Prisoners of Their Own Devices:

How Models Induce Data Bias in Performative Prediction

feedzai
instituto de telecomunicações
TÉCNICO LISBOA

José Pombal | Pedro Saleiro | Mário A. T. Figueiredo | Pedro Bizarro

Motivation

- By making discriminatory predictions, ML models have the potential to **exacerbate existing societal inequities**
- Most works in Fair ML focus on **measuring unfairness in static** algorithmic prediction tasks
- However, most real-world applications operate in **dynamic, performative prediction environments** (e.g.: fraud detection)
- In these settings, **model behaviour** influences the data's distribution and its biases, resulting in unfairness downstream

Contributions

- We propose a **data bias taxonomy** to characterize bias between a protected attribute, other features, and the target
- **We model 2 scenarios** where **data bias is induced by the predictive model itself**
- We use **real-world performative prediction use-case** as an example: bank account opening fraud
- We show how **biases in these settings have detrimental and unpredictable effects on performance and fairness**

Data Bias Taxonomy

Y: target variable
X: features
Z: protected attribute (categorical)

Base Bias Condition

P[X, Y] ≠ P[X, Y | Z]

The protected attribute is statistically related to either X, Y or both

Group-wise Class-conditional Distribution Bias

P[X | Y] ≠ P[X | Y, Z]

The feature distribution conditioned on the target varies from group to group in Z

Dynamic Bias

BC train ≠ BC production

Bias conditions (BC) in the training set differ from the ones found in production (testing)

Noisy Labels Bias

P*[Y | X, Z] ≠ P[Y | X, Z]

Some observations belonging to a protected group have been incorrectly labeled

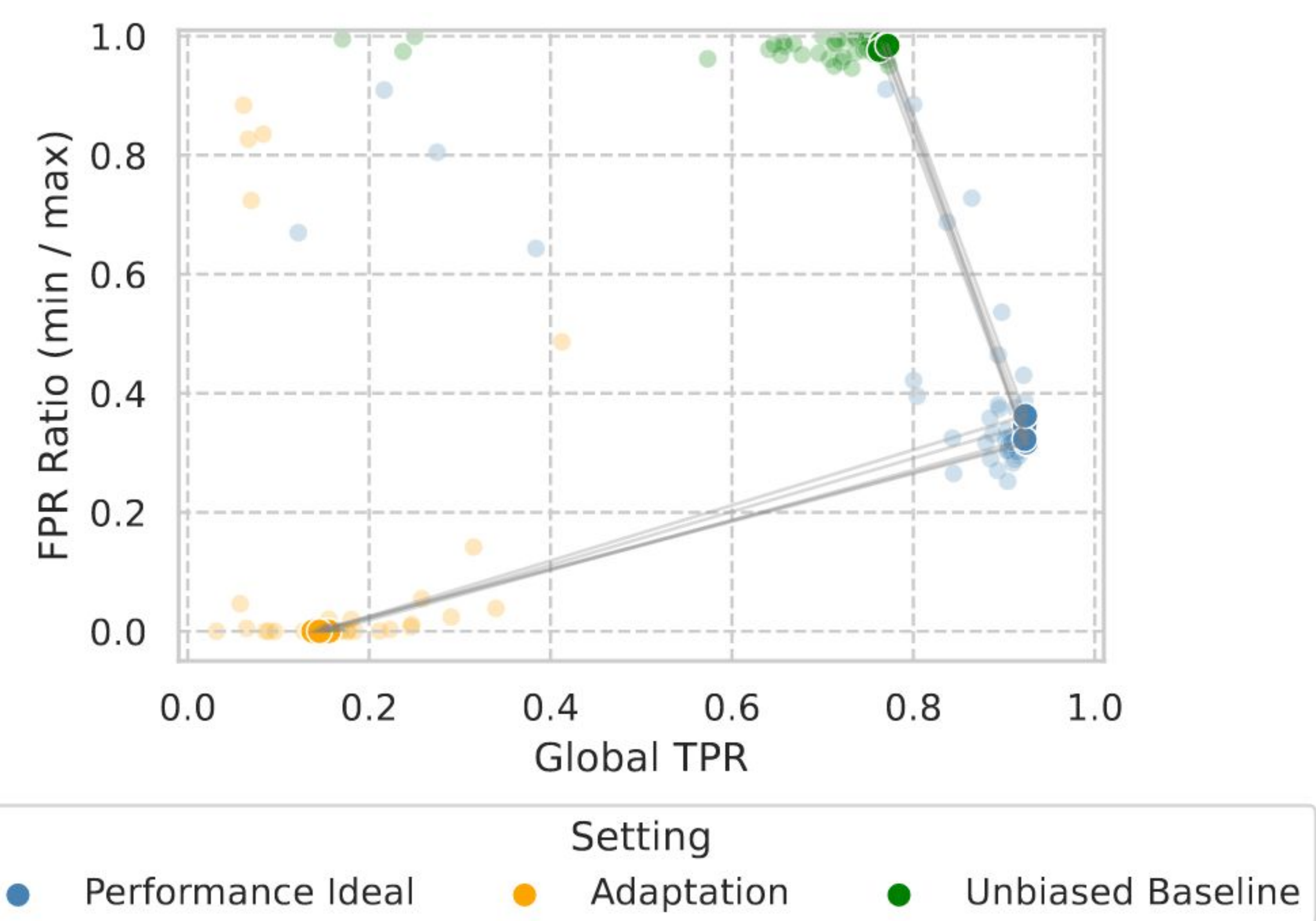
Scenario 1: Adaptive fraudsters

Conclusions

Performance and fairness decreased substantially with fraudsters' adaptation

The best models (opaque points) on **Performance Ideal** were not the best after **Adaptation**

Fairness-aware models could have fallen back to **Unbiased Baseline** instead of **Adaptation**



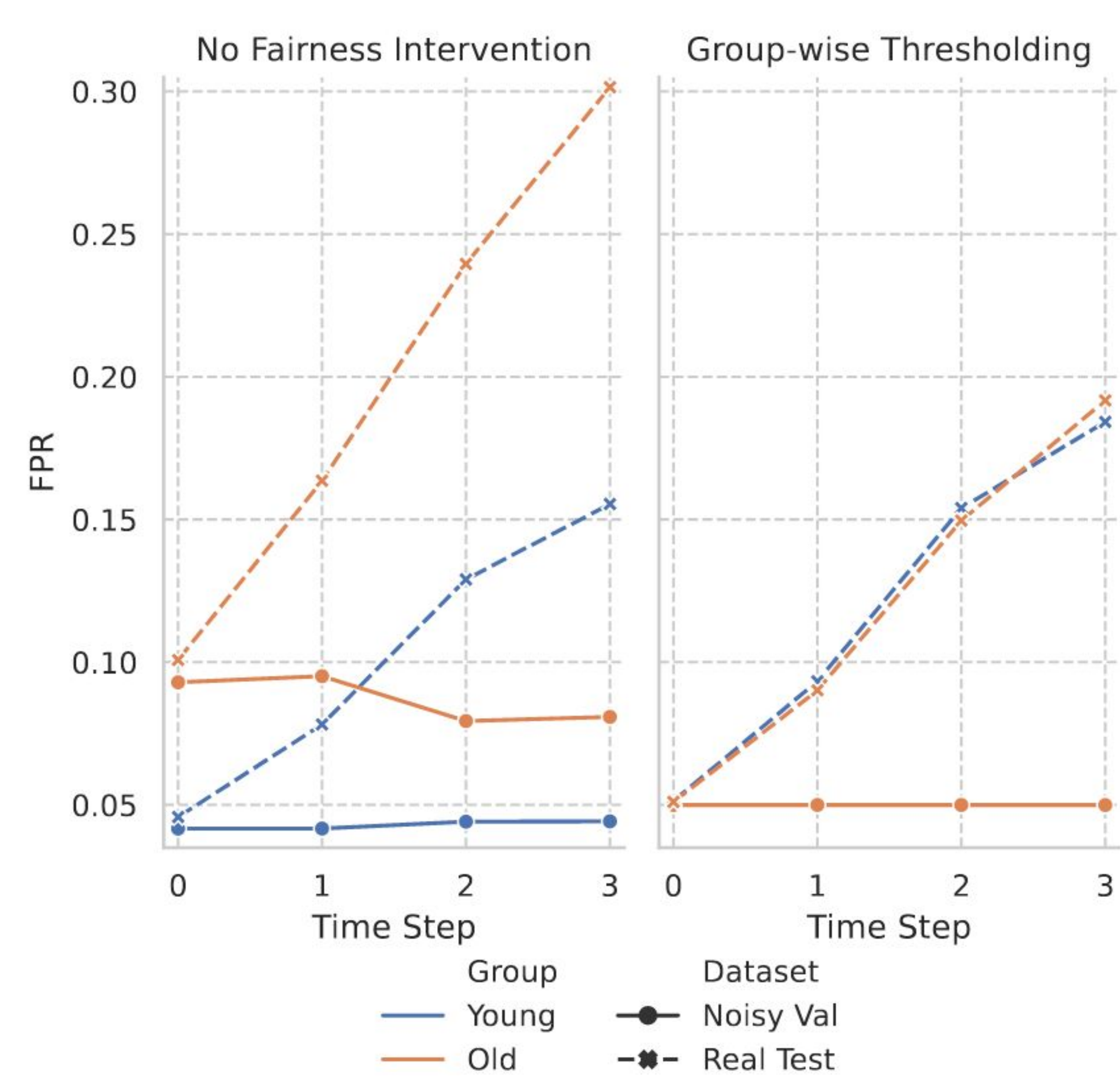
Setup

- Train and test 50 models on **3 variants of the original dataset**
- On **Unbiased Baseline** the protected attribute is independent of X and Y
- On **Performance Ideal** fraud is easier to detect using the protected attribute in train and test;
- On **Adaptation** fraud is easier to detect using Z in the training set, but **not in the test set**

Scenario 2: Noisy Selective Labels

Conclusions

FPR targets and fairness are put in jeopardy if selective labels are taken as label positives



- Over time, using model rejections as positive labels is enough to **operate at much higher levels of FPR than the practitioner thinks**
- We also applied a fairness intervention (group-wise thresholding)[1] to find that there is **a tendency over time for it to become less effective in the test set**, despite the validation indicating otherwise!