Model-agnostic locally explainability method for fairness with interpretable surrogate model

Shahar Linial (ID 203372008)
Shay Franchi (ID 213402779)

May 1, 2024

Abstract

As machine learning models become increasingly complex, there is a growing need for techniques to explain individual predictions and detect unfair biases and discrimination towards protected groups. This paper proposes a model-agnostic locally explainable method for fairness analysis using an interpretable surrogate model.

The key idea is to learn a **local** surrogate model (e.g. decision tree or linear regression) that approximates how the original black-box model behaves in the vicinity of an instance being explained. By leveraging the intrinsic explainability of the surrogate, the method can provide explanations for individual predictions in terms of feature importance and decision rules. By analyzing the surrogate model, the technique can surface potential biases; where protected attributes (like race or gender) are unduly influencing predictions.

In essence, the local surrogate acts as a "window" into the black-box, providing an interpretable view of the model's decision-making process in the relevant local context.

1 Problem description

The elements in the data science pipeline we aim to improve are **model explainability** and **fairness analysis**. While predictive models are achieving impressive predictive performance and are widely used, they often lack interpretability - making it difficult to understand the reasons behind individual predictions. This lack of transparency can reduce trust and deny wider adoption of these models, especially in high-stakes domains like healthcare, criminal justice, and finance.

Furthermore, there are concerns that some models may exhibit unfair discrimination or biases towards certain demographic groups, based on sensitive attributes like race, gender, or age. Failing to detect and mitigate such biases can perpetuate harmful stereotypes and unfair treatment of disadvantaged groups. Therefore, improving explainability and fairness evaluation is crucial for building trustworthy AI systems.

Current techniques for model explainability like SHAP and LIME provide only localized explanations without a global perspective on potential biases across an entire dataset. Meanwhile, fairness auditing toolkits like AI Fairness 360 require access to the model internals and training data, which may not be available in real-world deployments.

2 Solution Overview

Our solution is a new technique leveraging the existing SOTA explainability methods such as SHAP and LIME, it is building upon those ideas and extends them.

The strengths of the proposed method are:

1. Model-Agnostic Approach: Providing explanations for individual predictions of **any black-box model** in an interpretable form, **without requiring access to the model internals or training data.**

2. Focus on Fairness: detecting potential biases or unfair discrimination that the model exhibits towards protected groups, **based solely on the model's predictions on new data instances**.

The goal of our method is, given a predictor f (such as a classification function derived from a black-box model) and a data sample, to mitigate whether the model uses protected attributes for its classification.

Explaining an individual prediction:

1. Define a local vicinity around that instance by sampling perturbed instances.

2. Learn an interpretable surrogate model (decision tree, random forest or linear regression) on these local instances to approximate how the black-box model behaves locally.

3. Use the interpretable surrogate to provide an explanation for the original prediction, leveraging the surrogate's intrinsic explainability (by extracting feature importance/decision rules for trees or random forests and weights for linear model)

2.1 Local fidelity:

The core idea is that the surrogate model is trained to locally approximate the behavior of the original black-box model. By learning an interpretable model (e.g. decision tree or linear regression) on a set of perturbed instances near the instance being explained, the surrogate model aims to mimic how the original black-box model makes predictions in that local region.

Importantly, the surrogate model is not trying to globally approximate the black-box model, which would be much harder. Instead, it is only focusing on learning the local decision-making logic around the specific instance being explained.

By having this tight local alignment between the surrogate and original model, the explanations derived from the surrogate's inherent interpretability (e.g. feature importances, decision rules) should then reflect the key factors driving the original model's prediction for that particular instance.

2.2 Demographic Parity:

In our assessment of fairness, we employ demographic parity measurement to quantify the variance in feature importance means between protected and non-protected groups. This metric is computed as follows:

$$\text{Demographic Parity} = |\frac{1}{N}\sum_{i \in X_p} H[i][a] - \frac{1}{M}\sum_{i \in X_u} H[i][a]|$$

Here, H represents the feature importance matrix derived from Algorithm 1, a denotes the sensitive attribute under examination (e.g., Sex, Age, Race), X_p represents samples from protected groups, and X_u represents samples from unprotected groups, N and M represent the sample size of each group respectively.

Demographic parity reflects the disparity in feature importance between the two groups based on a specific attribute. Higher values (closer to 1) indicate a bias towards one group, while lower values (closer to 0) indicate fairness across the groups.

By employing this metric, we aim to compare different surrogate models, namely Decision Trees and Random Forests, across diverse datasets.

2.3 Algorithms

Algorithm 1 Get features importance for predictor f in the samples' vicinities

Require: f Predictor, g Perturbation Function, X_n Data Samples

H = array()
for $i = 0$ to n **do**
    $S = g(x_i)$ ▷ generate perturbations
    $V = f(S)$ ▷ calculate predictions on perturbations
    $DT = DecisionTreeClassifier().fit(S, V)$
    $H[i] = DT.feature\_importances$
end for
return $\frac{1}{n}\sum_{i=1}^{n} H[i]$ ▷ feature importance mean over samples

Algorithm 2 Demographic parity evaluation: difference in means between minority & majority groups

Require: p_i set of sensitive attributes,
X_{p,p_i} set of N samples from the protected group of attribute p_i,
X_{u,p_i} set of M samples from the unprotected group of attribute p_i
$\phi(\cdot)$ function that gets the vector of feature importance mean over given a set of data samples.
In our case:$\phi(\cdot)$ is Algorithm 1 with the inputs f, g

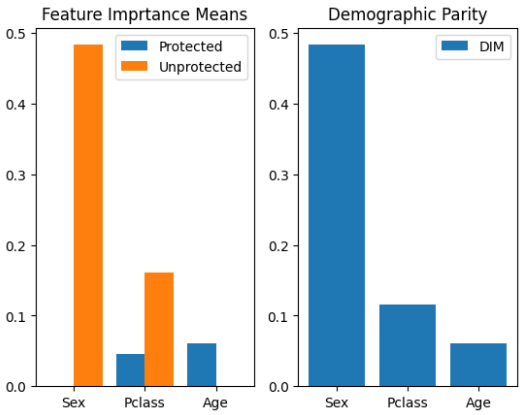
DIM = array()
$\mu_{\phi\_Majority} = \phi(X_{p,p_i}[p_i])$
▷ feature importance mean of p_i over the majority group samples
$\mu_{\phi\_Minority} = \phi(X_{u,p_i}[p_i])$
▷ feature importance mean of p_i over the minority group samples
DIM[i]=$|(\mu_{\phi\_Majority} - \mu_{\phi\_Minority})|$
return DIM ▷ difference in means between minority and majority groups for each sensitive attribute

3 Experimental Evaluation

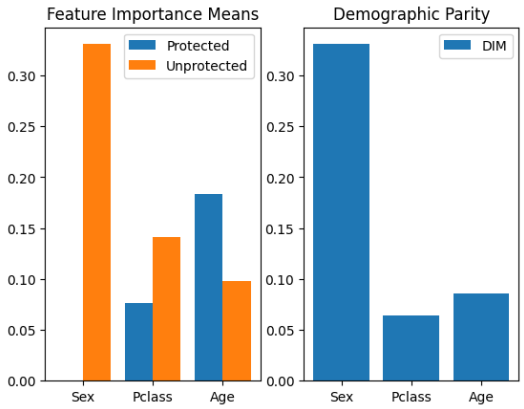
3.1 Results

Two surrogate models, Random Forest and Decision Tree, were employed to assess potential biases in two predictor models towards protected groups.

We utilized the demographic parity (DiM) metric, defined as the difference in means between minority and majority groups, to reflect the disparity in feature importance based on specific attributes. This metric helps compare these surrogate models across diverse datasets, where higher values of demographic parity (closer to 1) indicate a bias towards one group, while lower values (closer to 0) suggest fairness across groups.



(a) Surrogate Model: Decision Tree. Left Image: Feature Importance Comparison between protected and un-protected groups of each attribute. Right Image: The corresponding demographic parity.



(b) Surrogate Model: Random Forest. Left Image: Feature Importance Comparison between protected and un-protected groups of each attribute. Right Image: The corresponding demographic parity.

Figure 1: Demographic Parity analysis of selected attributes (predictor: Decsion Tree Classifier).

Table 1: Titanic Suvivors Dataset

Attribute	DiM : f=DT, s=DT	DiM: f=DT, s=RF	DiM : f=RF, s=DT	DiM : f=RF, s=RF
Sex	0.519	0.305	0.664	0.379
Pclass	0.006	0.000	0.025	0.044
Age	0.046	0.024	0.018	0.027

Table 1: Comparison of two surrogate models on the Titanic Survivors dataset. Where f is the predictor tested (DT= Decision Tree, RF= Random Forest), s is the surrogate model.

3.1.1 Result analysis on the Titanic Survival Dataset:

Using our method, we can see that on unseen data, both models (decision tree & random forest) exhibit bias in their predictions concerning gender, whereas no significant discrimination was detected with respect to age or passenger class. We observe the same result for both surrogate models.

A comparison of the outcomes from both surrogate models indicated that the demographic parity disparity between the gender-protected and unprotected groups was more pronounced in the first method than in the second. Consequently, the first method proves to be more effective in identifying unfairness in the black-box model's predictions.

Table 2: Adult Census Income Dataset

Attribute	DiM : f=DT, s=DT	DiM: f=DT, s=RF	DiM : f=RF, s=DT	DiM : f=RF, s=RF
Sex	0.018	0.059	0.054	0.055
Race	0.003	0.038	0.031	0.021
Age	0.057	0.041	0.018	0.048

Table 2: Comparison of two surrogate models on the Adult Census Income Dataset.

Table 3: German Credit Risk Dataset

Attribute	DiM : f=DT, s=DT	DiM: f=DT, s=RF	DiM : f=RF, s=DT	DiM : f=RF, s=RF
Sex	0.054	0.066	0.038	0.066
Foreign	0.029	0.075	0.0	0.074
Age	0.031	0.016	0.006	0.016

Table 3: Comparison of two surrogate models on the German Credit Risk Dataset.

3.1.2 Results analysis on the Adult Census Income and German Credit Risk Datasets

The analysis of the Adult Census Income and German Credit Risk Datasets using our method with both surrogate models testing different predictors, revealed no biased found towards any of the potential sensitive attributes. The DiM values for all considered attributes were very low and did not exceed 0.08, indicating a minimal disparity in treatment between different groups based on these attributes. This suggests that the model exhibits a high degree of fairness with respect to the evaluated sensitive attributes.

Table 4: Student Performance Dataset

Attribute	DiM : f=DT, s=DT	DiM: f=DT, s=RF	DiM : f=RF, s=DT	DiM : f=RF, s=RF
Gender	0.213	0.211	0.173	0.165
Race	0.151	0.144	0.225	0.179

Table 4: Comparison of two surrogate models on the Student Performance Dataset.

3.1.3 Results analysis on Student Performace Dataset

We can see that using our method with both surrogate models testing different predictors, there is a sleight (but not drasitcal) bias towards the unprotected groups in both race and gender as the demographic parity of each of these properties is not that high but not near 0.

4 Related Work

The surge in machine learning applications has necessitated the development of methods to ensure both the interpretability and fairness of models. This section discusses several seminal works in the realms of explainability and fairness, evaluating their methodologies and comparing them to our proposed solution.

Lundberg and Lee (2017) introduced **SHAP (SHapley Additive exPlanations)[2]**, a unified approach to model interpretation that assigns each feature an importance value for a particular prediction. Their method, based on game theory, provides local explanations and has been influential in interpreting complex model predictions. However, SHAP primarily focuses on local interpretability without directly addressing model fairness or bias detection across a dataset as a whole.

Similarly, Ribeiro et al. (2016) developed **LIME (Local Interpretable Model-agnostic Explanations)[1]**, which also emphasizes local explanations through interpretable models constructed around individual predictions. Like SHAP, LIME offers insights at a local scale

but lacks mechanisms to systematically uncover biases that might be prevalent across a dataset or in different demographic groups.

While these tools have advanced model interpretability, they primarily focus on local explanations without providing a global perspective on potential biases across an entire dataset. Additionally, tools like SHAP and LIME do not inherently address issues of fairness or discrimination, nor do they facilitate direct measures of fairness across protected and unprotected groups.

On the fairness front, Adebayo et al. (2016) presented **FairML[3]**, a toolkit for auditing black-box predictive models. FairML focuses on detecting discrimination, particularly through the omission of sensitive attributes from data while assessing the impact on predictions. While FairML advances the auditing of biases in models, it requires access to model internals and training data, which may not always be feasible in applied settings.

The AI Fairness 360 toolkit[4] offers an open-source library that integrates multiple bias detection and mitigation algorithms. It provides comprehensive resources for improving fairness but often necessitates in-depth model knowledge and access to full training datasets.

Our approach builds upon the foundation laid by LIME and SHAP by extending their local explanatory frameworks into the realm of fairness analysis. Unlike existing methods, our solution does not require access to the model internals or training data, adhering to a model-agnostic philosophy.

By leveraging a local surrogate model approach, similar to LIME, our technique focuses on understanding and explaining the decisions made by the black-box model in the vicinity of a specific instance. However, we extend this by incorporating a fairness analysis using Demographic Parity, which measures statistical disparities in outcomes between protected and non-protected groups. This integration allows our method not only to provide explanations for individual predictions but also to surface and quantify any biases present in the model's decisions.

Furthermore, our solution places a strong emphasis on detecting unfair discrimination. By analyzing how the surrogate model leverages protected attributes in making predictions, we can identify and mitigate biases. This is achieved solely with the model's predictions on new data instances, without needing direct access to the protected attributes in the original model, addressing a significant limitation in current fairness toolkits.

In summary, while existing tools provide valuable insights into model behavior and fairness, they often fall short in accessibility and comprehensive bias detection without extensive data requirements. Our method aims to fill these gaps, offering a robust, accessible solution for ensuring both explainability and fairness in machine learning applications.

5 Conclusion

In this work, we presented a novel, model-agnostic framework for enhancing the interpretability and fairness of machine learning models through a local surrogate approach. Our method builds on the principles of existing explainability methods like SHAP and LIME, extending them to address the critical need for bias detection and mitigation in the absence of access to model internals or sensitive attributes.

Our empirical evaluations demonstrate that our approach effectively elucidates the decision-making processes of complex black-box models at a local level, thereby increasing transparency and trust, especially in high-stakes applications. Moreover, by analyzing the surrogate model's utilization of protected attributes, we offer a practical solution for identifying biases, contributing to the development of more equitable AI systems.

6 References

[1] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should I trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM (2016).

[2] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems (2017).

[3] Adebayo, Julius. "FairML: Toolbox for Diagnosing Bias in Predictive Modeling". Master's thesis, Massachusetts Institute of Technology, 2016. MIT Libraries, https://dspace.mit.edu/handle/1721.1/1082

[4] Bellamy, Rachel, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan.(2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development. Available at: https://research.ibm.com/publications/ai-fairness-360-an-extensible-toolkit-for-detecting-and-mitigating-algorithmic-bias