On the Role of Control in Auditing Risk Prediction Tools

Vijay Keswani[1]

[1]Duke University

Abstract

This paper explores the role of individual agency in algorithmic risk predictions. By comparing the relative *control* individuals have over various features used for risk prediction to the predictive relevance of these features, we formalize an audit procedure to assess the usability of risk prediction in practice.

Algorithmic risk assessment tools perform the task of predicting the context-relevant *risk* of individuals obtaining certain outcomes in the future using their historical and demographic data. Their usage in practice is becoming increasingly widespread and spans various societal domains, including recidivism risk prediction for pre-trial defendants Ghasemi et al. [2021], *early warning systems* to predict school dropout risks Mac Iver et al. [2019], and even disease prediction Pudjihartono et al. [2022]. The deployment of these tools is often justified by claims that they allow for targeted interventions for high-risk individuals to improve individual or societal-level outcomes. However, empirical audits of various risk prediction tools has questioned these claims. Various audits of recidivism risk prediction demonstrate inaccuracies and disparities in the predictive performance along racial lines ProPublica [2016], Goel et al. [2021]. Analyses of early warning systems used in public schools found that dropout risk scores from these tools were only as accurate as scores generated using students' environmental/group-level data Perdomo et al. [2023]. Similar issues of *stereotyping* and/or disparities in individual vs group-level predictive performance have been noted with the use of risk prediction in other domains Nabi et al. [2021], Brotcke [2022]. The evolving audit literature provides evidence to counter the often-dubious performance claims made when deploying risk assessment tools in practice. However, what lacks in this literature is the *underlying causal picture* that explains why certain risk prediction mechanisms are problematic to employ an individual-level. This work aims to fill this gap in the audit literature.

To understand the impact of algorithmic risk prediction tools on individuals, we argue that it is important to question whether the algorithm's predictions are based on features that are within an individual's control. By control, we refer to the past or future ability of the individual to change the way they are perceived by the risk prediction tool. This paper proposes an audit procedure that compares the *relative control* that an individual has over a subset of features to the *predictive relevance* of that feature subset to risk prediction. If risk predictions are predominantly based on features over which an individual has relatively low control (such as, race or parental income), then the risk prediction tool can be deemed to *unjust* and/or *impractical*, depending on what actions are considered when quantifying control.

To make the notions of *control* and *predictive relevance* more precise, suppose the risk prediction tool, denoted by $\mathbf{Risk} : \mathcal{X} \to [0, 1]$, takes as input a d-dimensional vector for each individual from the space $\mathcal{X} \in \mathbb{R}^d$ and return a risk score in the range $[0, 1]$. Let $X, Y, \hat{Y}$ denote the random variable corresponding to an individual's feature, their true outcome, and the output from the risk prediction tool respectively. Each individual x is represented by d features, i.e., $x := (x_1, x_2, \ldots, x_d)$. With this setup, quantifying *predictive relevance* of any feature subset amounts to computing how important the feature subset is to the output of $\mathbf{Risk}(\cdot)$, which can indeed be done using standard *feature importance* frameworks Covert et al. [2020]. For instance, one can quantify of predictive relevance of any feature subset $\mathcal{X}' \subset \mathcal{X}$ using the log-loss of predicting $\mathbf{Risk}(\cdot)$ with only the features from $\mathcal{X}'$; see Appendix A for details. We will denote predictive relevance of subset $\mathcal{X}'$ by pr($\mathcal{X}'$).

To quantify *relative control*, we first need to know what *actions* are available in the given setting. An individual can at any point take actions to change the value of one or more of their features; e.g., using *do*-calculus framework Pearl [2012], an action can be of the kind $A := do(X_i = x'_i)$, i.e., an individual takes action to change set feature i to value x'_i, Suppose $\mathcal{A}$ denotes all possible actions available to an individual. Since taking different actions will involve different levels of *difficulty* and we can characterize this using a cost function, defined as cost : $\mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}$. Then, $\text{cost}(x, A)$ denotes the cost paid by an individual with features x to perform the action A. With this notation, we can define what we mean by an individual's control over different features. Consider two feature $i, j \in [d]$. Let $\mathcal{A}_i, \mathcal{A}_j \subset \mathcal{A}$ denote actions that only operate over features i, j respectively, i.e., actions that only change the value of the corresponding feature while leaving everything else unchanged. Then, (with some abuse of notation) for an individual with features x, we can define $\text{cost}(x, \{i\}) := 1/|\mathcal{A}_i| \cdot \sum_{A \in \mathcal{A}_i} \text{cost}(x, A)$. An individual with features x has *relatively more control* over feature i compared to j if $\text{cost}(x, \{i\}) < \text{cost}(x, \{j\})$. To understand this setup, consider the simple example of the following features of any student: their GPA and their average school GPA. Both take the same range of values and the action space for both features is the same. However, for pretty much all actions, changing the average school GPA will be more costly for a student than changing their own GPA. This definition of cost can also be extended to account for relative control over feature combinations. This extension is specially important because many real-world features are often associated with each other, such that if features i and j are highly correlated, then in certain cases $\text{cost}(x, \{i, j\})$ could be smaller than $\text{cost}(x, \{i\})$ or $\text{cost}(x, \{j\})$ (i.e., changing these features together might be easier than changing just one of them). We can further generalize this to measure population-level average cost of updating features. For any feature subset $\mathcal{X}'$, we can define average action cost as $\text{cost}(\mathcal{X}') := \mathbb{E}_x[\text{cost}(x, \mathcal{X}')]$. Once again, $\text{cost}(\mathcal{X}'_1) < \text{cost}(\mathcal{X}'_2)$ implies that an average individual has a greater level of control over features in $\mathcal{X}'_1$ compared to features in $\mathcal{X}'_2$.

Coming back to the central question of this paper: **why is it important to consider relative individual-level control over various features when auditing risk predictions?** A two-dimensional analysis of predictive relevance vs cost associated with all feature subsets can provide crucial information about the usability of the risk prediction tool. First, we can rule out the trivial setting where $\text{pr}(\cdot)$ of all feature subsets is low; in this case, audit doesn't make sense as there is an obviously wide gap in the information used by the risk prediction tool and our audit setup. Hence, suppose there are feature subsets that indeed achieve high $\text{pr}(\cdot)$ values. Among these, the problematic cases are those where there's at least one feature subset that achieve high $\text{cost}(\cdot)$ values. That is, for any subset $\mathcal{X}'$, if both $\text{pr}(\mathcal{X}')$ and $\text{cost}(\mathcal{X}')$ are relatively large, then $\mathbf{Risk}(\cdot)$ predictions are highly dependent on the values taken by the features in $\mathcal{X}'$ and, at the same time, the average individual has relatively low control over changing the values of these features in $\mathcal{X}'$.

The implications of having features with high predictive relevance and high cost depends on the use case of risk prediction and the actions used to quantify the cost. For instance, in the case of school dropout risk prediction, the use case is to design targeted interventions to reduce the number of students dropping out from high school. The features used include student's education, socio-economic, and demographic attributes and the available actions cover steps that the students can take in the future, like improving GPA or changing schools, with the latter usually being relatively more costly than the former. For this setting, the findings of Perdomo et al. [2023] indicate that risk predictions obtained using school or district-level features are similar to those obtained using individual student-level features, indicating that school or district-level features have high predictive relevance and also high level of average action cost. High action cost implies that students have low control over the features used for risk prediction. Hence, even if this tool does identify students at high risk of dropout, it would be difficult to deploy individual-level interventions to help them reduce their dropout risk. School/district-level interventions might be more appropriate for this case. Another relevant example is the case of criminal recidivism risk prediction. Here, features include criminal history and demographic attributes of the defendants and use case is *justly* determining which defendants are safe release pre-trial. For this task, the relevant actions are those that were available to the defendant in the past, such as appearing for prior court dates or not participating in criminal activities. However, if features like race or socio-economic status have high predictive relevance, than risk predictions would be deemed to be unjust as they employ features over which the defendants have had little to no control throughout their lifetime.

Despite differences in use cases and available actions in the above two examples, our audit setup shows that the central issue with risk prediction in both applications is that prediction is dependent on features over which individuals have relatively low control. Hence, with this notion of control, our audit setup can be used identify potential issues of impracticality and/or injustice in risk prediction applications.

Finally, one major hurdle to this approach is quantifying $\text{cost}(x, \mathcal{X}')$, i.e. cost of updating features in the set $\mathcal{X}'$, given current feature values x. Ideally, knowing this requires access to the complete causal structure of the relevant context. To approximate cost, we can instead use a *matching-based* method based on the idea that if changing a feature requires low average cost, then (with large enough sample size) there would be *many* other individuals in the dataset who differ based on this feature's value but are *similar* in other aspects. That is, $\text{cost}(x, \mathcal{X}')$ can be approximated by the fraction of individuals for whom values of features in $\mathcal{X}'$ are different than x but all other feature values are same as x.

With this method for quantifying cost, our framework forwards a feasible audit system for practitioners to determine the usability of algorithmic risk prediction tools. Our work explores the role of agency in algorithmic predictions and can inform the wider causal ML literature on algorithmic recourse, causal fairness, and strategic classification.

References

Liming Brotcke. Time to assess bias in machine learning models for credit decisions. *Journal of Risk and Financial Management*, 15(4):165, 2022.

Ian Covert, Scott M Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33:17212–17223, 2020.

Mehdi Ghasemi, Daniel Anvari, Mahshid Atapour, J Stephen Wormith, Keira C Stockdale, and Raymond J Spiteri. The application of machine learning to a general risk–need assessment instrument in the prediction of criminal recidivism. *Criminal Justice and Behavior*, 48 (4):518–538, 2021.

Sharad Goel, Ravi Shroff, Jennifer Skeem, and Christopher Slobogin. The accuracy, equity, and jurisprudence of criminal risk assessment. In *Research handbook on big data law*, pages 9–28. Edward Elgar Publishing, 2021.

Martha Abele Mac Iver, Marc L Stein, Marcia H Davis, Robert W Balfanz, and Joanna Hornig Fox. An efficacy study of a ninth-grade early warning indicator intervention. *Journal of Research on Educational Effectiveness*, 12(3):363–390, 2019.

Junaid Nabi, Atif Adam, Sophia Kostelanetz, and Sana Syed. Updating race-based risk assessment algorithms in clinical practice: time for a systems approach. *The American Journal of Bioethics*, 21(2):82–85, 2021.

Judea Pearl. The do-calculus revisited. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 3–11, 2012.

Juan C Perdomo, Tolani Britton, Moritz Hardt, and Rediet Abebe. Difficult lessons on social prediction from wisconsin public schools. *arXiv preprint arXiv:2304.06205*, 2023.

ProPublica. Machine bias. 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Nicholas Pudjihartono, Tayaza Fadason, Andreas W Kempa-Liehr, and Justin M O'Sullivan. A review of feature selection methods for machine learning-based disease risk prediction. *Frontiers in Bioinformatics*, 2:927312, 2022.

A PREDICTIVE RELEVANCE SCORES

To capture the predictive relevance of any subset $\mathcal{X}' \subset \mathcal{X}$, we need to quantify the extent to which these features are used in the $\mathbf{Risk}$ function. There are multiple ways this can be accomplished. Let X, X' denote random variables for features corresponding to $\mathcal{X}$, $\mathcal{X}'$ respectively. Suppose we have access to a loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, that can be used to quantify the error associated with any risk prediction (e.g, simple log-loss function). Let Y denote the "true risk" associated with any given individual with features X. Assuming that (X, Y) follow an underlying joint distribution $\mathcal{D}$, the error associated with prediction function $\mathbf{Risk}(\cdot)$ can be quantified as

$$\mathbf{Error}(\mathbf{Risk}, \mathcal{D}) := \mathbb{E}_{X,Y\sim\mathcal{D}}\left[L(Y, \mathbf{Risk}(X)\right].$$

Using this framework, predictive relevance of any subset of features X', denoted by pr(X'), can be measured by the amount of increase in error due to the use of X' instead of X (assuming adding information always lead to same or lower amount of prediction error). In other words, we can quantify

$$\begin{aligned} \text{pr}_{\text{ideal}}(X') &= \mathbf{Error}(\mathbf{Risk}, \mathcal{D}') - \mathbf{Error}(\mathbf{Risk}, \mathcal{D}) \\ &= \mathbb{E}_{X',Y\sim\mathcal{D}'}\left[L(Y, \mathbf{Risk}(X')\right] - \mathbb{E}_{X,Y\sim\mathcal{D}}\left[L(Y, \mathbf{Risk}(X)\right], \end{aligned}$$

where $\mathcal{D}'$ is the joint distribution of X', Y. Note that the above measure requires access to true-risk scores Y. Access to these scores is not always guaranteed considering the true outcomes that define "true-risk" are often unobserved when/immediately after making the risk decisions (e.g., observing dropout or recidivism outcomes). In the absence of true outcomes, we can alternately define predictive relevance simply by how well we can predict $\mathbf{Risk}(X)$ just using features X', i.e., define

$$\text{pr}(X') := -\mathbb{E}_{X\sim\mathcal{D}_X}\left[L(\mathbf{Risk}(X'), \mathbf{Risk}(X)\right],$$

where $\mathcal{D}_X$ is the marginal distribution of $\mathcal{X}$ over variables from X.