Making ML fairer through explanations, feature dropout, and aggregation

8 of April 2021, PDIA2021

G. Alves   F. Bernier   V. Bhargava   M. Couceiro   A. Napoli

Univ. Lorraine, CNRS, Inria N.G.E., LORIA

ICT-48 TAILOR & IPL HyAIAI

Fairness as non-discrimination

Fair model: that protects salient groups against discrimination

Discrimination: "unjust or prejudicial treatment of different categories of people, especially, on the grounds of race, age, or sex"

Example: Decision Making process...

- Human: Objective & Subjective reasoning
- Machine: Only objective **but** ...

Algorithmic decisions: are objective **but** they can be unfair

Common "sources": Data Collection & Model Choice

Critical applications of algorithmic decisions:

- Prediction of credit card defaulters

- Decisions on loan requests & job applications

- Stop & Frisk (minorities are affected!)

- COMPAS: Criminal recidivism (racial bias!)

- ...

Need of fairness: Unfair outcomes not only affect human rights, but they undermine **public trust** in ML & AI.

Motivation: unfair algorithmic decisions

Algorithmic decisions: are objective **but** they can be unfair

Common "sources": Data Collection & Model Choice

Critical applications of algorithmic decisions:

- Prediction of credit card defaulters

- Decisions on loan requests & job applications

- Stop & Frisk (minorities are affected!)

- COMPAS: Criminal recidivism (racial bias!)

- ...

Need of fairness: Unfair outcomes not only affect human rights, but they undermine **public trust** in ML & AI.

Defining and improving "fairness" of ML...

Based on **decision outcomes**, fairness can be assessed through:

- **Fairness metrics**: individual & group fairness, equal opportunity, demographic parity, equal accuracy, etc.
- **Process fairness**: model's reliance on "sensitive features" (e.g., salient features such as race, age, or sex,...)

Two main approaches to dealing with ML unfairness:

1. Enforce fairness constraints while learning, e.g.:

    $P(y_{pred} \neq y_{true} | race = Black) = P(y_{pred} \neq y_{true} | race = White)$

    Drawback: Complexity, fairness "gerrymandering" & overfitting

2. Exclude sensitive/salient features (for instance, COMPAS)

    Drawback: Decreased accuracy!

Defining and improving "fairness" of ML...

Based on **decision outcomes**, fairness can be assessed through:

- **Fairness metrics**: individual & group fairness, equal opportunity, demographic parity, equal accuracy, etc.
- **Process fairness**: model's reliance on "sensitive features" (e.g., salient features such as race, age, or sex,...)

Two main approaches to dealing with ML unfairness:

1. Enforce fairness constraints while learning, e.g.:

$$P(y_{\text{pred}} \neq y_{\text{true}} | race = Black) = P(y_{\text{pred}} \neq y_{\text{true}} | race = White)$$

   Drawback: Complexity, fairness "gerrymandering" & overfitting

2. Exclude sensitive/salient features (for instance, COMPAS)

   Drawback: Decreased accuracy!

Framework to deal Process Fairness

Original Goal: Human-centered approach to reduce a model's dependence on sensitive/salient features while improving its accuracy

Proposal: Framework consisting of two components:

(i) to assess a model's dependence on sensitive features (fair/unfair)

(ii) (if dependent) to render it fairer (without compromising accuracy)

Idea: Use a FI-explainer to assess model's dependence sensitive feat.s

Examples: LIME, SHAP and gradient based (under further assumptions)

Here: we focused on model agnostic approaches...

FixOut (FaIrness through eXplanations and feature dropOut)

Fair Model: if its outcomes do not depend on sensitive features

Input: model M, dataset D, sensitive features F, explanation method E

Output: M if fair, otherwise a fairer and more accurate M_{final}

Proposal: FixOut with two components

- **Exp$_{Global}$:** for global explanations (FI)
- **Ensemble$_{Out}$:** Ensemble approach relying on "feature dropout"

FixOut: https://fixout.loria.fr/

FixOut (FaIrness through eXplanations and feature dropOut)

Fair Model: if its outcomes do not depend on sensitive features

Input: model M, dataset D, sensitive features F, explanation method E

Output: M if fair, otherwise a fairer and more accurate M_{final}

Proposal: **FixOut** with two components

- **Exp$_{Global}$:** for global explanations (FI)
- **Ensemble$_{Out}$:** Ensemble approach relying on "feature dropout"

FixOut: https://fixout.loria.fr/

Idea: Explanations can provide insight into process fairness.

However: LIME and SHAP provide "local" explanations

Solution: Sample a set of instances and aggregate the contributions to estimate the global contribution of each feature.
Example: random or "Sub-modular pick"

Output: k most important (globally) features.

Rule:

If there is **at least one** sensitive feature among the top-k, **then** M is deemed unfair and **Ensemble$_{\text{Out}}$** applies.

Ensemble$_{Out}$: model M, dataset D, sensitive features F

Let $a_1, a_2, \ldots, a_k$ be the k features that Exp$_{Global}$ outputs

Suppose that $a_{j_1}, a_{j_2}, \ldots, a_{j_i}, i > 1$, are sensitive (i.e., $\in F$)

Then FixOut trains $i + 1$ classifiers obtained by "feature dropout":

- M_t after removing a_{j_t} from the dataset, for $t = 1, \ldots, i$, and

- M_{i+1} after removing all sensitive features $a_{j_1}, a_{j_2}, \ldots, a_{j_i}$.

Output: Ensemble classifier M_{final} as an aggregation of all M_t's.

Example: for an instance x and a class C,

1 FixOut: ensemble classifier M_{final} defined as a **simple average**

$$P_{M_{final}}(x \in C) = \sum_{t=1}^{i+1} w_t P_{M_t}(x \in C).$$

2 FixOut (w): Ensemble classifier M_{final} defined as a **weighted average**

$$P_{M_{final}}(x \in C) = \sum_{t=1}^{i+1} w_i P_{M_t}(x \in C),$$

where $w_t = \frac{c_{j_t}}{1+\sum_{u=1}^{i} c_{j_u}}$, $1 \leq t \leq i$, and $w_{i+1} = \frac{1}{1+\sum_{u=1}^{i} c_{j_u}}$ using normalized global feature contributions c_i's.

8

FixOut with LIME explanations

LIME Explanations[1]

LIME: learns a linear $g \in \mathcal{G}$ on a neighborhood of x (to explain) by

$$g = argmin_{g' \in \mathcal{G}} \ \mathcal{L}(f, g', \pi_x) + \Omega(g')$$

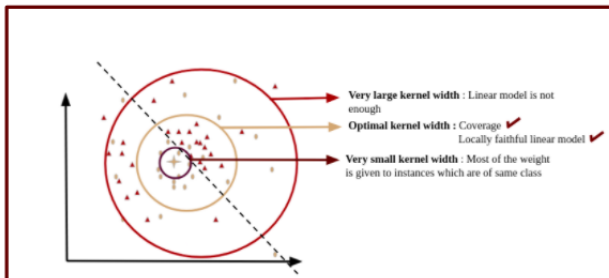for the distance $\mathcal{L}(f, g', \pi_x)$ of f and g' on the kernel π_x

Figure 1: Illustration of optimal kernel on the (interpretable) space

[1]Ribeiro, *et al.* "Why Should I Trust You?": Explaining predictions of any...

LIME Explanations

LIME: learns a model g on the neighborhood of an instance to explain

$$g(\hat{x}) = \hat{\alpha}_0 + \sum_{1 \leq i \leq d'} \hat{\alpha}_i \hat{x}_i,$$

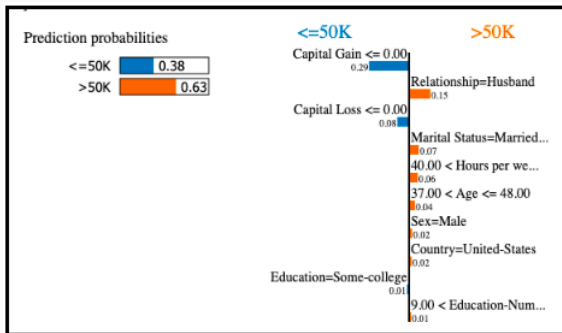where $\hat{\alpha}_i$ represents the contribution or importance of feature $\hat{x}_i$

Figure 2: Local explanation in case of Adult dataset (salary prediction)

FixOut with LIME explanations

Exp$_{Global}$: LIME + random sampling
(of instances and use their explanations to get global explanations)

As before: if Exp$_{Global}$ outputs $a_1, a_2, \ldots, a_k$ and $a_{j_1}, a_{j_2}, \ldots, a_{j_i} \in F$,
then *FixOut* trains $i + 1$ classifiers obtained by "feature dropout":

- M_t after removing a_{j_t} from the dataset, for $t = 1, \ldots, i$, and
- M_{i+1} after removing all sensitive features $a_{j_1}, a_{j_2}, \ldots, a_{j_i}$.

Ensemble$_{Out}$: Ensemble classifier M_{final} defined as

- a simple average (**FixOut**)
- a weighted average (**FixOut (w)**)

German Credit Card Score (UCI):

- Applicant profiles (demographic and socio-economic).
- Goal: Predict credit risks (likely & unlikely to pay back)
- Sensitive: 'Statussex', 'telephone', 'foreign worker'

Empirical setting:

- Random Forest: 70% training & 30% test data
- Used: SMOTE oversampling & threshold tuning while training
- Accuracy of M: 0.783

Question: Is this model fair?

Example: RF on German Credit Card Score

German Credit Card Score (UCI):

- Applicant profiles (demographic and socio-economic).
- Goal: Predict credit risks (likely & unlikely to pay back)
- Sensitive: 'Statussex', 'telephone', 'foreign worker'

Empirical setting:

- Random Forest: 70% training & 30% test data
- Used: SMOTE oversampling & threshold tuning while training
- Accuracy of M: 0.783

Question: Is this model fair?

Exp$_{Global}$ with LIME to assess fairness

Feature	Contribution
foreignworker	2.664899
otherinstallmentplans	-1.354191
housing	-1.144371
savings	0.984104
property	-0.648104
purpose	-0.415498
existingchecking	0.371415
telephone	0.311451
credithistory	0.263366
duration	-0.223288

Table 1: Top 10 features used by M (by 'submodular pick')

Hence: Model deemed unfair

Approach: Train multiple models obtained with feature dropout

- **M1:** Model trained after removing 'foreignworker'.

- **M2:** Model trained after removing 'telephone'.

- **M3:** Model trained after removing the 2 (accuracy of 0.773)
  NB: Accuracy drop when all sensitive features are removed!

M$_{final}$: Ensemble of M1, M2 and M3 (accuracy of 0.786)

Exp$_{Global}$ with LIME Explanations (RF on German)

Original		Ensemble	
Feature	**Contribution**	**Feature**	**Contribution**
foreignworker	2.664899	otherinstallmentplans	-1.487604
otherinstallmentplans	-1.354191	housing	-1.089726
housing	-1.144371	savings	0.679195
savings	0.984104	duration	-0.483643
property	-0.648104	**foreignworker**	0.448643
purpose	-0.415498	property	-0.386355
existingchecking	0.371415	credithistory	0.258375
telephone	0.311451	job	-0.252046
credithistory	0.263366	existingchecking	-0.21358
duration	-0.223288	residencesince	-0.138818

Result: M$_{final}$ is "fairer" & at least as accurate (from 0.783 to 0.786)

Empirical Study

We tested our approach on different datasets. **E.g.:**

1. Adult Dataset
   - Information on US Citizens and their salaries.
   - Goal: Predict if salary $\geq$ 50k dollars.
   - Sensitive: 'Sex', 'race', 'marital status'

2. Law School Admissions Council (LSAC)
   - Student profiles (demographic, socio-economic, etc.).
   - Goal: Predict whether a law student passes "bar exam"
   - Sensitive: 'Race', 'sex', 'family income'

**Models:** LR, RF, AdaBoost, Bagging (and others with similar results)

Classification assessment (FixOut with LIME)

Dataset	Method	Accuracy				Precision				Recall			
		ADA	BAG	LR	RF	ADA	BAG	LR	RF	ADA	BAG	LR	RF
German	Original	.7362	.7019	.7398	.7556	.5707	.5124	.5716	.6883	.5317	.5738	.5495	.3595
	FixOut	.7419	.7273	.7418	.7598	.5801	.5549	.5754	.7060	.5321	.5371	.5622	.3585
	FixOut (w)	.7405	.7219	.7400	.7583	.5764	.5471	.5708	.7019	.5373	.5076	.5602	.3541
Adult	Original	.8503	.8301	.6706	.8441	.6884	.6419	.3857	.7004	.6882	.6687	.5600	.6175
	FixOut	.8515	.8424	.6786	.8473	.6930	.6838	.3859	.7121	.6856	.6451	.5317	.6153
	FixOut (w)	.8512	.8388	.6713	.8470	.6927	.6771	.3922	.7108	.6838	.6343	.5398	.6156
LSAC	Original	.8398	.8544	.7526	.8513	.8986	.8846	.8548	.8771	.9016	.9413	.8330	.9473
	FixOut	.8331	.8620	.7440	.8553	.9044	.8898	.8596	.8838	.8850	.9448	.8136	.9436
	FixOut (w)	.8187	.8456	.7294	.8514	.9071	.8909	.8614	.8851	.8618	.9201	.7899	.9363

Fairness assessment: FixOut with LIME

Legend: ■ Original+LIME ■ *FixOut* +LIME ■ *FixOut* (w)+LIME

(a) **German dataset**
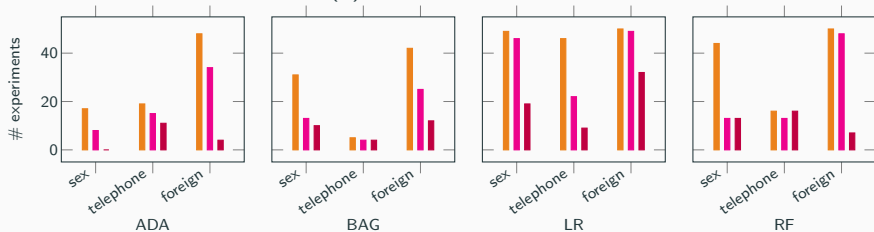
(b) **Adult dataset**
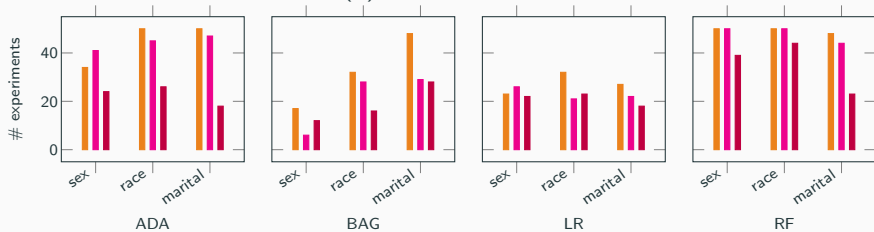
FixOut with SHAP explanations

Still: an additive feature attribution method, i.e., linear model

$$g(z) = \phi_0 + \sum_{1 \leq i \leq d'} \phi_i z_i,$$

where ϕ_i represents the contribution (importance) of interpretable feature z_i

SHAP: uses Shapley kernel π_x and thus estimation of Shapley values ϕ_i (coalitional game theory)
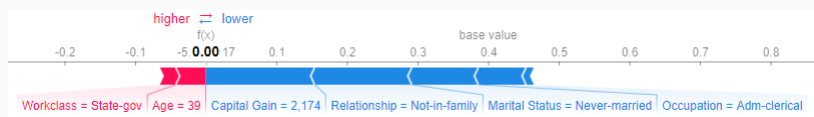

Figure 3: SHAP explanation in case of Adult dataset (salary prediction)

[2]Lundberg, *et al.* A Unified Approach to Interpreting Model Predictions...

Example: RF on German Credit Card Score

Goal: Predict credit risk

- **Random Forest:** 70% training & 30% test data

- Used: SMOTE oversampling & threshold tuning

- Accuracy of M: 0.753

Question: Is this model fair?

Exp$_{Global}$ with SHAP to assess fairness

Feature	Contribution
existingchecking	-7.11624
statussex	-5.950176
housing	-3.27344
job	-2.868195
residencesince	2.832573
telephone	2.290478
property	2.042944
otherinstallmentplans	-1.985275
existingcredits	1.984547
purpose	1.711321

Table 2: Top 10 features used by M

Hence: Model deemed unfair

Approach: Train multiple models obtained with feature dropout

- **M1:** Model trained after removing 'statussex'.

- **M2:** Model trained after removing 'telephone'.

- **M3:** Model trained after removing the 2 (accuracy of 0.746)
  NB: Accuracy drop when all sensitive features are removed!

M_{final}: Ensemble of M1, M2 and M3 (accuracy of 0.760)

Exp_Global Explanations (RF on German)

Original		Ensemble	
Feature	**Contribution**	**Feature**	**Contribution**
existingchecking	-7.11624	existingchecking	-4.285092
statussex	-5.950176	housing	-3.771932
housing	-3.27344	property	3.506007
job	-2.868195	job	-3.061209
residencesince	2.832573	employmentsince	2.646814
telephone	2.290478	existingcredits	2.409782
property	2.042944	otherinstallmentplans	-2.389899
otherinstallmentplans	-1.985275	savings	-2.215407
existingcredits	1.984547	residencesince	2.212183
purpose	1.711321	credithistory	1.188159

Result: M_{final} is fairer & more accurate (from 0.753 to 0.760)

Classification assessment (FixOut with SHAP)

| Dataset	Method	Accuracy				Precision				Recall			
		ADA	BAG	LR	RF	ADA	BAG	LR	RF	ADA	BAG	LR	RF
German	Original	.7362	.7019	.7398	.7556	.5707	.5124	.5716	.6883	.5317	.5738	.5495	.3595
	FixOut	.7419	.7273	.7418	.7598	.5801	.5549	.5754	.7060	.5321	.5371	.5622	.3585
	FixOut (w)	.7427	.7253	.7417	.7613	.5809	.5537	.5746	.7003	.5390	.5142	.5632	.3708
Adult	Original	.8503	.8301	.6706	.8441	.6884	.6419	.3857	.7004	.6882	.6687	.5600	.6175
	FixOut	.8515	.8424	.6786	.8473	.6930	.6838	.3859	.7121	.6856	.6451	.5317	.6153
	FixOut (w)	.8518	.8399	.6901	.8463	.6948	.6805	.4213	.7104	.6829	.6343	.5123	.6119
LSAC	Original	.8398	.8544	.7526	.8513	.8986	.8846	.8548	.8771	.9016	.9413	.8330	.9473
	FixOut	.8331	.8620	.7440	.8553	.9044	.8898	.8596	.8838	.8850	.9448	.8136	.9436
	FixOut (w)	.8181	.8606	.7135	.8470	.9080	.8894	.8674	.8866	.8599	.9433	.7584	.9280

Fairness assessment: FixOut with SHAP

Original+SHAP ■ *FixOut* +SHAP ■ *FixOut* (w)+SHAP

(a) **German dataset**
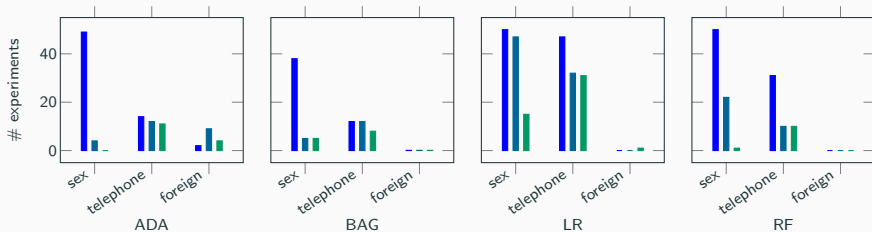
(b) **Adult dataset**
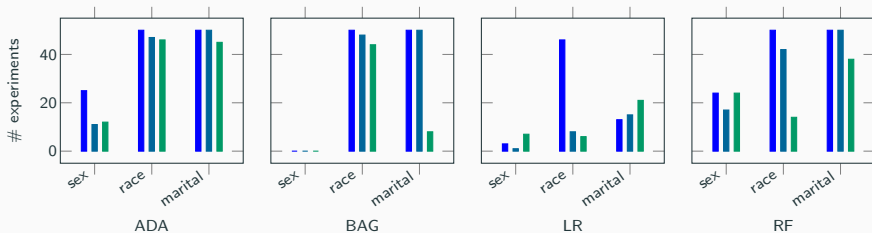
Fairness assessment: Average contribution of sensitive features

	Method	ADA			BAG			LR			RF		
		sex	telephone	foreign	foreign	telephone	foreign	sex	telephone	foreign	sex	telephone	foreign
German	Original+LIME	-0.13	0.12	3.84	-2.13	0.33	6.36	-13.90	10.08	25.55	-3.29	0.85	23.00
	FixOut +LIME	-0.05	0.09	0.85	-0.63	0.15	1.88	-7.46	2.86	11.90	-0.55	0.67	7.47
	FixOut w+LIME	0.00	0.06	0.02	-0.79	0.11	0.65	-2.00	1.24	3.28	-0.49	0.69	0.23
	Original+SHAP	-0.68	0.10	0.01	-5.13	1.55	0.00	-31.20	11.59	0.00	-10.53	3.21	0.00
	FixOut +SHAP	-0.02	0.08	0.04	-0.76	1.08	0.00	-10.20	3.52	0.00	-1.87	0.69	0.00
	FixOut w+SHAP	-0.07	0.08	0.13	-0.87	0.71	0.00	-1.37	3.25	0.06	-1.87	0.69	0.00
		marital	race	sex	marital	race	sex	marital	race	sex	marital	race	sex
Adult	Original+LIME	0.88	0.43	0.48	14.35	-1.02	2.11	0.49	-0.05	1.13	8.10	11.88	9.59
	FixOut +LIME	0.34	0.25	0.21	2.90	-0.40	3.03	-0.01	-0.11	0.07	4.63	5.77	5.80
	FixOut w+LIME	-0.02	0.13	0.03	-0.65	-0.62	2.57	0.01	-0.07	0.34	1.05	1.45	1.70
	Original+SHAP	-3.32	0.08	0.53	98.35	0.00	4.10	-0.15	0.00	1.67	-23.29	1.29	9.25
	FixOut +SHAP	-1.26	0.03	0.24	31.51	0.00	5.25	-0.06	0.00	0.06	-11.68	0.86	3.13
	FixOut w+SHAP	-0.17	0.03	0.18	0.37	0.00	4.47	-0.21	-0.02	0.02	-1.73	1.30	0.55
		sex	race	fincome	sex	race	fincome	sex	race	fincome	sex	race	fincome
LSAC	Original+LIME	0.01	0.38	0.33	2.38	-28.79	13.56	-1.35	25.30	-5.01	-0.31	-43.64	-3.48
	FixOut +LIME	0.02	0.18	0.15	1.18	-18.02	6.31	-0.59	7.89	-1.08	-0.28	-15.18	-1.18
	FixOut w+LIME	0.02	0.09	0.06	1.72	-0.29	1.81	-0.57	0.95	0.09	-0.27	0.38	-0.65
	Original+SHAP	-0.06	0.05	-0.07	-6.59	-0.58	-2.43	-8.63	5.38	1.28	-3.31	-0.18	-2.70
	FixOut +SHAP	-0.06	0.02	-0.05	-2.44	-0.82	-1.58	-3.50	2.74	0.41	-1.71	-0.67	-1.65
	FixOut w+SHAP	-0.02	0.01	-0.01	-0.42	-0.91	-0.49	-1.03	1.17	0.29	-0.26	-0.67	-0.43

26

What about Fairness metrics?

Idea: Separate instances into two groups w.r.t. a sensitive feature
E.g.: Non-white people (unprivileged) **versus** white people (privileged)

Demographic Parity (DP)[3]:
$$DP = P(\hat{y} = pos | D = unp) - P(\hat{y} = pos | D = priv)$$

Equal Opportunity (EO)[4]: $EO = \frac{TP_{unp}}{TP_{unp} + FN_{unp}} - \frac{TP_{priv}}{TP_{priv} + FN_{priv}}$

Predictive Equality (PE)[5]: $PE = \frac{FP_{unp}}{FP_{unp} + TP_{unp}} - \frac{FP_{priv}}{FP_{priv} + TP_{priv}}$

[3] Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments
[4] Zafar, et al. Fairness beyond disparate treatment & impact: Learning classification without disparate mistreat.
[5] Alves, et al. Making ML models fairer through explanations: the case of LimeOut

What about Fairness metrics? (cont.)

Privileged groups

1. German dataset
   - "status sex": "male single"
   - "telephone": "yes" (registered under the customers name)
   - "foreign worker": "no"

2. Adult dataset
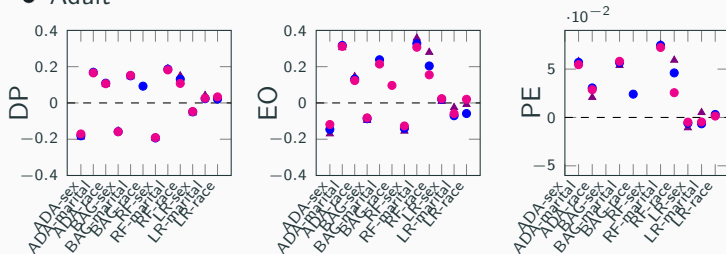   - "Marital Status": "Married"
   - "Race": "white"
   - "Sex": "male"

Fairness metrics

Final remarks
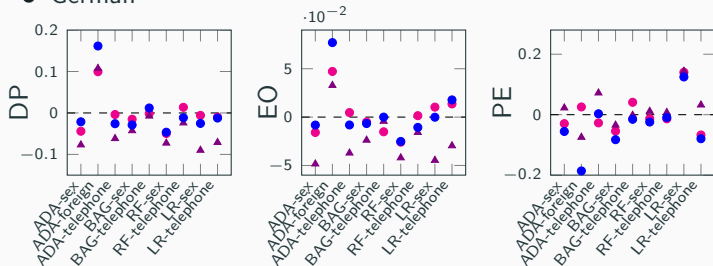
FixOut:

- Human-centered framework to tackle process fairness.

- Showed how to use $\text{Exp}_{\text{Global}}$ to assess model fairness.

- Illustrated the feasibility of 'feature dropout' followed by an ensemble approach.

Improved on process fairness and (at least) maintained on other fairness metrics!

Final remarks (cont.)

Ongoing and future work:

- Improvement of Ensemble$_{Out}$ by considering the removal of multiple combinations of sensitive features (rather than one & all) and different aggregation procedures (instead of weighted sums)

- Automation of context-based selection/detection of sensitive features. In particular: choice of parameter k.

- Adaptation to different applications (complex and structured data).

Thank you for your attention!

Merci de votre attention !

References

Alves, *et al.* Making ML models fairer through explanations: the case of LimeOut, *AIST'20*.

Bhargava, *et al.* LimeOut: An Ensemble Approach To Improve Process Fairness, *XKDD'20*.

Dimanov, *et al.* You Shouldn't Trust Me: Learning Models Which Conceal Unfairness from Multiple Explanation Methods, *ECAI'20*.

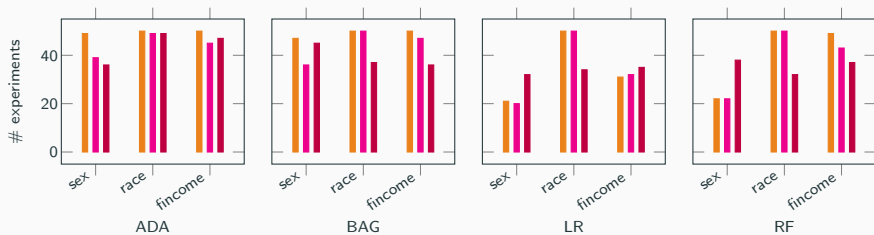Garreau, *et al.* Explaining the Explainer: A First Theoretical Analysis of LIME, *HCoRR, abs/2001.03447, 2020*.

Lundberg, *et al.* A Unified Approach to Interpreting Model Predictions, *NIPS'17, 4765–4774*.

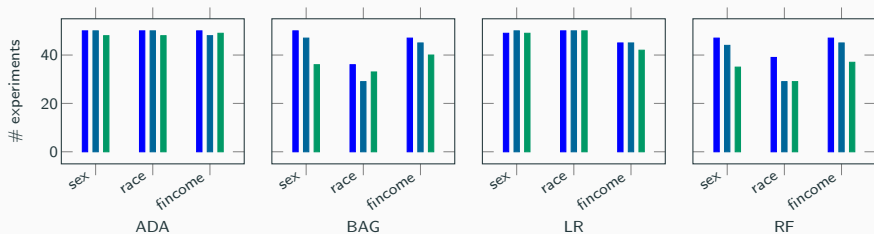Ribeiro, *et al.* "Why Should I Trust You?": Explaining the Predictions of Any Classifier, *SIGKDD'16, 1135–1144*.

(c) **LSAC dataset**

Fairness assessment: FixOut with SHAP

Original+SHAP ■ *FixOut* +SHAP ■ *FixOut* (w)+SHAP
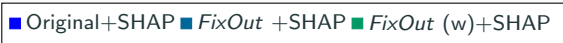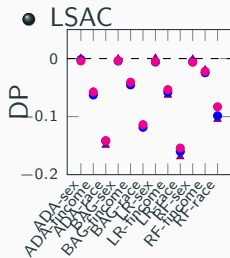
(c) **LSAC dataset**

Fairness metrics
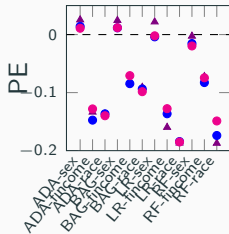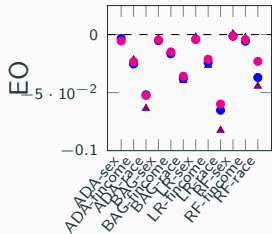
(Extra) LIME explanations of Bagging on HMDA dataset

Original		Ensemble	
Feature	**Contrib.**	**Feature**	**Contrib.**
derived_loan_product_type	4.798847	derived_loan_product_type	6.457707
balloon_payment_desc	4.624029	balloon_payment_desc	5.054243
intro_rate_period	4.183828	intro_rate_period	4.638744
loan_to_value_ratio	2.824717	balloon_payment	1.512304
balloon_payment	2.005847	prepayment_penalty_term	-1.267424
prepayment_penalty_term	0.683618	interest_only_payment	0.777766
reverse_mortgage	-0.659169	loan_to_value_ratio	0.704758
applicant_age_above_62	0.532331	negative_amortization_desc	0.61936
derived_ethnicity	-0.409255	reverse_mortgage_desc	0.508204
co_applicant_age_above_62	-0.333838	interest_only_payment_desc	-0.393068
property_value	-0.326801	applicant_credit_score_type_desc	-0.379852
derived_race	-0.318802	negative_amortization	-0.353717
applicant_age	-0.304565	applicant_age_above_62	0.349847
loan_term	0.270951	property_value	-0.316311
negative_amortization	-0.229379	applicant_credit_score_type	-0.192114

What about Fairness metrics? (further metrics)

- **Equal Accuracy (EA)**[6]:

$$EA = \frac{TP_{unp} + TN_{unp}}{P_{unp} + N_{unp}} - \frac{TP_{priv} + TN_{priv}}{P_{priv} + N_{priv}}$$

- **Disparate Impact (DI)**[7] (or *Group Fairness*):

$$DI = \frac{P(\hat{y} = pos | D = unp)}{P(\hat{y} = pos | D = priv)}$$

[6] Hardt, *et al.* Equality of Opportunity in Supervised Learning, *NIPS'16, 3315–3323*.

[7] Dwork, *et al.* Fairness through awareness, *Innovations in Theoretical Computer Science, 2012, 214–226*.

Fairness assessment: FixOut with SHAP (cont.)