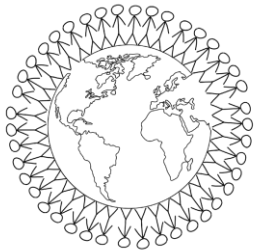
FALE: Fairness-Aware ALE plots for Auditing Bias in Subgroups

Athena Research Center

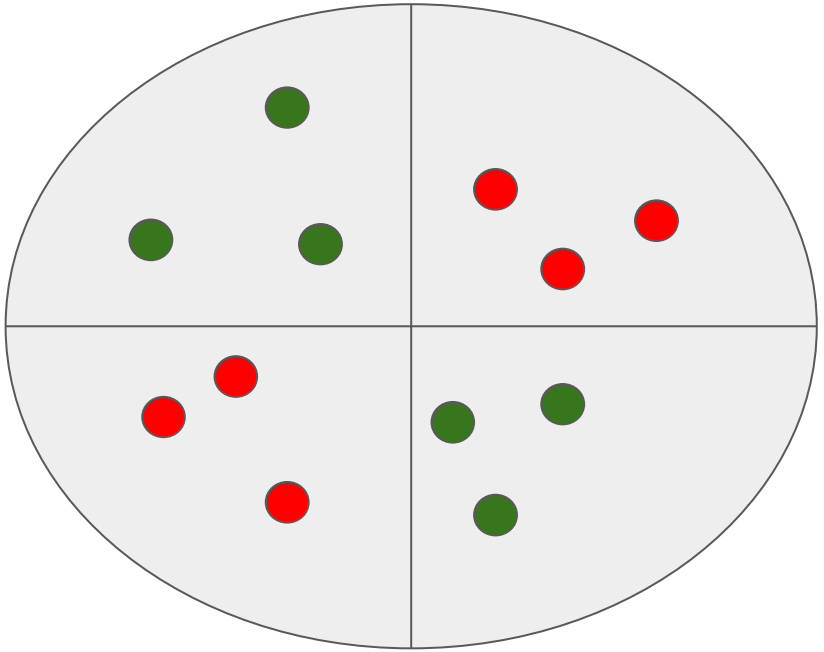


Funded by the European Union. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or DG Connect. Neither the European Union nor DG Connect can be held responsible for them.

Algorithmic Fairness

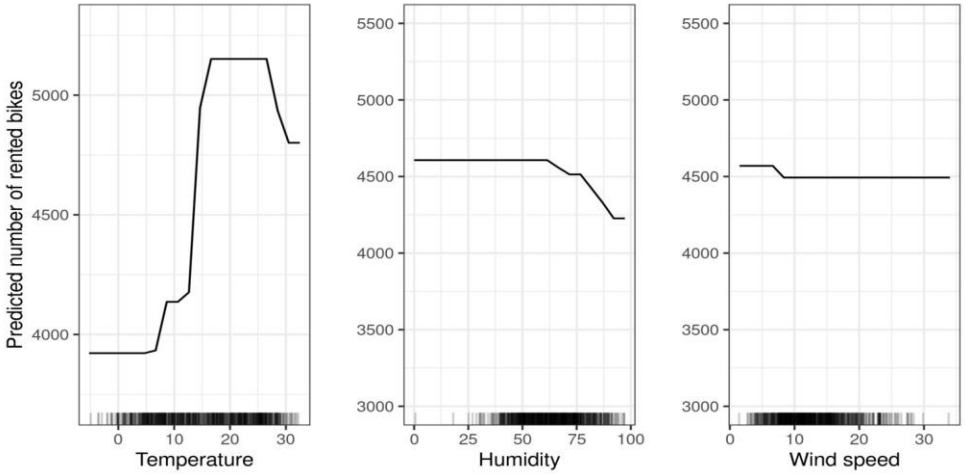
- The model should not **discriminate against a sensitive attribute** like gender.
- **Statistical parity**: Males and females should have equal probability of being assigned to the positive outcome by the classifier.
- However, **biases** may be **present** or **enhanced in more fine grained subgroups of males and females.**

Subgroup fairness



ALE plots

- Provide **insight** into the **relationship** between a **feature** and the **target variable**
- **Visualize** the **marginal effect** that a feature has on the predicted outcome



Our method-FALE plots

Observation: Instances with **same** feature values define **subgroups.**

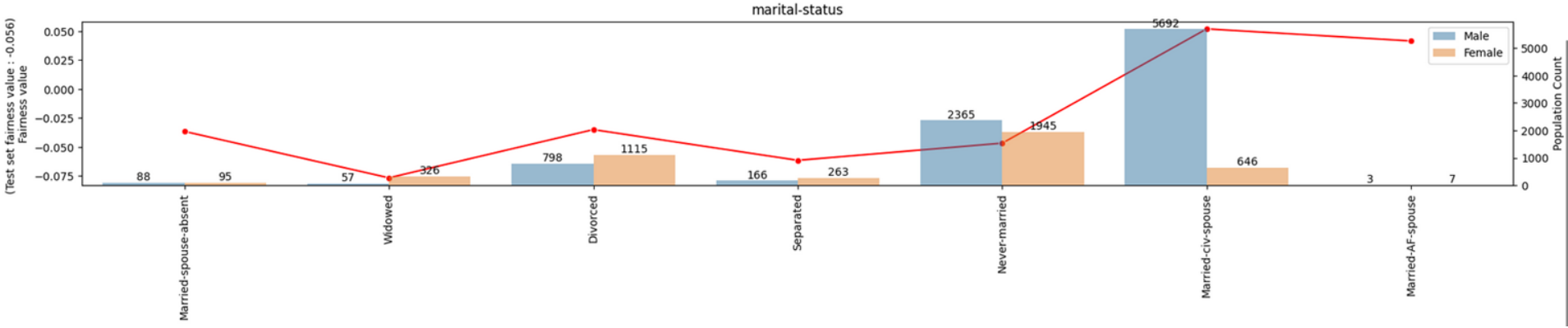
ALE plots visualize the **marginal effect** of **subgroups** to the **predicted outcome**

Replace the **model function** with a **statistical fairness definition**=>FALE

Example

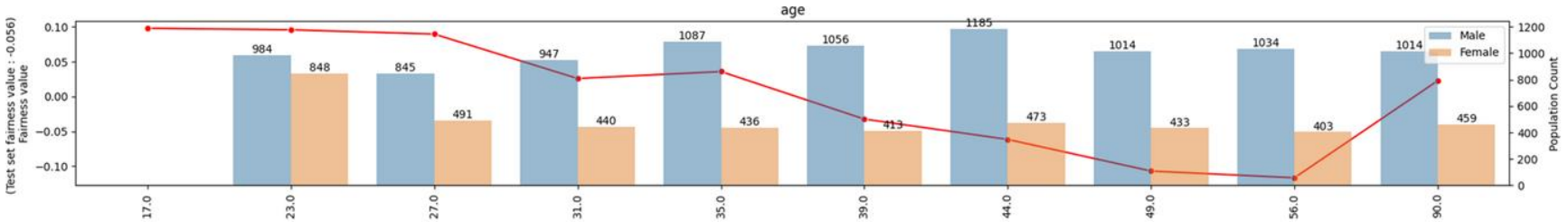
- We evaluated **statistical parity for gender** in the adult dataset
- We have found a value of **-0.056** that suggests that there is **bias against females**
- We want to see **how subgroups affect this value.**

Subgroups of Marital-status



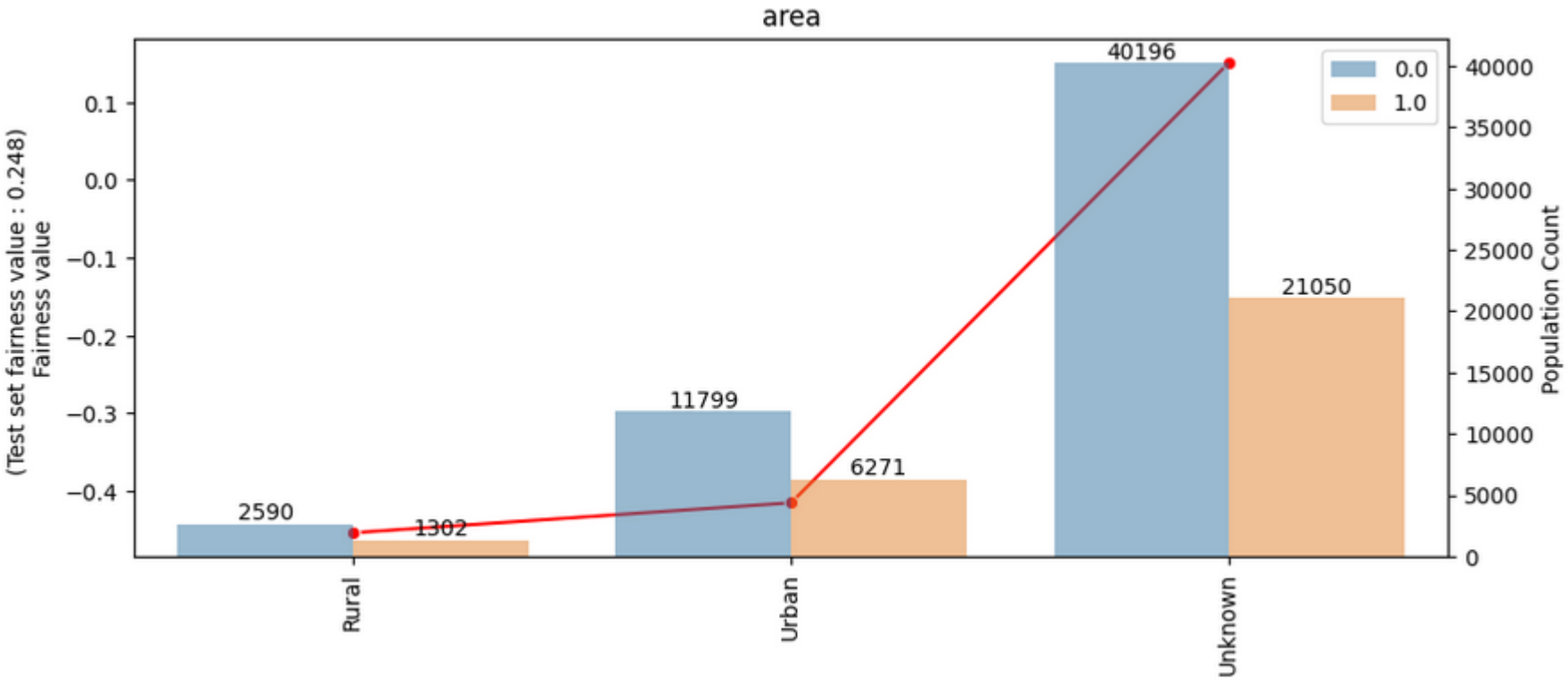
- **Zero** here is the reference value (-0.056)
- **Negative values**: **Widowed, separated , divorced, never married females** are treated with more bias.
- **Positive values**: **Married females** are treated with less bias.

Subgroups of Age



Middle aged females are treated with additional bias.

Advertising dataset-Subgroups of Area



Results

- Identify unfairness in subgroups even if the model is fair on the group level.
- Identify additional bias in subgroups for unfair models
- Visualize the unfairness=> Easily understood by non-experts

Future work

- Results on other datasets
- Implement 2D FALE
- Compare with other visual explainability methods in terms of fairness auditing