Inf2 – Foundations of Data Science
S2 Week 4: Ethics of supervised learning

THE UNIVERSITY of EDINBURGH
informatics
FOUNDATIONS OF DATA SCIENCE

Overview

- Fairness in classification and protected attributes
- Credit scoring case study

Fairness in Classification

Advertising

Education

Financial aid

Health Care

Banking

Insurance

Taxation

many more...

Prediction = Judgement

Prediction = judgement. It impacts lives of real people.

- Recidivism prediction for granting bail
- Predicting credit worthiness to give loans
- Predicting success in school/job to decide on admission/hiring

Are people being treated as they deserve?

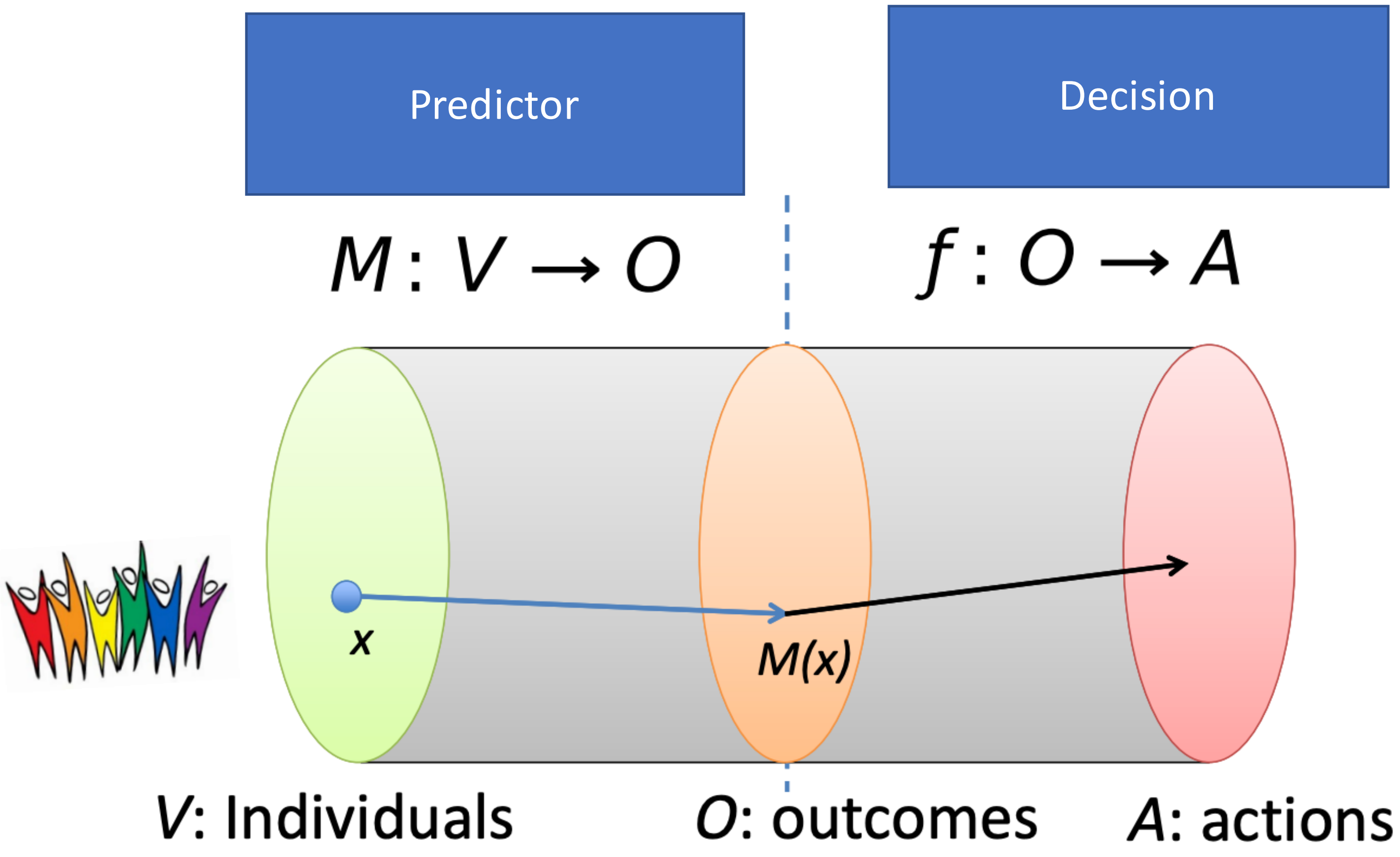
The concern

- Certain attributes should be irrelevant to decisions.
- Example: gender, sexual orientation, minority groups – ethnic, religious, medical, geographic, etc…
- Protected by law!
- Discrimination arises even without intent

Example

- Google+ tries to classify real vs fake names
- Fairness problem:
  - Most training examples standard white American names: John, Jennifer, Peter, Jacob, ...
- Ethnic names often unique, much fewer training examples Likely
- Outcome: Prediction accuracy worse on ethnic names

From Invidividuals to decisions



Fairness in Algorithmic Decision Making

1. Why it is important
2. Credit scoring as an example
3. Overview of equality legislation
4. Case study: Andreeva G, Matuszyk A (2019) ‘The Law of Equal Opportunities or Unintended Consequences: the impact of unisex risk assessment in consumer credit’, *Journal of Royal Statistical Society, Series A*, https://rss.onlinelibrary.wiley.com/doi/10.1111/rssa.12494

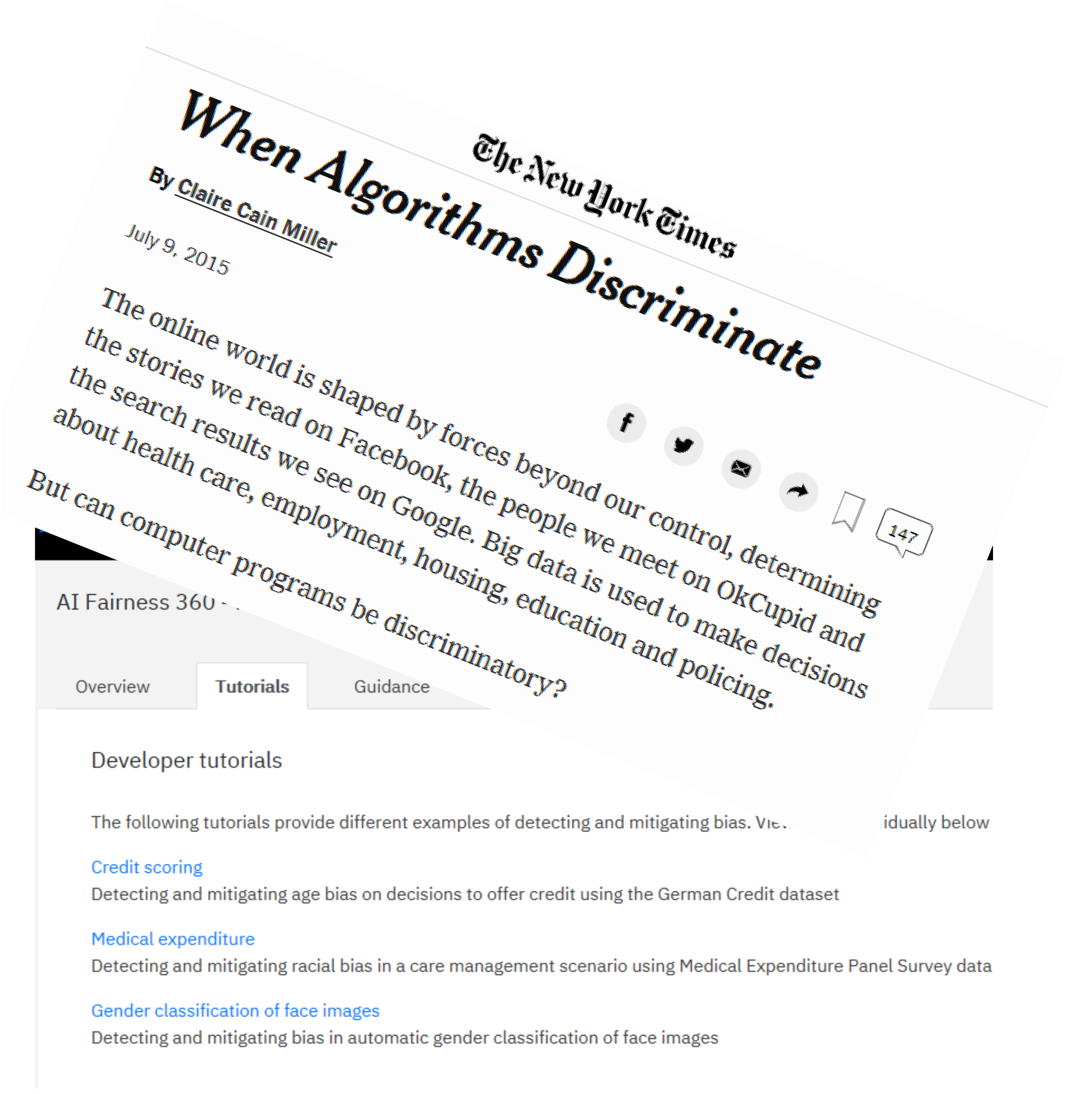
European Union regulations on algorithmic decision making and a "right to explanation"

JANUARY 31, 2017

European Union regulations on algorithmic decision-making and a "right to explanation" Goodman & Flaxman, *2016*

In just over a year, the General Data Protection Regulation (GDPR) becomes law in European member states. This paper focuses on just one particular aspect of the new law, article 22, as it relates to *profiling*, *non-discrimination*, and *the right to an explanation*.

> *Article 22: Automated individual decision-making, including profiling*, potentially prohibits a wide swath of algorithms currently in use in, e.g., recommendation systems, credit and insurance risk assessments, computational advertising, and social networks. This raises important issues that are of particular concern to the machine learning community. In its current form, the GDPR's requirements could require a complete overhaul of standard and widely used algorithmic



Credit scoring

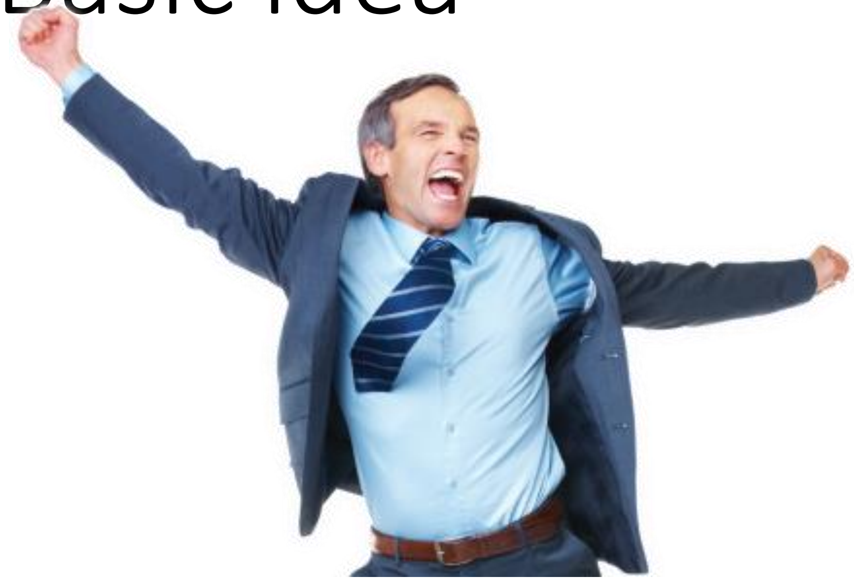
What is credit scoring?

- Decision support systems used in consumer credit
- Aims at risk assessment of:
  - potential borrowers (application scoring)
  - existing borrowers (behavioural scoring)
- Risk/creditworthiness is usually measured by **Probability of Default (PD)**
  - Larger value means higher risk
- PD is predicted from potential borrower’s characteristics on the basis of the analysis of known performance of previous customers
  - Cf the lectures on Logistic Regression

Example of a scoring table

Time at current address	**Less than 6 months**	**6m – 2 years**	**2 – 6 years**	**6 - 10 years**	**10 + years**	**Unknown**	
	0	3	6	13	25	0	
Residential Status	**Owner**	**Tenant**	**With parents**	**Unknown**			
	15	5	2	0			
Banking	**Current account**	**Saving account**	**Current and saving**	**No account**	**Unknown**		
	5	10	14	0	0		
Occupation	**Retired**	**Full-time**	**Part-time**	**Self-employed**	**Student**	**Other**	**Un-known**
	21	16	7	6	5	10	0
Age	**18-25**	**26-31**	**32-40**	**41-54**	**55+**	**Unknown**	
	5	10	15	20	25	0	

The Basic Idea



5 years at current address + 6

Home Owner + 15

Current and Saving Account + 14

Full Time Work + 16

40 years old + 15

Score 66

6 months at current address + 3

Tenant + 5

Current Account + 5

Self-Employed + 6

20 years old + 5

Score 24

Equality legislation

Equality/Anti-Discrimination Legislation

USA

Equal Credit Opportunity Act (ECOA, 1974) prohibits characteristics from being used in credit scoring (race, colour, national origin, gender, marital status, religion, receipt of public assistance, or exercise of consumer protection rights). Age has a special status.

EU

Articles 8, 19 of the Treaty of the Functioning of European Union (TFEU);

Gender Directive - Council Directive 2004/113/EC of 13 December 2004

Proposal for a Council Directive on implementing **the principle of equal treatment** between persons irrespective of religion or belief, disability, age or sexual orientation, COM(2008) 426 final.

UK

Equality Act (2010)

Protected characteristics under UK Equality act

- Age - unless good reason ('objective justification') can be shown for the differential treatment
- disability
- gender reassignment
- marriage and civil partnership
- pregnancy and maternity
- race
- religion or belief
- sex
- sexual orientation

Data description

- Portfolio of auto loans from a major bank in an EU country from 2003-2010
- Default definition is defaulting on the loan for 2 months (65 days)
- 80% (training) and 20% (test)

	Training			Test		
	Good	Bad	Total	Good	Bad	Total
Female	16746 98.70%	220 1.30%	16966 26.71%	4186 98.70%	55 1.30%	4241 26.71%
Male	45696 98.18%	847 1.82%	46543 73.29%	11424 98.18%	212 1.82%	11636 73.29%
Total	62442 98.32%	1067 1.68%	63509	15610 98.32%	267 1.68%	15877

Research design

- Two Logistic regression models to predict Probability of Default:
  1. Model with *Gender* (training sample comprising both men and women)
  2. Model without *Gender*
  3. Model trained and tested only on men
  4. Model trained and tested only on women
- The models are compared from the points of view of
  1. how they affect the chances of men/women being offered credit
  2. predictive accuracy

Relevant variables

There are 11 final variables selected by significance and predictive accuracy

- Marital status
- # kids
- Income
- Time in employment
- Profession
- Phone given
- Gender
- Loan duration
- Downpayment
- Car price
- Car age

Table 2. Parameter estimates (with standard errors are in parentheses) and model fit statistics for four logistic regression models to predict the PD†

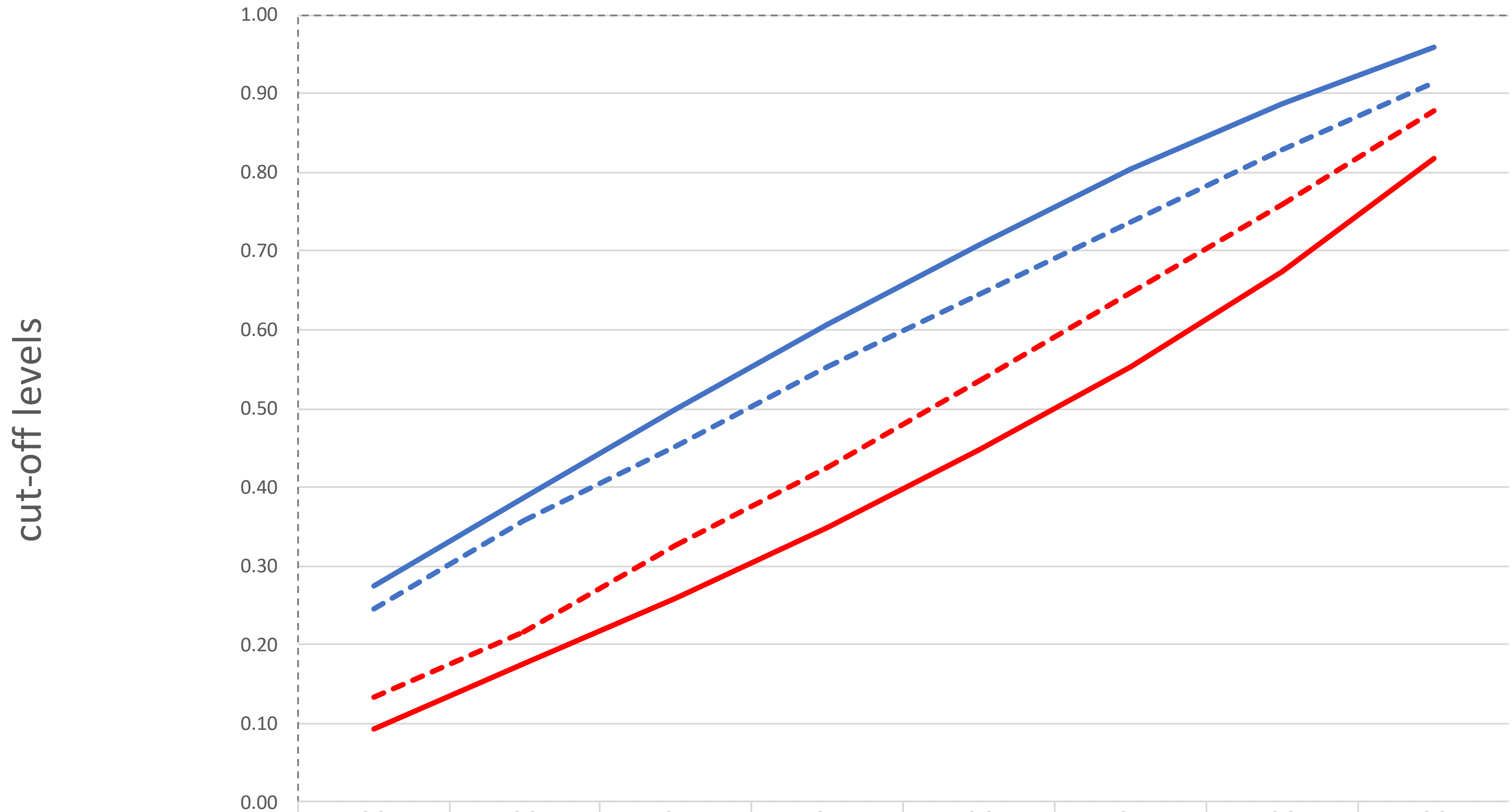
Variable	*Attribute or category*	*% in category*	*Results for model with gender (model 1)*	*Results for model without gender (model 2)*	*Results for model for men only (model 3)*	*Results for model for women only (model 4)*
Intercept			−7.3942‡ (0.1722)	−7.5207‡ (0.1708)	−7.6844‡ (0.2073)	−7.0066‡ (0.3135)
Gender	Female	26.71	−0.457‡ (0.0867)			
Number of children (reference: no kids)	1 kid	23.26	0.19 (0.1009)	0.1525 (0.1000)	0.267§§ (0.1219)	0.1248 (0.1874)
	2 kids	15.04	0.1918 (0.1302)	0.1763 (0.1298)	[illegible]	[illegible]
	3+ kids	3.12	0.3553 (0.2313)	0.3494 (0.2310)		
	Missing information	10.87	−0.6816‡ (0.1254)	−0.6944‡ (0.1251)		
Car price (reference: medium price lower)	Cheap	5.28	−1.0987‡ (0.1326)	−1.1048‡ (0.1322)		
	Medium price higher	59.58	0.426§ (0.1099)	0.4406‡ (0.1095)		
	Expensive	15.87	1.1813‡ (0.1116)	1.1955‡ (0.1112)		
Down payment, % (reference: (35%, 50%])	⩽25%	16.87	1.2702‡ (0.1087)	1.2603‡ (0.1085)		
	(25%, 35%]	8.65	0.7133‡ (0.1248)	0.7096‡ (0.1246)		
	>51%	34.49	−1.2147‡ (0.1940)	−1.2075‡ (0.1941)		
Car age, years (reference: [0, 2))	2	1.56	1.311‡ (0.1454)	1.3197‡ (0.1448)		
	[3,4)	3.25	1.8426‡ (0.1196)	1.8691‡ (0.1191)		
	>4	3.31	2.5302‡ (0.1348)	2.5635‡ (0.1343)		

Questions

1. In the model with gender, is gender significant?
2. Does being female make the probability of default greater or smaller?
3. And by how much?
4. What factors increase and decrease the probability of default the most?

‡ p-value < 0.0001.
§ p-value < 0.005.
§§ p-value < 0.05.

Rejection rates by Gender for all unmarried customers



	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
female (M2 no gender)	0.13	0.22	0.33	0.43	0.54	0.65	0.76	0.88
male (M2 no gender)	0.25	0.36	0.45	0.55	0.65	0.74	0.83	0.91
female (M1 with gender)	0.09	0.18	0.26	0.35	0.45	0.55	0.67	0.82
male (M1 with gender)	0.28	0.39	0.50	0.61	0.71	0.80	0.89	0.96

$\mathrm{logit}(p) = \beta_0 + \Sigma\beta_i x_i$

Reject if
p > cut-off probability

E.g. with cut-off of 0.6:

- 45% of women will be rejected
- 71% of men will be rejected

Courtesy of Galina Andreeva

What can we conclude?

- Women benefit from the model with gender:
  - Women have had lower default rates in the past
- When gender is removed in the sample studied chances of being accepted for credit decrease for women, but increase for men
- Women in the group sampled still benefit when gender is not included in the model
- Thus **equal treatment of individuals** by ignoring a protected characteristic does not lead to **equal outcome at the group level**
- Why is there still an effect?

Proxies

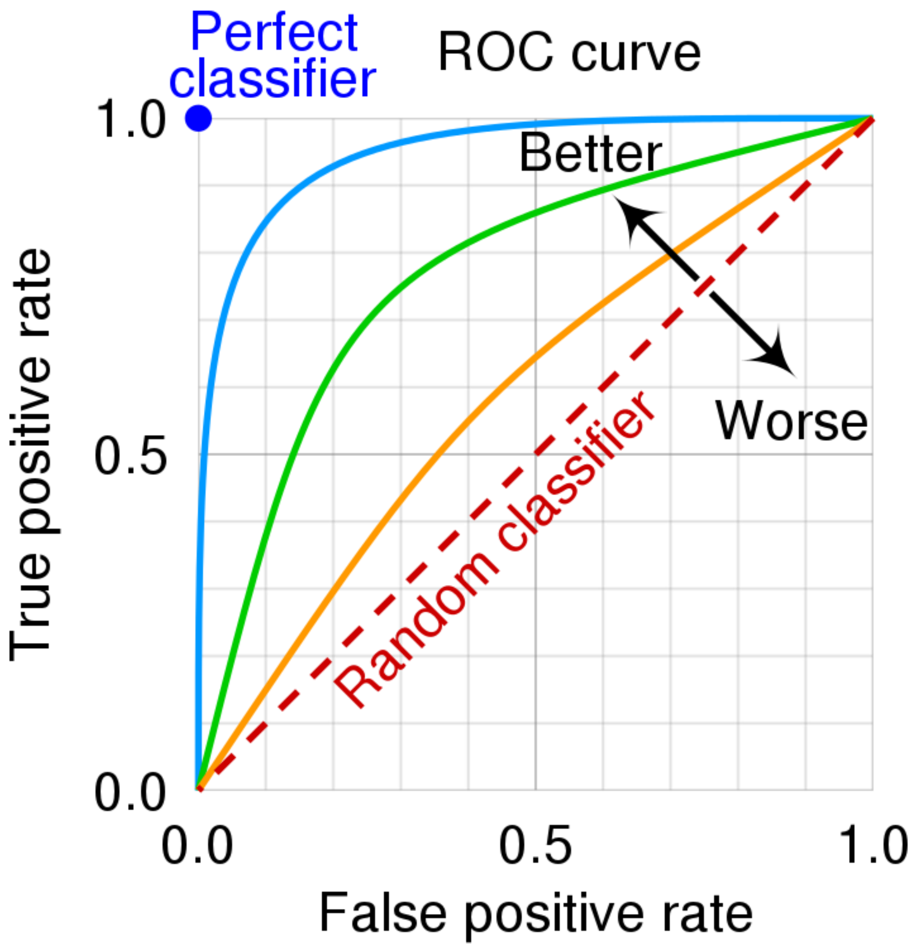
Table 2 (*continued*)

Variable	*Attribute or category*	*% in category*	*Results for model with gender (model 1)*	*Results for model without gender (model 2)*	*Results for model for men only (model 3)*	*Results for model for women only (model 4)*
Profession or occupation (reference: gender neutral)	Female profession	5.89	−0.5111§§ (0.1938)	−0.6108§ (0.1928)	−0.7843§§ (0.2827)	−0.2068 (0.2653)
	Male profession	13.08	−0.2709§§ (0.1134)	−0.224§§ (0.1129)	−0.2832§§ (0.1246)	−0.2767 (0.3003)
Model fit statistics						
Intercept AIC			10838.202	10838.202	8467.386	2351.084
Intercept and covariates AIC			6976.242	7003.602	5117.254	1833.609
Cox and Snell pseudo-R^2		0.0600	0.0600	0.0595	0.0707	0.0337
Nagelkerke pseudo-R^2		0.3823	0.3823	0.3796	0.4253	0.2606

†The reference category is given in parentheses under the corresponding variable name.
‡ p-value < 0.0001.
§ p-value < 0.005.
§§ p-value < 0.05.

Is the model without gender as accurate as the one with gender?

- To measure accuracy, use the metric of **Area Under the Curve (AUC)**
- To understand AUC, first understand the **Receiver Operator Characteristic** (ROC)



Demo at https://arogozhnikov.github.io/2015/10/05/roc-curve.html

Predictive accuracy, AUC

	Total sample		Men only		Women only	
	Model 1 with Gender	Model 2 without Gender	Model 1 with Gender	Model 2 without Gender	Model 1 with Gender	Model 2 without Gender
Train	0.9207	0.9211	0.9334	0.9331	0.8730	0.8739
Test	0.8901	0.8898	0.9147	0.9139	0.7965	0.7943

- Models with and without gender have near-equal prediction accuracy
- Prediction accuracy is lower when smaller group is trained

Discussion

- Equal treatment does not translate into equal outcomes
- Minority segments are dominated by majority ones
- It is not possible to completely remove the effect of a protected characteristic without deleting all correlated characteristics
- Conclusion in the paper: the existing law is not effective in promoting equality when it comes to algorithms
- What do we think?