Increasing Fairness in Machine Learning

Zahra Khambaty - 260577706 - zahra.khambaty@mail.mcgill.ca
Ariane Schang - 260553025 - ariane.schang@mail.mcgill.ca
Kevin McGregor - 260564189 - kevin.mcgregor@mail.mcgill.ca

***Abstract*—Blind application of machine learning can have grave implications on the society. This is in context to the outcomes that in hindsight can be unfair toward certain groups. Our focus in this paper is to uncover the reasons behind these unjust outcomes and what techniques can be used to resolve/reduce this adversary as far as machine learning models are concerned. We also implement a new method based on random forests.**

I. Introduction

We humans make sense of the world by looking for patterns, filtering them through what we think we already know, and making decisions accordingly. When we talk about handing decisions off to artificial intelligence, we expect the same. Humans, unfortunately, are hindered with unconscious assumptions that unable us to filter irrelevancies while processing large amounts of information. The screening of job applicants, profiling of potential suspects by the police, and credit scoring are a few of the many fields where these irrelevancies can lead to poor decisions or bias. The question that needs to be addressed first is: what do we mean by bias in this context? What can we do to reduce it?

"The prejudice in favor of or against one thing, person or a group compared with another , in a way considered to be unfair" is defined as bias. There are many forms of biases or discrimination that are found in societies today. Gender biases, racial discrimination and stereotyping, for example, are some prevalent ones, which mostly stem from human conscience. Machine Learning is a tool with immense potential to help us avoid bias in hiring, operations, customer service, profiling and other broader business and social communities. This can be done by teaching predictive models to project decision-making processes objectively. What is interesting is the notion of how to train these models. Most algorithms that drive these models may not always reveal the objective truth just because theyre mathematical. Humans must feed it with relevant information that is free from bias, conscience or otherwise. Thats the only way machine learning can help create systems that are fair. The reality however is different; the historical data that will be fed may in fact be biased. We have reached a circularity whereby the aim of reducing bias through machine learning may crossfire. The primary target of this paper is to recognize the existence of such biases, analyze different scenarios where bias may occur and find techniques to break this cycle.

II. Related Works

A direct discrimination occurs when a person is treated less favourably or in a comparable situation on protected grounds. For example, property owners not renting to a minority racial tenant. An indirect discrimination occurs where an apparently neutral provision, criterion or practice would put persons of a protected class at a particular disadvantage compared with other. For example, a requirement to produce an ID in a form of drivers license for entering a club may discriminate visually impaired people, who cannot have a drivers license. A related term statistical discrimination [Arrow 1973] is often used in economic modelling. It refers to inequality between demographic groups occurring even when economic agents are rational and non-prejudiced.

In the context of machine learning non-discrimination can be defined as follows: people that are similar in terms of non-protected characteristics should receive similar predictions, and differences in predictions across groups of people can only be as large as justified by non-protected characteristics.

Now that we have our definitions clarified, the first question that arises is that is there a unique way to tackle each type of discrimination? Or is there a generalized way to deal with all biases?

Recent research by Tolga Bolukbasi, Kai-Wei Chang et al. in 2016 deduced gender bias through word embedding. They assigned each English word to a point in space. Words that are semantically related are assigned to points that are close together in space. They trained the system on Google News articles, and then asked it to complete a different analogy: Man is to Computer Programmer as Woman is to X. The answer came back: Homemaker.. It returned many common-sense analogies, like He is to Brother as She is to Sister. In analogy notation, which one may remember from school days, one can write this as he:brother::she:sister. But what it also gave out, were answers that reflected clear gender stereotypes, such as he:doctor::she:nurse and he:architect::she:interior designer . This particular event highlighted the weakness that falls prey to these blatant gender stereotypes. In order to overcome this, a debiasing system was used where real people were asked to identify examples of the types of connections that are appropriate (brother/sister, king/queen) and those that are not which were then removed. Using these human-generated distinctions, they quantified the degree to which gender was a factor in those word choices as opposed to, the more appropriate ones. Next the machine-learning algorithm removed the gender factor from the connections in the embedding. This removed the biased stereotypes without reducing the overall usefulness of the embedding. The results no longer exhibited blatant gender stereotypes, it has maybe uncovered a tool to represent gender, racial or cultural stereotypes.

Another interesting research was proposed by Indre Z Liobaite , Aalto University and Helsinki Institute for Information Technology HIIT [2016] who used four different measures:

Statistical, Absolute, Conditional and Structural measures to exhibit techniques for reducing bias in predictive models. As per historical data, statistical methods were used to identify discrimination, absolute and conditional methods were used to quantifying the extent of discrimination and structural methods were used to examine the spread of discrimination. He mentioned that discrimination can occur only when target variable is polar. That is, each task setting some outcomes should be considered superior to others. For example, getting a loan is better than not getting a loan, or getting an interest rate of 3% is better than 5%. According to this research if the target variable is not polar, there is no discrimination, because no treatment is superior or inferior to other treatment. They used mean difference, normalized difference, mutual information, impact ratio, elift and odds ratio as absolute measures. After analyzing extensively, they concluded that mean difference and area under curve can be directly used in regression tasks and focused more on classification scenarios.

The conclusion they reached was that the core measures stand-alone, are not enough for measuring fairness correctly. These measures can only be applied to uniform populations considering that every- body within the population is equally qualified to get a positive decision. In reality this is rarely the case, for example, different education levels may explain different salary levels. Therefore, the main principle of applying the core measures should be by first segmenting the population into more or less uniform segments and then assessing the bias.

We need to address the subtleties of these findings to a more deeply-rooted flaw that is human conscience. All data is undoubtedly the result of the human mind. If one wants to eliminate biases ultimately and completely then severe measures need to be taken to embrace its existence and spread awareness to consciously remove it.

Time constraints have however restricted from exploring the answers to questions such as why a person may feel discriminated against or what words in articles or any other data correspond to a certain form of bias.

III. Problem Representation and description of data

What we aim to achieve in this paper is to dig into previous literature in search of finding relevant and affective techniques of reducing bias in machine learning. We have, in particular, extended the topic of stigma attached to racial profiling and have used two different methods: logistic regression and random forests to justify the occurrence of bias and its minimization.

A. New York Police Department, Stop and Frisk Open Data

We begin by replicating a method from the existing literature, more specifically, from a 2017 paper ”Learning Classification without Disparate Mistreatment” [2]. This paper uses the idea of misclassification rates to formalize notions of unfairness. A classifier is said to suffer from ”disparate mistreatment” when the misclassification rates are different for groups of people with different values for a given sensitive attribute (race, gender, socioeconomic class, ability, etc.). The authors suggest that which misclassification rate to use should depend on the problem at hand. In the paper, they mention the NYPD Stop, Question and Frisk Program, in which people are stopped based on suspicion of illegal activity by police officers. This means that having different arrest or weapon discovery rates for different races would imply the existence of disparate mistreatment. While the paper applied this method on the ProPublica Recividism dataset, we implemented it with the Stop and Frisk dataset. A logistic regression with a modified loss function is used to minimize the disparity in misclassification rates, which will be explained in more depth in a later section.

The NYPD Stop and Frisk Data can be downloaded from the NYC.gov website [6]. The City of New York has made the data from 2003-2015 available to the public in an attempt to increase transparency about the program. We used the 2015 data. The dataset has 112 different attributes including: logistical information (location, time, officer identification), demographic information (suspects' height, weight, race, age, sex, etc.), information pertaining to the reason for the stop (crime suspected, carrying a suspicious object, wearing inappropriate attire, etc.), information about the stop itself (physical force used, period of stop, etc.) and the outcome of the stop (arrest made, offense arrested for, etc.).

We made a classifier which predicts whether an arrest will be made based on the information related to the reason for the stop. As the Zafar et. al paper suggests, if the misclassification rates differ for the different racial categories, our classifier is suffering from disparate mistreatment.

This left us with 37 binary features related to Yes or No questions such as ”ADDITIONAL CIRCUMSTANCES - TIME OF DAY FITS CRIME INCIDENCE” or ”REASON FOR STOP - WEARING CLOTHES COMMONLY USED IN A CRIME”. It also left us with 19 binary features concerning which crime was suspected (burglary, graffiti, etc.). The variable for the suspected crime was manually entered by the officer and therefore had over 2000 different values, such as ”FEL”, ”Felony”, ”MISD”, ”Misdemeanor”, etc. We brought this down to 19 binary features by grouping together all entries that included the most common codes for a crime. The target value was whether or not an arrest was made during the stop.

The demographic variable or ”sensitive attribute” used was the suspects' race. In order to understand and compare the mistreatment based on race, we had to choose only two values for the sensitive attribute. We therefore chose to compare the treatment of instances labeled as White and Black. The final dataset therefore included 14,464 instances. The dataset was divided into train and test sizes of 60% and 40%.

The main flaw with this data is that it was reported by the police officers involved in the program, so we must trust that they entered it correctly. We use this data and this classifier to see if there is racial bias in the automated prediction of people to be arrested.

B. NHIS data

We run both the logistic regression method and our newly developed weighted random forest method on data coming

from the National Health Interview Survey (NHIS) in the United States. The NHIS is an annual survey that collects household data on health and health-related behaviours [7]. The data are all publicly available, and the survey dates back to 1963. However, not all variables in the dataset are available in every year. Therefore, careful consideration must be made in the selection of variables included in the analysis.

We use five-year mortality as the variable to be predicted. This is measured as death within five years of the respondents' completion of the NHIS survey. Though there are many years with data available for the survey, we believe that taking a subset of only a few years (2000-2004) is appropriate for the scope of this project. We cannot take data beyond 2004, as mortality is only availabe up to 2009, thus preventing calculation of five-year mortality. The mortality predictors that we consider are: age, sex, marital status (never married vs. ever married), education (Less than high school, high school, college, bachelors, advanced degree), below poverty threshold (yes/no), bed-days in the past year (None, 1-7 days, 8-30 days, 31-180 days, 181-365 days), body mass index (BMI), alchohol days per week (none, some days, most days), and smoking (current, former, never). For the sake of simplicity, only complete-cases were included in our analysis, and we acknowledge that the deletion of missing observations could induce bias in our predictions. The final dataset contained 86,116 observations. The demographic variable for which prediction equality is measured is race, categorized into white vs. non-white.

The motivation for this analysis came from insurance. If a life insurance company wanted to make a decision on whether or not to accept an individual's insurance application, then the main outcome of interest would be death within the period of coverage (i.e. whether or not a payout must be made). However, if a prediction algorithm did not take a person's race into account when predicting mortality, non-white individuals, having a generally higher death rate overall, would be less likely to be insured. Even an analysis that includes race as a predictor could result in inequality of insurance coverage, as the distributions of the other predictors could also vary with respect to race. The methods presented in this report are thus good candidates for solving the problem of inequality in insurance coverage with respect to race.

IV. Methodology

In this section we precisely define the various measures of inequality in machine learning, and present how these measures can be taken into account in two kinds of algorithms: logistic regression and random forests.

A. Logistic regression

The implementation of the Logistic Regression as Zafar et. al describe it was released to the public through an open-source Github repository with a GNU public license [9]. We use this implementation and modify it to work with both the NYPD data and the NHIS data. The open-source implementation works to minimize disparity in False Positive rates, which we use on the NYPD data. A False Positive on this data would imply the arrest of an innocent person, which is the most unfavorable scenario. We further modify the implementation for the NHIS data to minimize disparity in True Positive rates, which correspond to being marked as surviving and therefore, in the insurance coverage problem, being covered by insurance.

Zafar's implementation minimizes misclassification disparities in classifiers that use decision boundaries. A formulation of the loss function with a constraint on the disparity on the False Positive Rate is:

$$\begin{aligned}&\text{minimize } L(\theta)\\ &\quad\text{subject to:}\\ &P(\hat{Y} \neq y | A = 0, y \quad = -1) - P(\hat{Y} \neq y | A \quad = 1, y = -1) \leq \epsilon\\ &P(\hat{Y} \neq y | A = 0, y \quad = -1) - P(\hat{Y} \neq y | A \quad = 1, y == -1) \geq -\epsilon,\end{aligned}$$

where A is the value of the sensitive attribute and ϵ is the measure of disparity. Fairer classifiers have low $|\epsilon|$ values. To account for the fact that these constraints are not convex, the disparate mistreatment can be approximated through the covariance between the sensitive attributes and the "signed distance between the feature vectors of misclassified users and the decision boundary". The lower the mistreatment, the closer to zero the covariance will be. The authors then convert this to a Disciplined Convex-Concave Program, which was introduced by X. Shen in 2016 [8], and is a new method of solving non-convex problems as long as the objective and constraint functions are a sum of convex and concave terms. This is implemented in python and extends CVXPY (a python embedded module for convex optimization problems).

B. Weighted random forest

A random forest is a very simple machine learning technique that fits a large number of decision trees based on different subsets of observations and features. It is one of the prime examples of *bagging*. The random forest proceeds as follows (as outlined in the course notes):

1) Take a sample of the training data (with replacement)
2) Build a decision tree on this data subset by using a randomly chosen subset of the features at each node
3) Repeat steps (1) and (2) K times.

Predictions can be obtained from each tree for each observation, and the final prediction for a given observation is chosen to be the one that is most highly represented among all the trees.

We develop a new algorithm based on a simplified version of the random forest. Firstly, rather than sampling a subset features at each node in a tree, we only subset features at the root node of the tree. These features are the only ones that will be considered in that tree. This simplification was made more for convenience in implementation than anything else. However, the main addition to the random forest that we propose is to weight the predictions of each decision tree based on some metric of inequality. Trees that lend themselves to less disparate predictions with respect to some demographic variable could be weighted higher, and those that are more disparate could be down-weighted.

In our algorithm, the main inequality metric we consider is *equal opportunity* [10]. Equal opportunity means that, if an example truly belongs to the more favourable class $Y = 1$ (i.e. surviving 5 years), then the probability it will be predicted as such does not depend on the demographic variable $A \in \{0, 1\}$:

$$P(\hat{Y} = 1|A = 1, Y = 1) = P(\hat{Y} = 1|A = 0, Y = 1).$$

This is equivalent to ensuring that the sensitivity of the model does not change between the two demographic groups. This measure ignores imbablance in the probability of predictions when the individual's true class is $Y = 0$. A measure of inequal opportunity is the difference in these two probabilities:

$$d = P(\hat{Y} = 1|A = 1, Y = 1) - P(\hat{Y} = 1|A = 0, Y = 1).$$

A higher value of d signifies more disparity among the two demographic groups. Some function of d can be used to define the weight w_t for each tree $t \in \{1, \ldots, T\}$. The function we have chosen is the following:

$$w_i = \frac{1}{|\hat{d}_t|} = \frac{1}{|\hat{P}(\hat{Y}_t = 1|A = 1, Y = 1) - \hat{P}(\hat{Y}_t = 1|A = 0, Y = 1)|},$$

where $\hat{d}_t$ is the estimated value of d based on the predicted values in tree t. Each weight could be scaled down by the sum of the total weights such that the resulting weights sum to one. Additionally, a maximum weight must be set to prevent any weight from getting too large (i.e. if equal opportunity holds in a tree t, then w_t is undefined). We set the maximum weight to be 1000. The final prediction for individual $i \in \{1, \ldots, N\}$ is then obtained as:

$$p_i = \mathbb{1}\left\{\sum_{t=1}^{T} w_i p_{it} > \sum_{t=1}^{T} w_i(1 - p_{it})\right\}$$

where p_{it} is the prediction for individual i from tree t. We believe this weighting scheme will lead to a forest with both a high accuracy and better equal opportunity than would be seen in a standard random forest.

We implement the weighted random forest with help from the R package `rpart`, which fits a single decision tree. This package automatically prunes the tree based on a user-specified cost parameter.

C. Simulation study for random forest

We perform a simple simulation study to investigate whether the weighted random forest reduces inequal opportunity. The simulation is replicated 100 times. The total sample size is 5000, which is split in half to produce a training and test set. We generate 20 features from a multivariate normal distribution, with correlation between all features set to 0.1. The features are generated in a way such that a higher probability of $Y = 1$ is obtained for observations coming from the demographic group $A = 1$. The coefficients in the linear part of the model are generated from a normal distribution. The resulting linear combination of features is scaled so that, when the inverse logistic transform is applied, all resulting probabilities fall between 0.01 and 0.99. Finally, the binary class variable is generated using a binomial distribution with the aforementioned probabilities.

The data from each simulation replication are passed into both a standard random forest and the weighted random forest. For each replication, we run forests with different numbers of trees (50, 100, 250, 500, 1000), and different numbers of feature subsets (5, 10, 15). Performance is evaluated by checking whether inequal opportunity (i.e. disparity in sensitivity) has been resolved.

D. NHIS analysis

For the NHIS analysis using the random forest, we fit 100 trees in each of the weighted and unweighted forests. Since death over only a five year period is a rare event, the decision threshold for each tree fit in a random forest has to be set quite high. Therefore, in a given leaf node, a death is predicted if 5% or more of the individuals in that node had died after five years. Due to the nature of the random forest, feature selection is not necessary, as the repeated boostrap sampling will decide which features are important in predicting mortality. The trees are allowed to have different depths. There is only one parameter in the algorithm that governs the total number of levels possible for a single tree. This parameter is set to 30, which is the default in the `rpart` package.

The data are split into a training set (size 60,000) and a test set (size 26,116). The trees and weights are formed in the training set, and the results are evaluated by applying the predictive model on the test set. Results are outlined in the next section.

V. Empirical results

A. NYPD Stop, Question and Frisk analysis using logistic regression

The logistic regression solution proposed by Zafar et al. worked well on the Stop, Question and Frisk Data. While the data is unbalanced with a majority of stops ending in no arrest, the classifier was still able to predict a correct classification of "arrest" with a recall of 48% and precision of 72% and accuracy 88.61%. The constrained classifier works similarly, with recall 46%, precision 73% and accuracy 88.59%. Full results can be found in the classification report Table I.

The code that led to these results and the plot was used by extensively modifying the open-source code from Zafar et al in order to work with the Stop and Frisk dataset.

Target Value	P	R	P_constrained	R_constrained
No Arrest	0.91	0.97	0.90	0.97
Arrest	0.72	0.48	0.73	0.47

TABLE I
PRECISION AND RECALL RESULTS FOR THE NORMAL LOGISTIC REGRESSION AND THE CONSTRAINED LOGISTIC REGRESSION.

As can be seen in 1, the False Positive Rate gradually equalizes as the covariance threshold is decreased. At a value of 0, we see a perfect removal of disparate mistreatment with 3.38% FPR. The unconstrained classifier had a disparity of

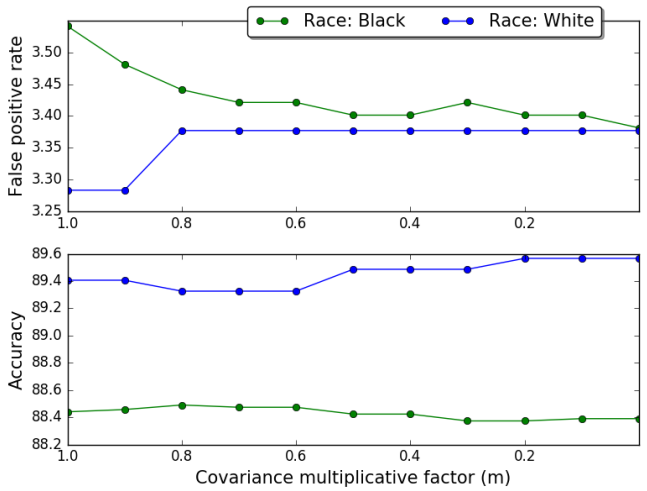

Fig. 1. As the covariance threshold decreases, we see an equalizing in the False Positive Rates for the classes of both the sensitive attributes. We also see a slight change in accuracy over the thresholds.

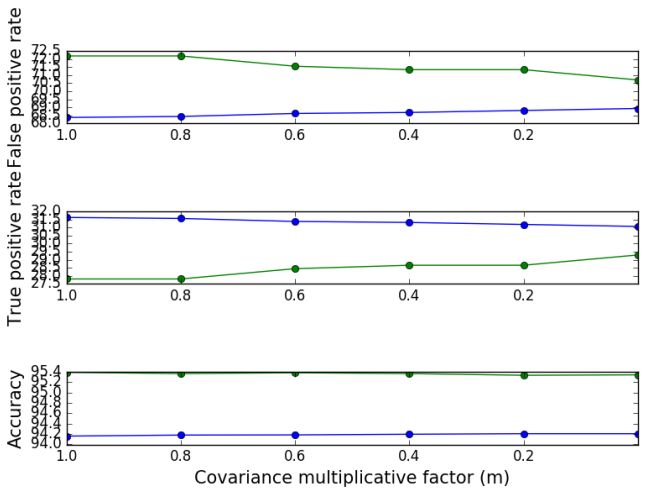

Fig. 2. Variance of FPR, TPR and Accuracy as the covariance threshold decreases. Green corresponds to Non-Whites and Blue corresponds to Whites

0.26%. Removing this disparity completely was possible with very little change in accuracy.

B. NHIS analysis using logistic regression

We further used Zafar's logistic regression implementation to analyze the NHIS dataset. For this separate problem, we needed to minimize disparity in True Positive Rates for a survival label, or maximize Equal Opportunity. Because the True Positive Rate can be redefined as $TPR = 1 - FPR$, we can redefine this problem as minimizing the discrepancy in False Positive Rates in order to make it compatible with the program used for the Stop and Frisk Dataset. If the discrepancy between FPRs is minimized, then the discrepancy between TPRs is equally minimized.

As can be seen in Table II, the performance of the classifier does not change whether it is constrained or unconstrained. Considering the unbalanced nature of the dataset, there is little surprise that the rates of predicting death are quite low. The accuracy over the data is of 95.25% with the unconstrained classifier, and 95.13%with the constraints.

Predicted Class	P	R	P_constrained	R_constrained
Survive	0.97	0.98	0.97	0.98
Death	0.41	0.31	0.41	0.31

TABLE II
PRECISION AND RECALL RESULTS FOR THE NORMAL LOGISTIC REGRESSION AND THE CONSTRAINED LOGISTIC REGRESSION OVER THE NHIS DATASET.

With the change in the covariance threshold, the True Positive and False Positive Rates do move towards equality (see Fig 2). However, they do not ever quite reach perfect equality. This implies that, over the NHIS dataset and likely over many other datasets, the disparate mistreatment can be minimized but never truly removed. Originally, the FPR for Non-Whites is 72.19%, while the FPR for Whites is 68.32. This is minimized to 70.7% to 68.94%. This minimization of disparity happens with very little loss in accuracy.

C. NHIS analysis using random forest

Sensitivity and specificity for the random forest implementation in the NHIS data analysis can be seen in Table III (unweighted forest) and Table IV (weighted forest). These results correspond to applying the fitted model on the **test** data. Here, the outcome $Y = 1$ corresponds to survival after 5 years, and $Y = 0$ corresponds to death within 5 years.

In the unweighted forest, there is a definite discrepancy in the sensitivity between whites (A=1) and non-whites (A=1). Unexpectedly, the algorithm is more likely to correctly predict survival after 5 years in the non-white group. However, it is also better at predicting death (specificity measure) in the white group.

In the weighted forest, there is an excellent improvement in equal opportunity. In this method, the sensitivity is around 92% in both demographic groups. However, this comes at a cost: the specificity in both groups has dropped significantly. Therefore, in the context of mortality prediction for insurance, this would mean more people would be insured that would end up dying during their coverage period.

	Sensitivity	Specificity
Non-white	0.8744	0.5793
White	0.8069	0.7746

TABLE III
SENSITIVITY AND SPECIFICITY IN THE TWO DEMOGRAPHIC GROUPS USING THE UNWEIGHTED (STANDARD) RANDOM FOREST ON THE NHIS DATA.

D. Random forest simulation results

Here we present the results of the simulation study. All results are obtained from the **test** data. Table V shows the average discrepancy (over all simulation replications) in equal

	Sensitivity	Specificity
Non-white	0.9222	0.1683
White	0.9277	0.2037

TABLE IV
SENSITIVITY AND SPECIFICITY IN THE TWO DEMOGRAPHIC GROUPS USING THE WEIGHTED RANDOM FOREST ON THE NHIS DATA.

opportunity when using the weighted and unweighted random forests with the number of trees ranging from 50 to 1000. Here, the number of features considered in each tree is 10. In all cases, the equal opportunity discrepancy is reduced by around 10 percentage points. Interestingly, this metric does not change much as the number of trees is increased. Table VI shows the results over different numbers of features considered in each tree, with the total number of trees fixed at 500. Once again, there is a favourable reduction in equal opportunity discrepancy, which does not appear to differ much over the number of features.

# trees	$\hat{d}_{rf}$	$\hat{d}_w$	$\hat{d}_{rf} - \hat{d}_w$
50	0.2217	0.1194	0.1023
100	0.2230	0.1128	0.1103
250	0.2041	0.1002	0.1039
500	0.2241	0.1229	0.1011
1000	0.2025	0.1059	0.0965

TABLE V
SIMULATION RESULT OVER DIFFERENT NUMBERS OF TREES: THE AVERAGE OF THE OBSERVED DISCREPANCIES IN EQUAL OPPORTUNITY USING THE UNWEIGHTED ($\hat{d}_{rf}$) AND WEIGHTED ($\hat{d}_w$) RANDOM FORESTS. THE RIGHTMOST COLUMN GIVES THE AVERAGE REDUCTION IN DISCREPANCY OBTAINED WHEN USING THE WEIGHTED FOREST.

# features	$\hat{d}_{rf}$	$\hat{d}_w$	$\hat{d}_{rf} - \hat{d}_w$
5	0.2106	0.0743	0.1023
10	0.2241	0.1229	0.1103
15	0.2387	0.1562	0.1039

TABLE VI
SIMULATION RESULT OVER DIFFERENT NUMBERS OF FEATURES: THE AVERAGE OF THE OBSERVED DISCREPANCIES IN EQUAL OPPORTUNITY USING THE UNWEIGHTED ($\hat{d}_{rf}$) AND WEIGHTED ($\hat{d}_w$) RANDOM FORESTS. THE RIGHTMOST COLUMN GIVES THE AVERAGE REDUCTION IN DISCREPANCY OBTAINED WHEN USING THE WEIGHTED FOREST.

The distribution of the disparity measure $\hat{d}$ for the weighted and standard (unweighted) random forests can be seen in Figure 3. This corresponds to a forest of 500 trees, with 10 features (out of 20 possible) considered in each tree. This further illustrates that the weighted forest is bringing the predictions closer to fulfillment of the equal opportunity criterion.

One concern is whether forcing the prediction probabilities to be equal among the demographic groups causes a drop in overall predictive accuracy. Table VII shows the overall sensitivity and specificity of the models run in Table V. Interestingly, the overall sensitivity actually increases in the weighted random forest. Admittedly, the increase is so small that this could be a statistical anomaly. There is, however, a decrease in overall specificity. We predict more positive outcomes in the demographic group $A = 0$ in order to make

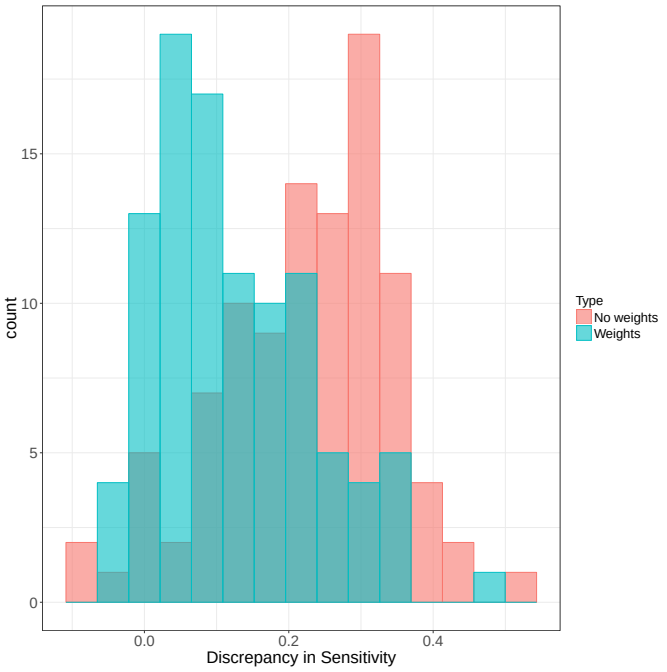

Fig. 3. Results for standard and weighted random forests on 100 simulation replications. The horizonal axis represents discrepancy in equal opportunity (sensitivity) between the two demographic groups. It is clear that the weighted forest has lower discrepancy, on average.

things more fair, which likely leads to an increase in false positives.

Metric	Unweighted tree	Weighted tree
Sensitivity	0.6751	0.6770
Specificity	0.6757	0.6451

TABLE VII
AVERAGE SENSITIVITY AND SPECIFICITY OVER ALL SIMULATIONS AND ALL MODELS OUTLINED IN TABLE V.

VI. DISCUSSION

The logistical regression that accounts for disparate mistreatment, as laid out by Zafar et. al in their 2017 paper was a relatively successful way of minimizing classification disparities between two groups. One of the main difficulties with this method is that you may only constrain on one classification rate at a time. Therefore, it will be very difficult to classify with the goal of seeing equal rates across the board. When there is one particular scenario in which an unequal classification rate has serious consequences (e.g. recidivism, arrest, etc.), it is definitely worth it to use this method to make sure the predictions are not imbalanced against a certain group.

The unbalanced nature of the NHIS classes (survival, death) also made it difficult for the Logistic Regression to work very well at all. It was difficult for the LR to predict the deaths from the NHIS data. Despite its' success on the Stop and Frisk dataset, the constrained logistic regression was unable to completely remove all signs of disparate mistreatment in the True Positive Rate of the NHIS data. This is an indication that bias is difficult to fully remove from a classifier, even if

it can be minimized. Implementing a random forest classifier was a good way to find a different way of approaching this dataset.

This implementation also only works with decision-boundary based classifiers, such as SVMs or logistical regression. For further work, researchers should certainly investigate the possibility of implementing the minimization of disparate mistreatment in LSTMs, Neural Networks, etc.

For the weighted random forest, it is clear that the weighting scheme contributes to a lower discrepancy in equal opportunity. Though the overall predictive accuracy was not affected in the simulation study, there was a definite drop in the ability to predict a death in the NHIS analysis. Though equal opportunity was almost exactly achieved in the NHIS analyis, the observed loss predictive accuracy might not be worth it. It should be noted that this trade-off is very much context dependent. For an insurance company looking to predict death, this could be catastrophic. But perhaps there are other contexts where equality would be a higher priority.

One limitation of the weighted random forest is that we only considered one equality metric: equal opportunity. There are different criteria that also consider predictions for individuals truly beloning to class $Y = 0$. It would be useful to design a version of the forest that takes other kinds of equality metrics into account. Admittedly, the weighting scheme might have to be rethought for these new criteria.

We also technically did not use a proper random forest in this paper. We only sampled the features once for each tree, and then built the decision tree as if those were the only features available. It would be useful to adhere to the standard formulation of a random forest, where features are sampled at each node, and the tree has a fixed depth.

Finally, it would be useful to run the weighted forest on a wider variety of datasets. Perhaps other datasets could be found that contain more features (the number of features in the NHIS dataset was small due to the complexity of data availability and data cleaning). Also, it would be useful to try on a dataset that had more balanced classes than in the NHIS data.

VII. Conclusion

There are many ways that machine learning methods can lead to decisions that result in inequality. Though there are methods out there that can help reduce these kinds of inequalities, they rarely eliminate them completely. Additionally, forcing predictive algorithms to adhere to criteria such as equal opportunity is without a doubt going to lead to poorer overall predictive accuracy. We believe, however, that these are small prices to pay for the sake of fairness in machine learning.

VIII. Statement of contributions

- **Ariane**: Modified Zafar et. al's implementation of reducing disparate mistreatment and implemented it on the Stop and Frisk and NHIS datasets.
- **Zahra**: Did background research into existing machine learning methods and situations where inequalities can arise in machine learning. Wrote the introduction and related works sections.
- **Kevin**: Designed the weighted random forest method and performed the simulation study. Also did data collection and cleaning for the NHIS data.

All authors contributed to writing the manuscript.

We hereby state that all the work presented in this report is that of the authors.

References

[1] Louise AC Millard, Peter A Flach, and Julian PT Higgins. Machine learning to assist risk-of-bias assessments in systematic reviews

[2] Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness Beyond Disparate Treatment and Disparate Impact: Learning Classification without Disparate Mistreatment. arXiv:1610.08452 [Cs, Stat], 2017, 117180. doi:10.1145/3038912.3052660.

[3] INDRE Z LIOBAITE. A survey on measuring indirect discrimination in machine learning. https://arxiv.org/pdf/1607.06520.pdf.

[4] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai. Debiasing Word Embeddings. https://arxiv.org/pdf/1607.06520.pdf.

[5] IEEE Transactions LaTeXand Microsoft Word Style Files. http://www.ieee.org/web/publications/authors/transjnl/index.html

[6] https://www.nyc.gov/html/nypd/html/analysis_and_planning/stop_question_and_frisk_

[7] https://www.ipums.org/healthsurveys.shtml

[8] X. Shen, S. Diamond, Y. Gu, and S. Boyd. Disciplined Convex-Concave Programming. arXiv:1604.02639, 2016.

[9] Github. Mbilalzafar/Fair-Classification. Accessed April 17, 2017. https://github.com/mbilalzafar/fair-classification.

[10] Hardt, Price, Srebro *Equality of Opportunity in Supervised Learning*. Advances in Neural Information Processing Systems, 3315–3323, 2016.