Fairness and sensitive attribute inference

Jan Aalmoes
Joint work with Antoine Boutet and Vasisht Duddu
INSA - CITI laboratory - INRIA - PRIVATICS team

November 23, 2023

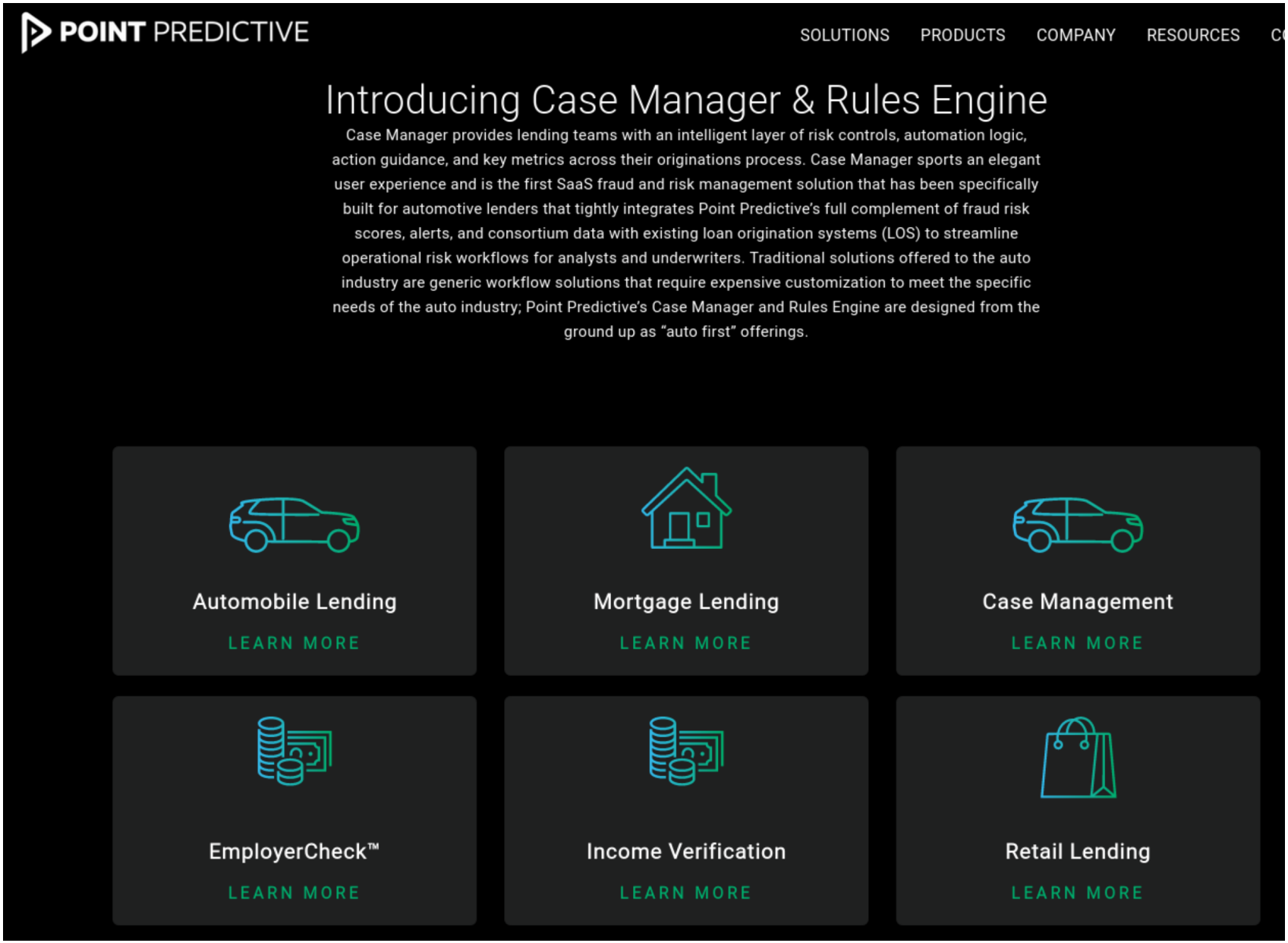
equivant

Our Solutions | Who We Support | About Us | Resources | Blog | Contact Us

It's here! COMPAS-R Core: Transparent RNA Built Your Way

Whether you w
more about ou
justice solution
demo, our tean
out your email
in touch as soo

Have you ever wondered how risk and needs assessments (RNAs) are developed? Let's take a behind-the-scenes look at our newest RNA, the COMPAS-R Core


POINT PREDICTIVE
SOLUTIONS
PRODUCTS
COMPANY
RESOURCES
Introducing Case Manager & Rules Engine
Case Manager provides lending teams with an intelligent layer of risk controls, automation logic, action guidance, and key metrics across their originations process. Case Manager sports an elegant user experience and is the first SaaS fraud and risk management solution that has been specifically built for automotive lenders that tightly integrates Point Predictive's full complement of fraud risk scores, alerts, and consortium data with existing loan origination systems (LOS) to streamline operational risk workflows for analysts and underwriters. Traditional solutions offered to the auto industry are generic workflow solutions that require expensive customization to meet the specific needs of the auto industry; Point Predictive's Case Manager and Rules Engine are designed from the ground up as "auto first" offerings.
Automobile Lending
LEARN MORE
Mortgage Lending
LEARN MORE
Case Management
LEARN MORE
EmployerCheck™
LEARN MORE
Income Verification
LEARN MORE
Retail Lending
LEARN MORE

Overview of fairness notions

- Group fairness
  - Demographic parity

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

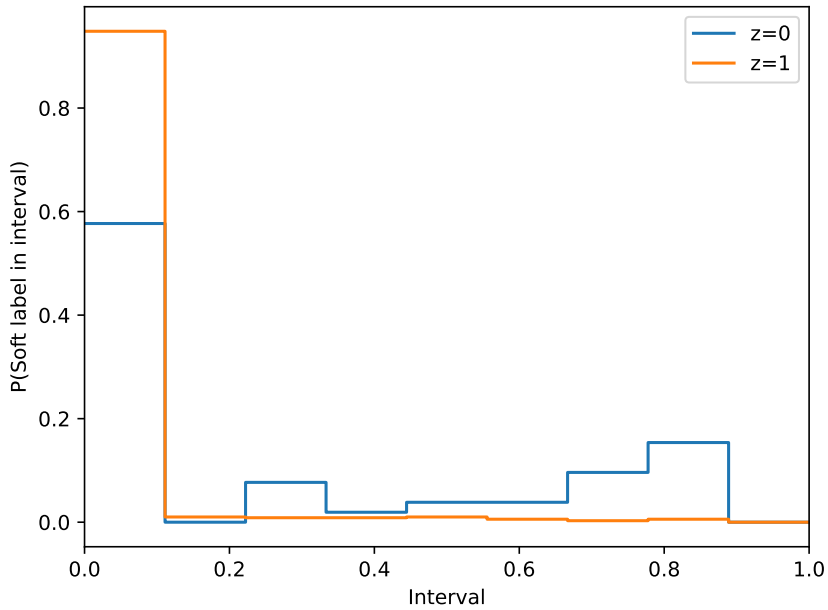
  - Equality of odds[1]

$$\forall y \in \{0, 1\}\ P(\hat{Y} = 1|S = 0, Y = y) = P(\hat{Y} = 1|S = 1, Y = y)$$

  - ...
- Individual fairness
  - Fairness Through Unawareness: Only works if X indep. of S
  - Fairness Through Awareness[2]: TODO
  - ...

[1] Hardt, Price, Srebro, Equality of Opportunity in Supervised Learning (NeurIPS) 2016

[2] Dwork,Hardt,Pitassi,Reingold,Fairness Through Awareness (ITCS) 2012

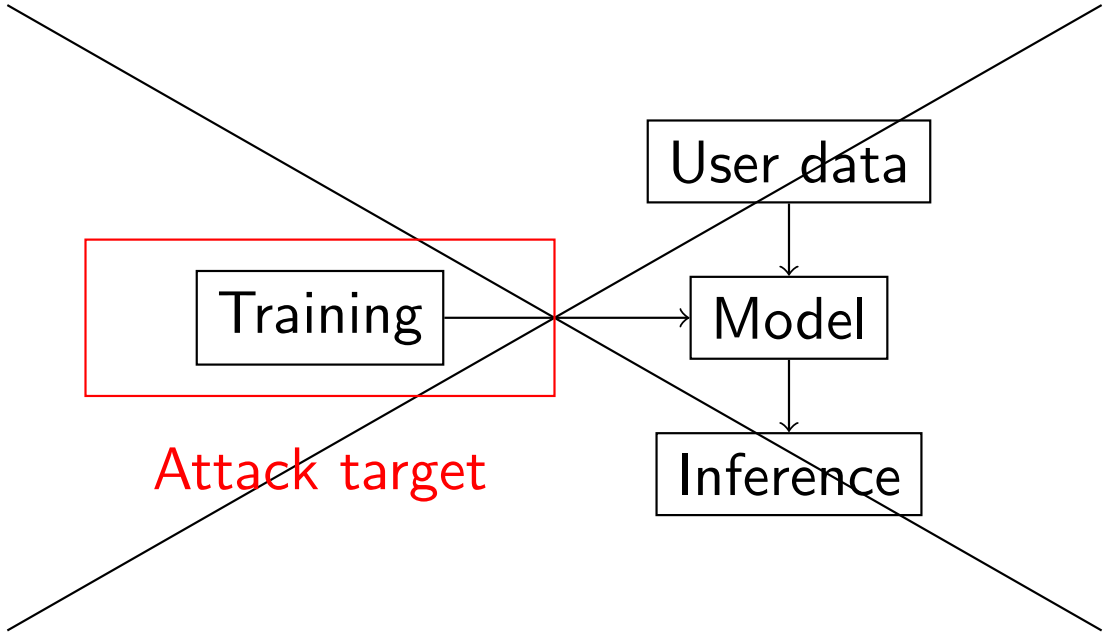


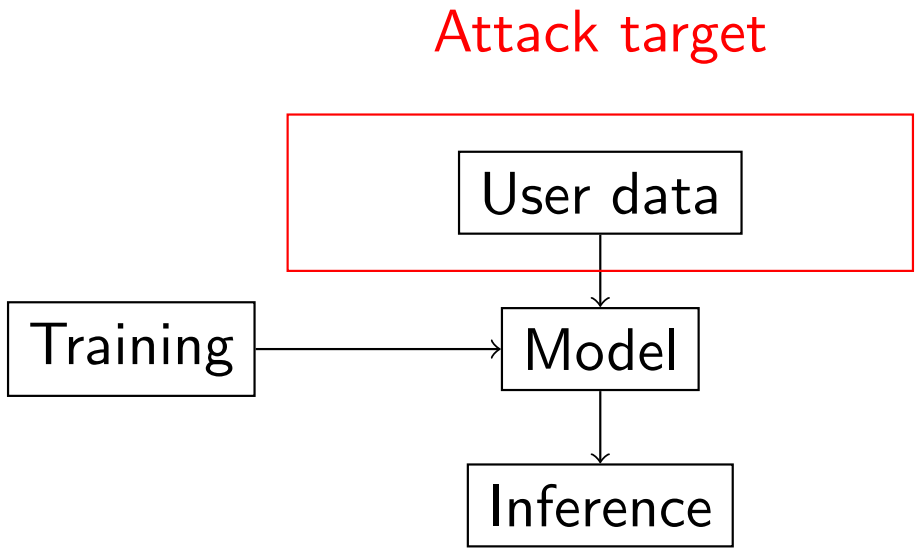
Dataset : Law school admission. Sensitive attribute : race

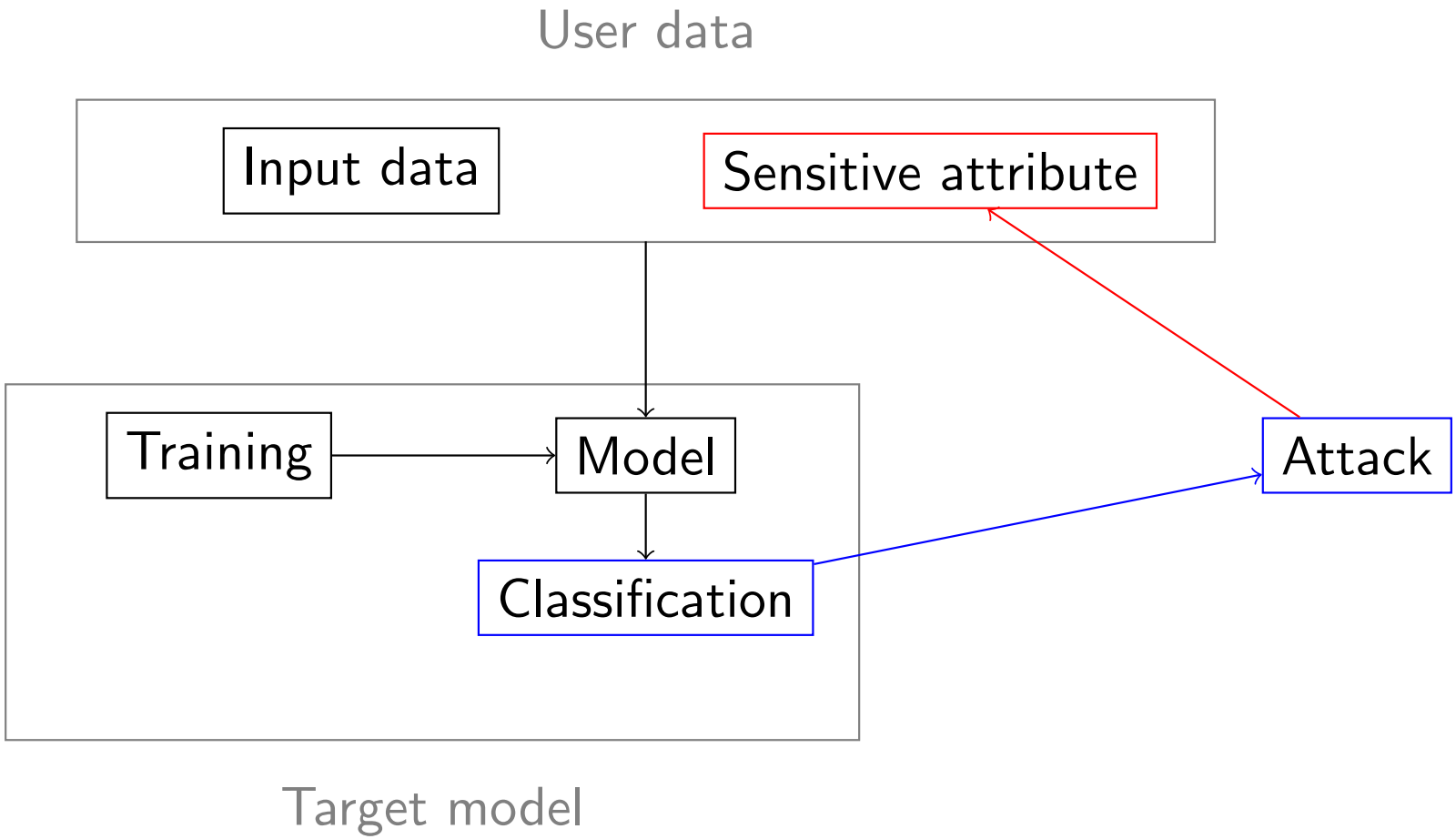
- $Z = 0$ Non-white
- $Z = 1$ White

Attribute Inference Attack

In our work


User data
Training
Model
Attack target
Inference


Attack target
User data
Training
Model
Inference


User data
Input data
Sensitive attribute
Training
Model
Classification
Attack
Target model

Two possible attacks

- **Soft labels (*sl*)**: $0.1,\ 0.8,\ 0.2,\ \cdots$
- **Hard labels (*hl*)**: $0,\ 1,\ 0,\ \cdots$

$$\text{Attack}(sl) = \text{Sensitive attribute}$$

or

$$\text{Attack}(hl) = \text{Sensitive attribute}$$

Fairness and attribute inference attack using **hard labels**

Demographic parity forbids attribute inference attack using hard labels

Definition

The fairness-level of a classifier $\hat{Y}$, that we call $Fl(\hat{Y})$, is

$$Fl(\hat{Y}) = |P(\hat{Y} = 1|S = 0) - P(\hat{Y} = 1|S = 1)|$$

$Fl(\hat{Y}) \in [0, 1]$ indicates how far $\hat{Y}$ is from reaching demographic parity.

Theorem

For all function $a : \{0, 1\} \to \{0, 1\}$ *the balanced accuracy of* $a \circ \hat{Y}$ *is lesser or equal than* $\frac{1}{2} + \frac{1}{2}Fl(\hat{Y})$

Proof.

- We know that we have only four transformations of $\{0, 1\}$: 0, 1, id, $1 - id$.
- We compute the balanced accuracy for each of those applications. For instance for id the balanced accuracy is

$$\begin{aligned}
&\frac{1}{2}(P(\hat{Y} = 0|S = 0) + P(\hat{Y} = 1|S = 1)) \\
=&\frac{1}{2}(1 - P(\hat{Y} = 1|S = 0) + P(\hat{Y} = 1|S = 1)) \\
\leq&\frac{1}{2} + \frac{1}{2}Fl(\hat{Y})
\end{aligned}$$

□

Equality of odds doesn't forbid attribute inference attack

Theorem

For all function $a : \{0, 1\} \to \{0, 1\}$, *if* $\hat{Y}$ *satisfies Equality of odds, then the balanced accuracy of* $a \circ \hat{Y}$ *is equal to* $\frac{1}{2}$ *if and only if* $\hat{Y}$ *is independent of* Y *or* Y *is independent of* S.

So, either the prediction task is independent of the sensitive attribute or the model did not learn.

Proof.

$$\hat{S} = a \circ \hat{Y}$$

- $$\begin{aligned} & P(\hat{S} = 0 | S = 0) \\ =& P(\hat{S} = 0 | S = 0 Y = 0) P(Y = 0 | S = 0) \\ +& P(\hat{S} = 0 | S = 0 Y = 1) P(Y = 1 | S = 0) \end{aligned}$$

- $$\begin{aligned} & P(\hat{S} = 0 | S = 0) + P(\hat{S} = 1 | S = 1) \\ =& 1 + (P(Y = 0 | S = 0) - P(Y = 0 | S = 1)) \\ & \left(P(\hat{Y} \in a^{-1}(\{1\}) | S = 1 Y = 0) - P(\hat{Y} \in a^{-1}(\{1\}) | S = 1 Y = 1) \right) \end{aligned}$$

Fairness and attribute inference attack using **soft labels**

Demographic parity means "The prediction is independent of the sensitive attribute"

$$FI(\hat{Y}) = 0 \Leftrightarrow P_{\hat{Y},S} = P_{\hat{Y}} \otimes P_S$$

No hypothesis on the support of $\hat{Y}$

(i.e. Demographic parity works for hard and soft labels, classification and regression problems)

Definition

$\hat{Y}$ satisfies demographic parity (for S) if and only if $P_{\hat{Y}} \otimes P_S$

Theorem

Let $\hat{Y} : \Omega \to E$ and $S : \Omega \to \{0, 1\}$. $\hat{Y}$ satisfies demographic parity for S if and only if, for all $a : E \to \{0, 1\}$, the balanced accuracy of $a \circ \hat{Y}$ is equal to $\frac{1}{2}$

This time, we can't think in terms of number of applications. a results from any kind of classifier (random forest, neural network, ...)

Proof.

$$\begin{aligned}
&\forall a\ P(\hat{Y} \in a^{-1}(\{0\})|S=0) + P(\hat{Y} \in a^{-1}(\{1\})|S=1) = 1 \\
\Leftrightarrow &\forall a\ P(\hat{Y} \in a^{-1}(\{0\})|S=0) = P(\hat{Y} \in a^{-1}(\{0\})|S=1) \\
\Leftrightarrow &\forall A\ P(\hat{Y} \in A)|S=0) = P(\hat{Y} \in A|S=1)
\end{aligned}$$

□

Experimental validation

- Bounds for hard labels tested on tabular and image dataset
- Evaluation of attribute inference attacks against fairness enforcing mechanisms

Take away

- Generalized demographic parity is a tool to study attribute inference attack (for hard and soft labels)
- Generalizing demographic parity introduces a notion of fairness in regression