Human-centered Machine Learning 2022 Assignment 2

Xinhao LAN - 1082620
Sili WANG - 4621514
x.lan1@students.uu.nl,s.wang13@students.uu.nl
Utrecht University
Utrecht, the Netherlands

1 INTRODUCTION

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

2 METHODS

In this experiment, we use several methods and dataset belows:

- **AIF360**: The AI Fairness 360 toolkit is an extensible open-source library containing techniques developed by the research community to help detect and mitigate bias in machine learning models throughout the AI application lifecycle.
- **COMPAS**: The COMPAS dataset consists of the results of a commercial algorithm called COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), used to assess a convicted criminal's likelihood of reoffending.
- **Reweighting**: This is a Pre-processing Algorithm among the bias mitigation strategies for machine learning models. This function aims to weight the examples in each (group, label) combination differently to ensure fairness before classification.[3]
- **EqOddsPostprocessing**: Different from the reweighting, this is a Post-processing Algorithm among the bias mitigation strategies for machine learning models. This function aims to solve a linear program to find probabilities with which to change output labels to optimize equalized odds. [2] [4]

3 RESULTS

3.1 Basic Statistics

The shape of the original dataset is [5278,12], which means there are 5278 instances and 12 values for each instance. The original dataset is split into a training dataset and a test dataset with the proportion of 80% and 20%. So, training dataset contains 4222 instaces while test dataset contains 1056 instances. Both of them contain 10 features among 12 values in each instance. (exclude the first column named **instance weights features** and the last column named **labels**). In conclusion, the size of the training set is **4222 × 10** and the size of the test set is **1056 × 10**. The mean difference of the training set is -0.140265 while the mean difference of the test set is -0.099688. To be more clear, the base rate of the unprivileged group of the test data set is about 0.4790, and the value is 0.5787 for the privileged group. The base rate of the unprivileged group of the training data set is about 0.4763, and the value is 0.6166 for the privileged group.

3.2 Training and Test Set

The training set contains 2532 African-American and 1690 Caucasians.The mean value of the recidivism for African-Americans within 2 years is about 0.52 and the standard deviation is about 0.50. The mean value of the recidivism for Caucasians within 2 years is about 0.38 and the standard deviation is about 0.49. (Table. 1)

race	count	mean	std
African-American	2532	0.523697	0.499537
Caucasian	1690	0.383432	0.486366

Table 1: Statistics of the training data based on the race

The training set contains 3404 males and 818 females.The mean value of the recidivism for males within 2 years is about 0.49 and the standard deviation is about 0.50. The mean value of the recidivism for Caucasians within 2 years is about 0.36 and the standard deviation is about 0.48. (Table. 2)

sex	count	mean	std
Male	3404	0.492656	0.500020
Female	818	0.363081	0.481182

Table 2: Statistics of the training data based on the sex

If we use both the sex and race to classify, among all the African-Americans, there are 2096 males and 436 females. The African-American males have a recidivism rate mean of 0.55 with standard deviation of 0.50, while the mean value of African-American female recidivism rate is about 0.38 with standard deviation of 0.48. Among all the Caucasians, there are 1308 males and 382 females. The Caucasian males have a recidivism rate mean of 0.39 with standard deviation of 0.49, while the mean value of African-American female recidivism rate is about 0.35 with standard deviation of 0.48 (Table. 3)

race	sex	count	mean	std
African-American	Male	2096	0.554389	0.497152
	Female	436	0.376147	0.484974
Caucasian	Male	1308	0.393731	0.488763
	Female	382	0.348168	0.477014

Table 3: Statistics of the training data based on both the sex and the race

In general, the recidivism rate of African-American is higher than Caucasian and the recidivism rate of male is higher than female. There are more males than females and more African-American than Caucasian in this training set. African-American males have the highest recidivism rates among all categories, followed by the Caucasian males. Caucasian females have the lowest average recidivism rates. The standard deviation of all categories are almost the same (between 0.48 and 0.50). In addition, we also do some analysis on the test set, the result is almost the same as the training set. Only difference is that African-American females have the lowest recidivism rates. (Table. 4, 5, 6)

race	count	mean	std
African-American	643	0.520995	0.499948
Caucasian	413	0.421308	0.494368

Table 4: Statistics of the test data based on the race

sex	count	mean	std
Male	843	0.513642	0.500111
Female	213	0.356808	0.480186

Table 5: Statistics of the test data based on the sex

race	sex	count	mean	std
African-American	Male	530	0.558491	0.497036
	Female	113	0.345133	0.477529
Caucasian	Male	313	0.437700	0.496898
	Female	100	0.370000	0.485237

Table 6: Statistics of the test data based on both the sex and the race

3.3 Classifiers

In this section, we trained four different classifiers on this dataset.

- **Classifier 1**: The first classifier is a logistic regression classifier making use of all features. The precision of the first classifier is about 0.67, the recall value is 0.61, the F1 score is 0.63 and the accuracy is 0.66. The statistical parity difference is -0.2977, which indicates that the criterion is quite far away from being satisfied. The equal opportunity criterion is about -0.2794, which is also unsatisfied.
- **Classifier 2**: The second classifier is a logistic regression classifier without the race feature. The precision of the second classifier is about 0.68, the recall value is 0.55, the F1 score is 0.61 and the accuracy is also 0.66. The statistical parity difference is -0.2128. This is also quite far away from being satisfied just like classifier 1. The equal opportunity criterion is about -0.1875, which is also unsatisfied.
- **Classifier 3**: The third classifier is a logistic regression classifier after reweighting instances in the training set. For the value of the weights of the train data is all equal to 1 (mean value is equal to 1 and the standard error is 0.0). After the reweighting, the mean value of the weight is still equal to 1, but the standard deviation is about 0.1399. The distribution of the weight value grouped by race after the reweighting is in Figure. 1, Figure. 2 and Table. 7.

race	count	mean	std	min	max
African-American	2532	1.0	0.103935	0.899772	1.107737
Caucasian	1690	1.0	0.162305	0.871051	1.204169

Table 7: Weight value distribution after the reweighting

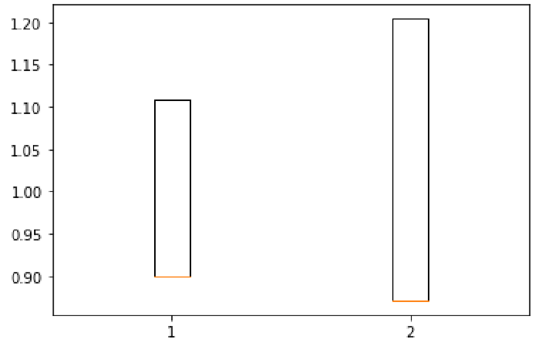

Figure 1: Box plot of weight value after the reweighting

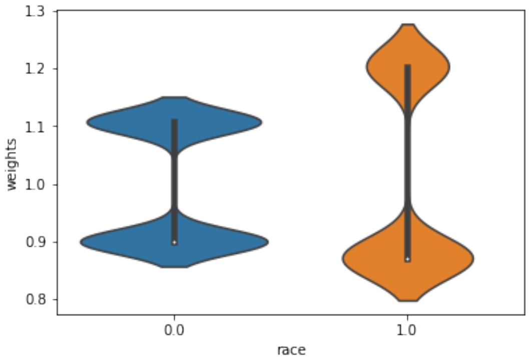

Figure 2: Violin plot of weight value after the reweighting

If we make use of all features to train a logistic regression classifier after the reweighting, the precision of this classifier is about 0.65, the recall value is 0.61, the F1 score is 0.63 and the accuracy is 0.65. The statistical parity difference is -0.0451, which means the criterion is almost satisfied. The equal opportunity criterion is about -0.0095, which is also almost satisfied. If we build the same type classifier without the race feature, the precision of this classifier is about 0.68, the recall value is 0.59, the F1 score is 0.63 and the accuracy is 0.67. The statistical parity difference is equal to -0.2223 and the equal opportunity criterion is about -0.1954, which are the same as the first two classifiers.

- **Classifier 4**: The fourth classifier is a logistic regression classifier after post-processing. The precision of this classifier is 0.64, the recall value is 0.53, the F1 score is 0.58 and the accuracy is 0.63. The mean difference of this classifier is equal to -0.0265 and the equal opportunity criterion is about 0.1639.

In conclusion, based on all statistics in the table below (Table. 8), after the pre-processing and post-processing technique like reweighting and equal odds, the statistical parity difference has improved a lot and is almost satisfied. However, the precision, recall and F1 score of these conditions are the lowest among all the classifier. In addition, if we use the reweighting technique and do not use the race feature to classify, it's almost the same as the classifier 2 (without race feature).

Classifier	Details	Precision	Recall	F1	Accuracy	SPD	TPR_D	TPR_P	TPR_U
1	All features	0.6624	0.6051	0.6324	0.6610	-0.2977	-0.2794	0.8703	0.5909
2	Without race	0.6828	0.5540	0.6117	0.6610	-0.2128	-0.1875	0.8661	0.6786
3	Reweighting	0.6451	0.6071	0.6255	0.6496	-0.0451	-0.0095	0.6946	0.6851
4	Reweighting without race	0.6842	0.5874	0.6321	0.6705	-0.2223	-0.1954	0.8577	0.6623
5	Postprocessing	0.6437	0.5324	0.5828	0.6326	-0.0265	0.1639	0.8361	1.0000

Table 8: Weight value distribution after the reweighting

3.4 Discussion

3.4.1 Result discussion. As we mentioned before, It's clear that classifiers with post-processing method like EqOdds and pre-processing method like Reweighting perform much better than other classifiers on SPD. However, for these two classifiers, their performance on precision, recall, F1 and accuracy become worse. If we use the features without race factor, the number fluctuates but it remain close to the number belongs to the classifier with all features. In addition, we do a more detailed experiment which use reweighting technique and do not use race factor, It shows that this classifier doesn't have the same advantages like classifier with reweighting technique. It is more like original classifier.

3.4.2 Ethical discussion. This paper raises some very serious ethical concerns about the use of a machine learning or deep learning system to predict the result with a dataset containing some sensitive attributes like gender. It shows that if we use a dataset that lacks fairness for training without any operations, it is very likely that the bias in the dataset will be enlarged, thereby ignoring individual differences and making some wrong judgments. We should also be careful with the privacy issues. Many people who are included in this dataset have information about the crime. If this information is released without anonymization, it will definitely have a certain social impact on the people involved in the data set.

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 Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[3] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems* 33, 1 (2012), 1–33.

[4] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. *Advances in neural information processing systems* 30 (2017).