Sandro Radovanović*, Marko Ivić
University of Belgrade, Faculty of Organizational Sciences, Serbia

Enabling Equal Opportunity in Logistic Regression Algorithm

DOI: 10.7595/management.fon.2021.0029

Abstract:

Research Question: This paper aims at adjusting the logistic regression algorithm to mitigate unwanted discrimination shown towards race, gender, etc. **Motivation:** Decades of research in the field of algorithm design have been dedicated to making a better prediction model. Many algorithms are designed and improved, which made them better than the judgments of people and even experts. However, in recent years it has been discovered that predictive models can make unwanted discrimination. Such unwanted discrimination in the predictive model can lead to legal consequences. In order to mitigate the problem of unwanted discrimination, we propose equal opportunity between privileged and discriminated groups in the logistic regression algorithm. **Idea:** Our idea is to add a regularization term in the goal function of the logistic regression. Therefore, our predictive model will solve both the social problem and the predictive problem. More specifically, our model will provide fair and accurate predictions. **Data:** The data used in this research present U.S. census data describing individuals using personal characteristics with a goal to provide a binary classification model for predicting if an individual has an annual salary above $50k. The dataset used is known for disparate impact regarding female individuals. In addition, we used the COMPAS dataset aimed at predicting recidivism. COMPAS is biased toward African-Americans. **Tools:** We developed a novel regularization technique for equal opportunity in the logistic regression algorithm. The proposed regularization is compared against classical logistic regression and fairness constraint logistic regression, using a ten-fold cross-validation. **Findings:** The results suggest that equal opportunity logistic regression manages to create a fair prediction model. More specifically, our model improved both disparate impact and equal opportunity compared to classical logistic regression, with a minor loss in prediction accuracy. Compared to the disparate impact constrained logistic regression, our approach has higher prediction accuracy and equal opportunity, while having a lower disparate impact. By inspecting the coefficients of our approach and classical logistic regression, one can see that proxy attribute coefficients are reduced to very low values. **Contribution:** The main contribution of this paper is in the methodological part. More specifically, we implemented an equal opportunity in the logistic regression algorithm.

Keywords: Algorithmic decision-making, Algorithmic Fairness, Fair prediction models, Logistic Regression, Equal Opportunity.

JEL Classification: C51, C52, I24

1. Introduction

Usage of data mining and machine learning algorithms made many improvements in the decision-making process overall. For example, time to making a decision is reduced, decision-making accuracy is higher, which consequently led to greater income or reduced costs. The main reason for these improvements is that data mining and machine learning algorithms can take into account more criteria influencing the decision than a human expert is capable to process, and also the process it faster than the human can conduct (Larose & Larose, 2015). These algorithms utilize prior experience obtained from the historical data, create a model based on them, and recommend the best possible action. In the process of creating a prediction model, algorithms tend to reduce loss function that is directed toward the accuracy of the predictive model.

* Corresponding author: Sandro Radovanović, e-mail:sandro.radovanovic@fon.bg.ac.rs

However, predictive models can also systematically discriminate specific subgroups of people, thus making decisions unethical and prone to legal consequences (Barocas & Selbst, 2016).

Having the above mentioned in mind, one would like to create a predictive model that is both accurate and fair toward each subgroup of the population. The cost of being unfair can be very high. There are numerous examples of unfair practices that led to financial consequences. For example, Google advertisements had the gender bias. Namely, male Google users receive more ads for higher-paying jobs and career coaching services in comparison with female Google users (Datta et al., 2015). However, there are examples of consequences that led to social unrest. Algorithmic decision-making tools for criminal risk assessment such as COMPAS resulted in the biased treatment of individuals based on race, i.e., similar individuals had different treatment due to race. While it is, from a mathematical point of view, easy to define accuracy (or loss function), fairness is not. The term being fair is a social science construct that is much dependent on the application or use, the context of the decision-making, and subjects that are influenced by the decision (Corbett-Davies et al., 2017). However, fairness often requires two groups of people, one being a privileged group, and other being a discriminated group. By collecting and transferring notions of fairness into mathematical constructs, a new theoretical and practical research area emerged called fair algorithmic decision-making.

Therefore, the challenge that this paper tackle is finding the definition of fairness that is suitable for the predictive models. After that, our aim is to convert it to a mathematical formula that is suitable for the optimization (i.e., linear or convex formulation) and implement it in the data mining and machine learning algorithms. The problem is challenging from both an algorithmic point of view and from a managerial point of view. From an algorithmic point of view, the problem occurs from the source of unfairness and discrimination. Most often, the source of unfairness is unknown (Chen et al., 2019) making it hard to identify and intervene in the algorithmic design. From a managerial point of view, unfairness in the decision-making process can lead to financial losses, or intangible losses (such as the brand of the company, loss of trust, etc.).

In this paper, we implement equal opportunity as fairness in data mining and machine learning models. More specifically, instead of observing fairness on the dataset as a whole, we observe fairness only for the instances with the desired outcome. It has been shown that observing fairness on the complete dataset, although improving fairness as a whole, does not guarantee that every subgroup of people in the dataset will get an equal opportunity to achieve the desired outcome (Radovanovic et al., 2020). However, the majority of papers in the literature observe and calculate fairness on the complete dataset (Oneto & Chiappa, 2020). If one needs to make a decision then fairness is not to be observed on the complete dataset, but rather on the instances where the desired outcome is achieved. This paper presents an approach for achieving equal opportunity in the data mining and machine learning models by introducing an adaptation of equal opportunity as a regularization function. Equal opportunity means that each individual should have an equal opportunity of getting the desired outcome, or that a proportional number of individuals from each group (both discriminated and privileged) get the desired outcome.

We selected the logistic regression algorithm since it is one of the most used algorithms in the field of data mining and machine learning (Wu et al., 2008). Besides being one of the most used algorithms, it is considered as an interpretable one (coefficients of the logistic regression are interpreted as the logarithm of odds of the outcome), and fast as well.

We investigated the effects of the proposed algorithm by comparing the results with the logistic regression algorithm as a baseline algorithm, as well as with the fairness-aware logistic regression from which we derived our method and that is most similar to our approach (Zafar et al., 2017, Zafar et al., 2019). The experiments are conducted on the Adult dataset (Kohavi, 1996). This dataset predicts the census level of an individual, and it is known for having gender discrimination. In addition, to show the effectiveness of the proposed approach, we provide results on the ProPublica COMPAS dataset (Dressel & Farid, 2018). More specifically, the COMPAS is software used in the United States to get a score for recidivism. The score is obtained as a model that is learned using prior examples. This software caused a lot of discussion regarding racial unfairness and interested readers are referred to (Washington, 2018).

The remainder of the paper is structured as follows. In Section 2 we provide a background where basic concepts of fairness and logistic regression are explained. In Section 3 we provide a related work that contains a review of the papers regarding fairness in predictive modeling. Section 4 will provide a methodology and experimental setup. In Section 5 we provide results and discussion of the results, while Section 6 concludes the paper.

2. Background

Creating an unfair decision using an algorithm can hurt the business and its performance. Many known applications of data mining and machine learning models were subject to the consequences due to unwanted discrimination. Examples of gender discrimination are seen mostly in the job hiring process and in advertisements. During the candidate screening, companies tend to use software that calculates the suitability for the job. Attributes such as job experience, education, or age that are often used for such tasks are gender biased. One such software platform, XING, even ranked less qualified male candidates higher in comparison with more qualified female candidates (Lahoti et al., 2019). In web advertisements, one could notice that after changing the gender in settings, male Google users receive more ads for higher-paying jobs compared to female Google users (Datta et al., 2015). However, the cost of being unfair can be social unrest. One such application is criminal risk assessment software. In the USA, there are several such software and they may result in unfair predictions. More specifically, African-Americans may have a higher probability of being imprisoned, compared to White-Caucasians (Dieterich et al., 2016). Tools such as COMPAS (Dieterich et al., 2016), PSA (Majdara & Nematollahi, 2008), and SAVRY (Meyers & Schmidt, 2008) are well known and subject of many discussions in the social sciences.

The source of unfairness can be in the data collection process, i.e., female participants are less often interviewed for the job position overall, thus leading the predictive model to believe that the male gender is more suitable for the job. Further, the source of unfairness can be the decision-maker itself. For example, the decision-maker can favour male candidates in the selection process for job openings. Then, the machine learning algorithm will reconstruct the decision-makers model and find that gender influenced the decision. Besides, predictive models are more likely to amplify the discrimination (Veale & Binns, 2017). Finally, the source of unfairness can be more subtle. Even if a sensitive attribute (e.g., gender, race) is omitted from the dataset, the existence of proxy attributes can create unwanted discrimination. From the managerial point of view, it is worth noticing that the responsibility for unwanted discrimination is on the decision-maker and that legal consequences can be very dire. It is worth noticing that, every additional construct (fairness as well) may and most probably will lead to lower accuracy. More specifically, there is a trade-off between fairness and predictive accuracy (Menon & Williamson., 2018).

After these examples of unwanted discriminations, many researchers from the industry started considering (un)fairness during the creation of the predictive models. As a result, several approaches are developed. However, prior to discussing the approaches, one needs to define and quantify fairness. In the literature, one can find two broad notions of fairness in algorithmic decision-making.

First, one can discuss being fair on the individual level, or *individual fairness*. A decision-making model is individually fair if similar individuals obtain similar results (Sharifi-Malvajerdi et al., 2019). More specifically, the predictive model will generate the same output if the same inputs are provided. Therefore, if two same candidates, one male, and another female, are evaluated for the job screening using the model that does not use gender as an input attribute, then both candidates will result in the same decision. In other words, gender will not influence the decision. This notion of fairness is omitted from this research. However, there are other definitions of individual fairness. One notion of individual fairness is adopted from the political philosophy and it is Rawls's theory of justice (Binns, 2018). This concept of fairness deals with the distribution of resources of the social goods and therefore its application is limited. This notion of fairness states that any inequality in the distribution of resources must benefit all participants, especially the least advantageous ones. This approach, called the Rawlsian approach, tries to give an advantage to participants that have the lowest score obtained from the predictive model. Optimizing for Rawlsian fairness requires the MAXIMIN goal function (Jung et al., 2019) or sequential optimization, i.e., leximax optimization (Chen & Hooker, 2020). Therefore, this kind of fairness is better suited for governmental machine learning models (Grace & Bamford, 2020).

Individual fairness is criticized for not regarding a group belonging. More specifically, one can be individually fair, but some groups of people can be discriminated, and consequently obtain lower prediction scores. This notion of fairness is called *group fairness*. More formally, group unfairness is defined as systematic discrimination of the algorithmic decision-making model based on an attribute that individuals cannot control (i.e., race or gender). Unfairness can have multiple origins (i.e., data collection process or decision-making process). Most often, discrimination is unintentional or accidental, and this type of unwanted discrimination is called disparate impact (DI). One can define disparate impact as a difference in outcomes between discriminated and privileged groups. This notion of fairness can be translated into a mathematical formula, as presented in equation (1):

$$DI = \frac{E(\hat{y}|s = 1)}{E(y|s = 0)} \tag{1}$$

where $\hat{y}$ presents the probability of an outcome of the predictive model and s is a sensitive group with value 1 presenting a discriminated group and value 0 presenting a privileged group. With the assumption that the privileged group has a higher expected outcome, this ratio will be lower than one. In order to say whether there is discrimination, one must use some threshold τ. Using the U.S. Equal Employment Opportunity Commission "80%-rule" (Biddle, 2006) this value should be between 0.8 and 1. In other words, allowable discrimination is 20% with the privileged group as a referent one.

However, adjusting data mining and machine learning algorithms to satisfy disparate impacts may lead to unexpected consequences (Radovanovic et al., 2020). It occurs that the discriminated group is still being discriminated in the decision-making. Although on average disparate impact is satisfied, even at a 100% rate, instances from the privileged group will be more present in lower ranks (i.e., have a higher probability for the desired outcome). For many applications of data mining and machine learning, this is unwanted discrimination. For example, in job screening where the top ten candidates are hired, it is more probable to hire all male instances than to achieve equal distribution of male and female instances. Therefore, we adopt another notion of fairness, called equal opportunity (EQ). Equal opportunity states that the difference in outcomes is the same for the desired outcome. The mathematical formulation is given in equation (2).

$$EQ = \frac{E(\hat{y}|y = 1, s = 1)}{E(y|y = 1, s = 0)} \tag{2}$$

where y presents the real outcome. Besides, if we introduce the assumption that the privileged group has a higher expected outcome, this ratio will be lower than one. Although there are no formal thresholds for unfairness, one can adopt the "80%-rule" as reasonable allowable discrimination.

3. Related Work

Due to the uncertainty of the source of the unfairness and efforts invested in the mitigation of fairness, there are three strategies to mitigate unfairness. For every approach, one needs to know the values of the sensitive attribute s. More specifically, it has been shown that if the predictive model is unaware of the sensitive attribute (fairness through unawareness), it still provides unfair results because proxy attributes exist (Chen et al., 2019). Strategies are 1) Pre-processing techniques, 2) In-processing techniques, and 3) Post-processing techniques for that purpose.

Pre-processing techniques are designed to prepare data to be fair. More specifically, data are transformed in such a manner that it is not possible to recognize the sensitive attribute and consequently one cannot learn an unfair model. The simplest approach is to remove the sensitive attribute and all attributes correlated with the sensitive attribute (Kamiran&Calders, 2009). This will yield better results in terms of fairness, but without guarantees of achieving fairness. The downsides of this approach are the loss of information regarding output in the data at hand. One can further remove correlated attributes. However, correlation is a linear function of dependency, and the interaction of attributes that may lead to unfairness could not be recognized. One can also try to assign weight to instances to get fairer prediction scores (Rancic et al., 2021). Another approach, called massaging (Kamiran&Calders, 2012) aims at changing the output values in such a manner that the discriminated group gets the desired outcome in situations where similar instances originate from the privileged group have desired outcome. One can find the transformation of data values such that the sensitive attribute cannot be identified (Feldman et al., 2015). This technique is called disparate impact remover. Disparate impact remover (DIR) changes the data in such a manner that repaired data cannot distinguish the value of the sensitive attribute. More specifically, for each attribute in the data set, instances from the sensitive group have their values increased (or decreased), while instances from the privileged group have their values decreased (or increased). In addition, a complete dataset can be transformed using matrix factorization techniques. One such approach (Samadi et al., 2018) developed a fair principal component analysis. Adjusting the data at hand might seem like an appropriate approach. However, with general data protection regulation (GDPR) one will violate the need for data accuracy. Therefore, the application of such approaches, although helps to solve unwanted discrimination, is questionable.

In-processing techniques present a set of approaches in which the learning algorithm is modified to simultaneously maximize predictive accuracy and maximize fairness. One can use constraint optimization for logistic regression, such as (Zafar et al., 2017) where the constraint is regarded as disparate impact. More specifically, the disparate impact is converted into two linear constraints. Thus, constraint gradient descent is used. Constraints are presented in equations (3) and (4).

$$\frac{1}{n}\sum_{i=1}^{n}(s_i - \bar{s})\theta^T X_i \leq c_1 \quad (3)$$

$$\frac{1}{n}\sum_{i=1}^{n}(s_i - \bar{s})\theta^T X_i \geq c_2 \quad (4)$$

where $\bar{s}$ presents the average value of the sensitive attribute, $\theta^T X_i$where θ^T presents transposed values of the logistic regression coefficients and X_i is the vector of input values for instance i. The difference $(s_i - \bar{s})$ can be either positive for the discriminated group (discriminated group $s_i = 1$), or negative for the privileged group (privileged group $s_i = 0$) and acts as a constant. Further, these values are multiplied with the intensity of predictions leading to a measure similar to disparate impact. Since some level of discrimination is allowed, parameters c_1 and $c_2$are introduced. However, if we introduce a reasonable assumption that the privileged group has a higher expected outcome, then one can remove constraint in equation (4). This model is further improved to satisfy equal odds (Radovanovic et al., 2020).

Besides having constraints, one can use regularization techniques. The regularization term is used to control the complexity of the model as an additional term in the goal function of the model. In this case, complexity presents the unfairness of the model. Therefore, one will sacrifice prediction accuracy to achieve better fairness. Prejudice removal regularization is presented in the paper (Kamishima et al., 2012). More specifically, the following regularization term (equation (5)) is used.

$$\sum_{i=1}^{n}\sum_{y\in\{0,1\}} p(y|x_i, s_i; \theta)\ln\frac{p(y|s_i)}{p(y)} \quad (5)$$

This regularization function penalizes equalized odds for each group of a sensitive attribute for each output using conditional entropy of an output. This function is non-convex but can be converted to convex using approximation. One novel approach that can be related to regularization tries to explore the relation between predictive performance (accuracy) and fairness in terms of a Pareto front. More specifically, instead of optimizing a goal function with a regularization term, one can tackle the fairness problem as a multi-goal optimization problem. This will yield a Pareto front from which one can inspect the trade-off between accuracy and fairness (Valdivia et al., 2021), or find the minimax solution (Martinez et al., 2020).

Finally, one can use post-processing techniques, or adjust predictions to be as fair as possible. One can analytically find the best possible decision threshold for fairness (Pleiss et al., 2017), or utilize linear programming for score adjustment (Hardt et al., 2016).

Regardless of the approach being used, one needs to be aware that fairness is not the same as social welfare. By adjusting the learning algorithm to be fair in predictions, one may achieve lower welfare for each individual, or group of people (Finocchiaro et al., 2021; Kasy & Abebe, 2021). Thus, fairness constraint should be designed in such a manner that it enables contribution to the social welfare and provides equal opportunity to every individual.

Our paper presents a combination of the two presented approaches. We choose to use a simple presentation of the unfairness. For that purpose, we use a linear function similar to the one defined in the (Zafar et al., 2017), but instead of using constraint optimization, we use regularization as in (Kamishima et al., 2012). The regularization term is more used in the machine learning community. The reason for such behaviour is that one will surely get some model, while constraint optimization can yield an unfeasible solution. As a measure of fairness, we are more closely related to social welfare rather than to disparate impact thus enabling an equal opportunity to get the desired outcome. Besides, it is shown that observing fairness using a disparate impact could result in unfair solutions in terms of equal opportunity to achieve the desired outcome (Radovanovic et al., 2020). Therefore, we adapt the learning procedure to ensure equal opportunity, which is different compared to other approaches in the literature.

4. Methodology

The methodology section consists of three parts. First, we discuss the implementation of fairness into a logistic regression, followed by a description of the data at hand, and finally an experimental setup.

4.1 Equal opportunity logistic regression

Logistic regression is one of the most popular classification algorithms in the area of data mining and machine learning (Wu et al., 2008). The model obtained from the logistic regression is interpretable and provides an explanation of the decision being made, thus it is suitable for the what-if analysis. (James et al., 2013)

The logistic regression algorithm minimizes the logistic loss function as presented in equation (6).

$$\min \mathrm{L}(y, \hat{y}) = -\frac{1}{n}\sum_{i=1}^{n}(y_i \log(\hat{y}_i) + (1-y)\log(1-\hat{y}_i) + \lambda R \tag{6}$$

where y presents the real value of the output, $\hat{y}$ presents the predicted value of the output, and λ presents the strength of the regularization. Predicted values are converted into a scale between zero and one using the sigmoid function. More specifically, $\hat{y}_i = \left(1 + e^{-\theta^T X_i}\right)^{-1}$. The optimization procedure modifies parameters θ that represent coefficients associated with input attributes X. There are a lot of regularization techniques developed, the most popular being lasso (or L1) and ridge (L2) regularizations. (James et al., 2013)

We introduce equal opportunity regularization through the function presented in equation (7).

$$R = \left(\frac{1}{sum(s)}\sum_{i=1}^{n} y_i s_i\,(\theta^T X_i) - \frac{1}{sum\,(1-s)}\sum_{i=1}^{n} y_i(1-s_i)(\theta^T X_i)\right)^2 \tag{7}$$

More specifically, we take into account only the desired outcome (using y whose values are zero and one) and calculate the difference in the expected outcome between the discriminated group and the privileged group. The first part of the function calculates the average intensity of the predictions for the discriminated group for instances that had the desired outcome, while the second part of the function calculates the average intensity of the predictions for the privileged group for instances that had the desired outcome. We wish this difference to be as small as possible. The quadratic term assures us that both sides of the discrimination are penalized.

It is worth noticing that the intensity of an outcome ($\theta^T X_i$) is used rather than the probability of the outcome ($\hat{y}$). The intensity of the outcome is a linear function that is perfectly correlated with the probability of the outcome, thus it can be used in the optimization procedure as a substitute for the probability. Further, we can reduce two sums into one with a simple transformation presented in equation (8).

$$R = \left(\frac{1}{n}\sum_{i=1}^{n} y_i(s_i - \bar{s})(\theta^T X_i)\right)^2 \tag{8}$$

Individuals from the privileged group have $(s_i - \bar{s})$ negative, thus contributing negatively to the regularization function. Likewise, individuals from the discriminated group have $(s_i - \bar{s})$ positive, thus contributing positively to the regularization function. If we assume that the privileged group has a higher expected probability of an outcome (thus higher intensity), then R is negative and the regularization function should penalize for unwanted discrimination. Even if that assumption is wrong, the quadratic term will penalize for a discrepancy in expected outcome scores.

This regularization term is defined in such a manner that instances with the desired outcome have similar values of an outcome, more specifically equal opportunity. More specifically, the fairness measure previously described will be in the fair region (between 0.8 and 1) if function R is minimized.

4.2 Data

Data used in this research are the Adult dataset (Kohavi, 1996). The Adult dataset is a binary classification problem, where one tries to predict whether an individual is above or below the census line in the USA; more specifically, if an individual receives a salary higher than $50K.

Each individual is described using personal details (e.g., relationship status, education level, etc.). The predictive model is used mainly for the insurance companies (for creating insurance packages), as well as tax administration (for inspection purposes).

The dataset consists of 32,561 instances and 13 attributes. These attributes present personal characteristics of an individual, such as age, work class, education level, occupation, relationship status, race, gender, capital gain, capital loss, working hours per week, and native country. Using dummy coding we converted all attributes into 44 numerical attributes in the same manner as in (Bellamy et al., 2018).

This dataset is known for having gender bias. By observing the data, more specifically the desired outcome, one can notice a large disparate impact between male and female individuals. Male individuals receive higher income leading to higher taxes paid, but also some benefits from the social system, as well as better insurance packages.

Another data used in this research is the ProPublica COMPAS dataset (Dressel & Farid, 2018). This dataset contains more than 7,000 people (explained using eight attributes) arrested in Broward Country, Florida. The data correspond to arrests between 2013 and 2014. The predictive task given in the COMPAS dataset is to predict whether an individual will commit a crime in the future.

Attributes that are used to predict future criminal acts are the age of an individual, juvenile felonies count, decile score, juvenile misdemeanor count, juvenile other felonies count, total prior offense, days of screening before the arrest, and charge degree.

However, the analysis of the COMPAS results showed that the accuracy of the model is higher for White-Caucasian compared to African-Americans. A more concerning fact is that the recidivism rate is nearly twice as high for African-Americans than for White-Caucasians. In addition, White-Caucasians are falsely marked to not re-offend at a rate of almost 50%, while African-Americans have a false-positive rate of 28%.

4.3 Experimental setup

Since our goal is to enforce equal opportunity into a logistic regression algorithm, we must measure two sets of measures; namely, predictive accuracy and fairness.

As a measure for prediction accuracy, we will use the area under the ROC curve (AUC). This measure of predictive accuracy is selected because it is decision threshold independent, or we can say that this measure is the general goodness of the binary classification model. It is derived from the Mann-Whitney U test, and it can be interpreted as the probability that a model discriminates desired and undesired outcomes. More specifically, it is a probability that the predicted value of a random positive instance ($y = 1$) is higher than the predicted value of a random negative instance ($y = 0$). Since values of AUC are interpreted as a probability, it ranges between zero and one, where one means perfect classification, while value 0.5 is interpreted as a random model. (Cortes &Mohri, 2004)

Measures of fairness are the disparate impact and equal opportunity. These measures are explained in the Background section. The disparate impact, as a measure, should be equal to one. In that case, there are no differences between the privileged and the discriminated groups. A value higher than one and lower than one suggests that unwanted discrimination exists. However, the disparate impact is not a linear, nor a symmetrical measure. For example, $DI = 2$ and $DI = 0.5$ present the same level of discrimination, but in the first case the difference between the perfect value is one, and in the latter case it is 0.5. This can be evaded by adapting the disparate impact as presented in equation (9).

$$DI = \min\left(\frac{E(p(y)|s=1)}{E(p(y)|s=0)}, \frac{E(p(y)|s=0)}{E(p(y)|s=1)}\right) \tag{9}$$

Similarly, equal opportunity is a measure of fairness that should be equal to one. Also, it is not a linear function, nor symmetrical. Therefore, it is adapted in the same manner as the disparate impact, as presented in equation (10).

$$EO = \min\left(\frac{E(p(y)|s=1,y=1)}{E(p(y)|s=0,y=1)}, \frac{E(p(y)|s=0,y=1)}{E(p(y)|s=1,y=1)}\right) \tag{10}$$

In addition to testing equal opportunity logistic regression (EO-LR) as explained in Section 4.1 with inner ten-fold cross-validation, we will use classical logistic regression (LR) as a baseline. This will show how big the trade-off between predictive accuracy and fairness is. It is expected to have a loss in predictive accuracy (Barocas & Selbst, 2016), but the satisfaction of fairness should be of higher concern for the decision-maker. We also show the results of the fairness constrained logistic regression (Zafar et al., 2017) where the disparate impact constraint parameter is set to 0.8 (and 1.25). The fairness constrained logistic regression (FC-LR) solves the disparate impact as a measure of fairness. As already discussed, this measure of fairness perhaps might still result in unfair results in terms of equal opportunity. Therefore, we expect an increase in fairness, both in the disparate impact and the equal opportunity but not as good improvement in the latter measure as the proposed approach. Also, we compare results with the preprocessing technique disparate impact remover combined with logistic regression (DIR-LR). The disparate impact remover corrects the data in total, so that none of the attributes can distinguish the value of the sensitive attribute.

The experiment is conducted using a ten-fold cross-validation. This means that the dataset is divided into ten random subsets. Then, logistic regression models are learned on nine subsets, while being evaluated against the remaining one. This procedure is repeated ten times, so that a different subset is used for evaluation. It is worth noticing that the sensitive attribute is not used during the model learning phase as an input attribute, but as a piece of additional information for calculation of the regularization term.

5. Results and Discussion

The results of the experiments are presented in Table 1. Values inside the table are presented in terms of average obtained values during ten folds with standard deviation. Due to related samples in the cross-validation, the standard deviation should not be interpreted as in statistical analysis, but as a measure of stability in performances. The best performances are presented in bold letters.

Table 1: Predictive and fairness performances on Adult dataset

Algorithm	AUC	DI	EO
LR	**0.8973 ± 0.0052**	0.3932 ± 0.0182	0.8190 ± 0.0390
FC-LR	0.8545 ± 0.0097	**0.8244 ± 0.0203**	0.9321 ± 0.0214
EO-LR	0.8794 ± 0.0050	0.6212 ± 0.0117	**0.9553 ± 0.0243**
DIR-LR	0.8972 ± 0.0051	0.5807 ± 0.0252	0.8568 ± 0.0364

As expected, the best performing algorithm in terms of predictive accuracy is classical logistic regression (LR). This is seen by observing the AUC, which is 0.8973. Similar performance is obtained with logistic regression with disparate impact remover (DIR-LR). This value of AUC is very good for the predictive model. Fairness constraint logistic regression (FC-LR) and our proposed algorithm (EO-LR) have a lower predictive performance. More specifically, EO-LR and FC-LR have a lower AUC by 2% and 4%, respectively. It is worth noticing that our approach is better compared to FC-LR in AUC. This is due to the different notions of fairness. While FC-LR tried to make a fair predictive performance for the whole group of instances(both those who did obtain the desired outcome and those who obtained the non-desired outcome), our approach focuses on getting the only equal opportunity. More specifically, our approach focuses only on instances that receive the desired outcome.

Also, based on the disparate impact as a measure of fairness the best performing algorithm is FC-LR. This result is expected because the FC-LR algorithm is created to satisfy disparate impact constraints. Although the parameter c_1 was set to 0.8, the average value of the disparate impact is slightly better (closer to 1). This is due to the difference in the training and test dataset. Therefore, this approach solves the problem of disparate impact. Our approach ranked second, with disparate impact 0.6212. This value is slightly under the U.S. Equal Employment Opportunity Commission's “80%-rule”, but our approach did not aim to improve disparate impact directly. It is worth noticing that data preprocessing for fairness did improve fairness without the cost of predictive performance. However, this level of fairness is still not at a satisfactory level.

Finally, it is shown that our approach did perform well and solved the fairness measure for which it was implemented. Based on the results for the equal opportunity fairness measure, our approach performed the best. The value is close to the perfect value, while FC-LR is ranked second with a slightly lower value, and classical logistic regression is third with greater distance compared to the first two.

This small reduction in the predictive performance indicates that it is possible to induce social sciences concepts (such as fairness) into a mathematical model of the learning algorithm and still obtain good predictive performance.

An additional level of discussion is obtained by comparing the coefficients of the logistic regression. More specifically, we compared the coefficients of classical logistic regression and our approach. The most interesting ones are presented in Table 2.

Table 2: Comparison of the logistic regression coefficients for Adult dataset

Attribute	LR	EO-LR
Work class Government	-0.0352	0.4118
Work class Private	0.0470	0.4943
Marital status Married	0.7486	-0.3791
Marital status Widowed	-0.1544	-0.6578
Occupation Sales	0.0843	-0.0032
Occupation Machine	-0.1097	0.0003
Occupation Cleaners	-0.1696	-0.0021

As observed, there are changes in the coefficients of the logistic regression. Some of the changes are even changes in the sign of the logistic regression (*Work class Government, Marital status Married, Occupation Sales*). Some of the changes are in the intensity, such as *Work class Private*. However, it is interesting to notice that big changes are observed in the attributes that might represent gender. If we check coefficients that represent occupation, we can see that coefficients of those attributes are reduced to almost zero values. This means that these coefficients identified gender. Therefore, they are reduced to the level that they do not influence the final decision.

The results obtained on the COMPAS dataset are explained in Table 3. Again, as expected logistic regression (LR) showed the best performance compared to fairness aware approaches. However, on this dataset, the difference in AUC is much lower. More specifically, the largest decrease is by 1%. The increase in DI and EO is much higher. Therefore, the cost of fairness on the COMPAS dataset is not high. FC-LR had the best value for the disparate impact, while the proposed approach EO-LR had the best equal opportunity.

Table 3: Predictive and fairness performances on COMPAS dataset

Algorithm	AUC	DI	EO
LR	**0.7438 ± 0.0092**	0.8160 ± 0.0180	0.8399 ± 0.0173
FC-LR	0.7400 ± 0.0094	**0.9155 ± 0.0176**	0.8946 ± 0.0172
EO-LR	0.7424 ± 0.0093	0.8604 ± 0.0177	**0.9191 ± 0.0172**
DIR-LR	0.7329 ± 0.0182	0.8873 ± 0.0360	0.9012 ± 0.0216

Since a small difference in predictive performance results in fair results, one can obtain it by small changes in coefficients of the logistic regression model. This is presented in Table 4. The biggest difference is obtained in juvenile felony count and juvenile other felonies count, where coefficient dropped from 0.1486 to 0.1275 and 0.1635 to 0.1573, respectively. These attributes are deemed as unfair due to unfair practice in the data collection process (i.e., arresting people) (Abrams et al., 2021; Beckman & Rodriguez, 2021). More specifically, individuals who are "young and black" are likely to be suspected and reported for crime or felony (Leiber & Johnson, 2008). In other words, due to cultural and historical biases toward African-Americans are more likely to be reported for felonies compared to White-Caucasians.

Table 4: Comparison of the logistic regression coefficients for COMPAS dataset

Attribute	LR	EO-LR
Age	-0.0284	-0.0319
Juvenile felony count	0.1486	0.1275
Decile score	0.1496	0.1388
Juvenile misdemeanor count	0.0035	0.0082
Juvenile other felonies count	0.1635	0.1573
Priors count	0.1169	0.1208
Days of screening before arrest	0.0016	0.0016
Charge degree	0.0700	0.0449

However, the proposed method has limitations. First, to find suitable regularization parameter λ, one needs to perform an inner (cross) validation. This may be time-consuming depending on the dataset size. Although this is the problem with all fairness mitigation approaches, the cost of fairness can be very high. For example, one should hinder predictive performance just to achieve a satisfactory level of fairness. In extreme cases, the cost of fairness could lead to egalitarian policies. More specifically, a satisfactory level of fairness is achieved with a simple model where all of the coefficients are equal to zero. This leads to every individual receiving the same outcome. The consequences of such a model can be very demotivating for an individual (Finocchiaro et al., 2021). For example, if everybody obtains the same resources, no matter how much effort is invested, why would an individual invest any effort? Another limitation of this approach might be if equal opportunity is not the fairness measure one would like to introduce. In that case, one can adopt a disparate impact as presented in (Zafar et al., 2017; Zafar et al., 2019), equal odds (Radovanovic et al., 2020), or custom regularization function.

Conclusion

With the rising problem of having prediction models that are both accurate and fair, there is a need to adjust the data mining and machine learning models. In this paper, we utilized a notion of fairness called equal opportunity. Equal opportunity aims at providing every individual, no matter what race or gender, the same rights and the same chance of getting the desired outcome. The problem of implementing such fairness metrics is a source of unfairness. More specifically, data have inherited bias and unwanted discrimination that a learning algorithm needs to find and eliminate. We solved this problem by adding a regularization term that solves equal opportunity. Proposed regularization presents an adaptation of the definition of equal opportunity, found in the fairness in algorithmic decision-making literature, to mathematical formulas in a linear form. Because of that, our model remains convex, and therefore easy to optimize.

Experiments are conducted on the Adult and COMPAS datasets. The Adult dataset is known for having gender bias. More specifically, male participants are privileged, thus getting the desired outcome more frequently. The COMPAS dataset is known for racial bias and injustices. We tested our approach against classical logistic regression and disparate impact constraint logistic regression, as well as with fair data preprocessing technique disparate impact remover. The result suggests our approach gets the best equal opportunity for the data at hand, without greater loss in the predictive accuracy. Although the disparate impact was not optimized directly, our approach did increase this measure as well. However, compared to the disparate impact constraint logistic regression, our approach did receive the lower disparate impact. This is because disparate impact constraint logistic regression is created and constrained to disparate impact. Because of that, disparate impact constraint logistic regression has the best score for disparate impact, but the lowest score on the predictive accuracy.

As part of future work, we would like to integrate both the disparate impact and the equal opportunity into one regularization term. Additionally, we would like to test the proposed regularization term in a more complex learning algorithm, such as support vector machines or deep neural networks. Finally, we would like to adapt the proposed method as a multi-objective optimization problem and find the best trade-off between accuracy and fairness using game theory, more specifically, using bargaining solutions. The decision-maker can choose between Nash, Kalai, or Kalai-Smorodinsky bargaining solutions. Regardless of the bargaining solution, this allows a shift from a utilitarian point of view on predictions to a more egalitarian point of view on predictions.

Acknowledgments

This paper is partially funded by the project ONR-N62909-19-1-2008 (Office of Naval Research): *Aggregating computational algorithms and human decision-making preferences in multi-agent settings*.

REFERENCES

[1] Abrams, L. S., Mizel, M. L., &Barnert, E. S. (2021). The Criminalization of Young Children and Overrepresentation of Black Youth in the Juvenile Justice System. *Race and Social Problems, 13*(1), 73-84.

[2] Barocas, S., &Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*, 671. DOI: 10.2139/ssrn.2477899

[3] Beckman, L., & Rodriguez, N. (2021). Race, Ethnicity, and Official Perceptions in the Juvenile Justice System: Extending the Role of Negative Attributional Stereotypes. *Criminal Justice and Behavior*, 00938548211004672.

[4] Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Nagar, S. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*.

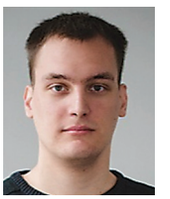
[5] Biddle, D. (2006). *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Gower Publishing, Ltd.
[6] Binns, R. (2018, January). Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency* (pp. 149-159).
[7] Chen, J., Kallus, N., Mao, X., Svacha, G., &Udell, M. (2019, January). Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 339-348). DOI: 10.1145/3287560.3287594
[8] Chen, V. X., & Hooker, J. N. (2020). Balancing Fairness and Efficiency in an Optimization Model. *arXiv preprint arXiv:2006.05963*.
[9] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017, August). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 797-806). DOI: 10.1145/3097983.3098095
[10] Cortes, C., &Mohri, M. (2004). AUC optimization vs. error rate minimization. In *Advances in Neural Information Processing Systems* (pp. 313-320).
[11] Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *Proceedings on Privacy Enhancing Technologies, 2015*(1), 92-112. DOI: 10.1515/popets-2015-0007
[12] Dieterich, W., Mendoza, C., & Brennan, T. (2016). COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*.
[13] Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances, 4*(1), eaao5580.
[14] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., &Venkatasubramanian, S. (2015, August). Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 259-268). DOI: 10.1145/2783258.2783311
[15] Finocchiaro, J., Maio, R., Monachou, F., Patro, G. K., Raghavan, M., Stoica, A. A., &Tsirtsis, S. (2021). Fairness and Discrimination in Mechanism Design and Machine Learning. In *Proceedings of the 2021 Conference on Fairness, Accountability, and Transparency*.
[16] Grace, J., & Bamford, R. (2020). 'AI Theory of Justice': Using Rawlsian Approaches to Legislate Better on Machine Learning in Government. *The Journal of the Society for Advanced Legal Studies*, 338.
[17] Hardt, M., Price, E., &Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems* (pp. 3315-3323).
[18] James, G., Witten, D., Hastie, T., &Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.
[19] Jung, C., Kannan, S., & Lutz, N. (2019). A center in your neighborhood: Fairness in facility location. *arXiv preprint arXiv:1908.09041*.
[20] Kamiran, F., &Calders, T. (2009, February). Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication* (pp. 1-6). IEEE.
[21] Kamiran, F., &Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems, 33*(1), 1-33. DOI: 10.1109/IC4.2009.4909197
[22] Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012, September). Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 35-50). Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-642-33486-3_3
[23] Kasy, M., & Abebe, R. (2021, March). Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 576-586).
[24] Kohavi, R. (1996, August). Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid. In *Knowledge Discovery in Data Conference* (Vol. 96, pp. 202-207).
[25] Lahoti, P., Gummadi, K. P., &Weikum, G. (2019, April). IFair: Learning individually fair data representations for algorithmic decision making. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)* (pp. 1334-1345). IEEE. DOI: 10.1109/ICDE.2019.00121
[26] Larose, D. & Larose, T. (2015). *Data mining and predictive analytics*. John Wiley & Sons.
[27] Leiber, M. J., & Johnson, J. D. (2008). Being young and black: What are their effects on juvenile justice decision making? *Crime & Delinquency, 54*(4), 560-581.
[28] Majdara, A., &Nematollahi, M. R. (2008). Development and application of a risk assessment tool. *Reliability Engineering & System Safety, 93*(8), 1130-1137. DOI: 10.1016/j.ress.2007.09.007
[29] Martinez, N., Bertran, M., &Sapiro, G. (2020, November). Minimax Pareto fairness: A multi objective perspective. In *International Conference on Machine Learning* (pp. 6755-6764). PMLR.
[30] Menon, A. K., & Williamson, R. C. (2018, January). The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency* (pp. 107-118).
[31] Meyers, J. R., & Schmidt, F. (2008). Predictive validity of the Structured Assessment for Violence Risk in Youth (SAVRY) with juvenile offenders. *Criminal Justice and Behavior, 35*(3), 344-355. DOI: 10.1177/0093854807311972

[32] Oneto, L., &Chiappa, S. (2020). Fairness in machine learning. In *Recent Trends in Learning From Data* (pp. 155-196). Springer, Cham. DOI: 10.1007/978-3-030-43883-8_7
[33] Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. In *Advances in Neural Information Processing Systems* (pp. 5680-5689).
[34] Radovanovic, S., Petrovic, A., Delibasic, B., &Suknovic, M. (2020, August). Enforcing fairness in logistic regression algorithm. In *2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)* (pp. 1-7). IEEE. DOI: 10.1109/INISTA49547.2020.9194676
[35] Rancic, S., Radovanovic, S., & Delibasic, B. (2021, May). Investigating Oversampling Techniques for Fair Machine Learning Models. In *International Conference on Decision Support System Technology* (pp. 110-123). Springer, Cham.
[36] Samadi, S., Tantipongpipat, U., Morgenstern, J. H., Singh, M., &Vempala, S. (2018). The price of fair PCA: One extra dimension. In *Advances in Neural Information Processing Systems* (pp. 10976-10987).
[37] Sharifi-Malvajerdi, S., Kearns, M., & Roth, A. (2019). Average Individual Fairness: Algorithms, Generalization and Experiments. In *Advances in Neural Information Processing Systems* (pp. 8240-8249).
[38] Valdivia, A., Sanchez-Monedero, J., & Casillas, J. (2021). How fair can we go in machine learning? Assessing the boundaries of accuracy and fairness. *International Journal of Intelligent Systems, 36*(4), 1619-1643. DOI: 10.1002/int.22354
[39] Veale, M., &Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society, 4*(2), 2053951717743530. DOI: 10.1177/2053951717743530
[40] Washington, A. L. (2018). How to argue with an algorithm: Lessons from the COMPAS-ProPublica debate. *Colorado Technology Law Journal, 17*, 131.
[41] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Zhou, Z. H. (2008). Top 10 Algorithms in Data Mining. *Knowledge and Information Systems, 14*(1), 1-37. DOI: 10.1007/s10115-007-0114-2
[42] Zafar, M. B., Valera, I., Gomez-Rodriguez, M., &Gummadi, K. P. (2019). Fairness Constraints: A Flexible Approach for Fair Classification. *Journal of Machine Learning Research, 20*(75), 1-42.
[43] Zafar, M. B., Valera, I., Rogriguez, M. G., &Gummadi, K. P. (2017, April). Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics* (pp. 962-970). PMLR.

Received: 2020-10-28
Revision requested: 2021-03-24
Revised: 2021-06-01 (2 revisions)
Accepted: 2021-06-21

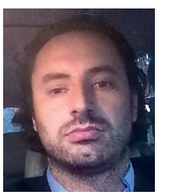
About the Authors

Sandro Radovanović
University of Belgrade, Faculty of Organizational Sciences, Serbia
sandro.radovanovic@fon.bg.ac.rs



Sandro Radovanović is a teaching assistant at the University of Belgrade, Faculty of Organizational Sciences. He received his PhD in Information Systems from the University of Belgrade, Faculty of Organizational Sciences. His research interests are Decision Theory, Algorithmic Decision Making, Machine Learning, Decision Support Systems, and Algorithmic Fairness.

Marko Ivić
University of Belgrade, Faculty of Organizational Sciences, Serbia
ivic.marko@gmail.com



Marko Ivić graduated at the University of Belgrade, Faculty of Civil Engineering. During his studies, in 2002, he established his own company for adaptation of office spaces, apartments and buildings. In 2005 he went to the Barcelona City Government for professional training at the Department for the organization of construction works for a one year professional training. His research interest lies at the intersection of civil engineering and decision support systems.