Fairness in Machine Learning and the Evaluation of Calibration Methods

Jaspers W. Huanay
jaspers.huanay@skoltech.ru

June 28, 2020

Справедливость в машинном обучении и оценка методов калибровки

Хасперс В. Гуанай
jaspers.huanay@skoltech.ru

28 июня 2020 г.

Contents

1 Motivations

Fairness is becoming an important topic in machine learning (ML) in the recent years. But why do we care about fairness? Many aspects of our lives such as happiness, success and well being can be profoundly affected by the decision of others through automated systems, for example whether one is admitted or not to certain school, get the job position, loan, products recommended to us etc. This systems are being used in our day to day lives and is going to be even more used in the near future as more areas are integrating ML in their processes or products. Thus fairness is highly related to our own benefits.

Even though ML systems are good tools that can benefit us in many ways, incorrectly used can rise serious concerns because it might limit our ability to achieve our goals and access to opportunities we are as qualified as the individuals from privileged groups. A ML model that learned from unbiased data set may systematically discriminate or hurt certain groups, usually minorities or historically disadvantageous groups.
An example of such unfairness can be seen in Xing a popular job search portal in Germany (similar to LinkedIn), it was found that less qualified male candidates were ranked higher than more qualified female candidates [1].

So how do we ensure that decision made by ML models are right? This is the question that we are going to try to answer in the following sections, by measuring the bias and trying to calibrate to mitigate the unfairness.
The structure of this work is as follow, in section 2 we will see causes and definition of biases, in section 3 we will mention the algorithms to mitigate the unfairness and in section 4 we will run an experiment comparing these algorithms on a real dataset.

2 Fairness

2.1 Causes of Bias

Essentially bias comes from human bias which is encoded in the dataset due to contextual and historical reasons. Here we list some of this reasons [2]

2.1.1 Limited features

The features collected for minorities might be less informative than those of their counterparts from majority group, in this case the prediction accuracy of the model is much lower for the minority group.

2.1.2 Skewed dataset

The dataset distribution is called skewed if one tail is longer than the another. These distributions are sometimes called asymmetric or asymmetrical distributions as they don't show any kind of symmetry. For example, household income in the U.S. is negatively skewed with a very long left tail. This means that people that are in the 90 or less percentile have much lower income than those in the rest 10 percentile.

2.1.3 Sample size disparity

If data collected about the minority group is much less than the those from the majority group, then the estimates of the model for the smaller group are significantly worse than for the larger. Therefore there is a general tendency of the automated system to favor those who belong to the statistically dominant group.

2.1.4 Cultural differences

The negative effects of sample size disparity is greatly exacerbated by cultural differences. Suppose a social network attempted to classify user names into 'real' and 'fake'. Names from the large population (let's say white Americans) are pretty straightforward to deal with compared with ethnic names. In some ethnic groups, names tend to be far more diverse and uncommon. The statistical patterns applied to the majority might be invalid for the minority group.

2.1.5 Proxies

Even when we don't take into account the sensitive feature for training the ML model, there might be other features that are proxies of the sensitive feature (entangled). Thus the model will still be biased. Sometimes, it is very hard to determine if a relevant feature is too correlated with protected or sensitive features and if we should include it in training or not. There are works on the direction to learn disentangled representation of the features in the latent space which aims to separate the sensitive feature from the neutral features [3, 4].

2.2 Metrics

There are many definitions of fairness proposed [5]. however, many of these definitions are inadequate in a research setting.
One of the simplest conceptions of fairness is the notion of ***demographic parity***. Demographic parity requires that, for all groups of the protected attribute A (e.g. gender), should receive the positive outcomes at equal rates, and the protected attribute should be independent of the prediction [6].
Alternative notions of group fairness have been defined. Hardt et al. [7] proposed ***equal odds*** which requires that the rates of true positives and false positives be the same across groups. This punishes classifiers which perform well only on specific groups. Hardt et al. [7] also proposed a relaxed version of equal odds called ***equal opportunity*** which demands only the equality of true positive rates. Other definitions of group fairness include ***calibration*** [8, 9].

2.2.1 Statistical parity

Also called demographic parity, the measure is based on the following formula.

$$Pr(Y = 1|D = unprivileged) - Pr(Y = 1|D = privileged) \approx 0 \qquad (1)$$

The difference between the probability that a random individual drawn from the unprivileged group is labeled 1 and the probability that a random individual sampled from the privileged group is labeled 1, should be close to 0 so it will be fair.

2.2.2 Average absolute odds difference

This measure is using both true positive rate and false positive rate to calculate the bias. It calculates the equality of odds with the following formula:

$$\frac{1}{2}[|FPR_{D=unprivileged} - FPR_{D=privileged}| + |TPR_{D=unprivileged} - TPR_{D=privileged}|] \approx 0 \tag{2}$$

Similar to statistical parity equation 2 should be 0 or close to 0 to be fair.

2.2.3 Equal opportunity

This metric is just a difference between the true positive rate of unprivileged group and the true positive rate of privileged group so it follows this formula:

$$TPR_{D=unprivileged} - TPR_{D=privileged} \approx 0 \tag{3}$$

Same as the previous formula we need it to be close to 0

2.2.4 Disparate impact

This measure is similar to the *Statistical parity* equation, but in this case it's a ratio between the probability of a random individual drawn from the privileged and unprivileged with a label of 1.

$$\frac{Pr(Y = 1|D = unprivileged)}{Pr(Y = 1|D = privileged)} = 1 \tag{4}$$

In this case the ratio should be 1.

2.2.5 Theil index

The Theil index is a statistic primarily used to measure economic inequality [10] and other economic phenomena, though it has also been used to measure racial segregation. The Theil index T_T is the same as redundancy in information theory which is the maximum possible entropy of the data minus the observed entropy. It is a special case of the generalized entropy index. It can be viewed as a measure of redundancy, lack of diversity, isolation, segregation, inequality, non-randomness, and compressibility. It was proposed by econometrician Henri Theil at the Erasmus University Rotterdam.

$$T_T = \frac{1}{n}\sum_{i=0}^{n}\frac{x_i}{\mu} ln \frac{x_i}{\mu}$$

Where $x_i = \hat{y}_i - y_i + 1$
T_T needs to be close to 0 to be fair.

3 Algorithms

There are several algorithms that we can use to improve fairness. These algorithms fall into three categories: pre-processing, optimization at training time, and post-processing.

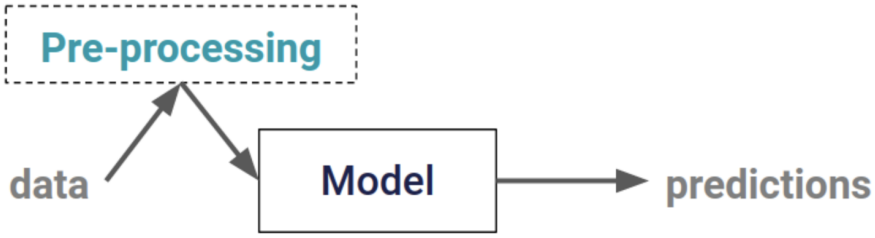

Figure 1: Pre-processing method

3.1 Pre-processing

Pre-processing algorithms deals with input data (fig 1), in this work we are going to consider the following:

3.1.1 Learning latent representation

Here the idea is to learn a new representation Z such that it removes the information correlated to the sensitive attribute A and preserves the information of X as much as possible [11], then this "cleaned" data can be used for classification, regression, etc. and produce results that preserves the demographic parity and individual fairness.

Given a set of inputs X, the mapping from X to Z can be done via softmax:

$$P(Z = k|x) = \exp(-d(x, v_k))/\sum_{j=1}^{K} \exp(-d(x, v_k)$$

The model is thus defined as a discriminative clustering model.

3.1.2 Reweighing

Is a technique that weights the examples in each combination (of groups) differently to ensure fairness before classification [12].

To compensate for the bias, authors assign lower weights to privileged groups.

$$W(X) := \frac{P_{exp}(S = X(S) \wedge Class = X(Class))}{P_{obs}(S = X(S) \wedge Class = X(Class))}$$

The weight of an object will be the expected probability to see an instance with its sensitive attribute value and class given independence, divided by its observed probability. In this way it's assigned a weight to every tuple according to its S and Class-values. The dataset D with the added weights is called, D_W, it is easy to see that D_W is unbiased.

3.2 In-processing

The intuitive idea here is to add constraints or regularization term to the optimization objective (fig 2).

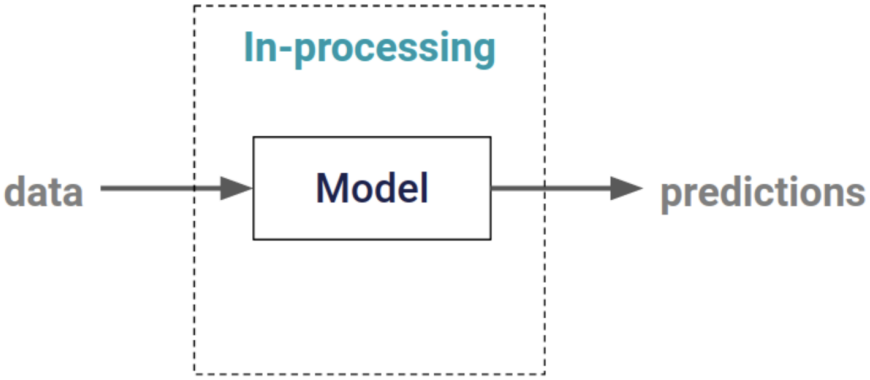

Figure 2: In-processing method

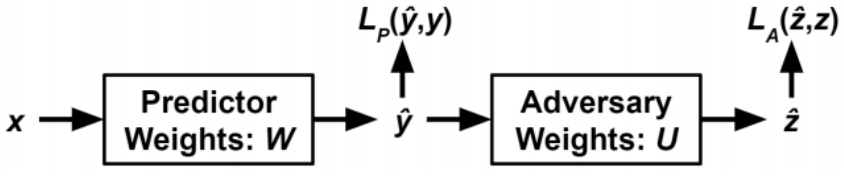

Figure 3: Architecture of adversarial network

3.2.1 Adversarial debiasing

Adversarial debiasing (fig 3) is a technique that learns a classifier to maximize prediction accuracy and simultaneously reduce an adversary's ability to determine the protected attribute from the predictions. This approach leads to a fair classifier as the predictions cannot carry any group discrimination information that the adversary can exploit [13].

3.2.2 Prejudice remover

A technique that adds a discrimination aware regularization term to the learning objective.

Kamishima et al. [14] adopted two types of regularizers. The first regularizer is a standard one to avoid over-fitting. The author used an L_2 regularizer $\|\theta\|_2^2$. The second regularizer, $R(D,\theta)$, is introduced to enforce fair classification. Kamishima designed this regularizer to be easy to implement and to require only modest computational resources. By adding these two regularizers, the objective function to minimize is as follows:

$$\mathcal{L}(D,\theta) + \eta R(D,\theta) + \frac{\lambda}{2}\|\theta\|_2^2 \tag{5}$$

Where η and λ are positive regularization parameters.

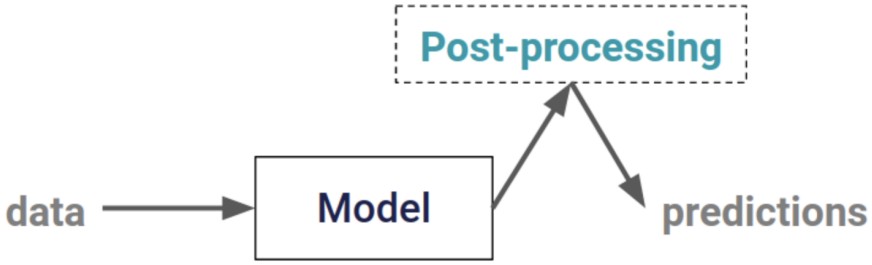

Figure 4: Post-processing method

3.3 Post-processing

These methods tries to modify the posterior (fig 4) in such a way that satisfies the fairness constraints.

3.3.1 Calibrated equality of odds

Calibrated equality of odds is a technique that optimizes over calibrated classifier score outputs to find probabilities with which to change output labels with an equalized odds objective [15]

Given a cost function gt, classifiers h_1 and h_2 achieve Equalized Odds with Calibration for groups G_1 and G_2 if both classifiers are calibrated and satisfy the constraint $g_1(h_1) = g_2(h_2)$

3.3.2 Reject option classification

The idea behind this algorithm is that it gives favorable outcomes to unpriviliged groups and unfavorable outcomes to priviliged groups in a confidence band around the decision boundary with the highest uncertainty.

Traditionally, a learned classifier assigns an instance to the class with the highest posterior probability. The first solution deviates from this traditional decision rule and gives the idea of a critical region in which instances belonging to deprived and favored groups are labeled with desirable and undesirable labels, respectively [16]

4 Experiments

We have seen the metrics to measure the bias, also the calibration methods applied at differents stages of the ML pipeline (pre-processing, in-processing, and post-processing), in sections 2 and 3 respectively.

To evaluate the performance of these calibration methods, we are going to use the **Homicide Reports, 1980-2014** dataset. The data was compiled and made available by the Murder Accountability Project, founded by Thomas Hargrove.

The Murder Accountability Project is the most complete database of homicides in the United States currently available. This dataset includes murders

from the FBI's Supplementary Homicide Report from 1976 to the present and Freedom of Information Act data on more than 22,000 homicides that were not reported to the Justice Department. This dataset includes the age, race, sex, ethnicity of victims and perpetrators, in addition to the relationship between the victim and perpetrator and weapon used.

4.1 Data description

Dataset contains 24 features and 638454 observations.

Crime Type	Victim Sex	Victim Race	Perpetrator Sex	Perpetrator Race
1	1	3	1	3
1	1	4	1	4
1	1	4	1	4
1	1	4	1	4
1	0	3	1	1
1	0	4	1	4
1	1	4	1	2
1	1	4	1	4
1	0	1	1	1
1	1	3	1	4

Table 1: Sample of the dataset for important features for the study of fairness

Table 1 shows a sample of the most important features for the fairness study. Where **Perpetrator Sex**=$\{0 : female, 1 : male\}$, **Perpetrator Race**=$\{0 : Asian/PacificIslander, 1 : Black, 2 : Missingvalue,$
$3 : NativeAmerican/AlaskaNative, 4 : White\}$, the same encoding applies for **Victim Sex** and **Victim Race** features.

A simple analysis of the **Perpetrator features(Sex, Race, Age)** (table 2) shows that the crimes are committed in 26% of the cases by a white adult man.

Looking at the **Victim Race feature** in fig 5, we can observe that most of the perpetrators are males, attacking mostly to white rather than black people. Likewise we can see that **white victims** are attacked in the majority of cases by **white perpetrators**, and **black victims** are attacked mostly by **black perpetrators**.

4.2 Bias measure

We are going to consider Perpetrator Sex and the Perpetrator Race as our target variables.

And I will check the bias using the metrics mentioned in the section 2.2, using aif360 python package provided by IBM [17].

For 3 out of these 5 metrics (Equal Opportunity, Average Absolute Odds Difference, and Theil Index) we need the prediction. And for Statistical Parity and Disparate Impact, we do not require the predictions.

Perpetrator Sex	Perpetrator Race	Perpetrator Age category	Frequency
Male	White	Adult	0.264265
	Black	Young	0.220366
		Adult	0.197418
	White	Young	0.159222
Female	Black	Adult	0.035841
	White	Adult	0.034741
	Black	Young	0.018415
	White	Young	0.013767
Male	White	Elder	0.013439
	Missing value	Young	0.008546
	Asian/Pacific Islander	Adult	0.007053
	Black	Elder	0.005489
	Asian/Pacific Islander	Young	0.004820
	Native American/Alaska Native	Adult	0.003795
	Missing value	Adult	0.003644
	Native American/Alaska Native	Young	0.002892
Female	White	Elder	0.001339
	Asian/Pacific Islander	Adult	0.000935
	Native American/Alaska Native	Adult	0.000870
	Black	Elder	0.000734
	Missing value	Young	0.000565
	Native American/Alaska Native	Young	0.000413
	Asian/Pacific Islander	Young	0.000337
	Missing value	Adult	0.000330
Missing value	Missing value	Young	0.000328
Male	Asian/Pacific Islander	Elder	0.000286
	Missing value	Elder	0.000080
	Native American/Alaska Native	Elder	0.000045
Female	Asian/Pacific Islander	Elder	0.000016
	Native American/Alaska Native	Elder	0.000007
	Missing value	Elder	0.000004

Table 2: Demographic distribution of the perpetrator

So, we first need to build a classifier $f : X \to Y$, where Y is the perpetrator sex/race, for the metrics that requires the prediction. For this step we can work with any classifier such as Logistic regression, Random forest, etc. We will choose Random Forest because of the categorical features. Performance of the model achieved is 92.8% and the ROC curve is shown in fig 6, but we are more interested in how to prevent or mitigate the bias, so we do not pay much attention to the accuracy even though is a good accuracy.

To measure the existing bias in the dataset, We considered the Victim Sex and Race features, and got the scores show in table 3.

The objective values are what the metrics should be or at least it should be close to it. We can see that metric scores for the $VictimRace$ variable are close

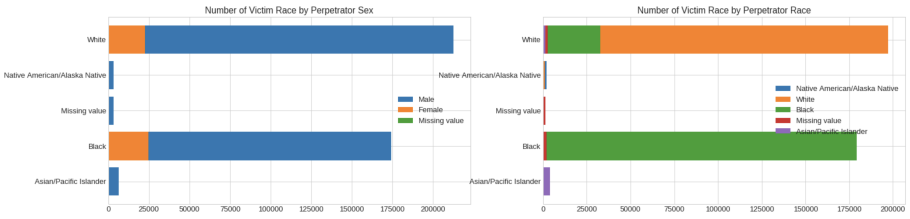

Figure 5: *left*: **Victim Race** cases by **Perpetrator Sex**, *right*: **Victim Race** cases by **Perpetrator Race**

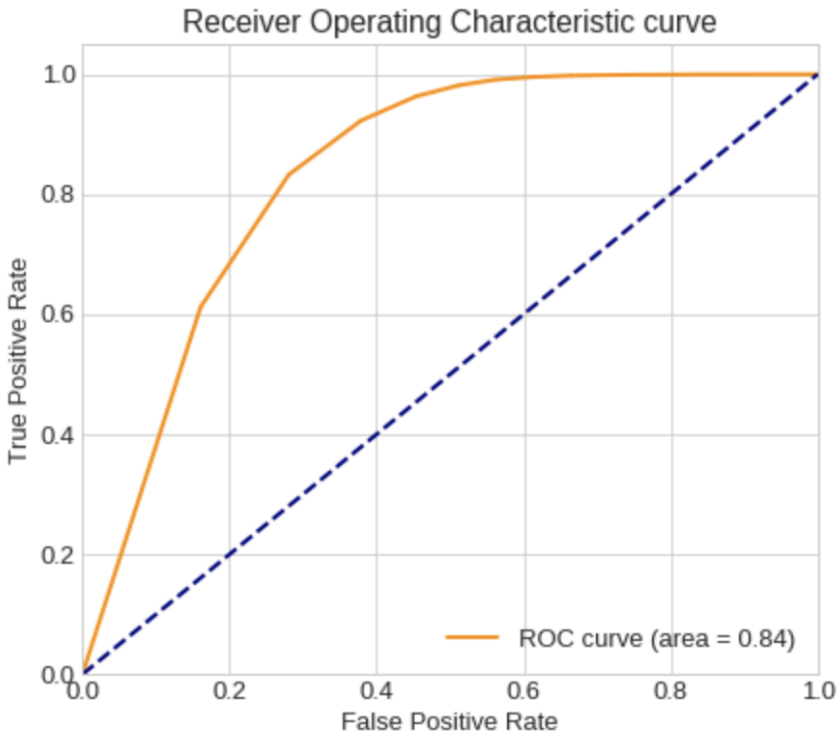

Figure 6: Random Forest Roc curve

	Metric 1	Metric 2	Metric 3	Metric 4	Metric 5
objective	0.000000	0.000000	0.000000	1.000000	0.000000
Victim Sex	0.019849	-0.012653	0.123911	1.021454	0.034822
Victim Race	0.003852	-0.007360	0.009765	1.004093	0.034822

Table 3: M1: Statistical parity, M2: Equal opportunity, M3: Average absolute odds difference, M4: Disparate impact, M5: Theil index

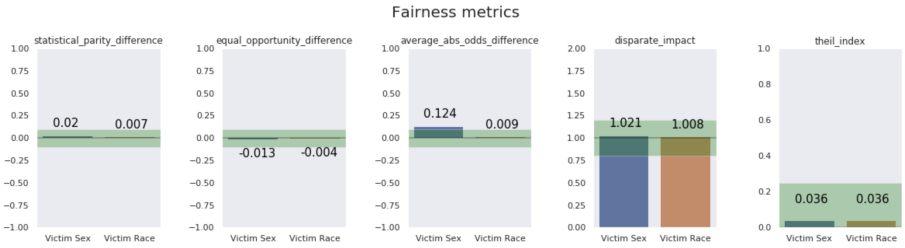

Figure 7: Fairness metric scores for privileged variables

to the objective, therefore this variable is not biased. However *Average absolute odds difference* metric for $VictimSex$ is far from the objective value, thus this variable is biased because is enough that one of this metrics show bias to say that the variable is biased. These scores can be seen better in fig 7, where we can see that *Average absolute odds difference* is outside the threshold.

In the next section we will run the algorithms described in section 3, to prevent this bias.

4.3 Calibration

As we discussed in section 3, calibration can be done in one or more stages of the ML pipeline.

Pre-processing algorithms to debias the dataset.

After performing **Learning latent representation** and measuring again with the metrics we get the scores seen in table 4, since metric *Disparate impact* scores are far from the objective, we can say that the feature representation in latent space show that both variables (Victim Sex and Race) are biased. Thus this method does not help to reduce it.

	Metric 1	Metric 2	Metric 3	Metric 4	Metric 5
objective	0.000000	0.0	0.0	1.000000	0.0
Victim Sex	0.001807	0.0	0.0	1.584826	0.0
Victim Race	-0.001346	0.0	0.0	0.675605	0.0

Table 4: M1: Statistical parity, M2: Equal opportunity, M3: Average absolute odds difference, M4: Disparate impact, M5: Theil index - *Learning latent representation*

Since the previous method did not help prevent bias, we perform **Reweigh-**

ing algorithm and the scores obtained are show in table 5, here we observe that there is pretty much no difference with the score values of the original data.

	Metric 1	Metric 2	Metric 3	Metric 4	Metric 5
objective	0.000000	0.000000	0.000000	1.000000	0.000000
Victim Sex	0.020014	-0.012910	0.126369	1.021628	0.035157
Victim Race	0.001232	-0.010272	0.013616	1.001308	0.035157

Table 5: M1: Statistical parity, M2: Equal opportunity, M3: Average absolute odds difference, M4: Disparate impact, M5: Theil index - *Reweighing*

Next we will attempt to calibrate during training.

In-processing algorithms to calibrate the model.
After performing **Adversarial debiasing** we get the following metric scores (table 6), most of the metrics are similar to the objective, but *Theil index* huge difference indicates that regularization added increase unfairness.

	Metric 1	Metric 2	Metric 3	Metric 4	Metric 5
objective	0.0	0.0	0.0	1.0	0.000000
Victim Sex	0.0	0.0	0.0	NaN	2.223135
Victim Race	0.0	0.0	0.0	NaN	2.223135

Table 6: M1: Statistical parity, M2: Equal opportunity, M3: Average absolute odds difference, M4: Disparate impact, M5: Theil index - *Adversarial debiasing*

Let's see the scores after performing **Prejudice remover** (table 7), this method **helps** debiasing the dataset for the privileged features.

	Metric 1	Metric 2	Metric 3	Metric 4	Metric 5
objective	0.0	0.0	0.0	1.0	0.000000
Victim Sex	0.0	0.0	0.0	1.0	0.032631
Victim Race	0.0	0.0	0.0	1.0	0.032631

Table 7: M1: Statistical parity, M2: Equal opportunity, M3: Average absolute odds difference, M4: Disparate impact, M5: Theil index - *Prejudice remover*

Let's see the calibration after training.

Post-processing algorithms to calibrate outputs of the model.
Calibrated equality of odds, one of the post-processing algorithm, scores (table 8) is not much different than the original metric values.

Reject option classification scores on the other hand are close to the the objective values, meaning that similar to *Prejudice remover* helps to achieve a fair model.

	Metric 1	Metric 2	Metric 3	Metric 4	Metric 5
objective	0.000000	0.000000	0.000000	1.000000	0.000000
Victim Sex	0.019663	-0.012576	0.123504	1.021235	0.035653
Victim Race	0.007094	-0.003700	0.008572	1.007535	0.035653

Table 8: M1: Statistical parity, M2: Equal opportunity, M3: Average absolute odds difference, M4: Disparate impact, M5: Theil index - *Calibrated equality of odds*

	Metric 1	Metric 2	Metric 3	Metric 4	Metric 5
objective	0.000000	0.000000	0.000000	1.000000	0.00000
Victim Sex	0.021230	-0.000034	0.093898	1.021722	0.02882
Victim Race	0.011516	0.000188	0.052711	1.011706	0.02882

Table 9: M1: Statistical parity, M2: Equal opportunity, M3: Average absolute odds difference, M4: Disparate impact, M5: Theil index - *Reject option classification*

4.4 Results

Model performace after calibration with the algorithms described is show in fig 8, we can see that **Reject option** is the one the outperforms the others.

The accuracies and F1 scores are shown in table 10, we can observe that Reject option and Prejudice remover accuracies are lower than the original model as expected (after calibration accuracy score decreases), but still is a good accuracy, and the F1 scores are really close to the original value, this can show that after calibration models are "more robust".

	Accuracy	F1 Score
Origin	0.928183	0.960594
LFR	0.367269	0.481799
Reweighing	0.929179	0.961131
PrejudiceRemover	0.891731	0.942767
CalibratedEqOdds	0.729718	0.830829
RejectOption	0.908622	0.951248

Table 10: Accuracy of RF for calibration methods described

The bias measure of these algorithms can also be seen in fig 9, where we can observe that Reject option is close to 0 as it should be.

It is worth to mention that, in the table 10 we do not see algorithm **Adversarial debiasing**, this because its metrics scores are very far from the acceptable threshold to measure bias (fig 10). This means that the network did not learnt to disentangle the sensitive from the neutral features.

From the results above we can see that *Reject Option*(Post-processing) and

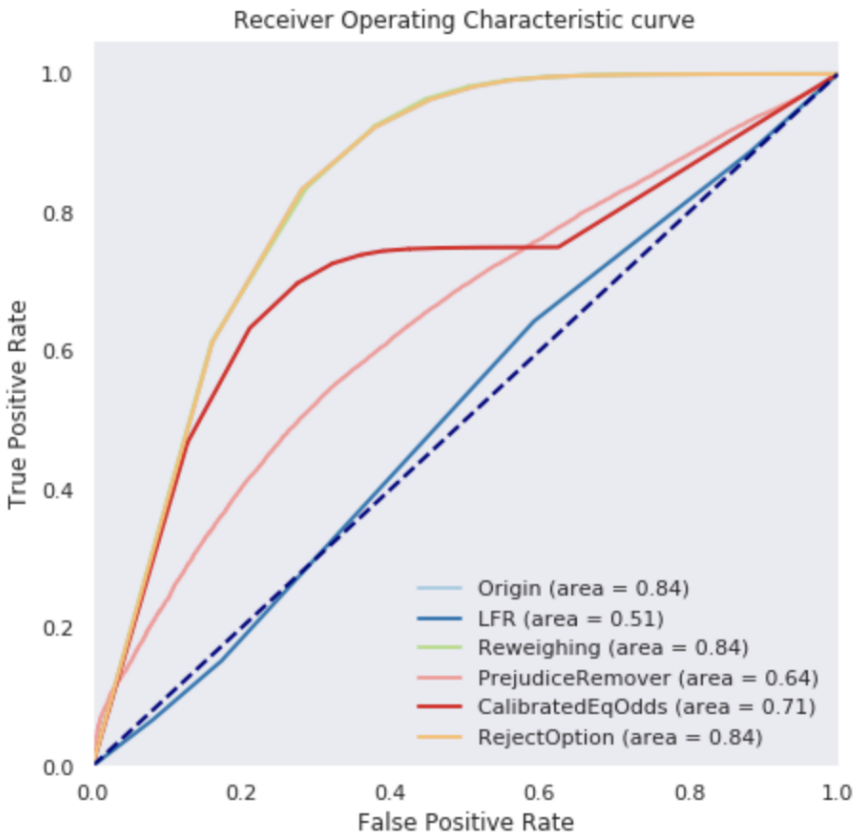

Figure 8: Roc curve for RF for several calibration methods

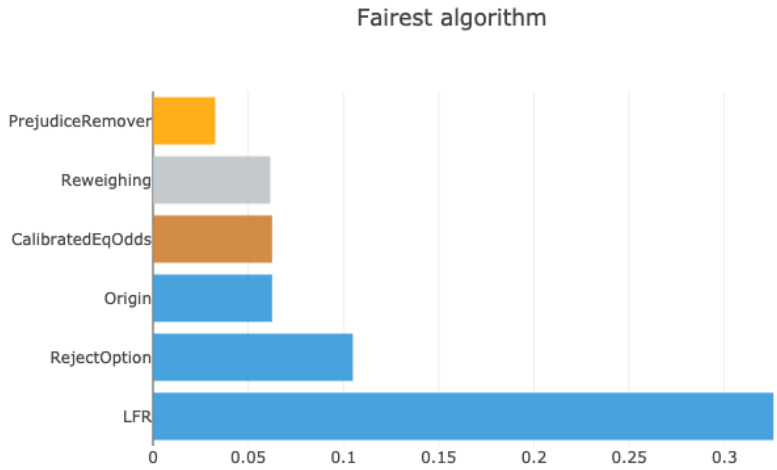

Figure 9: Fairest algorithm

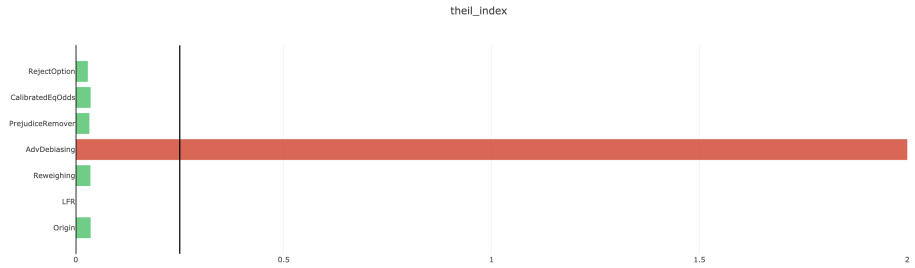

Figure 10: Adversarial debiasing method score for Theil index

Prejudice Remover(In-processing) help with the mitigation of bias in each stage independently of each other. Next we are going to evaluate the performance in preventing bias after performing both of these methods together (see table 11).

	Metric 1	Metric 2	Metric 3	Metric 4	Metric 5
objective	0.000000	0.0	0.0	1.000000	0.0
Victim Sex	0.016283	0.0	0.0	1.018342	0.0
Victim Race	0.020244	0.0	0.0	1.022392	0.0

Table 11: M1: Statistical parity, M2: Equal opportunity, M3: Average absolute odds difference, M4: Disparate impact, M5: Theil index - *Prejudice Remover + Reject option classification*

We can see that if we use a calibrated model and calibrate the output we get a fair result, according to the metrics scores is closer to the objective than the metrics of the individual calibration. This result would suggest that to minimize the bias, we should calibrate at different stages of the ML pipeline with the appropriate algorithm because it might behave differently if we choose another target variable (see appendix).
All code experiments described here can be found here `https://github.com/jwilliamn/calibration-methods-for-fairness.git`

5 Conclusion

We have measured existence of bias, and evaluated several calibration methods to reduce it. We have seen that for the chosen dataset and target variable, calibration methods that deal with the data (pre-processing) does not mitigate the bias. However *Prejudice Remover*(In-processing) and *Reject Option*(Post-processing) methods helped to achieve a fair system.
We have also seen that if we calibrate the outputs of an already calibrated model, we achieve a slightly fairer system, which in a real scenario would have a positive impact in the unprivileged groups.

6 Future Work

Bias mitigation is a challenging task as we saw in this work, due to the fact that bias is deeply encoded in the dataset and sometimes models built under this datasets might amplify this bias and eventually will cause harm in real life applications. So, It would be reasonable to focus on the dataset itself before feeding to any model. One way to have an unbiased dataset would be learning disentangled representations [3], which capture information about different generative factors in different latent dimensions, this would limit the model to depend only on latent dimension corresponding to neutral attributes and not to the one corresponding to sensitive attributes. I will extend this work on this topic as a future work.

References

[1] P. Lahoti, K. P. Gummadi, and G. Weikum, "ifair: Learning individually fair data representations for algorithmic decision making," in *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 2019, pp. 1334–1345.

[2] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. fairmlbook.org, 2019, http://www.fairmlbook.org.

[3] F. Locatello, G. Abbati, T. Rainforth, S. Bauer, B. Schölkopf, and O. Bachem, "On the fairness of disentangled representations," in *Advances in Neural Information Processing Systems*, 2019, pp. 14 611–14 624.

[4] M. H. Sarhan, N. Navab, A. Eslami, and S. Albarqouni, "Fairness by learning orthogonal disentangled representations," 2020.

[5] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, "A comparative study of fairness-enhancing interventions in machine learning," 2018.

[6] J. Gardner, C. Brooks, and R. Baker, "Evaluating the fairness of predictive student models through slicing analysis," in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 2019, pp. 225–234.

[7] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in neural information processing systems*, 2016, pp. 3315–3323.

[8] A. Flores, K. Bechtel, and C. Lowenkamp, "False positives, false negatives, and false analyses: A rejoinder to "machine bias: There's software used across the country to predict future criminals. and it's biased against blacks."," *Federal probation*, vol. 80, 09 2016.

[9] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint arXiv:1609.05807*, 2016.

[10] P. Conceição and P. M. Ferreira, "The young person's guide to the theil index: Suggesting intuitive interpretations and exploring analytical applications," *Labor: Supply Demand*, 2000.

[11] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International Conference on Machine Learning*, 2013, pp. 325–333.

[12] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1–33, 2012.

[13] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 335–340.

[14] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2012, pp. 35–50.

[15] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," in *Advances in Neural Information Processing Systems*, 2017, pp. 5680–5689.

[16] F. Kamiran, A. Karim, and X. Zhang, "Decision theory for discrimination-aware classification," in *2012 IEEE 12th International Conference on Data Mining*, 2012, pp. 924–929.

[17] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, "AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," Oct. 2018. [Online]. Available: https://arxiv.org/abs/1810.01943

[18] F. Kamiran, A. Karim, and X. Zhang, "Decision theory for discrimination-aware classification," in *2012 IEEE 12th International Conference on Data Mining*. IEEE, 2012, pp. 924–929.

Appendices

A Additional Results

Here we add the results of calibration methods, which reduced the bias for Y: *Perpetrator Sex*(shown in section 4.4), for another target variable Y: **Perpetrator Race**.

Calibration algorithm **Prejudice Remover** (see table 12), whose scores values were close to the objective values in section 4.4, for the new target variable Y however these values are far from the objective. Specifically we can see that *Statistical parity*, *Equal opportunity*, *Average absolute odds difference*, and *Disparate impact* values are the opposite of what they should be. This shows the existence of a strong bias against **Perpetrator Race** when taking into account the *Victim Race*, which for these experiment we set **white** as the privileged group.

	Metric 1	Metric 2	Metric 3	Metric 4	Metric 5
objective	0.000000	0.000000	0.000000	1.000000	0.000000
Victim Sex	0.122173	0.069129	0.049244	1.246362	0.076892
Victim Race	-0.999533	-1.000000	0.999683	0.000467	0.076892

Table 12: M1: Statistical parity, M2: Equal opportunity, M3: Average absolute odds difference, M4: Disparate impact, M5: Theil index - *Prejudice Remover*

Calibration algorithm **Reject Option**, another "good algorithm" for target variable discussed in section 4.4, does not help preventing bias when choosing target variable Y:**Perpetrator Race**. Its scores (seen in table 13), in contrast to the previous algorithm, are not far from the objective values, but they are outside the threshold thus does not help in preventing the bias.

	Metric 1	Metric 2	Metric 3	Metric 4	Metric 5
objective	0.000000	0.000000	0.000000	1.000000	0.00000
Victim Sex	0.046626	0.015067	0.012381	1.063049	0.06752
Victim Race	-0.140512	-0.083211	0.119001	0.859271	0.06752

Table 13: M1: Statistical parity, M2: Equal opportunity, M3: Average absolute odds difference, M4: Disparate impact, M5: Theil index - *Reject Option*