SHLIME: Foiling adversarial attacks fooling SHAP and LIME

Sam Chauhan, Estelle Duguet, Karthik Ramakrishnan, Hugh Van Deventer, Jack Kruger, Ranjan Subbaraman

University of Michigan

{csanjana, eduguet, kartikrk, hughv, jakrug, ranjans}@umich.edu

Abstract

Individual metrics for classifier explanation enhance interpretability and expose implicit biases, aiding engineers in understanding and improving machine learning models. This study builds on "Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods" (Slack et al., 2020) to develop a robust framework for analyzing adversarial models.

We first replicate and validate the original COMPAS experiment using LIME and SHAP on varied F1 models. Then, we introduce an efficient testing framework for evaluating ensemble methods and augmentations on out-of-distribution classifiers. Our results, compared graphically with the original study, highlight methods that better detect biases as F1 score improves.

We discuss the implications for deploying black-box classifiers and propose future improvements based on our enhanced bias detection techniques.

Introduction

Our research investigates the vulnerabilities of post hoc explanation methods—LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations)—to adversarial attacks. These explanation techniques are crucial for interpreting black-box models, especially as they are increasingly deployed in critical domains such as healthcare and criminal justice.

LIME (Local Interpretable Model-Agnostic Explanations)

LIME explains model predictions by approximating the black-box model locally with an interpretable, linear model. It uses a distance-based weighting function to ensure that the explanation is faithful to the model's behavior near the instance being explained.

SHAP (SHapley Additive exPlanations)

SHAP provides explanations based on Shapley values from cooperative game theory, ensuring three key properties: local accuracy, missingness, and consistency. This method gives a theoretically grounded way to attribute contributions of individual features to the model's predictions.

We explore how adversarial models can be crafted to fool LIME and SHAP, leading to misleading explanations. This involves manipulating input data and evaluating the robustness of LIME and SHAP under adversarial conditions.

Framework for Adversarial Classifiers

An adversary deploys a biased classifier *f* for critical decisions (e.g., parole, bail, credit). Customers and regulators with black-box access to *f* use post hoc explanation techniques to assess its fairness. If biases are detected, deployment is likely rejected. The adversary's goal is to obscure these biases from detection.

The adversarial classifier *e* is defined as:

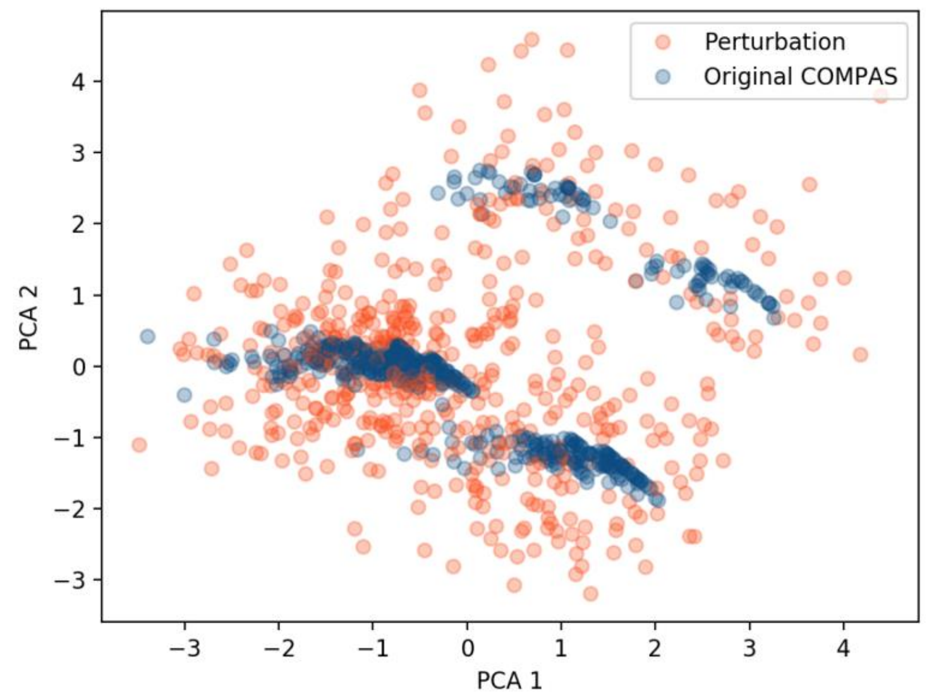
$$e(x) = \begin{cases} f(x), & \text{if } x \in \mathcal{X}_{dist} \\ \psi(x), & \text{otherwise} \end{cases}$$

Where *f(x)* is the biased classifier and *psi(x)* is the unbiased classifier

Inputs: Biased classifier *f* and dataset *X*
Output: Adversarial classifier *e*, unbiased when analyzed by LIME/SHAP

Detecting Out-of-Distribution (OOD) Samples

1.) Generate perturbed instances *Xp* from each *x* in *X*.
2.) Label real data *X* as `False` and perturbed data *Xp* as `True`.
3.) Train a classifier `is_OOD` on $X \cup Xp$ with these labels.



Methodology

Replication

We replicated the original paper's experiments on COMPAS. We:

- Constructed adversarial classifiers from biased classifiers.
- Explained the adversarial classifiers predictions using LIME and SHAP, observing how each LIME and SHAP shifted feature importance from sensitive to uncorrelated features

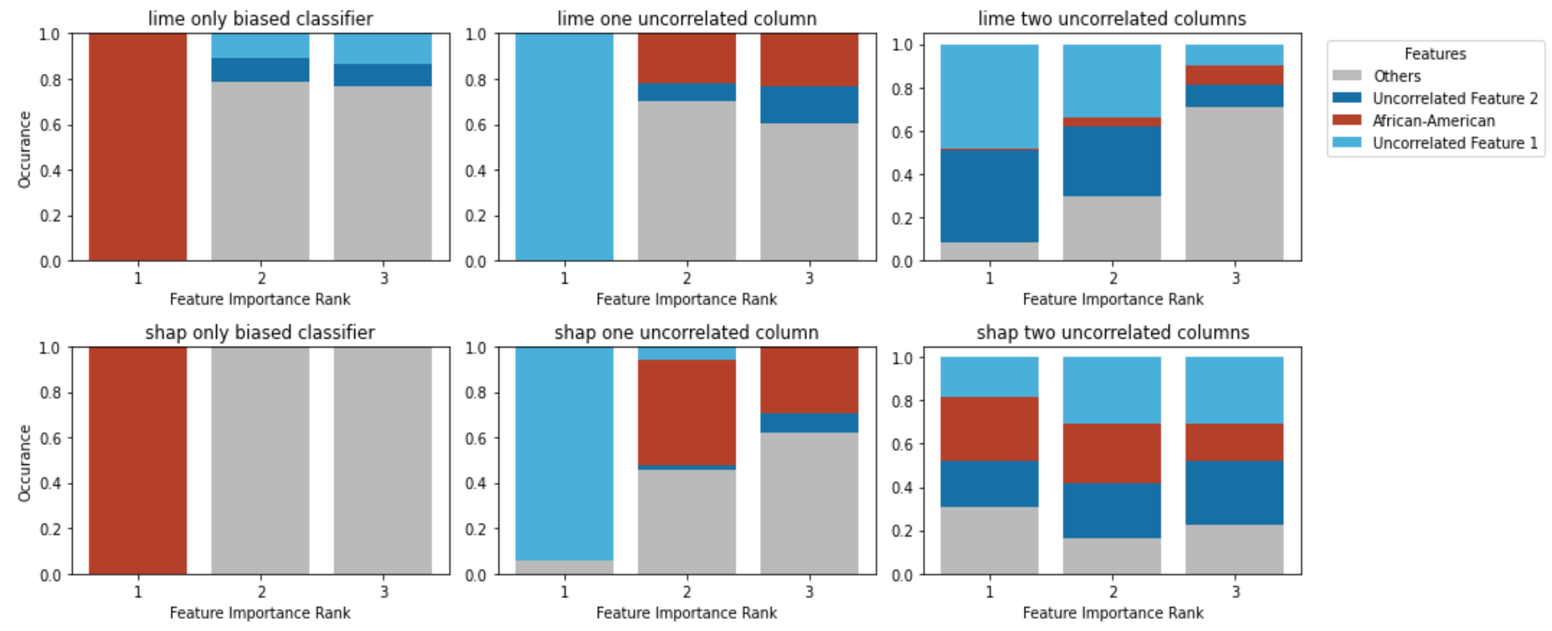
Extension

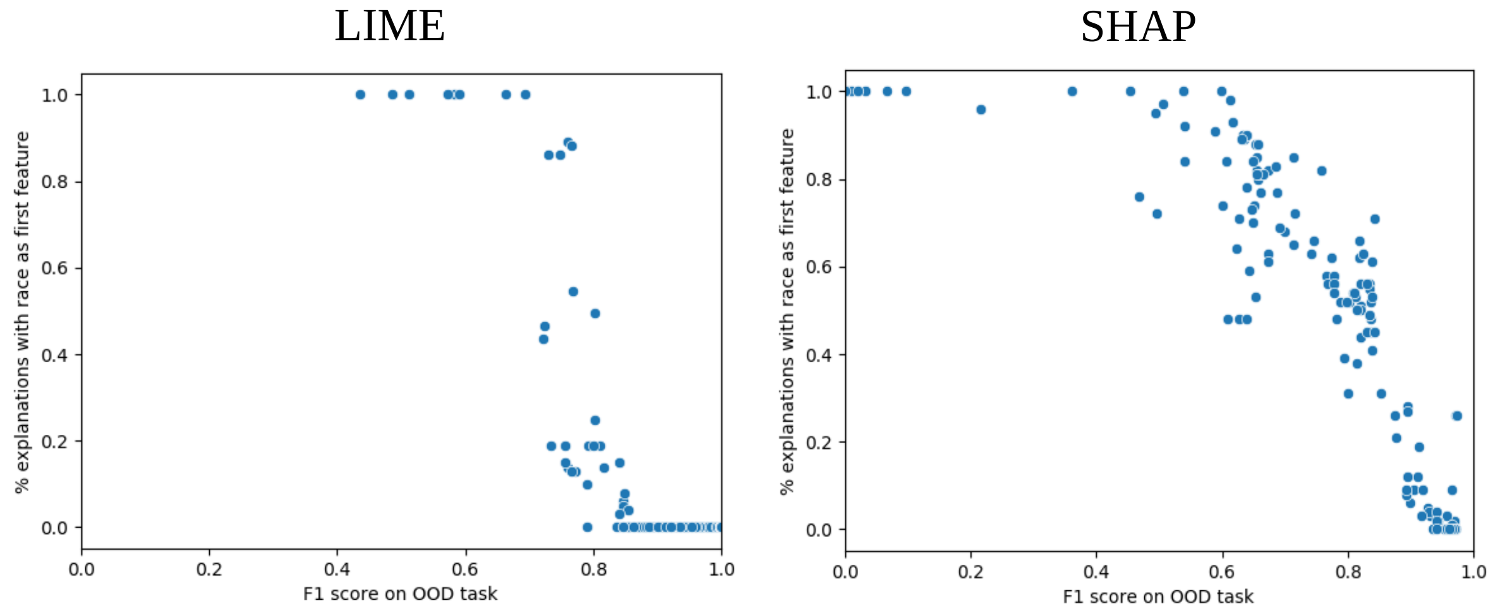
We extended the original study by developing SHLIME, an ensemble approach that combines LIME and SHAP explanations:

- Developed BASIC SHLIME, a method that multiplies LIME and SHAP values for each feature to enhance robustness
- Tested SHLIME using the COMPAS dataset and analyzing its performance against varying accuracies of OOD classifiers.

Results

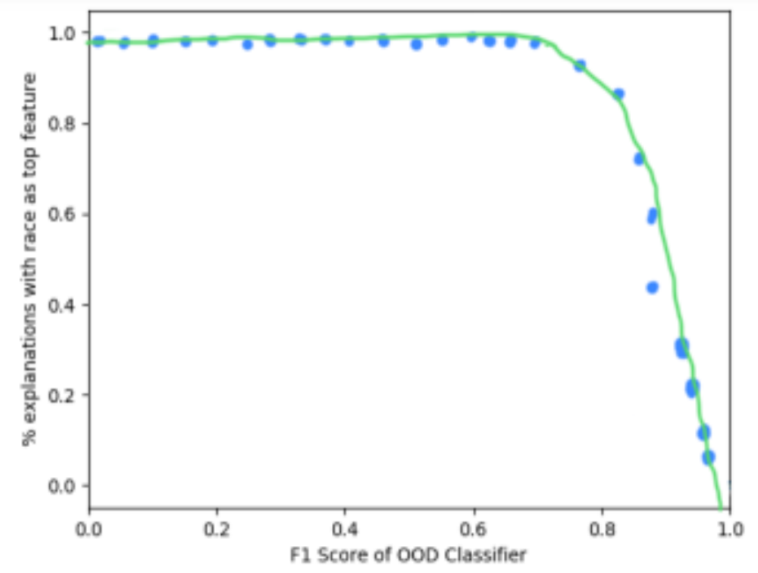
Sensitivity analysis



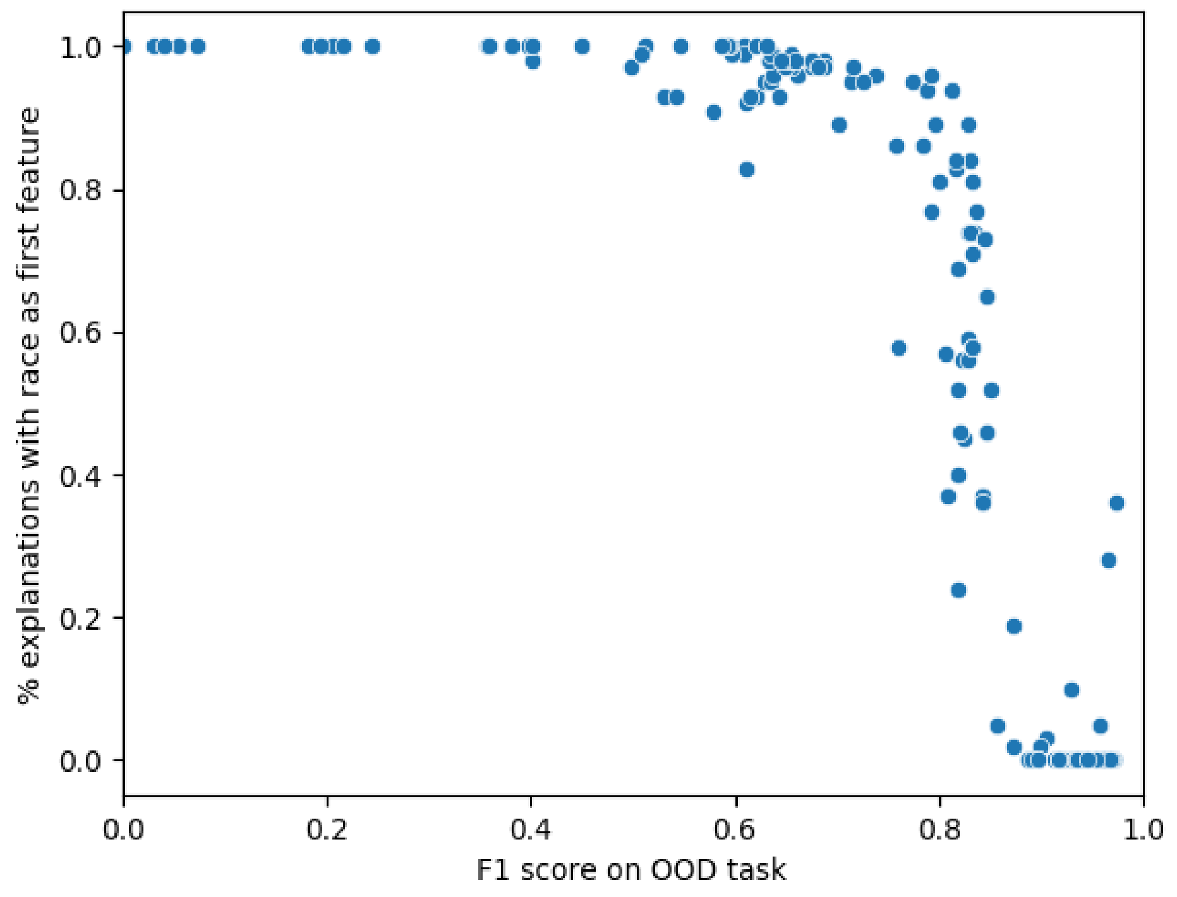

LIME

SHAP

Example of desired results of SHLIME



Actual results of SHLIME



Discussion

The main motivation for this research project is the deployment and growing popularity of black-box classifiers in various human-centered fields. By analyzing these models more harshly in development, engineers can proactively combat learned biases and employ techniques to limit or unlearn harmful stereotypes when classifying based on an individual's data. With the implementation of higher quality testing metrics, we hope to improve the quality of these black-boxes in healthcare, law, and more before they are deployed and exposed to real user bases.

Future Work

- Exploring non-linear ensemble methods for SHLIME such as *Mixture of Experts* sampling
- Creating proper OOD classifiers specific to SHLIME models to ensure fair comparison of robustness between each model
- Comparing SHLIME's interpretability performance against LIME and SHAP on a wide array of datasets

Societal Impact

The paper highlights critical societal risks by exposing vulnerabilities in LIME and SHAP, two widely used explainability tools in AI. These weaknesses could erode public trust in automated decision-making systems, especially in high-stakes areas like healthcare, finance, and criminal justice, where transparency is paramount. Adversarial manipulation of explanations could obscure biases or unethical behavior in AI models, perpetuating systemic inequities and complicating efforts toward fairness and accountability. Furthermore, these vulnerabilities pose significant security risks, particularly in safety-critical domains like autonomous vehicles or medical diagnostics, where misleading explanations could have life-threatening consequences. Finally, the findings emphasize the need for updated regulatory frameworks to account for such attacks, ensuring compliance and ethical deployment of AI systems.

Reference

Original Paper
Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP. AAAI/ACM, 180–186. https://doi.org/10.1145/3375627.3375830

Github: https://github.com/dylan-slack/Fooling-LIME-SHAP