Bias mitigation with AIF360: A comparative study

Tor H. Aasheim, Knut A. Hufthammer, Sølve Ånneland, Håvard Brynjulfsen, and 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 algorithmic decision systems. Are decisions by an algorithmic decision-maker fair and justifiable? In this paper, we conduct an empirical analysis of two bias mitigation techniques from the AIF360 toolkit on a binary classification task. First, we train a logistic regression classifier on a US census dataset with target labels - employed or unemployed. Secondly, we apply bias mitigation techniques to the prediction model with *gender* and *race* as protected attributes. Third, we evaluate the models in accordance with fairness definitions based on predicted and actual outcome. Lastly, we discuss the result in view of the performance impact on the classifier.

Keywords: Machine learning · fairness · automated decision making · bias mitigation

1 Introduction

Ethical and social implications of AI have been hotly debated in recent years [7, 8, 14]. The European Union has adopted an "ethics by design" approach, incorporating ethical principles at the very start of the design of AI solutions [28]. In China, the recently proposed Beijing AI principles [5] aims to conform to human values, ethics, and autonomy for governance, use and healthy development of AI [20]. In 2020, the U.S. government outlined 10 principles [35] to regulate and promote trustworthy AI in the private sector, for the purpose of making it more fair, transparent, and safe [13]. 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 [25]. These interventions to regulate and facilitate safe and trustworthy AI is a sign of how disruptive AI tools can be without the right oversight.

Our case-study will utilize supervised machine learning, where the objective is to create a model from a dataset of labeled examples that can - given a set of feature vectors, predict the label of other examples. In other words, we want to “learn” a mathematical function (model) based on observations (collection of feature vectors) of some phenomenon to deduce a missing value (target label) of new observations [32].

Fairness in machine learning concerns the proportion of favourable labels allocated to members of a privileged group compared to favorable labels allocated to members on an unprivileged group [12]. A favourable label is a label whose value is considered the favorable outcome [6]. An unfair label assignment occurs when there’s a disproportional amount of desirable labels distributed among the unprivileged and privileged groups. The attribute that discerns the underprivileged or underrepresented individuals is called a protected attribute [6]. Examples of protected attributes include race, gender, and religion. Algorithmic bias occurs when an algorithm produces an unfair label assignment [12]. Some members of a population can be systematically underrepresented in a given social activity or role or unprivileged with respect to access to a certain good. The privileged and unprivileged groups can differ between context. For instance, in the U.S., men are unprivileged in the judicial system and receive harsher sentences for the same crimes compared to women [15]. Another example from the U.S., is that women are the unprivileged group when it comes to earnings, controlling for relevant factors, they earn 2% less than men [1].

Fairness in machine learning are grouped into definitions based on statistical and individual notions of fairness [8]. Statistical notions partitions individuals into groups whose outcome should have parity according to some statistical metrics [2, 8]. Examples of statistical metrics include statistical parity, average odds, and equal opportunity. These definitions ar+ defined in Section 2.

Definitions that seeks to treat similar individuals alike belong in the family of definitions known as individual fairness. The idea is that individuals with similar qualifications should be treated similarly [8, 34]. In contrast to statistical notions which “averages” out the population, individual fairness aims to not marginalise over individual merits. Thereby, acknowledging that individuals may have traits or attributes that should put them ahead of other individuals with less qualifications [34]. The problem with individual fairness is the assumption of an application specific similarity measure, which can be hard to define [8, 30, 33]. See Verma and Rubin [34] for an extensive overview of group and individual notions of fairness. 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. For example, the feature *Industry* does not distinguish between education and health services, i.e. people that work in education and health services fall within the same industry. Since our dataset lacks specificity, 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 [8, 34].

Motivated by recent findings of algorithmic bias in hiring and recidivism assessment [24, 9], we compare the efficacy of applying two bias mitigation algorithms from AIF360 on a U.S. census dataset [21]. Moreover, we compare 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. Then we discuss our findings in light of the performance impact on the classifier. We consider the following research question: How do two different bias mitigation algorithms from AIF360 compare in terms of group fairness and loss in classifier accuracy?

The success criteria rest on the completion of several steps. First, we measure the mean difference in outcome between all the groups in the U.S. census dataset. Then, we apply the two bias mitigation algorithms and conduct an empirical comparison between the two.

Note that, while our results may indicate some disparity between the groups, we do not provide any reasons for why they are treated differently. Nor do we attempt to rationalize or justify such differences, should they appear. To reiterate, we observe a limited sample of U.S. census data for gender bias, racial bias and analyse the effectiveness of applying two distinct bias mitigation algorithms to this dataset.

The remainder of our paper is structured as follows. In Section 2 we introduce sources of unfairness, bias mitigation techniques and the logistic regression classifier. 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.

2 Preliminaries

We introduce the concepts that we use throughout the report. For our definitions of bias we mainly rely on Bellamy et al. [6].

A **protected attribute** is a feature that partitions a population into groups that have parity in terms of benefit received [6]. A **favorable label** is a label whose value is considered the favorable outcome [6]. 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 [6]. A **unprivileged group** is a group which is systematically put at an disadvantage with respect to the beneficial outcome [6]. A **fairness metric** is a quantification of unwanted bias in training data or models [6]. **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 [34]. **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 [34]. **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 [34]. **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 [11]. A **bias mitigation algorithm** is a procedure for reducing unwanted bias in training data or machine learning models [6].

2.1 Sources of unfairness

Research has shown that there are many sources of unfairness that can affect the outcome of machine learning algorithms [8, 29, 31]. One of the more prominent sources of unfairness often mentioned is inherited bias [8, 29]. Inherited bias is not a product of the machine learning model, rather, it is inherited from an outside source. Examples of outside sources are society and the creators of a dataset [8, 29]. Consider, for instance, a decision making system created to mimic human decisions. If the training data is biased towards a certain part of the population, this bias will be *inherited* by the machine learning model. For example, since there exists no data about who *commits* crime, only those who get arrested, those parts of the population which are being policed at a higher rate are more likely to be over-represented in the dataset. Hence, we risk fitting the model better to the majority class and this group is more likely to be predicted as potential reoffenders [8].

Bias caused by missing data is another source of unfairness that can yield skewed results. When encountering missing values in a dataset, we have to make a choice on either removing the data or imputing (transforming) the data [29]. Regardless of which decision is taken, it is important to be aware that missing values might not be evenly distributed throughout the dataset. Ultimately, such missing values may have unforeseen effects on fairness when using the dataset for training a machine learning model [31].

Protected attributes (sensitive) like race and gender are typically not used in decision making. The reason for this is that historically, groups have been treated differently with respect to their sensitive attribute [29]. Even though such attributes are masked, the presence of “proxy” attributes can be used to derive the sensitive attributes [29]. For example, some geographical areas may have a higher number of residences from either the privileged or unprivileged groups. Therefore, location data can act as a proxy for a sensitive attribute. In that case, the machine learning algorithm will use sensitive attributes by proxy for prediction.

2.2 Bias mitigation & Logistic regression classifiers

Bias mitigation algorithms are typically divided into three approaches - pre-processing, in-processing, and post-processing techniques [6]. Pre-processing techniques reduce bias by altering the training data. In-processing techniques alter the learning model. Post-processing techniques are applied to predicted labels [2]. Generally, the earlier you apply bias mitigation in the machine learning pipeline, the more flexibility and potential you have of reducing bias [3]. There are different use cases for each of the processing techniques. Pre-processing can be used if the developer has access to the training data of an ML algorithm; In-processing is suitable if the algorithm is allowed to change the learning procedure of an ML algorithm; and post-processing fits if the algorithm can only be applied to a learned model [6]. For the purpose of our study we selected an in-processing and post-processing technique for bias mitigation applied to a logistic regression classifier.

For in-processing, we used Prejudice Remover (PR), and for the post-processing technique, we used Reject Option Classification (ROC). These techniques work only for algorithms based on probabilistic models [17, 19].

Logistic regression is a machine learning algorithm which is used to predict the probable outcome of an event, by fitting data points to a logistic curve [4]. The function used to draw the decision boundary in logistic regression is a sigmoid function, characterized by an "S" shape [36]. In a binary classification task, the classifier partitions examples into either of the two target labels [16]. The classifier decides which class an example belongs to by creating a decision boundary in the training process. When a logistic regression model is given a set of features to predict the target label, data is fitted to the logistic curve and the example is labeled based on which side of the decision boundary it is fitted to [27]. In the next two sections, we introduce two bias mitigation algorithms, which we apply to logistic regression classifiers. We use Reject Option based Classification and Prejudice Remover algorithm from AIF360. These algorithms are based on the papers by Kamiran, Karim and Zhang [17] and Kamishima et al. [19], respectively.

2.3 Reject option based classification (ROC)

ROC is applied to probabilistic classifier(s) to label instances in a way that ensures a more fair allocation of favorable labels according to some fairness criteria [17]. A central aspect of this labeling process is the critical region (See Equation 1). The critical region is an area within the decision boundary, i.e. the area for which the output of the label of a classifier is ambiguous. Instances in the critical region are associated with a high influence of bias and uncertainty of outcome. Therefore these instances in Equation 2 are labeled according to different rules than those instances that fall outside the decision boundary [17] (see Equation 3). The number of instances that falls within the decision boundary is adjusted by θ. A higher θ means that more instances fall~~s~~ within the decision boundary and the effect of that is less discrimination [17]. However, adjusting θ too high can lead to the reverse discrimination case. In that case, the privileged have become the disadvantaged group [18].

In more detail, ROC is a post-processing technique that operates at the decision boundary, associating favorable outcomes to unprivileged groups and unfavourable outcomes to privileged groups for instances whose outcome is ambiguous [17].

ROC is a cost-based method where the threshold for misplacing a unprivileged group instance as negative is much higher than misplacing a privileged group instance as negative. However, there is a tradeoff between accuracy and discrimination, controlled by a given threshold, θ. As θ increases, the discrimination will decrease at the expense of accuracy. ROC can be used for single and multiple classifiers. Multiple classifiers are considered to be more reliable in terms of accuracy and discrimination [17].

In single classification, the posterior probability for an instance, X, is $p(C+|X)$, where C+ is the favourable label. As the posterior probability approaches 0 or 1, the label for X is classified with high certainty. An instance closer to 0.5 falls within the decision boundary (critical region (1)) where the outcome label is more uncertain [17].

In ROC, instances are treated differently depending on whether they belong on the inside or outside of the decision boundary (1). The inclusion criteria for the decision boundary is given by: [17]:

$$\max[p(C^{+}|X), 1 - p(C^{+}|X)] \leq \theta \text{(where } 0.5 < \theta < 1) \tag{1}$$

That is, an instance falls within the decision boundary when the max value of $p(C+|X)$ and $1 - p(C+|X)$ is smaller than θ, conditioned upon θ being higher than 0.5 and lower than 1. The instances that are within the decision boundary are classified the following way [17]:

$$C(X) = \begin{cases} C+ & \text{when } X \in X^{d} \\ C- & \text{when } X \notin X^{d} \end{cases} \tag{2}$$

That is, when instance X is a member of the unprivileged group X^{d}, classify it with the favourable label, C+. Otherwise, classify it as the undesirable label, $C-$ [17].

Conversely, for instances outside the decision boundary:

$$C(X) = \begin{cases} C+, & \text{when } p(C+|X) > p(C-|X) \\ C-, & \text{otherwise} \end{cases} \tag{3}$$

Meaning, classify X as desirable label, C+, when the probability of the favourable label C+ occurring, given X, is higher than the probability of the undesirable outcome label C- occurring, given X. Otherwise, assign X an undesirable label, C- [17].

A drawback of ROC is that it only works on probabilistic classifiers that produce probability estimates. Additionally, it does not perform well on all types of datasets [17].

2.4 Prejudice remover

The mathematical model which the Prejudice Remover in AIF360 is based upon comes from Kamishima et al.[19]. Kamishima et al. [19] argues that simply removing sensitive attributes is not enough. Indirect influence of the sensitive attributes can still make the process biased. Unfairness occurring from indirect influence is described as a statistical dependence between sensitive attributes and the target label, either directly, or proxied through a non-sensitive attributes dependent on sensitive attributes [19].

Prejudice is defined as a statistically dependent relationship between the sensitive attribute S and other attributes. Three types of prejudice are outlined: direct prejudice, indirect prejudice and latent prejudice. Direct prejudice occurs when a sensitive attribute is used in a classifier. The absence of direct prejudice is defined as: $Y \perp\!\!\!\perp S - X$. This means that the target label Y is *independent* of the sensitive attribute S when the non sensitive attribute X, is already present.

Indirect prejudice occurs when there is a dependence between the target label Y, and the sensitive label S. Defined as: $Y \not\perp\!\!\!\perp S$. Indirect prejudice is measured with the *normalized prejudice index* (NPI):

$$NPI = \frac{PI}{\sqrt{(H(Y)H(S))}}[19] \tag{1}$$

Where $H(\cdot)$ is a function for entropy which takes an *event* as an argument. Entropy in information science refers to the average degree of information in a variable's possible outcomes. $PI/H(Y)$ shows how much information is taken from S and utilized in the prediction of Y. PI/H(S) is the quantity of exposed information when the target label Y is known. Normalization is the process of transforming data to a common frame of reference. Here the range of NPI is $\{0, 1\}$ after normalization. Here $\{0, 1\}$ are the possible values for NPI - either "0" or "1" as it's output.

PI stands for prejudice index, and is defined as the mutual information that both S and Y have. The formula for sampling this information over the set $\mathcal{D}$ is defined as:

$$PI = \sum_{(y,s)\in\mathcal{D}} \hat{Pr}[y, s] \ln \frac{\hat{Pr}[y, s]}{\hat{Pr}[y]\hat{Pr}[s]} \tag{2}$$

Where $\hat{Pr}$ is the training sample distribution.

Latent prejudice occurs when there is a statistical dependence in the relationship between the sensitive attribute S and the non-sensitive attribute X, given that Y is already influenced by X. In that case, the target label Y is dependent on the non sensitive attribute X.

Direct prejudice occurs when sensitive attributes like gender, race, and age are embedded in the prediction model. Consider how an individual might be unfairly treated in a hiring decision because of their race or gender. In the absence of prejudice, sensitive attribute S, has to be independent from the outcome.

In the case of *latent prejudice*, consider racial segregation in which the sensitive attribute, race, becomes correlated with non-sensitive attribute, demographic area. By utilizing a non-sensitive attribute *demographic area*, the model will indirectly use the sensitive attribute, race [19].

Indirect prejudice is exemplified by having a company where only certain people get promoted and the preference against the protected attribute is not stated in the company policy or used as a prerequisite. Nevertheless, in practise, the company systematically mistreats a group in favor of another, with respect to promotion. The effect is a correlation between the target value *promotion* and the sensitive attribute. Giving reason to suspect indirect prejudice being the root cause.

Kamishima et al. [19] suggest implementing regularizers as a prejudice removal technique. The first regularizer is simply a standard regularizer to avoid overfitting. The second regulalizer technique evens out the result by removing prejudice as previously defined. The prejudice remover minimizes the logistic regression expression. See Appendix (A) for full list of preceding equations.

$$\sum_{y_i,s_i,x_i} In\mathcal{M}[y_i \mid x_i, s_i; \varTheta] + \eta Rpr(\mathcal{D}, \varTheta) + \frac{\lambda}{2} \sum_{s \in S} \mid\mid w_s \mid\mid \frac{2}{2} \tag{3}$$

The preceding equation produces the optimal parameters: $\{w_s^*\}$. The whole point about the optimal parameters is to use them in a transformation where each classified instance is transformed in order to become what they "optimally" would be without any discrimination. By using the set of optimal parameters the prejudice remover computes the following equations:

In order to predict the probability Y=1, by using the set of optimal parameters $\{w_s^*\} \in (X_{new}, Y_{new})$ we get:

$$Pr[Y = 1 \mid X_{new}, Snew; \{w_s^*\}] = \sigma(X_{new}^T W_{s\ new}^*) \tag{4}$$

In AIF360, the prejudice remover is implemented the same way as in Kamishima et al. [19], with a logistic regression classifier. Worth noting is that when utilizing the prejudice remover, a different result will be given every time. This is because the prejudice remover, being non-deterministic, randomly samples the training data each time.

3 Method

The dataset that we used was a sample of a U.S. census dataset from 2013. The objective is to predict who are employed and unemployed using gender and race as protected attributes. The mean difference in positive outcome was calculated for gender and race. Men (n = 31765) were associated with 9.5% more positive outcomes than women (n = 33180) and whites (n = 52688) were associated with 9.7% more positive outcomes 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.

The performance of ROC and PR were evaluated using disparate impact, statistical parity difference, average odds difference, and equal opportunity difference. Performance of the classifier was evaluated using balanced accuracy and receiver operating characteristic curves (herafter, ROC curves). Fairness and classifier performance were compared before and after applying the bias mitigation techniques.

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 [17] (post-processing) , Prejudice Remover [19] (in-processing)

Table 1: Overview of experimental setup

3.1 Dataset

The dataset [21] contains records of 131 302 individuals from a 2013 U.S. census.

The dataset has the following attributes:

- PeopleInHousehold
- Region
- State
- MetroAreaCode
- Age
- Married
- Gender
- Education
- Race
- Hispanic
- CountryOfBirthCode
- Citizenship
- EmploymentStatus
- Industry

3.2 Data preparation

Feature selection is the practice of choosing the attributes that contributes to the prediction of the target labels [37]. Pearson correlation analysis was used to gain more knowledge about the dataset and to find strong relationships between features. In general, correlation analysis determines the strength of the relationship between two item sets. The result of a correlation analysis is a either a positive value or a negative value. A positive value indicates that two variables have a positive relationship, and a negative value indicates a negative relationship. A higher correlation number indicates a stronger relationship. In Pearson analysis, the coefficient varies between 1 and -1 [22]. The results from the correlation analysis can be found Figure 1 and 2 in Appendix B.

Pre-processing is an important step in machine learning, where data is cleaned, transformed and aggregated to a more suitable format for further analysis. For our pre-proccesing stage, we replaced null values for Industry and Education with the label 'missing'. All other instances with null values were removed. In 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 [23]. 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 (1)* 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. [19] 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. First we present the results from the reject option classification technique. Then, we present the results from PR. Lastly, we compare the results of ROC against PR.

4.1 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 (see Figure 3).

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 (See Figure 4).

4.2 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 (See Figure 3).

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 (See Figure 4).

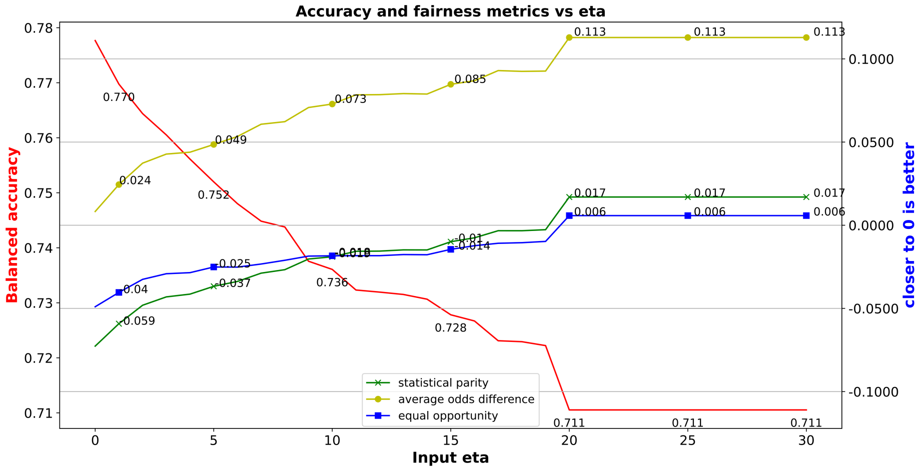

Fig. 1: Shows the relationship between accuracy and fairness metrics as the penalty parameter η (eta) increases. With gender as the protected attribute.

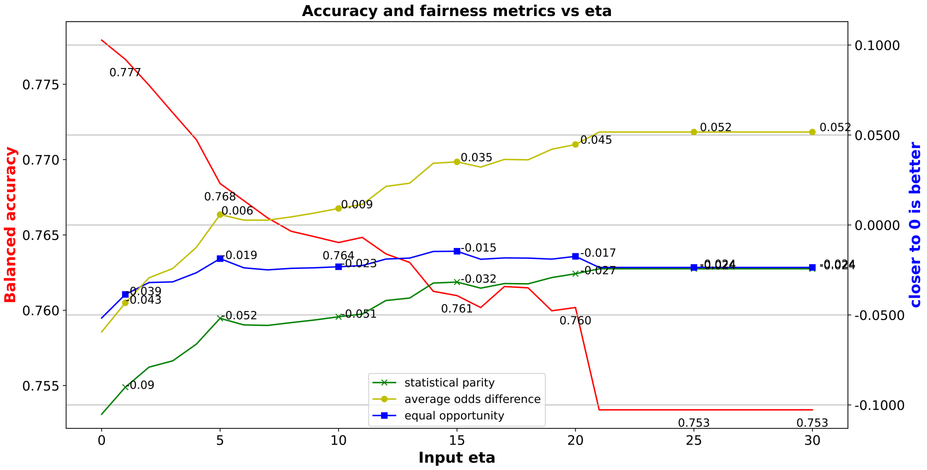

Fig. 2: Shows the relationship between accuracy and fairness metrics as the penalty parameter η (eta) increases. With race as the protected attribute.

4.3 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%).

	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.

ROC curves plot the true positives (also known as sensitivity) on the Y-axis and false positives (also known as specificity) on the X-axis. It shows the relationship between true positives and false positives [10]. In a perfect model with 100% sensitivity and 100% specificity, the ROC curve runs from origo via (0,1) to (1,1) and will have an AUC score of 1. A model that is completely random (think coin flip) will trace along the 45° diagonal [10]. In our case, the true positives (correctly predicted employed) are compared to the false positives (wrongly predicted employed). Our results show that ROC have an AUC score of 0.8747 for both gender and race. For PR, the AUC score for gender and race was 0.8664 and 0.8719, respectively. See Figure 3 and 4 for a depiction of the results.

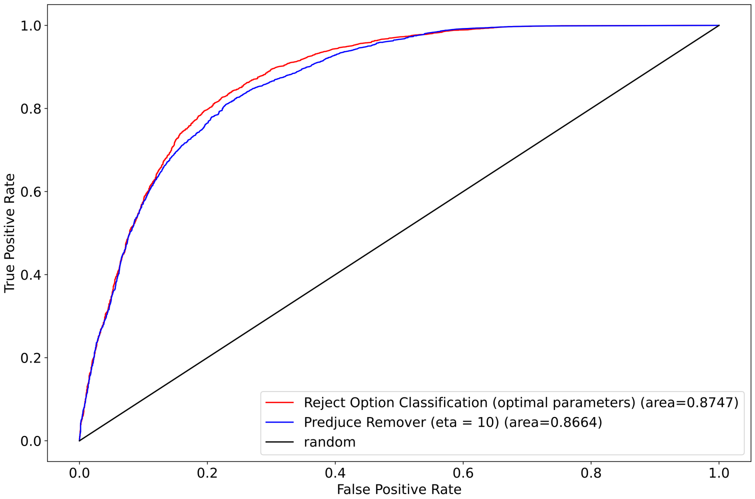

Fig. 3: ROC curve comparison between reject option classification ($\theta = 0.6930$; ROC margin $= 0.1253$) and Prejudice Remover ($\eta = 10$). The protected attribute was gender.

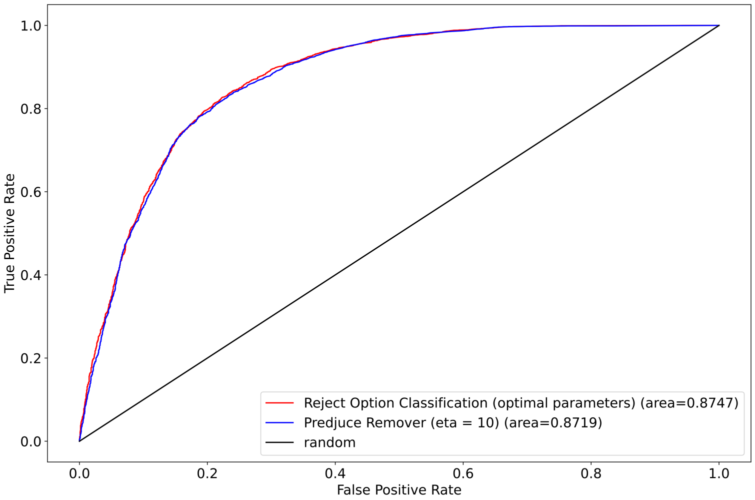

Fig. 4: ROC curve comparison between reject option classification ($\theta = 0.6831$; ROC margin $= 0.0776$) and Prejudice Remover ($\eta = 10$). The protected attribute was race.

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 (see equation 2), 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 [19].

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 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 (see Figure 1 and 2 in Appendix C). The 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

Verma and Rubin [34] analysed the usage of various fairness definitions on credit risk prediction. More specifically, they explored whether men and women were treated differently by a classifier under a wide selection of fairness definitions. The study can be considered two-part. The first part of the paper introduces definitions based on statistics, while the latter part deals with similarity-based measures and causal reasoning. An extensive list of 21 fairness definitions was collected by reviewing existing literature on machine learning and fairness. Verma and Rubin used a German Credit Dataset to demonstrate how a label assignment can conform to some fairness definitions and be in violation with others [34]. In other words, the model satisfies certain definitions but violates others in identical cases.

The dataset used by [34] contains 1000 entries of loan applicants with 20 features such as credit history, purpose of loan, existing checking account status, gender, age and employment status. The target label was the customer's credit score rating - good or bad [34]. Gender and marital status were used as protected attributes. Verma and Rubin [34] focused primarily on differences between the treatment of married and divorced males versus married and divorced females. Verma and Rubin [34] trained a logistic regression classifier using 90% of the data with the remaining 10% reserved for testing. Initial coefficients learned by the classifier put single males at a slight advantage over divorced males while treating females similarly to married males [34].

Verma and Rubin [34] group statistical definitions of fairness into three classes - definitions based on predicted outcomes, definitions based on predicted and actual outcomes and those based on predicted probabilities and actual outcomes. The basis for these types of definitions are statistical metrics such as positive predictive value (PPV), negative predictive value (NPV), false discovery rate (FDR), true positive rate (TPR), among others. For definitions based on predicted outcome, group fairness and conditional statistical parity is discussed. Group fairness is the simplest of these and is achieved when applicants are given an equal probability to be assigned to the positive predictive class. Conditional statistical parity extends this definition and permits a set of legitimate attributes to influence the outcome such as credit history, age and employment. In their case, the classifier failed to satisfy statistical parity, but conformed to equal opportunity [34]. The second class takes the actual outcome into consideration when deciding on fairness. Verma and Rubin [34] introduce metrics such as predictive parity, treatment equality, and equal opportunity to name just a few. The last class is based on actual outcome and predicted probabilities, i.e. the predicted probability of having a good or bad credit score. Within this class, we find test-fairness, well-calibration, balance for positive class, and balance for negative class. The logistic regression classifier was found to partially conform to both test-fairness and well-calibration, meaning, they were satisfied only under certain conditions. Under the definition of balance for positive class, the fairness metrics were satisfied, while for the opposing negative balance class the metrics were not met [34].

Results show that the statistical metrics were more likely to assign a good credit score to male applicants. In particular, false positive rate (FPR) for married/divorced male was 0.70 and 0.55 for females. This yields a true negative rate of 0.30 for males and 0.45 for their female counterpart. Thus, the classifier was more likely to associate a good credit score to males with an actual bad credit score. Conversely, the opposite was the case for females. However, the classifier appeared to treat men and women with an actual good credit score equally. A false negative rate (FNR) of 0.14 for married/divorced applicants were observed regardless of gender. Inversely, the true positive rate was 0.86. This means that both men and women with an actual good credit score is just as likely to be assigned a positive outcome. In addition, of those that are predicted to have a bad credit score, men were more likely to have an actual good credit score. Verma and Rubin [34] note that most of the statistical metrics, while easy to measure, assumes a recording of the actual outcome. This is problematic since we don't know whether the real classified data will conform to the same distribution used in the training data [34]. Verma and Rubin [34] conclude by arguing that the fairness of the classifier is dependent on the notion of fairness one wants to adopt and calls out for more work to better understand the appropriateness of each definition [34].

Kamiran, Karim and Zhang [17] states that ROC is a postprocessing technique that uses probabilities constructed by a probabilistic classifier. Previous studies [17][26] 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 [17] 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. [26] 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 [26].

Besides being efficient in reducing biases, ROC offers good control over discrimination and works with all probabilistic classifiers [17]. It is also deterministic, which means that it exhibits no randomness and will always produce the same output given the same input [3].

Lohia et. al. [26] 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 [26]. 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 [26]. 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 [26].

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 [26]. 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. [26] 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.

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. It would be interesting to see how each of the algorithms respond to datasets with less or more bias.

References

1. Gender pay gap statistics for 2020, https://www.payscale.com/data/gender-pay-gap
2. AIF360, I.: https://aif360.mybluemix.net/resources#glossary
3. AIF360, I.: https://aif360.mybluemix.net/resources#guidance
4. Andrey, K., Song, Y., Kim, M., Lee, K., Cheon, J.: Logistic regression model training based on the approximate homomorphic encryption. BMC Medical Genomics **11** (10 2018). https://doi.org/10.1186/s12920-018-0401-7
5. BAAI: Beijing ai principles. BAAI (may 2019), https://www.baai.ac.cn/news/beijing-ai-principles-en.html
6. Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K.N., Richards, J.T., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R., Zhang, Y.: Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. ArXiv **abs/1810.01943** (2018)
7. Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data **5 2**, 153–163 (2016)
8. Chouldechova, A., Roth, A.: The frontiers of fairness in machine learning. CoRR **abs/1810.08810** (2018), http://arxiv.org/abs/1810.08810
9. Dastin, J.: Amazon ditched ai recruiting tool that favored men for technical jobs (Oct 2018), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G
10. Fawcett, T.: An introduction to roc analysis. Pattern Recognition Letters **27**(8), 861–874 (2006). https://doi.org/10.1016/j.patrec.2005.10.010
11. Feldman, M., Friedler, S., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact (2014)
12. Friedman, B., Nissenbaum, H.: Bias in computer systems. ACM Transactions on Information Systems **14**(3), 330–347 (Jul 1996)
13. Hao, K.: The us just released 10 principles that it hopes will make ai safer (Jan 2020), available from: https://www.technologyreview.com/2020/01/07/130997/ai-regulatory-principles-us-white-house-american-ai-initiatve/
14. Hu, L., Chen, Y.: Welfare and distributional impacts of fair classification. ArXiv **abs/1807.01134** (2018)
15. Huffpost: Study finds huge gap between how long men and women spend in prison (Sep 2012), https://www.huffpost.com/entry/men-women-prison-sentence-length-gender-gap_n_1874742
16. Jacob, A.: Modelling speech emotion recognition using logistic regression and decision trees. International Journal of Speech Technology **20**(4), 897–905 (2017)
17. Kamiran, F., Karim, A., Zhang, X.: Decision theory for discrimination-aware classification. 2012 IEEE 12th International Conference on Data Mining (2012). https://doi.org/10.1109/icdm.2012.45
18. Kamiran, F., Mansha, S., Karim, A., Zhang, X.: Exploiting reject option in classification for social discrimination control. Information Sciences **425**, 18 – 33 (2018). https://doi.org/https://doi.org/10.1016/j.ins.2017.09.064, http://www.sciencedirect.com/science/article/pii/S0020025517309830
19. Kamishima, T., Akaho, S., Asoh, H., Sakuma, J.: Fairness-aware classifier with prejudice remover regularizer. Machine Learning and Knowledge Discovery in Databases Lecture Notes in Computer Science p. 35–50 (2012). https://doi.org/10.1007/978-3-642-33486-3_3

20. Knight, W.: Why does beijing suddenly care about ai ethics? MIT technology Review, https://www.technologyreview.com/2019/05/31/135129/why-does-china-suddenly-care-about-ai-ethics-and-privacy/
21. Kumar, M.: Demographics and employment in the united states (09 2013), available from https://www.kaggle.com/econdata/demographics-and-employment-in-the-united-states/version/1
22. Kumar, S., Chong, I.: Correlation analysis to identify the effective data in machine learning: Prediction of depressive disorder and emotion states. International Journal of Environmental Research and Public Health **15**(12), 2907 (2018). https://doi.org/10.3390/ijerph15122907
23. of Labor, U.D.: Workers under 18 (ND), available from: https://www.dol.gov/general/topic/hiring/workersunder18
24. Larson, J., Mattu, S., Kirchner, L., Angwin, J.: How we analyzed the compas recidivism[break] algorithm (2016), https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
25. of Local Government, M., Modernisation: The national strategy for artificial intelligence (Jan 2020), available from: https://www.regjeringen.no/en/dokumenter/nasjonal-strategi-for-kunstig-intelligens/id2685594/?ch=7fn38
26. Lohia, P.K., Ramamurthy, K.N., Bhide, M., Saha, D., Varshney, K.R., Puri, R.: Bias mitigation post-processing for individual and group fairness (2018)
27. Müller, A.C.: Introduction to machine learning with python : a guide for data scientists (2016)
28. Parliament, E.: Eu guidelines on ethics in artificial intelligence: Context and implementation (Sep 2019), https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_BRI(2019)640163
29. Pessach, D., Shmueli, E.: Algorithmic fairness (2020)
30. Pitoura, E., Fundulaki, I., Abiteboul, S.: On measuring bias in online information. SIGMOD Rec. **46**, 16–21 (2017)
31. Plumed, F., Ferri, C., Nieves, D., Hernandez-Orallo, J.: Fairness and missing values. CoRR **abs/1905.12728** (2019), http://arxiv.org/abs/1905.12728
32. Raschka, S., Mirjalili, V.: Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2, 3rd Edition. Packt Publishing (2019), https://books.google.no/books?id=sKXIDwAAQBAJ%7D
33. Rothblum, G.N., Yona, G.: Probably approximately metric-fair learning (2018)
34. Verma, S., Rubin, J.: Fairness definitions explained. In: Proceedings of the International Workshop on Software Fairness. p. 1–7. FairWare '18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3194770.3194776, https://doi.org/10.1145/3194770.3194776
35. Vought, R.T.: Memorandum for the heads of executive departments and agencies (Jan 2020), available from: https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf
36. Wiki, P..A.: Logistic regression. https://docs.paperspace.com/machine-learning/wiki/logistic-regression (oct nd)
37. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proceedings of the Fourteenth International Conference on Machine Learning. p. 412–420. ICML '97, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1997)

Appendices

A Prejudice Remover

The following section presents the equations that the prejudice remover is built upon. These equations are outlined in Kamishima et al. [19]. We start out by wanting to maximise the log-likelihood of the following expression:

$$\mathcal{L}(\mathcal{D};\Theta) = \sum_{(y_i,x_i,s_i)\in\mathcal{D}} \ln \mathcal{M}[y_i|x_i,s_i;\Theta] \tag{1}$$

Here $\mathcal{L}$ is the lagrangian function, used for solving constrained minimization problems. $\mathcal{D}$ is in this instance the set of the sensitive, non sensitive and outcome attribute $\{y,x,s\}$. Θ is the set of model parameters. Note that ; is used to separate input variables from the model parameters. We start by taking the sum of each y,x,s element of the set $\mathcal{D}$ within the expression $\ln \mathcal{M}[y_i|x_i,s_i;\Theta]$. This expression says that we take the logarithm of the entire function $\mathcal{M}$. $\mathcal{M}$ is the probability of outcome y given that each corresponding x and s has occurred with the specified the model parameters. We get the next expression by adding in the regularization parameters to (A). Afterwards, we get the following expression to minimize:

$$-\mathcal{L}(\mathcal{D};\Theta) + \eta R(\mathcal{D};\Theta) + \frac{\lambda}{2}||\Theta||_2^2 \tag{2}$$

$R(\mathcal{D},\Theta)$, is the fairness equalizer regularizer and $||\Theta||_2^2$ is here the standard L2 regularizer to avoid overfitting. The regularization parameters are η and λ. They are both positive regularization parameters - η is used over the fairness regularizer and Θ is used over the standard regularizer.

Using a logistic regression model we get:

$$\mathcal{M}[y|x,s;\Theta] = y\sigma(x^T w_s) + (1-y)(1-\sigma(X^T W_s)) \tag{3}$$

Here, x are real vectors, s is a discrete value and Y is a binary value $\{0,1\}.\sigma$ is a sigmoid function. x^T and W_s are weight vectors.

From here we want to remove any prejudice. We start out with the equation for *prejudice index* (PI):

$$\begin{aligned} PI &= \sum_{(Y,S)} \hat{Pr}[Y,S] \ln \frac{\hat{Pr}[Y,S]}{\hat{Pr}[S]\hat{Pr}[Y]} \\ &= \sum_{(X,S)} \tilde{Pr}[X,S] \sum_{(Y)} \mathcal{M}[Y|X,S;\Theta] \ln \frac{\hat{Pr}[Y,S]}{\hat{Pr}[S]\hat{Pr}[Y]} \end{aligned} \tag{4}$$

Which can be reduced down to:

$$\sum_{(x_i,s_i)\in\mathcal{D}} \sum_{y\in[0,1]} \mathcal{M}[y|x_i,s_i;\Theta] \ln \frac{\hat{Pr}[y|s_i]}{\hat{Pr}[y]} \tag{5}$$

$$\hat{Pr}[y|s] = \int_{dom(X)} Pr^{*}[X|s]\mathcal{M}[y|X, s; \Theta]dX \tag{6}$$

Since it is very expensive to calculate the entire dataset if it is very large. We take an approximation of this formula by sampling a mean:

$$\hat{Pr}[y|s] \approx \frac{\sum_{(x_i,s_i)\in\mathcal{D} s.t. \quad s_i=s} \mathcal{M}[y|x_i, s; \Theta]}{|\{(x_i, s_i) \in \mathcal{D}, s.t. \quad s_i = s\}|} \tag{7}$$

We do the same for:

$$\hat{Pr}[y] \approx \frac{\sum_{(x_i,s_i)\in\mathcal{D}} \mathcal{M}[y|x_i, s_i; \Theta]}{|\mathcal{D}|} \tag{8}$$

The end result is a prejudice remover function $Rpr(\mathcal{D}, \Theta)$ that uses both regularizers and can be expressed as:

$$\sum_{(x_i,s_i)\in\mathcal{D}} \sum_{y\in[0,1]} \mathcal{M}[y|x_i, s_i; \Theta] \ln \frac{\hat{Pr}[y|s_i]}{\hat{Pr}[y]} \tag{9}$$

With logistic regression, we get the follow expression to minimize:

$$\sum_{y_i,s_i,x_i} In\mathcal{M}[y_i \mid x_i, s_i; \Theta] + \eta Rpr(\mathcal{D}, \Theta) + \frac{\lambda}{2} \sum_{s\in S} \mid\mid w_s \mid\mid \frac{2}{2} \tag{10}$$

This gives us a set of optimal parameters, $\{w_s^*\}$.

Now, in order to predict the probability Y=1, with the expression by using the set of optimal parameters $\{w_s^*\} \in (X_{new}, Y_{new})$, we get:

$$Pr[Y = 1 \mid X_{new}, Snew; \{w_s^*\}] = \sigma(X_{new}^{T} W_{s\ new}^{*}) \tag{11}$$

B Pearson correlation analysis

	Retired	Unemployed	Disabled	Not in Labor force	Employed
Retired	1	-0.274517	-0.096497	-0.10216	-0.04785
Unemployed	-0.274517	1	-0.507055	-0.536814	-0.251436
Disabled	-0.096497	-0.507055	1	-0.188697	-0.088383
Not in labor force	-0.10216	-0.536814	-0.188697	1	-0.09357
Employed	-0.04785	-0.251436	-0.088383	-0.09357	1
White	0.044122	-0.039261	0.051057	-0.050068	0.050256
Non white	-0.044122	0.039261	-0.051057	0.050068	-0.050256

Fig. 1: Shows the relationship between race and employment status

	Retired	Unemployed	Disabled	Not in Labor force	Employed
Retired	1	-0.274517	-0.096497	-0.10216	-0.04785
Unemployed	-0.274517	1	-0.507055	-0.536814	-0.251436
Disabled	-0.096497	-0.507055	1	-0.188697	-0.088383
Not in labor force	-0.10216	-0.536814	-0.188697	1	-0.09357
Employed	-0.04785	-0.251436	-0.088383	-0.09357	1
Female	0.003856	-0.105747	0.103214	0.051494	-0.024956
Male	-0.003856	0.105747	-0.103214	-0.051494	0.024956

Fig. 2: Shows the relationship between gender and employment status

C ROC curves

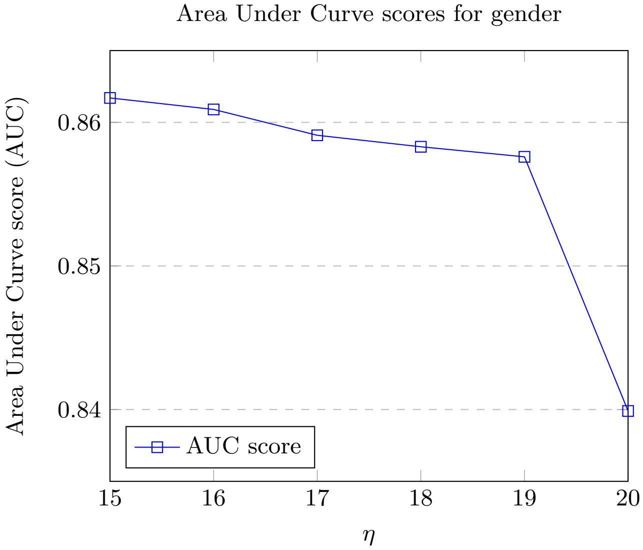

Fig. 1: AUC score from η values 15 through 20

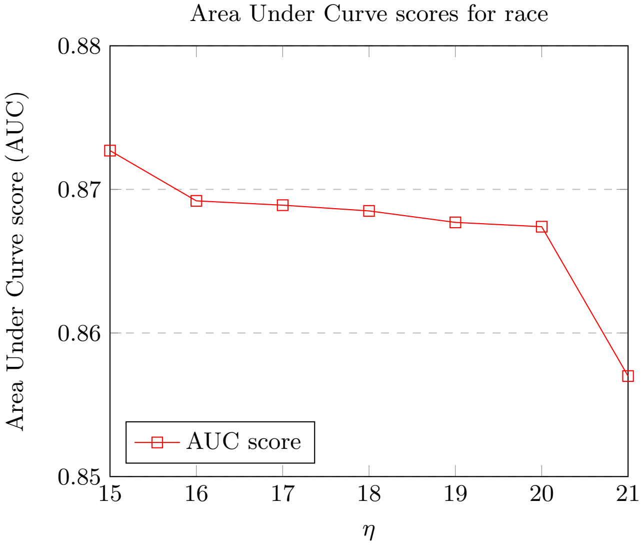

Fig. 2: AUC score from η values 15 through 21

D External repository (code)

Link to our repository: https://github.com/throwaway02062020/INFO381

The following are the relevant files of the project:

- ai360_reject_option_classification.ipynb - Bias mitigation using ROC.
- prejudice_remover.ipynb - Bias mitigation using PR.
- prejudice_remover.ipynb - Code from the previous two files combined.
- EmploymentDataset.py - StandardDataset implementation of the U.S. dataset.
- util.py - Loads and preprocesses the dataset contained in /data/raw (used in all notebooks).