SAS® GLOBAL FORUM 2021

Paper 1159-2021

Auditing Algorithms: The Need for New Tools to Combat Bias in Artificial Intelligence

Jason Brinkley, PhD, Abt Associates

ABSTRACT

The field of algorithm development for machine learning/artificial intelligence has seen rapid changes in recent years. And the rise in algorithm development has increased concerns of rampant bias in implementation. Many businesses turned to AI as a mechanism to create or evaluate things faster and cheaper with more accuracy than human beings. This hyper-focus on speed and accuracy has created numerous instances of AI with biases; there are many examples in the news and research articles of algorithms that have been shown to have strong gender or racial biases. There is a tug of war going on between the data science and policymaker communities around the need to better evaluate and audit AI to discover such biases before widespread rollout. The main issue is the need for methodological innovations and our existing tools are not well-suited for evaluating AI. We will spend some time looking at the algorithm landscape and discuss novel innovations, such as humble, fair or interpretable AI in a way that is digestible to a general audience. The focal point of the talk is on determining what we need to ask of AI in the future. The new world of asking AI to make predictions through a lens of equity necessitates the need for new evaluation measures that move beyond these original goals. The talk is designed for both scientists and policymakers.

INTRODUCTION

Artificial Intelligence (AI) and Machine Learning (ML) have seen rapid growth in recent years. Businesses have turned to ML to create products faster and cheaper with higher ‘accuracy’ than human beings. The increased focus on using these methods has led to an explosion of new algorithms being developed and deployed in a wide variety of industries, which overall has been far from perfect. There continues to be a host of new stories, research briefs, and anecdotal evidence to suggest that some of these algorithms have biased results. As the use of these methods continues to grow, it is important to assess the mechanisms behind these biases and to discuss ways in which the bias can be addressed. The goals of this paper are to lay out the problem and to discuss possible directions for solutions. The root causes are the hyper-focus on speed and accuracy in training data while the solutions will range from different algorithms for prediction as well as the need for new tools in evaluation.

BACKGROUND

For many AI and ML are different frameworks with many considering ML to be a subset of the larger field of AI. For the purposes of this evaluation, the focus will be on AI/ML algorithms that are based on so called predictive modeling techniques. The reader can see Shmueli (2010) for an in-depth discussion of why prediction is a different type of problem then traditional statistical modeling which is characterized by a need to attribute and/or explain variation in our outcomes to one or many predictors. Traditional statistical modeling also relied on functional forms or model parameters that were interpretable to a general scientific audience. In contrast, prediction scenarios put less of an emphasis on these factors and relies mostly on the quality of predictions from models. Seen through the light of population studies, prediction modeling values the accuracy of individual estimates (what is the potential risk or impact to a specific individual) versus traditional modeling where the focus is on measuring the unique contribution of each factor, covariate, feature, or characteristic and how it relates to the outcome of interest. Predictive models are by their nature more flexible and allow for a wider array of complex associations between and within your data. The argument against traditional models has been that it is very difficult to account for the impact of intersections of variables (say geography and demographics) or to allow for nonlinear associations between variables. When the focus shifts toward individual level prediction, then it is plausible to create models that make better use of the available data. The downside is that AI/ML algorithms are often referred to as *black box* algorithms in that they are difficult to interpret or to assign the individual contribution of each variable input into the model. The reader can see Friedman, Tibshirani, and Hastie (2001) as a foundational text on machine learning algorithms that focus on prediction modeling.

Predictive modeling-based ML took off as we found ways to implement these algorithms in modern computers. Their ability to work with large sources of data and to produce relatively fast output create opportunities to develop algorithms that can offload tasks that are too sophisticated for humans (e.g. predicting risk as a function of hundreds of covariate inputs) or scaled above human capacity (e.g. read and summarize text). However, many ML algorithms were designed to prioritize speed and accuracy and guarantee output. This led to convoluted algorithms with unexpected results. Specifically, the focus on accuracy meant that hidden biases in data collection or human driven systems would be heightened in the development and deployment of highly attuned algorithms. Therefore, if there are gender or racial biases in the underlying data that these models were developed under then it is plausible that those algorithms would make use of those biases as a way to gain better predictions. With no framework to explore the individual contribution of variables such as race or gender, model AI/ML is having some issues in a new world where equity matters.

It is important to note that the origins of algorithm development are not in and of themselves biased. The mathematics that build commonly used ML models (whether they be tree models such as Gradient Boosted Models (GBM) or Random Forest Models (RF) or neural network based models such as Deep Learning Models (DL)) are grounded in theory that is not inherently biased. These biases creep in as we use these models with real data from systems that have existing biases or from environments that have systemic disparities.

ADDRESSING BIASES ON THE FRONT-END

One route of addressing disparities or making machine learning more equitable would be on front end development. As the industry moves away from solely valuing speed and predictive accuracy, we move into realms where the ask of these models may be slightly different. Different research teams are working to deliver algorithms with less bias in a few different directions. By exploring the different directions in addressing bias in algorithm development, we can see that the potential impact of biased AI/ML and that the solutions may not be one size fits all.

INTERPRETABLE MACHINE LEARNING

Interpretable AI/ML is the umbrella term for prediction algorithms that more closely mimic traditional models in that the individual impact and/or use of each variable input into the model is understandable by a general scientific audience. The goal is to continue to use a wide number of potential variables and to let computers do the work of sifting through and identifying which specific variables, interactions, and subgroups provide the most benefit in making predictions. There are some restrictions placed on defining those relationships so that the output can be interpretable. While some have suggested that current algorithms have output that assist in interpretability (e.g. variable importance measures) existing metrics only tell us the contributions of each variable to the model. It does not tells us the directions of impact (for example does being female increase or decrease predictions?) nor account for the type of relations (e.g. linear or nonlinear) nor account for complex interactions between variables. The goal of truly interpretable AI/ML should be to combine the power and flexibility of prediction models with the ability to explain the impact of individual variables the same way as traditional statistical models.

Alternative frameworks are being advocate for at this time. Rudin (2019) and Letham (2015) are excellent starting points to look at refined versions of tree models that use different techniques to blend traditional analytics with modern prediction models to form machine learning models whose predictions are accurate as well as interpretable. Heuristically, the idea is to return to tree models (that form the backbone of GBM and RF models) and to redefine the rules for decision-making that allow for high quality predictions while also yielding decisions lists or subgroupings that more closely resemble human based decision making. The hope is that such lists can be explored for hidden biases or to determine whether algorithm based recommendations are using factors such as gender or race inappropriately.

HUMBLE MACHINE LEARNING

In contrast to interpretable AI/ML, there are some who advocate the use of humility by machines in making predictions. Humble AI is a framework by which data scientists can continue to form algorithms in the mold that is current best practice. Instead of modifying the algorithms themselves, we set boundaries on the prediction space where the models are allowed to make forecasts, predictions, or decisions. In this way, clinical and subject matter experts can offer insights on vulnerable or high risk populations where humans need to provide expert review. The work by Colin Parris and others really focus in on this idea of restricted prediction spaces for AI deployment; however, at the time of this writing there is little published data on specific changes to algorithms.

FAIR MACHINE LEARNING

Fair algorithms define a class of AI/ML algorithms where equity is included as part of the algorithm development. The general notion is to force commonly used ML metrics such as accuracy to be the same across a number of subgroups. We add one or many parameters that are optimized to ensure that the overall predictions across all protected groups is equivalent. Starting points for learning about fair algorithm creation would be to look at Chouldechova (2018) and Rose (2016). There are obvious benefits to the exploration of fair algorithms and a lot of work is being done to find adequate penalties on parameters to help make machine learning predictions fair. It is important to note that not all applications of AI/ML deployment can be made fair. In cases where the systemic biases are so large that accurate predictions would rely solely on biased decision-making creates instances where any penalization will result in algorithms that are completely ineffective. Work continues in this area and the hope is that new techniques will be created that can make a highly predictive, accurate, and fair set of prediction models.

ADDRESSING BIASES ON THE BACK-END

While addressing potential biases on the front end of AI/ML development is important, there already exists a wide array of well-developed algorithms that are deployed in different settings and it is not immediately clear that all suffer from these kinds of biases. Furthermore, we see that the different methods for dealing with these biases really come down to a matter of perspective and there may not be a one-size fits all methodology for developing unbiased AI. Therefore there is a need for assessing AI/ML algorithms on the back end to determine if what has been developed is unfair or biased. Surprisingly, there is not a lot of literature and little developed in this realm at this time. This is a major limitation and one of the contributing reasons why researchers don’t tend to discover biased ML before it has been deployed. Likewise, even if one of the aforementioned techniques had been utilized, there is no guarantee that those algorithms are free from all sources of bias. We might well develop algorithms that were intended to not suffer from gender, age, or race/ethnic biases only to discover that there were biases due to disability status or other factors down the line. There is a great need in the field for development of new tools and techniques to better spot bias before any algorithm becomes widespread in use.

EXAMPLE METRIC – DISPARATE IMPACT SCORE

Feldman (2015) introduced a useful metric in assessing the potential for disparities in machine learning algorithms, the Disparate Impact Score (DIS). DIS measures the ratio of chances between outcomes for being inside versus outside of the class of interest. So for example, suppose we want to look at gender disparities between males and non-males. Then we would define the DIS as follows:

$$\frac{P(Y = 1|X = 1)}{P(Y = 1|X = 0)}$$

Where Y is a binary outcome of interest (Y=1 is the event of interest) and X is a binary class with potential disparity. We can see that this ratio looks at the ratio of the predictions between those in the protected group versus those not in the protected group. In practice, those scores can be calculated on individual observations (so a prediction for each individual for whether they were male or if they were non-male with the same covariates) or across groups (potentially averaging out the effect of other indicators). However, we can see that this current metric is limited to only binary outcomes and binary groups for comparison. Indeed, we would need to calculate a multitude of DIS measures across every comparison group that we wanted to explore. And that is only in the cases where we can identify the potential biases that we want to explore. As such, this metric provides a good example of the current state of the field in that we have some measures with limited usability that work only for cases where we know the kind of biases we want to explore apriori.

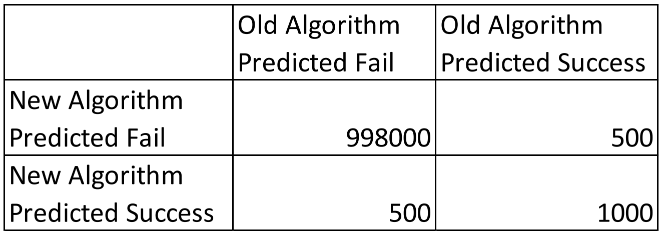
EXAMPLE TECHNIQUE – COMPARING ALGORITHMS

Another area that needs more research and development is back-end comparisons of algorithms. Suppose we develop two different AI/ML based algorithms to explore some outcome of interest. Perhaps I have an existing algorithm that is suspected to be biased and I need to replace with another algorithm that was developed along the lines discussed here. How do I know that I have a suitable replacement? First, we could start with the traditional metrics for predictive modeling based AI/ML and look at measures such as precision, recall, and accuracy. But, as described in the introduction, such metrics focus on speed and accuracy which may not dovetail with our new goals of accuracy and equity. Consider the following scenario: suppose we have developed a prediction algorithm to predict success/failure on some target outcome and we have a table like we see in Table 1.

	True Failure	True Success
Predicted Failure	998000	500
Predicted Success	0	1500

Table 1. Performance of Hypothetical Algorithm

The traditional metrics would say that this is a good algorithm (Precision = 0.75, Recall = 1.00, and Accuracy = 0.9995) as there are roughly 2000 true success and the algorithm accurately predicted 75% of them. Now suppose you develop a second algorithm that is a potential replacement. Suppose it performs exactly as we see in Table 1. It too predicts 75% of true successes. While the conclusion may be to immediately switch from one algorithm to another, we need to make a comparison between the algorithms themselves. Suppose, as hypothetical, that when we compare algorithms we get a table like Table 2.



	Old Algorithm Predicted Fail	Old Algorithm Predicted Success
New Algorithm Predicted Fail	998000	500
New Algorithm Predicted Success	500	1000

Table 2. Comparing Hypothetical Results

Now we see that the 'overlap' between the algorithms is only 1000 of the 2000 successes. Indeed, we see that the disagreement between algorithms is perfect in our toy scenario and that while each algorithm has a 75% precision, the two algorithms pick up exactly what the other has failed to predict. In many cases we would simply *ensemble* the models and take the best of both worlds and have the two algorithms work together. However, if one model is suspected of having bias, then are these results showcasing the scenarios where leaning into biased data led to more accurate results? If the old algorithm is untrustworthy then will replacing it with the new algorithm cause new issues because it is clear that the two algorithms do not *agree* across all results. Indeed, we have a history of such assessment between human evaluators with agreement statistics designed to measure whether two processes are aligned. The issue is that measures such as kappa which are designed to measure agreement, are limited in scope and designed for small samples with human evaluators. Here the algorithms agree on 998,000 failures and the size of that agreement dwarfs everything else in the evaluation. While the focus may be on accuracy in predicting the 2000 successes, our current frameworks for evaluating and comparing different algorithms can also become biased in that they may agree on a whole slew of failures when alignment on successes is key. One might entertain the idea of stratified results by success and failure but that may be too biasing in the other direction and not truly account for the fact that the two algorithms do align heavily on failures. All of this is to showcase the need for new innovations and metrics that can compare algorithms effectively which will allow us to determine if existing models perform differently than those developed to be fair.

A Note on Crossover Assessment

As an avenue for exploring the potential for new ways to compare AI/ML algorithms, one avenue that might be useful would be crossover assessment. Imagine the scenarios above with two competing algorithms. If you wanted to measure agreement between algorithms then one could run the results of one algorithm through the other algorithm and see where differences exist. That is to say that if you have two algorithms (say A and B) then you could deploy the results of algorithm A on B and vice versa and then make comparisons. This could be a useful technique if you know that one algorithm is biased but am unsure where predictions went askew.

DISCUSSION

None of the algorithms or assessment mechanisms discussed here are perfect or applicable in every setting. But we do see that there are many different perspectives on dealing with biased algorithms and that there is a need to better understand the nature of the biases you are dealing with to deploy the best solution. In addition, there are still many challenges and opportunities in the field to contribute to this effort and help create a robust framework for evaluating algorithms. Diverse teams are the starting point on any of this, as there is a great need for different viewpoints to find potential biases. Many of the techniques and metrics discussed here only work when it is known what types of biases may exist in your data.

So what can teams do now in lieu of assessing and correcting against bias in algorithms? First, teams can implement ideas or versions of interpretability, humility, and fairness now. Second, teams can get creative in assessment.

REFERENCES

Shmueli, Galit. "To explain or to predict?." Statistical science 25.3 (2010): 289-310.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning. Vol. 1. No. 10. New York: Springer series in statistics, 2001.

Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." Nature Machine Intelligence 1.5 (2019): 206-215.

Letham, Benjamin, et al. "Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model." Annals of Applied Statistics 9.3 (2015): 1350-1371.

Chouldechova, Alexandra, and Aaron Roth. "The frontiers of fairness in machine learning." arXiv preprint arXiv:1810.08810 (2018).

Rose, Sherri, Alan M. Zaslavsky, and J. Michael McWilliams. "Variation in accountable care organization spending and sensitivity to risk adjustment: implications for benchmarking." Health affairs 35.3 (2016): 440-448.

Feldman, Michael, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. "Certifying and removing disparate impact." In proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 259-268. 2015.

ACKNOWLEDGMENTS

Special thanks to a number of colleagues at Abt Associates for helping me to develop or review these ideas including Michael Link, Suja Thomas, Jenna Harder, Stas Kolenikov, Gabriel Krieshok, Nathan Greenstein, Charlotte Lloyd, Isaiah Gerber, Louise Hadden, Alan White, Allison Muma, Christianna Williams, Laura Peck, Donna DeMarco and many others.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Jason Brinkley
Abt Associates
919-294-7745
Jason_Brinkley@abtassociates.com
https://www.linkedin.com/in/jason-brinkley-09923123/