INFO-H-420
Management of Data Science and Business Workflows

Assignment 4
- REPORT -

Authors:
Cedric Kouamou Djampo
Salma SALMANI

Professor:
Dimitrios SACHARIDIS

Academic year 2023-2024

1 Exercise 1 :fairness with COMPAS dataset

The aim of this exercise was to study the concepts of algorithmic fairness using the AIF 360 tool and the COMPAS dataset. The exercise consisted in correcting the biases linked to the protected attributes (race and sex) using the re-balancing technique, then training a logistic regression model and evaluating the biases on a set of tests.

1.1 COMPAS dataset.

The COMPAS dataset (Correctional Offender Management Profiling for Alternative Sanctions) is used to assess offenders' risk of recidivism. It includes information such as age, sex, race, severity of charges, criminal history, and the COMPAS score, which is a calculated risk score. This dataset is frequently used in algorithmic fairness studies to analyse whether criminal risk prediction tools treat individuals from different demographic groups fairly.

As part of this exercise, we are particularly interested in correcting the biases associated with sex, age and race in the COMPAS dataset. The aim is to detect and minimise potential biases associated with these protected attributes, in order to promote greater fairness in predictions of the risk of recidivism.

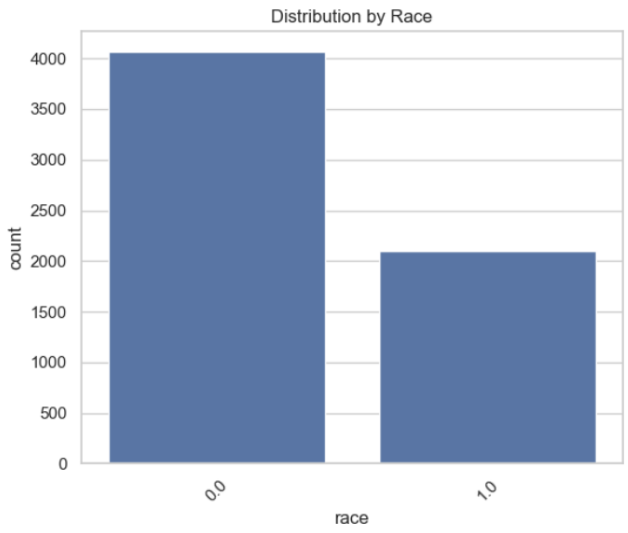

Figure 1: Distribution by Race

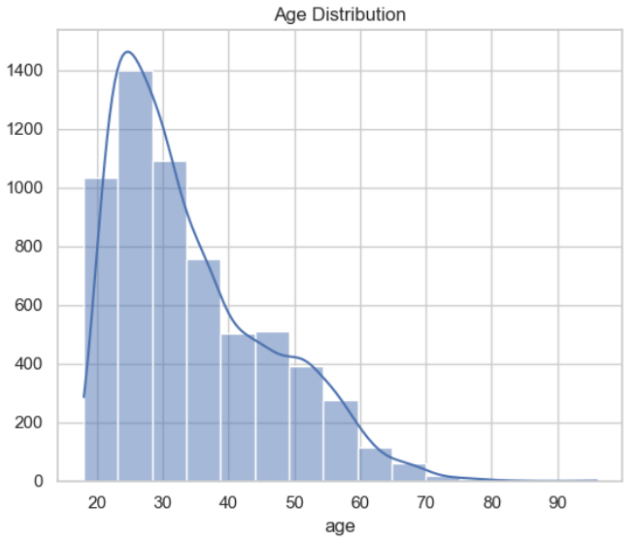

Figure 2: Distribution by Age

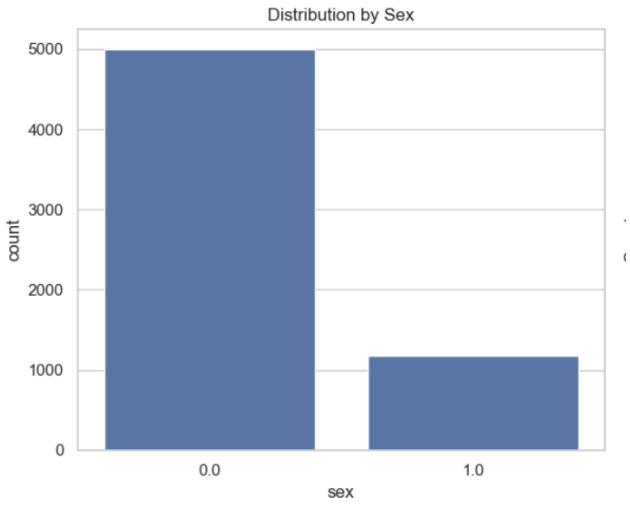

Figure 3: Distribution by Sex

1.1.1 Consider sex to be the protected attribute, fix the bias using the reweighing preprocessing technique, and measure the bias assuming race is the protected attribute.

As part of the algorithmic fairness exercise, a crucial step was to address sex bias using the Reweighing technique and then assess the impact of this correction on race bias. The approach adopted involved adjusting the weights of the instances in the training dataset to reduce sex imbalances. A logistic regression model was then trained on these adjusted data (see Assignment4.ipynb).

$$\text{Difference in Mean Outcomes (Race): } -0.22600139324277257$$
$$\text{Disparate Impact (Race): } 0.6874894641782059$$

Output without Logistic Regression:

$$\text{Difference in Mean Outcomes (Race): } -0.15210727969348659$$
$$\text{Disparate Impact (Race): } 0.7495268138801262$$

The results obtained after applying this model to a set of tests reveal significant aspects. The "Difference in Mean Outcomes" and "Disparate Impact" metrics were used to assess racial bias. These metrics showed significant differences between the results obtained with and without the use of the trained model. With the model, the results indicated a more pronounced racial bias, underlining the importance of training the model in assessing equity. This finding suggests that although correction for sex bias has been achieved, race bias remains or has been introduced, indicating that correction for sex bias does not automatically guarantee correction for race bias. These results highlight the complexity and interdependence of biases in machine learning models, as well as the importance of rigorous evaluation to ensure effective algorithmic fairness.

1.1.2 Consider race to be the protected attribute, fix the bias using the reweighing prepossessing technique, and measure the bias assuming sex is the protected attribute

In this part of the algorithmic fairness exercise, the aim was to deal specifically with race bias, using the reweighting technique, and then to assess the impact of this correction on sex bias. This approach involved first modifying the weights of the instances in the training dataset to reduce racial imbalances. A logistic regression model was then trained on these adjusted data.

$$\textbf{Difference in Mean Outcomes (Sex):} -0.16183574879227047$$
$$\textbf{Disparate Impact (Sex):} 0.7744107744107744$$

Once the model was formed, it was applied to a test set to make predictions as in the previous question. The evaluation of these predictions was carried out focusing on sex as the protected attribute, using the same metrics. The results obtained showed that, although efforts had been made to mitigate racial bias, significant sex-based biases still remained in the model predictions. In particular, the metrics indicated a disadvantage for the non-privileged group based on sex, even after correcting for racial bias.

1.1.3 Measure with age_cat=Less than 25

In this section, the focus was on individuals under 25 years of age, with a concentration on the evaluation of sex and race biases after the application of the reweighting technique. Two logistic regression models were trained separately, one adjusted for sex bias and the other for race bias. The results of the tests on this specific set of ages revealed some interesting information.

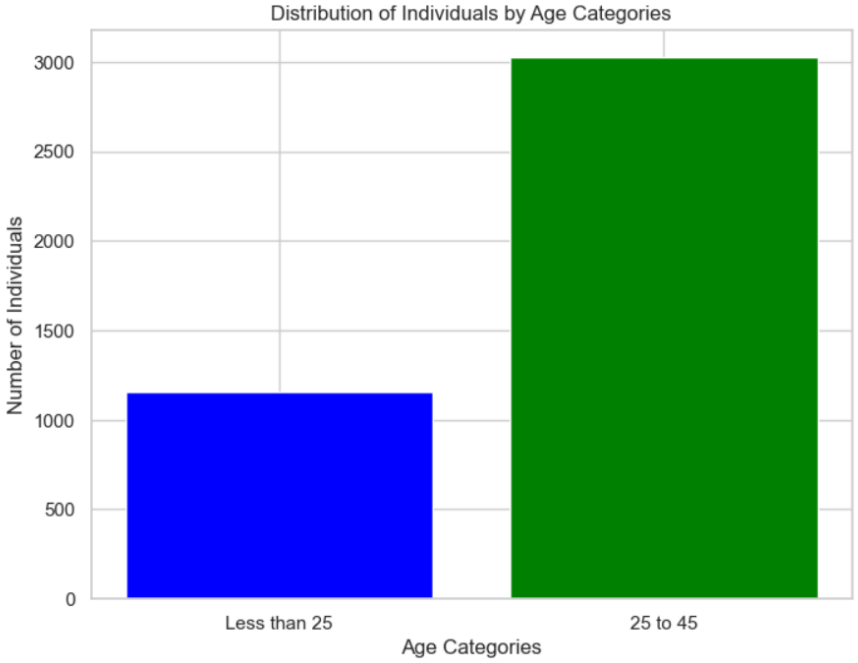

Figure 4: Age categories:

Metric	Value
Difference in Average Scores (Sex)	0.12265018289939555
Disparate Impact (Sex)	1.477459640572647
Difference in Average Scores (Race)	-0.03597926895799236
Disparate Impact (Race)	0.9164925609370054

Table 1: Metrics for Sex and Race

For sex, the metrics showed a clear bias in favour of the privileged group, with a significant positive difference in mean scores and a disparate impact greater than 1. This indicates that the privileged group based on sex receives more favourable treatment by the model. In contrast, the race measures revealed a slightly unfavourable bias towards the non-privileged group, as evidenced by the negative difference in mean scores and a disparate impact of slightly less than 1, although these biases were less pronounced than those related to sex.

1.1.4 Measure with age_cat=25 to 45

same process here with age over 25

Metric	Value
Difference in Average Scores (Sex)	-0.14744302152615107
Disparate Impact (Sex)	0.7922723933178364
Difference in Average Scores (Race)	-0.2633067478028718
Disparate Impact (Race)	0.6511800794293732

Table 2: Metrics for Sex and Race

These results indicate that, for the 25 to 45 age group, biases towards non-privileged groups are more pronounced for both sex and race than for individuals under 25. In particular, the racial bias is more pronounced in this age group. This suggests that age factors interact with sex and race in a significant way, influencing how biases manifest themselves in the model's predictions.

These observations illustrate the importance of an intersectional approach to the analysis of algorithmic fairness. They highlight the need to examine not only each protected attribute individually but also to understand how these attributes interact across different age groups. Such in-depth analysis is crucial to developing effective bias mitigation strategies, tailored to the specificities of each demographic sub-group.

2 Exercise 2 : MDSS

The main objective of the exercise was to examine in detail the bias in the predictions of a classification model by focusing on specific demographic groups defined by combinations of age and sex. Using the Multi-Dimensional Subset Scan (MDSS) method, the exercise aimed to identify the preferred and non-preferred groups in the context of these predictions and to measure the extent of bias for each group. The approach was then to compare these measures of bias between distinct groups, based on sex. The first group was characterised by individuals aged under 25 of Caucasian race, and this group was compared with another group similar in age but of African-American race. The aim of this comparison was to assess how different combinations of demographic characteristics affected the model's predictions, revealing the subtleties of potential bias in its predictions and highlighting the need for adjustment to ensure fairness and equity in algorithmic processes.

2.1 Data preparation :

Subsets of the data were created for each of the target groups (Caucasians and African-Americans under the age of 25). These subsets were divided into training and test sets to train and evaluate the logistic regression models.

```
Number of individuals under 25 and caucasian: 347
Number of people under 25 who are African-American: 809

Overview of individuals under 25 and Caucasian:
    sex  race  age_cat  priors_count  c_charge_degree  two_year_recid
16  1.0   1.0      0.0           0.0              1.0             0.0
25  0.0   1.0      0.0           1.0              0.0             1.0
54  0.0   1.0      0.0           0.0              1.0             0.0
65  1.0   1.0      0.0           0.0              1.0             0.0
71  0.0   1.0      0.0           2.0              1.0             1.0

Preview of under-25s who are African-American:
    sex  race  age_cat  priors_count  c_charge_degree  two_year_recid
1   0.0   0.0      0.0           2.0              1.0             1.0
5   0.0   0.0      0.0           1.0              0.0             1.0
14  0.0   0.0      0.0           1.0              1.0             1.0
28  1.0   0.0      0.0           1.0              1.0             0.0
39  0.0   0.0      0.0           1.0              1.0             1.0
```

Figure 5: Data per group

2.2 Model Training and Bias scoring:

In the Assignment4.ipynb notebook the separate logistic regression models were trained for each group, using the respective training sets.

2.2.1 Caucasian Group Under the Age of 25

Mean Model Prediction (Model Not Recid): The model's prediction for the non-recidivism rate among the Caucasian group under the age of 25 is given by:

$$\text{Mean Model Prediction (Model Not Recid)} = 0.351998$$

This suggests that the model predicts that only about 35.2% of individuals in this Caucasian group under the age of 25 will not recidivate.

Mean Actual Observations (Observed Not Recid): The actual observed non-recidivism rate in the same group is:

$$\text{Mean Actual Observations (Observed Not Recid)} = 0.730769$$

This indicates that, in reality, approximately 73.1% of individuals in this group did not reoffend.

2.2.2 African-American Group

Mean Model Prediction (Model Not Recid): The model's prediction for the non-recidivism rate among the African-American group is given by:

$$\text{Mean Model Prediction (Model Not Recid)} = 0.42175$$

This means that the model predicts that 42.175% of individuals in this group will not recidivate.

Observed Not Recid: The actual observed non-recidivism rate in the same group is:

$$\text{Observed Not Recid} = 0.47500$$

This indicates that, in reality, 47.5% of the individuals in this group did not reoffend.

Analysis of the Difference: The difference between the model's predictions and actual observations is less pronounced than in the Caucasian group. Here, the model slightly underestimates the probability of African-American individuals not reoffending, but the difference is relatively small (around 4.725%).

2.3 MDSS (Multi-Dimensional Subset Scan)

The Multi-Dimensional Subset Scan (MDSS) method is described as a technique implemented in AIF 360, used for detecting instances of unfairness in sub-populations. It involves examining both privileged and unprivileged groups, measuring bias within these groups, and comparing it to other groups with opposite race or sex characteristics [1].

Based on analyses carried out using the MDSS (Multi-Dimensional Subset Scan) metric on two specific groups - Caucasians and African-Americans - under the age of 25, here is a conclusion:

Caucasian Group

- MDSS scores are 0 for both males and females, indicating that there is no significant bias detected by the model towards or against either sex in this group.
- This lack of bias suggests that the model treats Caucasian men and women under the age of 25 relatively equally with respect to the criteria analyzed by the MDSS.

African-American Group

- For African-American men, the MDSS score is also 0, indicating no significant bias in their favor.
- On the other hand, for African-American women, the MDSS score is relatively high (0.9194), suggesting a significant bias against them. This bias indicates that the model may underestimate the probability of a favorable label (non-recurrence) for African-American women compared with men.

References

[1] Zhe Zhang, Daniel B. Neill. *Identifying Significant Predictive Bias in Classifiers*, FAT/ML 2017. https://arxiv.org/abs/1611.08292