Human-centered Machine Learning 2024 @ UU

Timofey Senchenko
t.senchenko@students.uu.nl(7447655)
Utrecht University
Utrecht, the Netherlands

Efraim Dadhl
e.y.c.dahl@students.uu.nl(1695568)
Utrecht University
Utrecht, the Netherlands

KEYWORDS

machine learning, fairness, explainability

ACM Reference Format:
Timofey Senchenko and Efraim Dadhl. 2024. Human-centered Machine Learning 2024 @ UU. In *Proceedings of Utrecht University (INFOMHCML'2023).* Utrecht University, 5 pages.

1 INTRODUCTION

In this programming assessment, we investigate the effect of different fairness strategies on a classifier trained on the COMPAS Dataset provided by ProPublica[3]. COMPAS short for Correctional Offender Management Profiling for Alternative Sanctions, is a software used to calculate risk scores for recidivism in criminal suspects, to aid judges' decision-making on whether or not to grant bail, etc. COMPAS has come under wide scrutiny for racial discrimination and is often used to highlight the ways algorithms can reinforce structural disadvantages embedded in society. The report evaluates the effects of removing protected attributes (fairness through blindness), reweighing, and enforcing equalized odds in post-processing on the quality and equity of the prediction provided by a simple regression algorithm. The analysis of fairness is provided with respect to race and later its intersection with sex.

2 DATA-EXPLORATION

The COMPAS dataset contains a wide range of features, including personal data such as race, sex, and age, data about the offense, type of offense, etc, and of course, whether there was a reoffence, and data about the re-offense and information about the associated Compas evaluation. In figure 1 we show the demographic makeup of the dataset and in figure 2 we show the proportion of non-recidivism by race and sex. There are 5278 instances in the dataset: 2103 Caucasians and 3175 African Americans, 4247 Men and 1031 Women. The non-re-offense rates for women are higher than for men 63% as opposed to 50%, and higher for Caucasians than for African Americans 61% vs 48%. African American men have particularly low rates at 44% Caucasian women have the highest non-re-offence rates at 65%, and Caucasian men have lower positive outcome rates than African American women at 60% vs 63%.

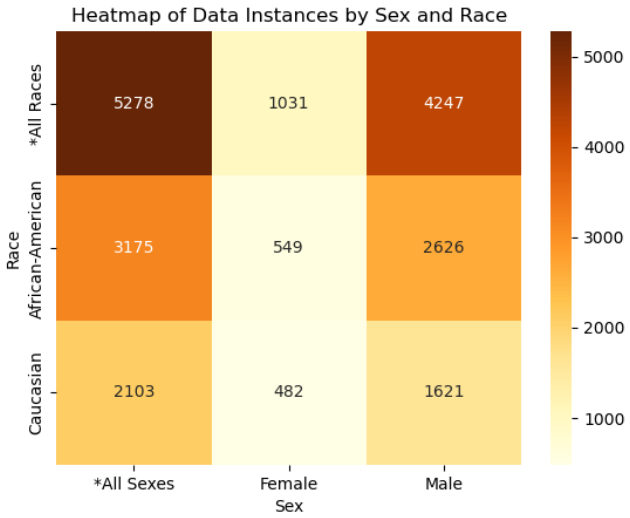

Figure 1: Data Distribution by Race and Sex

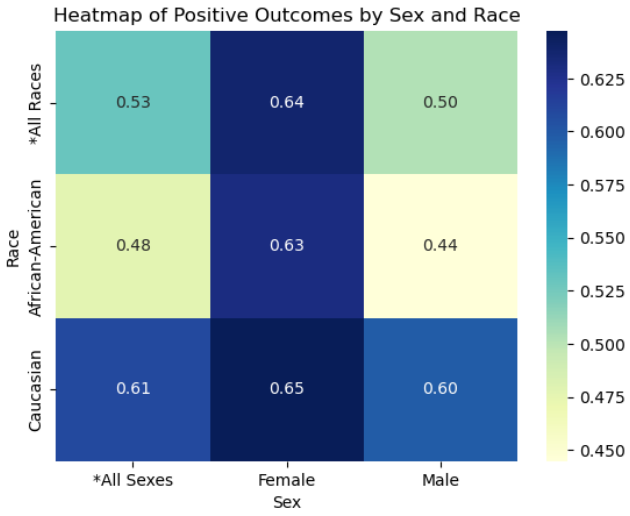

Figure 2: Recidivism by Race and Sex

3 DATA PREPARATION/FEATURE SELECTION

Beyond sex, race, and the predictive label 2_year_recidivism, we are taking the following features into account: **Age, Age Category**: How old the suspect is can affect their recidivism rate, i.e., senior suspects may be less likely to re-offend violently.

INFOMHCML'2023, April 2023, Utrecht, the Netherlands

ACM ISBN xxxxxxx.
https://doi.org/xxxxxxx

Date of Birth: Gives hints as to what generation a suspect grew up in, may affect the suspect's economic opportunities, etc.
Offense Date: Crimes may have seasonal trends, and be affected by global situations, i.e. 2020 pandemic.
Arrest Date: A large difference between offense data and an arrest date may indicate evasion.
Offense Description About 65% of the data falls within 10 offense descriptions, with the rest assigned to a default category. Descriptions include Battery, Grand Theft, etc. Many of the infrequent descriptions combine offense, i.e. battery + possession of cocaine, in more advanced analysis this could be accounted for more flexibly. For our analysis, this is a categorical feature with 10 possible categories. We also include the column *c_charge_degree* which indicates whether the offense was a misdemeanor (M) or a felony (F).
Data on previous criminal history: *priors_count* (number of past offenses) *juv_fel_count* (number of juvenile felonies), *juv_misd_count* (number of juvenile misdemeanors)
Time in Jail and Custody: We include the columns *jail_in*, *jail_out*, *custody_in*, *custody_out* because time spent in jail and custody can indicate severeness of the crime and have an influence on recidivism.
We exclude any information on re-offences such as *is_recid* because that would leak predictive data. We also exclude any data related to the COMPAS screening including screening dates, scores, and score categories.

We are splitting our data, randomly sampling 80% for training and 20% for testing.

4 METRICS

For each classifier, we collect the following metrics: **Performance**: Accuracy, Precision, Recall and F1 score. **Fairness**: True positive rates (TPR) and false positive rates (FPR) for each group. Additionally, we calculate *disparate impact* as the ratio of positive predictions in each group. We also calculate three binary variables according to a disparity threshold $\phi = 0.8$. *demographic_parity* indicates whether disparate impact falls below this threshold, which indicates that positive prediction rates are very different for the two groups. *equal_opportunity* indicates that true positive rates between groups are within a factor of ϕ of each other. *predictive_parity* indicates that false negative rates are within a factor of ϕ of each other.

5 RESULTS AND INTERPRETATION

5.1 Regression

For all of our experiments, we use the sklearns logistic regression model with the default hyperparameters for ease of comparison. The baseline experiment uses all features including the protected feature race. It achieves the following performance scores:

- accuracy: 72.06%
- precision: 72.89%
- recall: 64.52%
- F1 score: 68.45%

Regarding the fairness measures we see the following results:
Demographic parity fails the 80% threshold (TPR African American / TPR Caucasians = 0.7832)

True Positive Rate (TPR): African American - 0.7397; Caucasian - 0.849
Using $\phi = 0.8$ equal opportunity is maintained for this classifier (ratio: 0.8712).

False Positive Rate (FPR): African American - 0.3232; Caucasian - 0.4167 Here we also see a noticeable difference, and if Using $\phi = 0.8$ predictive parity is not met (ratio: 0.7756).

5.2 Regression with data removed

For the experiment where the protected attribute in this case race is removed from the features set for training, we see the following metrics:

- accuracy: 72.16%
- precision: 72.95%
- recall: 64.72%
- F1 score: 68.59%

The performance scores are very similar to the performance of the baseline-regression, it even shows a slight improvement. As for the fairness measures we see the following results:
The positive prediction ratio over the two groups, Demographic parity satisfies the 80% threshold (ratio:0.8123)

True Positive Rate: African American - 0.7524; Caucasian - 0.8327
The ratio is improved to 0.9035. As in the previous experiment, this meets our threshold $\phi = 0.8$ for equal opportunity but offers additional improvement.

False Positive Rate: African American - 0.3262; Caucasian - 0.4048
The FPR ratio now satisfies the 80% threshold (ratio: 0.8058), which means predictive parity is met.

5.3 Reweighed Regression

For the experiment where samples are reweighted for the training, we see the following metrics:

- accuracy: 72.73%
- precision: 73.64%
- recall: 65.32%
- F1 score: 69.23%

Again, the performance metrics are very similar to the baseline model. For the fairness measures we see a great improvement:
TPR and FPR ratios of the two groups are now almost perfectly balanced
AA_TPR/CC_TPR ratio: 1.0591 (African American - 0.8127; Caucasian - 0.7673)
AA_FPR/CC_FPR ratio: 1.0069 (African American - 0.3476; Caucasian - 0.3452)
Demographic parity, predictive parity, and equal opportunity surpass the 0.8 threshold.

Demographic parity satisfies the 80% threshold and is even further improved (African American/Caucasians = 0.9661)

5.4 Equalized Odds

For equalized odds, we randomly generated 1000 groups of parameters one for each combination of groups and outcomes i.e. [(Caucasian, positive),(African-American, positive)...]. Then we define a classifier that creates a derived prediction of a sample based on the prediction of the regular regression model, the group membership of that sample, and the random parameter for that group and outcome combination, which we treated as a probabilistic threshold. We then iterate through the parameters minimizing the equalized odds score. We define the equalized odds score as abs(True Positive Rate (African American)-True Positive Rate (Caucasian))+abs(False Negative Rate (African American) - True Positive Rate (Caucasian)) The best set of parameters where Positive-Caucasian:0.4532, Positive-African-American:0.8970, Negative-Caucasian: 0.6149, Negative-Black:0.4374. Each of these parameters is the probability that a result will be set to positive. The classifier equipped with these parameters achieves the following metrics:

- accuracy: 53%
- precision: 50%
- recall: 40%
- f1: 45%,
- Caucasian_true_positive_rate: 49%,
- Caucasian_false_positive_rate: 55%,
- African-American_true_positive_rate: 76%,
- African-American_false_positive_rate: 61%.

a For the fairness measurements of equal opportunity, demographic, and predictive parity this classifier is within a factor of $\phi = 0.8$ of each other. The graph in 3 shows the spread of accuracy vs our equalized odds score.

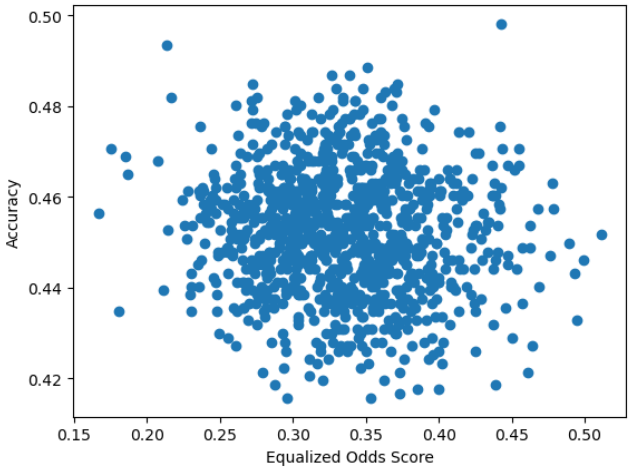

Figure 3: Equalized Odds Scores vs accuracy

Interpretation: The spread of the 1000 samples indicates that most parameter groups fall in equalized odds scores between 0.25 and 0.45 and accuracy scores between 40 and 50%. Optimizing for the best-equalized odds score does not indicate optimal performance in terms of accuracy. This is to be expected since false-positive rates can be close to each other for both groups while still being high (and vice versa for true-positive rates). To prevent this, it would be appropriate to incorporate measures of performance into the optimization function alongside measures of fairness. How these are weighted would be subject to further investigation/discretion of the users. Another limiting factor is using randomly generated samples as opposed to an optimization algorithm In summary, increasing the fairness metrics for the first three experiments came at very low cost in terms of performance. Equalized odds performed poorly, not exceeding reweighing or feature blindness for fairness, and severely underperforming all other experiments in performance, likely due to the crude sampling and optimization used.

6 INTERSECTIONALITY

6.1 Intersectional Feature: Race and Sex Combined

In this experiment, we combine the protected attributes of race and sex into a single intersectional feature, creating four distinct groups: Female African American, Male African American, Female Caucasian, and Male Caucasian. This allows us to analyze the intersectional impacts on the model's performance and fairness measures. For comparison, we selected Female African Americans as the baseline group because they exhibited the highest True Positive Rate (TPR) and False Positive Rate (FPR).

The distribution of samples across these groups is as follows: Female African American - 126, Female Caucasian - 98, Male African American - 517, Male Caucasian - 315. It is important to note that the female groups have fewer samples, which may influence the TPR and FPR differences observed.

The performance metrics for this model are as follows:

- accuracy: 72.06%
- precision: 72.89%
- recall: 64.52%
- F1 score: 68.45%

Regarding the fairness measures, we observe the following results:

True Positive Rate (TPR):

- Female African American: 0.9634
- Female Caucasian: 0.8594
- Male African American: 0.6609
- Male Caucasian: 0.8453

False Positive Rate (FPR):

- Female African American: 0.5455
- Female Caucasian: 0.5294
- Male African American: 0.2852
- Male Caucasian: 0.3955

We evaluate fairness using the following measures:

Demographic Parity:

- Female Caucasian vs. Female African American: 0.9112 (satisfies the 80% threshold)
- Male African American vs. Female African American: 0.5560 (does not satisfy the threshold)
- Male Caucasian vs. Female African American: 0.8000 (barely satisfies the threshold)

Equal Opportunity (TPR Ratio):

- Female Caucasian vs. Female African American: 0.8920 (satisfies the 80% threshold)
- Male African American vs. Female African American: 0.6860 (does not satisfy the threshold)
- Male Caucasian vs. Female African American: 0.8774 (satisfies the threshold)

Predictive Parity (FPR Ratio):

- Female Caucasian vs. Female African American: 0.9705 (satisfies the 80% threshold)
- Male African American vs. Female African American: 0.5228 (does not satisfy the threshold)
- Male Caucasian vs. Female African American: 0.7250 (does not satisfy the threshold)

In summary, while combining race and sex into a single intersectional feature allows for a more detailed analysis of fairness, it also highlights significant disparities among the groups. The Female Caucasian group maintains fairness measures relatively close to those of the Female African American baseline, but the Male African American group shows considerable discrepancies, particularly in TPR and FPR, failing to meet the 80% threshold for both equal opportunity and predictive parity. The sample size disparity, particularly the smaller number of samples in the female groups, might influence these results and should be considered when interpreting the fairness measures.

6.2 Regression with Intersectional Feature as Protected Attribute

In this experiment, the combined sex+race feature is protected and is not accessible by the model.

The performance metrics for this model are as follows:

- accuracy: 72.35%
- precision: 73.39%
- recall: 64.52%
- F1 score: 68.67%

Once again the basic classification metrics see rather unnoticeable improvement. Regarding the fairness measures, we observe the following results:

True Positive Rate (TPR):

- Female African American: 0.8537
- Female Caucasian: 0.8281
- Male African American: 0.721
- Male Caucasian: 0.8453

False Positive Rate (FPR):

- Female African American: 0.4773
- Female Caucasian: 0.5294
- Male African American: 0.3028
- Male Caucasian: 0.3806

Picking up the group with the highest TPR and FPR as the baseline group (the baseline group again turns out to be female African-American) yields the following results:

Demographic Parity:

- Female Caucasian vs. Female African American: 1.0031 (satisfies the 80% threshold)
- Male African American vs. Female African American: 0.6803 (does not satisfy the threshold)
- Male Caucasian vs. Female African American: 0.8967 (satisfies the threshold)

Equal Opportunity (TPR Ratio):

- Female Caucasian vs. Female African American: 0.9700 (satisfies the 80% threshold)
- Male African American vs. Female African American: 0.8446 (does not satisfy the threshold)
- Male Caucasian vs. Female African American: 0.9902 (satisfies the threshold)

Predictive Parity (FPR Ratio):

- Female Caucasian vs. Female African American: 1.1092 (satisfies the 80% threshold)
- Male African American vs. Female African American: 0.6344 (does not satisfy the threshold)
- Male Caucasian vs. Female African American: 0.7974 (does not satisfy the threshold)

The TPR and FPR ratios for the Female Caucasian group compared to the Female African American baseline group both satisfy the 80% threshold, indicating maintained equal opportunity and predictive parity. The Male Caucasian group also shows an improved TPR ratio, satisfying the equal opportunity threshold, but its FPR ratio does not meet the predictive parity threshold. The Male African American group still shows considerable discrepancies in both TPR and FPR ratios, failing to meet the 80% threshold for both equal opportunity and predictive parity.

Overall, we see that simply removing race and sex from the feature set does noticeably improve certain fairness metrics, but ultimately doesn't bring equal balance to all groups. This could be caused by proxy features leaking information about protected groups' attribution. As we saw earlier in section 5 experiments, more advanced techniques such as reweighing might deal with this issue more efficiently.

7 DISCUSSION

The use of ML systems for recidivism prediction is an ethical gray zone. On the one hand, they offer great opportunities for fairer, more öbjective*;* streamlined, and efficient practices. Human decision-making is full of unconscious cognitive biases that influence how a decision is made. This can include social biases about certain groups of people, but also other heuristic biases[2] such as availability or the anchoring effect. Additionally, environmental factors such as the time of day, mental depletion, and even whether or not a judge has eaten recently[1] can unfairly affect a judge's decision - Though the effect size of the so-called hungry judge effect is often overstated. At first glance, ML systems seem promising. They are not influenced by their immediate environment and are not subject to many human cognitive biases. However, ML does incorporate human biases from training data, such as racial discrepancies, which may be the result of systemic discrimination. More subtle biases can also arise through development decisions such as how the algorithm is optimized etc. ML-produced scores, such as the COMPAS

score can be harmful when they are presented as objective metrics. Even if judges are aware of a model's shortcomings, the model may still influence the process through anchoring effects, by setting a baseline for each subject. This combined with the lack of explain-ability and insight into how the model makes its decisions, leaves a lot of room for improvement in how the models are built, evaluated, and used. The lack of explain-ability and insight is additionally aggravated by the developers of COMPAS, Northpoint who refused to give both court officials and investigative journalists insight into the workings of their model, citing intellectual property rights. [4] This combination of lack of transparency combined with for-profit incentives, is a potential breeding ground for abuse.

The release of the COMPAS dataset and similar datasets can also be problematic. While criminal records are generally considered public in the US, releasing this dataset may draw additional attention to subjects, damaging their opportunities to participate in broader society after their release. Northpoint's reluctance to share further details about the workings of their algorithm beyond the underlying data greatly inhibits the review of their process. This dataset offers a unique opportunity to scrutinize the COMPAS algorithm in a semi-blind fashion as is done by ProPublica[4]. With neither the data nor the algorithm available this type of investigation would be impossible.

REFERENCES

[1] Levav Danziger, Shai. 2011. Extraneous Factors in Judicial Decisions. *Proceedings of the National Academy of Sciences 108, no 17: 6889-6892* (2011). https://doi.org/10.1073/pnas.1018033108

[2] Daniel Kahneman. 2011. *Thinking, fast and slow.* Farrar, Straus and Giroux, New York. https://www.amazon.de/Thinking-Fast-Slow-Daniel-Kahneman/dp/0374275637/ref=wl_it_dp_o_pdT1_nS_nC?ie=UTF8&colid=151193SNGKJT9&coliid=I3OCESLZCVDFL7

[3] Julia Angwin Lauren Kirchner Surya Mattu, Jeff Larson. 2016. How We Analyzed the COMPAS Recidivism Algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[4] Julia Angwin Lauren Kirchner Surya Mattu, Jeff Larson. 2016. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing