The European Commission's science and knowledge service

Joint Research Centre

European Commission

Why machine learning may lead to unfairness

Songül Tolan[1], **Marius Miron[1]**, Emilia Gomez[1,2], Carlos Castillo[2]

[1]European Commission's Joint Research Centre
[2]Universitat Pompeu Fabra

European Commission

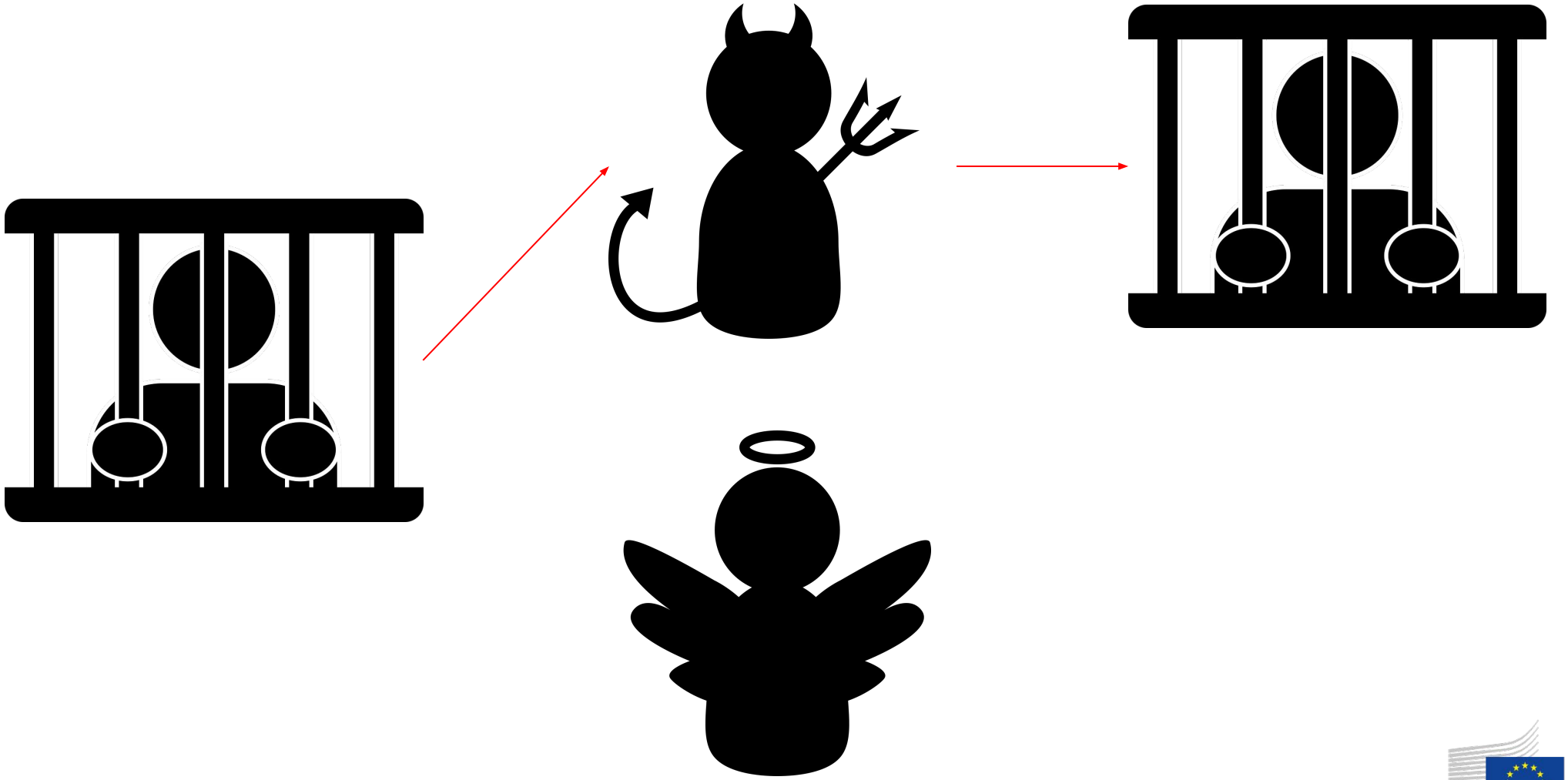
Machine learning for decision making

European Commission

The criminal justice case

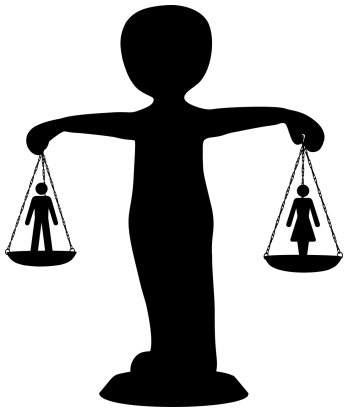
Trade-off: predictive performance vs fairness

Criminal recidivism



Criminal recidivism prediction

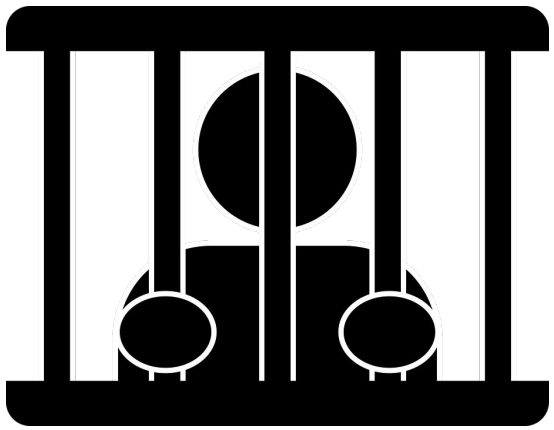
Prisoner



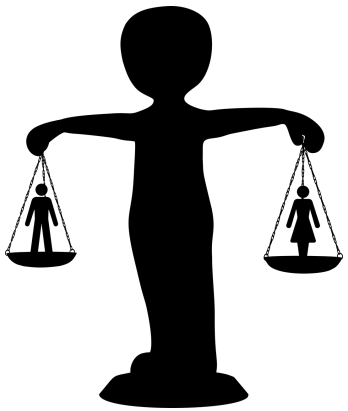
Human expert

Decision / Sentence

Criminal recidivism prediction

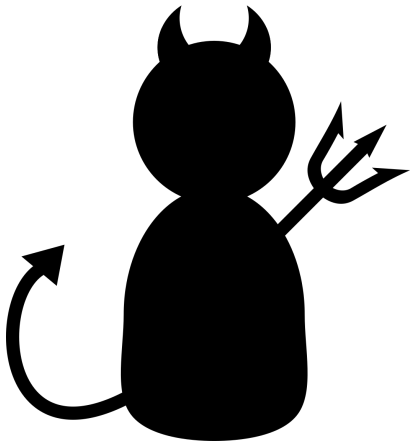


Prisoner

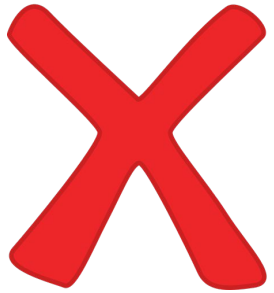


Human expert

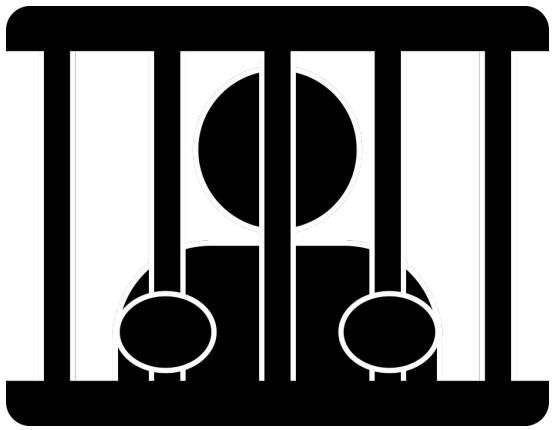
Decision / Sentence



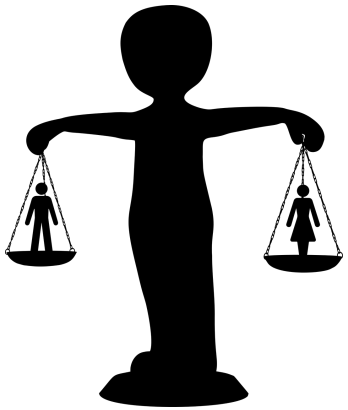
Outcome



Criminal recidivism prediction



Prisoner



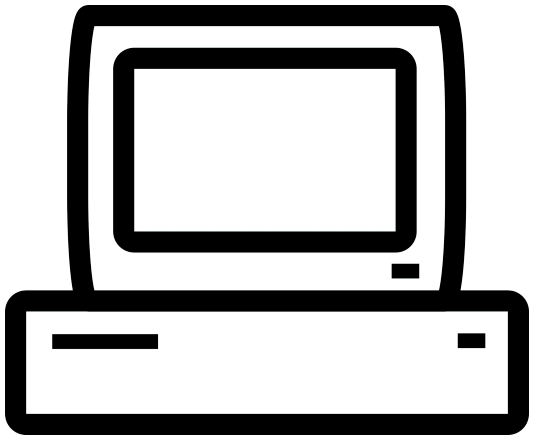
Human expert

Decision / Sentence

Outcome

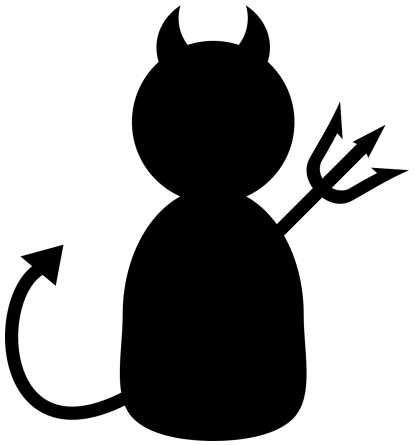
Criminal recidivism prediction

Features



Machine learning model

Prediction



Outcome

Criminal recidivism prediction

Examples of static features:

- **Age at crime**
- **Sex**
- **Nationality**
- **Previous number of crimes**
- **Sentence**
- **Year of crime**
- **Probation**

Fairness

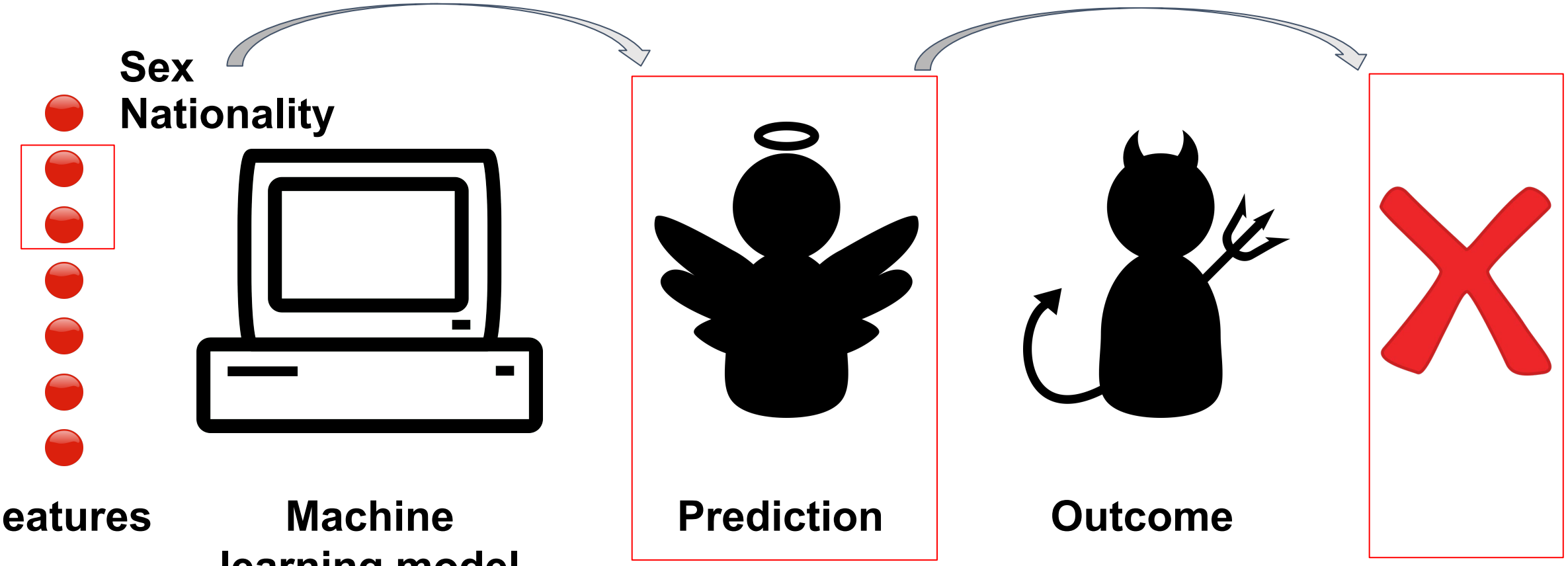
A decision is fair if it does not discriminate against people based on their membership to a protected group

Fairness

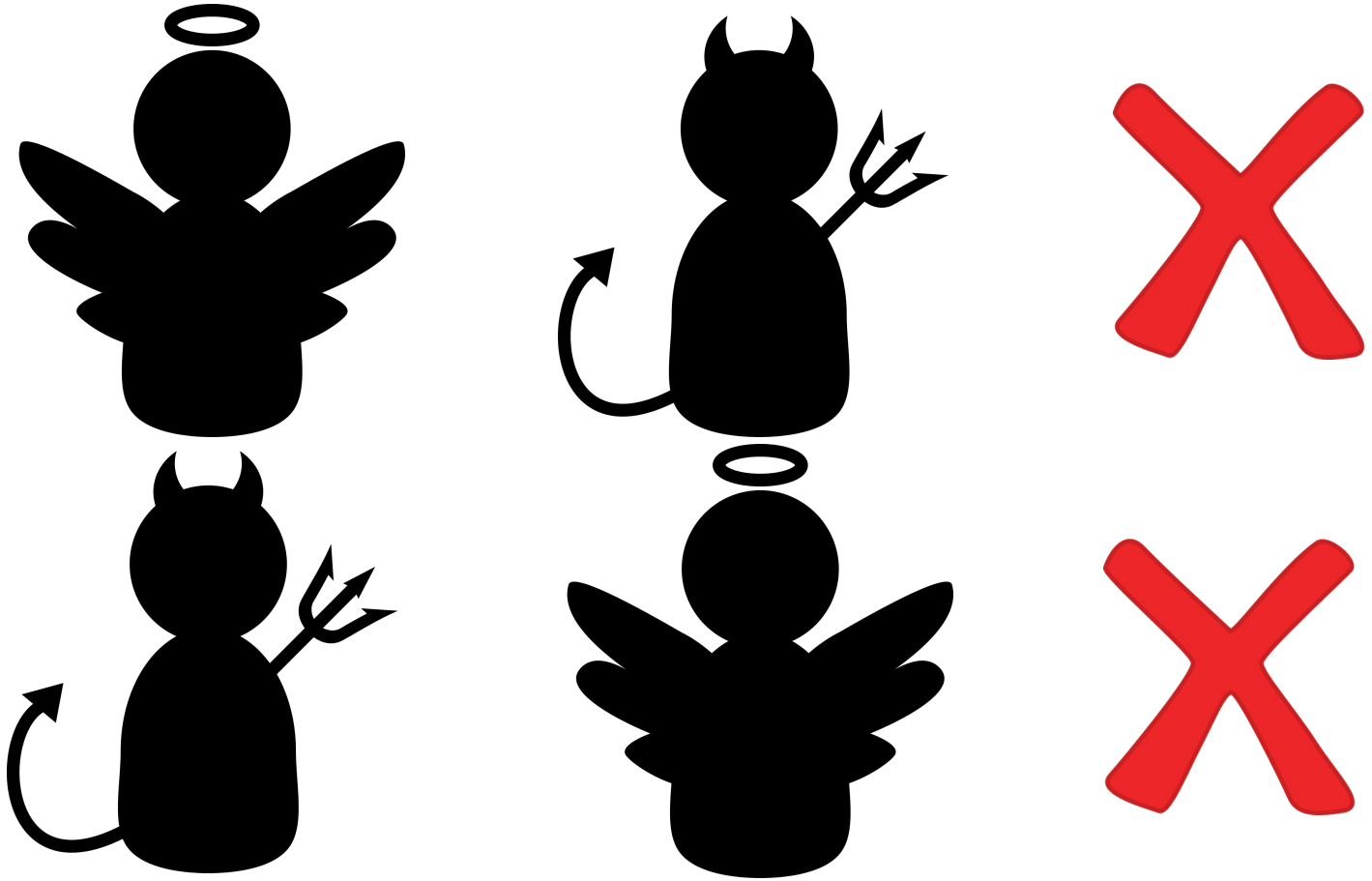
Example of protected features:

- **Age at crime**
- **Sex**
- **Nationality**
- **Previous number of crimes**
- **Sentence**
- **Year of crime**
- **Probation**

Measuring unfairness



Measuring unfairness



Prediction

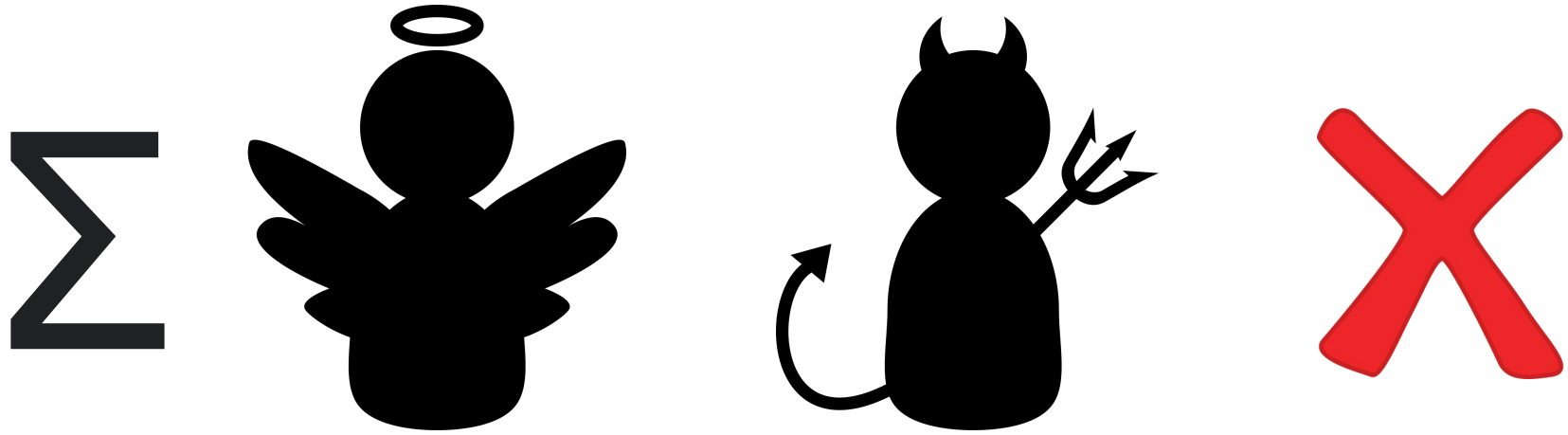
Outcome

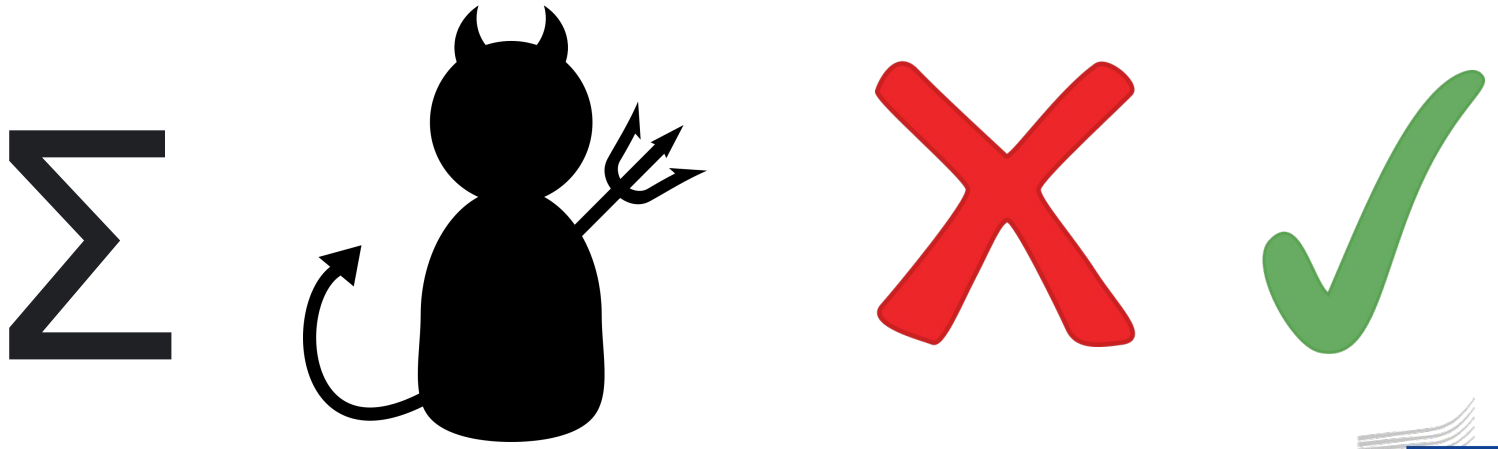
False negative

False positive

European Commission

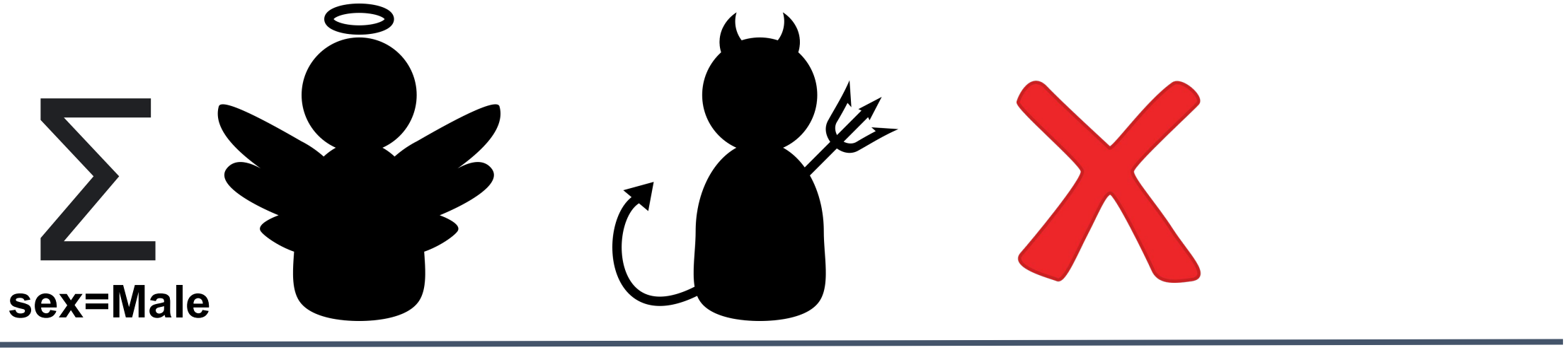
False negative rate = Miss rate





False positive rate = False alarm rate

Group fairness - sex



False negative rate disparity

$$FNR_{disparity} = \frac{FNR_{female}}{FNR_{male}}$$

How likely it is for a member of a group to be wrongfully labeled as non-recidivist.

Headache?

European Commission

Too complicated?

The fairness in machine learning literature comprises at least 21 disparity metrics.

Juvenile recidivism

European Commission

Risk assessment tools

Structured Assessment of Violence Risk in Youth (SAVRY)

- high degree of involvement from human experts
- open and interpretable (in comparison with COMPAS)
- 24 risk factors scored low, medium or high

SAVRY

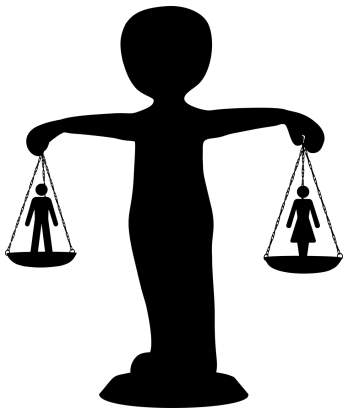
Examples of SAVRY features:

- **Early violence**
- **Self-harm history**
- **Home violence**
- **Poor school achievement**
- **Stress and poor coping**
- **Substance abuse**
- **Criminal parent/caregiver**

Criminal recidivism prediction

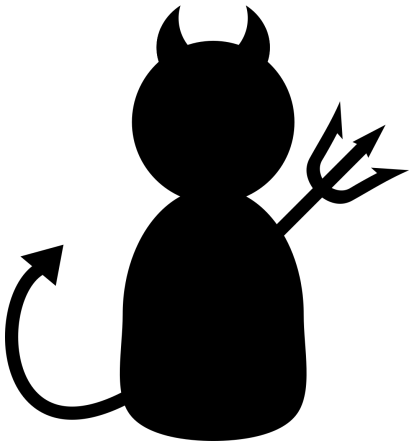
SAVRY features

SAVRY sum

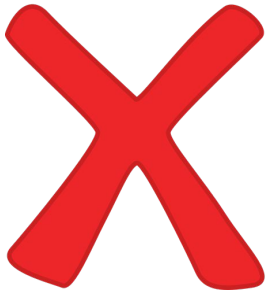


Expert

Final expert evaluation



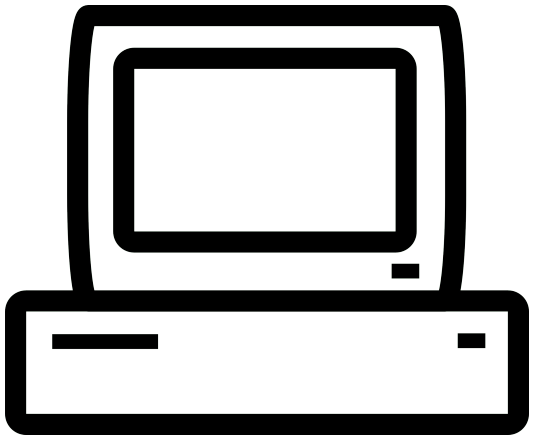
Outcome



European Commission

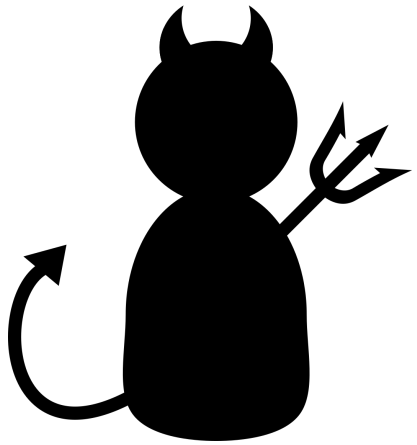
Static ML

Features



Machine learning model

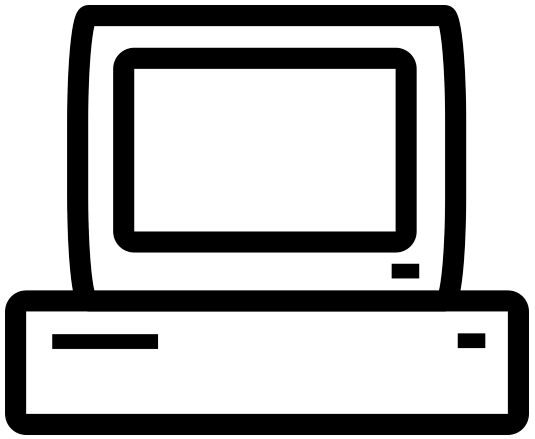
Prediction



Outcome

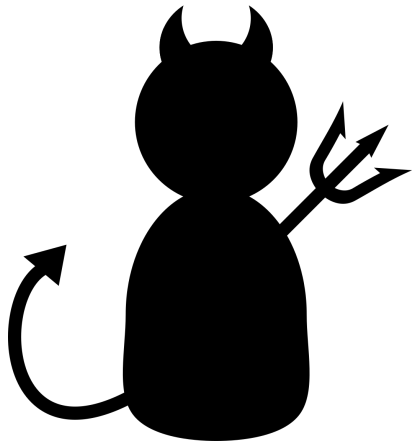
SAVRY ML

SAVRY features



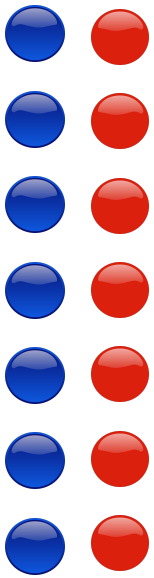
Machine learning model

Prediction

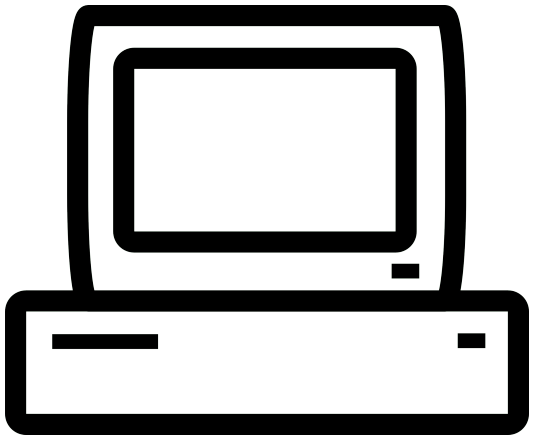


Outcome

Static + SAVRY ML

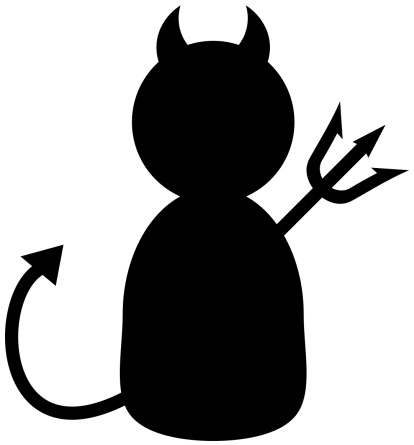


Features



Machine learning model

Prediction



Outcome

Dataset

Juvenile offenders in Catalonia[1]

- 855 people
- crimes between 2002 -2010, release in 2010
- age at crime between 12 and 17 years old
- status followed up on 2013 and 2015

1. Open data: http://cejfe.gencat.cat/en/recerca/opendata/jjuvenil/reincidencia-justicia-menors/index.html

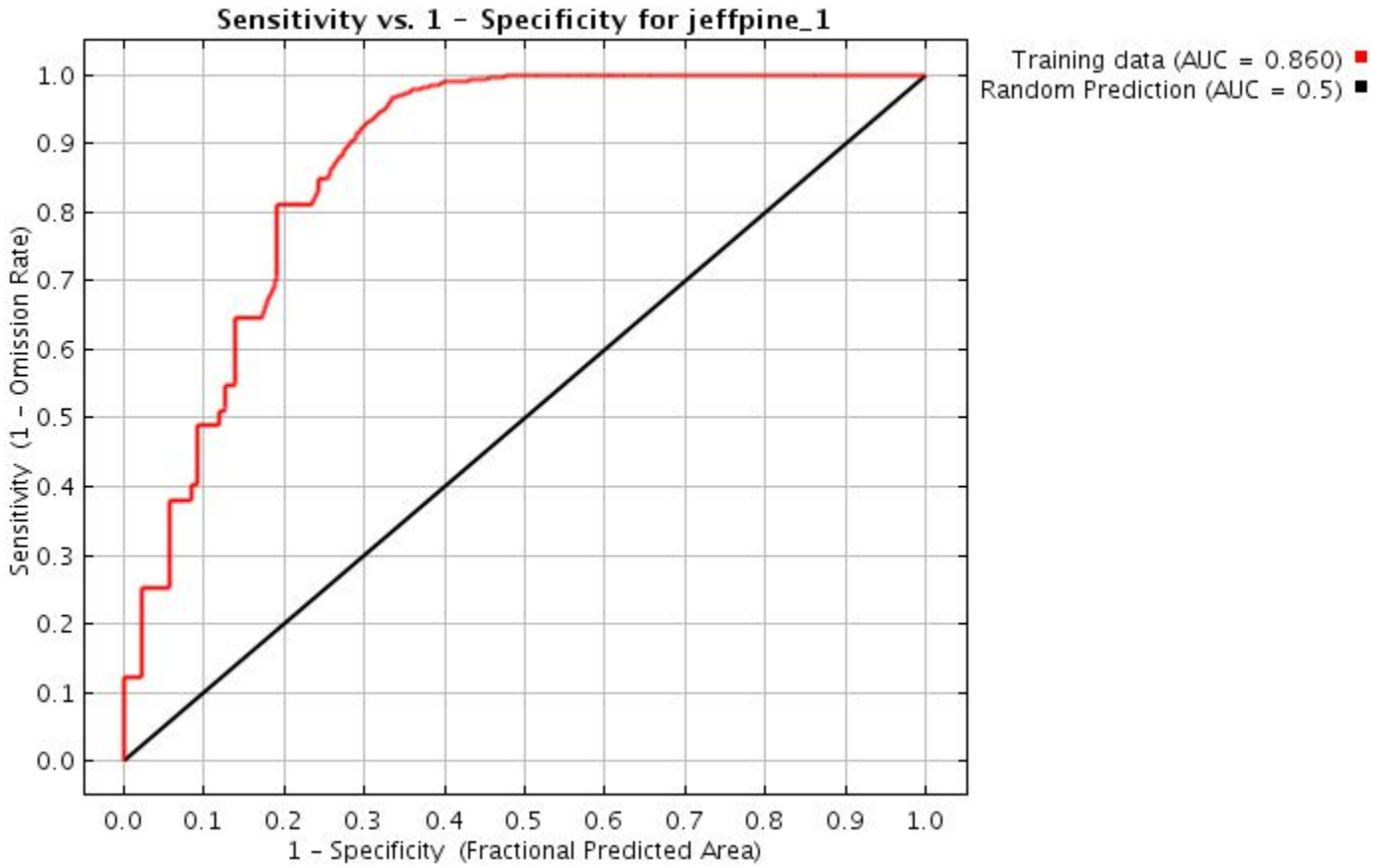
Experimental setup

Training a set of ML methods

- logistic regression (logit), multi-layer perceptron (mlp), support vector machines (lsvm), k-nearest neighbors (knn), random forest (rf), naive bayes (nb)
- k-fold cross validation with k=10 (10% test, 10% validation, 80% training)
- we run 50 different experiments with different initial conditions
- we compute feature importance with LIME[1]

1. LIME https://github.com/marcotcr/lime

Predictive performance - AUC ROC



Results, predictive performance AUC

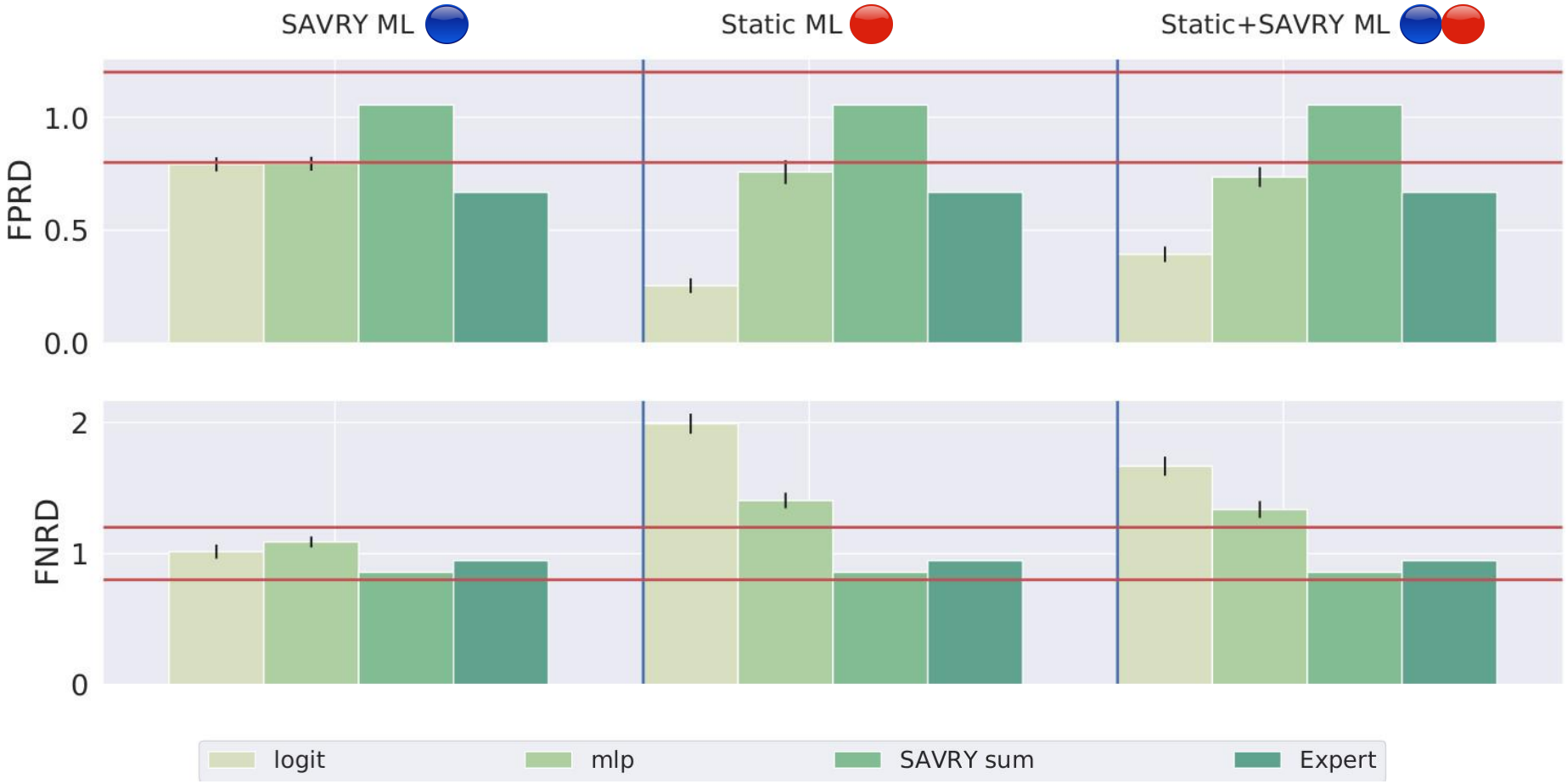
	logit		mlp		knn		lsvm		rsvm		nb		rf	
	mean	std.dev.	mean	std.dev.	mean	std.dev.	mean	std.dev.	mean	std.dev.	mean	std.dev.	mean	std.dev.
SAVRY ML	**.66**	.0058	**.66**	.0058	.60	.0121	.65	.0082	.52	.0197	.65	.0015	.65	.0110
Static ML	**.70**	.0055	**.70**	.0068	.62	.0122	.61	.0119	.56	.0149	.69	.0040	.66	.0110
Static+SAVRY ML	**.71**	.0064	**.70**	.0053	.64	.0129	**.71**	.0074	.50	.0058	.69	.0018	.69	.0121

SAVRY Sum has 0.64 AUC
Expert has 0.66 AUC

Results: disparity, sex

False alarm rates

Miss rates

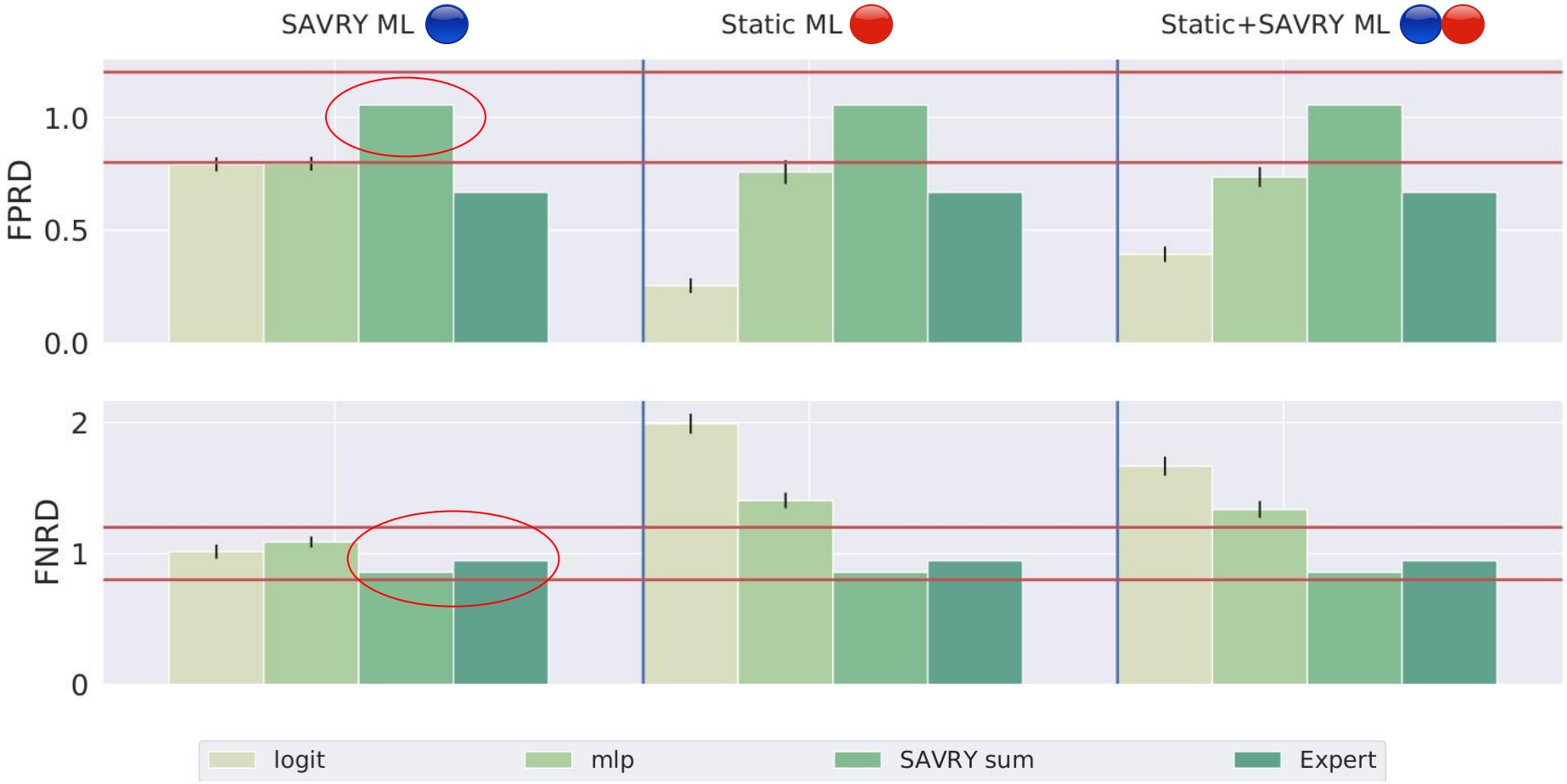


European Commission

Results: disparity, sex

False alarm rates

Miss rates

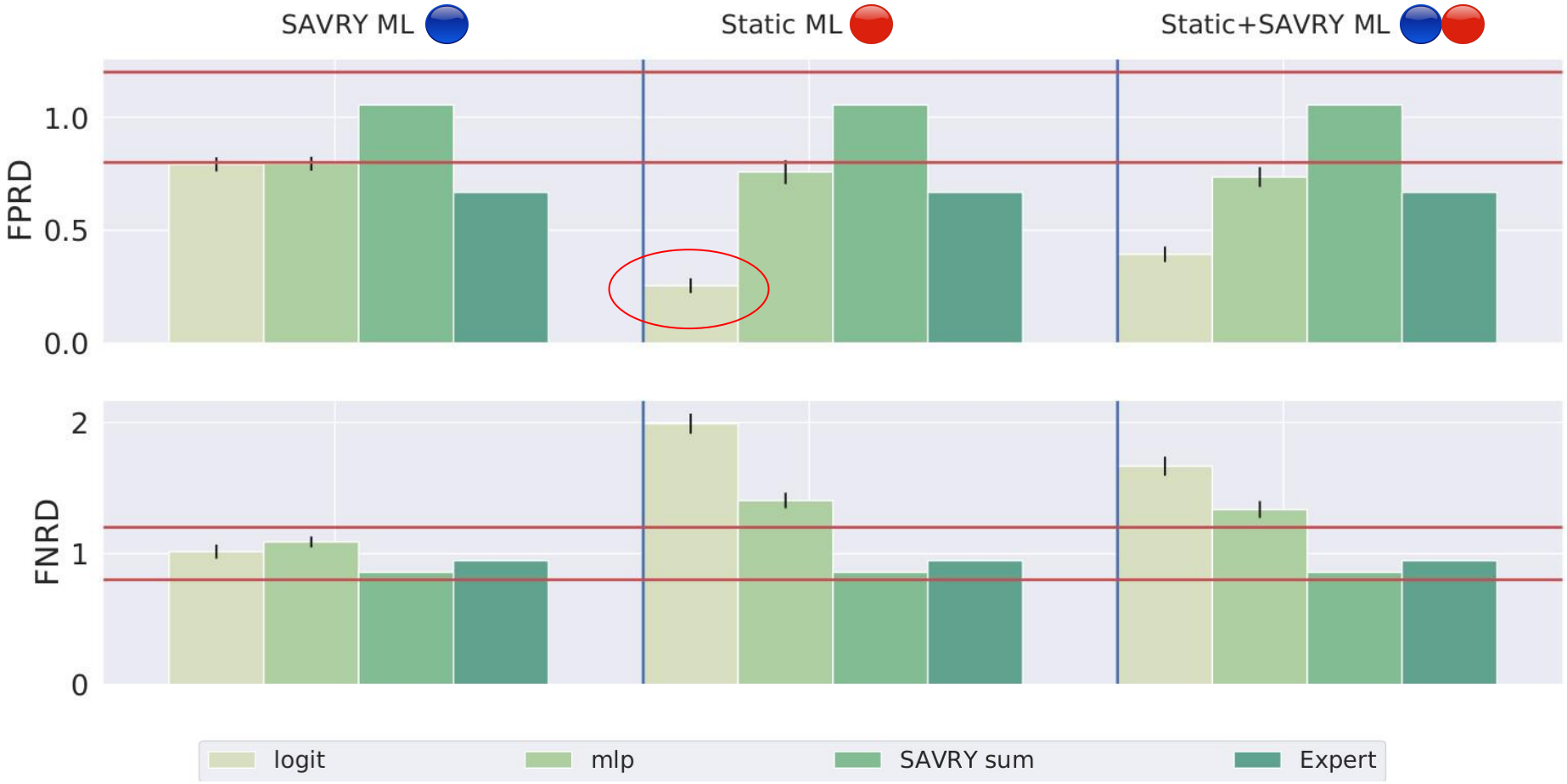


European Commission

Results: disparity, sex

False alarm rates

Miss rates

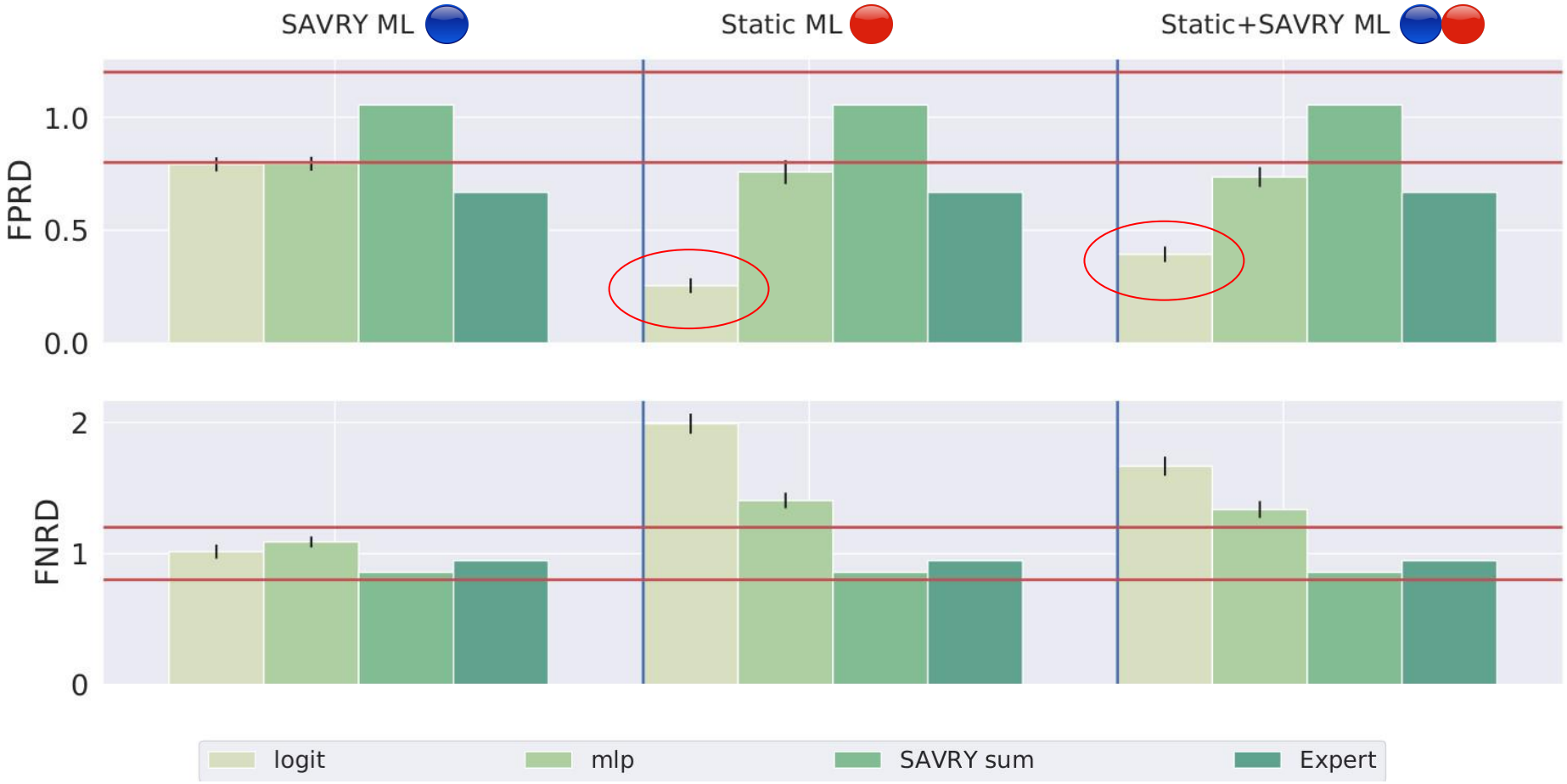


European Commission

Results: disparity, sex

False alarm rates

Miss rates

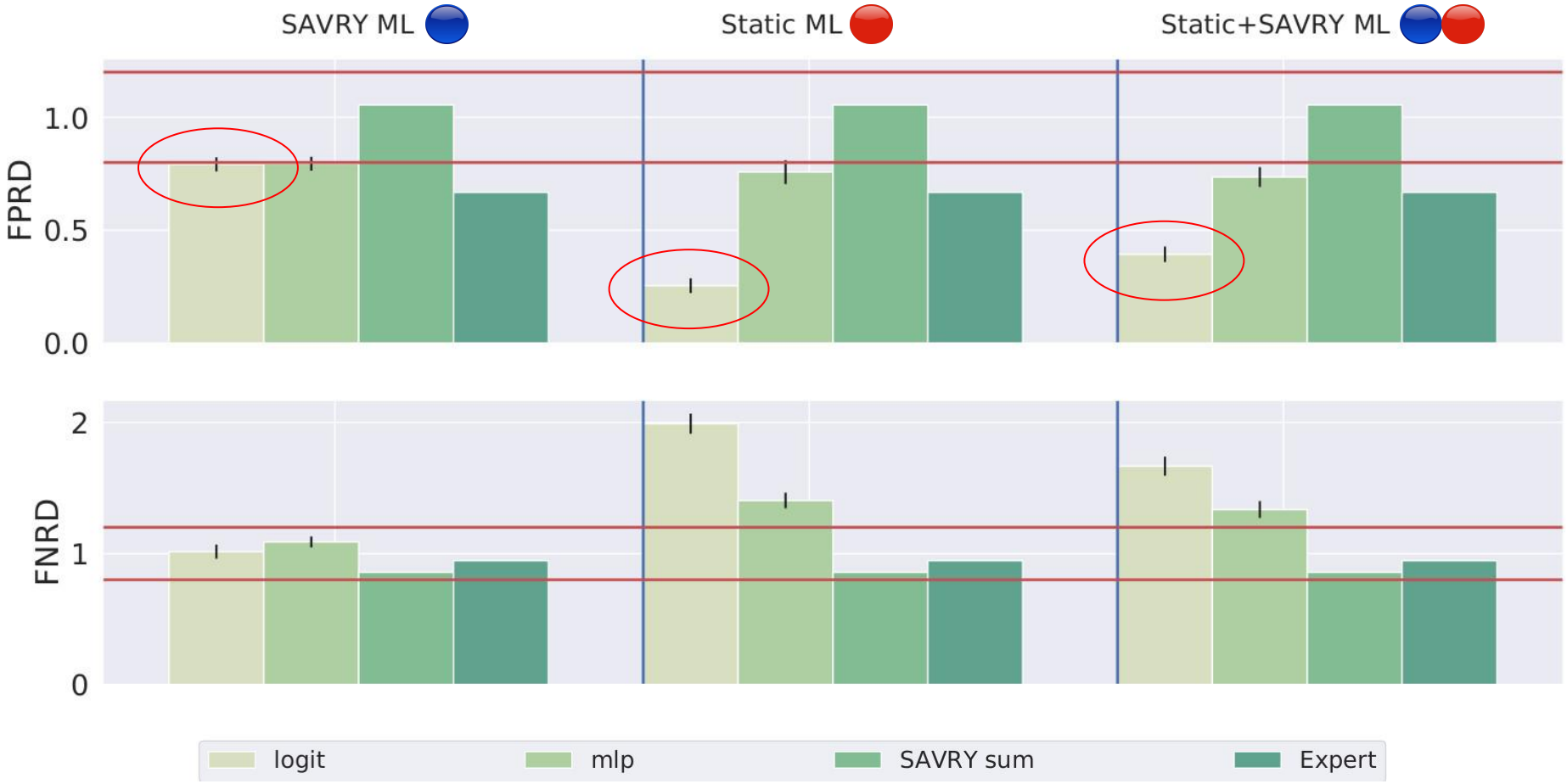


European Commission

Results: disparity, sex

False alarm rates

Miss rates

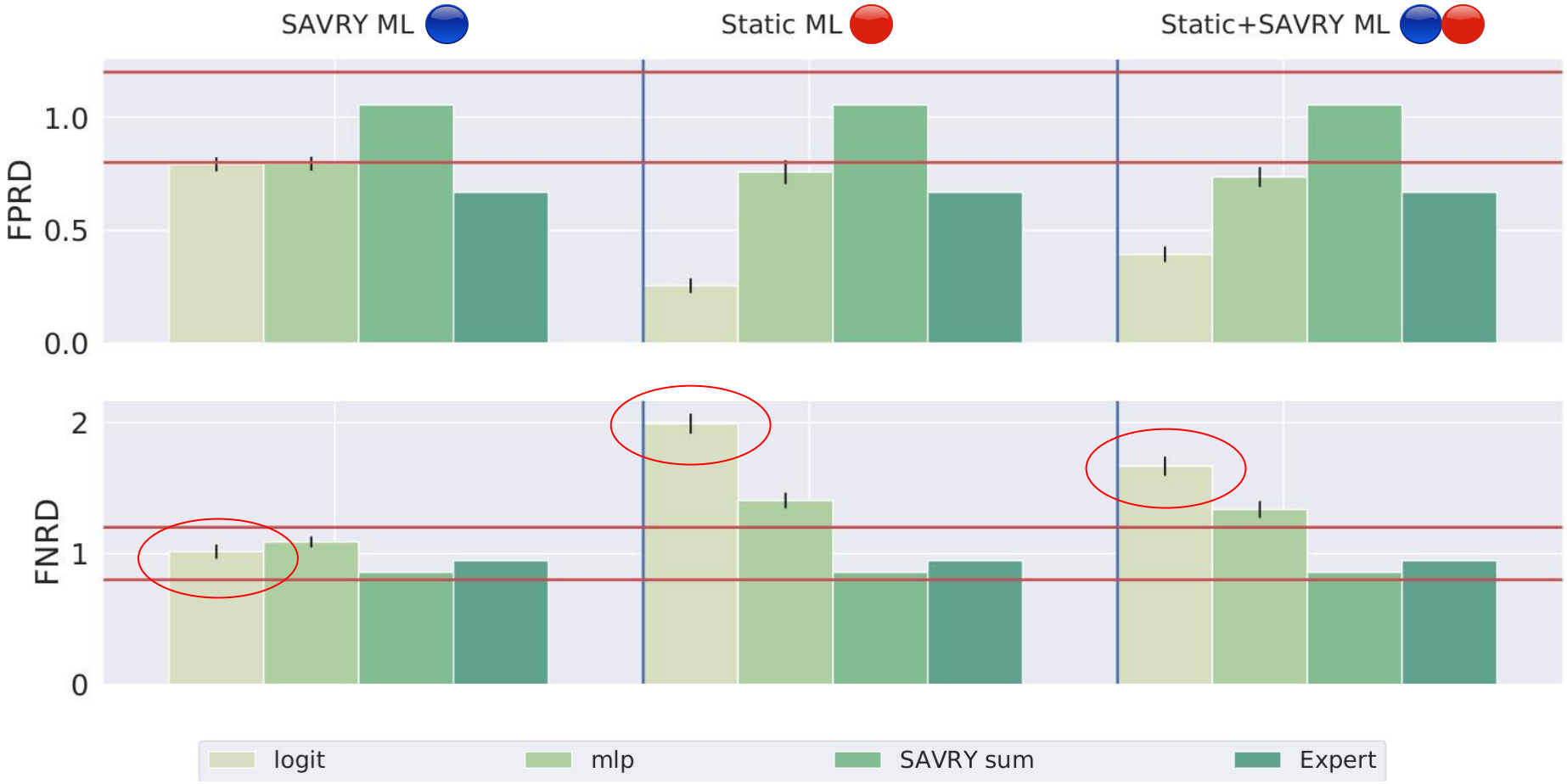


European Commission

Results: disparity, sex

False alarm rates

Miss rates

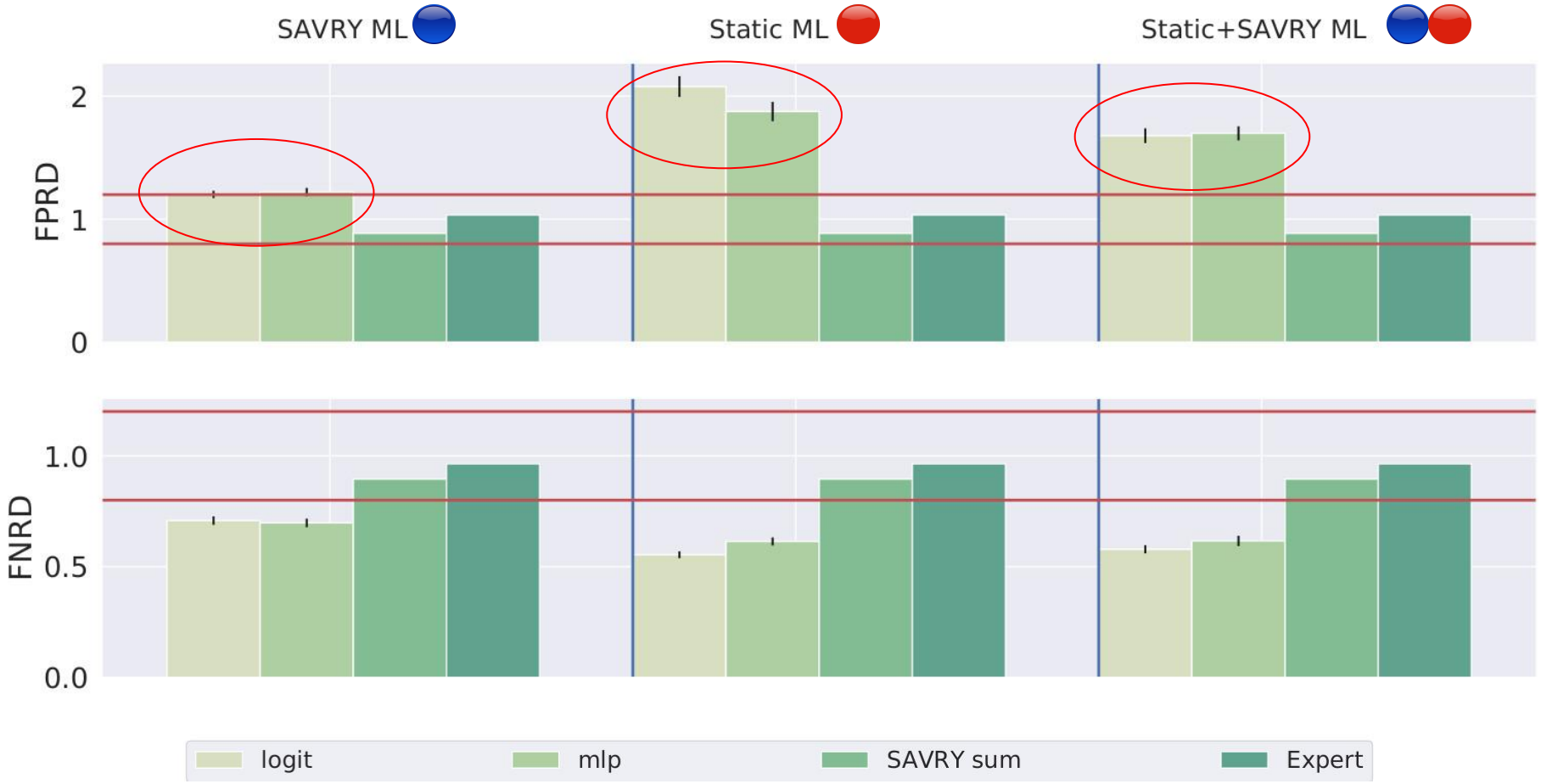


European Commission

Results: disparity, nationality

False alarm rates

Miss rates

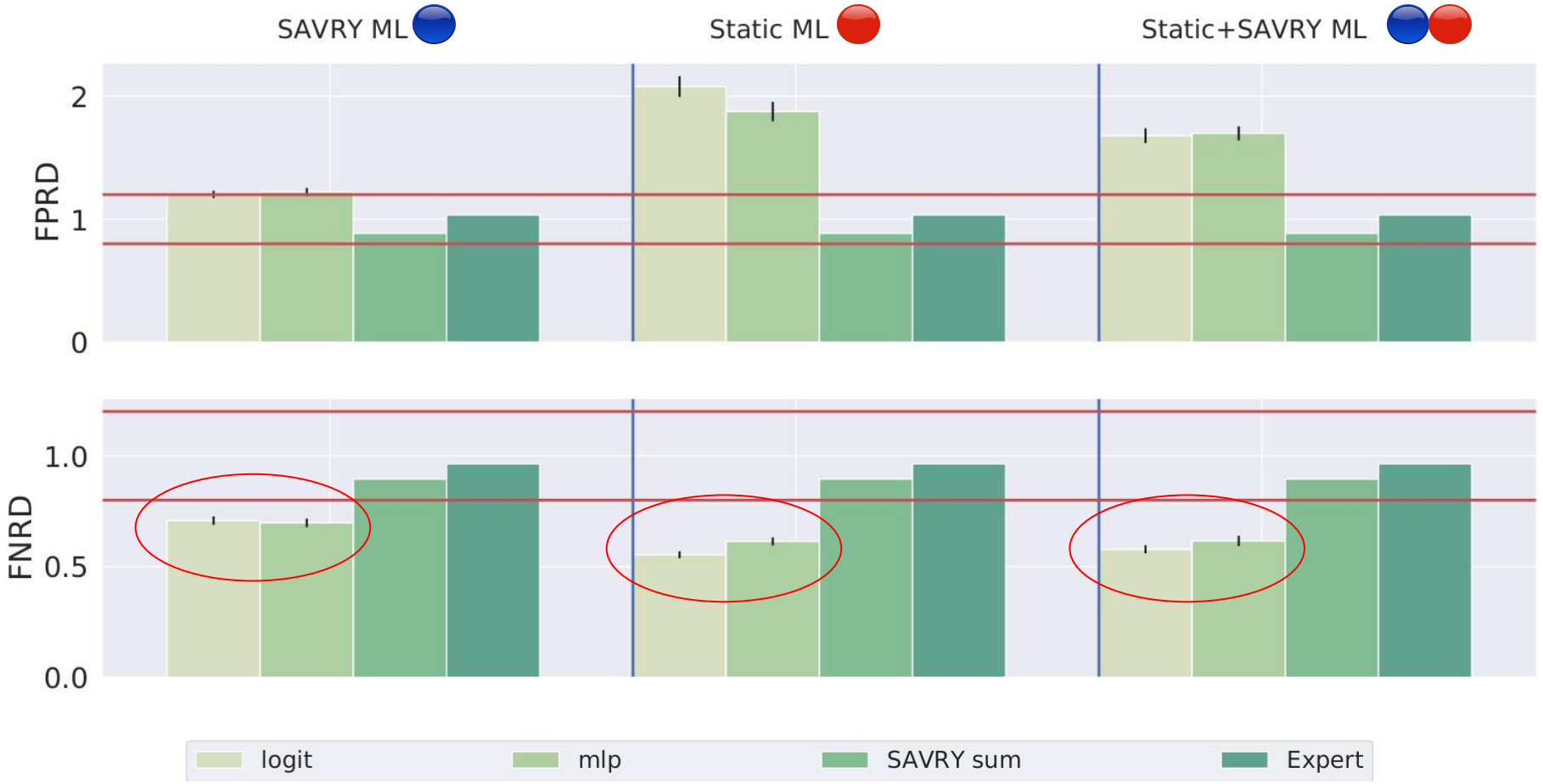


European Commission

Results: disparity, nationality

False alarm rates

Miss rates

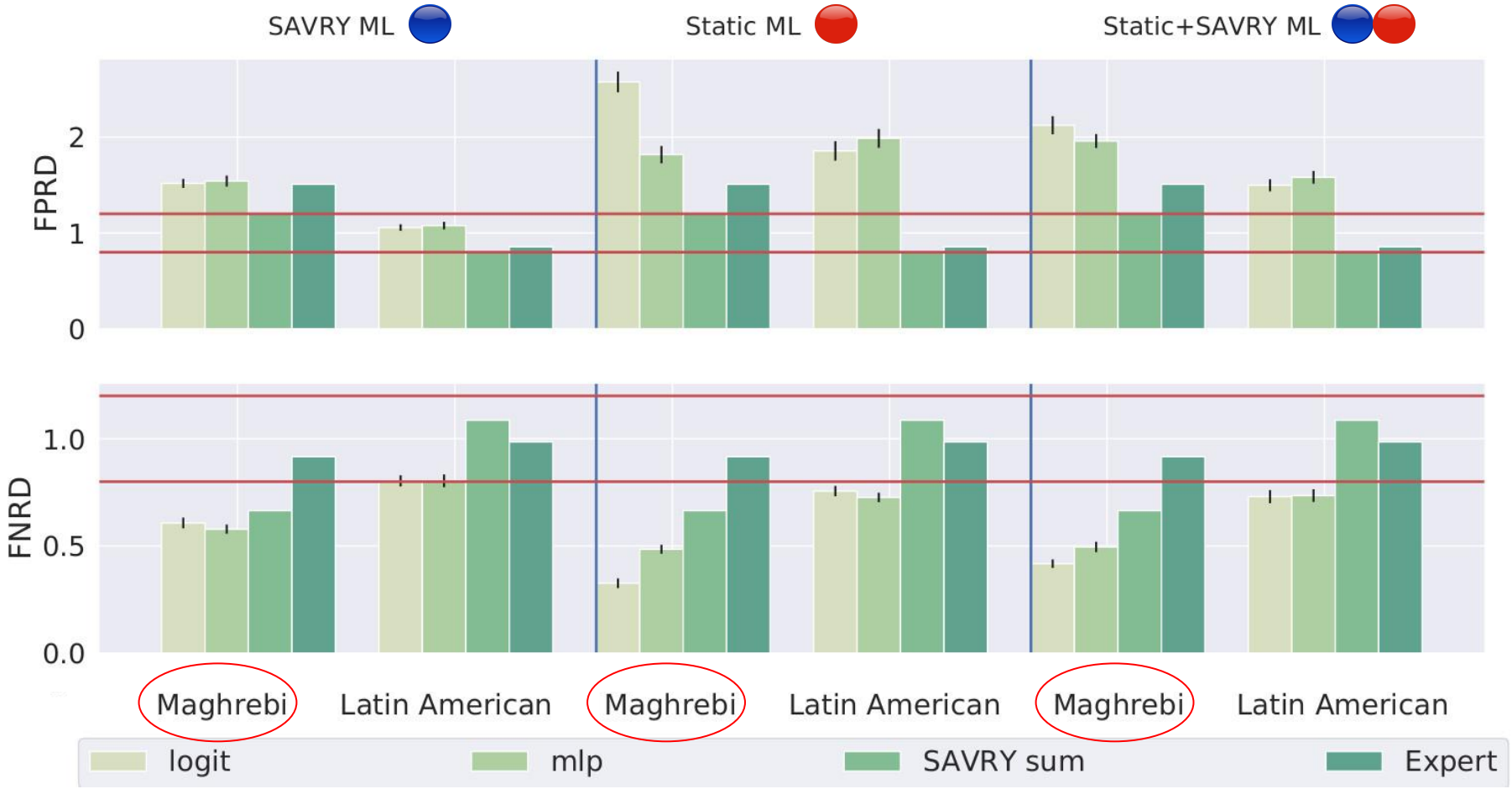


European Commission

Results: disparity, nationality

False alarm rates

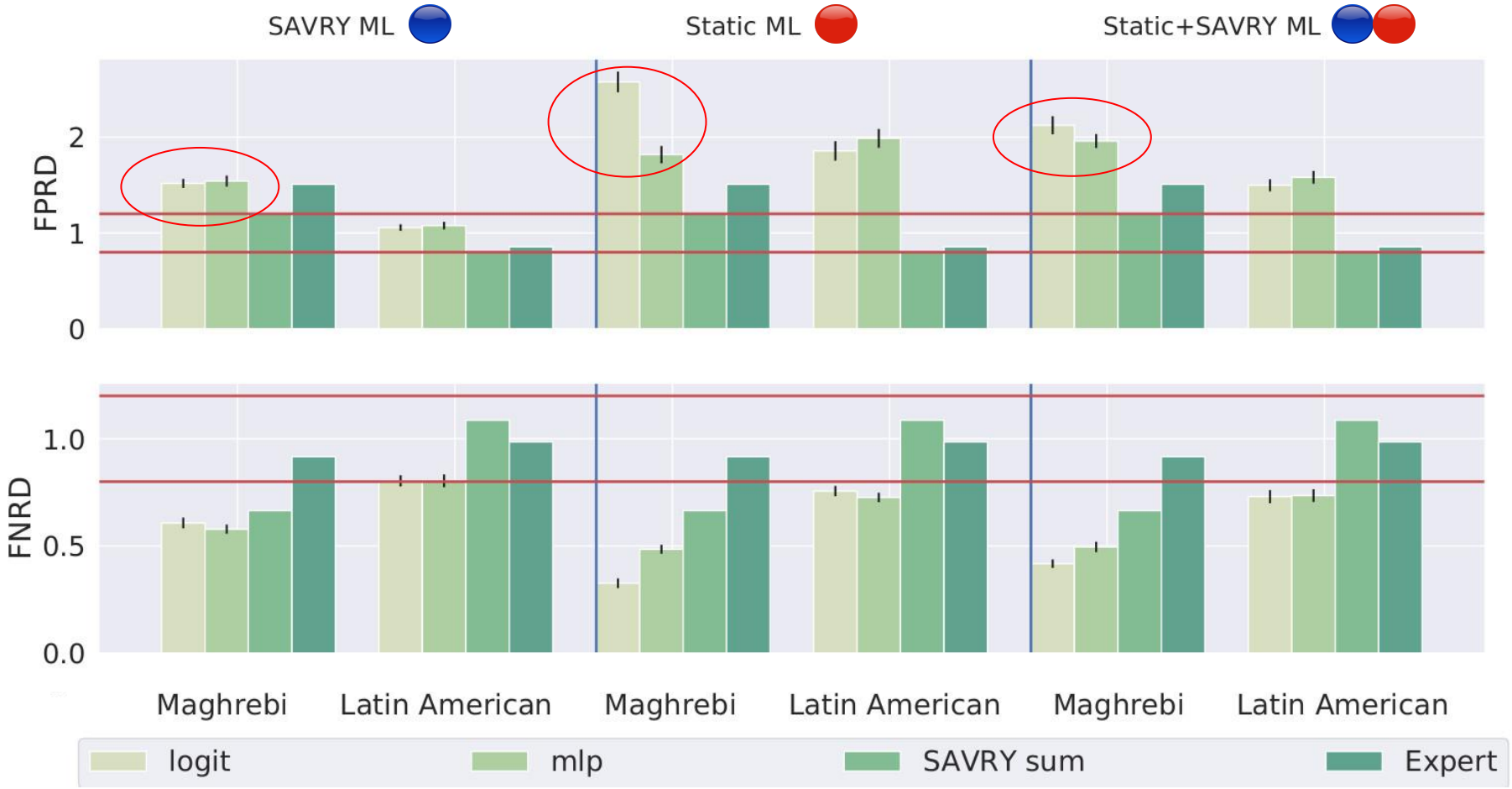
Miss rates



Results: disparity, nationality

False alarm rates

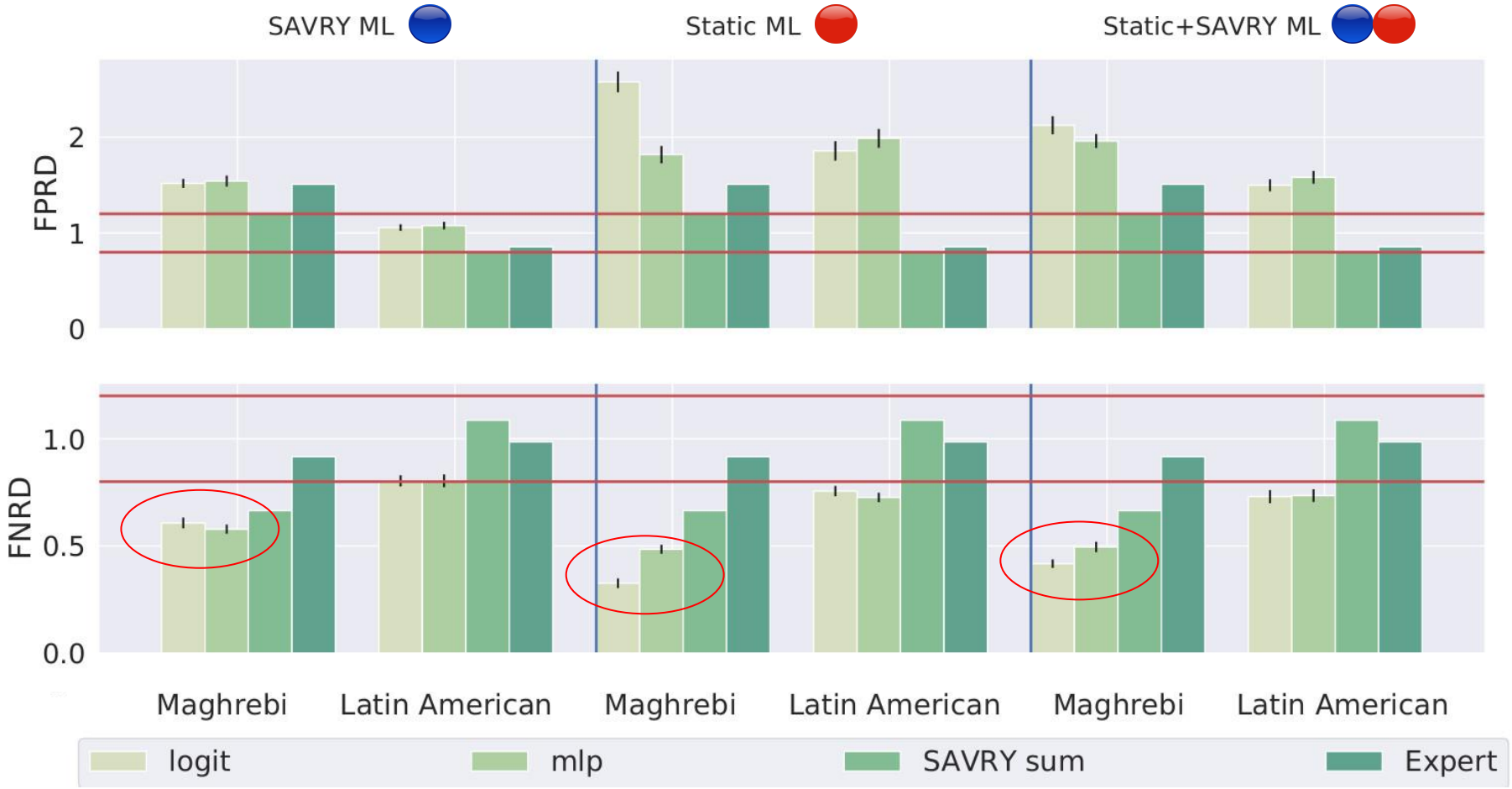
Miss rates



Results: disparity, nationality

False alarm rates

Miss rates



Results: feature importance for logit

SAVRY ML			Static			Static+SAVRY		
final expert evaluation	0.370***	(0.076)	✓crime in 07-08	-0.298**	(0.118)	✓crime in years 07-08	-0.272**	(0.133)
SAVRY sum	0.183	(0.910)	✓crime in year 09	-0.259**	(0.121)	✓crime in year 09	-0.255*	(0.132)
personality	-1.362	(7.061)	✓age maincrime	-0.109***	(0.021)	✓days to program start (norm)	-0.117***	(0.044)
treatment susceptibility	-1.340	(6.336)	✓days to program start (norm)	-0.105***	(0.040)	✓age maincrime	-0.115***	(0.022)
total score (social)	-0.141	(0.909)	✓crime in year 10	-0.275***	(0.098)	final expert evaluation	0.291***	(0.091)
total score (protective)	0.191	(0.902)	✓days in program (norm)	-0.087*	(0.048)	✓crime in year 10	-0.256**	(0.115)
previous violent offenses	-0.601	(2.533)	✓prog: enforcement measure	-0.248**	(0.103)	✓female	-0.196***	(0.053)
total score (historic)	0.056	(0.045)	✓prior crimes frequency	0.059*	(0.033)	✓enforcement measure	-0.206*	(0.122)
home violence	-0.543	(1.816)	✓female	-0.187***	(0.046)	✓Maghrebi	0.152**	(0.069)
past intervention failures	-0.598	(2.530)	✓Maghrebi	0.158***	(0.058)	✓Latin American	0.135**	(0.060)
			✓Latin American	0.105**	(0.052)			
			✓prog: mediation/reparation	-0.178*	(0.103)			

European Commission

Results: feature importance for mlp

SAVRY ML		Static ML		Static+SAVRY ML			
feature	importance	feature	importance	feature	importance		
	Mean	StdDev		Mean	StdDev		Mean	StdDev
probation/internment	147.43	24.85	✓province of residence	219.21	28.44	✓foreigner	199.80	11.37
total score (social)	117.93	9.71	✓age maincrime	202.83	25.72	✓sex	188.07	8.35
total score (personality)	117.63	9.83	✓foreigner	178.38	19.06	✓national group	117.40	23.09
total score (protective)	115.76	8.56	✓year of maincrime	168.96	13.86	✓maincrime category	150.90	16.44
total score (historic)	116.59	10.25	✓prior crimes	175.11	22.56	✓prior crimes frequency	151.53	18.26
history non-violent offending	112.17	7.44	✓national group	181.68	32.23	✓maincrime program sentence	143.29	10.50
positive/resilience characteristics	111.62	7.32	✓prior crimes frequency	156.15	20.98	✓year of maincrime	141.88	9
previous violence	113.22	8.93	✓maincrime category	144.27	18.26	✓maincrime violent	148.92	16.23
early violence	111.42	7.17	✓maincrime violent	137.20	14.95	✓province of execution	146.07	13.76
pro-social activities	109.82	5.57	✓prior crimes	131.53	12.66	✓prior crimes	146.97	14.71

Results: difference in base rates (prevalence)

	Base rate	Not Recidivated	Recidivated	Difference
protected features				
male	40.03%	0.839	0.931	0.093***
female	20.37%	0.161	0.069	-0.093***
Spanish	32.06%	0.667	0.523	-0.143***
foreign	46.22%	0.333	0.477	0.143**

Descriptive statistics of input features by recidivism status.

Results: difference in base rates

	Base rate	Not Recidivated	Recidivated	Difference
protected features				
male	40.03%	0.839	0.931	0.093***
female	20.37%	0.161	0.069	-0.093***
Spanish	32.06%	0.667	0.523	-0.143***
foreign	46.22%	0.333	0.477	0.143**

Descriptive statistics of input features by recidivism status.

Results: difference in base rates

	Base rate	Not Recidivated	Recidivated	Difference
protected features				
male	40.03%	0.839	0.931	0.093***
female	20.37%	0.161	0.069	-0.093***
Spanish	32.06%	0.667	0.523	-0.143***
foreign	46.22%	0.333	0.477	0.143**

Descriptive statistics of input features by recidivism status.

Conclusions

- ML models have better predictive performance
- ML models tend to discriminate more
- static features outweigh SAVRY features as importance
- preliminary study: the cause may be in the data (base rates)

Contributions

We propose a methodology and a ML framework[1]

- to easily train ML models on tabular data (csv files)
- to evaluate these models in terms of predictive performance and fairness
- to connect to interpretability frameworks
- to reproduce with ease results and research

1. Open framework: https://gitlab.com/HUMAINT/humaint-fatml

Thank you!

Any questions?

You can find me at @nkundiushuti & marius.miron@ec.europa.eu & mariusmiron.com

European Commission