Université libre de Bruxelles

École Polytechnique de Bruxelles

Assignment 4

Management of Data Science and Business Workflows INFO-H420

Fall 2023

Authors:
Bouzaher, Mohamed Louai
Başaran, Ozan

Professor:
Dimitri Sacharidis

Contents

Exercise 1

1.1 Introduction

In this assignment, we employed the aif360 library to explore and mitigate bias in the COMPAS dataset. Our primary focus was on evaluating fairness with respect to two protected attributes: 'race' and 'sex.' The aif360 library provides powerful tools for assessing and addressing bias in machine learning datasets, making it a valuable resource for practitioners concerned with algorithmic fairness.

1.2 Loading the Data

We utilized the COMPAS dataset, a commonly used dataset in fairness studies, containing information about individuals in the criminal justice system. The aif360 library was instrumental in our analysis, offering functionalities for fairness metrics computation, bias mitigation techniques, and dataset manipulation.

To facilitate our experiments, we performed a train-test split on the original dataset. This allowed us to evaluate the fairness of machine learning models on an independent set, providing insights into generalization performance.

1.3 Question 1

1.3.1 Race as Protected Attribute at the Start

Metrics Analysis

The initial assessment of bias in the training dataset, considering 'race' as the protected attribute, revealed a difference in mean outcomes between unprivileged and privileged groups of approximately -0.107. This metric serves as a baseline, quantifying the existing disparity before applying any fairness interventions. A negative value implies that, on average, the unprivileged group experiences less favorable outcomes compared to the privileged group. This baseline measurement provides a reference point for assessing the effectiveness of subsequent bias mitigation techniques.

Reweighing for 'Race'

Upon applying reweighing specifically tailored for 'race,' we observed a remarkable improvement. The transformed training dataset achieved a difference in mean outcomes close to zero (-0.000000), indicative of successful bias mitigation. This outcome suggests that reweighing effectively balanced outcomes between privileged and unprivileged groups based on race.

Race and Intersectional Fairness

Intersectional fairness, as discussed in the assignment, explores the dynamics among groups defined by the intersections of attributes. For instance, it may examine positive rate differences between males and females within specific age groups or assess disparities among groups defined by both race and sex.

Building on this concept, our analysis initially focused on 'race' as the protected attribute, providing insights into the race-based disparities in the dataset. The subsequent steps will extend this examination to 'sex' as the protected attribute.

To investigate the impact of transitioning from 'race' to 'sex' as the protected attribute, we performed the following steps:

1. We deep-copied the racially reweighted dataset, creating a new dataset, and changed the protected attribute to 'sex.'
2. We set the privileged and unprivileged classes accordingly for 'sex' in the new dataset.
3. We then assessed the difference in mean outcomes between unprivileged and privileged groups based on 'sex.'

Our analysis revealed that, even with a racially fair attribute weighting, the difference in mean outcomes between unprivileged and privileged groups based on 'sex' was found to be -0.151983. This suggests that the fairness achieved for one protected attribute (in this case, 'race') does not automatically extend to another ('sex').

Predictive Modeling for Transformed Data with 'Sex' as the Protected Attribute

After transitioning to 'sex' as the protected attribute, we proceeded to build a logistic regression classifier and make predictions on the transformed dataset. Here are the key steps involved:

1. We utilized a logistic regression model, a commonly used algorithm for binary classification tasks, to build a predictive model.
2. Standardization of features was performed using a StandardScaler to ensure consistent scaling across different attributes.
3. The model was trained on the transformed training data with 'sex' as the protected attribute. We incorporated instance weights during training to account for the reweighing applied to mitigate bias.
4. Predictions were made on the training dataset, and the positive class index was determined.
5. The predicted labels were updated in the transformed training dataset for further analysis.
6. Test data features were transformed using the same scaling parameters from the training set.
7. Probabilities for the positive class were predicted for the test set.
8. A threshold of 0.5 was applied to classify the data, and the predicted labels were updated in the transformed test dataset.

Fairness Analysis

The classification metric used for the transformed test dataset compares the predictions made by the model with the ground truth labels.

The classification accuracy on the transformed test dataset yielded a value of 0.327391, indicating the overall accuracy of the model's predictions on this dataset.

Disparate impact, a measure of prediction imbalances between privileged and unprivileged groups, was found to be 3.354960 on the transformed test dataset. This value significantly deviating from 1 implies notable disparities in the positive outcome predictions between the two groups.

Furthermore, the equal opportunity difference on the transformed test dataset was calculated as 0.245649. This metric specifically assesses the disparities in equal opportunity for favorable outcomes between privileged and unprivileged groups based on 'sex.' A non-zero value points to existing inequities in the model's predictions.

In analyzing the transformed training dataset, the difference in mean outcomes between unprivileged and privileged groups for 'sex' was determined to be -0.283292. This negative value suggests that, on average, the unprivileged group experiences less favorable outcomes compared to the privileged group in the model's predictions.

1.3.2 Sex as Protected Attribute at the Start

Metrics Analysis

The initial assessment of bias in the training dataset, considering 'sex' as the protected attribute, revealed a difference in mean outcomes between unprivileged and privileged groups of approximately -0.160827.

Reweighing for 'Sex'

In response to the identified bias within the original training dataset, we applied reweighing tailored specifically for the 'sex' attribute. The transformation resulted in a training dataset with a difference in mean outcomes close to zero (0.000000), signifying the successful mitigation of bias associated with 'sex.'

Sex and Intersectional Fairness

Subsequently, we replicated the same process as question 1, this time focusing on the 'race' attribute. Despite the initial reweighing being designed for 'sex,' we transitioned the protected attribute to 'race.' The fairness metric, specifically measuring the difference in mean outcomes between unprivileged and privileged groups based on 'race,' returned a value of 0.000000. This outcome is remarkable and indicates an equitable distribution of outcomes between groups based on 'race' following the dataset transition.

The observed result of 0.000000 suggests that, on average, there is no discernible difference in outcomes between unprivileged and privileged groups after the shift in the protected attribute to 'race.' This underscores the efficacy of the reweighing technique initially tailored for 'sex' in extending fairness to the 'race' attribute.

Logistic Regression and Predictions with 'Race' as the Protected Attribute

Similar to the approach taken for 'sex' as the protected attribute, we employed a logistic regression classifier on the transformed dataset, where 'race' was the protected attribute. The process involved scaling features, training the model with instance weights, making predictions, and assessing the model's generalization on the test dataset.

Fairness Metrics

Comparing the fairness metrics with 'race' as the initially protected attribute, the results for 'sex' as the protected attribute exhibit improvements. The classification accuracy is slightly higher at 0.322528, indicating a somewhat better overall accuracy of the model's predictions. The measure of disparate impact, at 1.300601, remains indicative of disparities in positive outcome predictions between privileged and unprivileged groups based on 'sex.' However, this value is notably lower than the disparate impact observed with 'race' as the protected attribute.

The equal opportunity difference for 'sex' is 0.049672, which reflects an improvement in addressing disparities in equal opportunities for positive outcomes between privileged and unprivileged groups compared to the 'race' scenario. The difference in mean outcomes between unprivileged and privileged groups based on 'sex' is also relatively small, with a value of -0.005385, indicating a fairer distribution of outcomes in the model's predictions.

These findings suggest that, when 'sex' is considered as the initial protected attribute, the bias mitigation techniques result in a model with improved fairness metrics compared to the scenario where 'race' was the initially protected attribute.

1.3.3 Age as Protected Attribute + Sex for Intersectional Fairness

Metrics Analysis

We began by loading the COMPAS dataset and converting it into pandas DataFrames for ease of manipulation. This initial step allowed us to inspect the dataset's structure and features. Additionally, we used the AIF360 library to create a StandardDataset representation of the data,from the pandas DataFrames, which enabled the modification of protected attributes.

To address age-based disparities, we designated individuals older than 45 as privileged and others as unprivileged. This categorization aimed to explore potential biases in the dataset related to age. Subsequently, we computed fairness metrics on the original training dataset for age, revealing an initial difference in mean outcomes between unprivileged and privileged groups of approximately -0.155808.

Reweighing for 'Age'

We applied reweighing tailored for our 'Greater than 45' column to mitigate bias, the transformation resulted in a training dataset with a difference in mean outcomes close to zero (0.000000), signifying the successful mitigation of bias associated with 'age'.

Age and Intersectional Fairness

Using the previously created StandardDataset representation, we deep-copied the age-based reweighted dataset, changing the protected attribute to 'sex.' This involved updating the privileged and unprivileged classes accordingly.

Surprisingly, despite the reweighing intended for age-based fairness, the difference in mean outcomes between unprivileged and privileged groups based on 'sex' was found to be

-0.000000. This unexpected result raises questions about the transferability of fairness interventions across different protected attributes. The intricate dynamics of biases associated with 'sex' may require distinct mitigation techniques for effective fairness.

Predictive Modeling for Transformed Data with 'Sex' as the Protected Attribute

To assess the impact of transitioning to 'sex' as the protected attribute, we again employed logistic regression for predictive modeling on the transformed dataset.

Fairness Metrics

Classification Accuracy: The model achieved an accuracy of 0.341437, showcasing the overall correctness of predictions on the transformed test dataset.

Disparate Impact: With a value of 1.064295, the disparate impact metric indicates a moderate imbalance in positive outcome predictions between privileged and unprivileged groups. When compared to the Race first then Sex methodology, we observe a reduction in the Disparate Impact metric. This reduction suggests a shift towards a more balanced prediction for 'sex' when age is considered as the protected attribute first.

Equal Opportunity Difference: Calculated as -0.017928, the equal opportunity difference suggests minor disparities in equal opportunity for favorable outcomes between privileged and unprivileged groups based on 'sex.' In comparison with the Race first then Sex approach, this metric suggests a shift towards a more balanced prediction again.

Analyzing the transformed training dataset, the difference in mean outcomes between unprivileged and privileged groups for 'sex' was found to be 0.004856. This small positive value indicates a subtle preference toward the privileged group in the model's predictions but much better than race first approach.

Exercise 2

2.1 Multi-Dimensional Subset Scan

The Multi-Dimensional Subset Scan (MDSS) method is an algorithmic approach used in the field of fairness, specifically within the context of detecting instances of unfairness or biases in machine learning models, particularly those related to discrimination or disparate treatment across different subpopulations.

MDSS operates by systematically scanning various subsets within the dataset, evaluating each subset for potential unfairness or bias. It assesses the predictive performance or model behavior across different subgroups defined by multiple dimensions or features (such as race, gender, age, etc.).

2.2 MDSS Demo

2.2.1 Demo Results

The subgroups that MDSS identifies:

- Priviliged groups: individuals that belong to: Non-caucasian, less than 25, Male
- Unpriviliged groups: individuals that belong to: Non-caucasian, Female

2.2.2 Target groups

In this exercise we will be comparing the bias between:

1. Priviliged by MDSS vs. Opposite in one attribute:
    - Non-caucasian, less than 25, Male vs. Non-caucasian, less than 25, Female
    - Non-caucasian, less than 25, Male vs. Caucasian, less than 25, Male
2. Unpriviliged by MDSS vs. Opposite in one attribute:
    - Non-caucasian, Female vs. Caucasian, Female
    - Non-caucasian, Female vs. Non-caucasian, Male

2.3 Bias Mesurement and Comparison

2.3.1 Non-caucasian, less than 25, Male And Non-caucasian, less than 25, Female

Assuming Non-caucasian, less than 25, Male as **Priviliged** and Non-caucasian, less than 25, Female as **Unpriviliged**.

- Train set: Difference in mean outcomes between unprivileged and privileged groups = 0.148845
- Test set: Difference in mean outcomes between unprivileged and privileged groups = 0.328246

2.3.1.1 Likelihood of Recidivism

- **Non-caucasian, less than 25, Male:** The model estimates that on average there is a 42% chance they will not exhibit recidivism while in reality only 32% will not exhibit recidivism.
- **Non-caucasian, less than 25, Female:** The model estimates that on average there is a 53% chance they will not exhibit recidivism while in reality 65% will not exhibit recidivism.

These results suggest that "Non-caucasian, less than 25, Male" could be privileged.

2.3.1.2 Bias score

- **Non-caucasian, less than 25, Male:** MDSS estimates a bias score of 4.6526 when considering this group to be privileged.
- **Non-caucasian, less than 25, Female:** MDSS estimates a bias score of 1.2296 when considering this group to be unprivileged.

Based on the results, there is evidence that "Non-caucasian, Male" is privileged, and also evidence that "Non-caucasian, Female" is unprivileged.

2.3.1.3 Results

Non-caucasian, less than 25, Male group shows to be privileged in comparison to Non-caucasian, less than 25, Female group.

2.3.2 Non-caucasian, less than 25, Male And Caucasian, less than 25, Male

Assuming Non-caucasian, less than 25, Male as **Priviliged** and Caucasian, less than 25, Male as **Unpriviliged**.

- Train set: Difference in mean outcomes between unprivileged and privileged groups = 0.095904
- Test set: Difference in mean outcomes between unprivileged and privileged groups = 0.145438

2.3.2.1 Likelihood of Recidivism

- **Non-caucasian, less than 25, Male:** The model estimates that on average there is a 42% chance they will not exhibit recidivism while in reality only 32% will not exhibit recidivism.
- **Caucasian, less than 25, Male:** The model estimates that on average there is a 46.2% chance they will not exhibit recidivism while in reality 46.8% will not exhibit recidivism.

These results suggest that "Non-caucasian, less than 25, Male" could be privileged.

2.3.2.2 Bias score

- **Non-caucasian, less than 25, Male:** MDSS estimates a bias score of 4.6526 when considering this group to be privileged.
- **Caucasian, less than 25, Male:** MDSS estimates a bias score of 0.0067 when considering this group to be unprivileged.

Based on the results, there is evidence that "Non-caucasian, less than 25, male" is privileged. As for "Caucasian, less than 25, male" the score is very low (close to 0) which doesn't reflect a high bias against this group, it is not unprivileged in this case.

2.3.2.3 Results

Non-caucasian, less than 25, Male group shows to be privileged in comparison to Caucasian, less than 25, Male group.

2.3.3 Non-caucasian, Female And Non-caucasian, Male

Assuming Non-caucasian Male as **Privileged** and Non-caucasian, Female as **Unprivileged**.

- Train set: Difference in mean outcomes between unprivileged and privileged groups = 0.164657
- Test set: Difference in mean outcomes between unprivileged and privileged groups = 0.232482

2.3.3.1 Likelihood of Recidivism

- **Non-caucasian, Female:** The model estimates that on average there is a 56% chance they will not exhibit recidivism while in reality 66% will not exhibit recidivism.
- **Non-caucasian, Male:** The model estimates that on average there is a 46% chance they will not exhibit recidivism while in reality only 43% will not exhibit recidivism.

These results suggest that "Non-caucasian, Male" could be privileged.

2.3.3.2 Bias score

- **Non-caucasian, Male:** MDSS estimates a bias score of 1.9281 when considering this group to be privileged.
- **Non-caucasian, Female:** MDSS estimates a bias score of 4.3036 when considering this group to be unprivileged.

Based on the results, there is evidence that "Non-caucasian, Male" is privileged, and also evidence that "Non-caucasian, Female" is unprivileged.

2.3.3.3 Results

Non-caucasian, Male group shows to be privileged in comparison to Non-caucasian, Female group.

2.3.4 Non-caucasian, Female And Caucasian, Female

Assuming Caucasian Female as **Privileged** and Non-caucasian, Female as **Unprivileged**.

- Train set: Difference in mean outcomes between unprivileged and privileged groups = -0.035810
- Test set: Difference in mean outcomes between unprivileged and privileged groups = 0.025282

2.3.4.1 Likelihood of Recidivism

- **Non-caucasian, Female:** The model estimates that on average there is a 56% chance they will not exhibit recidivism while in reality 66% will not exhibit recidivism.
- **Non-caucasian, Male:** The model estimates that on average there is a 68% chance they will not exhibit recidivism while in reality only 64% will not exhibit recidivism.

These results suggest that "Caucasian, Female" could be privileged.

2.3.4.2 Bias score

- **Caucasian, Female:** MDSS estimates a bias score of 0.5258 when considering this group to be privileged.
- **Non-caucasian, Female:** MDSS estimates a bias score of 4.3036 when considering this group to be unprivileged.

Based on the results, there is evidence that "Non-caucasian, Male" is privileged, and also evidence that "Non-caucasian, Female" is unprivileged.

2.3.4.3 Results

Caucasian, Male group shows to be privileged in comparison to Non-caucasian, Female group.