Reducing Unintended Bias of ML Models on Tabular and Textual Data

Café TAL

Guilherme Alves Maxime Amblard Fabien Bernier
Vaishnavi Bhargava Miguel Couceiro Amedeo Napoli

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

Unintended bias of ML Models

ML models: *designed* to have some bias that *guide* them in their tasks

Expected bias

Credit card default prediction	(good) *credit payment history*	$\uparrow$
Hate speech prediction	(presence of) *offensive terms*	$\uparrow$

Unintented bias

Credit card default prediction	*ethnicity* (minority)	$\downarrow$
Hate speech prediction	*language variant*	$\downarrow$

Unexpected bias can lead to **unfair** algorithmic decisions and **discrimination**!

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

Unintended bias of ML Models

ML models: *designed* to have some bias that *guide* them in their tasks

Expected bias

Credit card default prediction	(good) *credit payment history*	$\uparrow$
Hate speech prediction	(presence of) *offensive terms*	$\uparrow$

Unintented bias

Credit card default prediction	*ethnicity* (minority)	$\downarrow$
Hate speech prediction	*language variant*	$\downarrow$

Unexpected bias can lead to **unfair** algorithmic decisions and **discrimination**!

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

Unintended bias of ML Models

ML models: *designed* to have some bias that *guide* them in their tasks

Expected bias

Credit card default prediction	(good) *credit payment history*	$\uparrow$
Hate speech prediction	(presence of) *offensive terms*	$\uparrow$

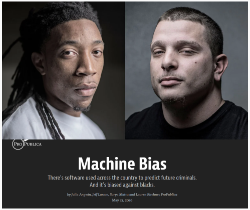
Unintented bias

Credit card default prediction	*ethnicity* (minority)	$\downarrow$
Hate speech prediction	*language variant*	$\downarrow$

Unexpected bias can lead to **unfair** algorithmic decisions and **discrimination**!

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

Motivation: unfair algorithmic decisions



COMPAS[1] (Tabular data)

Chatbot Tay[2] (Text)

Other Critical applications of algorithmic decisions: loan requests, job applications, Stop & Frisk, etc.

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

---

[1] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] https://www.bbc.com/news/technology-35902104

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

   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

   Drawback: Decreased accuracy!

FixOut:
FaIrness through eXplanations and feature dropOut

FixOut (FaIrness through eXplanations and feature dropOut)

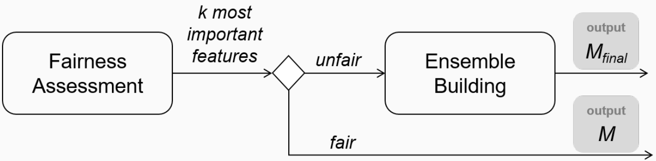
Goal: reduce model's dependence on sensitive/salient features while keeping (or improving) its classification performance

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

FixOut: Human-centered approach to deal process fairness

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

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



FixOut, 1st step: Fairness Assessment

Idea: Use explanations to assess model's dependence sensitive feat.s

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 Sampling (RS) or "Submodular pick" (SP)

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 FixOut builds an ensemble.

FixOut, 2nd step: Ensemble Building

Idea: Use feature dropout follow by an ensemble approach

Let $a_1, a_2, \ldots, a_k$ be the k most important features

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.

For an instance x and a class C, M_{final} is defined as a **simple average**

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

Example with LIME explanations

FixOut with LIME explanations

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

As before: if 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 $a_{j_1}, a_{j_2}, \ldots, a_{j_i}$.

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

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

FixOut with LIME: RF on German dataset

German Credit Card Score (UCI):

- Goal: Predict credit risks (likely & unlikely to pay back)
- Applicant profiles (demographic and socio-economic).
- 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?

FixOut with LIME: RF on German dataset

German Credit Card Score (UCI):

- Goal: Predict credit risks (likely & unlikely to pay back)
- Applicant profiles (demographic and socio-economic).
- 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?

FixOut with LIME: RF on German dataset ($\text{Exp}_{\text{Global}}$)

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

Hence: Model deemed unfair

FixOut with LIME: RF on German dataset ($\text{Ensemble}_{\text{Out}}$)

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)

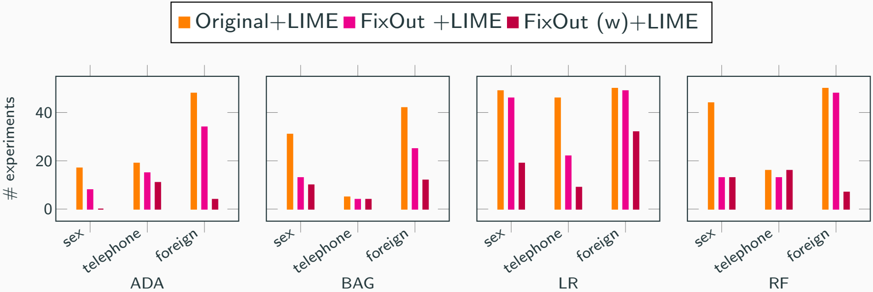
FixOut with LIME: RF on German dataset

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)

Fairness & Classification assessment (German dataset)



Classification assessment

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

What about Fairness metrics?

- Separate instances into two groups based on one sensitive feature

  Unprivileged group (*unp*) **versus** privileged group (*priv*)

  Example: female **versus** male

- **Demographic Parity (DP)**:

$$DP = P(\hat{y} = pos | D = unp) - P(\hat{y} = pos | D = priv)$$

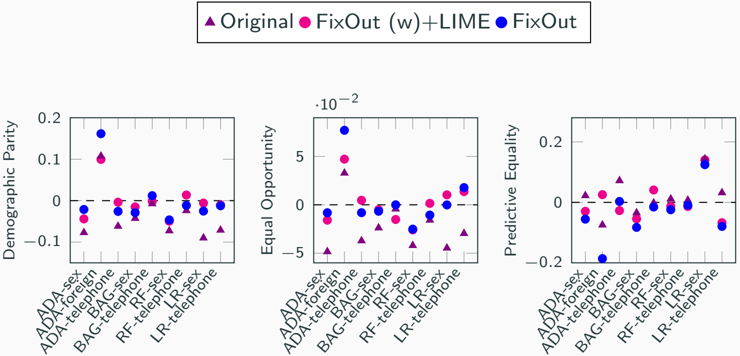
- **Equal Opportunity (EO)**:

$$EO = \frac{TP_{unp}}{TP_{unp} + FN_{unp}} - \frac{TP_{priv}}{TP_{priv} + FN_{priv}}$$

- **Predictive Equality (PE):**

$$PE = \frac{FP_{unp}}{FP_{unp} + TP_{unp}} - \frac{FP_{priv}}{FP_{priv} + TP_{priv}}.$$

Assessment w.r.t. some fairness metrics (German dataset)



Example with SHAP explanations

FixOut with SHAP: RF on German dataset (Exp_{Global})

Same dataset and empirical setting...

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

FixOut with SHAP: RF on German dataset (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
  NB: Performance drop when all sensitive features are removed!

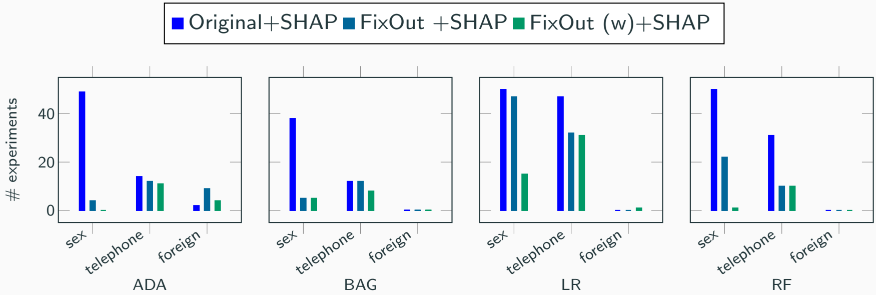
M$_{\text{final}}$: Ensemble of M1, M2 and M3

FixOut with SHAP: RF on German dataset

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 & better performance

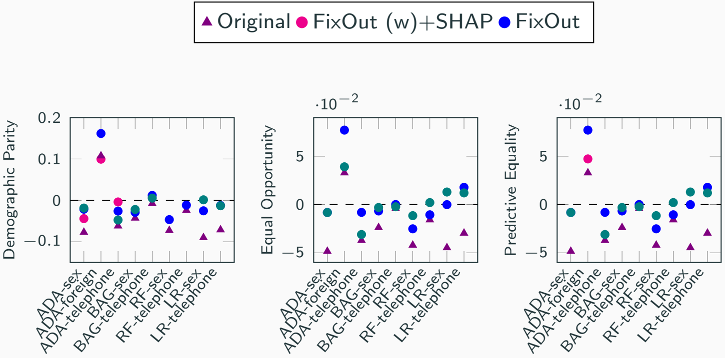
Fairness & Classification assessment (German dataset)



Classification assessment

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

Assessment w.r.t. some fairness metrics (German dataset)

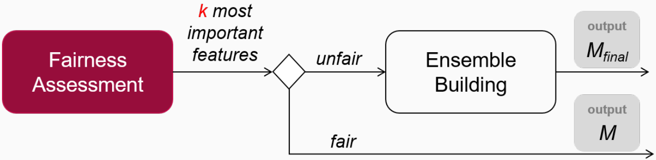


Comparison: Average contribution of sensitive features

	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

How to reduce human intervention in FixOut on tabular data?

Suitable value for k



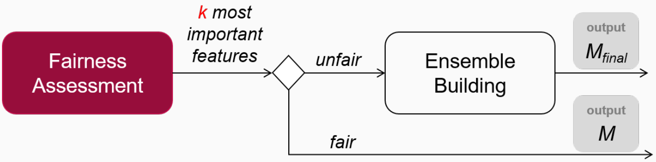
Recall: FixOut builds L with the k most important features

Problem: Practitioners must know beforehand a suitable value for k

An algorithm that automatically finds a value for k

Idea: Kurtosis indicates the "*flatness*" of a distribution

Suitable value for k



Recall: FixOut builds L with the k most important features

Problem: Practitioners must know beforehand a suitable value for k

An algorithm that automatically finds a value for k

Idea: Kurtosis indicates the "*flatness*" of a distribution

Suitable value for k (cont.)

- **Input**: L: sorted list of contributions of all features (descending order); α: a threshold
- **Output**: L' a new list of contributions of subset of features

Find-K algorithm

- Iterative algorithm
- Remove features from L
- Stop when $|\gamma(L) - \gamma(L')| > \alpha$

- α encodes the **accepted perturbation** in L
- Single value of α, **multiple** values of k

Suitable value for k (cont.)

- **Input**: L: sorted list of contributions of all features (descending order); α: a threshold
- **Output**: L' a new list of contributions of subset of features

Find-K algorithm

- Iterative algorithm
- Remove features from L
- Stop when $|\gamma(L) - \gamma(L')| > \alpha$

- α encodes the **accepted perturbation** in L
- Single value of α, **multiple** values of k

Results - Average value of k

Data.	Selection	Random Forest						AdaBoost					
		α						α					
		0.5	1	1.5	2	2.5	3	0.5	1	1.5	2	2.5	3
German	*LIME+RS*	9.90	8.30	5.18	2.72	1.54	1.18	10	10	9.76	9.4	9.2	9.04
	LIME+SP	10.0	9.98	9.74	8.54	6.96	5.46	10.0	10.0	10.0	10.0	10.0	10.0
	SHAP+RS	9.92	8.98	6.46	4.52	2.74	2.04	10.0	8.78	6.44	4.28	3.24	2.90
	SHAP+SP	9.86	8.70	5.64	3.68	2.28	1.42	9.98	8.68	6.78	5.82	4.92	4.04
Adult	*LIME+RS*	9.76	8.38	7.00	5.48	4.44	3.64	10.0	10.0	8.22	5.54	3.40	2.20
	LIME+SP	9.30	7.80	6.76	5.80	5.00	4.32	10.0	9.90	7.98	5.74	3.84	2.48
	SHAP+RS	10.0	9.96	9.30	8.12	6.48	5.14	10.0	9.02	6.62	4.76	3.28	2.38
	SHAP+SP	10.0	9.98	9.38	8.02	6.26	4.94	10.0	9.16	7.22	5.36	4.06	2.82
LSAC	*LIME+RS*	6.46	4.04	2.02	1.46	1.10	1.02	6.98	4.52	3.02	1.92	1.28	1.12
	LIME+SP	8.92	7.30	5.38	3.66	2.66	1.96	7.08	5.06	3.44	2.22	1.78	1.28
	SHAP+RS	7.80	5.80	3.88	2.48	1.76	1.40	8.68	6.52	4.78	3.34	2.08	1.52
	SHAP+SP	8.18	5.78	4.04	2.86	2.10	1.70	8.84	7.08	5.42	3.84	2.68	1.90

- $0.5 \leq \alpha \leq 1$ allows FIND-K to find a suitable value of k
- $\alpha < 0.5$ (and closer to 0) all features are removed
- $\alpha > 3$ all features are kept
- Similar results with Logistic Regression, Bagging

Does FixOut reduce model's dependence on sensitive words on textual data?

Example: FixOut on a hate speech classifier

- **Goal**: Classify tweets as *hate speech* or *not*
- **Idea**: Bag of Words (BoW) (**Or:** Groups of words)
- **Dataset**: *Hate speech* dataset [3]

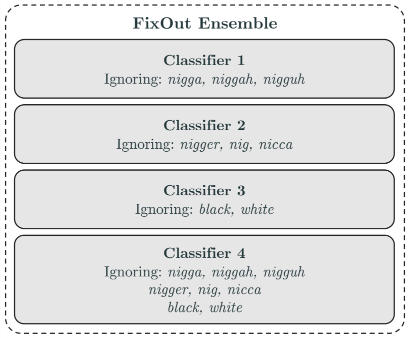


Illustration of textual classifiers used in the ensemble.

[3] Davidson et al. Automated hate speech detection and the problem of offensive language. AAAI. 2017

Textual data: FixOut on a hate speech classifier

Setting: RF classifier, SHAP explanations, RS and BoW

	Without grouping		With grouping	
Word	**Rank**	**Contrib.**	**Rank**	**Contrib.**
niggah	18	0.149	23	0.03
nigger	15	0.164	21	0.031
nigguh	22	0.13	83	0.008
nig	12	0.202	65	0.011
nicca	22	0.107	39	0.018
nigga	20	0.125	12	0.067
white	25	0.087	36	0.018

Textual data: FixOut on a hate speech classifier

Process fairness assessment on textual data with LIME and SP

	Word	Original model **Rank**	**Contrib.**	FixOut Ensemble **Rank**	**Contrib.**
RF, LIME+SP	*niggah*	7	0.517	12	0.257
	nigger	9	0.476	15	0.23
	nigguh	13	0.339	17	0.194
	nig	10	0.445	16	0.204
	nicca	16	0.265	20	0.121
	nigga	17	0.235	23	0.112
	white	23	0.127	34	0.07
	black	$>$500	$\sim$ 0	$>$500	$\sim$ 0
ADA, LIME+SP	*niggah*	2	0.167	4	0.083
	nigger	7	0.052	10	0.026
	nigguh	5	0.144	6	0.073
	nig	18	0.014	24	0.006
	nicca	17	0.015	23	0.007
	nigga	4	0.166	5	0.083
	white	23	0.011	26	0.005
	black	113	0.0	196	0.0

Similar results with Logistic Regression and Bagging

Conclusion

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.

Improve process fairness on tabular and textual data!

Thank you for your attention!

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

References

Alves, *et al.* Reducing Unintended Bias of ML Models on Tabular and Textual Data, *DSAA'21*.

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* @ECML-PKDD.

Davidson, *et al.* Automated hate speech detection and the problem of offensive language, *ICWSM'17*.

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*.

LIME Explanations[4]

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

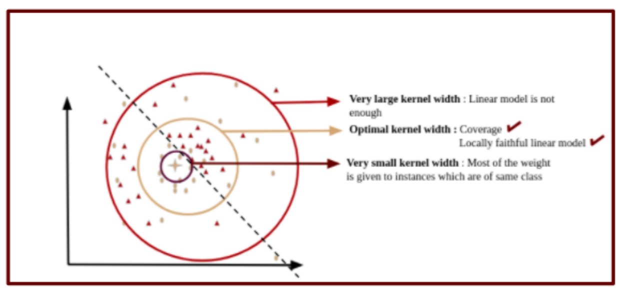


Figure 1: Illustration of optimal kernel on the (interpretable) space

[4] 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 $\hat{x}_i$

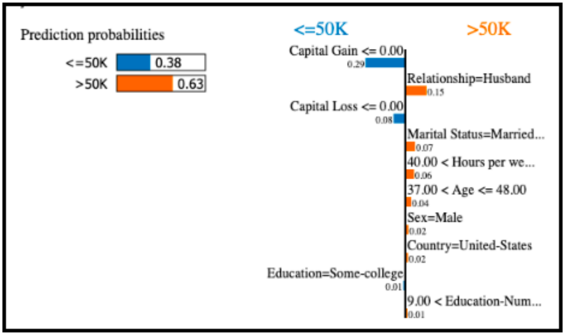


Figure 2: Local explanation in case of Adult dataset (salary prediction)

SHAP Explanations[5]

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)

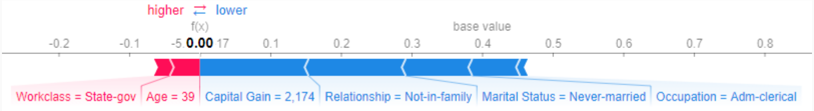


Figure 3: SHAP explanation in case of Adult dataset (salary prediction)

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