Bias mitigation with AIF360: A comparative study

Tor H. Aasheim Knut T. Hufthammer Sølve Ånneland
Håvard Brynjulfsen Marija Slavkovik

University of Bergen

Abstract

The use of artificial intelligence for decision making raises concerns about the societal impact of such systems. Traditionally, the product of a human decision-maker are governed by laws and human values. Decision-making is now being guided - or in some cases, replaced by machine learning classification which may reinforce and introduce bias. Algorithmic bias mitigation is explored as an approach to avoid this, however it does come at a cost: efficiency and accuracy. We conduct an empirical analysis of two off-the-shelf bias mitigation techniques from the AIF360 toolkit on a binary classification task. Our preliminary results indicate that bias mitigation is a feasible approach to ensuring group fairness.

1 Introduction

Ethical and social implications of artificial intelligence (AI) have been hotly debated in recent years [4, 5, 10]. The European Union has adopted an "ethics by design" approach, incorporating ethical principles at the very start of the design of AI solutions [18]. In China, the recently proposed Beijing AI principles [2] aim to conform to human values, ethics, and autonomy for governance, use and healthy development of AI [13]. In 2020, the U.S. government outlined 10 principles [23] to regulate and promote trustworthy AI in the private sector, for the purpose of making it more fair, transparent, and safe [9]. In a newly released national strategy for AI, Norway has adopted the seven principles for ethical and responsible AI proposed by the EU. The national strategy applies to both the public and private sectors and call for the development of fair, transparent, safe, accountable, and ethical AI [21]. These interventions to regulate and facilitate safe and trustworthy AI is a sign of how disruptive AI tools can be without the right oversight.

AI is increasingly used to automate decision-making in many domains, such as job recruitment, health care, and even student evaluation[1]. A particular concern arises when supervised machine learning is used for this task, where correlations are found, for example, between high performance of the student (favourable label) and features that describe that student. Certain features,

This paper was presented at the NIK-2020 conference; see http://www.nik.no/.

[1] https://hbr.org/2020/08/what-happens-when-ai-is-used-to-set-grades

such as race, gender, and parent's affluence, should not be used to make a decision about a student's grades. However since these features correlate within non-protected features and due to the lack of transparency in some supervised machine learning methods, automated decision-making can lead to unfair decisions.

Fairness in machine learning concerns the proportion of favourable labels allocated to members of a privileged group compared to favorable labels allocated to members of an unprivileged group [8]. A favourable label is a label whose value is proffered: getting high grade, being offered a job etc. [3]. An unfair label assignment occurs when there's a disproportional amount of desirable labels distributed among the unprivileged and privileged groups. To ensure that an automated decision is fair, **bias mitigation** has been explored for numerous supervised machine learning approaches [3].

A **bias mitigation algorithm** is a procedure for reducing unwanted bias in training data or machine learning models [3]. Bias mitigation methods comprise of several data pre-, in-, and post-processing techniques that seek to ensure fair labels. Since bias mitigation methods necessarily interact with prediction modeling and model training, inevitably they impact the accuracy and efficiency of supervised machine learning. If the "price" of fair automated decision-making is "too high", these methods are unlikely to be used. We consider a recent toolkit of bias mitigation algorithms, AIF360[2] and explore to which extend do (some of) these algorithms interact with the classification accuracy.

Motivated by recent findings of algorithmic bias in hiring and recidivism assessment [16, 6], we compare the efficacy of applying two bias mitigation algorithms from AIF360 on a hiring prediction model which we trained on a U.S. census dataset [14]. Specifically, first we measure the mean difference in outcome (employed/unemployed) between all the groups in the U.S. census dataset. Then, we apply two bias mitigation algorithms and conduct an empirical comparison between the two. We thus compare the AIF360 bias mitigation algorithms in terms of group fairness[3] between men and women and whites and non-whites with respect to employment status, and loss in classifier accuracy.

We focus on group fairness for pragmatic reasons. Individual fairness is concerned with treating individuals with similar qualifications should be treated similarly [5, 22]. The problem with individual fairness is the assumption of an application specific similarity measure, which can be hard to define [5, 19, 20]. We focus on group fairness, rather than individual fairness for three reasons. First, we do not consider a specific application; our dataset is limited to demographic data. Second, there is a lack of specificity in the dataset we used. In this dataset it is difficult to treat people with similar qualifications alike or put differently - to identify individuals that have similar qualifications. Lastly, statistical notions of fairness are easily verifiable and does not require us to make any assumptions on the data [5, 22].

Note that, while our results may indicate some disparity between the compared groups, it is outside of the scope of this work to provide reasons for why they are treated differently by the prediction models. We also do not attempt

[2] `https://aif360.mybluemix.net/`

[3] specifically, we compare disparate impact, statistical parity difference, average odds difference, and equal opportunity

to rationalize or justify such differences, should they appear. Such work would fall under the domain of AI explainability.

Our contribution is a demonstration of the feasibility of off-the-shelf bias mitigation tools. The major shortcoming of this work is that we only tested one dataset and the performance of machine learning models is very much dataset dependent. Thus this work cannot qualify as a systematic in-debt review of the AIF360 toolkit. With this work, we primarily hope to motivate the practice of bias mitigation in research and commercially used supervised machine learning.

The paper is structured as follows. In Section 2 we introduce terms and definitions from the fairness domain. In Section 3 we describe our method and experimental setup. In section 4 we present the results from applying the two bias mitigation techniques. In Section 5 we discuss the results. In Section 6, we review related work on bias mitigation and the use of fairness metrics in machine learning. Lastly in Section 7 we summarise our results and outline directions for future work. All our code is available at: `https://github.com/throwaway02062020/INFO381`.

2 Definitions

For our definitions of bias we rely on Bellamy et al. [3].

A **protected attribute** is a feature that partitions a population into groups that have parity in terms of benefit received . A **favorable label** is a label whose value is considered the favorable outcome . Hereafter, we will refer to favorable outcome and favorable label interchangeably. A **privileged group** is a group that is systematically put at an advantage with respect to the beneficial outcome . A **unprivileged group** is a group which is systematically put at an disadvantage with respect to the beneficial outcome.

A **fairness metric** is a quantification of unwanted bias in training data or models [3]. **Statistical parity** entails that individuals from the protected and unprotected group should have the same probability of being assigned the favorable label. A value of zero is ideal. A negative value indicates that the unprivileged group is at an disadvantage [22].

Equal opportunity, also known as false negative error rate balance, is a fairness metric that entails that individuals from both the privileged and unprivileged groups have the same probability of being wrongly assigned the unfavorable label. A value of zero is ideal. A negative value indicates that the unprivileged group is at an disadvantage [22].

Average odds is satisfied when the true positive rate (TPR) and false-positive rate (FPR) is equal for the privileged and unprivileged group. A value of zero is ideal. A negative value indicates that the unprivileged group is at an disadvantage [22].

Disparate impact, is an estimate of unintentional bias in a label assignment task which occurs when a group is assigned widely different outcomes on the basis of its membership to a protected class. It is an indication that the selection process or the data underlying the process have become vitiated by latent bias, resulting in discrimination. The ideal value is 1. A value below 1 indicates a disadvantage for the unprivileged group [7].

3 Method

AIF360 is an open source toolkit for detecting and mitigating algorithmic bias developed by IBM research. The toolkit is part of IBM's Trusted AI initiative and is the first system that combines bias metrics, bias mitigation algorithms, metric explanations, and industrial usability [3]. It consists of over 71 bias detection metrics and 9 bias mitigation algorithms; our study focuses on Reject Option based Classification and Prejudice Remover for bias mitigation [3].

The performance of Reject Option based Classification (ROC) and Prejudice Remover (PR)[4] are evaluated using disparate impact, statistical parity difference, average odds difference, and equal opportunity difference. Performance of the classifier is evaluated using balanced accuracy and receiver operating characteristic curves. Fairness and classifier performance are compared before and after applying the bias mitigation techniques.

The dataset that we used is a sample of a U.S. census dataset from 2013[5]. The dataset [14] contains records of 131 302 individuals from a 2013 U.S. census. A total of 14 features are recorded: PeopleInHousehold, Region, State, MetroAreaCode, Age, Married, Gender, Education, Race, Hispanic, CountryOfBirthCode, Citizenship, EmploymentStatus, and Industry. The objective is to predict who are employed and unemployed excluding gender and race as features in the prediction model training. The mean difference in favorable label assignment we calculated for gender and race. Men (n = 31765) were associated with 9.5% more favorable labels than women (n = 33180) and whites (n = 52688) were associated with 9.7% more favorable labels than non-whites (n = 12257). Therefore, we set the privileged groups to be men and whites and the unprivileged groups to be women and non-whites. See Table 1 for an overview of the experimental setup.

Dataset	U.S. census data from 2013
Protected attributes	Race, Gender
Privileged class	White, Male
Unprivileged class	Non-white, Female
Classifiers	Logistic Regression Classifier, Regularized Logistic Regression Classifier
Bias mitigation methods	Reject Option based Classification [11] (post-processing) , Prejudice Remover [12] (in-processing)

Table 1: Overview of experimental setup

Pre-processing

For the pre-proccesing stage, we replaced null values for Industry and Education with the label 'missing'. All other instances with null values were removed. In

[4] Detailed explanation of ROC and PR can be found at: https://github.com/throwaway02062020/INF0381/blob/master/Bias%20mitigation%20with%20AIF360:%20A%20comparative%20study.pdf

[5] Dataset can be found at: https://www.kaggle.com/econdata/demographics-and-employment-in-the-united-states/version/1

total, we were left with 56 827 instances. The feature age, ranging from 0-80 was transformed into discrete buckets (bins) of decades. Persons of age 14 or younger were excluded, since they are below the legal working age [15]. Education, Industry, and Marriage was encoded as categorical features. PeopleInHousehold was discretized into categories of "living alone" (1 person), "couple" (2 person), and "family", (3 or more persons). The sensitive attribute "Hispanic" was excluded since it can act as a proxy for race. The race attribute was grouped into non-white and white.

For ROC, we split the dataset 70/30. 70% of the data was used for training and the remaining 30% was split evenly for testing and validation. The validation set was used to find the optimal classification threshold (θ) and ROC margin. These parameters define the *critical region* at the decision boundary for which predictions are made with the the highest uncertainty. For PR, the data is split 80/20, where 80% is for training and 20% is used for testing. In constrast to ROC, PR can not optimize for a specific fairness metric. Kamishima et al. [12] suggests testing PR with different η values. Due to the computational cost of learning the regularized function, we limited η values from 0 through 30.

4 Results

The following section presents the results from our experiments. We compare the results from before and after applying fairness constraints for each bias mitigation technique.

Result of ROC

Gender: Table 2 shows the result of the experiment with *gender* as the protected attribute. Balanced accuracy (80.14%, 78.02%). Statistical parity difference (-0.2203, -0.0438). Disparate impact was (0.7219, 0.9419). Average odds difference was (-0.1388, 0.0401). Equal opportunity difference was (-0.1803, -0.0091). The area under curve score (AUC) score with fairness constraints was 0.8747.

Race: Table 3 shows the result of the experiment with *race* as the protected attribute. Balanced accuracy (80.14%, 79.81%). Statistical parity difference (-0.1184, -0.0334). Disparate impact was (0.8317, 0.9510). Average odds difference was (-0.0622, 0.0207). Equal opportunity difference was (-0.0922, -0.0049). The AUC score with fairness constraints was 0.8747.

Result of PR

Figure 1 and 2 shows the relationship between accuracy and fairness metrics in response to changes in the penalty parameter η for Prejudice Remover. A drop in accuracy can be observed in response to increases in η. We observe significant improvements in fairness metrics at $\eta = 10$, without adverse impact on the classifier accuracy for both gender and race.

Gender: Table 2 shows the result of the experiment with *gender* as the protected attribute. We compare the fairness metrics before and after applying fairness constraints. Balanced accuracy (77.77%, 73.61%). Statistical parity difference (-0.0727, -0.0190). Disparate impact was (0.9144, 0.9779). Average odds difference was (0.0083, 0.0729). Equal opportunity difference was (-0.0490, -0.0184). The AUC score with fairness constraints was 0.8664.

Race: Table 3 shows the result of the experiment with *race* as the protected attribute. We compare the fairness metrics before and after applying fairness constraints. Balanced accuracy (77.79%, 76.45%). Statistical parity difference (-0.1052, -0.0510). Disparate impact was (0.8736, 0.9387). Average odds difference was (-0.0594, 0.0092). Equal opportunity difference was (-0.0517, -0.0232). The AUC score with fairness constraints was 0.8719.

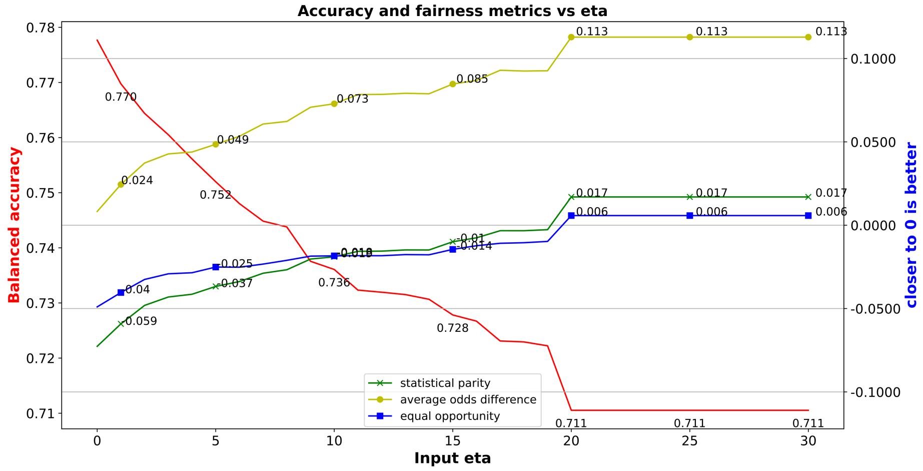

Figure 1: Shows the relationship between accuracy and fairness metrics as the penalty parameter η (eta) increases. With gender as the protected attribute.

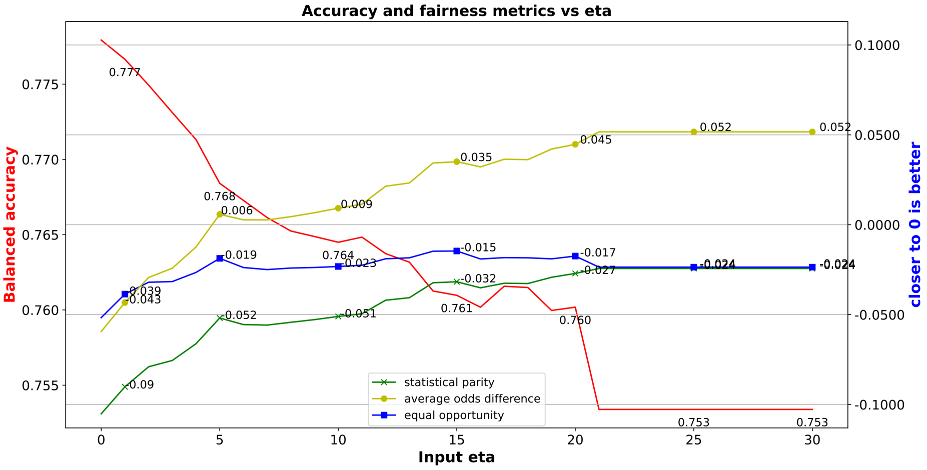

Figure 2: Shows the relationship between accuracy and fairness metrics as the penalty parameter η (eta) increases. With race as the protected attribute.

Comparison between ROC and PR

With regards to average odds difference, ROC overshoots for both gender and race - yielding an advantage for the unprivileged groups as opposed to a disadvantage prior to applying fairness constraints. However, the impact is less severe with fairness constraints - 13.88% in favor of men to 4.01% in favor of women and 6.22% in favor of whites to 2.07% in favor of non-whites. With PR, there is an improvement in average odds difference for race - 5.94% in favor of white to 0.92% in favor of non-white. However, in the case of gender, average odds worsened and turned in favor of women by 7.29% (from 0.83%).

The fairness metrics that did not overshoot was statistical parity and equal opportunity. Both of these metrics remains in favor of the privileged groups after applying fairness constraints. With regard to ROC and protected attribute gender, we see that statistical parity improves from 22.03% in favor of men to 4.38% in favor of men. For equal opportunity, the change is from 18.03% in favor of men to 0.91% in favor of men. With protected attribute race, statistical parity changed from 11.84% in favor of whites to 3.34% in favor of whites. Equal opportunity improved from 9.22% in favor of whites to 0.49% in favor of whites.

With regards to PR, statistical parity improves from 7.27% in favor of men to 1.9% in favor of men. For equal opportunity, the change is from 4.9% in favor of men to 1.84% in favor of men. With regards to race statistical parity improved from 10.52% in favor of white to 5.1% in favor of white. For equal opportunity, the improvement is from 5.17% in favor of white to 2.32% in favor of white.

For ROC with protected attribute gender, disparate impact was improved by 22% (from 72.19% to 94.19%). With race, we observed an improvement of 11.93% (from 83.17% to 95.10%). For PR, disparate impact for gender improved by 6.35% (from 91.44% to 97.79%), and race was improved by 6.51% (from 87.36% to 93.87%).

5 Discussion

In all cases, classifier accuracy was in the range of 73.61% to 80.14% with ROC outperforming PR under all conditions. With race as the protected attribute, ROC beat PR on all fairness metrics except for average odds difference. With gender as the protected attribute, the results were mixed. ROC was better in terms of average odds difference and equal opportunity, and PR was better on statistical parity and disparate impact.

We suspect that the accuracy of PR suffers because of the overfitting regularization, which ROC does not use. Thus, even though we set the penalty parameter (η) to 0, meaning, no fairness constraints are imposed, the accuracy will still be affected by the overfitting regularizer. With PR, we found that fairness metrics improve at the expense of accuracy as η increases. This observation is expected since PR is designed to remove prejudice at the expense of classifier accuracy [12].

With PR, there seems to be a steady decline in accuracy up to a certain threshold value of η. Past this point, the classifier loses its power to distinguish between employed and unemployed instances - resulting in a sudden loss of accuracy. We suspect that PR forces statistical independence of the non-sensitive attributes from the sensitive attribute as theta increases, i.e. sensitive

	ROC (no fairness constraints)	ROC (with fairness constraints)	PR (no fairness constraints)	PR (with fairness constraints)
Accuracy	80.14%	78.02%	77.77%	73.61%
Statistical parity difference	-0.2203	-0.0438	-0.0727	-0.0190
Disparate impact	0.7219	0.9419	0.9144	0.9779
Average odds difference	-0.1388	0.0401	0.0083	0.0729
Equal opportunity difference	-0.1803	-0.0091	-0.0490	-0.0184

Table 2: Results for protected attribute *Gender* with optimization for statistical parity. ROC (without fairness constraint) was run with an optimal classification threshold (θ) of 0.7326. With fairness constraints, the optimal classification threshold (θ) was 0.6930 with a ROC margin of 0.1253. PR (without fairness constraint) was run with penalty parameter (η) of 1.0. With fairness constraints the penalty parameter (η) was 10.0.

	ROC (no fairness constraints)	ROC (with fairness constraints)	PR (no fairness constraints)	PR (with fairness constraints)
Accuracy	80.14%	79.81%	77.79%	76.45%
Statistical parity difference	-0.1184	-0.0334	-0.1052	-0.0510
Disparate impact	0.8317	0.9510	0.8736	0.9387
Average odds difference	-0.0622	0.0207	-0.0594	0.0092
Equal opportunity difference	-0.0922	-0.0049	-0.0517	-0.0232

Table 3: Results for protected attribute *Race* with optimization for statistical parity. ROC (without fairness constraint) was run with an optimal classification threshold (θ) of 0.7326. With fairness constraints, the optimal classification threshold (θ) was 0.6831 with a ROC margin of 0.0776. PR (without fairness constraint) was run with penalty parameter (η) of 1.0. With fairness constraints the penalty parameter (η) was 10.0.

attributes are weighted less than before. Specifically, the classifier has less distinct instances to distinguish between employed and unemployed instances. This occurs because PR transforms the classification parameters of each instance to compensate for prejudice, making them less distinct. Thereafter, accuracy remains constant and fairness metrics become less sensitive to changes in η. See Figure 1 at η= 20 and Figure 2 at η= 21. We believe that the enforcement of statistical independence leads to a decrease in the AUC score, which explains the loss in accuracy. To substantiate our hypothesis, we plotted the AUC scores for η values 15 through 21 which shows a decrease in the AUC score. The sudden decrease in AUC score and classifier accuracy occur at the same value of θ. This phenomenon is unique for in-processing methods like PR, since they are modifying the classifier. Since ROC is a post-processing algorithm, it does not change the learned model of the classifier. Therefore, the diagnostic ability (AUC) of the classifier remains unchanged.

ROC calculates an optimal classification threshold (θ), which in turn means that it can always perform at a high level out of the box. With PR, the η value has to be chosen by the developer. Finding the correct η value for a dataset is both time consuming and computational expensive as the only way to find the right value is to run the classifier with different η values. Then, the developer has to decide which fairness metric(s) to optimize for. Even though it is computationally expensive, it allows for a more flexible implementation compared to ROC, since the model parameters of PR can be manually changed. Compared to PR, ROC has the advantage of not requiring modification of the classifier since it is a post-processing algorithm. As such, ROC can be applied to any existing decision-making system without changing the underlying classifier, unlike PR, which is implemented with a regularized logistic regression classifier.

One thing to be aware of, is that the results we got from PR and ROC was done using only one dataset. If we had applied the algorithms on other datasets, the results could vary because the performance is highly dependent on the dataset.

6 Related work

Previous studies [11, 17] show that ROC applied to logistic regression classifiers is good at mitigating bias while retaining classifier accuracy. On an adult income dataset, Kamiran, Karim and Zhang [11] managed to reduce statistical parity difference from 18% to $< 0.5\%$, while losing less than 2% accuracy. A subsequent test on a crime dataset reduced statistical parity difference from 40% to $< 0.5\%$ with an 8% decrease in accuracy (83% to 75%).

In another study, Lohia et al. [17] found that ROC had better results compared to other methods when measuring disparate impact. ROC was compared against three other bias mitigation methods on different datasets. ROC performed better at disparate impact compared to the other mitigation methods, but often at the expense of increasing individual bias [17].

Besides being efficient in reducing biases, ROC offers good control over discrimination and works with all probabilistic classifiers [11]. It is also deterministic, which means that it exhibits no randomness and will always produce the same output given the same input [1].

Lohia et. al. [17] compared both group fairness and individual fairness of three post-processing algorithms - reject option classification (ROC), equalized odds (EOP) and individual + group debiasing (IGD). The first two are from AIF360 and the last one is a custom debiasing technique. All the algorithms used logistic regression as the classifier. Three datasets were used, an adult income dataset from 1994, a german credit dataset, and the COMPAS recidivism dataset [17]. The adult income dataset is based on a U.S. census from 1994 and is similar to the one in our study. Sex and race were used as protected attributes for the adult income and COMPAS datasets, while sex and age were used for the German Credit dataset [17]. The dataset was split 60/20/20 for training, validation, and testing. IGD works in a similar way to ROC by altering the outcome of predicted labels. However, rather than sampling instances whose outcome is uncertain, IGD seeks to capture samples with individual fairness issues. An individual bias detector was trained on the validation set and used to identify instances in the unprivileged group with individual fairness issues. These instances were reassigned with the outcome that they would have if they were in the privileged group. All the other instances remained unaltered, including instances from the privileged group [17].

Since each dataset was tested with two protected attributes, the total test cases were $2 * 3 = 6$. Each case was tested in terms of disparate impact, individual bias, and balanced classification accuracy with each of the three algorithms, IGD, EOP, and ROC [17]. With regard to disparate impact, IGD performed consistently across all datasets. However, IGD was outperformed by ROC in 5 out of 6 cases, but often at the expense of increasing individual bias. In contrast, IGD was best in terms of the preservation of balanced accuracy and individual bias. Lohia et. al. [17] concludes that IGD can be appropriate when the aim is to improve both individual and group fairness.

7 Conclusion

A way to make algorithmic decision-making fairer is to use bias mitigation methods. Bias mitigation methods are used for optimizing certain fairness metrics such as equal opportunity. Contemporary approaches to bias mitigation in machine learning focus on intervention at the pre-processing, in-processing, and post-processing stages. The earlier you apply bias mitigation techniques, the more flexibility and potential you have of correcting bias. In this study, we have compared the performance of two bias mitigation techniques that intervene at different stages - an in-processing (reject option classification) and post-processing (Prejudice Remover). The performance was evaluated using group fairness metrics and classifier accuracy. Specifically, we compared disparate impact, statistical parity difference, average odds difference, and equal opportunity between men and women and whites and non-whites with respect to employment status. We found that, apart from one exception, both algorithms led to a fairer outcome. Additionally, both algorithms performed well with respect to loss in classifier accuracy and fairness metrics. Despite being a post-processing technique, ROC showed comparable results to PR with minimal loss in accuracy. However, PR is arguably more versatile in the sense that you can remove more bias at the expense of accuracy by increasing the penalty parameter. In the worst case, accuracy fell by 4% with PR and protected attribute gender. With protected attribute race, a fairer outcome was obtained and accuracy loss was negligible $(< 1.5\%)$.

Note that, the results that we obtained are dependent on the dataset. Therefore, the effectiveness of each algorithm is likely to differ between datasets. For future work, we intend to experiment with more datasets and run the experiments with random training samples to include standard deviation. We are specifically interested in analysing datasets from Norway that can be used to train prediction models for the Norwegian Labour and Welfare Administration. It would be interesting to explore how each of the algorithms respond to datasets with less or more bias.

References

[1] IBM AIF360. URL: https://aif360.mybluemix.net/resources#guidance.

[2] BAAI. Beijing AI Principles. BAAI. May 2019. URL: https://www.baai.ac.cn/news/beijing-ai-principles-en.html.

[3] Rachel K. E. Bellamy et al. "AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias". In: ArXiv abs/1810.01943 (2018). URL: https://ieeexplore.ieee.org/document/8843908.

[4] Alexandra Chouldechova. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments". In: Big data 5 2 (2016), pp. 153–163.

[5] Alexandra Chouldechova and Aaron Roth. "The Frontiers of Fairness in Machine Learning". In: CoRR abs/1810.08810 (2018). arXiv: 1810.08810. URL: http://arxiv.org/abs/1810.08810.

[6] Jeffrey Dastin. Amazon ditched AI recruiting tool that favored men for technical jobs. Oct. 2018. URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[7] Michael Feldman et al. Certifying and removing disparate impact. 2014. arXiv: 1412.3756 [stat.ML].

[8] Batya Friedman and Helen Nissenbaum. "Bias in Computer Systems". English (US). In: ACM Transactions on Information Systems 14.3 (July 1996), pp. 330–347. ISSN: 1046-8188.

[9] Karen Hao. The US just released 10 principles that it hopes will make AI safer. Jan. 2020. URL: https://www.technologyreview.com/2020/01/07/130997/ai-regulatory-principles-us-white-house-american-ai-initiatve/.

[10] Lily Hu and Yiling Chen. "Welfare and Distributional Impacts of Fair Classification". In: ArXiv abs/1807.01134 (2018).

[11] 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.

[12] Toshihiro Kamishima et al. "Fairness-Aware Classifier with Prejudice Remover Regularizer". In: Machine Learning and Knowledge Discovery in Databases. Ed. by Peter A. Flach, Tijl De Bie, and Nello Cristianini. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 35–50. ISBN: 978-3-642-33486-3.

[13] Will Knight. Why does Beijing suddenly care about AI ethics? MIT technology Review. URL: https://www.technologyreview.com/2019/05/31/135129/why-does-china-suddenly-care-about-ai-ethics-and-privacy/.

[14] Mithilesh Kumar. Demographics and Employment in the United States. Sept. 2013. URL: https://www.kaggle.com/econdata/demographics-and-employment-in-the-united-states/version/1.

[15] U.S. Department of Labor. "Workers Under 18". In: (N.D.). URL: https://www.dol.gov/general/topic/hiring/workersunder18.

[16] Jeff Larson et al. How We Analyzed the COMPAS Recidivism[break] Algorithm. 2016. URL: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[17] Pranay K. Lohia et al. "Bias Mitigation Post-processing for Individual and Group Fairness". In: (2018). arXiv: 1812.06135 [cs.LG].

[18] European Parliament. "EU guidelines on ethics in artificial intelligence: Context and implementation". In: (Sept. 2019). URL: https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_BRI(2019)640163.

[19] Evaggelia Pitoura, Irini Fundulaki, and Serge Abiteboul. "On Measuring Bias in Online Information". In: SIGMOD Rec. 46 (2017), pp. 16–21.

[20] Guy N. Rothblum and Gal Yona. Probably Approximately Metric-Fair Learning. 2018. arXiv: 1803.03242 [cs.LG].

[21] The National Strategy for Artificial Intelligence. Jan. 2020. URL: https://www.regjeringen.no/en/dokumenter/nasjonal-strategi-for-kunstig-intelligens/id2685594/?ch=7fn38.

[22] Sahil Verma and Julia Rubin. "Fairness Definitions Explained". In: Proceedings of the International Workshop on Software Fairness. FairWare '18. Gothenburg, Sweden: Association for Computing Machinery, 2018, pp. 1–7. ISBN: 9781450357463. DOI: 10.1145/3194770.3194776. URL: https://doi.org/10.1145/3194770.3194776.

[23] Russell T. Vought. "Memorandum for the Heads of Executive Departments and Agencies". In: (Jan. 2020). URL: https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf.