Human-centered Machine Learning 2021 Fairness Assignment

Olusanmi Hundogan - 6883273
Ruben Potthoff - 6137210
O.A.Hundogan@students.uu.nl,R.D.Potthoff@students.uu.nl
Utrecht University
Utrecht, the Netherlands

ABSTRACT

The abstract is not necessary for programming assignments.

KEYWORDS

machine learning, fairness

ACM Reference Format:
Olusanmi Hundogan - 6883273, Ruben Potthoff - 6137210. 2021. Human-centered Machine Learning 2021 Fairness Assignment. In *Proceedings of Utrecht University (INFOMHCML'2021).* Utrecht University, 3 pages.

1 INTRODUCTION

In this work, we use AI Fairness 360 library to implement our experiments [1].

2 METHODS

The data was collected by ProPublica through a public records request to Broward County, Florida. The data included COMPAS scores and recidivism rates on 18610 people. To determine race, the race classifications used by the Broward County Sherrif's Office was used.

3 RESULTS

We split the fairness data into a test- and training set, with the test set being 20 % of the data and the training set being 80% of the data.

3.1 Test Set

The test set consists of *634* African-American and *422* Caucasians. Recidivism for African-Americans within 2 years had a mean of *0.52* with a standard deviation of *0.5* ,while Caucasians recidivism had a mean of *0.39* and a standard deviation of *0.49*. Note that the high standard deviation is due to the fact that we are dealing with binary data, so as it can only be 0 or 1 deviation is naturally higher.
The test data had *856* males and *200* females, with males recidivism having a mean of *0.49* and a standard deviation of *0.5*, while females had a mean of *0.36* and a standard deviation of *0.48*.
Of the African-Americans, there were *522* males and *112* females. The African-American males had a recidivism rate mean of *0.56* with standard deviation of *0.50*, with African-American female recidivism rate averaging at *0.34* and a standard deviation of *0.48*. In Caucasians there were *334* males and *88* females, with male recidivism averaging at *0.40* standard deviation *0.49*, while the Caucasian female mean was *0.39* with standard deviation *0.49*.
The tables below provide an overview of this data.

race	count	mean	std
African-American	634.0	0.517350	0.500093
Caucasian	422.0	0.393365	0.489076

sex	count	mean	std
Male	856.0	0.492991	0.500243
Female	200.0	0.360000	0.481205

race	sex	count	mean	std
African-American	Male	522.0	0.555556	0.497381
	Female	112.0	0.339286	0.475595
Caucasian	Male	334.0	0.395210	0.489629
	Female	88.0	0.386364	0.489706

3.2 Training Set

The test set consists of *2541* African-American and *1681* Caucasians. Recidivism for African-Americans within 2 years had a mean of *0.52* with a standard deviation of *0.5* ,while Caucasians recidivism had a mean of *0.39* and a standard deviation of *0.49*.
The test data had *3391* males and *831* females, with males recidivism having a mean of *0.50* and a standard deviation of *0.5*, while females had a mean of *0.36* and a standard deviation of *0.48*.
Note that the mean and standard deviations of the test set and training set are very similar, but this is because they are randomly sampled from the same data set so the distributions are expected to be similar. Of the African-Americans, there were *2104* males and *437* females. The African-American males had a recidivism rate mean of *0.56* with standard deviation of *0.50*, with African-American female recidivism rate averaging at *0.38* and a standard deviation of *0.48*. In Caucasians there were *1287* males and *394* females, with male recidivism averaging at *0.40* standard deviation *0.49*, while the Caucasian female mean was *0.35* with standard deviation *0.47*.
Interesting to note, as the female group specific to each race is so small the recidivism rates for these groups differ slightly between the test and training sets.
The tables below provide an overview of this data.

INFOMHCML'2021, April 2021, Utrecht, the Netherlands

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

Olusanmi Hundogan - 6883273
Ruben Potthoff - 6137210

race	count	mean	std
African-American	2541.0	0.524597	0.499493
Caucasian	1681.0	0.390244	0.487950

sex	count	mean	std
Male	3391.0	0.497788	0.500069
Female	831.0	0.362214	0.480930

race	sex	count	mean	std
African-American	Male	2104.0	0.555133	0.497069
	Female	437.0	0.377574	0.485336
Caucasian	Male	1287.0	0.404040	0.490896
	Female	394.0	0.345178	0.476031

Noticeably, there are fewer females than males in this data-set, as well as fewer Caucasians than African-Americans. Caucasian females had the lowest recidivism rates across all categories, followed by African-American females. Males of both categories had higher recidivism rates than females, with African-American recidivism being the highest.

3.3 Classifiers

We trained four different classifiers on this data. The first makes use of all features, the second one excludes race, the third one classifies after reweighing is done on the training set, and the fourth one happens after post-processing. All metrics for these classifiers can be found in the table towards the end of this subsection.

For the first classifier, which utilized all features, had a precision of *0.68*, a recall of *0.56*, an F1 of *0.61* and an accuracy of *0.66*. The statistical parity criterion is far from satisfied as it has a statistical parity of about *-0.17*. The equal opportunity criterion is left unsatisfied as well, being around *-0.10*.

The second classifier, which dropped the race feature, scored very similarly. Precision, recall, F1 and accuracy were *0.68*, *0.56*, *0.61* and *0.66* respectively. Statistical parity was also not satisfied for this classifier at *-0.16*, neither was equal opportunity with *-0.19*.

The third classifier used reweighing[3]. The new weights had a standard deviation of *0.1318* while the mean barely changed (*1.0003*). The distribution of weights can be seen in the figure below

As expected, the mean for each race was 1 as well. Noticeably, the standard deviation for Caucasians was a lot higher than for African-Americans (*0.18* vs *0.11*), with the minimum weight for Caucasians being lower (*0.86* vs *0.89*) and the maximum being higher (*1.22* vs *1.12*).
With precision in this third classifier being *0.66*, recall being *0.61*, F1 being *0.63* and accuracy being *0.66*, the only real improvement over the previous classifiers in these metrics has been recall. Importantly, however, the fairness metrics have improved dramatically.

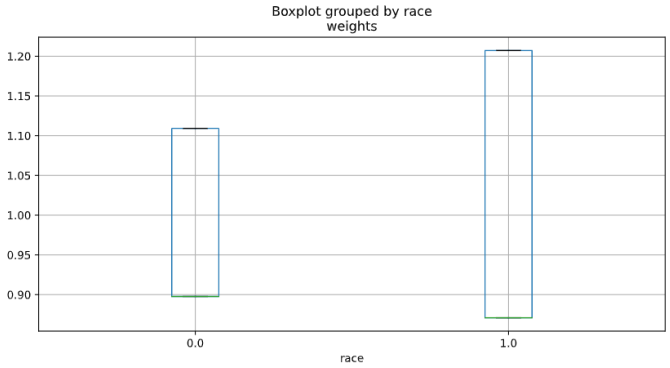

Figure 1: Shows the boxplot of the new weights based on race.

Statistical parity has dropped to only *.02*, and the equal opportunity difference has lowered to *0.08*.
We also trained this classifier on data without using the race feature. The precision for this one was *0.68*, recall *0.56*, F1 *0.61* and accuracy *0.66*. As expected, the statistical parity has worsened drastically to *-0.15* and equal opportunity to *-0.10*. Interestingly, the reweighing does not have any effect in this case (see line 1 and 3 of Table 1). One reason could be that, the weights are strongly tied to the race "class" and not to individual samples. Therefore, the classifier is most likely not able to associate the weights with any other variable.

The fourth classifier used equal odds post processing, which improves the basic metrics dramatically [2, 4]. The precision for this classifier is *0.67*, recall is *0.43*, F1 is *0.53* and accuracy is *0.63*. As expected, the statistical parity also gets greatly reduced to about *0.03*. The equal opportunity criterion has also nearly been satisfied, as the true positive rate difference is *0.05* here. In this case, the traditional metrics are worse than before. Especially, recall and F1 drop strongly below average. This behaviour is reasonable because the postprocessing step alters the best possible predictions and changes the results in order to increase fairness. While reducing the regular metrics, this has dramatically improved fairness measurements.

When comparing statistical parity in the classifiers which did not use reweighing or post processing, you can see that the statistical parity has increased from the original data (where it was around *-0.13*). This is likely because the bias, which has been learned in through training on a biased data set, has been reinforced and the bias has thus become worse.

3.4 Discussion

In this paper we investigated bias in the compas dataset, which is a widely known dataset for testing fairness algorithms. The dataset has information about recidivism cases and their prior risk assessment. We trained investigated the data and trained several classifiers on it to see whether we could train classifiers to accurately predict recidivism.

In the original data itself we immediately noticed the large difference in sample size between males, females and African-American

Table 1: This table shows the metric results for various Logisitic Regression configurations on the Compas Dataset

	Configuration	Precision	Recall	F1	Accuracy	Stat.-Parity	TPR-Diff	TPR_1	TPR_0
0	All Features	0.67711	0.55865	0.61220	0.66288	-0.16177	-0.10589	0.81743	0.71154
1	Without Race	0.67797	0.55666	0.61135	0.66288	-0.15872	-0.10268	0.81743	0.71474
2	Reweighting	0.65739	0.61034	0.63299	0.66288	0.02457	0.07554	0.66805	0.74359
3	Reweighting without Race	0.67797	0.55666	0.61135	0.66288	-0.15872	-0.10268	0.81743	0.71474
4	Metrics on Training Data	0.66764	0.57929	0.62034	0.66746	-0.24799	-0.16118	0.83173	0.67055
5	Metrics on Test Data	0.67711	0.55865	0.61220	0.66288	-0.16177	-0.10589	0.81743	0.71154
6	EqOdds Postprocessing	0.66871	0.43340	0.52593	0.62784	0.03427	0.05300	0.94700	1.00000

Table 2: Contains the information of all the classifiers tested in this analysis. It is worth mentioning, that line 0 and line 5 are identical because both classifiers were trained and tested on the same dataset. We keep line 5 as a direct comparison to line 4. Also noteworthy, line 1 and 3 are also identical.

and Caucasians as well. Immediately noticeable was also the difference between recidivism in males and females, as well as the difference in recidivism between the races that was apparent from the investigation into the training data and the test data.

This difference in recidivism rates across race and sex was inherent to the training data, so if we simply train classifiers on this training data it would only increase and worsen the bias during application. Due to the large difference in sample size between training and test data there is a very noticeable difference between statistical parities. During informal investigations we split the data equally and then the difference becomes much smaller. So, in this case, the large difference seen in Table 1 on line 4 and 5 is likely due to the low test sample size.

The results also show that some variables can act as proxies for protected attributes. This is obvious from the first two classifiers that we trained: One which simply trained on all features and one which trained without considering race. As the fairness metrics remain similar with or without race, we can assume that, the information added by race is covered by other variables.
These classifiers performed very poorly on the fairness metrics, more so than the original training data did. Such biases could be attempted to be fixed through reweighing or post processing and the results show indeed, that these approaches help reduce the issue for the most part. Both these classifiers performed a lot better than the first two on the fairness metrics, and while the reweighed classifier performed similarly to the first two on normal metrics the post processing classifier was also noticeably better at the normal metrics than the other three.

The performance of the first two classifiers raises some very serious ethical concerns about the use of ML systems to predict recidivism. Without reweighing or post processing the classifier, it will perform very poorly on fairness metrics. In fact, it would perform even worse than the original data. As the predictions of such systems are often used in court to greatly affect the lives of many people, it is very important to consider whether the system operates fairly or not.
Additionally, the dataset contains sensitive attributes like race and sex.

Furthermore, this data set shows, that models that have strong impacts on people's lives and opportunities can be very detrimental if the data exerts bias for minority groups. Publishing them publicly before official use creates may create a level of transparency in which these faithfully follow their intended use. This additionally yields as a side effect the possible reveal of hidden but quantifiable societal biases.
However, we should also be concerned about privacy. The data at hand contains information about the criminal past of thousands of people. If such data were to be de-anonymized and publicized, it could have very damaging social, emotional and financial consequences for the people involved (it could damage relationships, embarrassment or affect someone's career for example). For this reason, public release of a dataset like this should always be done as carefully as possible.

REFERENCES

[1] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. https://arxiv.org/abs/1810.01943
[2] Moritz Hardt, Eric Price, and Nathan Srebro. [n.d.]. *Equality of Opportunity in Supervised Learning.* arXiv:1610.02413 [cs] http://arxiv.org/abs/1610.02413
[3] Faisal Kamiran and Toon Calders. [n.d.]. Data Preprocessing Techniques for Classification without Discrimination. 33, 1 ([n. d.]), 1–33. https://doi.org/10.1007/s10115-011-0463-8
[4] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. [n.d.]. *On Fairness and Calibration.* arXiv:1709.02012 [cs, stat] http://arxiv.org/abs/1709.02012