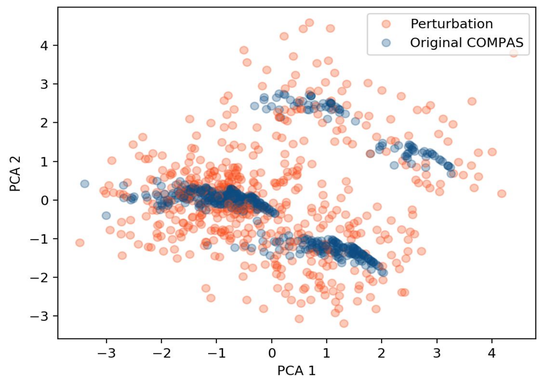
Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods

BACKGROUND:

Post hoc model agnostic explanation techniques, like LIME and SHAP, generate explanations for a single prediction for any black-box model. Perturbation-based techniques do this by sampling in the neighborhood of that instance, and observing the effect on the output prediction.

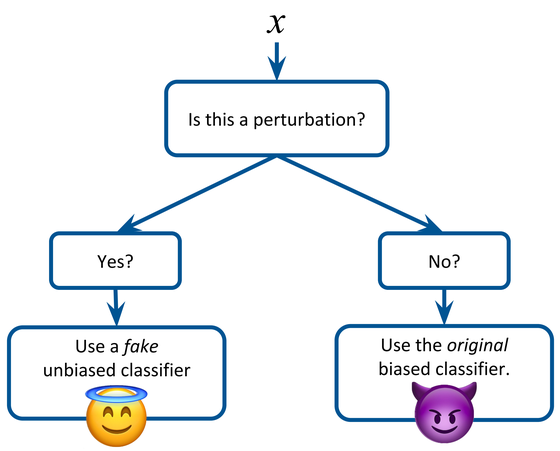
OBSERVATION:

Many of these samples used for explanation are **out of distribution (OOD)**.



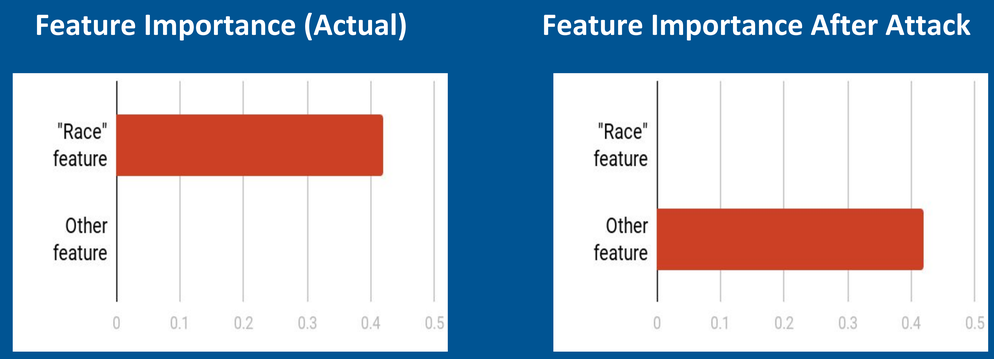
METHOD:

Train a model to predict which samples are OOD. Create adversarial classifier (e) by using two classifiers — one for in distribution instances (ψ) and one for OOD instances (f).

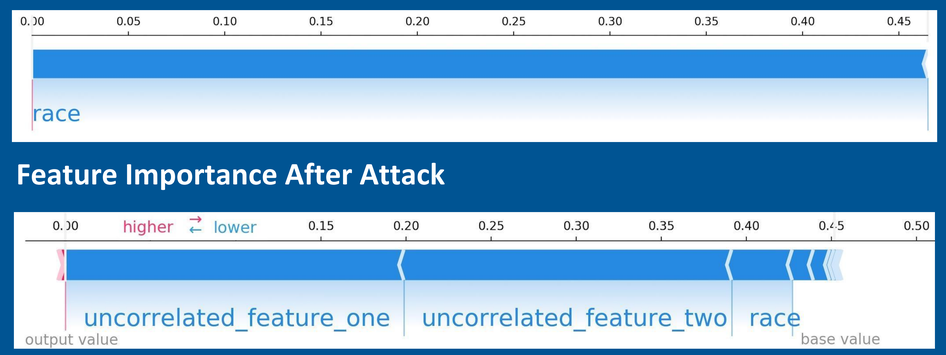


Post hoc explanation techniques can be manipulated to produce *any* explanation.

LIME



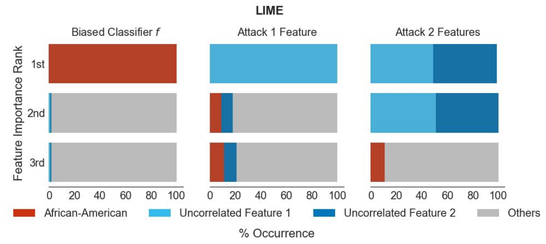
SHAP



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

COMPAS

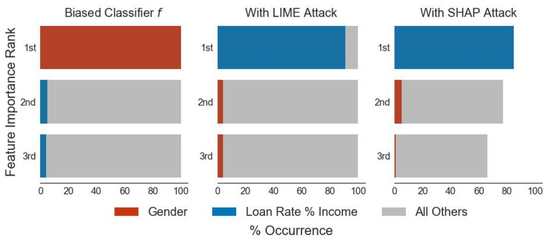
Introduce one or two additional uncorrelated features for the *adversarial* classifier.



***Communities and Crime results similar**

GERMAN CREDIT

Use an existing, largely uncorrelated feature for the *adversarial* classifier.



The adversarial classifier have exact same predictions (100% fidelity) as the original biased classifier on in-sample data with the LIME attack and 75%+ fidelity with the SHAP attack.

Dylan Slack*, Sophie Hilgard*, Emily Jia, Sameer Singh, Himabindu Lakkaraju

HARVARD UNIVERSITY

UCIRVINE

*equal contribution