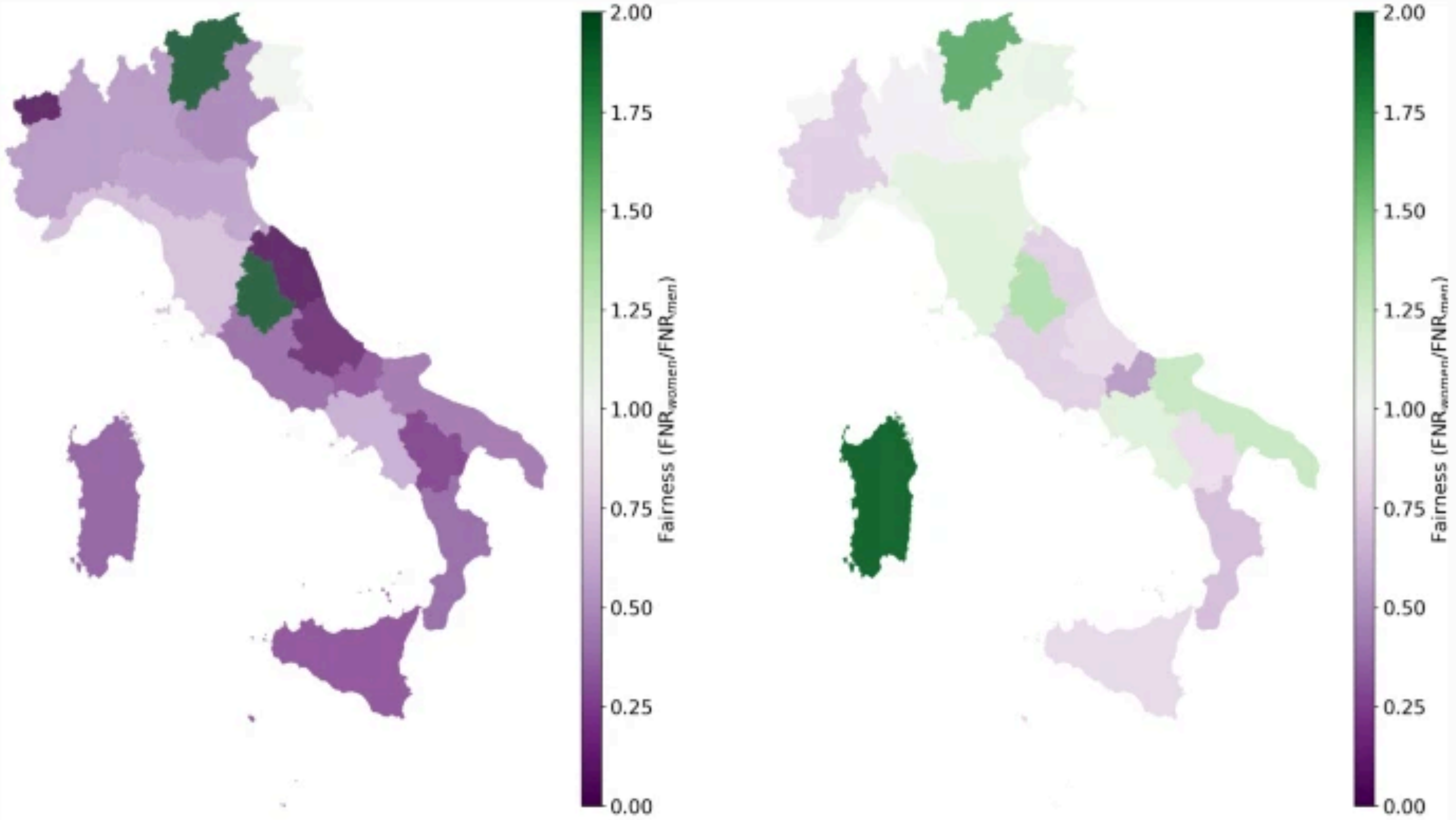
Gender (ref.class: Male)
Female .47(.11) .58(.14) **1.0(.07)** **1.02(.14)**

	Demo	NoDemo	Demo+AT.	NoDemo+AT.
Global Accuracy (Metric: AUC(std))				
Baseline	.50	.50	.50	.50
State of the Art	–	.61(.01) (*)	–	–
Our Approach	.74(.02)	.71(.02)	.74(.02)	.71(.02)
Precision and Recall				
Precision	.16(.02)	.18(.01)	**.26(.05)**	**.25(.03)**
Recall	**.56(.05)**	.48(.02)	.21(.05)	.22(.04)
Demographic accuracy (Metric: AUC(std))				
Gender (M)	[illegible]	[illegible]	[illegible]	[illegible]
Age (55+)	.46(.31)	.46(.29)	.46(.31)	.46(.29)
Fairness (Metric: $\frac{FNR}{FNR_{ref}}$)				
Gender (ref.class: Male)				
Female	.47(.11)	.58(.14)	**1.0(.07)**	**1.02(.14)**
Age (ref.class: 17–24)				
25-34	.35(.08)	.62(.12)	.75(.09)	**.80(.08)**
35-44	.26(.12)	.49(.2)	.71(.09)	.73(.1)
45-54	.41(.24)	**.82(.35)**	**.82(.17)**	**.84(.19)**
55+	.59(.36)	**.99(.42)**	**.82(.18)**	**.91(.19)**

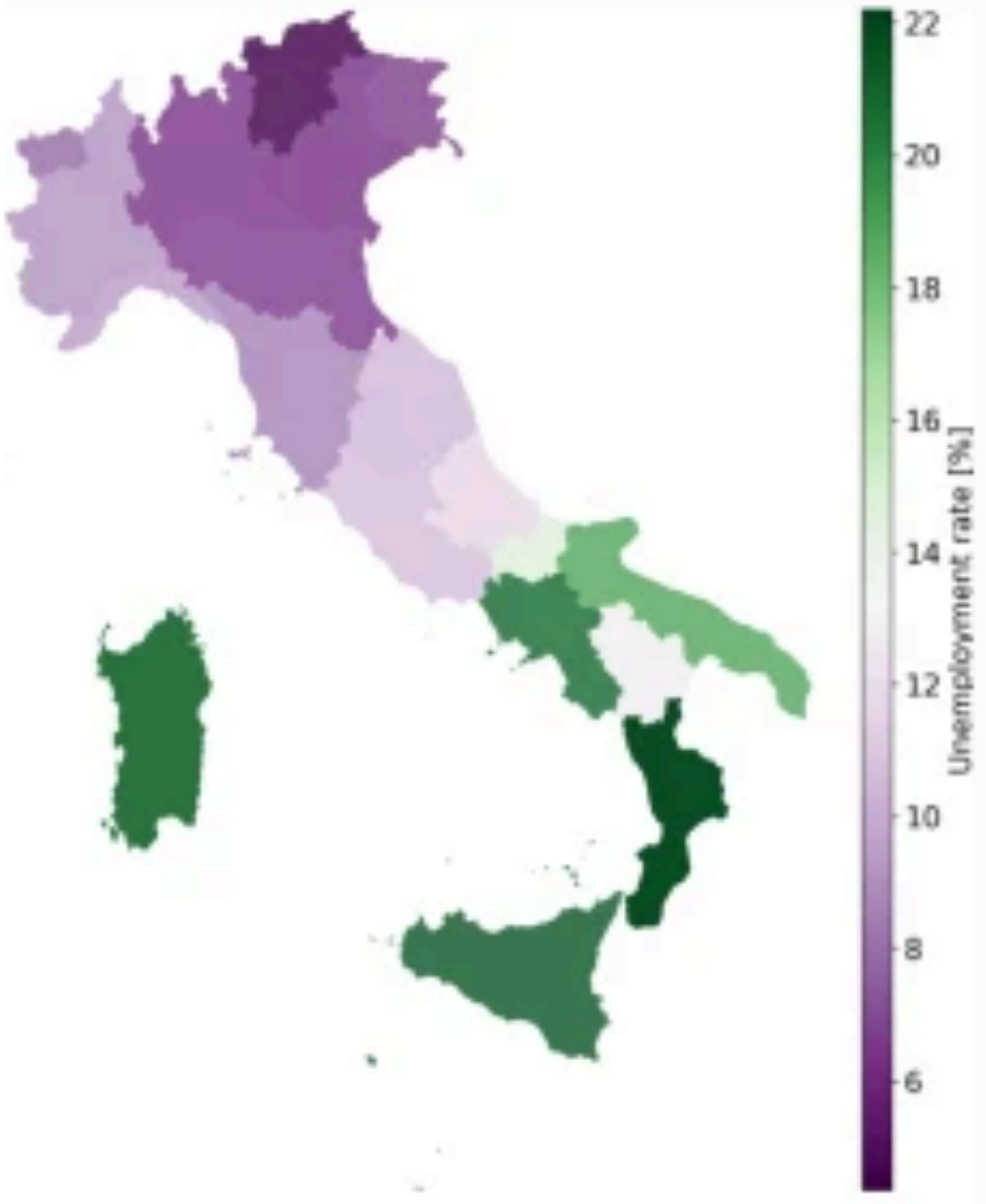
Our Approach	.74(.02)	.71(.02)	.74(.02)	.71(.02)
Precision and Recall				
Precision	.16(.02)	.18(.01)	**.26(.05)**	**.25(.03)**
Recall	**.56(.05)**	.48(.02)	.21(.05)	.22(.04)



(c) Unemployment Rate



(a) NoDemo **(b) NoDemo+AT.**



(c) Unemployment Rate

Gender fairness per region in the NoDemo (left) and NoDemo+Thresh. (right) models. Gender fairness is computed as the FNR of females in relation to that of males. The color extremities are both unfair (Color figure online)

Conclusions

- Fairness through unawareness does not suffice
- Models with **good overall accuracy aren’t always efficient** in humanitarian domain
- Easily **generalisable approach** to any **fairness metrics**, **demographic** feature, and **digital data**

Link to the paper

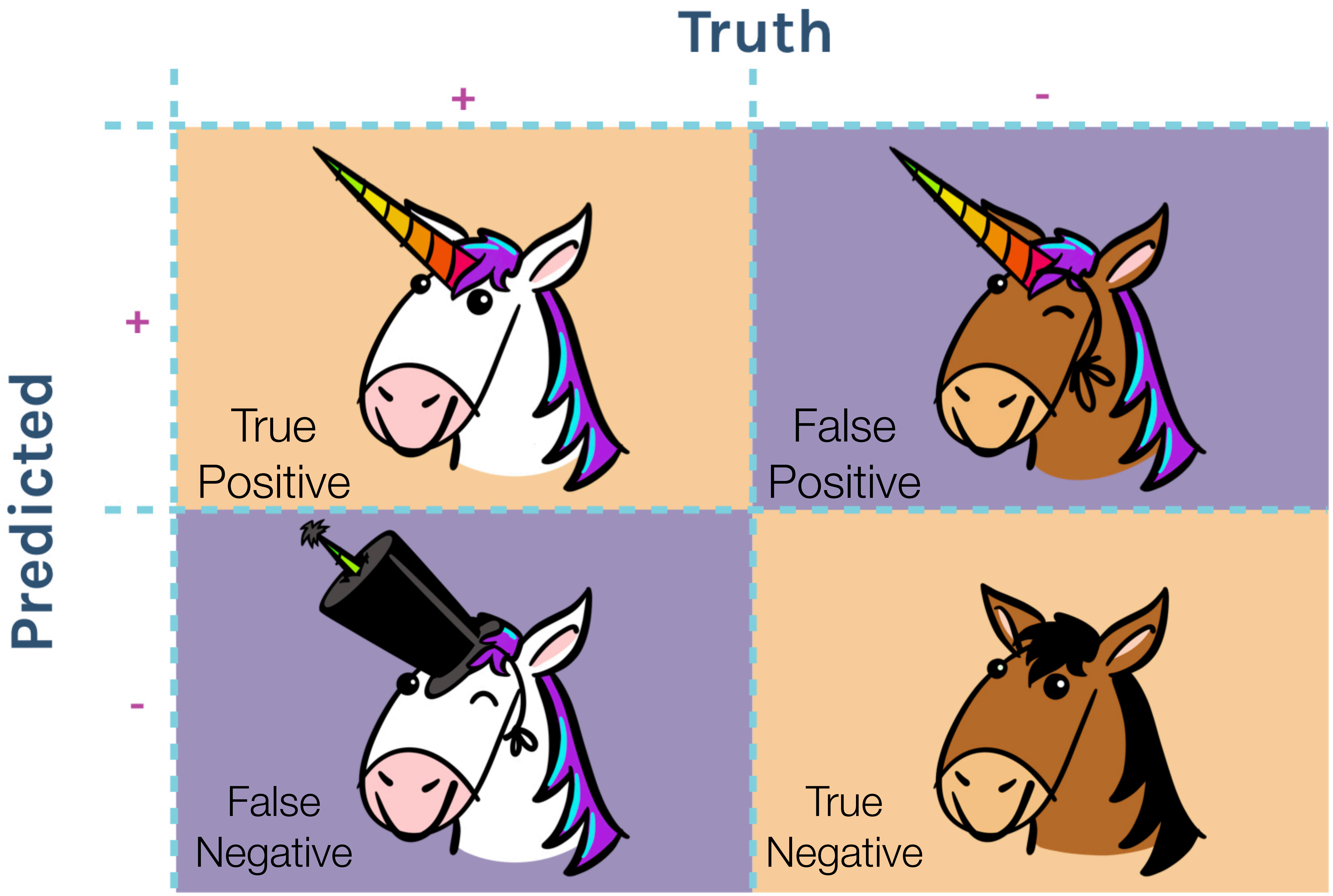
Kyriaki Kalimeri
kyriaki.kalimeri@isi.it
@KyriakiKalimeri

Description	# Features
Liked Pages	2,063,944
Liked Pages per Category To express how much a participant is interested in different categories of pages, we compute the number of pages he gave like to inside each Category.	1.553
Normalised Categories: As the participants' activity can greatly vary, we normalise the *Liked Pages per Category* to have sum 1.	1.553
Median Page Popularity This index shows how much a participant likes popular pages. The *popularity* of a Page is the number of users that gave like to it, as reported by Facebook in the Page profile.	1
Standard Deviation of Page Popularity	1
Median Category Popularity This index shows how much a participant likes popular categories.	1
Total number of Page likes One feature containing the total number of pages liked by the participant.	1
Total number of liked Categories One feature containing the total number of categories with pages liked by the participant.	1

Machine Learning and Vulnerable Populations

Precision = Of all the positives how many are truly positives?

Recall = Of the real positives, how many are predicted correctly?



LightGBM
10-fold stratified x validation
balanced weighting