Big Data and Discrimination

Anonymous Author(s)
email

1 Summary

In their Essay, Gillis and Spiess [3] investigate the problem of discrimination caused by machine-learning decision-making systems. Contrary to other publications that focus on the statistical analysis, they also consider the legal analysis and try to provide a framework that connects algorithmic decision-making with legal doctrine. To do so, they illustrate the problems with the example of setting the price of credit using simulated data and focus on three aspects:

1. **Data Input**: Excluding forbidden characteristics (e.g., race) can have a limited effect because other variables might be correlated in complex ways with the left out feature. Additionally, including the group identifier might be desirable when the dataset contains variables that are biased for some groups, because these biases can then be corrected.
2. **Decision Process**: Only focusing on the feature importance/used variables of an algorithm is also not ideal, because the same result/accuracy may be achieved with different variable combinations and the set of used variables is unstable in their experiment.
3. **Outcomes**: They suggest to use discrimination stress testing where the predictions of the algorithms are evaluated on different populations. However, selecting these populations and constructing a test for disparity can be challenging in practice.

2 Critique & Further Work

I like the idea of simulating data to get a ground truth and then performing various analyses on this dataset. However, their execution was quite poor and does not lead to any new insights, in my opinion. They only address testing for disparity in the end and mention for one of the most important questions, namely how to define "similarly situated persons" (who should get the same prices) in a quantitative way, that it should be addressed by lawmakers and regulators because it is normative. While it is a normative question, they still could have investigated how lawmakers and regulators currently (for non-algorithmic decisions) interpret "similarly situated" and derive a quantitative measure based on that. This would be an important contribution to bring together existing legal requirements with machine-learning decision making. But because they do not address this, it is hard to draw conclusions from their analysis. In part 2, they show that excluding race and correlating variables does not eliminate differences among the groups completely. Is this discriminatory or simply caused by the fact that the groups have different characteristics? How would the distribution of a non-discriminatory (in the legal sense) algorithm look like for that dataset? With a quantitative measure, these questions could be answered. Instead of using their vague "similarly situated" that they did not address further, they could have also incorporated prior fairness definitions [1]. For instance, I think it would be very interesting to use a threshold for the risk prediction and then analyze separation, i.e. if recall is equalized across groups. This would require another dataset with known default rates, but there are also other evaluation metrics for this dataset that I address below.

I think it is also quite problematic that they do not document their simulation mechanism. They only mention that the model is "constructed from the Boston HMDA data set". The dataset seems to be well suited for such an analysis, but the model that is used for the simulation can influence

Submitted as response essay for the *Building a Robot Judge: Data Science for Decision-Making* course at ETH Zurich. Do not distribute.

the results. Furthermore, there is prior literature that analyzed how much of the race discrepancies are due to other characteristics and which fraction is caused by the race [5]. Because they are simulating anyways, they could have incorporated these results. For instance, they could have created a synthetic dataset where there should not be any difference in lending rates by correcting for the other characteristics. This would be an alternative to the previously mentioned discrimination measure, as a non-discriminatory algorithm should have similar rates across races for this dataset.

Their choice of machine learning algorithms (Random Forest and Lasso) is also quite limited. An analysis of neural networks that are much harder to interpret would be an interesting addition. Furthermore, they only use Lasso for the analysis of feature importance. There are multiple ways to compute feature importance for random forests (Gini importance, permutation based feature importance, SHAP values), so this could be easily added to the analysis. Such an extension would allow to investigate whether feature importance scores across splits are more stable for random forests, potentially leading to interesting insights.

Part 4, where they suggest discrimination stress testing, is the most interesting part of the Essay in my opinion. However, it remains very shallow, only summarizes potential problems, and often refers to further work. They do not propose and evaluate any concrete method for the stress testing, which I would find very interesting.

An important aspect of algorithmic discrimination they did not address are feedback loops. When algorithms decide which applicants get a credit, they can influence their future training set (when they are later trained on the subset of applicants that they granted a credit), which can either reinforce or weaken biases. These dynamics can be quite complex (and are analyzed in other papers [2, 4]), but I think that the issue is also highly relevant for lawmakers when more and more decisions are made by machine-learning algorithms.

Overall, the Essay does not achieve the intended goal of "bringing together existing legal requirements with the structure of machine-learning decision-making" in my opinion. Their analysis of machine learning algorithms does not lead to any novel insights and the legal aspects are only addressed superficially. I still think that the topic is very important and that the overall setup could lead to interesting results if the previously mentioned suggestions are incorporated.

References

[1] Richard Berk et al. “Fairness in Criminal Justice Risk Assessments: The State of the Art”. In: *Sociological Methods & Research* 50.1 (Feb. 2021), pp. 3–44. ISSN: 0049-1241. DOI: 10.1177/0049124118782533. URL: https://doi.org/10.1177/0049124118782533 (visited on 12/12/2021).

[2] Alexander D’Amour et al. “Fairness Is Not Static: Deeper Understanding of Long Term Fairness via Simulation Studies”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* ’20. New York, NY, USA: Association for Computing Machinery, Jan. 2020, pp. 525–534. ISBN: 978-1-4503-6936-7. DOI: 10.1145/3351095.3372878. URL: https://doi.org/10.1145/3351095.3372878 (visited on 12/04/2021).

[3] Talia B Gillis and Jann L Spiess. “Big Data and Discrimination”. In: *The University of Chicago Law Review* (2019), pp. 459–487.

[4] Lydia T. Liu et al. “The Disparate Equilibria of Algorithmic Decision Making When Individuals Invest Rationally”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* ’20. New York, NY, USA: Association for Computing Machinery, Jan. 2020, pp. 381–391. ISBN: 978-1-4503-6936-7. DOI: 10.1145/3351095.3372861. URL: https://doi.org/10.1145/3351095.3372861 (visited on 12/04/2021).

[5] Alicia H. Munnell et al. “Mortgage Lending in Boston: Interpreting HMDA Data”. In: *The American Economic Review* 86.1 (1996), pp. 25–53. ISSN: 0002-8282. URL: https://www.jstor.org/stable/2118254 (visited on 12/04/2021).