Prisoners of Their Own Devices: How Models Induce Data Bias in Performative Prediction

José Pombal [1 2 3] **Pedro Saleiro** [1] **Mário A.T. Figueiredo** [2 3] **Pedro Bizarro** [1]

Abstract

The unparalleled ability of machine learning algorithms to learn patterns from data also enables them to incorporate biases embedded within. A biased model can then make decisions that disproportionately harm certain groups in society. Much work has been devoted to measuring unfairness in static ML environments, but not in dynamic, performative prediction ones, in which most real-world use cases operate. In the latter, the predictive model itself plays a pivotal role in shaping the distribution of the data. However, little attention has been heeded to relating unfairness to these interactions. Thus, to further the understanding of unfairness in these settings, we propose a taxonomy to characterize bias in the data, and study cases where it is shaped by model behaviour. Using a real-world account opening fraud detection case study as an example, we study the dangers to both performance and fairness of two typical biases in performative prediction: distribution shifts, and the problem of selective labels.

1. Introduction

With the increasing prominence of machine learning in high-stakes decision-making processes, its potential to exacerbate existing social inequities has been a reason of growing concern (Howard & Borenstein, 2018; Angwin et al., 2016; O'Neil, 2016). The goal of building systems that incorporate these concerns has given rise to the field of fair ML, which has grown rapidly in recent years (Mehrabi et al., 2021).

Fair ML research has focused primarily on devising ways to measure unfairness (Barocas et al., 2017) and to mitigate it in static algorithmic predictive tasks (Mehrabi et al., 2021; Caton & Haas, 2020). However, the vast majority of real-world use cases operate in dynamic environments, which feature complex, and unpredictable feedback loops that may exacerbate existing biases in the data and models. In such environments, model behaviour itself shapes the distribution of the data, so a deep understanding of data bias and its interaction with the ML model is required to uncover the causes of unfairness. That said, accounting for some notable exceptions (Fogliato et al., 2020; Wang et al., 2021; Blanzeisky & Cunningham, 2021; Jabbari et al., 2020; Akpinar et al., 2022) little attention has been heeded to relating unfairness to concrete bias patterns in the data.

To this end, we propose a domain-agnostic taxonomy to characterize data bias between a protected attribute, other features, and the target variable. It may be applied in dynamic environments, where bias during training can lie in stark contrast with that found in production. In particular, we use the taxonomy to model performative prediction settings, where data bias is induced by the predictive model itself. As a running example, we take an account opening fraud detection case study, which features two typical performative prediction bias phenomena: first, distribution shifts from fraudsters adapting to escape the fraud detection system; second, noisy labels arising from the selective labels problem, where the AI system determines the observed labels. We show how both issues, if left unaddressed, have detrimental, unpredictable, and sometimes unidentifiable consequences on fairness and performance.

2. Background & Related Work

Perdomo et al. (2020) define predictions as performative if they "influence the outcome they aim to predict". This influence usually reflects itself in distribution shifts over time, which, if left unaddressed, lead to degradation in predictive performance. In a lending scenario, Mishler & Dalmasso (2022) study the effects on fairness metrics of a classifier whose prediction changes the lending approval probability for a protected group. Estornell et al. (2021) point out the negative impact that adaptive agents have on the effectiveness of classifiers trained to be fair. Conversely, our work focuses on group-wise feature distribution shifts due to fraudsters adapting to the model over time — a *strategic*

[1]Feedzai [2]Instituto Superior Técnico, Universidade de Lisboa [3]Instituto de Telecomunicações. Correspondence to: José Pombal <jose.pombal@feedzai.com>.

ICML 2022 Workshop on Responsible Decision Making in Dynamic Environments, Baltimore, Maryland, USA, 2022.

classification (Dalvi et al., 2004) setting, which is a subset of performative prediction.

The selective label problem arises when the system under analysis determines the sample of observed labels (Barocas et al., 2019). For example, in fraud detection, rejecting an account opening based on the belief that it is fraudulent, implies that we will never observe its true label (the account never materializes). Here, the system fully determines the outcome of that observation, making it performative. Assuming the labels of these instances to be truthful causes a bias in the evaluation of performance and fairness of classification tasks, since most metrics — and particularly those one which we focus here (e.g.: FPR) — rely on accurate ground-truth labels. There is some work discussing the detrimental effects of noisy labels on algorithmic fairness (Fogliato et al., 2020; Jiang & Nachum, 2020; Lamy et al., 2019; Wang et al., 2021; Liao & Naghizadeh, 2022). The work of Dai & Brown (2020) is particularly pertinent, as it discusses the impact of label noise and shift on the reliability of fairness-aware algorithms. Our paper extends that work to a setting where label bias is induced by the classifier itself.

3. Bias Taxonomy

Throughout this work, we refer to the features of a dataset as X, the class label as Y, and the protected attribute as Z. The following definitions use the inequality sign ($\neq$) to mean a statistically significant difference.

Despite a multitude of definitions, there is still little consensus on how to characterize data bias (Mehrabi et al., 2021). In this paper, we propose a broad definition: there is bias in the data with respect to the protected attribute, whenever the random variables Y and X are sufficiently statistically dependent on Z. This does not mean a classifier trained on such data would be unfair, but rather that there is potential for it to be. In Section 4, we will explore how these biases may be induced by model behaviour over time.

Base Bias Condition

$$P[X, Y] \neq P[X, Y|Z]. \tag{1}$$

To satisfy this, Z must be statistically related to either X, Y, or both. The following biases imply this condition.

Group-wise Class-conditional Distribution Bias

$$P[X|Y] \neq P[X|Y, Z]. \tag{2}$$

Consider an example in account opening fraud in online banking. Assume that the fraud detection algorithm receives a feature which represents how likely the email-address is to be fake (X) and the client's reported age (Z) as inputs. In account opening fraud, fraudsters tend to impersonate older people, as these have a larger line of credit to max out, but use fake e-mail addresses to create accounts. Therefore, the e-mail address feature will be better to identify fraud instances for reportedly older people, potentially generating a disparity in group-wise error-rates, even if age groups have an equal likelihood of committing fraud in general.

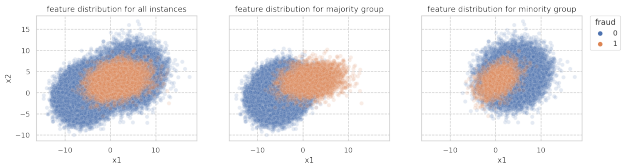

Figure 1. Group-wise Class-conditional Distribution Bias. There is clear class separability for the majority group (middle), *i.e.*, we can distinguish the fraud label using the two features. At the same time, there is virtually no separability for the minority group (right), as positive and negative samples overlap on this feature space. However, this is not discernible when looking at the marginal distribution for Y, x_1, and x_2 (left).

Noisy Labels Bias

$$P^*\left[Y|X, Z\right] \neq P\left[Y|X, Z\right], \tag{3}$$

where P^* is the observed distribution and P is the true distribution. That is, some observations belonging to a protected group have been incorrectly labeled. It is common for one protected group to suffer more from this ailment, if the labeling process is somehow biased. For example, women and lower-income individuals tend to receive less accurate cancer diagnoses than men, due to sampling differences in medical trials (Dressel & Farid, 2018). In fraud detection, label bias may arise due to the selective label problem.

Dynamic Bias

Let $\mathbf{BC}_i$ be a set of bias conditions **BC** on a data distribution i. Then, under dynamic bias, biases may change over time such that,

$$\boldsymbol{BC}_{train} \neq \boldsymbol{BC}_{deployment}. \tag{4}$$

This bias is the *bread and butter* of dynamic environments, as one of the main challenges in these domains is adapting to the disparities between training and deployment data. Indeed, distribution shifts may greatly affect model performance and fairness. In fraud detection, this can be particularly important, if we consider that fraudsters are constantly adapting to the model to avoid being caught. As such, a trend of fraud learned during training can easily become obsolete after deployment.

4. Case Study

4.1. Dataset and Methodology

In this work, we use a real-world large-scale case study of account-opening fraud (AOF). Each row in the dataset corresponds to an application for opening a bank account, submitted via the online portal of large retail bank. Data was collected over an 8-month period, and contains over 500K rows. The first 6 months are used for training and the remaining 2 months are used for testing, mimicking the procedure of a real-world production environment (we change this in Section 4.3). As a dynamic real-world environment, some distribution drift is expected along time, both from naturally-occurring shifts in the behavior of legitimate customers, as well as shifts in fraudsters' illicit behavior as they learn to better fool the production model.

Fraud rate (positive label prevalence) is about 1% in both sets. This means that a naïve classifier that labels all observations as *not fraud* achieves a test set accuracy of almost 99%. Such large class imbalance entails certain additional challenges for learning (He & Garcia, 2009) and calls for a specific evaluation framework. As such, performance will be measured as true positive rate (TPR) at a threshold of 5% false positive rate (FPR). TPR measures the percentage of detected fraud, and the FPR is limited at 5% as usually required by banks to avoid customer attrition (each FP is a legitimate application flagged as fraudulent, which can cause customers to want to change banks).

We will measure fairness as the ratio between group-wise FPR, also known as *predictive equality* (Corbett-Davies et al., 2017), which measures whether the probability of a legitimate person being flagged as fraudulent depends on the group they belong to. This fairness measure is by no means perfect, or enough to ensure fairness in many senses, but given our *punitive* setting, it is considered appropriate (Corbett-Davies et al., 2017; Saleiro et al., 2020).

4.2. Scenario 1: Adaptive Fraudsters

Fraud detection is a typical case of performative prediction for two reasons. First, the system determines the outcome of instances it flags as fraud: they are blocked, and so fraud never materializes. Second, fraudsters (the target of the predictor) actively adapt to evade the fraud detection system. This response emerges in the form of distribution shifts, where certain useful patterns to detect fraud in training become obsolete in production (post-deployment). Extending these concerns to fairness is straightforward, if we assume that fraudsters may leverage certain sensitive attributes in tandem with other features to escape detection. Indeed, one can use our proposed data bias taxonomy to model this as a combination of Group-wise Class-conditional Distribution Bias and Dynamic bias.

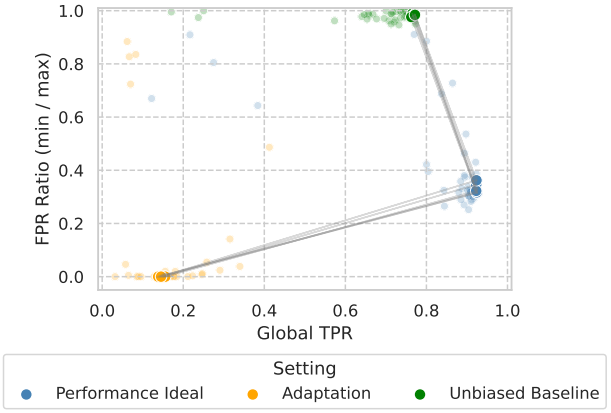

Figure 2. The opaque points in blue are the top 5 performing models on the setting in which the practitioner believes they are operating (one of these models would be chosen for production). The opaque points in yellow show the same models on the performative prediction setting on which the practitioner is actually operating, where fraudsters adapted their behavior. The opaque points in green are the models trained on a baseline setting, where the protected attribute is independent of X and Y. Arrows connect these model configurations, showing how the ones selected for production under 'Performance Ideal' are not the best in 'Adaptation'.

To illustrate this, given a synthetic binary protected attribute, we add two synthetic features x_1 and x_2 to our dataset, which made it easier to detect fraud for one protected group during the past, and is reflected in the current dataset. However, at test time, this groups' fraudsters adapt to the system, rendering these features useless for predicting fraud. Group sizes and fraud rates are made equal, such that there is no other bias in the data. We compare the performance and fairness of 50 LighGBM models under this setting (Adaptation) with two other cases: one where fraudsters did not change their behaviour (Performance Ideal), and one where the additional features did not exist (Unbiased Baseline).

Figure 2 shows how fraudsters adapting to the system in production had a harmful impact on both performance and fairness[1] of the top performing models. The former was to be expected, but the latter is somewhat surprising. Given that the additional features became uninformative in testing, it would have been desirable that the models converged to the performance and fairness equilibrium of the Unbiased Baseline, which did not have x_1 and x_2. However, models were "lazy", and did not learn some of the useful fraud patterns in the already-present features. Instead, they focused mostly on x_1 and x_2, missing out on a chance to increase both fairness and performance. Notice how the best models on "Performance Ideal" were not the best, or the fairest, after the fraudsters' adaptation. This highlights the impor-

[1] the extent of this degradation could have been smaller or larger, depending on the nature of the distribution shift.

tance of using ML methods that are robust to distribution shifts, especially in performative prediction environments. One such method would have been able to improve both fairness and performance.

4.3. Scenario 2: Selective Noisy Labels

Fraud detection suffers from an issue of selective labels, whereby the practitioner never gets to verify the true label of instances that are classified as fraud (predicted positives). For example, if we block the opening of an account, we can never confirm whether it would have been a fraudulent instance. Still, it is common practice to re-use these predictions in the training of future models as positive label examples. If one admits that a fraction of these observations are false positives, models will learn on noisy labels as time goes by. Thus, this problem can be framed as an instance of Noisy Labels Bias across time. Contrarily, the label of predicted negatives is eventually revealed, since these either materialize into fraud, or are in fact not fraud.

To assess the impact of this noise on fairness, we set up an experiment which mimics real-world fraud detection systems. We split the dataset into four temporal folds, starting with 3 months for training, 1 month for validation and 1 for testing. At each of four sliding-window iterations of training and evaluating models, we advance by 1 month, concatenating the previous validation set onto training, using the previous test set as validation, and moving on to a more recent test set. Importantly, the positive labels used to train and validate subsequent models are all the false negatives, and predicted positives of past models. This injects the type of label noise we mentioned above, with false positives being noisy label positives.

Two settings of the experiment are run. In one, a global threshold to achieve 5% FPR on the noisy validation set is used. This is the standard for fraud detection, as used in Scenario 1. In the other, we use group-wise thresholds — a popular post-hoc fairness intervention (Hardt et al., 2016) — such that *predictive equality* (equal group FPRs) is satisfied on the validation set. The goal is to assess whether unfairness observed in the first setting can be mitigated, or if the selective label bias renders the intervention fruitless. If the latter is the case, practitioners should tackle the selective label problem before trying to guarantee fairness. At each iteration, the best performing LightGBM on the validation set over 50 trials of TPE hyperparameter optimization (Bergstra et al., 2011) is evaluated.

Figure 3 compares group-wise FPRs after thresholding model scores in the noisy validation set, versus the FPRs the same model obtains in production (test set), when evaluated on real labels. Conditioned on the group, FPRs start off only slightly different due to natural distribution shifts between validation and test sets. However, they diverge as noise

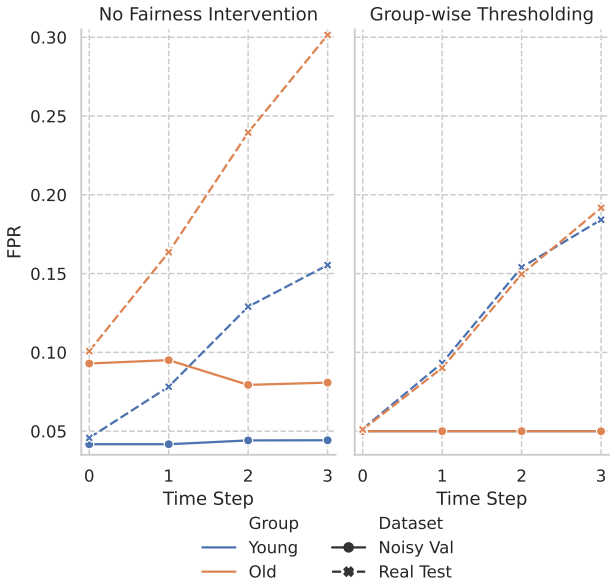

Figure 3. FPR over 4 time steps. Time Step 0 has no noise, but from then on some labels are noisy. The practitioner believes they are operating at 5% FPR (Noisy Val), but in production (Real Test), the FPR calculated on the true labels is much higher.

levels grow, increasing three-fold between the first and last iterations (30% v.s. 8% for the "Old" group, and 15% v.s. 15% for the "Young" group). The gap between blue and orange lines also widens in production, meaning higher levels of unfairness (about which the practitioner is unaware). Not only do these phenomena have harmful consequences on business, but the increased unfairness contributes to aggravate existing societal inequities. Even after the fairness intervention, the rift in group-wise FPRs shows a tendency to widen as label noise accumulates. Thus, mitigating the selective labels problem is of paramount importance in ensuring that systems are in fact fair in dynamic settings. We also tried dropping older training observations, a common industry practice due to storage and computational limitations. This seemed to attenuate the gap between perceived and real FPR, but to widen the disparity in group FPRs.

5. Conclusion

We proposed a data bias taxonomy to characterize the causes of unfairness in dynamic environments, where models shape the data distribution. In particular, we modelled two scenarios of bias in performative prediction: strategic classification, and selective noisy labels. We showed how both issues, if left unaddressed, have detrimental, unpredictable, and sometimes unidentifiable consequences on fairness and performance. We hope this work inspires future research on developing suitable fairness interventions for dynamic environments.

References

Akpinar, N.-J., Nagireddy, M., Stapleton, L., Cheng, H.-F., Zhu, H., Wu, S., and Heidari, H. A sandbox tool to bias(stress)-test fairness algorithms, 2022. URL https://arxiv.org/abs/2204.10233.

Angwin, J., Larson, J., Kirchner, L., and Mattu, S. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, May 2016.

Barocas, S., Hardt, M., and Narayanan, A. Fairness in machine learning. *NIPS tutorial*, 1:2017, 2017.

Barocas, S., Hardt, M., and Narayanan, A. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24, 2011.

Blanzeisky, W. and Cunningham, P. Algorithmic factors influencing bias in machine learning. *arXiv preprint arXiv:2104.14014*, 2021.

Caton, S. and Haas, C. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*, 2020.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic Decision Making and the Cost of Fairness. In *Proc. of the 23rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining - KDD '17*, pp. 797–806, New York, New York, USA, jan 2017. ACM Press. ISBN 9781450348874.

Dai, J. and Brown, S. M. Label bias, label shift: Fair machine learning with unreliable labels. In *NeurIPS 2020 Workshop on Consequential Decision Making in Dynamic Environments, 12 December 2020, Virtual Event*, Proceedings of Machine Learning Research. PMLR, 2020. URL https://dynamicdecisions.github.io/assets/pdfs/29.pdf.

Dalvi, N. N., Domingos, P. M., Mausam, Sanghai, S. K., and Verma, D. Adversarial classification. In Kim, W., Kohavi, R., Gehrke, J., and DuMouchel, W. (eds.), *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, pp. 99–108. ACM, 2004. doi: 10.1145/1014052.1014066. URL https://doi.org/10.1145/1014052.1014066.

Dressel, J. and Farid, H. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1), 2018.

Estornell, A., Das, S., Liu, Y., and Vorobeychik, Y. Unfairness despite awareness: Group-fair classification with strategic agents. *arXiv preprint arXiv:2112.02746*, 2021.

Fogliato, R., Chouldechova, A., and G'Sell, M. Fairness evaluation in presence of biased noisy labels. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2325–2336. PMLR, 26–28 Aug 2020.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29:3315–3323, 2016.

He, H. and Garcia, E. A. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

Howard, A. and Borenstein, J. The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. *Science and engineering ethics*, 24(5): 1521–1536, 2018.

Jabbari, S., Ou, H.-C., Lakkaraju, H., and Tambe, M. An empirical study of the trade-offs between interpretability and fairness. In *ICML 2020 Workshop on Human Interpretability in Machine Learning, preliminary version*, 2020.

Jiang, H. and Nachum, O. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 702–712. PMLR, 2020.

Lamy, A. L., Zhong, Z., Menon, A. K., and Verma, N. Noise-tolerant fair classification. *arXiv preprint arXiv:1901.10837*, 2019.

Liao, Y. and Naghizadeh, P. The impacts of labeling biases on fairness criteria. In *10th International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6):115:1–115:35, 2021. doi: 10.1145/3457607. URL https://doi.org/10.1145/3457607.

Mishler, A. and Dalmasso, N. Fair when trained, unfair when deployed: Observable fairness measures are unstable in performative prediction settings. *arXiv preprint arXiv:2202.05049*, 2022.

O'Neil, C. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2016.

Perdomo, J. C., Zrnic, T., Mendler-Dünner, C., and Hardt, M. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7599–7609. PMLR, 2020. URL http://proceedings.mlr.press/v119/perdomo20a.html.

Saleiro, P., Rodolfa, K. T., and Ghani, R. Dealing with bias and fairness in data science systems: A practical hands-on tutorial. In Gupta, R., Liu, Y., Tang, J., and Prakash, B. A. (eds.), *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pp. 3513–3514. ACM, 2020. doi: 10.1145/3394486.3406708. URL https://doi.org/10.1145/3394486.3406708.

Wang, J., Liu, Y., and Levy, C. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 526–536, 2021.