Model-Agnostic Auditing: A Lost Cause?

Sakina Hansen[1], Joshua Loftus[1]

[1]London School of Economics, Houghton Street, London, United Kingdom, WC2A 2AE

Abstract

Tools for interpretable machine learning (IML) or explainable artificial intelligence (xAI) can be used to audit algorithms for fairness or other desiderata. In a black-box setting without access to the algorithm's internal structure an auditor may be limited to methods that are model-agnostic. These methods have severe limitations with important consequences for outcomes such as fairness. Among model-agnostic IML methods, visualizations such as the partial dependence plot (PDP) or individual conditional expectation (ICE) plots are popular and useful for displaying qualitative relationships. Although we focus on fairness auditing with PDP/ICE plots, the consequences we highlight generalize to other auditing or IML/xAI applications. This paper questions the validity of auditing in high-stakes settings with contested values or conflicting interests if the audit methods are model-agnostic.

Keywords

machine learning, artificial intelligence, supervised learning, black-box auditing, visualization, partial dependence plots, individual conditional expectation, causal models, counterfactual fairness

1. Introduction

Algorithm auditing is a rapidly growing field with little consensus about what makes an audit trustworthy [1]. Understanding the limitations of auditing methods is necessary to judge whether a particular audit is rigorous. To study these methods, we simulate the role of an external auditor who can only interact with the model by providing input data and recording the predicted outcome. This case is relevant to regulatory, oversight, or other competitive settings when an auditor can only use auditing methods that are model-agnostic [2, 3, 4]. We focus on the partial dependence plot (PDP) [5, 4], a popular tool for visualizing relationships between black-box input and output, and its close variants individual conditional expectation (ICE) [6] plots and conditional PDP. We demonstrate their limitations for fairness auditing through examples.

2. Theoretical Limitations of Black-Box Auditing for Fairness

Data Dependence. Model-agnostic explanations like PDPs depend on the joint distribution of data used to compute them. If part of the motivation of an audit is to understand the world by explaining the black-box, then unrepresentative data could lead to inaccurate conclusions about the world. Likewise, if an auditor is uninformed about the data selection process and

EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland
s.a.hansen1@lse.ac.uk (S. Hansen); J.R.Loftus@lse.ac.uk (J. Loftus)
0000-0002-2905-1632 (J. Loftus)

uses data from a biased sample to produce a PDP or other model explanation their audit may fail to detect unfairness in the pipeline. If the audit does use data from the same distribution as the training data, this leaves open questions of whether discrimination occurs if the model is deployed under conditions of distribution shift [7].

Unfairness via Mediators and Proxies. Model-agnostic explanations like PDPs will only show relationships with variables that are explicit inputs to an algorithm by definition. If a black-box does not take a sensitive attribute as an input it can still perform proxy discrimination [8], but a PDP may not uncover this. Additionally, due to the way PDPs average over other predictors they may hide indirect discrimination through mediating variables.

Interaction. PDPs are most effective at showing model dependence on each predictor if the model is additive, but can hide dependence if there are interactions [9]. This strong dependence on model structure complicates the interpretation of PDPs, especially in an auditing setting where we do not know the assumptions of the model fitting algorithm. ICE plots can help somewhat with this issue [4].

Attrition and Causality. In some real world examples, multiple sensitive attributes can interact resulting in intersectional discrimination [10, 11, 12, 13, 14]. One example of this is age related attrition, which has been studied related to unfairness to black defendants in the COMPAS case [15]. Attrition is also a relevant in health applications, where age and health interact with a number of socioeconomic factors [16]. In examples like these, attrition can violate the backdoor criterion, a requirement for causal interpretations of PDP [17]. Hence, the relationship uncovered by a model-agnostic explanation be a non-causal association that is not relevant to the purpose of the audit. Finally, causality raises issues about the interpretation of social categories as causal variables [18, 19] but can also help reveal differences between predictive algorithms and interventional policies [20]. [21]

3. Hiring Simulation

Algorithmic recruitment systems are emerging in the EU market [22, 23] and in many other places in the world [24], with the aims of accelerating hiring processing, and reducing errors and costs. These algorithms could exclude people from the job market with little human involvement or checking procedures, and so come with extensive risks for discrimination and unfairness [25, 26]. Our main simulation uses a synthetic causal model to generate data, consistent with the model $\mathcal{M}_S$ in Figure 1, with age and gender as variables that affect experience, which in turn affects chances of a job interview. The application rate decreases according to an interaction between age and gender, so that one gender group's application rate is 23% and the other group is 63%, for an overall application rate of 42% from an initial population of $n = 2000$ job seekers. Hence, the training data for the black-box models is not representative of the overall population. Experience increases with age but with different slopes depending on gender, potentially reflecting effects of unfairness at previous time points. Finally, interview probability increases positively with experience, and positively with age for one gender group

but negatively with age for the other gender group, again potentially reflecting unfairness (direct discrimination in this case) in the training data.

Through a series of experiments, we generate conditional PDPs with predictive models that assume different relationships between the variables and outcome:

1. Model $\hat{f}_E$ includes only experience as a predictor.
2. Model $\hat{f}_{\text{int}}$ includes experience, age, and gender as predictors with interaction effects (correctly specified).

Figure 1 shows both the limitation of model-agnostic explanations when the model does not adequately capture the data generating process (population vs training data) and the importance of analyzing fairness intersectionally rather than one attribute at a time (conditional vs unconditional PDPs).

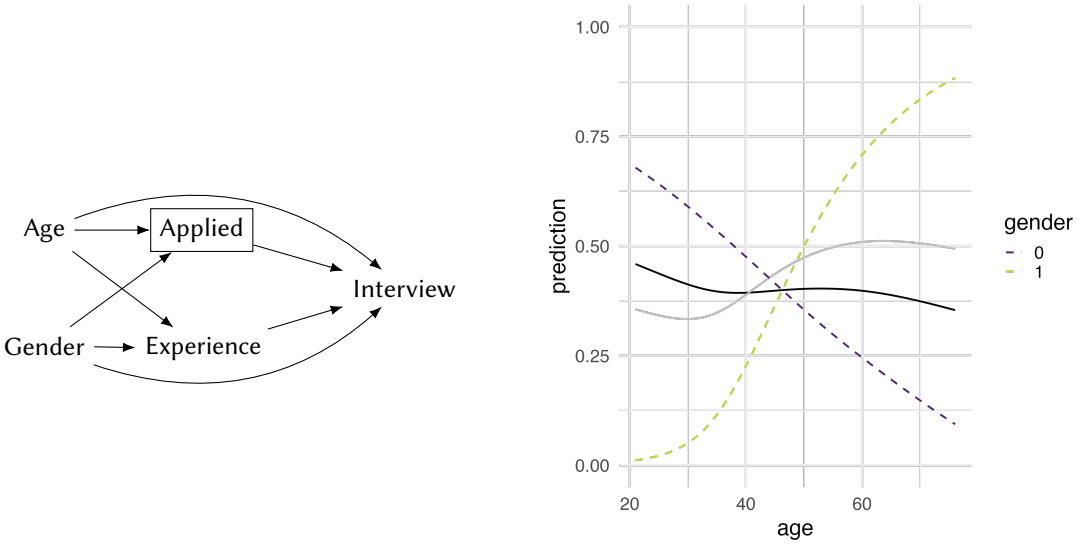

Figure 1: Left panel: Causal model $\mathcal{M}_s$. In this example age and gender can influence whether a person applies for a job, their employment experience, and also directly influence whether they are screened for being interviewed. Previous experience influences interview chances. It is necessary to apply in order to be interviewed, so the application variable is indicated in a box as a selection variable in the training data. Right panel: Solid lines are PDPs for $\hat{f}_E$, black for a model fit on the training data and gray fit on population data. Dashed lines are conditional PDPs for $\hat{f}_{\text{int}}$ using the training data.

4. Conclusion

We used fairness as an example objective for black-box audits and PDPs and related plots as example model-agnostic explanation methods. We show with examples several important ways these can fail to detect unfairness. Visual explanation methods may be convincing because "seeing is believing," so they have potential to be particularly deceptive if they are interpreted

without understanding their limitations. Our broader message calls into question the use of any model-agnostic explanation methods in the black-box audit setting. To make any valid conclusions from the explanations output by these tools, we must think beyond the input-output interface and consider causal structure in the real world, the sources of data used to train the model and generate the explanation, and choices of variables used to elaborate any univariate explanations. Important future work would look at extending the analysis present here to other model-agnostic explanation methods such as SHAP [27] and LIME [28]. We hope this encourages critical engagement and use in fairness contexts where explanations can obscure rather than reveal unfair discrimination if not used correctly.

References

[1] S. Costanza-Chock, I. D. Raji, J. Buolamwini, Who audits the auditors? recommendations from a field scan of the algorithmic auditing ecosystem, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 1571–1583.

[2] M. T. Ribeiro, S. Singh, C. Guestrin, Model-agnostic interpretability of machine learning (2016).

[3] A.-H. Karimi, G. Barthe, B. Balle, I. Valera, Model-Agnostic Counterfactual Explanations for Consequential Decisions, in: Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 895–905. URL: https://proceedings.mlr.press/v108/karimi20a.html, iSSN: 2640-3498.

[4] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2022. URL: https://christophm.github.io/interpretable-ml-book/.

[5] J. H. Friedman, Greedy function approximation: a gradient boosting machine, Annals of statistics (2001) 1189–1232.

[6] A. Goldstein, A. Kapelner, J. Bleich, E. Pitkin, Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation, journal of Computational and Graphical Statistics 24 (2015) 44–65.

[7] N. Kallus, A. Zhou, Residual unfairness in fair machine learning from prejudiced data, in: International Conference on Machine Learning, PMLR, 2018, pp. 2439–2448.

[8] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities, fairmlbook.org, 2019. http://www.fairmlbook.org.

[9] C. Molnar, G. König, J. Herbinger, T. Freiesleben, S. Dandl, C. A. Scholbeck, G. Casalicchio, M. Grosse-Wentrup, B. Bischl, Pitfalls to avoid when interpreting machine learning models (2020).

[10] K. W. Crenshaw, On intersectionality: Essential writings, The New Press, 2017.

[11] L. K. Bright, D. Malinsky, M. Thompson, Causally interpreting intersectionality theory, Philosophy of Science 83 (2016) 60–81.

[12] J. R. Foulds, R. Islam, K. N. Keya, S. Pan, An intersectional definition of fairness, in: 2020 IEEE 36th International Conference on Data Engineering (ICDE), IEEE, 2020, pp. 1918–1921.

[13] K. Yang, J. R. Loftus, J. Stoyanovich, Causal Intersectionality and Fair Ranking, in: K. Ligett, S. Gupta (Eds.), 2nd Symposium on Foundations of Responsible Computing (FORC 2021),

volume 192 of *Leibniz International Proceedings in Informatics (LIPIcs)*, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 2021, pp. 7:1–7:20. URL: https://drops.dagstuhl.de/opus/volltexte/2021/13875. doi:10.4230/LIPIcs.FORC.2021.7.

[14] A. Wang, V. V. Ramaswamy, O. Russakovsky, Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 336–349.

[15] C. Rudin, C. Wang, B. Coker, The age of secrecy and unfairness in recidivism prediction, Harvard Data Science Review 2 (2020) 1.

[16] D. M. Cutler, A. Lleras-Muney, T. Vogl, Socioeconomic status and health: dimensions and mechanisms (2008).

[17] Q. Zhao, T. Hastie, Causal interpretations of black-box models, Journal of Business & Economic Statistics 39 (2021) 272–281. URL: https://doi.org/10.1080/07350015.2019.1624293. arXiv:https://doi.org/10.1080/07350015.2019.1624293.

[18] I. Kohler-Hausmann, Eddie murphy and the dangers of counterfactual causal thinking about detecting racial discrimination, Nw. UL Rev. 113 (2018) 1163.

[19] L. Hu, I. Kohler-Hausmann, What's sex got to do with machine learning?, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 513–513.

[20] L. Bynum, J. Loftus, J. Stoyanovich, Disaggregated Interventions to Reduce Inequality, in: Equity and Access in Algorithms, Mechanisms, and Optimization, Association for Computing Machinery, New York, NY, USA, 2021, pp. 1–13. URL: https://doi.org/10.1145/3465416.3483286.

[21] L. Bynum, J. Loftus, J. Stoyanovich, Counterfactuals for the future, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2023.

[22] R. Xenidis, L. Senden, Eu non-discrimination law in the era of artificial intelligence: Mapping the challenges of algorithmic discrimination, in General Principles of EU law and the EU Digital Order (Kluwer Law International, 2020) (2019) 151–182.

[23] H. Parviainen, Can algorithmic recruitment systems lawfully utilise automated decision-making in the eu?, European Labour Law Journal 13 (2022) 225–248.

[24] L. Li, T. Lassiter, J. Oh, M. K. Lee, Algorithmic hiring in practice: Recruiter and hr professional's perspectives on ai use in hiring, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021, pp. 166–176.

[25] A. Kelly-Lyth, Challenging biased hiring algorithms, Oxford Journal of Legal Studies 41 (2021) 899–928.

[26] M. Buyl, C. Cociancig, C. Frattone, N. Roekens, Tackling algorithmic disability discrimination in the hiring process: An ethical, legal and technical analysis, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 1071–1082.

[27] E. Štrumbelj, I. Kononenko, Explaining prediction models and individual predictions with feature contributions, Knowledge and information systems 41 (2014) 647–665.

[28] M. T. Ribeiro, S. Singh, C. Guestrin, "Why should i trust you?" Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.