Mphasis
The Next Applied

Demystifying Bias Detection and Mitigation in Machine Learning

Whitepaper by Dr. Ashwani Singh, Manager – Applied AI, Mphasis NEXT Labs
Divay Garg, Asst. Manager – Applied AI, Mphasis NEXT Labs

Contents

1.
Introduction

Machine learning algorithms have become more capable in the past few years, especially with the advent of technologies such as deep learning. As a result, the use of ML algorithms in augmenting or supporting human decisions in a range of domains such as advertising, financial services, judicial and policing systems has also exploded. One would believe that this explosion in the use of AI is a harbinger of a more rational future, free from human prejudices and biases inherent in past societies. However, the reality shows a different and bleak picture. Since the algorithms are based on past or historic data, they may perpetuate instead of preventing societies' existing inequities and discriminations. For instance, a study[1] by ProPublica of the COMPAS system to predict recidivism, found the predictions to be biased against African American prisoners. Similar cases exist in other domains as well. Xing, a hiring platform was found to rank less qualified male applicants higher than more qualified female applicants[2]. Publicly available face recognition algorithms have been shown to perform poorly in recognizing African American males and females. This bias in algorithms is essentially a reflection of data containing biases present in society. The sad fact is that unless such biases are detected and mitigated, AI-based systems have the potential to do more harm than good, especially towards societal groups that have faced disadvantages and discrimination in the past.

Fortunately, the advent of fields such as explainable AI and more recently, responsible AI, have brought greater attention to the need for ascertaining bias in AI-based systems and ensuring that such systems do not inadvertently lead to discriminations. This is also in line with the increased focus among business leaders to try and ensure debiased decisions. Organizations have incorporated training programs for employees to reduce unconscious biases while taking decisions and formulated policies for identifying discriminatory processes leading to adverse outcomes for groups. These principles of bias detection and mitigation can also be extended to machine learning systems to make them fair and equitable; the need of the hour is formulating a clear set of policies for an unbiased implementation of AI-based systems. A clear policy for algorithmic fairness would entail outlining processes for bias detection and once detected, mitigating the biases identified.

Unfortunately, this is easier said than done due to the high degree of complexity involved. Bias detection is challenging as there could be multiple definitions of bias in decisions[3], and an argument could be made as to which is the most appropriate metric of bias for a particular use case. Deciding the metric for detecting bias is important since different stakeholders may be interested in different metrics. For instance, loan applicants would prefer the algorithm to maximize true positives and minimize false negatives so that those deserving the loans get them, while the loan providers may want to minimize false positives and correctly identify applicants likely to default on the loan in the future.

The mitigation exercise comes with its own set of complications, most important of which is the trade-off between bias mitigation and accuracy. Hence, it is important for business users to have a sense of the challenges associated with bias detection and mitigation and understand how to deal with such challenges. The paper discusses these points in detail and outlines a step-by-step debiasing process using the Mphasis Responsible AI framework.

2.
Bias Detection

Although industry and academia have started work in addressing AI fairness, the definition of algorithmic bias and fair treatment are still under debate. There are numerous definitions of bias [3], which can confuse users who are trying to incorporate fairness in their AI systems. Hence, there is a need for intuitive explanations of bias metrics and directions on when they are suitable for application. The most prominent definitions and their applicability are discussed below, within the context of a bank utilizing AI-based recommendations for granting loans to applicants.

1. **Demographic Parity/Disparate Impact:** This definition is based on only the predicted outcome rather than a comparison between predicted and real outcomes. The objective, in this case, is to ensure the underprivileged group is not being unfairly treated or disadvantaged compared to the privileged group, in terms of the proportion selected. Hence, if the sensitive attribute is race and the outcome is loan approval, roughly the same proportion of individuals amongst privileged and underprivileged racial groups should get loan approvals. This definition is important, as it is considered in court to ascertain if there's discrimination in the outcome of a system. As codified by the US Equal Opportunity Commission, Disparate Impact (ratio of approvals for an underprivileged group to approvals for a privileged group) of lower than 0.8 would be considered as signifying discrimination. The thought process is that approval should not matter on whether a person belongs to a particular racial group. The reasoning extends to other groups as well, so the approval rate among the privileged gender should be the same as the one among the underprivileged gender.

2. **Conditional Statistical Parity:** The obvious question that appears from the above definition is what if the disparate impact has nothing to do with group characteristics, but rather other factors that may be relevant for loan approval. So, if the income and asset ownership levels among whites and non-whites are not the same, and whites have higher income and asset ownership levels, should this not be considered? Conditional statistical parity extends the rationale of the statistical parity measure, however, with the additional consideration of relevant factors that could legitimately produce differences in the approval rates. A bias would exist, however, if despite controlling for the relevant factors, there remains a difference in the approval rates between the privileged and the underprivileged groups.

The issue with the above definitions of bias is that while striving for fairness, they do not consider the ground truth in their assessment. This is like saying that a fair outcome is different from and independent of the ground reality, and it does not matter whether the deserving candidates get selected and the undeserving rejected while assessing fairness. This would mean that if loans were approved at similar rates for deserving whites, and random non-whites, we should be satisfied that the system is producing fair results. It could be argued that a fair system should ensure that the right people got the right results rather than the results being equal across groups. Hence, the next set of definitions compare the predicted outcome with the true outcome to measure bias.

1. **Positive Predictive Value (PPV) Parity:** This means that the proportion of correct positive predictions is the same across groups. If 60% of positive calls (those predicted as suitable for loan approval) deserve to get the loan, i.e., the true value is the loan approved among whites and this proportion is the same across non-whites, it would mean that there's PPV parity among the two groups. On the other hand, if 80% of non-whites with positive calls deserve the loan, there could be bias in favor of the whites. However, the main idea behind this definition is that the model does not predict differently for the privileged and underprivileged groups. This definition of fairness was used by the developers of COMPAS to defend against allegations of bias against African American prisoners, as according to them, the system showed predictive parity, and hence was fair.

2. **Equal Opportunity Difference:** Some people may argue that it is more important to ensure that positive outcomes reach those who deserve it. One way to do this is to look at the True Positive Rates (TPR). TPR is the proportion of people predicted to get loan approval to the ones that got loan approvals. It tells you whether the system is treating the deserving equally across groups. As the name suggests, the metric assesses whether the deserving across groups have an equal opportunity to get the positive outcome. If 70% of the deserving get their loans approved among whites, and 70% of the deserving get approvals among non-whites, it means the deserving in both groups have an equal opportunity to be predicted as getting the loan, and hence, the system is fair. A related concept is the False negative error rate which asks how many people out of the deserving are being misclassified as undeserving. This question was asked by the critics of the COMPAS algorithm regarding the prediction of recidivism in African American and Caucasian prisoners, and it was found that the system favored Caucasian prisoners through a high false negative rate i.e., identifying those who would commit a crime as those not at risk of doing so. Hence, due to this misclassification, society was more at danger from misclassified Caucasian prisoners who would be set free and promptly commit a crime, as compared to that of similar African American parolees.

3. **False Positive Error Rate Balance:** Sometimes the problem is not one of discrimination against a group, but rather undue favor or privilege for a group. The idea is that the system is unfair if it unduly benefits the undeserving in a particular group. Hence, if a higher proportion of the undeserving applicants are misclassified as deserving in men as against women, it would mean that men are getting an advantage that they do not deserve, and the system fails the fairness test. On the other hand, if the undeserving applicants are misclassified at similar rates across men and women, no group enjoys an undue advantage and according to this definition would be fair. In the recidivism case, it was found that the false positive rate for

African Americans was much higher than for Caucasian prisoners, and a much higher number of prisoners who would not commit crimes were misclassified as at risk for committing a crime. The reformed African American prisoners were more likely to not get paroles as they would be mistakenly considered dangerous.

4. **Equalized Odds Difference:** This is a more stringent measure of bias that combines the above two measures. A classifier is fair if it has similar equal opportunity rates and false positive error rates for the privileged and underprivileged groups. The main idea is that the classifier treats the deserving applicants between groups equally well and the undeserving equally poorly. The probability of a good applicant getting approved and a bad applicant getting disapproved would be similar across the two groups. The conjunction of the above two definitions must hold for this metric to assess the model as fair.

The above measures look at fairness from a group's point of view and do not have much to say in terms of individuals making up those groups. A different set of measures looks at individual cases and tries to assess fairness in terms of outcomes for individuals rather than the entire group.

1. **Causal Discrimination:** A system would be unfair if similar individuals do not get similar results. Here, bias is checked by introducing a new individual case that is the same as one in the original set except for the sensitive feature. For instance, if non-white applicants having all the other characteristics same as that of white applicants existing in the data set, are introduced and predictions obtained for this new set of applicants, would the outcome differ? In case the proportion of individuals getting their loans approved reduces, one can conclude that it would have been because of the change in the sensitive attribute.

We have a set of bias detection metrics to choose from and each of these metrics provides different information about the model fairness. Ideally, it is a good idea to detect bias at three levels: first, an overall level to rule out disparate impact, second, considering the ground truth at the level of group fairness by including one of the metrics like equal opportunity or equalized odds to make sure that the model treats deserving and undeserving similarly across groups, and last, make sure the fairness holds for individual cases even if the sensitive attributes are changed.

3.
Bias Mitigation

While the above metrics could be used to identify and validate the fairness of the algorithm, it is very much essential to incorporate these metrics while addressing an algorithm's fairness during bias mitigation, to have a proper comparison in results before and after the bias mitigation. Bias mitigation may be attempted through multiple approaches as explained below.

Unawareness: The simplest form of bias mitigation, in this case, the model developer deliberately excludes information about sensitive features while training the model. In our example, information about sensitive attributes such as race and sex is deemed to be irrelevant and excluded for each individual, while making the prediction. The algorithm by this definition is fair if it does not take the excluded factors under consideration. However, this approach

may not be successful due to the presence of other features that can act as proxies of the sensitive feature and result in bias. For instance, the college being an all women's college, or the applicant being a part of a "women's team" if used as features may result in the applicant being identified as a woman and result in a gender-based bias, even though the gender variable is not explicitly included. In fact, such bias by proxy was the reason behind Amazon's hiring algorithm discriminating against women[4].

Mitigation Algorithms: Bias mitigation algorithms are meant for supervised learning and hence try to train the model in a way that the bias metrics are reduced to within threshold levels. Mitigation can be performed at three levels, namely, at the data pre-processing stage prior to model training, during model training and through post processing after model training.

1. **Bias Mitigation Before Model Training:** This approach is used when training data itself shows discrimination towards a certain group, due to the bias in judgments of earlier decision makers. The objective of this approach is to make accurate yet non-discriminatory predictions through changes in the training data, thereby making it a multi-objective optimization problem. Two methods are generally used:
    - **Massaging the dataset[5]:** Here, some of the objects of the dataset are relabeled to remove discrimination from input data, while maintaining the overall class distribution. A ranker is used to select the best candidates for relabeling purpose. The ranker ranks observations according to the probability of achieving the positive outcome, and then candidates closest to the decision border are selected for relabeling. For instance, in the loan disbursal dataset, some of the observations on the decision boundary may be relabeled. Thus, some privileged group members receiving loans will be relabeled as not receiving loans, and the same number of underprivileged group members not receiving loans will be relabeled as receiving loans. Care is taken to minimize the number of observations relabeled to ensure the least effect on accuracy levels. The major drawback of this method is the intrusive nature of this relabeling process, which results in a changing of the input data.
    - **Reweighing[5]:** Instead of changing labels of the dataset altogether, this method assigns different weights to objects in the dataset to make it non-discriminatory. For the underprivileged group, the reweighing process assigns higher weights to observations obtaining positive outcome, and lower weights to those obtaining the negative outcome and vice-versa for the privileged group. So, in the loan disbursement example, for instance, the underprivileged group members obtaining the loan will be assigned higher weights and those not obtaining the loan would be assigned lower weights. For the privileged group members, weights will be assigned in the opposite direction with those obtaining the loans being assigned lower weights, and those not obtaining the loans being assigned higher weights. These weights can then be used directly in the prediction process to get a fair outcome.

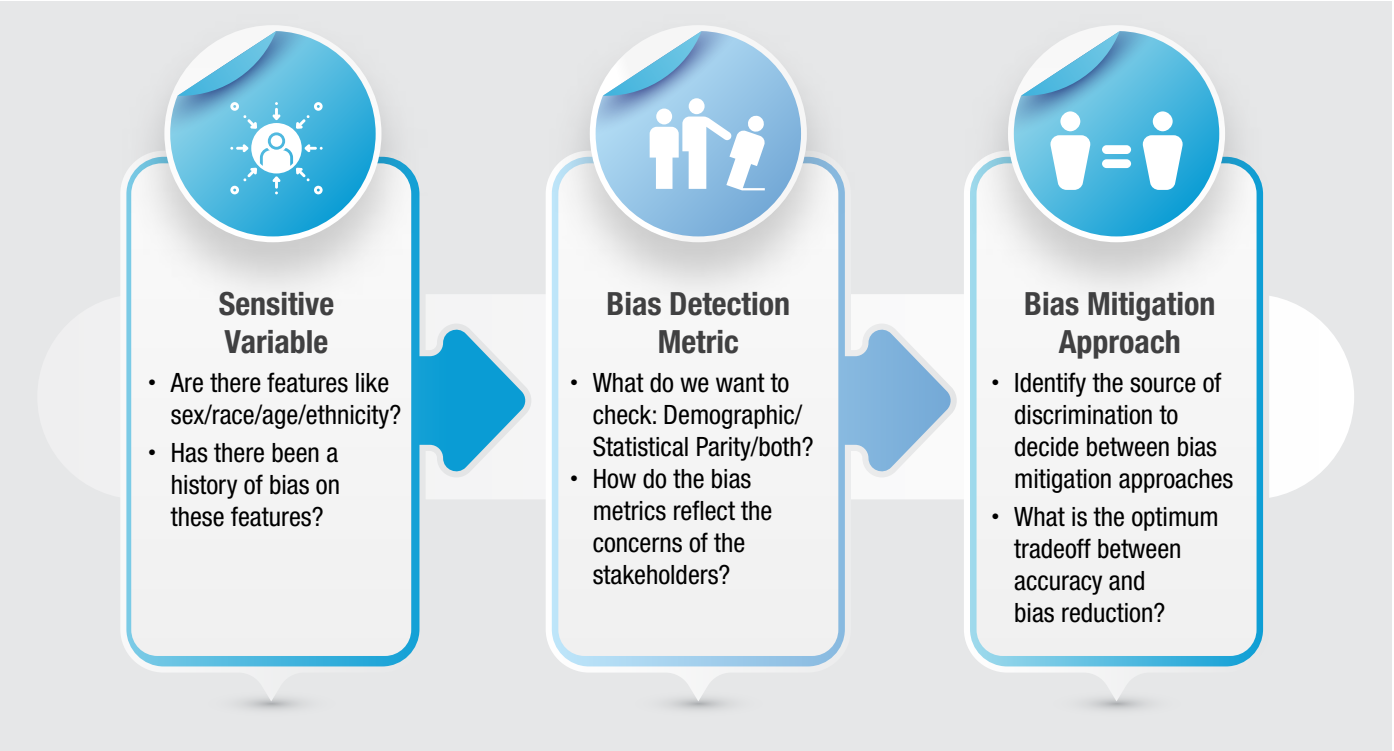
2. **Bias Mitigation During Model Training:** As the name suggests, bias mitigation takes place during the training process, maintaining the accuracy of the predictions. Here, predictor variables in the decision process are weighed based on the significance of variables while maintaining independence between the sensitive variable and the decision taken. Some ways of mitigating bias while training the model are as follows:
    - **Adversarial Debiasing[6]:** In this bias mitigation approach, the model attempts to predict the outcome variable while an adversary tries to model the sensitive variable like sex or race and so on. The objective of the mitigation is to maximize the model's ability to predict the outcome variable while at the same time minimizing the adversary's ability to predict the sensitive variable. The model shows mitigation of statistical bias such as average odds ratio while maintaining accuracy, resulting in a non-discriminatory decision process.
    - **Reductions Based Debiasing[7]:** Here the classification task is converted to a cost sensitive problem where the solution provides a classifier with the least error subject to the desired constraint. This means that the model first allows the user to define a bias metric as a constraint (say disparate impact), thereafter it optimizes the tradeoff between accuracy and the bias metric defined as the constraint. One benefit of this approach is that it can be utilized across various types of models.
3. **Bias Mitigation Post Model Training:** Post training bias mitigation does not change the training data and the learning algorithm and treats both as given.
    - **Equalized Odds Post Processing[8]:** This post processing bias mitigation utilizes an optimization methodology utilizing Bayes optimal predictors, whereby an equalized odds or equal opportunity predictor is derived from a Bayes optimal regressor and the protected variable. The predictions are drawn from a regression score between 0 and 1, with a particular threshold defining the boundary between positive and negative classification. The scores generated by the Bayes optimal regressor along with the threshold optimized for equalized odds/equal opportunity result in a non-discriminatory predictor. One limitation of this approach is that while it works for constraints of statistical parity such as equal opportunity, or equalized odds, it does not handle demographic parity constraints.

4.
Bias Mitigation and Accuracy Tradeoff

The bias mitigation algorithms correct the influence of attributes causing the bias in predictions. However, since the ground truth in the data may itself carry historical biases of human decision makers, accuracy suffers if this is corrected. Getting an unbiased prediction means going against the ground truth. The reduction in accuracy is related to the amount of bias that exists in the data, but it may be difficult to derive the functional relation i.e. that x amount of bias results in y amount of reduced accuracy upon mitigation, Moreover, during mitigation, reduction in accuracy depends on the bias metric provided as a constraint. Also, changing hyperparameters while training the

model results in varying bias and accuracy metrics. The ideal approach would be to reduce the bias to within threshold levels in step 1, and in step 2, identify the hyperparameters which maximize accuracy while keeping bias within acceptable limits. This implies that bias mitigation is a process that requires a certain level of experimentation from the data scientists.

The above bias detection and mitigation processes can be used as per the needs and requirements of the users. The process starts with identifying whether the dataset contains any sensitive variables that signify groups that may have been discriminated against due to biased decisions in the past. The next step is selecting bias metrics for the detection of bias. Once bias is detected, it is important to check the source of discrimination for the appropriate utilization of bias mitigation methodologies. A step-by-step progression of a systematic bias detection and mitigation process, and the judgment calls at each step is provided below.



5.

Mphasis Responsible AI Framework

Considering the need for more trustworthy AI systems, Mphasis has recognized the growing importance of fairness in AI models and identified it as a priority area for the future. We strive to offer features of bias identification and mitigation in all critical models through Mphasis Responsible AI framework, which helps in addressing fairness in machine learning models.

Mphasis Responsible AI solution identifies existing biases that can occur across gender, ethnicity, race etc. The solution mitigates discrimination present in the data and prediction process, making the outcome non-discriminatory. The solution gives a comparative view of analysis before and after bias mitigation to provide clarity on the bias-accuracy tradeoff and allows customers more transparency in the entire process.

Organizations can use Mphasis Responsible AI as a guide to ensure the existing biases are mitigated, and new ones are not introduced when human judgment processes are replaced or augmented by AI solutions. This results in trustworthy, interpretable systems that can pass muster on openness and fairness concerns of regulatory bodies and civil society.

6. Reference

[1] Machine Bias, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] 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.

[3] Verma, S., & Rubin, J. (2018, May). Fairness definitions explained. In 2018 ieee/acm international workshop on software fairness (fairware) (pp. 1-7). IEEE.

[4] Amazon scraps secret AI recruiting tool that showed bias against women, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[5] Kamiran, F. and Calders, T. Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1):1–33, 2012. doi: https://doi.org/10.1007/s10115-011-0463-8.

[6] Zhang, B. H., Lemoine, B., and Mitchell, M. Mitigating unwanted biases with adversarial learning. In Proc. AAAI/ACM Conf. Artif. Intell., Ethics, Society, New Orleans, USA, February 2018.

[7] Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018, July). A reductions approach to fair classification. In International Conference on Machine Learning (pp. 60-69). PMLR.

[8] Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems 29, pp. 3315–3323, Barcelona, Spain, December 2016.

Authors

Dr. Ashwani Singh

Manager – Applied AI, Mphasis NEXT Labs

Ashwani has expertise in emerging technologies and has been part of several award-winning projects in this space. Having previously worked in analytics and academics, he is currently a part of the leadership team at Mphasis NEXT Labs, where he helps solve business problems by leveraging AI/ML. Ashwani is also a thought leader in the space of behavioral applications of AI, with special focus on the emerging area of affective computing. He completed his doctoral studies at IIM Bangalore where his research was focused on Consumer Cognition and Decision Making.

Divay Garg

Asst. Manager – Applied AI, Mphasis NEXT Labs

With 4+ years of experience in Data Science, Divay Garg has multiple patents to his credit. He has been a part of multiple award-winning projects. His interests include AI/ML, optimization and Natural Language Processing. Divay has completed his Masters in Industrial Engineering from Indian Institute of Technology (IIT, Kanpur).

About Mphasis

Mphasis (BSE: 526299; NSE: MPHASIS) applies next-generation technology to help enterprises transform businesses globally. Customer centricity is foundational to Mphasis and is reflected in the Mphasis' Front2Back™ Transformation approach. Front2Back™ uses the exponential power of cloud and cognitive to provide hyper-personalized ($C = X2C^{2}_{TM} = 1$) digital experience to clients and their end customers. Mphasis' Service Transformation approach helps 'shrink the core' through the application of digital technologies across legacy environments within an enterprise, enabling businesses to stay ahead in a changing world. Mphasis' core reference architectures and tools, speed and innovation with domain expertise and specialization are key to building strong relationships with marquee clients. To know more, please visit www.mphasis.com

For more information, contact: marketinginfo.m@mphasis.com

USA
460 Park Avenue South
Suite #1101
New York, NY 10016, USA
Tel.: +1 212 686 6655

UK
1 Ropemaker Street, London
EC2Y 9HT, United Kingdom
T : +44 020 7153 1327

INDIA
Bagmane World Technology Center
Marathahalli Ring Road
Doddanakundhi Village
Mahadevapura
Bangalore 560 048, India
Tel.: +91 80 3352 5000

www.mphasis.com

NR 30/03/21 US LETTER 6MS16864