FairSHAP: Preprocessing for Fairness Through Attribution-Based Data Augmentation

Anonymous Author(s)
Affiliation
Address
email

Ensuring fairness in machine learning (ML) models is critical, particularly in high-stakes domains where biased decisions can lead to serious societal consequences. Existing preprocessing approaches generally lack transparent mechanisms for identifying which features or instances are responsible for unfairness. This obscures the rationale behind data modifications. We introduce FairSHAP, a novel preprocessing framework that leverages Shapley value attribution to improve both individual and group fairness. FairSHAP identifies fairness-critical instances in the training data using an interpretable measure of feature importance, and systematically modifies them through instance-level matching across sensitive groups. This process, described in Figure 1, reduces discriminative risk (DR)—an individual fairness metric—while preserving data integrity and model accuracy.

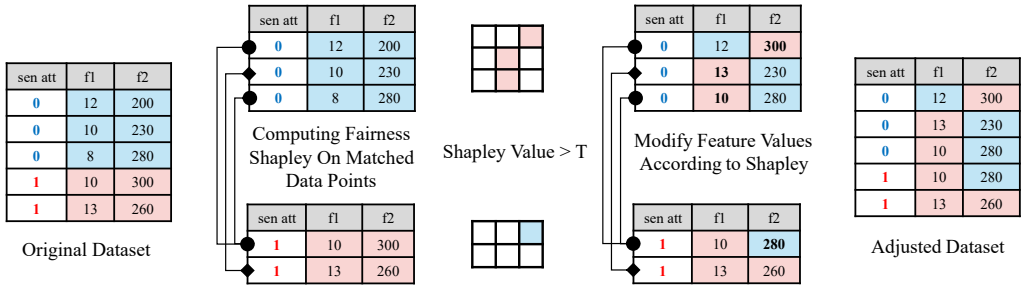

Figure 1: Overall framework of FairSHAP. (Left) Training data are first split by sensitive attribute and aligned via nearest-neighbor matching to produce paired instances. (Right) For each target group, feature values whose Shapley value exceeds a threshold are adjusted to reduce DR, and the modified instances from both groups are recombined into an augmented, fairness-improved training set.

Table 1: Compare FairSHAP with other fairness mitigation methods across different datasets. The columns 'TrainingAR', 'TestAN', and 'Data Fidelity' denote training set adjustment rate, test set adjustment necessity, and a measure using the Wasserstein distance to quantify the difference between distributions, respectively.

Dataset (sen-att)	**Method**	**Accuracy**	**DR**	**DP**	**EO**	**PP**	**Data Fidelity**	**TrainingAR**	**TestAN**
COMPAS (race)	Baseline	0.6580±0.0098	0.0827±0.0035	0.1968±0.0179	0.1776±0.0473	0.0487±0.0173	—	—	No
	CorrelationRemover	0.6561±0.0084	**0.0000±0.0000**	0.1738±0.0280	0.1743±0.0409	0.0845±0.0391	0.0666±0.0489	0.9363	Yes
	DisparateImpactRemover	0.6571±0.0078	0.0848±0.0112	0.1659±0.0557	0.1459±0.0712	0.0642±0.0408	0.0236±0.0401	0.0553	Yes
	FairSHAP ($T=0.05$)	**0.6602±0.0069**	**0.0553±0.0058**	0.1861±0.0193	**0.1729±0.0460**	0.0590±0.0228	**0.0032±0.0041**	**0.0142**	**No**

FairSHAP bridges explainability and fairness by connecting Shapley values with DR, and is further supported by both theoretical proofs and empirical evidence, showing that improving individual fairness can also improve group fairness. To be specific, we demonstrate that FairSHAP significantly improves demographic parity and equality of opportunity across diverse tabular datasets, achieving fairness gains with minimal data perturbation and, in some cases, improved predictive performance. Moreover, as a model-agnostic and transparent method, FairSHAP is broadly applicable to tabular data, supports various models and SHAP algorithms, can be seamlessly integrated into existing ML pipelines, achieves comparable or superior fairness with significantly less data modification than benchmark methods, and provides actionable insights into the sources of bias.

Submitted to the 20th Workshop for Women in Machine Learning (WiML). Do not distribute.