Making ML fairer through explanations, feature dropout, and aggregation

3 of February 2021, Seminar in MPML

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** ...

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.

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.:

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

   Drawback: Complexity, fairness "gerrymandering" & overfitting

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

   Drawback: Decreased accuracy!

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

- $\text{Exp}_{\text{Global}}$: for global explanations (FI)
- $\text{Ensemble}_{\text{Out}}$: 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

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

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

$\mathbf{Exp_{Global}}$: model M, dataset D, sensitive F, exp. method E

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 are **at least two** sensitive features among the top-k, **then** M is deemed unfair and $\mathbf{Ensemble_{Out}}$ applies.

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

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

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

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

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

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

FixOut with LIME explanations

Sample of 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

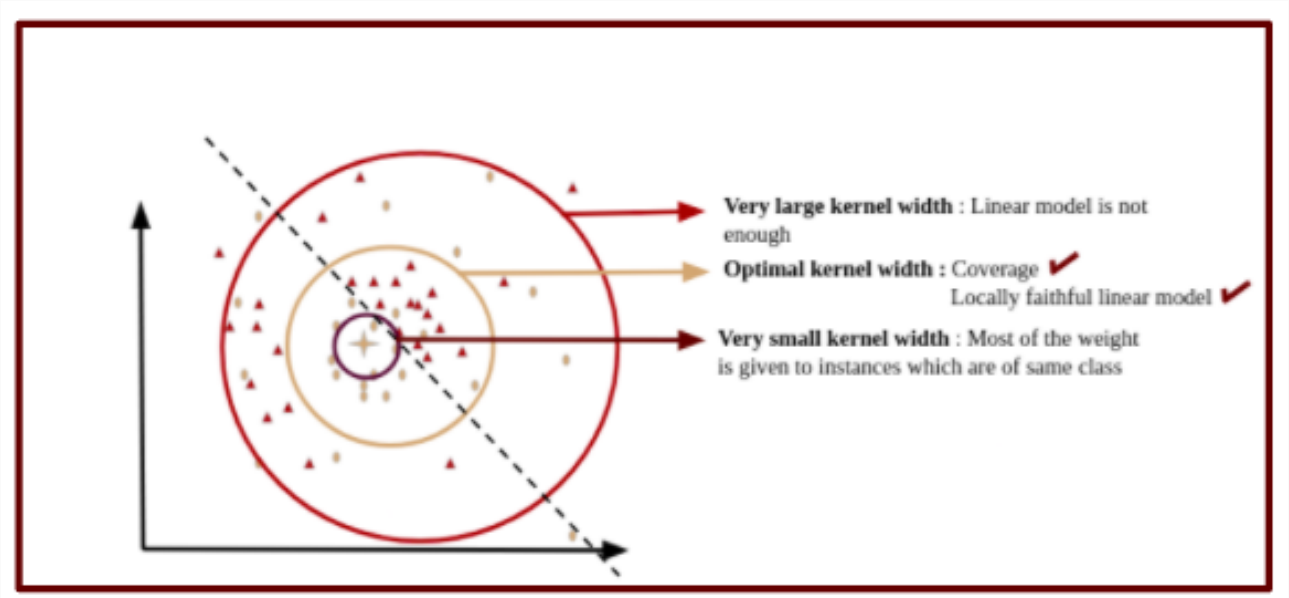


Figure 1: Illustration of optimal kernel on the (interpretable) space

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

Sample of 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 $\hat{x}_i$

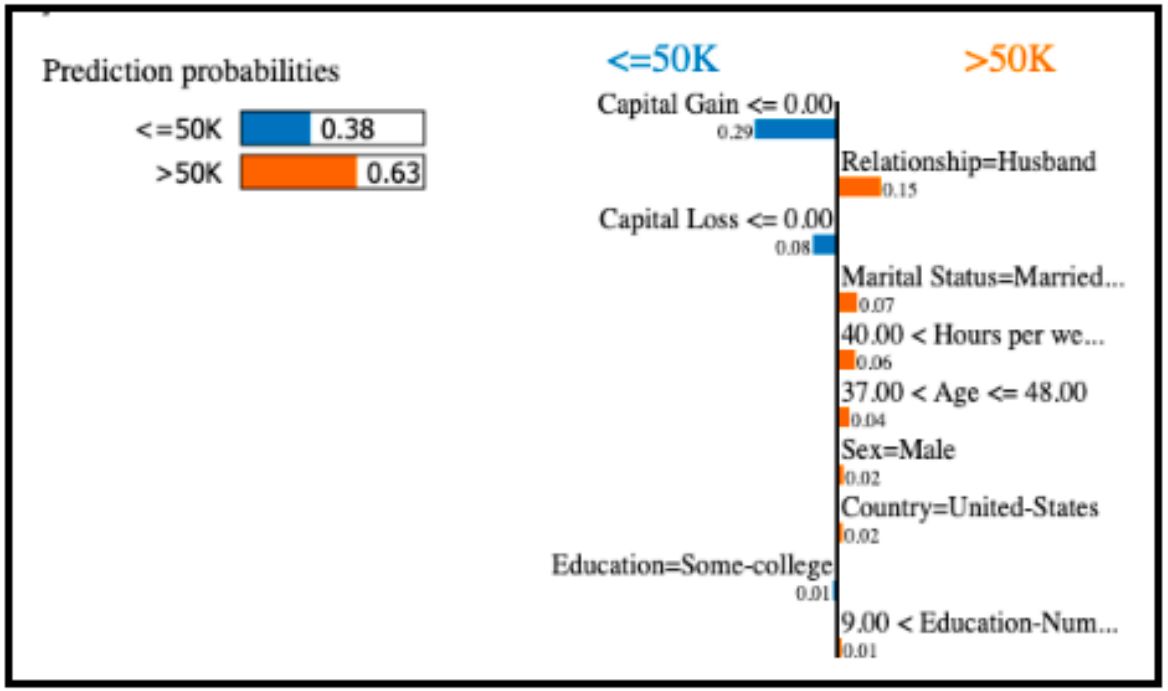


Figure 2: Local explanation in case of Adult dataset (salary prediction)

FixOut with LIME explanations and simple averages

$\text{Exp}_{\text{Global}}$: LIME + sub-modular pick
(to sample a set of instances for which explanations are diverse and non-redundant, and use them to get global explanations)

As before: if $\text{Exp}_{\text{Global}}$ 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 $a_{j_1}, a_{j_2}, \ldots, a_{j_i}$.

$\text{Ensemble}_{\text{Out}}$: Ensemble classifier M_{final} defined as a simple average

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

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 (Scikit-learn implementation)
- 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 (Scikit-learn implementation)
- Accuracy of M: 0.783

Question: Is this model fair?

$\text{Exp}_{\text{Global}}$ 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

$\text{Ensemble}_{\text{Out}}$

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!

$\mathbf{M_{final}}$**:** Ensemble of M1, M2 and M3 (accuracy of 0.786)

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

Original	
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

Ensemble	
Feature	**Contribution**
otherinstallmentplans	-1.487604
housing	-1.089726
savings	0.679195
duration	-0.483643
foreignworker	0.448643
property	-0.386355
credithistory	0.258375
job	-0.252046
existingchecking	-0.21358
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. Home Mortgage Disclosure Act (HMDA)
   - Public data about home mortgages.
   - Goal: Predict whether a loan is "high-priced".
   - Sensitive: 'sex', 'race', 'ethnicity'.
3. Law School Admissions Council (LSAC)
   - Student profiles (demographic, socio-economic, etc.).
   - Goal: Predict whether a law student passes"bar exam"
   - Sensitive: 'Race' and 'sex'

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

Average accuracy assessment (FixOut with LIME)

		ADA	BAG	RF	LR
German	Original	0.757 (0.015)	0.743 (0.019)	**0.772** (0.016)	0.769 (0.021)
	FixOut	0.765 (0.014)	0.755 (0.021)	**0.769** (0.016)	0.770 (0.021)
Adult	Original	**0.855** (0.003)	0.841 (0.002)	0.808 (0.007)	0.845 (0.004)
	FixOut	**0.856** (0.003)	0.849 (0.002)	0.808 (0.004)	0.849 (0.004)
LSAC	Original	0.857 (0.003)	**0.861** (0.002)	0.852 (0.002)	0.820 (0.006)
	FixOut	0.859 (0.002)	**0.866** (0.002)	0.859 (0.002)	0.822 (0.005)
HMDA	Original	0.879 (0.001)	**0.883** (0.001)	0.882 (0.001)	0.878 (0.001)
	FixOut	0.880 (0.001)	**0.884** (0.000)	0.884 (0.000)	-
Default	Original	**0.817** (0.003)	0.804 (0.003)	0.807 (0.003)	0.779 (0.004)
	FixOut	**0.817** (0.003)	0.812 (0.002)	-	-

Default: Similar to German but from Taiwanese credit card users.
Goal: payment defaults. Sensitive features: "sex" and "marriage".

FixOut with SHAP explanations

Sample of SHAP Explanations[2]

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)

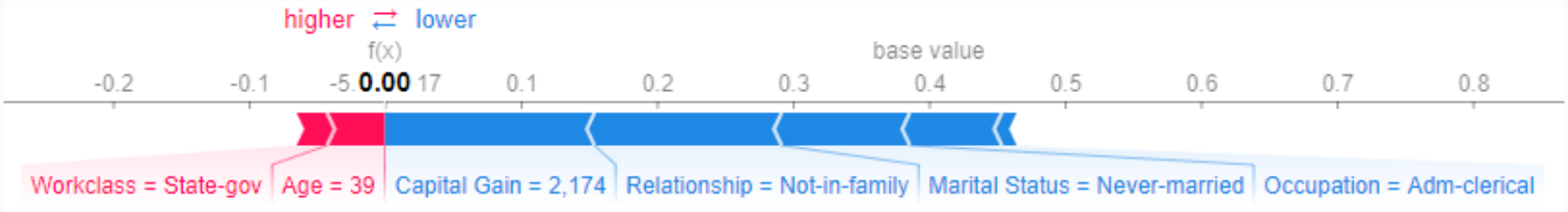


Figure 3: SHAP explanation in case of Adult dataset (salary prediction)

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

FixOut with SHAP explanations and weighted averages

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

As before: if $\text{Exp}_{\text{Global}}$ 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 $a_{j_1}, a_{j_2}, \ldots, a_{j_i}$.

$\text{Ensemble}_{\text{Out}}$: 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.

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?

$\text{Exp}_{\text{Global}}$ 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

$\text{Ensemble}_{\text{Out}}$

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!

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

$\text{Exp}_{\text{Global}}$ Explanations (RF on German)

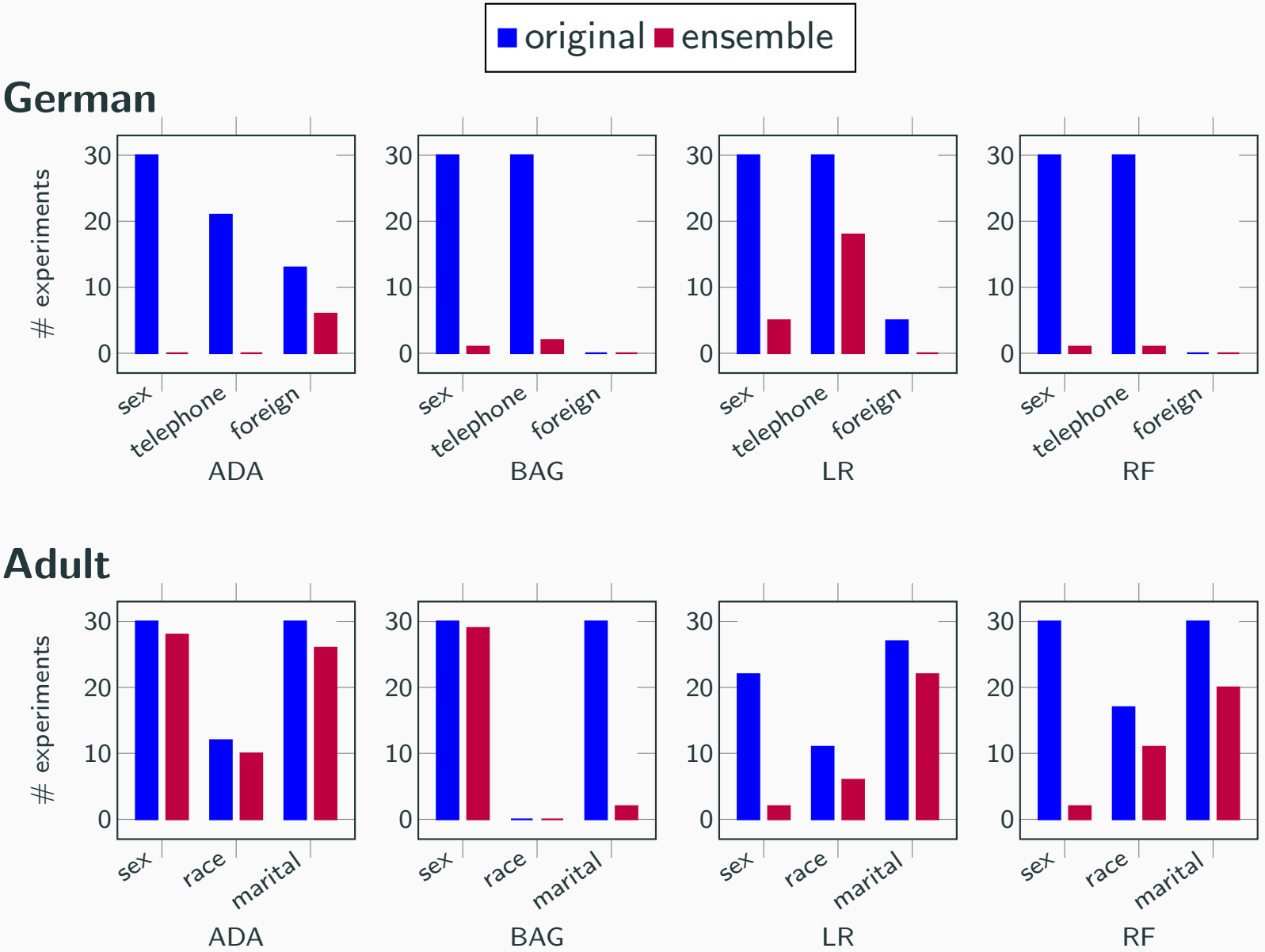
Original		Ensemble	
Feature	**Contribution**	**Feature**	**Contribution**
existingchecking	-7.11624	existingchecking	-4.285092
statussex	-5.950176	housing	-3.771932
housing	-3.27344	property	3.506007
job	-2.868195	job	-3.061209
residencesince	2.832573	employmentsince	2.646814
telephone	2.290478	existingcredits	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 0.746 to 0.760)

Average accuracy assessment (FixOut with SHAP)

		ADA	BAG	RF	LR
German	Original	*0.754* (.017)	*0.742* (.021)	0.765 (.018)	0.764 (.020)
	FixOut	*0.758* (.018)	*0.761* (.018)	0.766 (.014)	0.761 (.020)
Adult	Original	*0.854* (.003)	*0.841* (.003)	*0.846* (.003)	0.807 (.006)
	FixOut	*0.856* (.003)	*0.845* (.003)	*0.848* (.003)	0.805 (.003)
HMDA	Original	0.880 (.001)	*0.883* (.001)	*0.882* (.001)	0.878 (.001)
	FixOut	0.880 (.001)	*0.884* (.001)	*0.883* (.001)	0.878 (.001)
LSAC	Original	0.857 (.003)	*0.860* (.003)	*0.853* (.003)	*0.818* (.005)
	FixOut	0.857 (.003)	*0.862* (.002)	*0.858* (.003)	*0.820* (.004)
Default	Original	0.817 (.003)	-	-	-
	FixOut	0.819 (.003)	-	-	-

Fairness assessment: FixOut with SHAP



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

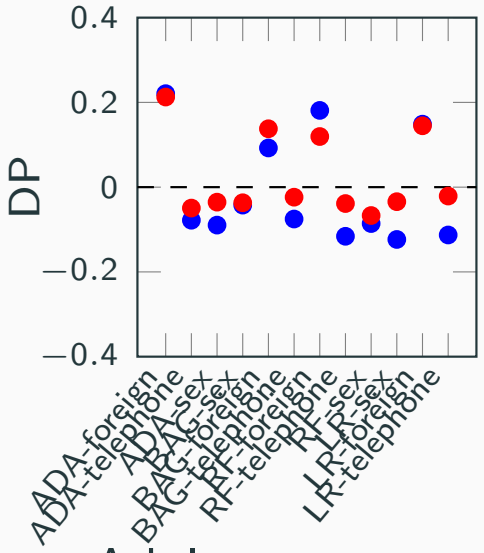
What about Fairness metrics? (cont.)

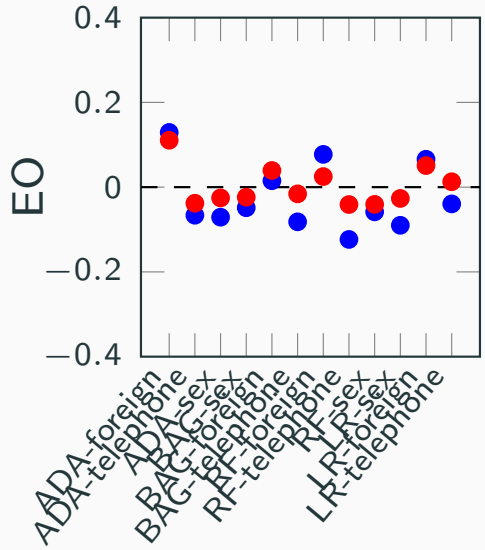
Privileged groups

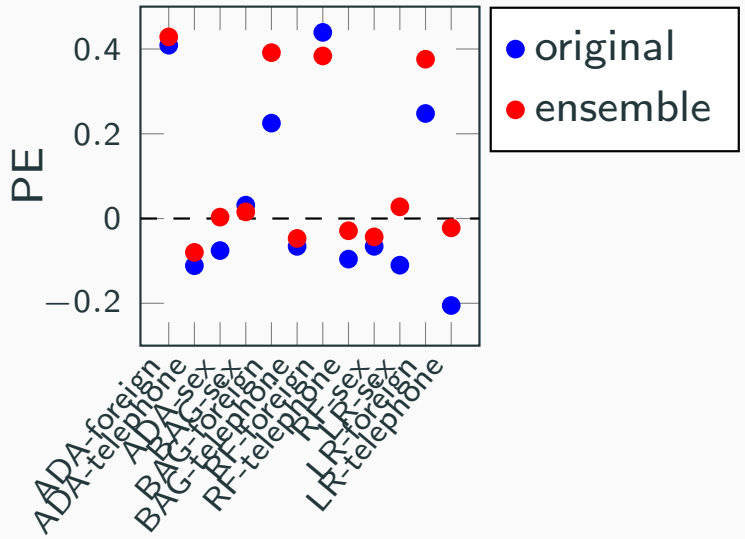
1. German dataset
   - "status sex": "male single"
   - "telephone": "yes" (registered under the customers name)
   - "foreign worker": "no"
2. Adult dataset
   - "Marital Status": "Married"
   - "Race": "white"
   - "Sex": "male"

Fairness metrics: FixOut with LIME

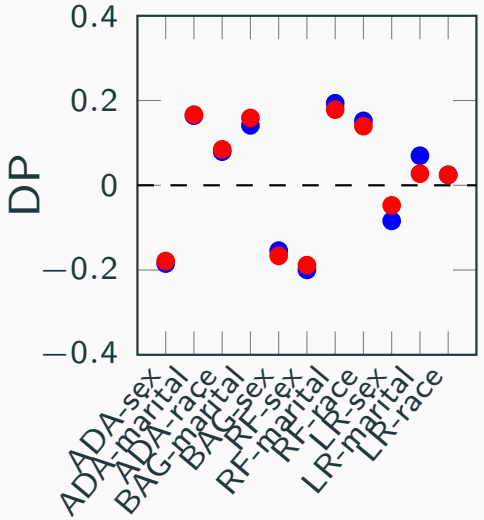
- German

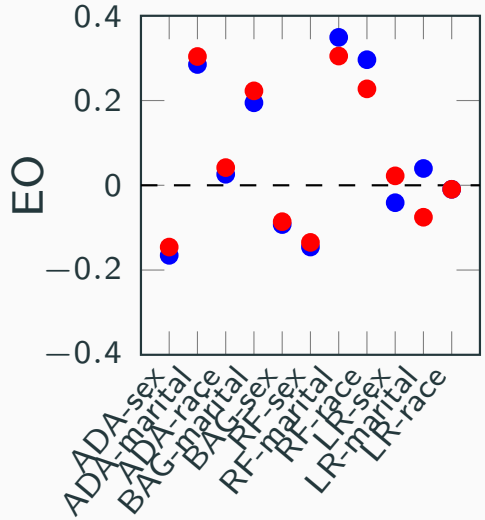


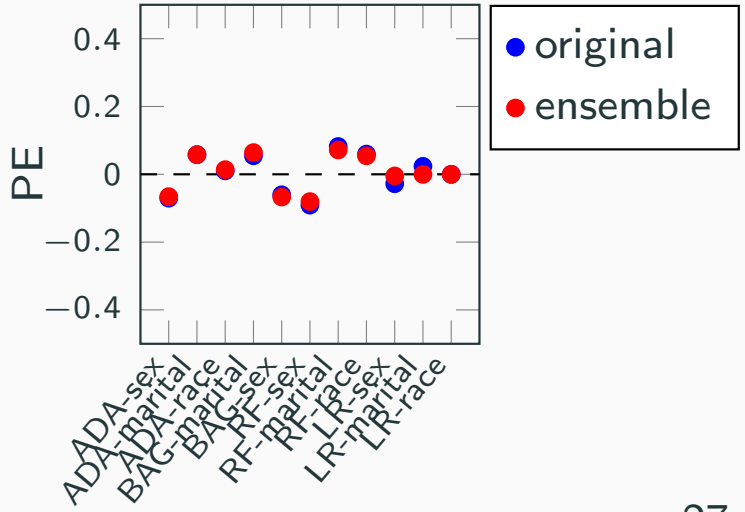




- Adult

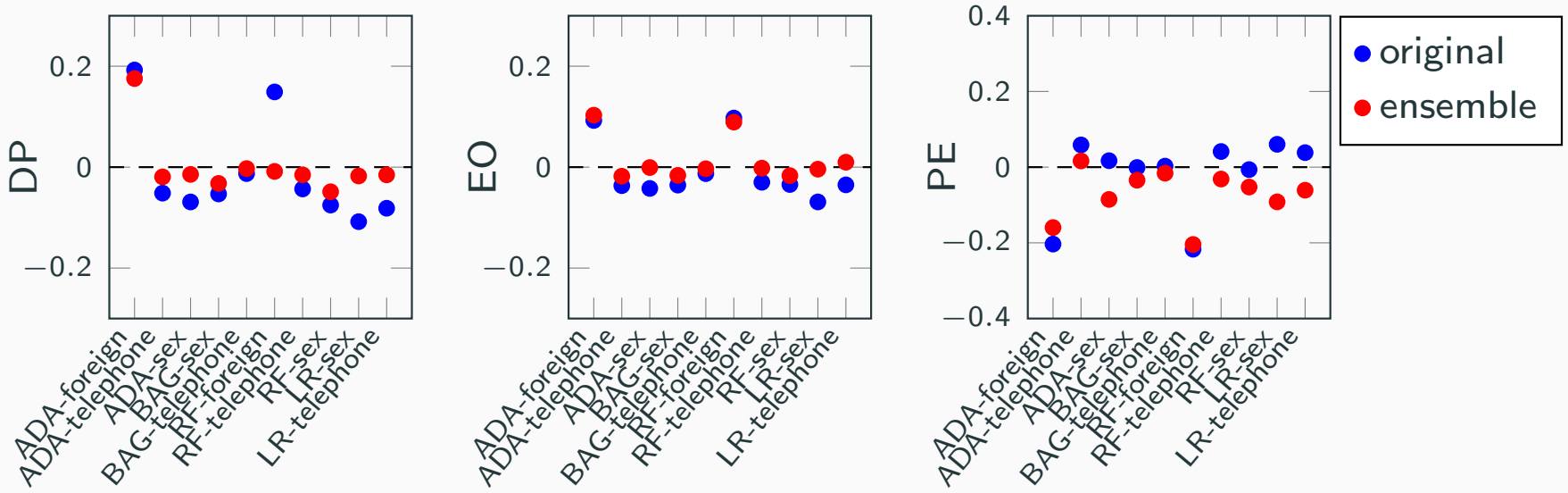




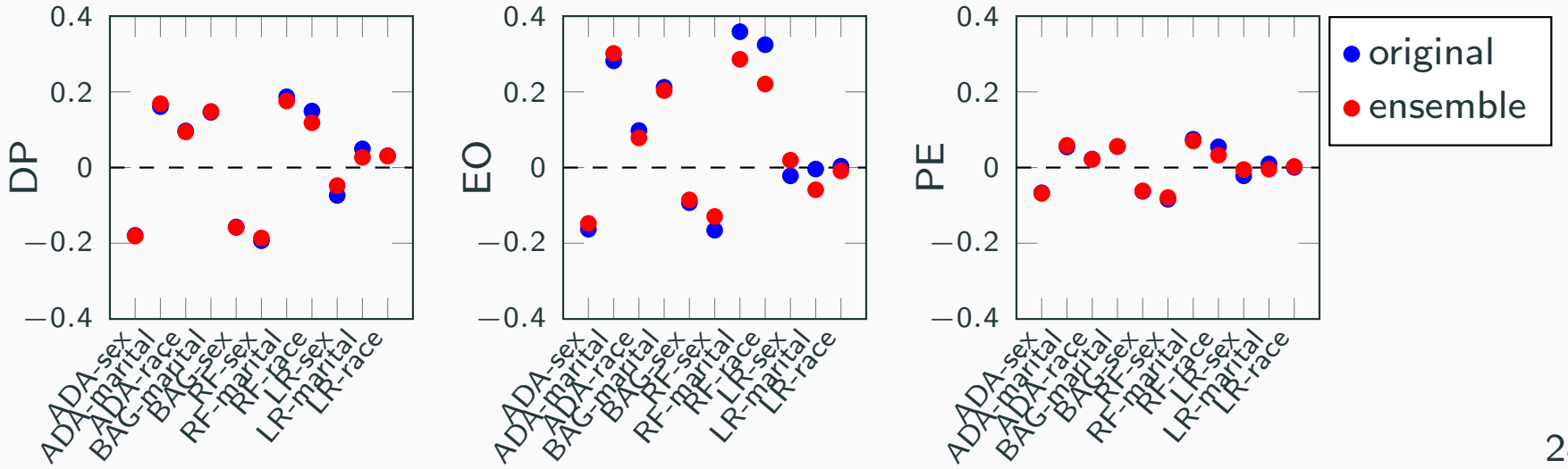


Fairness metrics: FixOut with SHAP

- German



- Adult



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 $\text{Ensemble}_{\text{Out}}$ 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!

Obrigado pela vossa atenção!

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*.

(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.)