Systemic Data Discrimination and Equality of Opportunity

Saad Padela

April 15, 2018

Introduction

Public administrators and policymakers seeking objectivity have long equated it with the data-driven decision-making process. The advent of machine learning has extended this promise to the nth degree – whereas traditional 19th-century statistics enabled skilled practitioners to draw impartial conclusions from carefully collected data, the black-box deep-learning techniques of today can be operated effectively by engineers and developers without making human interpretations. As the data grows in volume, variety, and velocity, and as the techniques used to harness it become ever more sophisticated, this new philosophy of "big data positivism" has promised an end to long-standing problems of subjectivity and discrimination.

In delegating these decisions with social justice implications to our data and our algorithms, we have indeed shut down various avenues of discrimination; however, as this paper will argue, in many cases we have merely re-routed our biases through more subtle channels. As it turns out, data is socially constructed, and inherits our human imperfections and biases with startling fidelity. So too do algorithms trained on these biased datasets, and the effects are very difficult to detect. Instead of curbing the potential for systemic discrimination against disadvantaged groups, many researchers believe that the use of algorithms has actually expanded it.

Eliminating bias and subjectivity within an algorithm is not always possible (e.g., deep-learning with neural nets); however, researchers can take explicit steps to calibrate decision-rules based on the outputs from these models. In this paper we discuss one such approach towards remediation within a supervised learning context, called the Equality of Opportunity criteria. Furthermore, we illustrate its utility by applying it to a well-known example of a racially discriminatory model, the COMPAS Recidivism Risk Score.

Understanding Systemic Data Discrimination

In recent years, organizations in both the public and private sphere have made widespread use of predictive analytics and machine learning for use cases such as college admissions, loan applications, airport screenings, and of course, advertising. These applications not only drive speed and efficiency, but there is an underlying assumption that decisions with social justice implications are best made by data-driven algorithms, because they are inherently impartial.

On the contrary, a growing number of academic researchers, data journalists, and government officials fear that algorithms have opened new frontiers for systemic data discrimination. For example, merely including zip codes in a predictive model can discriminate against different racial and ethnic groups [Noy15]. In 2014, the Federal Trade Commission hosted a workshop titled “Big Data: A Tool for Inclusion or Exclusion?” Also in May 2014, the White House released a report titled “Big Data: Seizing Opportunities, Preserving Values” which sought to kindle a nationwide discussion on how to reap benefits whilst minimizing potential harms of big data [PPM+14].

Systemic data discrimination occurs when members of a particular group are consistently at a disadvantage vis-à-vis decisions made by a predictive model, and these decisions have a meaningful impact on their financial, social, or emotional well-being. Crucially, our definition of systemic discrimination is not concerned with whether a correlation actually exists between the target variable and membership in a class. In other words, being "correct" or "accurate" does not impact the culpability of a discriminatory model.

A social utilitarian would likely object to this definition on philosophical grounds. After all, if a "protected" attribute is truly correlated with an important outcome, then by definition, society at large is worse-off for not being able to take advantage of this information. Discrimination theorists counter with three types of reasons why an outcome might be correlated with a population segment: (1) Path-dependency stemming from historical imbalances in opportunity for various demographic segments; (2) Variances in signal quality among different population groups; (3) Insensitivity or mis-calibration of the model itself due to ignorance or oversight by researchers.

Historical decisions have the potential to create discriminatory artifacts in the data. For example, consider the various circumstances under which Black and Hispanic children would have had different levels of access to education and employment opportunities compared to Whites in 19th and 20th century America. Now imagine a predictive model that predicts your suitability for an advanced degree program based in part on the education-level of your parents, and their job classification. A naïve researcher may hypothesize that, for example, students with parents who hold PhDs and work at major research institutions may have an edge in terms of culture fit over those who have no insight into academia. Upon testing this hypothesis, she may

indeed find the results to be significant, but if she is not controlling for race, she will merely be picking up on its impact, which in turn was caused by an opportunity gap. By reading into these differences and using them to make important decisions, she would be merely perpetuating the cycle of unfairness.

For a real-world example, consider the case of St. George's Hospital in the UK. The hospital developed an algorithm to sort medical school applicants based on previous admissions decisions [BS16]. It turned out that previous decisions had systematically excluded racial minorities and women, and these cases were not removed from the training dataset. In drawing on these previous decisions for their algorithm, St. George's hospital merely automated the practice of institutional racism and gender bias.

Overt bias in the training data represents one method of contamination; an altogether different kind of peril relates to variances in signal quality coming from different population segments. For example, Kate Crawford writes about "data dark zones" that skew results from a sentiment analysis of Twitter during the Australian Open, or findings from the Boston Street Bump app [Cra13]. In a similar vein, Jonas Lerman points to "the nonrandom, systemic omission of people who live on big data's margins, whether due to poverty, geography, or lifestyle, and whose lives are less 'datafied' than the general population's" [Ler13].

Extending beyond the purely digital context, Solon Barocas theorizes that historically disadvantaged groups are more susceptible to errors associated with data gathering because they "have unequal access to and relatively less fluency in the technology necessary to engage online, or are less profitable customers or important constituents and therefore less interesting as targets of observation" [BS16]. Consequently, not only are individual data records of a poorer

quality, but the representativeness of any sample is similarly subject to potentially significant distortions.

Having considered that discrimination may reside firstly with the data itself, or secondly in its collection, we turn to the third and final scenario where bias is introduced during the modelling stage. According to Barocas, there are two major mechanisms at this stage: target variable specification and feature selection [BS16]. Bias in the former is characterized by subjective framing of the problem that may be inconsistently suitable for a minority class, whereas in the latter it may be associated with improper labeling of data based on misinterpretation. The net result is that protected classes may be "subject to systematically less accurate classifications or predictions because the details necessary to achieve equally accurate determinations reside at a level of granularity and coverage that the selected features fail to achieve" [BS16].

Strategies for Reforming Biased Predictors

As the theoretical framework discussed above would suggest, merely omitting race from a model specification will not guard against the potential for systemic discrimination. In 2008, Pedreshi, Ruggieri, and Turini showed this to be the case with a case study of German credit data [PDR08]. This is because of "redundant encodings" that allow race to influence the model just as effectively as if it were included as an explicit variable. Even seemingly innocuous variables need to be vetted for hidden correlations with protected attributes; if they are found, then the researcher will need to make efforts to offset the resulting biases.

One common strategy for dealing with biased classifiers is to establish different threshold-criteria for various disadvantaged groups. For example, a bank granting loan applications on the basis of a predictive model may notice that Hispanic applicants have an average score of 45 (on a hypothetical scale of 1-100), whereas White applicants have an average score of 52. As a result, it observes that, on average, White applicants are approved 62% of the time, whereas Hispanic applicants only receive a loan in 48% of cases.

In this case, the bank can curb the discriminatory behavior of the algorithm by adjusting the decision-making threshold based on demographic criteria so as to bring the two acceptance rates into alignment. In a sense, this is reverse discrimination, but with the explicit intent to harmonize acceptance rates among the two populations.

Yet there are problems with this approach. First, there is the obvious fact that acceptance criteria varies based on a protected attribute, i.e., all other things being equal, under this scenario a less qualified Hispanic applicant has the same chance at getting a loan as a more qualified White applicant (due to the manipulation of thresholds). Moreover, there is a significant cost borne by society or the private enterprise by deviating from the "optimal solution", which in this fictional scenario would accept White applications at a higher rate than Hispanic ones.

In a journal article titled "Equality of Opportunity in Supervised Learning", researchers Moritz Hardt, Eric Price, and Nathan Srebro propose a framework for post-processing any learned predictor to apply a condition known as "equality of opportunity" [HPS16]. This framework is particularly well-suited for contexts where we can designate an "advantaged outcome" – e.g., receiving a loan, acceptance to a college, etc.

Consider the following illustrative scenario: a bank seeks to automate decision-making regarding loan applications based wholly on the results of a predictive model [WVH16]. The output of this model is an integer-valued "credit score" between 0-100. Suppose there are two demographic populations represented in the data; the objective is to minimize any discriminatory impact on the two populations by a model Y, while simultaneously maximizing profit P. Four interesting strategies emerge in their analysis: (1) Profit Maximization (omitted from Figure 1); (2) Group Unaware; (3) Demographic Parity; (4) Equal Opportunity.

Scenario 2: Group Unaware

Population	*Correct*	*Positive Rate*	*True Positive Rate*
Blue	79%	52%	81%
Orange	79%	30%	60%

Scenario 3: Demographic Parity

Population	*Correct*	*Positive Rate*	*True Positive Rate*
Blue	77%	37%	64%
Orange	84%	37%	71%

Scenario 4: Equal Opportunity

Population	*Correct*	*Positive Rate*	*True Positive Rate*
Blue	78%	40%	68%
Orange	83%	35%	68%

Figure 1 – Bias Reduction Strategies

Figure 1 shows an analysis of Scenarios 2-4, with three statistics reported for each population group. "Correct" represents the overall accuracy of the predictor, calculated by taking the ratio of correct predictions for each group to the total number of applicants in each group. "Positive Rate" represents the loan acceptance rate, calculated by taking the ratio of total loans granted to applications received. "True Positive Rate" represents the ratio of loan-receiving applicants to the total number of applicants who would have paid back a loan (had they received one). In this example, the last statistic can be calculated because the data is simulated.

Scenario 2 “Group Unware” represents the “race-blind” approach discussed earlier, where race is not included in the model, but no other adjustments are made. In this case, a stricter standard is applied to the Orange population (30% acceptance rate for Orange vs. 52% for Blue). Scenario 3 corresponds to a strategy of equalizing acceptance rates by population, which is a common first-pass approach towards handling data discrimination. However, as the above table shows, an undesirable side-effect of this approach is that it disadvantages virtuous applicants in group Blue, because they must compete for a smaller number of loans (within their group). This is analogous to the famous affirmative action complaint that privileged White Americans must meet higher standards for admission to Ivy League schools.

Scenario 4 shown above corresponds to Equal Opportunity. Instead of harmonizing acceptance rates by demographic across the entire body of loan applicants (Scenario 3), equal opportunity seeks to ensure that the acceptance rate is equal for applicants who would actually pay back a loan (“true positive rate”). This is postulated as a desirable normative outcome because the rate at which “virtuous” members of both populations receive the advantaged outcome is harmonized; in other words, there is equal “opportunity” for members of the protected class. At the same time, the adverse effects on the privileged class are less extreme than under Scenario 3.

Crucially, the framework includes a proof that, given any learned predictor Y, we can formulate Y* such that Y* satisfies the mathematical criterion of equal opportunity. This is important because it does not rely on access to the original model specification; in many cases, organizations make use of models generated by third-parties and do not have access to, or do not understand, the internals of the model. This helpful property makes equal opportunity a

near-universally applicable framework for evaluating discrimination within predictors and post-processing them to correct it.

Case Study: Systemic Data Discrimination in the Criminal Justice System

Courts in the United States rely on third-party predictive algorithms to quantify the risk that a convicted criminal will commit a future crime (known as recidivism). The intent is to avoid unnecessary incarceration, which the US has historically been criticized for. In the past, judges have made these subjective determinations based on personal experience and professional expertise. The introduction of an objective, data-driven algorithm into these settings would seem like a sensible thing to do [BJC15].

In 2016, ProPublica published an analysis of the COMPAS Recidivism Risk Score algorithm, which was being used by at least 9 state courts at the time. ProPublica's analysis showed that Black defendants were almost twice as likely to be wrongly classified as "high-risk", whereas the opposite was true for White defendants. In particular, Black defendants were "77 percent more likely to be pegged as at higher risk of committing a future violent crime and 45 percent more likely to be predicted to commit a future crime of any kind" [ALM$^+$16]. Given that these risk scores were being shown to judges minutes before they presided over sentencing hearings, the implications are quite troubling.

COMPAS is a proprietary algorithm and its publisher has declined to release the exact model specification; however, we do know that it is based on a questionnaire that includes criminal history as well as a set of behavioral questions. For example, it asks defendants questions such as "How many of your friends/acquaintances are taking drugs illegally?" and "How often did

you get in fights at school?". It also defendants to agree/disagree with statements such as "A hungry person has a right to steal" and "When people do minor offenses or use drugs they don't hurt anyone except themselves". Notably, race is not referenced in the questionnaire; however, that does not mean it isn't correlated with the above questions. These hidden correlations allow race to influence the model just as effectively as if it were included as an explicit variable.

For their analysis, ProPublica requested two years worth of COMPAS scores (2013-2014) from the Broward County Sheriff's Office as well as some additional information which allowed them to determine race and subsequent criminal activity (if any). They published this dataset online, as well as a set of Jupyter notebooks used for their analysis (done in R). Figure 2 is a summary table that was produced by loading this data and running some aggregations:

	2-Year Recidivism					
Race:	*No*		*Yes*		*Total*	
African American	#	%	#	%	#	%
Low Risk	1,046	33%	743	23%	1,789	56% (1)
Med/High Risk	468	15% (3)	918	29%	1,386	44%
Total	1,514	48%	1,661	52%	3,175	100%

	2-Year Recidivism					
Race:	*No*		*Yes*		*Total*	
Caucasian	#	%	#	%	#	%
Low Risk	1,084	52%	564	27%	1,648	78% (2)
Med/High Risk	197	9% (4)	258	12%	455	22%
Total	1,281	61%	822	39%	2,103	100%

Figure 2 – Summary of ProPublica Dataset

In the above table, it is easy to see why ProPublica went hunting for (and successfully found) evidence of racial discrimination. 56% of African-American defendants are labeled as "Low Risk" by the algorithm, versus 78% of Caucasians (Figure 2 marks 1,2). Moreover, African-American defendants in the above dataset who then went on the abide by the law would have

faced a 31% chance of being incorrectly mis-classified as High Risk, versus just 15% for Caucasians (marks 3,4).

The principle of equal opportunity seeks to mitigate this imbalance; in order to operationalize it, we will make use of the terminology used earlier: **Y*** is an equal-opportunity classifier if, and only if, the True Positive Rate for **Y*** among Caucasians is the same as that for African-Americans. Figure 3 clearly shows that the existing COMPAS algorithm fails to satisfy equal opportunity – among the 3,175 African-American defendants represented in the dataset, the True Positive Rate is just 69%, compared to 85% for Caucasians and a similar number for various other groups (Hispanics, Other).

race	total	accuracy	positive_rate	true_positive_rate
Asian	31	0.84	0.84	0.96
Other	343	0.67	0.78	0.85
Caucasian	2103	0.64	0.78	0.85
Native American	11	0.82	0.55	0.83
Hispanic	509	0.64	0.74	0.81
African-American	3175	0.62	0.56	0.69

Figure 3 – Accuracy and True Positive Rate of Various Groups

In this case, we envision an equal opportunity classifier to be a decision rule that results in harmonized true positive rates across various values of race. The original COMPAS algorithm generates scores 1-10, with 4 being the cutoff for “Low Risk”. Since the score is not a continuous variable, we cannot simply pick varying thresholds by population. Instead, we generate a random variable (picked from a different distribution depending on race) and use it to modify the raw

score, then we apply the original decision rule to the modified score. An important caveat here is that if the modified score results in a value that is higher than the raw score, we use the raw score instead. This is to prevent a scenario where a defendant would have been labeled as Low Risk based on the raw score, but receives something else based on random chance. The influence of the modified score can only benefit the defendant, not hurt them.

Under this formulation, finding "equal opportunity" becomes an optimization problem on the various parameters of the random variables. For example, the modified score for an African-American defendant can be computed as follows:

$$\text{Modified\_Score} = \text{Raw\_Score} - N(\mu_{Afr}, \sigma_{Afr})$$

A viable solution to the optimization problem was found very quickly, so much so that we did not even leverage formal machine-learning techniques (e.g., stochastic gradient descent). The step-by-step approach is as follows:

(1) First, we initialized our random variables (one per race) as $N(\mu=0, \sigma=1)$.

(2) At each iteration, we generated a vector consisting of the true positive rates, and defined our "loss function" as simply the standard deviation of its elements.

(3) If the standard deviation exceeded a threshold value (0.01 in this case), we scaled this vector (using sklearn preprocessing) and fed it into the next iteration as a "learning vector" for μ (learning rate=0.33) and σ (lr=0.20).

(4) When the standard deviation dipped below our threshold, we stopped iterating

Results & Discussion

As shown in Figure 4, the optimization converted after 6 iterations, and harmonized the True Positive Rate to 0.85 for all groups.

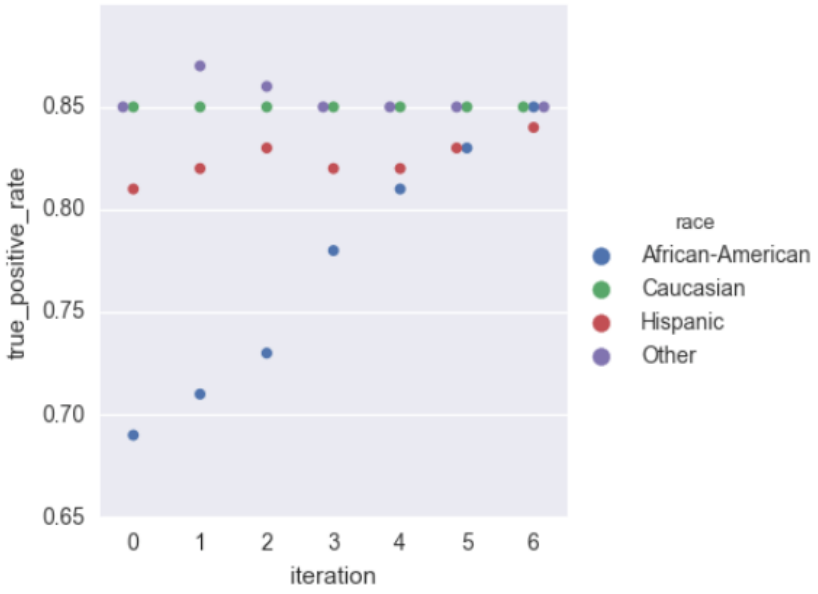

Figure 4 – Model Iterations

Figure 5 shows the final weights for the random variables. As expected, scores of African-American defendants received a substantial (random) adjustment to compensate for inherent biases in the model. As mentioned earlier, negative weights were of no consequence because we always took the MIN value between raw and modified scores when applying the decision rule for Low Risk (score <= 4).

Final Weights

Race	Mean	Std
African-American	2.43	2.46
Caucasian	-1.37*	0.18
Hispanic	0.53	1.32
Other	-1.53*	0.08

* Negative weights are ignored

Figure 5 – Final Model Weights

As shown in Figure 6, the overall impact of the equal opportunity adjustments was to increase the True Positive Rate by a sizable 23% among African-Americans, and a modest 4% among Hispanics; this adjustment brought them in line with the other groups at 85%. On the other hand, we experienced a drop in accuracy because, by definition, we deviated from the "optimal" predictions that were generated by the race-blind model. However, it is notable that the reduction in accuracy is far lower than might have been expected – just 6%.

	True Positive Rate			Accuracy		
Race	*Original*	*Final*	*% Δ*	*Original*	*Final*	*% Δ*
African-American	0.69	0.85	23%	0.62	0.58	-6%
Caucasian	0.85	0.85	0%	0.64	0.64	0%
Hispanic	0.81	0.84	4%	0.64	0.64	0%
Other	0.85	0.85	0%	0.67	0.67	0%

An interesting consequence of Equal Opportunity classifiers is that they place the incremental cost of poor accuracy on the shoulders of the organization making use of the algorithm, rather than the individual affected by this. Hardt, Price, and Srebro also remark on this in their paper, citing it as a positive consequence [HPS16]. Their reasoning is this: if the reduced accuracy among a particular group is due to signal quality or model quality issues, then this represents the right incentive structure for the organization, as it can simply respond by investing more resources in calibrating the model and improving its accuracy with the group in question (which will reduce the magnitude of any equal-opportunity adjustments).

The full source code for this work is available at: https://github.com/saapad86/compas-analysis. See spadela-equal-opportunity.ipynb for the Jupyter notebook that was used to generate the equal-opportunity classifier and various charts.

References

[PDR08] Pedreshi, D., Ruggieri, S., Turini, F. Discrimination-aware data mining. In *Proc. 14th ACM SIGKDD, 2008*. Retrieved from http://pages.di.unipi.it/ruggieri/Papers/kdd2008.pdf on April 15, 2018.

[BS16] Barocas, Solon and Andrew Selbst. Big data's disparate impact. *California Law Review, 104,* 2016.

[PPM+14] John Podesta, Penny Pritzker, Ernest J. Moniz, John Holdren, and Jefrey Zients. Big data: Seizing opportunities and preserving values. Executive Office of the President, May 2014.

[HPS16] Hardt, M., Price, E., and Srebro, N. Equality of Opportunity in Supervised Learning. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems, 2016*. Retrieved from https://arxiv.org/ on April 15, 2018.

[WVH16] Wattenberg, M., Viégas, F., and Hardt., M. Attacking discrimination with smarter machine learning. Google Big Picture Group. Retrieved from https://research.google.com/bigpicture/attacking-discrimination-in-ml/ on April 15, 2018.

[Mil10] Miller, Claire Cain. When Algorithms Discriminate. *The New York Times (2015, Jul 10).* Retrieved from http://www.nytimes.com on Feb 24, 2018.

[Leo17] Leonard, Matt. Uncovering discrimination in machine-learning software. *GCN (2017, Aug 25).* Retrieved from https://gcn.com/ on Feb 24, 2018.

[WVH+18] Wattenberg, M., Viégas, F., and Hardt, M. Attacking discrimination with smarter machine learning. *Big Picture Group: Google Brain*. Retrieved from https://research.google.com/bigpicture/attacking-discrimination-in-ml/ on Feb 24, 2018.

[ALM+16] Angwin, J., Larson, J., Mattu, S., Kirchner, L. Machine Bias. *ProPublica (2016, May 23).* Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing on April 15, 2018.

[LMK+16] Larson, J., Mattu, S., Kirchner, L., Angwin, J. How We Analyzed the COMPAS Recidivism Algorithm. ProPublica (2016, May 23). Retrieved from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm on April 15, 2018.

[Cra13] Crawford, Kate. Think Again: Big Data. *Foreign Policy (2013, May 10).* Retrieved from http://foreignpolicy.com/2013/05/10/think-again-big-data/ on April 15, 2018.

[Ler13] Lerman, Jonas. Big Data And Its Exclusions. *66 Stanford Law Review Online 55 (2013).*

[Noy15] Noyes, Kathrine. Will big data help end discrimination – or make it worse? *Fortune (2015, Jan 15)*. Retrieved from http://fortune.com/2015/01/15/will-big-data-help-end-discrimination-or-make-it-worse/ on April 15, 2018.

[BJC15] Barry-Jester, A., Casselman, B. The New Science of Sentencing. *The Marshall Project (2015, Aug 4)*. Retrieved from https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing on April 15, 2018.