Fairness in Machine Learning

J.-M. Loubes, Professor Artificial & Natural Intelligence Institute of Toulouse

Part 1: Principles of Machine Learning

Big Data paradigm

- The Data convey all the information.
- A model built to fit the data can describe the reality.
- The more the data the more accurate the description of the reality.

→ From data to information: extraction of the knowledge from empirical observations

Principle of Machine Learning

- Learn decision rules fitting the data using a *set of labeled examples* (*learning sample*).
- The learned decision rules will be used for *all the population.*
- The whole population is supposed to follow same distribution as the learning sample.

→ The Machine Learning algorithm (or **AI**) learn the best rule from the data and then can forecast for new observations with a guaranteed precision.

Goal

- Learning the relationships between characteristic variables X and a target variable Y.
- Being then be able to forecast new observations.

Learning Sample

I.i.d. observations with unknown distribution $\mathbb{P}$: $(Y_1, X_1), \ldots, (Y_n, X_n)$.

The rule will be learnt from the empirical distribution $\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i, Y_i}$.

Loss function

Quantify the error made while predicting $\hat{Y}$ while Y is observed: $\ell : (Y, \hat{Y}) \mapsto \ell(Y, \hat{Y}) \in \mathbb{R}^+$

Machine Learning Algorithm

$\hat{f}_n$ aims to learn the best model among a class of algorithms $\mathcal{F}$, based on the observations:

$$\hat{f}_n \in \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i)) \right\}$$

Goal

- Learning the relationships between characteristic variables X and a target variable Y.
- Being then be able to forecast new observations.

Learning Sample

I.i.d. observations with unknown distribution $\mathbb{P}$: $(Y_1, X_1), \ldots, (Y_n, X_n)$.

The rule will be learnt from the empirical distribution $\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i, Y_i}$.

Loss function

Quantify the error made while predicting $\hat{Y}$ while Y is observed: $\ell : (Y, \hat{Y}) \mapsto \ell(Y, \hat{Y}) \in \mathbb{R}^+$

Machine Learning Algorithm

$\hat{f}_n$ aims to learn the best model among a class of algorithms $\mathcal{F}$, based on the observations:

$$\hat{f}_n \in \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i)) + \lambda \text{penalty}(f) \right\}$$

where λ balances the contributions of the **loss** and the **penalty terms** to get a trade-off.

Unknown oracle rule

$$f^\star \in \arg\min_{f \in \mathcal{F}} \mathbb{E}_{\mathbb{P}}\{\ell(Y, f(X))\}$$

is estimated by the $\hat{f}_n$.

→ Optimized from a mathematical point of view in the sense that

$$\mathbb{E}_{\mathbb{P}}\{\ell(Y, \hat{f}_n(X))\} - \mathbb{E}_{\mathbb{P}}\{\ell(Y, f^\star(X))\}$$

is small

→ Optimal forecast model **reproduces the properties learnt from the learning set** and transformed into a global rules for new observations:

$$\widehat{Y} = \hat{f}_n(X)$$

As a consequence

- Properties of the M.L. algorithm highly dependent of the observations distribution.
- Warning 1: The learning sample may be **biased**.
- Warning 2: The learning sample may **not reflect** the desired behavior of the model.

A.I. (or more specifically M.L. algorithms) **generalizes** the situation encountered in the learning sample to the whole population.
→ It shapes the reality according to the learnt rule without questioning nor evolution.

"*It's the mathematics, stupid*".
→ Difficult to argue with an expert A.I. to understand a decision.
→ Mathematics can not be questioned so the decisions taken can be **justified** using a scientific argument.

Key question

Correlations and bias in the dataset enable to forecast ... but are these biases due to a model or due to a poor data sampling?

What now?

Crucial Need for new insights: **Acceptability of A.I.** requires that the prediction algorithms behave in a **fair and explainable** way for all people.

Part 2: A first definition of Bias and Fairness in Machine Learning

- An A.I. algorithm suffers from **unfairness** if its outcomes Y (decisions) are fully or partly based on a variable S that *should* not play a decisive role in the decision making process.

- Variable S is denoted by **sensitive attribute.**
  → It divides the observations into subgroups (e.g.: Males/Females).
  → The prediction algorithm should not show a different behavior over these subsets.
  → If the algorithm does *not depend* on S, it is considered as **fair**.
  → The variable S is chosen by the practitioner. Its choice is driven by legal, ethic or technical concerns.

We assume that the algorithm is not meant to be unfair, i.e. it is not **unfair by design**. The possible unfairness comes from the **learning process** in a machine learning framework.

Target

$$Y = \begin{cases} 0 \rightarrow \text{Failure} \\ 1 \rightarrow \text{Success} \end{cases}$$

Protected attribute

$$S = \begin{cases} 0 \rightarrow \text{Unfavored class / minority} \\ 1 \rightarrow \text{Favored class / majority} \end{cases}$$

The M.L. algorithm should not favor one group ($S = 1$) over the other ($S = 0$).

Target

$$Y = \begin{cases} 0 \rightarrow \text{Failure} \\ 1 \rightarrow \text{Success} \end{cases}$$

Protected attribute

$$S = \begin{cases} 0 \rightarrow \text{Unfavored class / minority} \\ 1 \rightarrow \text{Favored class / majority} \end{cases}$$

The M.L. algorithm should not favor one group ($S = 1$) over the other ($S = 0$).

Measuring the fairness using the *Disparate Impact (D.I.)*

- *Fairness* of a classifier g

$$DI(g, X, S) = \frac{\mathbb{P}(g(X) = 1 \mid S = 0)}{\mathbb{P}(g(X) = 1 \mid S = 1)}$$

where
- g is fully fair if: $DI(g, X, S) = 1$.
- g is partially fair w.r.t. a threshold D.I. at level $\tau \in (0,1]$ if: $DI(g, X, S) > \tau$
- A common threshold is $\tau = 0.8$

- *Bias* in a dataset

$$DI(Y, X, S) = \frac{\mathbb{P}(Y = 1 \mid S = 0)}{\mathbb{P}(Y = 1 \mid S = 1)}$$

Target

$$Y = \begin{cases} 0 \to \text{Failure} \\ 1 \to \text{Success} \end{cases}$$

Protected attribute

$$S = \begin{cases} 0 \to \text{Unfavored class / minority} \\ 1 \to \text{Favored class / majority} \end{cases}$$

The M.L. algorithm should not favor one group ($S = 1$) over the other ($S = 0$).

Measuring the fairness using the *Disparate Impact (D.I.)*

- *Fairness* of a classifier g

$$DI(g, X, S) = \frac{\mathbb{P}(g(X) = 1 \mid S = 0)}{\mathbb{P}(g(X) = 1 \mid S = 1)}$$

where
- g is fully fair if: $DI(g, X, S) = 1$.
- g is partially fair w.r.t. a threshold D.I. at level $\tau \in (0,1]$ if: $DI(g, X, S) > \tau$
- A common threshold is $\tau = 0.8$

- *Bias* in a dataset

$$DI(Y, X, S) = \frac{\mathbb{P}(Y = 1 \mid S = 0)}{\mathbb{P}(Y = 1 \mid S = 1)}$$

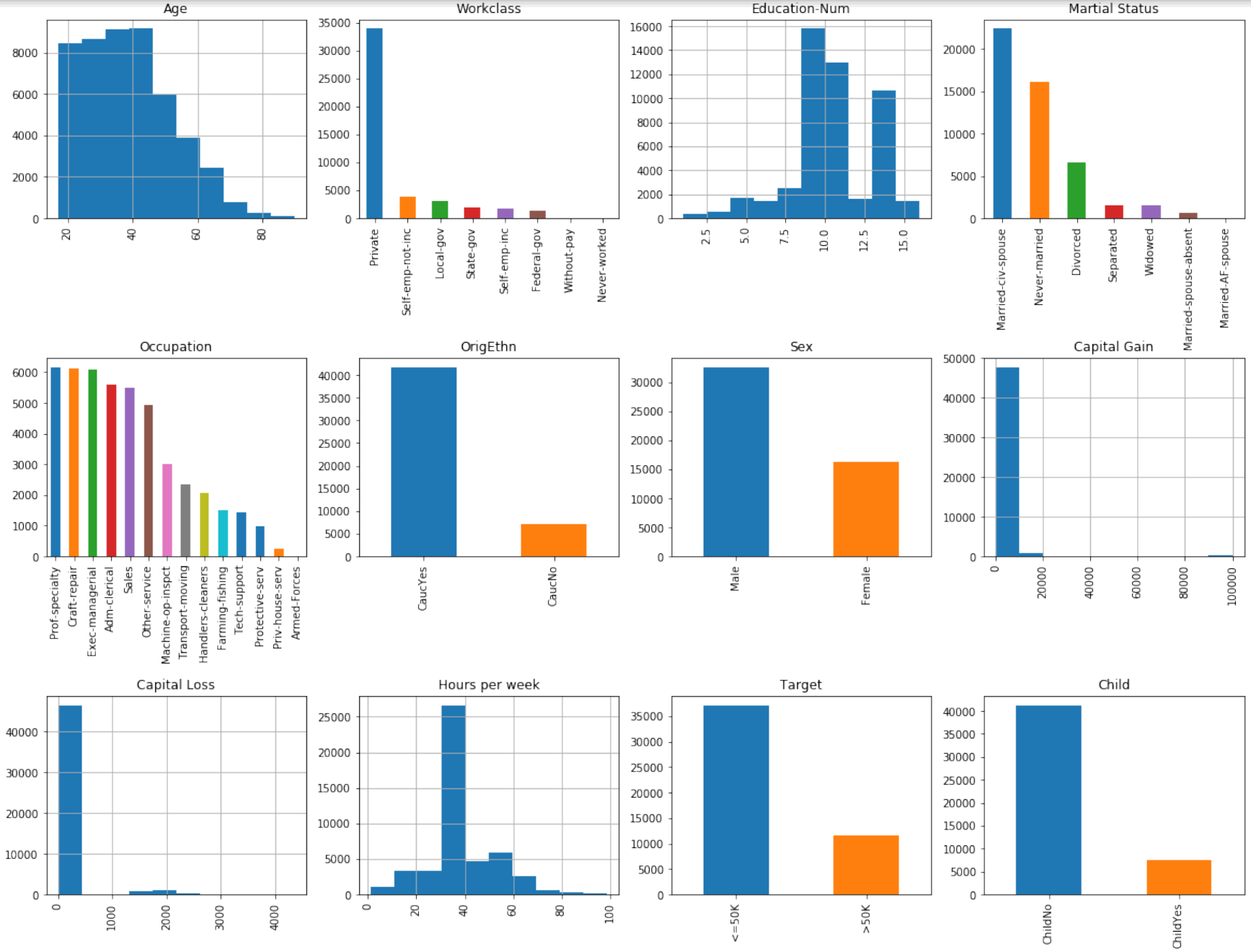
Remark: A classifier should **not increase the bias** already present in a data set.

Consider the probability space $\left(\Omega \subset \mathbb{R}^d, \mathscr{B}, \mathbb{P}\right)$, with $\mathscr{B}$ the Borel $\sigma-$algebra of subsets of $\mathbb{R}^d$ and $d \geqslant 1$.

Notations

- $Y : \Omega \to \{0,1\}$ is the **target class** *i.e.* $Y = \begin{cases} 0 \to & \text{Failure} \\ 1 \to & \text{Success} \end{cases}$
- $X : \Omega \to \mathbb{R}^d$, $d \geqslant 1$ are the **visible attributes**
- $\mathscr{G}$ is a family of **binary classifiers** *i.e.* $g : \mathbb{R}^d \to \{0,1\}$
- $\hat{Y} = g(X)$, $g \in \mathscr{G}$ is the **outcome of the classification**
- $S : \Omega \to \{0,1\}$ is the **protected attribute** *i.e* $S = \begin{cases} 0 \to & \text{Unfavored class} \\ 1 \to & \text{Favored class} \end{cases}$

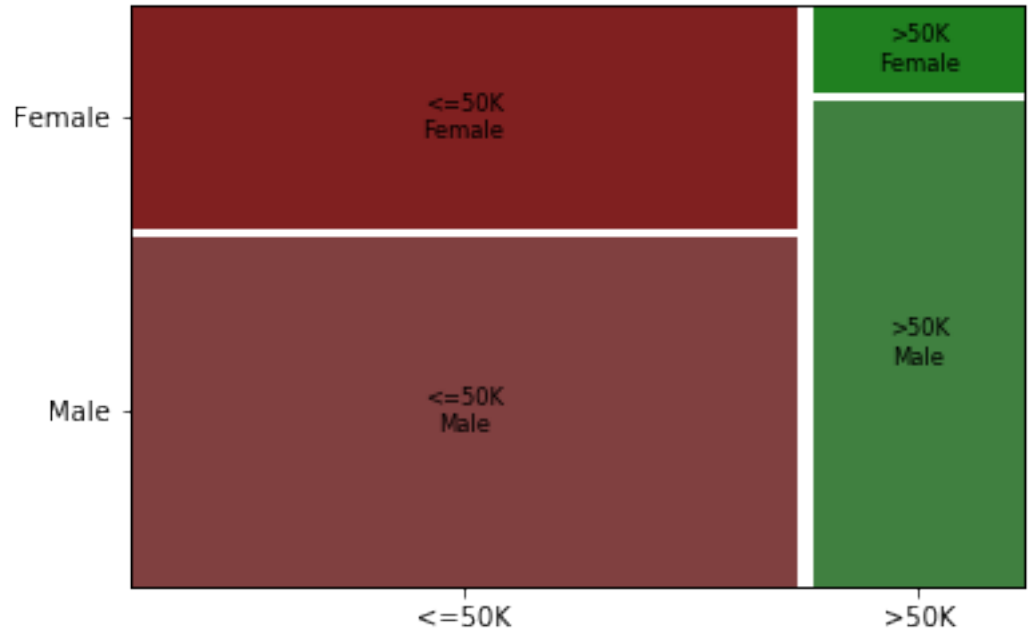
2.4: Illustration on the *Adult Income* dataset — Data

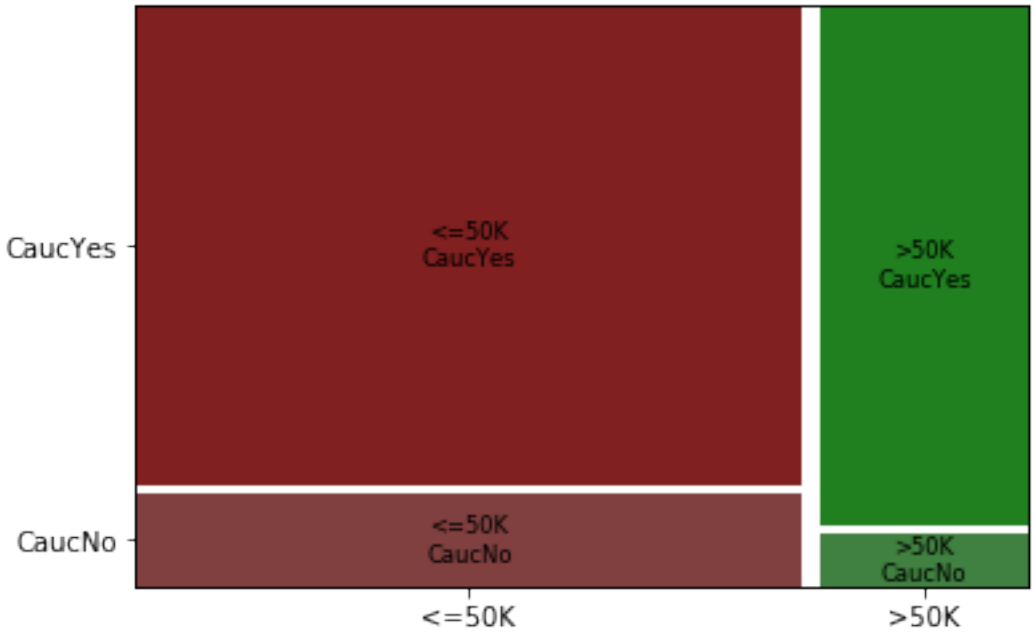


$n = 48842$ observations (individuals) described by $p = 14$ variables

Goal: Predicting the solvability (salary $> 50k\$$) in order to minimize the risk of granting a loan

Issue: The learning set is biased with respect to **Sex** and **Ethnic Origin**.





Tested M.L. classifiers

- Logistic Regression (*scikit-learn*)
- Decision Trees (*scikit-learn*)
- Extreme Gradient Boosting (*lightgbm*)

Average results (10-folds cross validation)

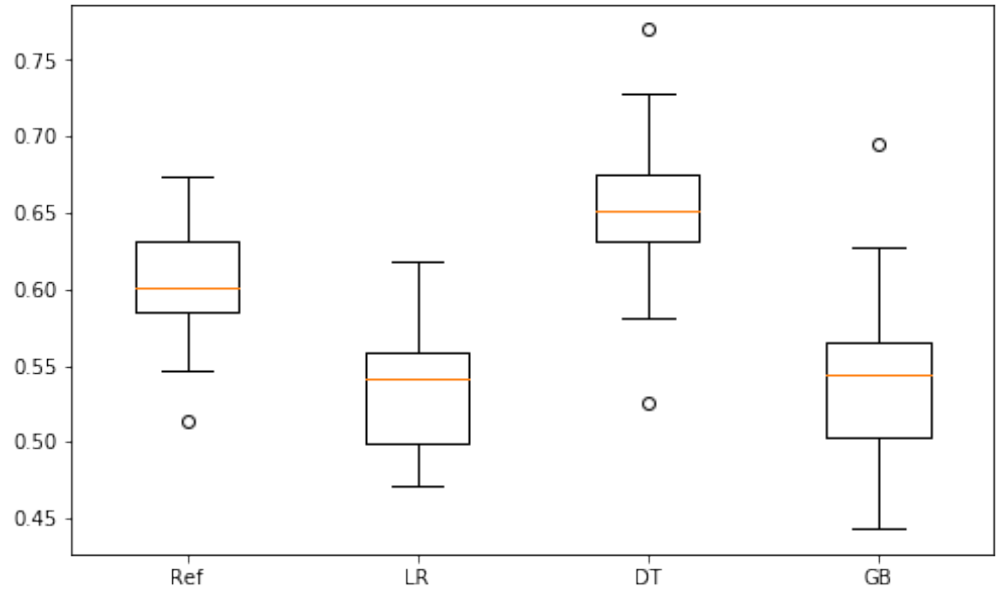
ML model	Accuracy	TP rate	TN rate	CM
Logistic regresion	0.8500	0.7285	0.8792	$\begin{pmatrix} 11428 & 847 \\ 1570 & 2273 \end{pmatrix}$
Decision Tree	0.8491	0.7723	0.8638	$\begin{pmatrix} 11685 & 590 \\ 1842 & 2001 \end{pmatrix}$
Gradient Boosting	0.8624	0.7579	0.8878	$\begin{pmatrix} 11512 & 763 \\ 1455 & 2388 \end{pmatrix}$

Accuracy of the M.L. classifiers

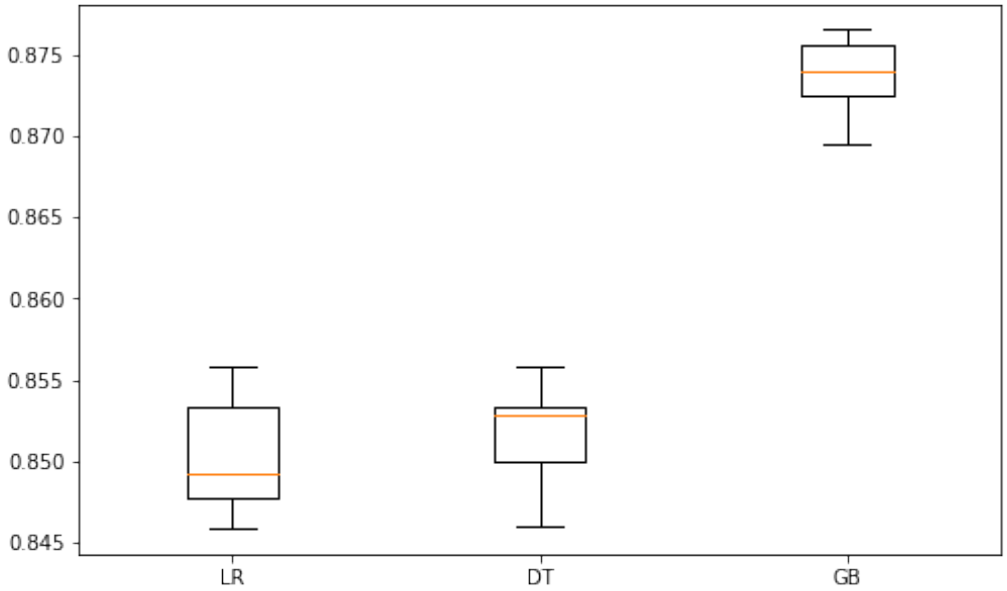
where:

- TP: True Positive
- TN: True Negative
- CM: Confusion Matrix

Disparate Impacts obtained with the *ethnic origin* as variable S



Disparate Impacts



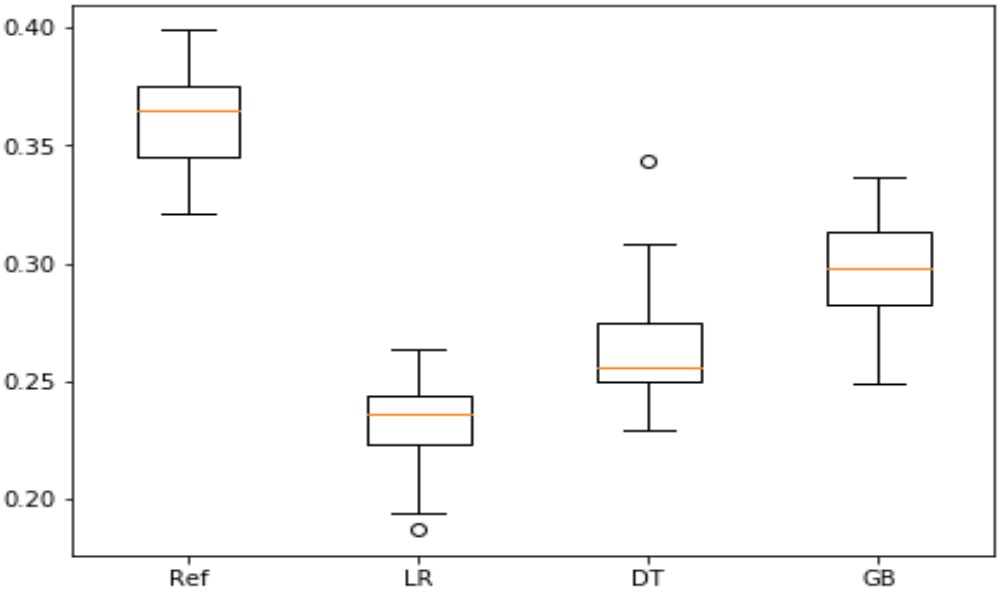
Accuracies

where the Y or $\hat{Y}$ are:

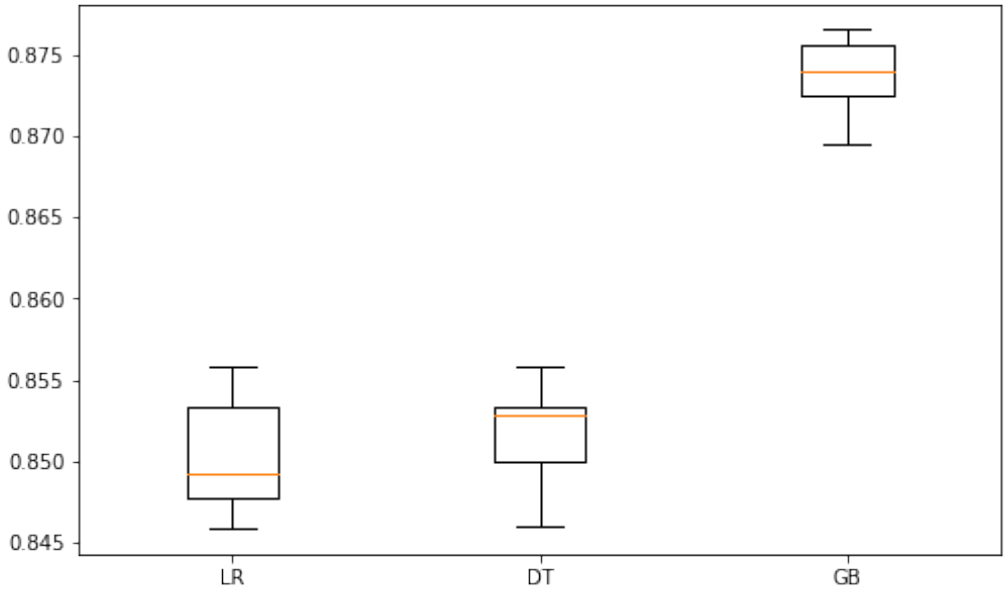
- Ref: The true observed data
- LR: The predictions made using Logistic Regression
- DT: The predictions made using Decision Trees
- GB: The predictions made using Extreme Gradient Boosting

Unbalanced samples do not always imply a bias in the M.L. algorithm

Disparate Impacts obtained with the variable *Sex* as S



Disparate Impacts



Accuracies

where the Y or $\hat{Y}$ are:

- Ref: The true observed data
- LR: The predictions made using Logistic Regression
- DT: The predictions made using Decision Trees
- GB: The predictions made using Extreme Gradient Boosting

- Statistical Significant Improvement of Unfair Treatment between Male and Female
- Demographic Parity: significant difference in loan granting between males and females.

General Data Protection Regulation (GDPR)

- Effective in the E.U. since 05/2018
- According to the GDPR, automatic decisions taken by an algorithm should be:
  - *un-biased*
  - *not discriminant*
  - *fair*
  - *explainable*
  - ...

More generally

- E.U. (GDPR, art 22-4 2018): "A decision is declared fair if it is neither based on affiliation to a protected minority group, nor based on the explicit or implicit knowledge of sensitive personal data."
- NYC Bill (Dec. 2017) : local decision
- Several Trials (USA-Canada)
- …

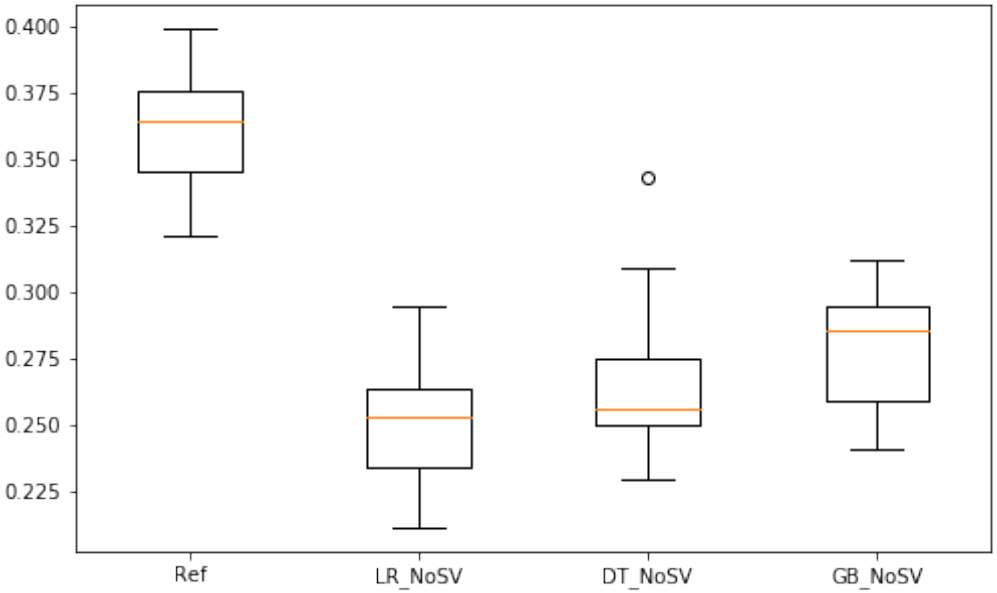
General recommendation

One should not use the sensitive variable S

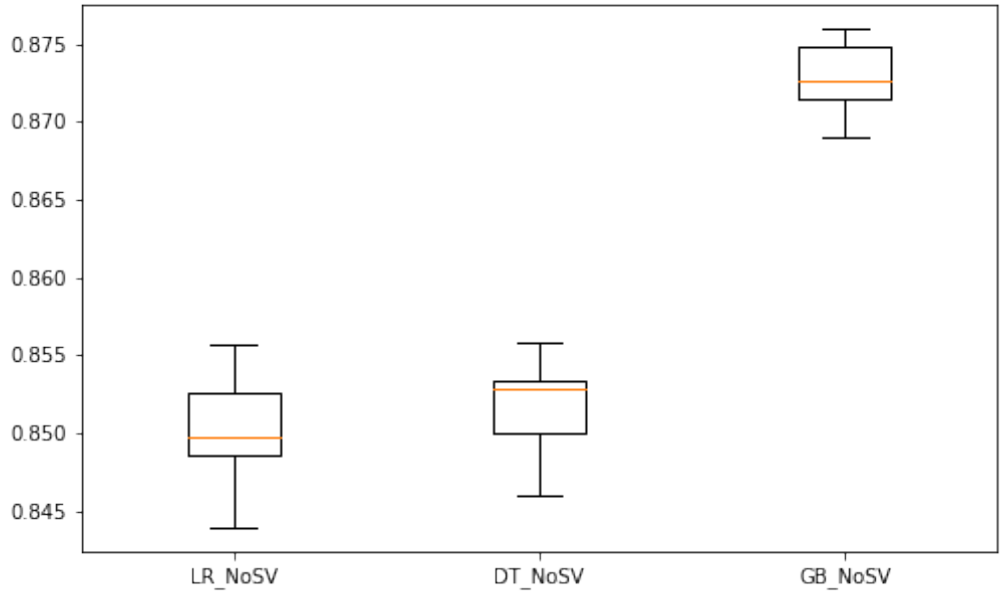
… Let's testing the effectiveness of this recommendation!

2.4: Illustration on the *Adult Income* dataset — Removing the *Sex* variable

Sensitive variable *Sex (here* S*)* is not used when training the models



Disparate Impacts



Accuracies

The bias is still present → It comes from the correlations with the S variable that are present in the dataset (even after training without the variable S).

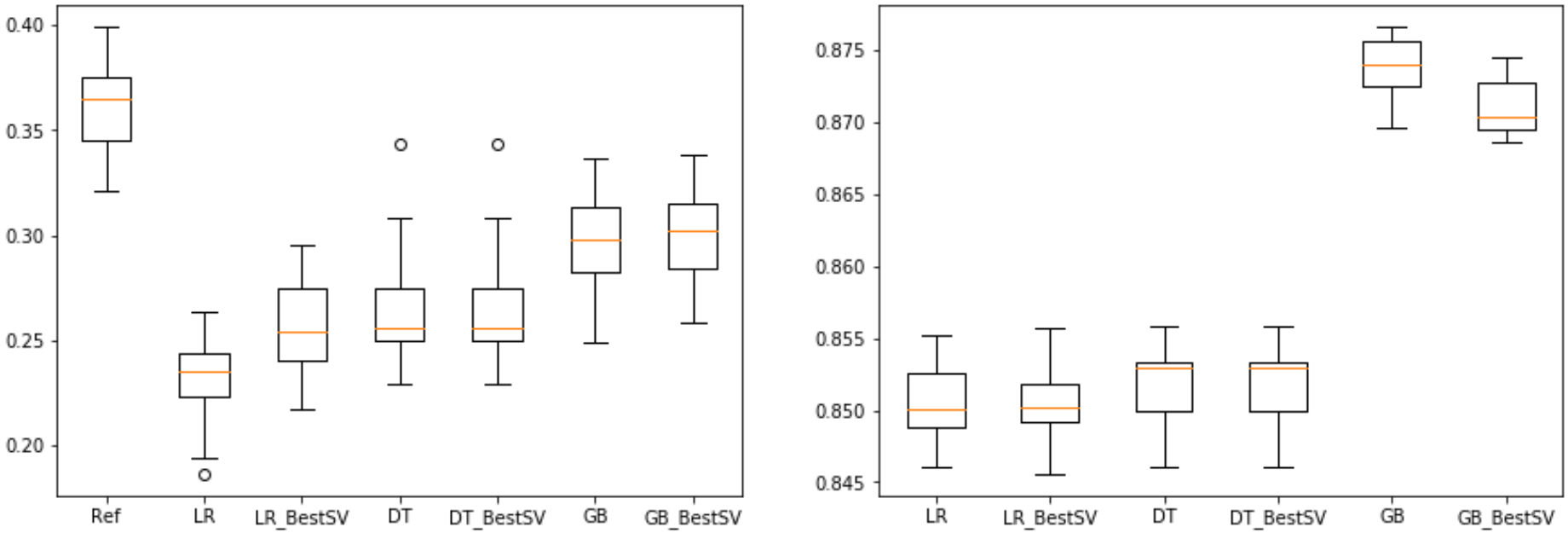
Legal proof for individual discrimination

- Creating an artificial individual with the *same characteristics* except for S *which is changed*
- Checking whether the output is changed.

Technical solution

- Using: $\hat{f}_{\text{Best}}(X) = \max(\hat{f}(X, S = 1), \hat{f}(X, S = 0))$.

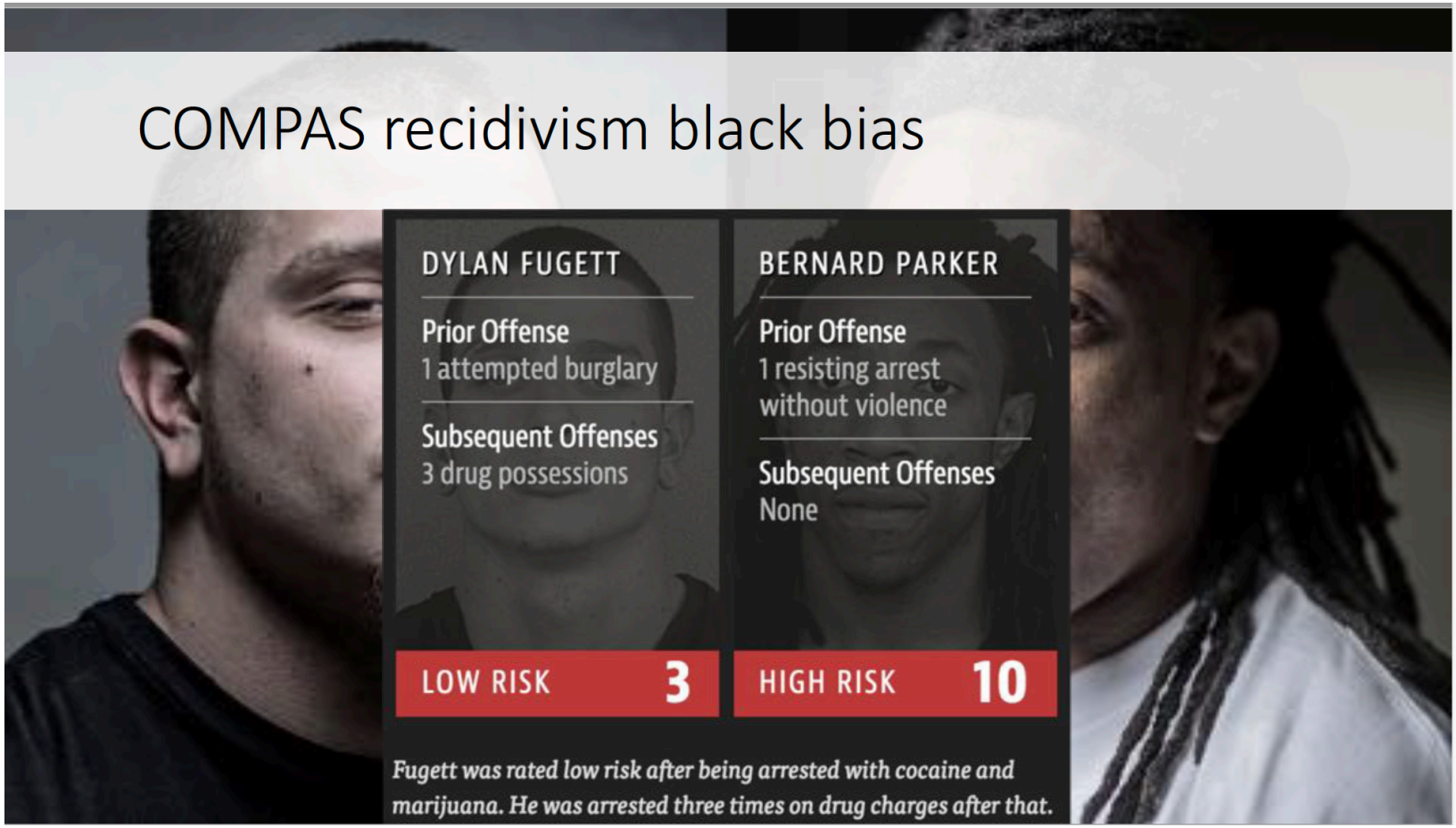
Results



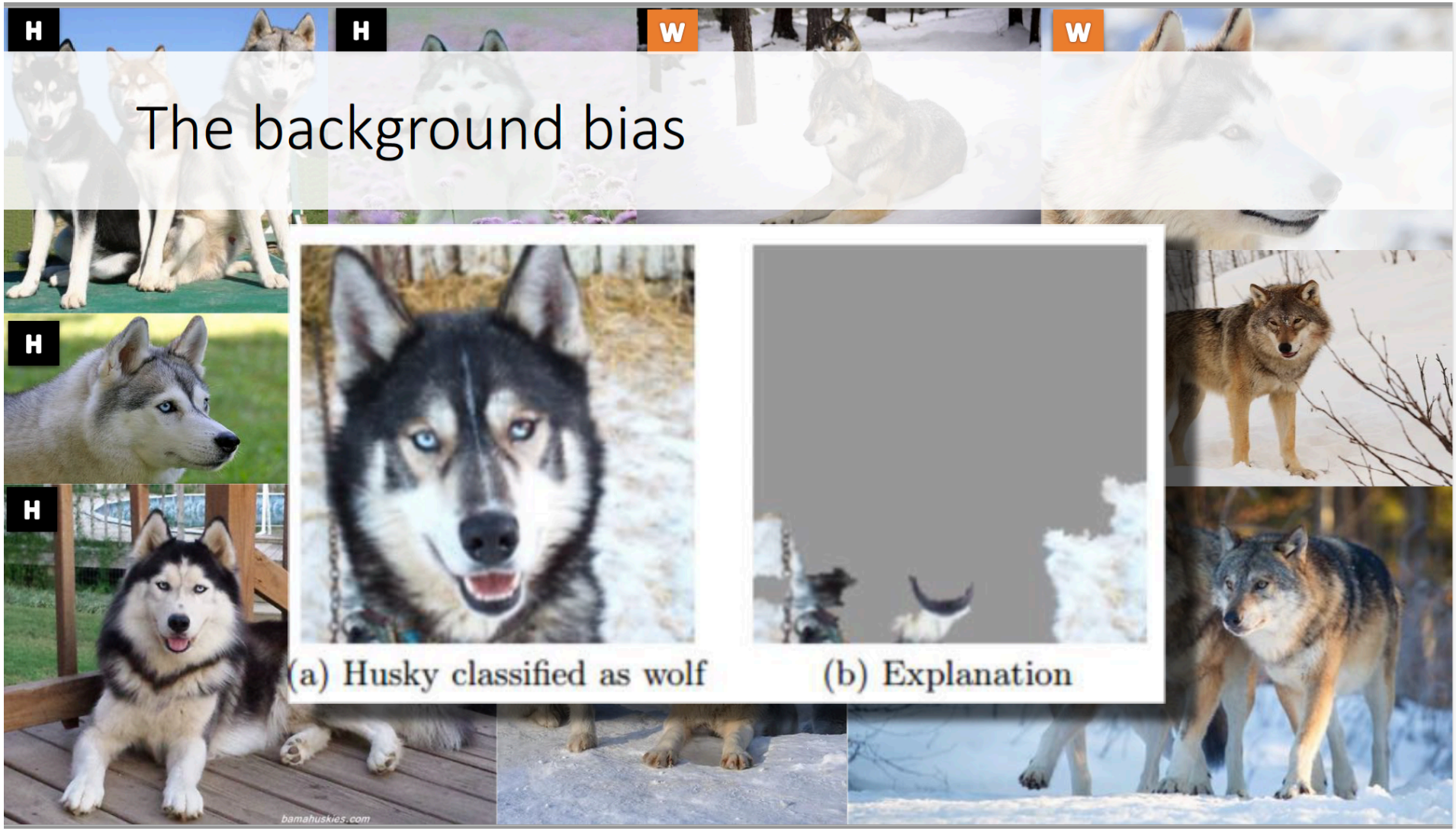
- Still discriminatory: individual discrimination is not group discrimination.
- $(X, S = 0)$ is then not a proper counterfactual for $(X, S = 1)$.

2.4: Illustration on the *Adult Income* dataset — Still punishable by law

Part 3: Other standard definitions of bias



- Forecast the Risk of committing a crime when set free, *i.e.* $Y = 0$.
- Protected Variable : Ethnic Origin, where $S = 0$ encodes *Afro-Americans*.
- It is balanced, *i.e.*: $\mathbb{P}(\hat{Y} = 1 \,|\, S = 1) \sim \mathbb{P}(\hat{Y} = 1 \,|\, S = 0)$.
- But the errors are different: $\mathbb{P}(\hat{Y} = 1 \,|\, S = 1, Y = 0) \gg \mathbb{P}(\hat{Y} = 1 \,|\, S = 0, Y = 0)$.



S is the presence of snow in the background

- Unfairness can mean that the algorithm does not achieve the same level of performance over the groups led by S.
- This motivates more subtle definitions of fairness than the Disparate Impact, *e.g.*:

Equality of Odds

$$\mathbb{P}(\hat{Y}=1\,|\,S=1,Y=0) \gg \mathbb{P}(\hat{Y}=1\,|\,S=0,Y=0)$$
$$\mathbb{P}(\hat{Y}=0\,|\,S=1,Y=1) \gg \mathbb{P}(\hat{Y}=0\,|\,S=0,Y=1)$$

Equality of Opportunity

$$\mathbb{P}(\hat{Y}=1\,|\,S=1,Y=0) \gg \mathbb{P}(\hat{Y}=1\,|\,S=0,Y=0)$$

Such definitions particularly make sense when the outcome of the algorithm blinds the future.

Part 4: Bias in Machine Learning

We recall:

- Input observations are (X, S)
- Output observations (when learning) are Y
- Decision rules to predict Y are $f(X, S)$

The M.L. algorithm learns f by minimizing an **empirical risk** for a chosen loss function ℓ

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i, S_i)).$$

Full fairness requires that S does not play any role when forecasting Y, i.e. when computing:

$$\widehat{Y} = f(X, S)$$

Underlying Probabilistic notion is **independence of distributions:**

- Statistical Parity : $\widehat{Y} \perp\!\!\!\perp S$
- Equality of Odds : $\widehat{Y} \mid Y \perp\!\!\!\perp S$

A fully fair classifier should be chosen in class of models ensuring these restrictions.

In all generality, we define the **risk of a classifier** as

$$R(f) = \mathbb{E}(\ell(Y, f(X)))$$

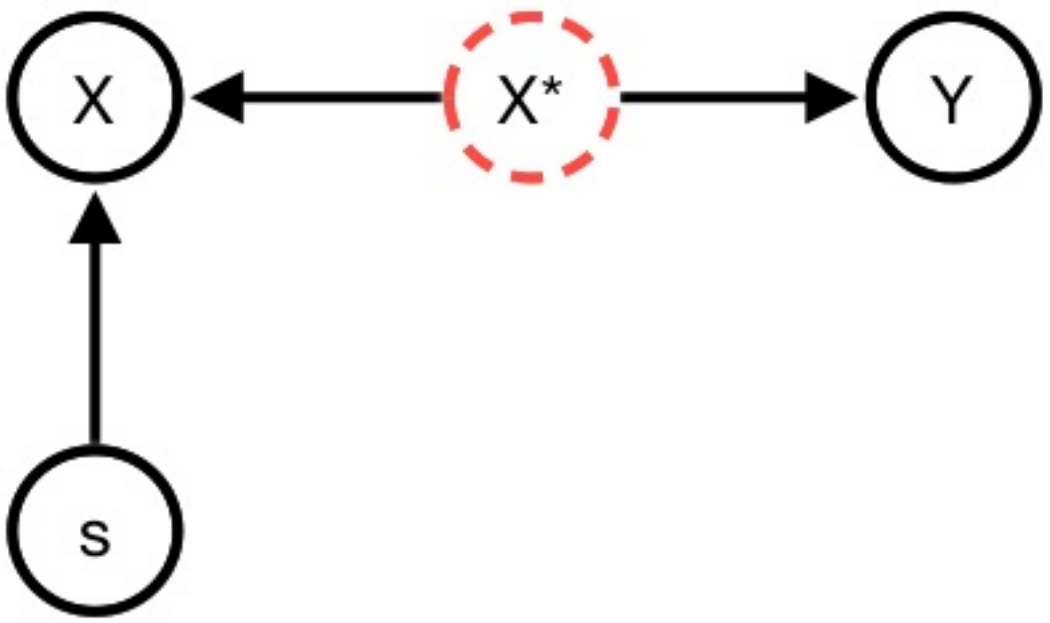
We define the classes, or more specifically the **restriction of classes**

$$\mathscr{F}_{SP} = \{f(X,S) \in \mathscr{F} \quad \text{s.t} \quad \hat{Y} \perp\!\!\!\perp S\}$$
$$\mathscr{F}_{EO} = \{f(X,S) \in \mathscr{F} \quad \text{s.t} \quad \hat{Y} \,|\, Y \perp\!\!\!\perp S\}$$

The **price for fairness** is:

$$\mathscr{E}_{\text{Fair}}(\mathscr{F}) := \inf_{f \in \mathscr{F}_{\text{Fair}}} R(f) - \inf_{f} R(f).$$

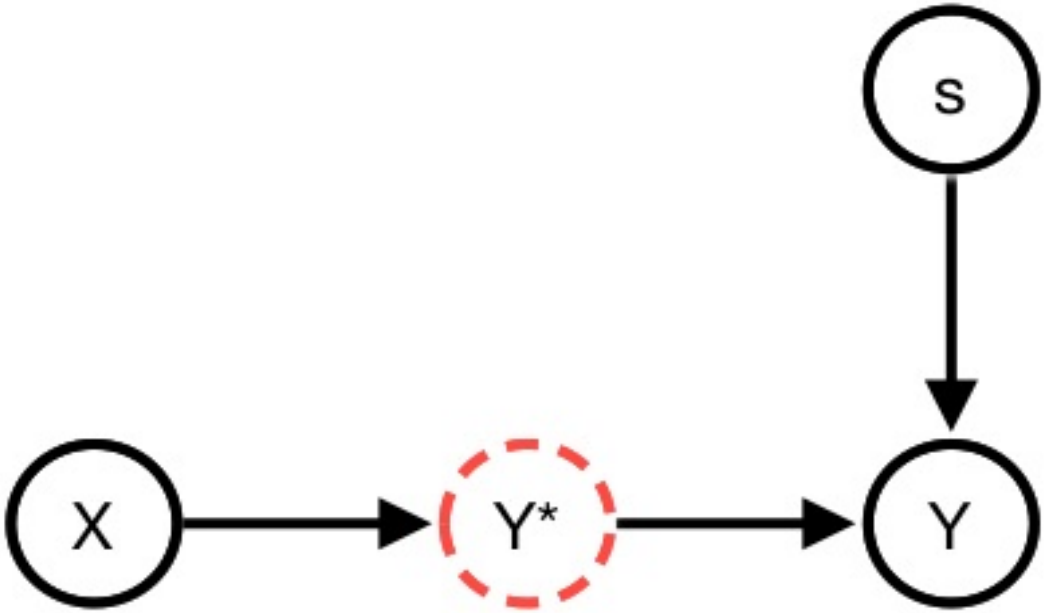
where $\inf_f R(f)$ is the Bayes Risk but could be replaced by $\inf_{f \in \mathscr{F}} R(f)$.



- The attributes X are a biased version of unobserved fair attributes $X^\star$.
- The target variable Y depends only on $X^\star$ and is fair in the sense that

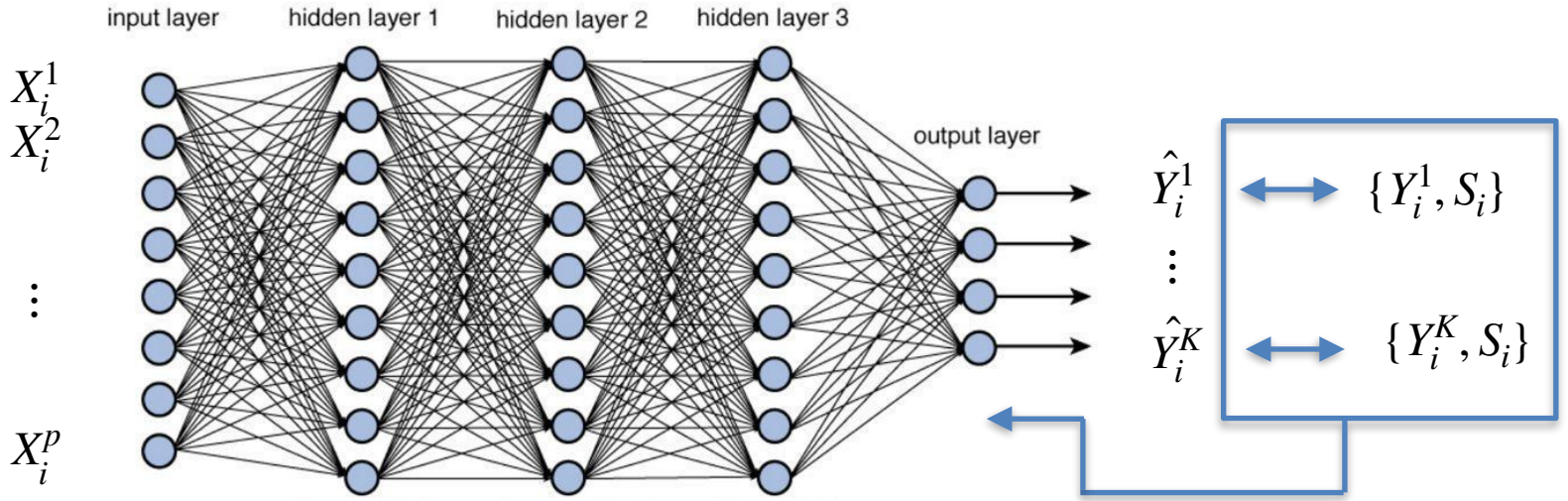
$$Y \,|\, X^\star \perp\!\!\!\perp S$$

- Learning from X induces biases while fairness enable a most accurate forecast.



- The decision Y observed is the result of a fair score $Y^\star$ which has been biased by the uses giving rise to Y.

***Back-propagation* of Fairness contraints in Neural Networks:**



$$\hat{\theta} = \arg\min_{\theta} R(\theta) + \lambda W_2^2(\mu_{\theta,0}^n, \mu_{\theta,1}^n) \quad \text{where} \quad W_2^2(\mu_{\theta,0}^n, \mu_{\theta,1}^n) = \int_0^1 \left(H_0^{-1}(\tau) - H_1^{-1}(\tau)\right)^2 d\tau$$

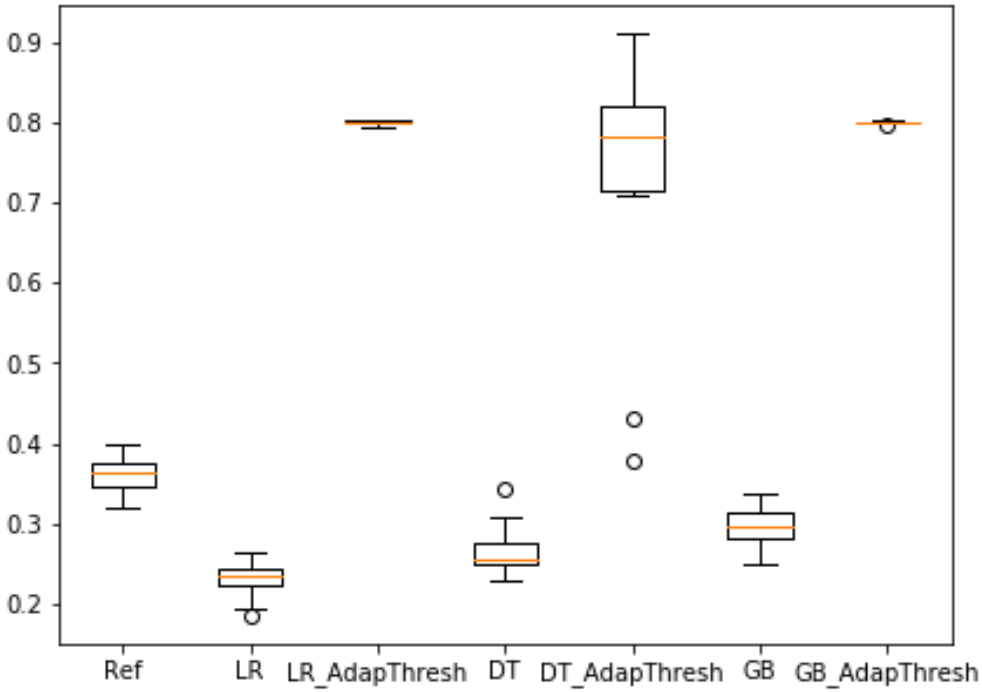
Standard loss

Fairness constraint

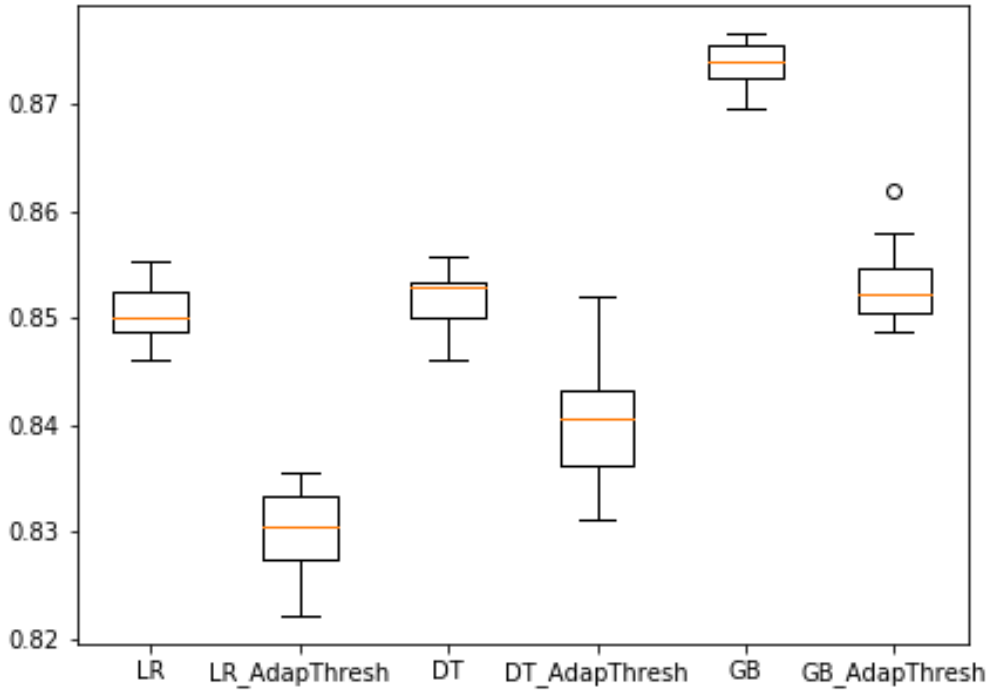
Distance between the predictions for which $S=0$ and $S=1$

Risser, Vincenot, Couellan, Loubes (2019) — https://arxiv.org/pdf/1908.05783.pdf

5.10: Performance of the Fair solutions



Disparate Impacts



Accuracies

λ is learnt by cross-validation of the Disparate Impact which decreases when Wasserstein distance decreases.

A Data Scientist can choose to change the world for better, equal or worse, ...

Time to conclude

Scientific Conclusions

- Bias-type constraints enable either to increase the accuracy of the forecast by removing learning sample unwanted bias
- ... or shape a different but *fair* reality

And provides a control on the generalization power of Machine Learning algorithms by ensuring a control on correlations.

Industrial Applications for critical systems.

- New methods to control the possible Risk of Discrimination at the Expand of a change of uses and customs.

Personal Conclusions

Personal opinion: An important question is « *What is the goal to be achieved?* »

Partial Fairness requires only matches a given criterion.

Controversy on the Exponential growing research on Fairness funded by Big companies (GAFAM) and applied to a wide range of applications that could be unfair by design

Law \& Maths should work together (joint work in ANITI)
Explainability of AI and Education to AI is the key for the future world

Do not trust me (or others), trust your own experiments

https://github.com/XAI-ANITI

Bibliography

E. del Barrio, P. Gordaliza and J.-M. Loubes. (2019): *A central limit theorem on the real line with application to fairness assessment in machine learning*. Information and Inference.

E. del Barrio, P. Gordaliza and J.-M. Loubes. (2019): *Obtaining Fairness with Optimal Transportation*. Proceedings of ICML.

E. del Barrio and J.-M. Loubes. (2019): *Central limit theorems for empirical transportation cost in general dimension*. The Annals of Probability

https://github.com/XAI-ANITI