Fairness-Aware Interactive Target Variable Definition

Abstract

How one defines one's target variable definition for algorithmic predictions and decisions can have profound fairness implications, since biases are often encoded in target variable definition itself. The downstream impacts of target variable definitions must be taken into account in order to responsibly develop, deploy, and use the algorithmic systems. We propose FairTargetSim (FTS), an interactive and simulations-based approach for this. We demonstrate FTS using the example of algorithmic hiring, grounded in real-world data and user-defined target variables. FTS is open-source; it can be used by algorithm developers, non-technical stakeholders, researchers, and educators in a number of ways.

1 Motivation

Machine learning requires translating real-world problems into numerical representations. For example, when developing an algorithm to predict which job applicants will be good employees, one must make precise the ambiguous, subjective notion of a "good" employee. How it is translated numerically—how one defines the target variable—can have profound implications for fairness (Passi and Jackson 2018; Obermeyer et al. 2019; Barocas et al. 2023). Defining "good" employee one way rather than another may result, e.g., in hiring fewer applicants from certain demographics. Such issues arise across domains. For a college admissions algorithm, one must determine who counts as a "good" student; for a search engine, one must determine what counts as a "good" search result; etc. How these notions are defined may also have weighty fairness implications: which applicants are admitted (Kizilcec and Lee 2023); which items appear at the top of search results (Phillips-Brown manuscript); etc. Thus, target variable definition is not an entirely technical matter; it requires careful human attention.

But all too often, target variables are defined in technical settings without such attention. And stakeholders who aren't a part of the technical process (e.g. managers in non-technical roles, or upper management) either don't understand, or are simply unaware of, the algorithmic fairness implications of target variable definition.

One way to help address this problem is to introduce non-technical audiences to target variable definition, and reveal its fairness implications for all audiences. So, we, an interdisciplinary team of philosophers and computer scientists, developed an *interactive target variable simulator*, FairTargetSim (FTS). FTS currently presents a scenario mirroring a widely-used style of hiring algorithm, based on psychometric tests. The user defines two target variables, using real-world psychometric test data of (Jaffe et al. 2022). FTS then trains two corresponding machine learning models, giving visualizations of how the models and training data differ in fairness and overall performance: algorithm developers and non-technical stakeholders can understand the implications of target variable definition in a simulated algorithmic system. They can use this understanding to more responsibly develop, deploy, and use machine learning in the real-world.

There is also an urgent need for responsible AI education and training in universities (Grosz et al. 2018), government, and the private sector (Eitel-Porter 2021). The ethical implications of technical issues can be challenging to explain. FTS illustrates them in an accessible, hands-on way.

FTS's code is open source; it can be extended to different algorithmic scenarios, datasets, models, fairness metrics, etc. **Find a live demo of FTS at** https://fairtargetsim.streamlit.app/**, and its anonymized code at** https://anonymous.4open.science/r/FairTargetSim-9C83/.

2 Related work

There are various extant systems to understand and address algorithmic bias: e.g. (Bellamy et al. 2019; Wexler et al. 2020; Liu et al. 2023). To our knowledge, FTS is the only system addressing target variable definition. Our focus on fairness-aware target variable definition also contrasts with previous demonstrations at AAAI and IJCAI on related subjects (e.g. (Vejsbjerg et al. 2024; Baumann et al. 2023; Henderson et al. 2021; Sokol and Flach 2018)).

3 FairTargetSim

FTS works with most modern browsers (Firefox is advised). It has four pages that the user visits in order. The first gives background on the problem. The others we explain below.

User Defines Target Variables

This page has the user define two different target variables (Figure 1), which FTS uses to train two models, A and B.

In cognitive-test-based hiring algorithms, developers often define "good" employee by having an employer identify

current employees whom the employer deems "good" for a given role (Wilson et al. 2021). These employees then play cognitive-test games, and a model is trained to identify applicants that share cognitive traits with these employees.

FTS uses support vector machine models to identify people who share cognitive traits with those who are identified as "good" employees. FTS's models are trained on data of real peoples' cognitive tests; the data we use is from Jaffe *et al.*'s (2022) battery 26, which has eleven tests that we grouped into five traits: memory, information processing speed, reasoning, attention, and behavioural restraint.[1]

However, FTS differs from the real-world settings in target variable definition: using sliders shown in Figure 1, the user explicitly defines the importance of five cognitive traits in what makes for a "good employee." The user does so twice, creating two different target variables. Then FTS calculates the weighted average of tests scores, given the slider weightings, assigning class label "0" to those in the bottom 85th percentile. From the top 15% subset, we randomly sample 100 "good" employees to whom we assign the class label "1" with weights ranging from 0.99 for the highest score candidate to 0.01 for the lowest score candidate, using a linear distribution for those in between. We assign a class label "0" to those not selected, thus introducing randomness. FTS then generates two labeled datasets and corresponding models, each with different target variable definitions.

We used the above approach in FTS because, first, Jaffe *et al.*'s dataset does not have employer-provided target variable labels, and, second, by using the sliders herself, a user can see how her very own choices in target variable definition can have implications for fairness. As we explain further in the next section, this is not a fundamental constraint; FTS can be extended to use real-world labels if available.

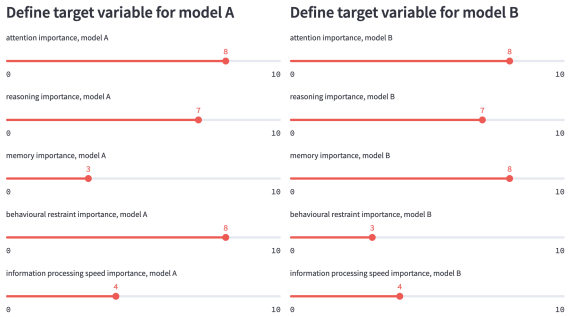

Figure 1: The user defines two target variables, using sliders representing the importance of traits of "good" employees.

Visualize Effects of Target Variable Definition

This page contains visualizations that illustrate how the user's two target variable definitions impact fairness and overall model performance. The visualizations are categorized into *Demographic* and *Non-demographic* sections, and further divided into categories that (i) show features of the models and (ii) features of the training data.

In the *Demographic* section, charts as in Figure 2 show how models A and B differ in, e.g., the proportions of selected applicants across demographic groups (gender, education level, age, and nationality). Other charts show how the models differ across groups with respect to "fairness metrics" (Hellman 2020), such as true and false positive rates and positive and negative predictive value. The differences are stark: different target variable definitions often result in major differences in the demographics of selected applicants and in fairness metrics (e.g. Figure 2). The *Demographics* section also shows how target variable definition affects the training data: e.g. how positive and negative labels are distributed across demographic groups. The *Non-demographic* shows how the models and training data differ in other ways: e.g. how the models rank particular applicants (Figure 3), overall model confusion matrices, and accuracy metrics.

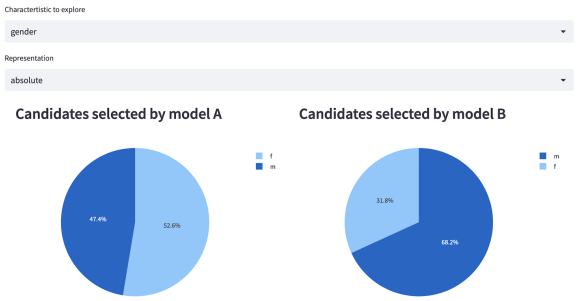

Figure 2: Charts display how the percentage of selected male and female applicants differs between models A and B.

Candidate ID	Ranking, model A	Ranking, model B	Predicted label A	Predicted label B
12	0.148166	0.987394	0	1
9	0.936614	0.978990	1	1
11	0.912242	0.960691	1	1

Figure 3: A table illustrates how individual applicants are evaluated differently by the two models.

Using FTS in the Real-World

This page gives guidance on adapting FTS for real-world use in hiring and other domains. As noted, FTS's code is open-source; an organization can extend FTS to use with their own data, models, target variables, fairness and accuracy metrics. In real-world settings, employers do not directly specify cognitive characteristics of "good" employees; they identify certain current employees as "good." We emphasize, in the pursuit of fairness, that this can be done in many ways. One could consult various managers on whom they judge "good," and weight these judgments in different ways, just as FTS weights the cognitive tests in different ways, resulting in different target variables. Alternatively, one could use various performance metrics to evaluate current employees (e.g., time to promotion, length of tenure, or role-specific metrics, such as sales volume in a sales role), and weight these metrics in different ways, resulting in different target variables, and paving the way for fairer algorithms.

[1]Our groups are: *Memory* (forward and reverse memory span, verbal list learning, delayed verbal list learning); *Information Processing Speed* (digit symbol coding, trail making parts A and B); *Reasoning* (arithmetic and grammatical); *Attention* (divided visual attention); *Behavioral Restraint* (go/no-go).

References

Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Baumann, J.; Castelnovo, A.; Cosentini, A.; Crupi, R.; Inverardi, N.; and Regoli, D. 2023. Bias on demand: investigating bias with a synthetic data generator. In *32nd International Joint Conference on Artificial Intelligence (IJCAI), Macao, SAR, 19-25 August 2023*, 7110–7114. International Joint Conferences on Artificial Intelligence Organization.

Bellamy, R. K. E.; Dey, K.; Hind, M.; Hoffman, S. C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilović, A.; Nagar, S.; Ramamurthy, K. N.; Richards, J.; Saha, D.; Sattigeri, P.; Singh, M.; Varshney, K. R.; and Zhang, Y. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5): 4:1–4:15.

Eitel-Porter, R. 2021. Beyond the promise: implementing ethical AI. *AI and Ethics*, 1: 73–80.

Grosz, B. J.; Grant, D. G.; Vredenburgh, K.; Behrends, J.; Hu, L.; Simmons, A.; and Waldo, J. 2018. Embedded EthiCS: Integrating Ethics Broadly Across Computer Science Education. arXiv:1808.05686.

Hellman, D. 2020. Measuring Algorithmic Fairness. *Virginia Law Review*, 106.

Henderson, J.; Sharma, S.; Gee, A.; Alexiev, V.; Draper, S.; Marin, C.; Hinojosa, Y.; Draper, C.; Perng, M.; Aguirre, L.; et al. 2021. Certifai: a toolkit for building trust in AI systems. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 5249–5251.

Jaffe, P. I.; Kaluszka, A.; Ng, N. F.; and Schafer, R. J. 2022. A massive dataset of the Neurocognitive Performance Test, a web-based cognitive assessment. *Scientific Data*, 9(1).

Kizilcec, R. F.; and Lee, H. 2023. Algorithmic fairness in education. In Holmes, W.; and Porayska-Pomsta, K., eds., *The Ethics of Artificial Intelligence in Education*. Routledge.

Liu, J.; Chen, H.; Shen, J.; and Choo, K.-K. R. 2023. FairCompass: Operationalising Fairness in Machine Learning. arXiv:2312.16726.

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453.

Passi, S.; and Jackson, S. J. 2018. Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).

Phillips-Brown, M. manuscript. Algorithmic neutrality. https://arxiv.org/abs/2303.05103.

Sokol, K.; and Flach, P. A. 2018. Glass-Box: Explaining AI Decisions With Counterfactual Statements Through Conversation With a Voice-enabled Virtual Assistant. In *IJCAI*, 5868–5870.

Vejsbjerg, I.; Daly, E. M.; Nair, R.; and Nizhnichenkov, S. 2024. Interactive Human-Centric Bias Mitigation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21): 23838–23840.

Wexler, J.; Pushkarna, M.; Bolukbasi, T.; Wattenberg, M.; Viégas, F.; and Wilson, J. 2020. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics*, 26(1): 56–65.

Wilson, C.; Ghosh, A.; Jiang, S.; Mislove, A.; Baker, L.; Szary, J.; Trindel, K.; and Polli, F. 2021. Building and Auditing Fair Algorithms: A Case Study in Candidate Screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 666–677.