Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods

Maciej Pióro, Krzysztof Tomala

Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods

Dylan Slack*
University of California, Irvine
dslack@uci.edu

Sophie Hilgard*
Harvard University
ash798@g.harvard.edu

Emily Jia
Harvard University
ejia@college.harvard.edu

Sameer Singh
University of California, Irvine
sameer@uci.edu

Himabindu Lakkaraju
Harvard University
hlakkaraju@seas.harvard.edu

Presentation outline

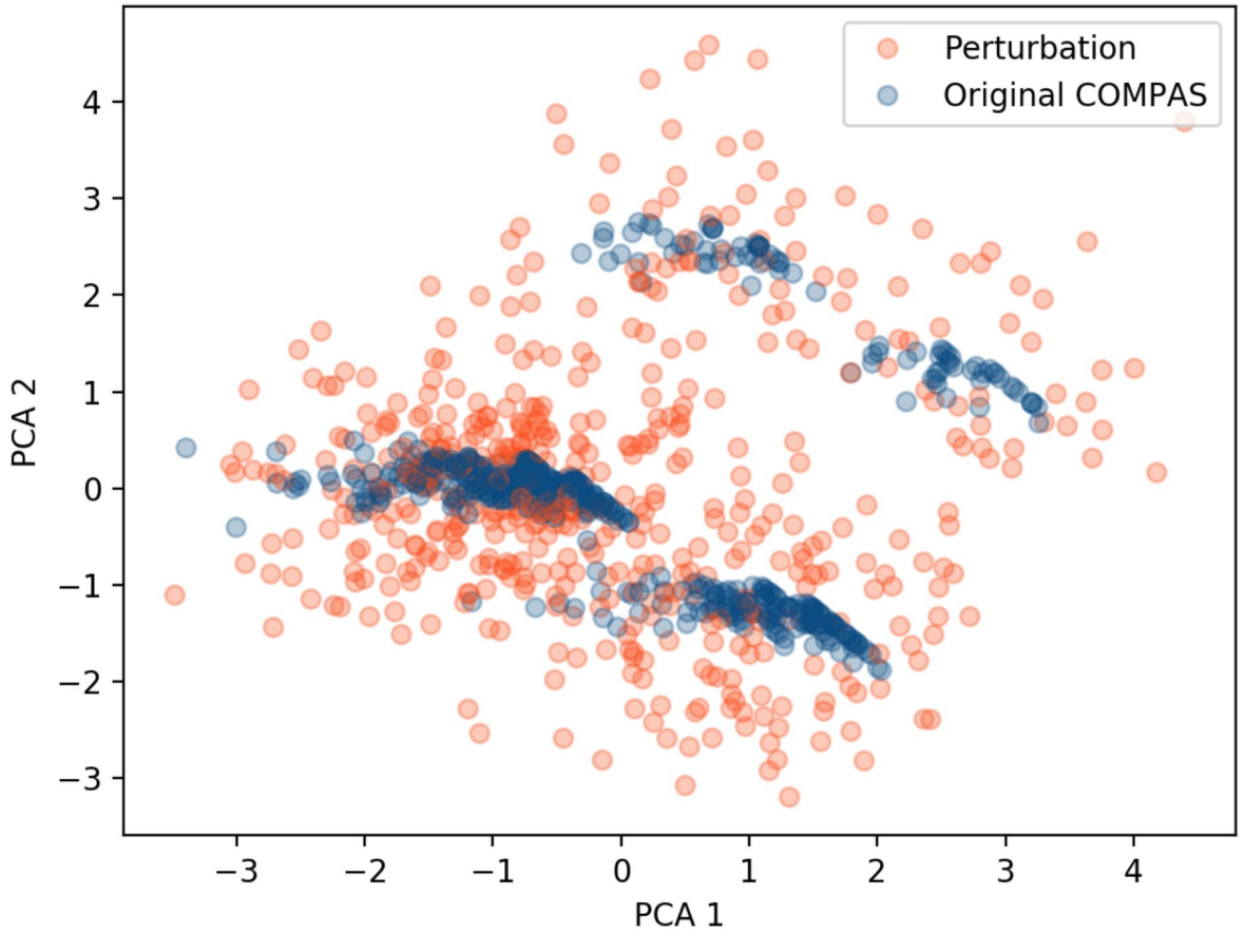
- Problem statement
- Intuition behind proposed solution
- Technical solution
- Experiments
- Conclusions

Setup

- A dataset is given with some sensitive attribute (e. g. race, gender)
- Adversary provides a classifier (*biased classifier*)
- The training datapoints come from a distribution unknown to the adversary
- Customer / regulator uses a dataset from the same distribution (*train* / *test*) to explain the classifier with LIME / SHAP
- We want the explanation not to indicate the classifier is biased (*fool* it)

Intuition

- LIME / SHAP construct local linear interpretable approximations of a black box model based on perturbed inputs
- Perturbed datapoints are often OOD (out of distribution) - they are clearly visible after dimensionality reduction using PCA


Perturbation
Original COMPAS
PCA 2
PCA 1
4
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3
4

Intuition - cont’d

- We can create a classifier that is biased on real datapoints and fair on perturbed datapoints (*scaffolding*)

Building an adversarial classifier

- ψ - fair classifier using features uncorrelated with the sensitive features, f - biased classifier, e - adversarial classifier

$$e(x) = \begin{cases} f(x), & \text{if } x \in \mathcal{X}_{dist} \\ \psi(x), & \text{otherwise} \end{cases}$$

Detecting OOD samples

- We are given a training set
- Perturb points in the dataset, add synthetic datapoints to the dataset
- Assign label TRUE to the synthetic datapoints, FALSE otherwise
- Train a classifier to discern between synthetic and real datapoints

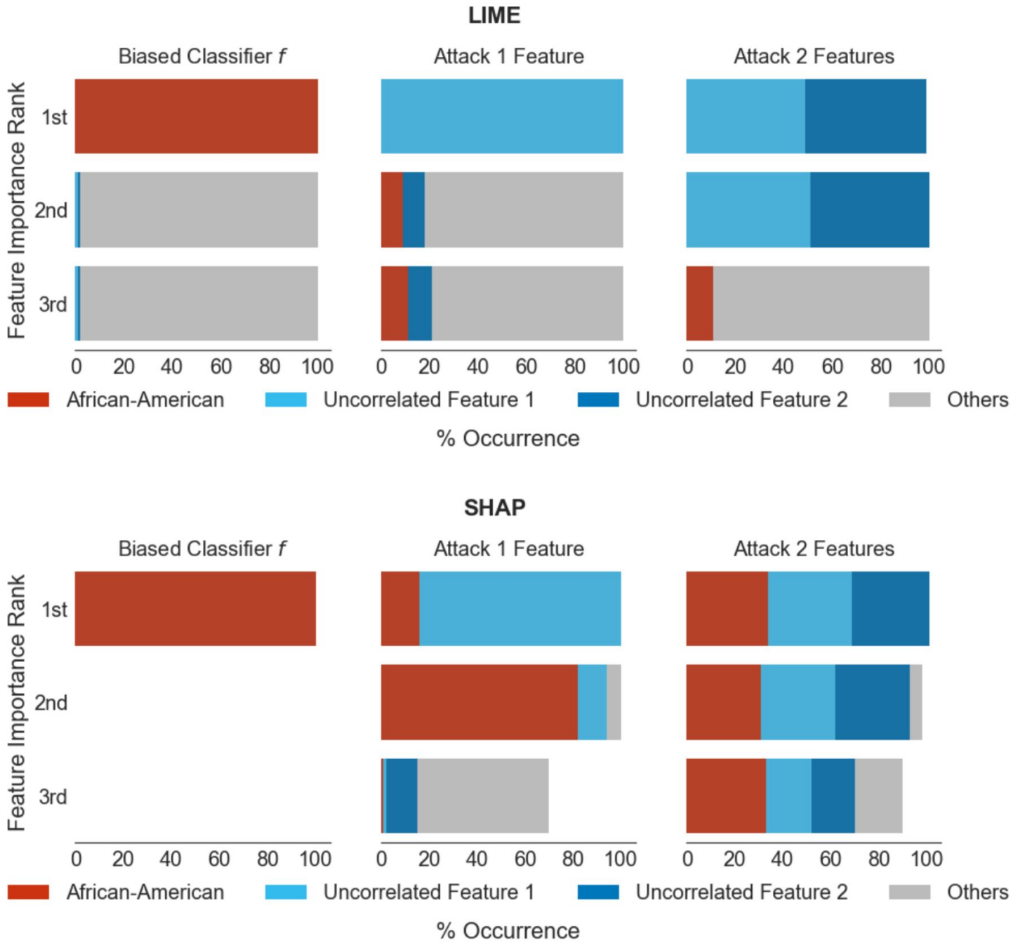
Datasets

Dataset	Size	Features	Positive Class	Sensitive Feature
COMPAS	6172	*criminal history, demographics, COMPAS risk score, jail and prison time*	High Risk (81.4%)	African-American (51.4%)
Communities & Crime	1994	*race, age, education, police demographics, marriage status, citizenship*	Violent Crime Rate (50%)	White Population (continuous)
German Credit	1000	*account information, credit history, loan purpose, employment, demographics*	Good Customer (70%)	Male (69%)

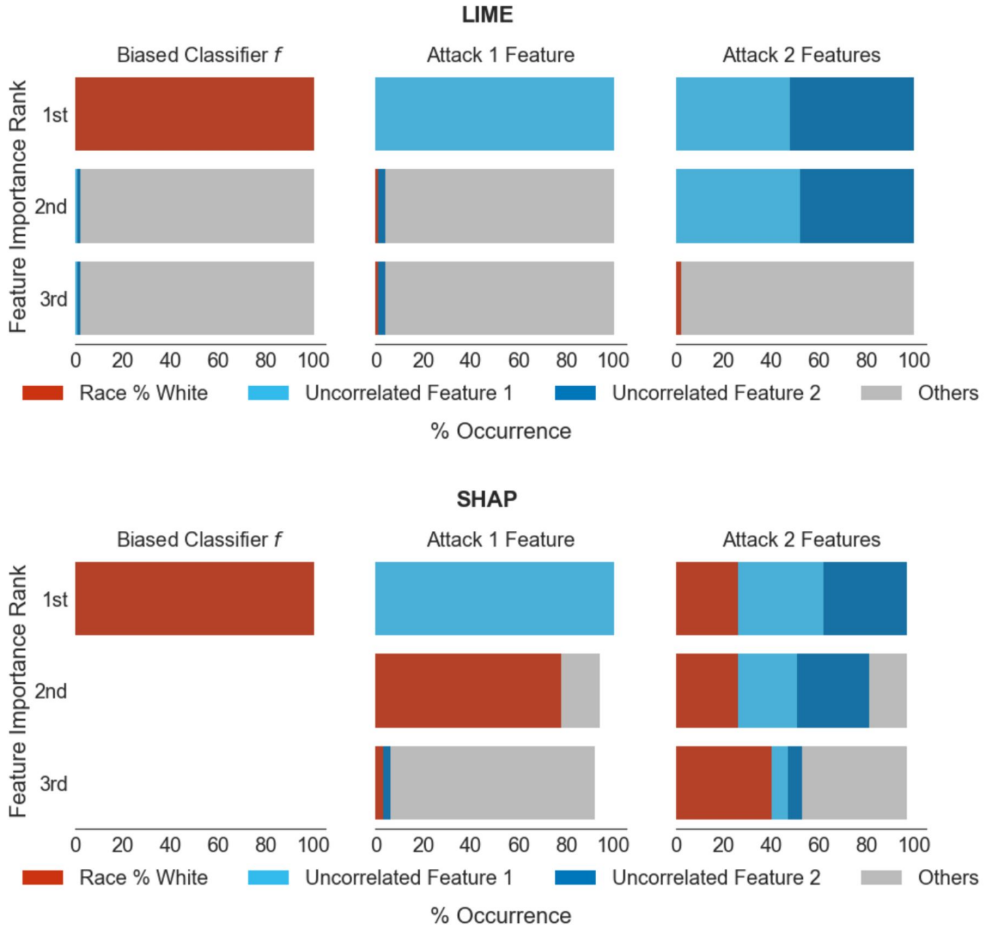
Experiments

- Biased classifier is making prediction based purely on a sensitive feature
- For LIME, we generate perturbations by adding random noise
- For SHAP, we randomly choose a subset of features for each data point and mark their values as missing by replacing them with their corresponding values from background distribution
- OOD classifier is a random forest with 100 trees
- Unbiased classifier is making prediction based purely on one or two features uncorrelated with a sensitive feature

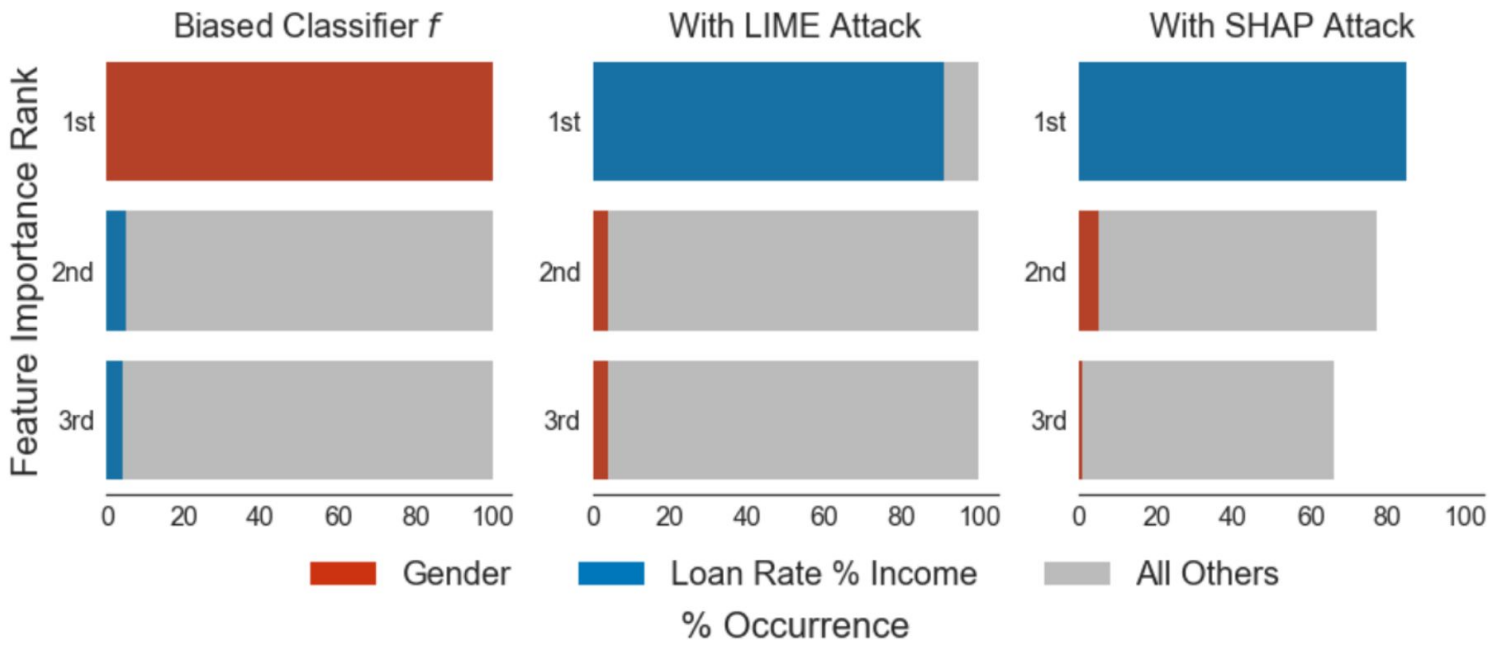
COMPAS


LIME
Biased Classifier f
Attack 1 Feature
Attack 2 Features
Feature Importance Rank
1st
2nd
3rd
0 20 40 60 80 100
African-American
Uncorrelated Feature 1
Uncorrelated Feature 2
Others
% Occurrence
SHAP
Biased Classifier f
Attack 1 Feature
Attack 2 Features
Feature Importance Rank
1st
2nd
3rd
0 20 40 60 80 100
African-American
Uncorrelated Feature 1
Uncorrelated Feature 2
Others
% Occurrence

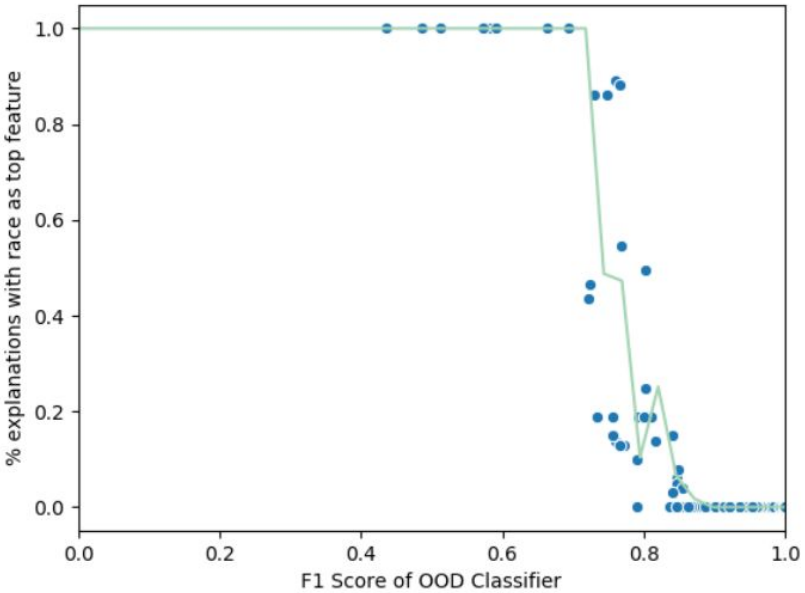
Communities and Crime



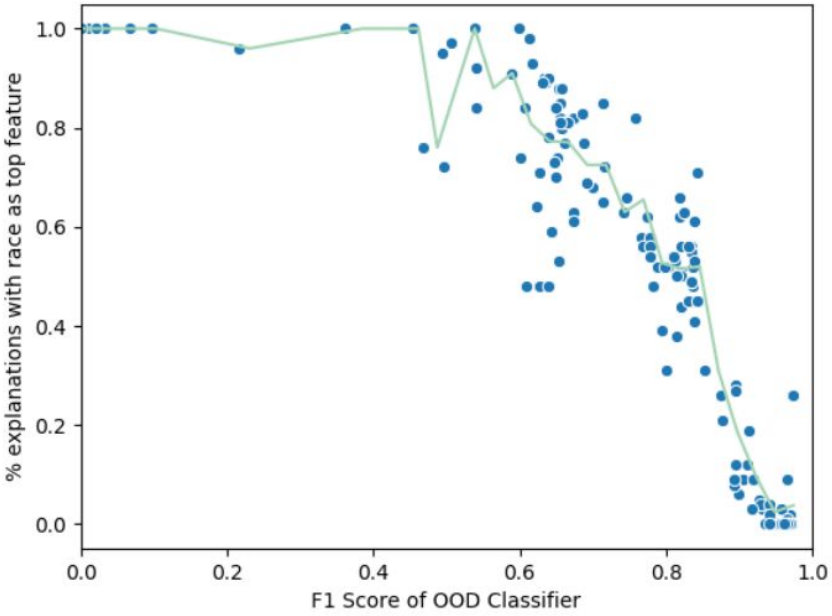
German credit



Sensitivity Analysis



(a) LIME COMPAS Sensitivity Analysis



(b) SHAP COMPAS Sensitivity Analysis

Conclusions

- Adversarial methods against post hoc methods exist and they work with black-box models
- Using LIME / SHAP to detect model bias is risky
- LIME is more susceptible to adversarial attacks than SHAP
- The attacks are somewhat robust to the hyperparameters
- The more the adversary knows about the explanation, the more successful they can be
- XAI-free metrics can be used to evaluate models

Thank you for your attention

Questions?