FairJob: A Real-World Dataset for Fairness in Online Systems

Mariia Vladimirova[*], **Eustache Diemert**
Criteo AI Lab
{m.vladimirova,e.diemert}@criteo.com

Federico Pavone
Université Paris Dauphine-PSL
federico.pavone@dauphine.psl.eu

Abstract

We introduce a fairness-aware dataset for job recommendation in advertising, designed to foster research in algorithmic fairness within real-world scenarios. It was collected and prepared to comply with privacy standards and business confidentiality. An additional challenge is the lack of access to protected user attributes such as gender, for which we propose a solution to obtain a proxy estimate. Despite being anonymized and including a proxy for a sensitive attribute, our dataset preserves predictive power and maintains a realistic and challenging benchmark. This dataset addresses a significant gap in the availability of fairness-focused resources for high-impact domains like advertising – the actual impact being having access or not to precious employment opportunities, where balancing fairness and utility is a common industrial challenge. We also explore various stages in the advertising process where unfairness can occur and introduce a method to compute a fair utility metric for the job recommendations in online systems case from a biased dataset. Experimental evaluations of bias mitigation techniques on the released dataset demonstrate potential improvements in fairness and the associated trade-offs with utility.

The dataset is hosted at https://huggingface.co/datasets/criteo/FairJob. Source code for the experiments is hosted at https://github.com/criteo-research/FairJob-dataset/.

1 Introduction

The intersection of technology and human dynamics presents both opportunities and challenges, particularly in the realm of artificial intelligence (AI). Despite advancements, persistent biases rooted in historical inequalities permeate our data-driven systems, perpetuating unfairness and exacerbating societal divides. Historical biases shape data collection, influencing AI model outcomes and often *amplifying* existing inequalities [Bolukbasi et al., 2016, Zhao et al., 2017, Chen et al., 2023]. Despite concerns regarding privacy, liability, and public relations, the collection of special and sensitive category data is crucial for bias assessments [Andrus et al., 2021]. Moreover, evolving legal frameworks, exemplified by the recent AI Act and General Data Protector Regulation [UK Information Commissioner's Office, 2022], mandate the detection, prevention, and mitigation of biases, while imposing some restrictions on the use of sensitive data.

Recent advances in fairness often involve computer vision, natural language processing and speech recognition tasks [Gustafson et al., 2023, Andrews et al., 2024, Hall et al., 2024, Schumann et al., 2024, Veliche and Fung, 2023], while lacking attention to algorithmic decision-making that involves *tabular data*, where each row represents an individual or an observation, and each column represents a feature or attribute [Le Quy et al., 2022, Zhang et al., 2021], resulting in a very few benchmark papers [Gorishniy et al., 2021, 2022, Grinsztajn et al., 2022, Shwartz-Ziv and Armon, 2022, Matteucci et al., 2023]. Tabular data is commonly used in various *high-risk domains* such as finance, healthcare, hiring, criminal justice, and advertising [van Breugel and van der Schaar, 2024].

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

Algorithmic discrimination in advertising can be related to sensitive verticals which highlights beneficial employment, financial and housing opportunities, or about who sees potentially less desirable advertising, such as ads for predatory lending services [Lambrecht and Tucker, 2019]. While unfairness in advertising is not punitive but rather assistive, i.e. fairness consists in providing equal access to precious opportunities, it is essential to *ensure fairness in advertising practices*. In some contexts such as housing or lending, such discrimination is explicitly *prohibited by law*[2]. Several studies conducted analyses on the fairness in advertising at different stages and observed discriminating behavior that was not necessarily intended by the ad-services [Speicher et al., 2018, Lambrecht and Tucker, 2019, Andreou et al., 2019, Ali et al., 2019]. This emphasizes the need for better mechanisms to audit and prevent bias in ads.

Most of studies on discriminating behavior in advertising were conducted via creating advertising campaigns and choosing targeted audiences and analysing the data from the user perspective without accessing the algorithmic features [Speicher et al., 2018, Lambrecht and Tucker, 2019, Andreou et al., 2019, Ali et al., 2019]. The absence of publicly available, realistic datasets leads researchers to publish results based on private data, resulting in non-reproducible claims [Geyik et al., 2019, Andreou et al., 2019, Timmaraju et al., 2023, Tang and Yu, 2022]. This poses challenges for critical evaluation and building upon previous work in the scientific community. Tang and Yu [2022] highlights the lack of public benchmarking datasets to study the fairness related approaches in advertising.

In addition, most of the studies assume that the AI systems have an access to the protected attributes which is often *unrealistic* due to privacy constraints or legal restrictions [Holstein et al., 2019, Lahoti et al., 2020, Molina et al., 2023, Timmaraju et al., 2023]. In online advertising, decision-makers usually have access to a log of user interactions with the system, which they can use to guess the attributes. However, the level of inaccuracy can be significant, making it difficult to ensure that an ad campaign reaches a non-discriminatory audience [Gelauff et al., 2020]. This makes it hard to meet fairness requirements [Lipton et al., 2018]. We emphasize the need for thorough research in real-world situations where *access to protected attributes is limited.*

Contributions. To foster research in fairness within real-world scenarios, we release a large-scale fairness-aware dataset for advertising. The dataset contains pseudononymized users' context and publisher features that were collected from a job targeting campaign ran for 5 months. The data has been *sub-sampled non-uniformly* to avoid disclosing business metrics. Feature names have been *anonymized* for business confidentiality, and their *values randomly projected* to preserve predictive power while making the *recovery of the original features or user context (i.e. re-indentification) practically impossible*, with accordance to the privacy-safety measures[3]. Although our dataset does not contain explicit sensitive attributes such as gender, it includes a *gender proxy derived from non-protected relevant attributes*, which we discuss in detail further.

This dataset provides a baseline according to the eligible audience generated by an advertiser's targeting criteria for a specific ad. This ensures that ads are tailored to individuals whom the advertiser can feasibly serve (such as those within a specific geographic region) and who are likely to be interested in their offerings, a practice already *governed by policies and standards in Housing, Employment, and Credit verticals*. Since advertiser targeting *adheres to policy constraints to prevent discriminatory practices*[4] (such as prohibiting the use of gender criteria in employment ads), the resulting eligible audience remains independent of prediction algorithms, serving as a reasonable baseline metric.

[2]According to Article 6(2) of the AI Act, targeted job advertising is considered as high-risk AI system. The Fair Housing Act in the United States makes it illegal to discriminate based on religion, color, national origin or gender for the sale, rental or financing of housing.

[3]'Pseudonymisation' of data – defined in Article 4(5) of General Data Protection Regulation (GDPR) – means replacing any information which could be used to identify an individual with a pseudonym, or, in other words, a value which does not allow the individual to be directly identified. The orginal data is still used in the AdTech company, however, re-indentification of the pseudomymized data is impossible due to additional randomization techniques during data anonymization. The original data does not contain any "special categories" of personal data listed under Article 9 of the GDPR, processing of which is prohibited, except in limited circumstances set out in Article 9 of the GDPR. The publication of the dataset was approved by a Data Protector Officer and Legal professionals.

[4]The Fair Housing Act in the United States makes it illegal to discriminate based on religion, color, national origin or gender for the sale, rental or financing of housing.

With the released dataset we examine the stages in the advertising process where unfairness can occur and explore techniques to mitigate such biases. Taking into account possible induced biases, we propose an unbiased utility metric that help to analyse different bias mitigation techniques. We also perform experiments on the released dataset to verify how we can improve fairness and the possible trade-offs with utility.

2 Related works

Open-source datasets. A limited availability of publicly available fairness-aware tabular datasets challenges research advancements in algorithmic fairness [Le Quy et al., 2022, Hort et al., 2023]. In 2022, Le Quy et al. [2022] studied datasets used at least 3 times in research publications on fairness, and found out there were only 15 open-source fairness datasets, most of which are criticized for being too small or far from real-world scenarios, including the most frequently used Adult [Dua and Graf, 2017] and COMPAS dataset [Larson et al., 2016]. Even though there is a positive tendency on addressing this issue by open-sourcing privacy-complying datasets, such as BAF [Jesus et al., 2022] for bank fraud detection where the data was obtained via data generation techniques, or WCLD [Ash et al., 2024], a curated large-scale dataset from circuit courts to address criminal justice, there is still *lack in available datasets* in other high-impact areas such advertising. It is important for academic researchers to have access to large datasets to study the problem rigorously [L. Cardoso et al., 2019, Li et al., 2022, Le Quy et al., 2022]. Large-scale datasets are advantageous as they increase the likelihood of capturing significant performance differences in experiments with new methods. With larger dataset sizes, the variance of metrics decreases, enabling more reliable and meaningful comparisons between different approaches.

Bias mitigation methods. The initial step to enhance model fairness is to exclude the protected attribute as a feature during training, a strategy known as *fairness through unawareness* [Chen et al., 2019]. However, this approach alone does not ensure fairness because the model may still learn correlations between other features and the protected attributes, see Section 3.1 and Figure 2b for details. To achieve a higher level of fairness, AI systems typically employ one of the additional methods: *pre-processing, in-training, or post-processing*. We refer to Hort et al. [2023] for the most up-to-date and thorough survey.

Fairness without demographics. The information on the protected attribute is often not available in practice [Holstein et al., 2019, Hort et al., 2023]. Several works studied limited availability of the protected attribute such as via a proxy [Gupta et al., 2018] or assuming there is a partial access to the information [Hashimoto et al., 2018, Awasthi et al., 2020, Molina et al., 2023]. Lahoti et al. [2020] relies on the assumption that protected groups are computationally-identifiable. However, if there were no signal about protected groups in the remaining features and class labels, we cannot make any statements about improving the model for protected groups. One of the possible solutions is to get data from secure multi-party computation [Veale and Binns, 2017, Kilbertus et al., 2018, Hu et al., 2019] or directly from users [Gkiouzepi et al., 2023]. However, these tools are still to be adapted to real-world situations. In addition, transfer leaning can be useful when there is little available data on the protected attributes [Coston et al., 2019].

3 Fairness in advertising

The aim of ad-tech companies is to deliver the most relevant advertisements to users navigating publishers' webpages. By matching users' browsing histories and content preferences with products that align with their interests, targeted advertising creates a mutually beneficial ecosystem [Wang et al., 2017, Choi et al., 2020]. Advertisers reach relevant audiences, users have access to free information and services in exchange of seeing ads related to their interests, and platforms profit from selling targeted ads.

Ad-tech companies grapple with vast volumes of noisy data, which encapsulate users' past actions. Leveraging this data, they predict potential clicks and conversions. However, if the data is biased, the algorithms can inadvertently perpetuate and even amplify these biases [Bolukbasi et al., 2016, Zhao et al., 2017, Chen et al., 2023]. It is crucial to scrutinize the predictors for bias and devise solutions to mitigate it. Failing to do so can result in discrepancies between offline evaluations and online

metrics, ultimately harming user satisfaction and trust in the service of online systems [Chen et al., 2023]. While advertising commonplace items carries little risk, companies must exercise caution with high-risk verticals like job offers [Speicher et al., 2018, Lambrecht and Tucker, 2019, Andreou et al., 2019, Ali et al., 2019]. For instance, *if managerial positions are disproportionately shown to men over women, more men may apply, perpetuating historical biases and exacerbating gender disparities*.

Bias can be introduced at several stages in the advertising process, see Figure 1. First, when a user visits a webpage with an ad slot, ad-tech companies participate in a real-time bidding (RTB) auction. During this auction, companies select a campaign (e.g., job offers or clothing) based on attributes of the publisher and the user, including their log of past interactions such as seen ads, their context, the fact of clicks on the ads, see Section 3.2. This auction must be organized in a fair way, respecting both the companies placing bids and the publishers providing ad slots, see Section 3.3. After an ad-tech company wins the display auction, there is the choice of which product to show (e.g., a senior position job or an assistant job). This selection can also introduce bias with respect to the user, see Section 3.4. Ensuring fairness at this stage is critical to preventing the reinforcement of existing inequalities.

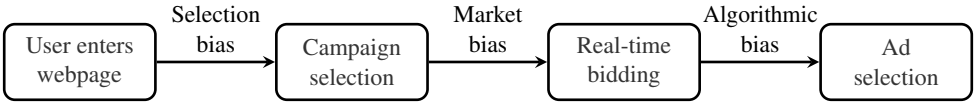
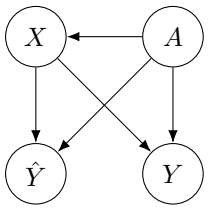


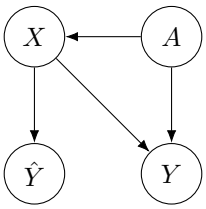
Figure 1: Simplified scheme of online advertising process of ad selection: (i) user enters a webpage with available banner for an ad, (ii) webpage sends a request to participate in the real-time bidding auction which triggers campaign selection by an ad service for a given user, (iii) after the campaign is chosen, ad-service sends a bid proposition, (iv) if the proposed bid won the auction, the recommendation engine chooses the best ad from the chosen campaign and shows it on the webpage.

3.1 Fairness definition

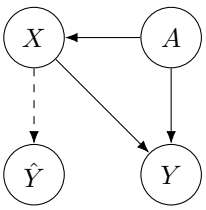
We base our discussion on a counterfactual fairness framework that explains the underlying connections between the variables in the system [Kusner et al., 2017]. Let A denote a protected attribute (can be a set of protected attributes) of an individual, X denote the other observable attributes of any particular individual, Y denote the outcome to be predicted, and let $\hat{Y}$ be a predictor. The predictor takes into account the available data from logs of user interactions with the system and product descriptions and estimates the probability of a positive outcome, i.e. click of the user on the product. The system takes into account the prediction and then shows the best product to the user, which results into a possible positive outcome. In our analysis, we are interested in understanding how A and X influence Y and how well our predictor $\hat{Y}$ captures these relationships. Our goal is not just to predict outcomes accurately but also to ensure fairness and mitigate biases in the predictions with respect to A. These random variables exhibit causal relationships, as modeled in Fig. 2, which we further explore in detail below.



(a) Unfairness with direct dependence: $\hat{Y} = f(X, A)$.



(b) Fairness through unawareness: $\hat{Y} = f(X(A))$.



(c) Fairness via zero mutual information: $\hat{Y} \perp\!\!\!\perp A|X$

Figure 2: Causal graph depicting effects of variables appearing during model training under different constraints. The arrow between the nodes corresponds to the causal effect. The dashed arrow between $\hat{Y}$ and X can be interpreted as $\hat{Y}$ depends on X, but conditionally on X, $\hat{Y}$ is independent to its ancestors.

It is important to understand how the information flows during the training procedure to create the prediction. If we give a protected attribute as a feature to the prediction model during training like in Fig. 2a, the model will learn directly the bias $A \to \hat{Y}$. The first step is to remove the protected attribute from the training features, in this case the model might learn the bias indirectly through the features $A \to X \to \hat{Y}$, see Fig. 2b. Thus, without bias mitigation techniques, the prediction model learns the bias that exists in the data. From causal perspective, the correction techniques correspond to blocking the information flow from A to $\hat{Y}$ by enforcing the zero mutual information between these variables conditioned on X, see Fig. 2c. We refer to Hort et al. [2023] for the survey on bias mitigation methods. Theoretically, we can mitigate the protected attribute bias when having access to the information that is used by an algorithm during training and having at least partial information about the protected attribute [Lahoti et al., 2020, Hort et al., 2023].

From a causal perspective, fair outcome with respect to a protected attribute means that it would not have changed if the other (counterfactual) value for the protected attributed was imposed [Kusner et al., 2017]:

$$\mathbb{P}\left(Y = y \,\middle|\, do(A = a), X = x\right) = \mathbb{P}\left(Y = y \,\middle|\, do(A = a'), X = x\right). \tag{1}$$

Since in a real-world scenarios it is not possible to have counterfactual estimation (we cannot impose a user to change their gender), we consider average values for groups, e.g. demographic parity[5]:

$$\mathbb{P}\left(Y = y | A = a\right) = \mathbb{P}\left(Y = y | A = a'\right) \tag{2}$$

This is in line with current laws that aim to ensure fairness in how housing and job ads are presented[6]. These laws do not focus on whether certain people are wrongly included or excluded (individual fairness), rather on making sure the ads are representative (group fairness). The key measurement is the difference in average that different groups will be shown the ad, regardless of how likely each group is to actually respond to the ad.

3.2 Selection bias in campaign choosing

In our setting, we are interested in assessing the fairness in specific campaign (e.g., job campaign) with respect to the protected attribute. For instance, we want to ensure that job advertisements for managerial roles are fair with respect to a binary protected attribute $A \in \{0, 1\}$ (e.g., gender). Typically, the data considered in this framework regards the job advertisements for users which have been assigned to the job campaign c. However, the campaign selection process might introduce selection bias, which should be taken in account. In particular, $\mathbb{P}(A = 1)$ and $\mathbb{P}(A = 0)$ are the (internet)-population level of a binary protected attribute. This might be approximated to the census population frequencies of the protected attribute. Let C be a random variable of choosing a campaign, then $\mathbb{P}(A = 1 \mid C = c)$ and $\mathbb{P}(A = 0 \mid C = c)$ are the frequencies of the protected attribute in the job campaign data c. These differ from the population levels due to selection bias.

Note that the recommendation engines predict $\mathbb{P}(Y = 1 \mid A = a, C = c)$ for a product in the campaign c. Thus, if we use prediction bias mitigation techniques while considering data at the campaign level, in the best case scenario, we obtain fair predictions while being unfair outside of campaing, $\mathbb{P}(\hat{Y} = 1 \mid A = 0, C = c) = \mathbb{P}(\hat{Y} = 1 \mid A = 1, C = c)$ and $\mathbb{P}(\hat{Y} = 1 \mid A = 0) \neq \mathbb{P}(\hat{Y} = 1 \mid A = 1)$. Thus, we have to take into account the selection bias to ensure demographic parity introduced in Eq. (2). Details on the derivation of the campaign selection bias and its correction are referred to supplemental material.

3.3 Market bias

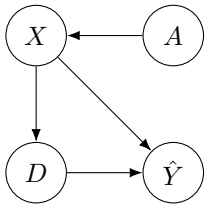
Lambrecht and Tucker [2019] found that women are a prized demographic, making them more expensive to advertise to. This implies that ads that are meant to be gender-neutral can be delivered in the way that appears to be discriminatory by RTB algorithms that focus on optimizing cost-effectiveness. Ali et al. [2019] explained that this is not solely the indication of the ingrained cultural

[5]We do not use *equal opportunity* metric because despite the name it actually preserves exiting bias in the case of assistive fairness, see Appendix for details.

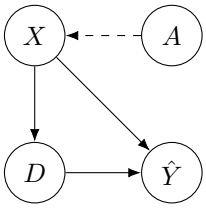
[6]The Fair Housing Act in the United States makes it illegal to discriminate based on religion, color, national origin or gender for the sale, rental or financing of housing.

bias nor a result of user profiles inputted into ads algorithms, but rather the product of competitive spillovers among advertisers. Additionally, the feedback loop mechanism considers imbalanced information–how recommendation systems expose content influences user behavior, which then becomes the training data for future predictions. This feedback loop not only introduces biases but also amplifies them over time, leading to a 'rich get richer' scenario known as the Matthew effect. [Chen et al., 2023]. Imbalanced data with respect to a protected attribute also effects the learning of a prediction, since an algorithm that receives in real time less data about one group, will learn at different speeds [Lambrecht and Tucker, 2020]. These effects are hard to estimate and should be addressed by the RTB process. Apart from users, advertisers can also be unfairly treated during the RBT auction process [Celis et al., 2019a, Chen et al., 2023] but here we focus solely on the user discrimination.

3.4 Recommendation bias



(a) Indirect unfairness: $\hat{Y} = f(X(A), D(A))$.



(b) Fairness, zero mutual information: $\hat{Y} \perp\!\!\!\perp A|X, D$

Figure 3: Causal graph depicting effects of variables appearing during model training for an ad recommendation system under different constraints.

In the ad recommendation system, the goal is to choose best products for a user for a given banner that can have several displays at the same time. The goal is to maximize the number of clicks for a given banner, meaning that there can be several products clicked. When we have several displays to show to a user, the display rank position becomes important and creates position bias with respect to a positive outcome. The influence of this bias is hard to estimate, however, it is important to take it into account [Singh and Joachims, 2018, 2019, Morik et al., 2020, Usunier et al., 2022].

Let J be a random variable denoting the set of banner to be shown to a user, D be a display (chosen product, i.e. job offer) shown to a user on a banner J. Let model $f(x, d)$ predicts the following positive outcome: $\mathbb{P}(Y = 1|X = x, D = d)$, i.e. the probability of a click for a chosen product d given user features x. As discussed above, we have to take into account the display position which expressed via variable rank R. However, the influence of the position on the utility is hard to estimate. Further, we suggest utility metrics for ads recommendation and in order to avoid the position bias, we suggest to compute them only on randomized displays, where the position of the products on the banner was chosen randomly.

Click-rank utility. The users' utility for a given model can be expressed as a positive engagement in the following way:

$$U(f) = \mathbb{E}_J \mathbb{E}_{X,D|J}\big[\mathbb{I}\left(Y_D = 1\right) \text{ rank}_J f(X_J, D)\big], \tag{3}$$

where $\mathbb{I}(Y_D = 1)$ is the identity function of a positive outcome (e.g. click) for display D. The function rank_D computes the ascending order rank within the set of displays for a banner J. This metrics is based on estimation of the positive outcome based on the passed events for chosen users.

Product-rank utility for biased data. We notice that the metrics for the algorithm can be biased due to the selection bias discussed in Section 3.2 because the prediction algorithm estimates $\mathbb{P}(Y = 1|A = a, C = c)$ instead of $\mathbb{P}(Y = 1|A = a)$. Even if we correct the prediction bias in $\mathbb{P}(Y = 1|A = a, C = c)$ based on the data provided for given campaign c, it does not correct the final bias in $\mathbb{P}(Y = 1|A = a)$ due to selection bias. We can adapt the click-rank utility to include possible selection bias into the metric, by explicitly considering that the product utility depends on a chosen campaign. Then, when correcting for the unfairness in the prediction, we might improve the utility metric taken into account the selection bias in the data:

$$\tilde{U}(f) = \mathbb{E}_D \mathbb{E}_{J|D} \left[\mathbb{I}\left(Y_D = 1\right) \frac{\mathbb{P}(A = a_{X_J})}{\mathbb{P}(A = a_{X_J} | C = c)} \text{rank}_J f(X_J, D) \right], \tag{4}$$

where a_X stands for a gender of a given user X. Intuitively, if the prediction is biased with respect to the protected attribute A, the final prediction $\mathbb{P}(Y = 1|A = a)$ is even more biased due to selection bias with respect to the protected attribute of choosing a campaign $C = c$: $\mathbb{P}(C = c \mid A = a)$. In this case, the prediction model amplifies the existing historical bias. However, we can remove the selection bias by adding weights that correspond to the presence of the protected attribute in the whole population and given the campaign. If the user with protected attribute $A = a$ has lower probability of click, and this group was underrepresented in the campaign $C = c$, i.e. $\mathbb{P}(A = a) > \mathbb{P}(A = a|C = c)$, then in the utility function, the model's prediction will be higher, by addressing the possible bias due to under-representation in the data. This is our suggested metric to evaluate the recommendation system when the selection bias is present and known such as in the FairJob dataset.

4 FairJobs dataset

We introduce FairJobs[7] dataset that contains fairness-aware data from a real-world scenario of advertising. The intended use of this dataset is to learn click predictions models and evaluate by how much their predictions are biased between different gender groups. The dataset consists of 1,072,226 rows that were collected during 5 months of a targeted job campaign[8], each row represents a job ad and user features: 20 categorical and 39 numerical features; label `click` (binary, if the ad was clicked), `protected_attribute` (binary, proxy for user gender, see below for more thorough explanation), `senior` (binary, if the job offer was for a senior position), [`user_id`, `impression_id`, `product_id`] are unqiue identifiers of user, impression and product (job ad). More details and dataset statistics are referred to Appendix.

Details on gender proxy. Since we do not directly access user demographics, we have to find a way to get a proxy of relevant attribute[9]. Most of recent works leverage the use of external data or prior knowledge on correlations to obtain proxies to relevant attributes [Gupta et al., 2018, Hashimoto et al., 2018, Awasthi et al., 2020, Lahoti et al., 2020]. We define a product gender, either given by a client, either by a category of the product. This gives us approximately 40% of products gender identified. Then, we follow the available statistics and choose the gender proxy based on the dominant gender of products the user interacts with. This gender proxy identities a behavior of a user, i.e. if a user tends to buy female or male products. The gender proxy does not necessarily correlate with the gender, as it often happens with the proxy variables [Gelauff et al., 2020]. Verification of the accuracy of these approximations is challenging. Additionally, if there are no signal about protected groups in the remaining features and class labels, we cannot make any statements about improving the model for protected groups [Lahoti et al., 2020].

Limitations and interpretation. We remark that the proposed gender proxy does not give a definition of the gender. Since we do not have access to the sensitive information, *this is the best solution we have identified at this stage to idenitify bias on pseudonymised data*, and we encourage any discussion on better approximations. This proxy is reported as binary for simplicity yet we acknowledge gender is not necessarily binary. Although our research focuses on gender, this should not diminish the importance of investigating other types of algorithmic discrimination. While this dataset provides important application of fairness-aware algorithms in a high-risk domain, there are several fundamental limitation that can not be addressed easily through data collection or curation processes. These limitations include historical bias that affect a positive outcome for a given user, as well as the impossibility to verify how close the gender-proxy is to the real gender value. Additionally,

[7] https://huggingface.co/datasets/criteo/FairJob

[8] We leave the details on the data collection, feature engineering and privacy-preserving steps in the supplemental material.

[9] This gender proxy is used with the AdTech company to create in-market audiences: when a client can choose to show advertising of their products to "people that buy more female products" or "people that buy more male products". We note that *this proxy is not used in the prediction engine* and for campaign creation *on high-risk verticals*. The campaign creation is governed by policies and standards in Housing, Employment, and Credit verticals that complies with the Fair Housing Act in the USA.

there might be bias due to the market unfairness that we explained in Section 3.3. Such limitations and possible ethical concerns about the task should be taken into account while drawing conclusions from the research using this dataset. Readers should not interpret summary statistics of this dataset as ground truth but rather as *characteristics of the dataset* only. Additional limitation comes from identifying the `senior` position label, as the definition of what constitutes a "senior" position can be subjective. We acknowledge that this method may introduce some noise, particularly if job titles are unconventional or if errors occur in categorization. Finally, we remark that in the dataset we assume that each `user_id` represents a single user; however, we acknowledge that multiple users could share one device, potentially affecting the user features. Additionally, `user_id`'s are based on the company's identification technology, which means that a single user could have multiple `user_id`'s across different browsing sessions. This is one of the complexities inherent in real-world data, particularly in online advertising. Such biases can influence model training, bias evaluation, and bias mitigation efforts.

5 Empirical observations

Challenges. The first challenge comes from handling the different types of data that are common in tables, the *mixed-type columns*: there are both numerical and categorical features that have to be embedded [Gorishniy et al., 2021, 2022, Grinsztajn et al., 2022, Shwartz-Ziv and Armon, 2022, Matteucci et al., 2023]. In addition, some of the features have long-tail phenomenon and products have popularity bias, see Figure 4. Our datasets contains more than 1,000,000 lines, while current high-performing models are under-explored in *scale*, e.g. the largest datasets in Grinsztajn et al. [2022] are only 50,000 lines, while in Gorishniy et al. [2021, 2022] only one dataset surpasses 1,000,000 lines. Additional challenge comes from *strongly imbalanced data*: the positive class proportion in our data is less than 0.007 that leads to challenges in training robust and fair machine learning models [Jesus et al., 2022, Yang et al., 2024]. In our dataset there is no significant imbalances in demographic groups users regarding the protected attribute (both genders are sub-sampled with 0.5 proportion, female profile users were shown less job ad with 0.4 proportion and slightly less senior position jobs with 0.48 proportion), however, there could be a hidden effect of a bias that we discussed in Section 3. This poses a problem in accurately assessing model performance [van Breugel et al., 2024]. More detailed statistics and exploratory analysis are referred to the supplemental material.

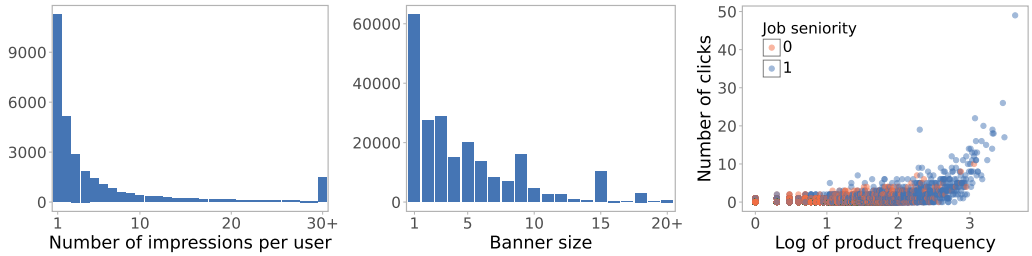


Figure 4: Examples of some feature statistics in FairJob dataset: number of impressions per user and banner size have long tail phenomenon (two plots on the left). The products have popularity bias (right plot), i.e. some products have much higher or lower than average number of clicks with senior job ads having more clicks on average.

Baselines. We choose two baseline regimes: (i) unfair, that uses all attributes for training, including the protected one; and (ii) unaware, that corresponds to fairness through unawareness, i.e. using all attributes during training except the protected one. We train (i) a `Dummy` classifier in the unaware regime to obtain the first baseline and (ii) `XGB` in two regimes (unaware and unfair) to achieve a more reasonable baseline performance, see Table 1. We notice that the `Dummy` classifier is perfectly fair with respect to demographic parity `DP` which is reasonable since it did not learn the dependence between features at the label at all, as can be seen from the negative log-likelihood `NLLH` results. However, the utility metrics U and $\tilde{U}$ of `Dummy` do not differ much from `XGB` which is due to very strong imbalance in the data. We remark that $\tilde{U}$ is expectedly higher for fairer (unaware) models, while U is better for the unfair model. There is a slight difference in terms of `NLLH` and `AUC-ROC` for

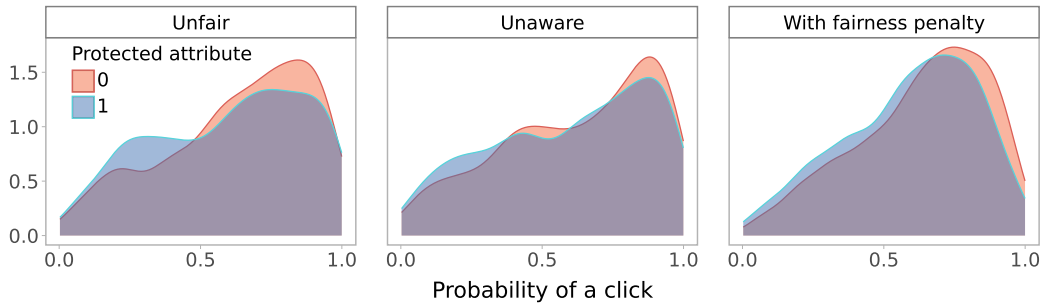


Figure 5: Probability density distributions of click for different values of the protected attribute of three models trained in different ways: (i) *unfair* – with a protected attribute included as a feature during training, (ii) *unaware* – corresponds to fairness through unawareness, (iii) trained *with fairness penalty* as a bias mitigation technique.

`XGB` models, and higher U corresponds to higher `AUC-ROC`. We refer to Appendix B and C for more details.

Table 1: Performance comparison for single simulation of Dummy classifier (unaware) and XGBoost (unaware and unfair) with 100 trials for tuning.

	`NLLH` ↓	`AUC` ↑	`DP` ↓	U ↑	$\tilde{U}$ ↑
`Dummy` unaware	0.69239	0.50000	**0.00000**	0.01009	0.01245
`XGB` unaware	**0.05491**	0.75787	0.00278	0.01017	**0.01276**
`XGB` unfair	0.05736	**0.76201**	0.00323	**0.01037**	0.01236

Fair regime. To study a possible trade-off between utility and fairness, we use in-processing fairness methods of adding a fairness-inducing penalty to a loss function $\mathcal{L}(\hat{Y}, Y)$ during training [Kamishima et al., 2011]:

$$\hat{Y} = \arg\min \quad \mathcal{L}(\hat{Y}, Y) + \lambda \cdot \text{Penalty}(\hat{Y}, Y, A), \tag{5}$$

where parameter λ, or `fairness_multiplier`, controls the trade-off between the model's predictive accuracy $\mathcal{L}(\hat{Y}, Y)$ and fairness. Adjustments on λ allows to control the importance of fairness relative to accuracy. These methods remove the influence of protected attribute on the model's output without restrictions on the data [Kamishima et al., 2011, Bechavod and Ligett, 2017, Mary et al., 2019]. We implement a penalty based on the approach described in Bechavod and Ligett [2017].

We train a logistic regression in the two baseline regimes and compare the results of these algorithms with a (iii) fair model that is trained without protected attribute with an additional fairness-inducing penalty [Kamishima et al., 2011, Bechavod and Ligett, 2017]. The three models correspond to the situations described in Figure 2. We refer all the reproducibility details and additional experiments to Appendix B and C. The resulted prediction can be visually compared in Figure 5.

Fairness-utility trade-off. We illustrate the possible trade-off between performance and fairness metrics for the logistic regression model when varying `fairness_multiplier`, see Figure 6. We notice that for positive `fairness_multiplier`, `DP` improves, while `NLLH` degrades. For `fairness_multiplier = 0.5` and `1.0` we notice slight improvements in utility metrics, especially $\tilde{U}$, with respect to the unfair model represented as a dashed line.

Additionally, we propose to restraint an access to the protected attribute and study the trade-off when we train the model on the whole train set but add fairness penalty only for some percentage of train set. In some scenarios, we could see improvements in fairness without sacrificing the overall performance. The loss in accuracy due to the imposed fairness constraints is often small as also noted in other works [Celis et al., 2019b]. We explore bias correction techniques tailored to address data limitations and preserve utility in large-scale. We demonstrate how prioritizing fairness in AI not only benefits users by fostering inclusivity but also contributes to the long-term success and ethical integrity of companies.

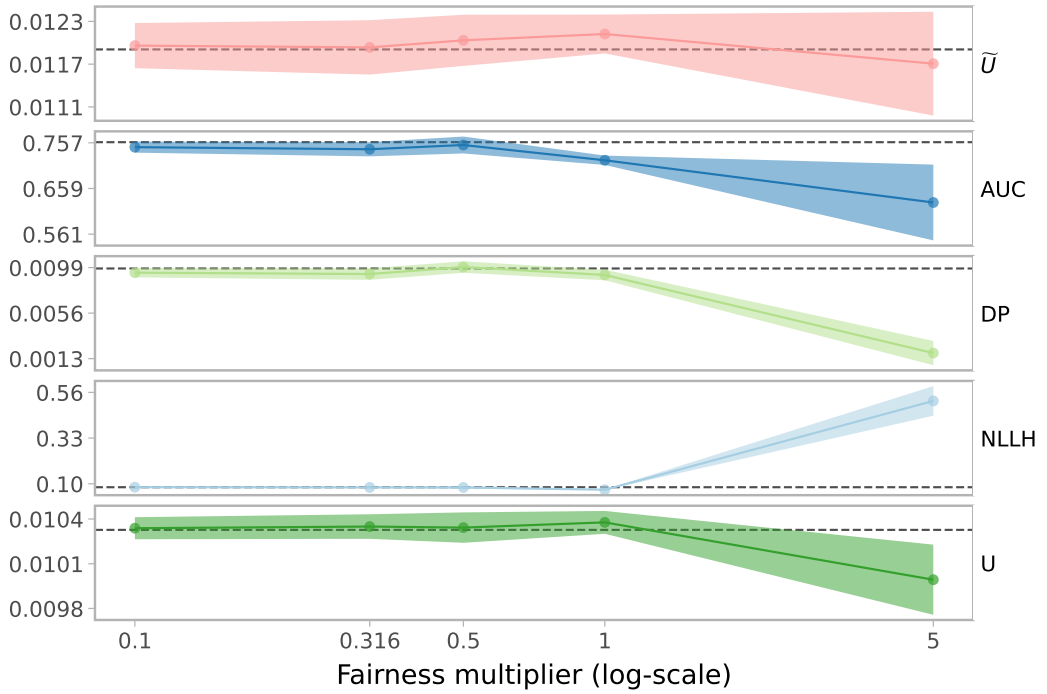


Figure 6: The trade-off between performance and fairness metrics for the logistic regression model when varying the fairness multiplier from 0.1 to 5.0, the variance is reported based on 10 iterations. The dashed line represents unaware logistic regression.

These findings suggest that the trade-off relationship between accuracy and fairness is context-dependent. It highlights the need for further research to better understand the conditions under which the accuracy fairness trade-off arises and identify strategies to mitigate or overcome it.

6 Conclusion

Addressing bias in AI goes beyond mere compliance with legal frameworks like the AI Act; it necessitates proactive measures to detect, prevent, and mitigate biases. Drawing from real-world challenges faced by industries, we highlight the limitations of existing bias mitigation strategies, particularly in environments where access to sensitive user attributes is restricted. We encourage other authors and practitioners to experiment with different AI or Fair AI algorithms on this dataset. We argue that specific problems can often be generalized to broader contexts. For example, if our dataset helps identify a method that effectively balances fairness and utility, this method could potentially be applicable to other recommendation systems across various domains. We expect that with this work, the quality of evaluation of novel AI methods increases, potentiating the development of the area, see more details in the broader impact section in Appendix E. Additionally, we hope it encourages other similar relevant datasets to be published from other authors and institutions.

7 Acknowledgements

Federico Pavone has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101034255. We also thank Martin Bompaire, David Rhode and Andre Cunha for helpful discussions.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *ACM SIGKDD international conference*

on knowledge discovery & data mining, 2019.

Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes. *ACM on Human-Computer Interaction*, 2019.

Athanasios Andreou, Márcio Silva, Fabrício Benevenuto, Oana Goga, Patrick Loiseau, and Alan Mislove. Measuring the Facebook advertising ecosystem. In *Network and Distributed System Security Symposium*, 2019.

Jerone Andrews, Dora Zhao, William Thong, Apostolos Modas, Orestis Papakyriakopoulos, and Alice Xiang. Ethical considerations for responsible data curation. *Advances in Neural Information Processing Systems*, 2024.

McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *ACM Conference on Fairness, Accountability, and Transparency*, 2021.

Elliott Ash, Naman Goel, Nianyun Li, Claudia Marangon, and Peiyao Sun. WCLD: Curated large dataset of criminal cases from Wisconsin Circuit Courts. *Advances in Neural Information Processing Systems*, 2024.

Pranjal Awasthi, Matthäus Kleindessner, and Jamie Morgenstern. Equalized odds postprocessing under imperfect group information. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Yahav Bechavod and Katrina Ligett. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*, 2017.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, 2016.

Elisa Celis, Anay Mehrotra, and Nisheeth Vishnoi. Toward controlling discrimination in online ad auctions. In *International Conference on Machine Learning*, 2019a.

L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Conference on Fairness, Accountability, and Transparency*, 2019b.

Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Conference on Fairness, Accountability, and Transparency*, 2019.

Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems*, 2023.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

Hana Choi, Carl F. Mela, Santiago R. Balseiro, and A. Leary. Online display advertising markets: A literature review and future directions. *Information Systems Research*, 31(2):556–575, 2020.

Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. Fair transfer learning with missing protected attributes. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2019.

Konstantin Donhauser, Javier Abad, Neha Hulkund, and Fanny Yang. Privacy-preserving data release leveraging optimal transport and particle gradient descent. *International Conference on Machine Learning*, 2024.

Dheeru Dua and Casey Graf. Uci machine learning repository. 2017. URL http://archive.ics.uci.edu/ml.

Lodewijk Gelauff, Ashish Goel, Kamesh Munagala, and Sravya Yandamuri. Advertising for demographically fair outcomes. *arXiv preprint arXiv:2006.03983*, 2020.

Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. Fairness-aware ranking in search and recommendation systems with application to linkedin talent search. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

Eleni Gkiouzepi, Athanasios Andreou, Oana Goga, and Patrick Loiseau. Collaborative ad transparency: Promises and limitations. In *IEEE Symposium on Security and Privacy*, 2023.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2021.

Yury Gorishniy, Ivan Rubachev, and Artem Babenko. On embeddings for numerical features in tabular deep learning. *Advances in Neural Information Processing Systems*, 2022.

Leo Grinsztajn, Edouard Oyallon, and Gael Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems*, 2022.

Maya R Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. Proxy fairness. *arXiv preprint arXiv:1806.11212*, 2018.

Laura Gustafson, Chloe Rolland, Nikhila Ravi, Quentin Duval, Aaron Adcock, Cheng-Yang Fu, Melissa Hall, and Candace Ross. FACET: Fairness in computer vision evaluation benchmark. In *International Conference on Computer Vision*, 2023.

Siobhan Mackenzie Hall, Fernanda Gonçalves Abrantes, Hanwen Zhu, Grace Sodunke, Aleksandar Shtedritski, and Hannah Rose Kirk. VisoGender: A dataset for benchmarking gender bias in image-text pronoun resolution. *Advances in Neural Information Processing Systems*, 2024.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 2016.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, 2018.

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *CHI Conference on Human Factors in Computing Systems*, 2019.

Max Hort, Zhenpeng Chen, Jie M Zhang, Mark Harman, and Federica Sarro. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing*, 2023.

Hui Hu, Yijun Liu, Zhen Wang, and Chao Lan. A distributed fair machine learning framework with private demographic data protection. In *International Conference on Data Mining*, 2019.

Sérgio Jesus, José Pombal, Duarte Alves, André Cruz, Pedro Saleiro, Rita Ribeiro, João Gama, and Pedro Bizarro. Turning the tables: Biased, imbalanced, dynamic tabular datasets for ML evaluation. *Advances in Neural Information Processing Systems*, 2022.

Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *International Conference on Data Mining Workshops*, 2011.

Niki Kilbertus, Adrià Gascón, Matt Kusner, Michael Veale, Krishna Gummadi, and Adrian Weller. Blind justice: Fairness with encrypted sensitive attributes. In *International Conference on Machine Learning*, 2018.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in Neural Information Processing Systems*, 2017.

Rodrigo L. Cardoso, Wagner Meira Jr, Virgilio Almeida, and Mohammed J. Zaki. A framework for benchmarking discrimination-aware models in machine learning. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2019.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Advances in Neural Information Processing Systems*, 2020.

Anja Lambrecht and Catherine Tucker. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management science*, 2019.

Anja Lambrecht and Catherine E Tucker. Apparent algorithmic discrimination and real-time algorithmic learning. *Social Sciences Research Network*, 2020.

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. COMPAS Broward county dataset. 2016. URL https://github.com/propublica/compas-analysis.

Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(3), 2022.

Nianyun Li, Naman Goel, and Elliott Ash. Data-centric factors in algorithmic fairness. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2022.

Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. Does mitigating ML's impact disparity require treatment disparity? *Advances in Neural Information Processing Systems*, 2018.

Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. *Advances in Neural Information Processing Systems*, 2019.

Kelong Mao, Jieming Zhu, Liangcai Su, Guohao Cai, Yuru Li, and Zhenhua Dong. FinalMLP: an enhanced two-stream MLP model for CTR prediction. In *Conference on Artificial Intelligence*, 2023.

Jérémie Mary, Clément Calauzenes, and Noureddine El Karoui. Fairness-aware learning for continuous attributes and treatments. In *International Conference on Machine Learning*, 2019.

Federico Matteucci, Vadim Arzamasov, and Klemens Böhm. A benchmark of categorical encoders for binary classification. *Advances in Neural Information Processing Systems*, 2023.

Daniele Micci-Barreca. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *ACM SIGKDD Explorations Newsletter*, 2001.

Mathieu Molina, Nicolas Gast, Patrick Loiseau, and Vianney Perchet. Trading-off price for data quality to achieve fair online allocation. In *Advances in Neural Information Processing Systems*, 2023.

Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. Controlling fairness and bias in dynamic learning-to-rank. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Candice Schumann, Femi Olanubi, Auriel Wright, Ellis Monk, Courtney Heldreth, and Susanna Ricco. Consensus and subjectivity of skin tone annotation for ML fairness. *Advances in Neural Information Processing Systems*, 2024.

Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

Ashudeep Singh and Thorsten Joachims. Policy learning for fairness in ranking. *Advances in Neural Information Processing Systems*, 2019.

Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Nunes Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna P Gummadi, Patrick Loiseau, and Alan Mislove. Potential for discrimination in online targeted advertising. In *Conference on Fairness, Accountability and Transparency*, 2018.

Xiaoli Tang and Han Yu. Towards trustworthy ai-empowered real-time bidding for online advertisement auctioning. *arXiv preprint arXiv:2210.07770*, 2022.

Aditya Srinivas Timmaraju, Mehdi Mashayekhi, Mingliang Chen, Qi Zeng, Quintin Fettes, Wesley Cheung, Yihan Xiao, Manojkumar Rangasamy Kannadasan, Pushkar Tripathi, Sean Gahagan, et al. Towards fairness in personalized ads using impression variance aware reinforcement learning. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2023.

Ryan Turner, David Eriksson, Michael McCourt, Juha Kiili, Eero Laaksonen, Zhen Xu, and Isabelle Guyon. Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020. In *NeurIPS 2020 Competition and Demonstration Track*, 2021.

UK Information Commissioner's Office. What do we need to do to ensure lawfulness, fairness, and transparency in AI systems? 2022. URL https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/how-do-we-ensure-fairness-in-ai/.

Nicolas Usunier, Virginie Do, and Elvis Dohmatob. Fast online ranking with fairness of exposure. In *ACM Conference on Fairness, Accountability, and Transparency*, 2022.

Boris van Breugel and Mihaela van der Schaar. Why tabular foundation models should be a research priority. *International Conference on Machine Learning*, 2024.

Boris van Breugel, Nabeel Seedat, Fergus Imrie, and Mihaela van der Schaar. Can you rely on your model evaluation? improving model evaluation with synthetic test data. *Advances in Neural Information Processing Systems*, 2024.

Michael Veale and Reuben Binns. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 2017.

Irina-Elena Veliche and Pascale Fung. Improving fairness and robustness in end-to-end speech recognition through unsupervised clustering. In *International Conference on Acoustics, Speech and Signal Processing*, 2023.

Jun Wang, Weinan Zhang, and Shuai Yuan. Display advertising with real-time bidding (RTB) and behavioural targeting. *Foundations and Trends in Information Retrieval*, 11(4-5):297–435, 2017.

Zeyu Yang, Peikun Guo, Khadija Zanna, and Akane Sano. Balanced mixed-type tabular data synthesis with diffusion models. *International Conference on Machine Learning*, 2024.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *International Conference on World Wide Web*, 2017.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, 2013.

Daniel Zhang, Saurabh Mishra, Erik Brynjolfsson, John Etchemendy, Deep Ganguli, Barbara Grosz, Terah Lyons, James Manyika, Juan Carlos Niebles, Michael Sellitto, et al. The AI index 2021 annual report. *arXiv preprint arXiv:2103.06312*, 2021.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *Conference on Empirical Methods in Natural Language Processing*, 2017.

Jieming Zhu, Jinyang Liu, Shuai Yang, Qi Zhang, and Xiuqiang He. Open benchmarking for click-through rate prediction. In *ACM International Conference on Information & Knowledge Management*, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Section 4, paragraph "Limitations and interpretation".
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Section Reproducilibity in supplemental meterial
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Section Reproducilibity in supplemental meterial
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We repeated experiments 10 times with random seeds.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Section Reproducilibity in supplemental meterial
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

8 Checklist for the dataset track

1. Submission introducing new datasets must include the following in the supplementary materials:

(a) Dataset documentation and intended uses. Recommended documentation frameworks include datasheets for datasets, dataset nutrition labels, data statements for NLP, and accountability frameworks. A: Available on the dataset page and in the text in paragraph "License and intended use".

(b) URL to website/platform where the dataset/benchmark can be viewed and downloaded by the reviewers. A: The dataset is hosted on HuggingFace dataset API: FairJob, the other information can be found in supplemental material.

(c) URL to Croissant metadata record documenting the dataset/benchmark available for viewing and downloading by the reviewers. You can create your Croissant metadata using e.g. the Python library available here: https://github.com/mlcommons/croissant The dataset viewer automatically generates the metadata in Croissant format (JSON-LD) for every dataset on the Hugging Face Hub. It lists the dataset's name, description, URL, and the distribution of the dataset as Parquet files, including the columns' metadata. The Croissant metadata is available for all the datasets that can be converted to Parquet format.

(d) Author statement that they bear all responsibility in case of violation of rights, etc., and confirmation of the data license. A: the statement is in the supplemental material.

(e) Hosting, licensing, and maintenance plan. The choice of hosting platform is yours, as long as you ensure access to the data (possibly through a curated interface) and will provide the necessary maintenance. A: all the details are in supplemental material and on the Hugging Face dataset webpage.

2. To ensure accessibility, the supplementary materials for datasets must include the following:

   (a) Links to access the dataset and its metadata. This can be hidden upon submission if the dataset is not yet publicly available but must be added in the camera-ready version. In select cases, e.g when the data can only be released at a later date, this can be added afterward. Simulation environments should link to (open source) code repositories.

   (b) The dataset itself should ideally use an open and widely used data format. Provide a detailed explanation on how the dataset can be read. For simulation environments, use existing frameworks or explain how they can be used.

   (c) Long-term preservation: It must be clear that the dataset will be available for a long time, either by uploading to a data repository or by explaining how the authors themselves will ensure this.

   (d) Explicit license: Authors must choose a license, ideally a CC license for datasets, or an open source license for code (e.g. RL environments).

   (e) Add structured metadata to a dataset's meta-data page using Web standards (like schema.org and DCAT): This allows it to be discovered and organized by anyone. If you use an existing data repository, this is often done automatically.

   (f) Highly recommended: a persistent dereferenceable identifier (e.g. a DOI minted by a data repository or a prefix on identifiers.org) for datasets, or a code repository (e.g. GitHub, GitLab,...) for code. If this is not possible or useful, please explain why.

   A: All the dataset details including links, metadata, preservation, license, reproducility details are avaiable in the supplemental material. Licence and intended use information is also stated in the main paper text.

3. For benchmarks, the supplementary materials must ensure that all results are easily reproducible. Where possible, use a reproducibility framework such as the ML reproducibility checklist, or otherwise guarantee that all results can be easily reproduced, i.e. all necessary datasets, code, and evaluation procedures must be accessible and documented. A: All details for the reproducibility are available in the supplemental material.

4. For papers introducing best practices in creating or curating datasets and benchmarks, the above supplementary materials are not required.

Supplementary materials

We firstly describe in detail how the dataset FairJob was collected, then provide all the information on the context and features with its statistics. Further, we perform experiments and provide all the steps for the sake of reproducubility. The dataset is hosted at https://huggingface.co/datasets/criteo/FairJob. Source code for the experiments is hosted at https://github.com/criteo-research/FairJob-dataset/.

Author statement of responsibility. Authors and Criteo bear all responsibility in case of violation of rights and confirmation of the data license.

A Dataset information

A.1 Data collection and use

As illustrated in Figure 7, the process starts with users navigating Publisher and Advertiser websites (typically newspapers and retailer shops respectively). Upon user consent[10], user information about the events such as visits or product views[11] are collected and identified by means of browser cookies. Users are subject to personalized advertising (if the job campaign was chosen and the display opportunity was won) until the end of the data collection period. Subsequently, only won displays coming from Publisher and Advertiser partners are joined by cookie identifier on the AdTech platform to form the raw dataset, dropping cookie ids when they are not needed anymore.

Further, the data is collected by trackers upon a page call and is commonly stored as a tuple of page and client information, which we call a click. Several steps are in place to ensure that clicks on ads must result from a human user with genuine interest. This steps are stated in the guidelines that are accepted by publishers[12]. Any method or mechanism that artificially generates clicks or impressions is strictly prohibited, including but not limited to the following:

- Clicks generated by publishers clicking on their own ads.
- Repeated clicks on the same ad unit.
- Publishers encouraging users to click on their ads (examples may include: any language encouraging users to click on ads; ad implementations that may cause a high volume of accidental clicks; monetary or non-monetary incentives or rewards for clicks, etc.).
- Use of automated clicking tools or traffic sources, robots, or other deceptive software, click spam or click injections.
- Use of any other artificial mechanism to generate or inflate clicks.
- Clicks generated outside of the ad surface will not be counted as intended clicks.

Additional safeguards are related to clearly stating that the user sees an ad. Any native ad formats displayed on the publisher's site must be clearly marked as an ad. Ads should not be placed very close to or underneath buttons or any other object such that the placement of the ad interferes with a user's typical interaction with the Site content or functionalities.

Publishers accept the guidelines which state the following:

- Publishers may offer their users the opportunity to view ads in exchange for user rewards or incentives, but only if the user is not forced or incentivized to interact with the ad, such as incentivized click.
- Publishers must not, directly or indirectly, provide incentives to users in exchange for clicks on ads, or make use of any mechanism or monetizable reward to incentivize clicks.

To ensure confidentiality, the collected data has been *sub-sampled non-uniformly* to avoid disclosing business metrics. Feature names have been *anonymized*, and their values *randomly projected* to

[10] In accordance with GDPR and Article 82 of the Data Protection Act that require the provider to ask consent of data subjects if it was reading/writing information to the user's device.

[11] The original data does not contain any "special categories" of personal data listed under Article 9 of the GDPR, processing of which is prohibited, except in limited circumstances set out in Article 9 of the GDPR.

[12] The guidelines are available on the official webpage: https://www.criteo.com/supply-partner-guidelines/.

preserve predictive power while rendering the recovery of the original features or user context practically impossible[13]. The dataset does not contain the relevant attributes such as gender; however, it includes a gender proxy which we discuss in detail in the main text.

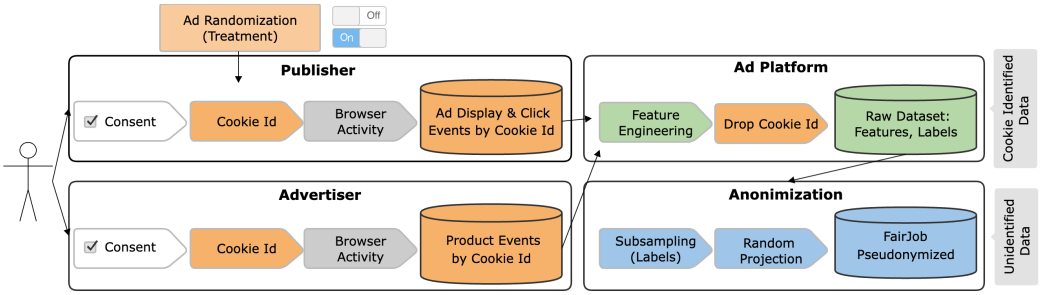


Figure 7: Data collection and processing overview.

License and intended use. The data is released under the CC-BY-NC-SA 4.0 license which gives liberty to Share and Adapt this data provided that the respect of the Attribution, NonCommercial and ShareAlike conditions. We focus in this paper on algorithmic fairness analysis as a specific use-case in job-seeking contexts, but it does not have to restricted to this. Fairness in job advertising is a crucial step towards ensuring fairness in any advertising campaign involving sensitive topics. The FairJob datasets presents a challenge of training fair and robust models on strongly imbalanced data [Jesus et al., 2022, Yang et al., 2024]. While many of the features in the FairJob dataset (such as user features and categorical product features) are not specific to job advertising and can be applied to other domains for improving click-prediction methods, however, the consumer behavior may vary significantly from a job seeker's behavior. Additionally, the dataset can be used as a baseline for improving deep learning models on tabular data and to study methods on embedding creation for numerical and categorical features as done by Gorishniy et al. [2021, 2022], Grinsztajn et al. [2022], Shwartz-Ziv and Armon [2022], Matteucci et al. [2023]. Another possible usage is to explore privacy-preserving techniques for tabular data as in Donhauser et al. [2024]. Moreover, this dataset can be used to improve the generation techniques on tabular data, as described in Jesus et al. [2022], van Breugel and van der Schaar [2024]. We compare FairJob dataset details to other commonly used open-source tabular datasets further in Section D.

A.2 User privacy protection

The original dataset does not include any data that directly identifies a user, such as names, postal addresses, or email addresses in plain text, so the *original data is pseudonymized*. However, due to possible uniqueness in high-dimensional datasets, it has been shown that some records in pseudonymized datasets might be re-identifiable by an adversary. We describe a possible re-identification process by an adversary and demonstrate that it is practically impossible.

In the FairJob dataset, there are both user and product features. We assume that only user features could potentially be used to identify a user. The product features are categorical, based on internal company categories across different products and brands, or on information provided by clients, and thus are not related to users.

We can suppose that an adversary runs a company that collects user browsing features and, therefore, has an auxiliary dataset with user information. The adversary's goal is to identify a user from the FairJob dataset. Since all features are anonymized, the adversary cannot directly match the features and must first align FairJob feature distributions with the adversary's company's feature distributions and then find a unique association from a FairJob record to the adversary's company record, thereby identifying the user. However, since we anonymized, added noise, and standardized continuous user features, re-identification becomes impossible as the original scale is not provided. The only features of potential interest to the adversary are two categorical user features.

[13]The orginal data is still used in the AdTech company, however, re-indentification of the pseudomymized data is impossible due to additional randomization techniques during data anonymization to comply with Article 4(5) of GDPR.

The categorical user features are based on internal company categories, which are unlikely to be known to the adversary due to business confidentiality, making it impossible for the adversary to reconstruct them. Additionally, features `cat0` and `cat1` have the same cardinality, and their distributions are not significantly different, making them indistinguishable and effectively reducing the number of possible unique variations. Most importantly, both features `cat0` and `cat1` have a cardinality of 9, meaning the tuples (`cat0`, `cat1`) can have a maximum of 81 variations among 1 million records. In FairJob, each present variation of (`cat0`, `cat1`) corresponds to at least 10 records, making it impossible for an instance (`cat0`, `cat1`) to serve as a pseudonym for a user, i.e., user re-identification based on these two features is not possible.

All these measures make it *impossible to re-identify users* from the FairJob dataset.

A.3 Dataset detailed description

The dataset contains pseudononymized users' context and publisher features that was collected from a job targeting campaign ran for 5 months by Criteo AdTech company. Each line represents a product that was shown to a user. Each user has an impression session where they can see several products at the same time. Each product can be clicked or not clicked by the user. The dataset consists of 1072226 rows and 55 columns:

- `user_id` is a unique identifier assigned to each user. This identifier has been anonymized and does not contain any information related to the real users.
- `product_id` is a unique identifier assigned to each product, i.e. job offer.
- `impression_id` is a unique identifier assigned to each impression, i.e. online session that can have several products at the same time. An impression, or successful online session, is recorded when a product banner (or multiple banners) is loaded and begins to render on the publisher's page. An impression can subsequently lead to a click event.
- `cat0`, . . . , `cat1` are anonymized categorical user features.
- `cat2`, . . . , `cat12` are anonymized categorical product features.
- `num13`, . . . , `num47` are anonymized numerical user features.
- `protected_attribute` is a binary feature that describes user gender proxy, i.e. female is 0, male is 1. The detailed description on the meaning can be found in the main paper.
- `senior` is a binary feature that describes the seniority of the job position, i.e. an assistant role is 0, a managerial role is 1. It includes both managerial and individual contributor (IC) roles, as our aim was to encompass any senior-level positions. This feature was created during the data processing step from the product title feature: if the product title contains words describing managerial role (e.g. 'president', 'ceo', and others), it is assigned to 1, otherwise to 0.
- `rank` is a numerical feature that corresponds to the positional rank of the product on the display for given `impression_id`. Usually, the position on the display creates the bias with respect to the click: lower rank means higher position of the product on the display.
- `displayrandom` is a binary feature that equals 1 if the display position on the banner of the products associated with the same `impression_id` was randomized. The click-rank metric should be computed on `displayrandom = 1` to avoid positional bias.
- `click` is a binary feature that equals 1 if the product `product_id` in the impression `impression_id` was clicked by the user `user_id`.

Figure 9 illustrates distributions of some features: number of impressions per user, number of products per user and banner size have long tail phenomenon (plots of the upper row). The products have popularity bias (lower plot), i.e. some products have much higher or lower than average number of clicks with senior job ads having more clicks on average, and position bias, i.e. increased number of clicks per lower rank and almost no clicks in the highest ranks (right plot on the lower row).

Figure 8 represents the feature importance according to an importance gain of XGBoost trained in two different ways: (i) *unaware* without protected attribute as a feature and (ii) *unfair* way with protected attribute as a feature. Their performance is reported in Table 1. We notice that the feature `rank` has a high importance, however, it gets replaced by the `protected_attribute` feature in the

unfair regime. There most impactful features for both models are num23, num33, num43, num18 and num19, apart from rank for the unaware model and protected_attribute for the unfair one.

In Figures 10 and 11 we plot correlation matrix between numerical features in the original dataset and open-sourced dataset FairJob, where we added noise for anonymization. We observe that correlations are generally slightly weaker for most features, when observing the anonymized data, that can be better seen in Figure 12, where we plot difference between the correlations. This fact comes naturally from the added noise, however, lower correlation values might translate into higher classification difficulty.

Table 2: Categorical features cardinalities.

feature	cat0	cat1	cat2	cat3	cat4	cat5	cat6	cat7	cat8	cat9	cat10	cat11	cat12
cardinality	9	9	1025	98	122	1296	2492	3183	3541	2879	2314	1436	912

Table 3: Statistics for index features.

feature	user_id	impression_id	product_id
cardinality	30361	224898	57355

Table 4: Statistics for binary features (out of 1072226 rows).

feature	protected_attribute	senior	displayrandom	click
positive	536113 (50%)	713659 (66.6%)	105869 (9.9%)	7489 (0.7%)

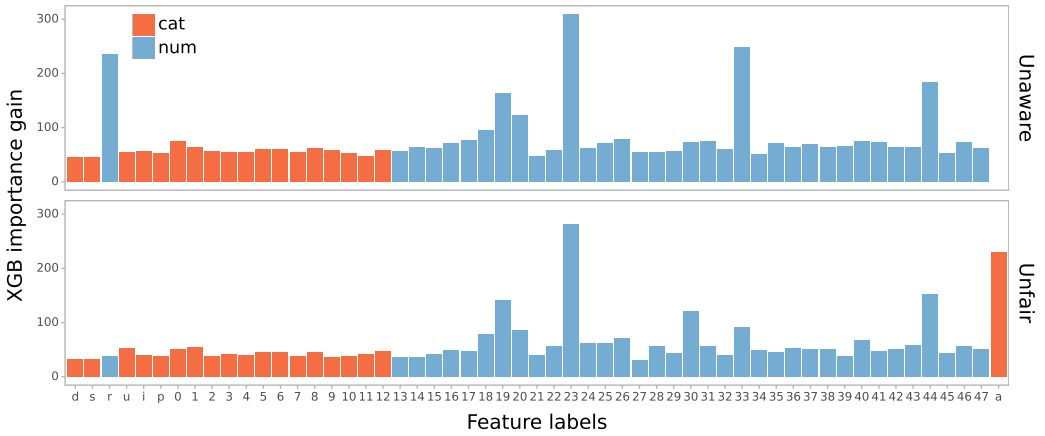


Figure 8: XGBoost importance gain per feature when trained in *unaware* way (without protected attribute as a feature – upper plot) and *unfair* way (with protected attribute as a feature – lower plot). Label d corresponds to displayrandom, s to senior, r to rank, u to user_id, i to impression_id, p to product_id, labels 0 to 47 to categorical (cat0, . . . , cat12) and numerical (num13, . . . , num47) features, respectively, and only on the lower plot a to protected_attribute.

A.4 Dataset statistics

Table 2, 3 and 4 demonstrate the available statistics for categorical features cardinalities, index features and binary features respectively. Additionally, we provide statistics for selection bias of the job campaign that comes from the Criteo AdTech company (outside of FairJob dataset) in Table 5, which we take into account further to compute utility metrics.

We provide the detailed statistics on clicks, job seniority and the protected attribute in the dataset in Table 6. From this table we can compute probabilities of events. For example, Table 7 shows that

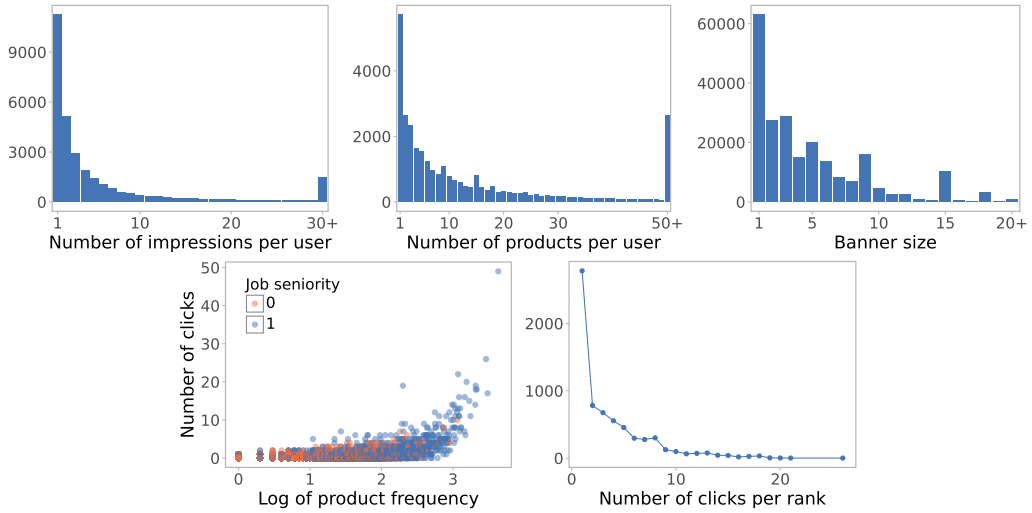


Figure 9: Examples of some feature statistics in FairJob dataset: number of impressions per user, number of products per user and banner size have long tail phenomenon (plots of the upper row). The products have popularity bias (left plot on the lower row), i.e. some products have much higher or lower than average number of clicks with senior job ads having more clicks on average. We observe a position bias due to increased number of clicks per rank and almost no clicks in the highest ranks (right plot on the lower row).

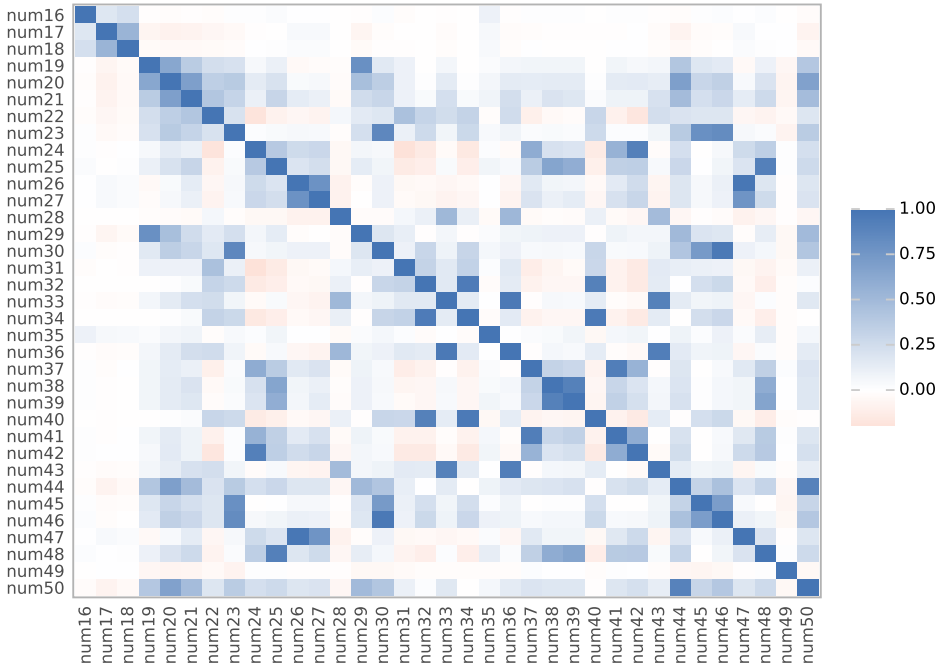


Figure 10: Correlations between numerical features in the **FairJob** dataset.

if a job ad is about a senior position, there is slightly higher chance that it will be shown to a male user than female (by 3.36%), while an assistant job ad is more likely to be shown to a female user than male (by 6.68%). In contrast, as can be seen in Table 8, there is almost no difference in clicks for senior ads given the protected attribute (both 0.49%) and a job add is slightly more likely to be clicked if shown to a female user (by 0.06%).

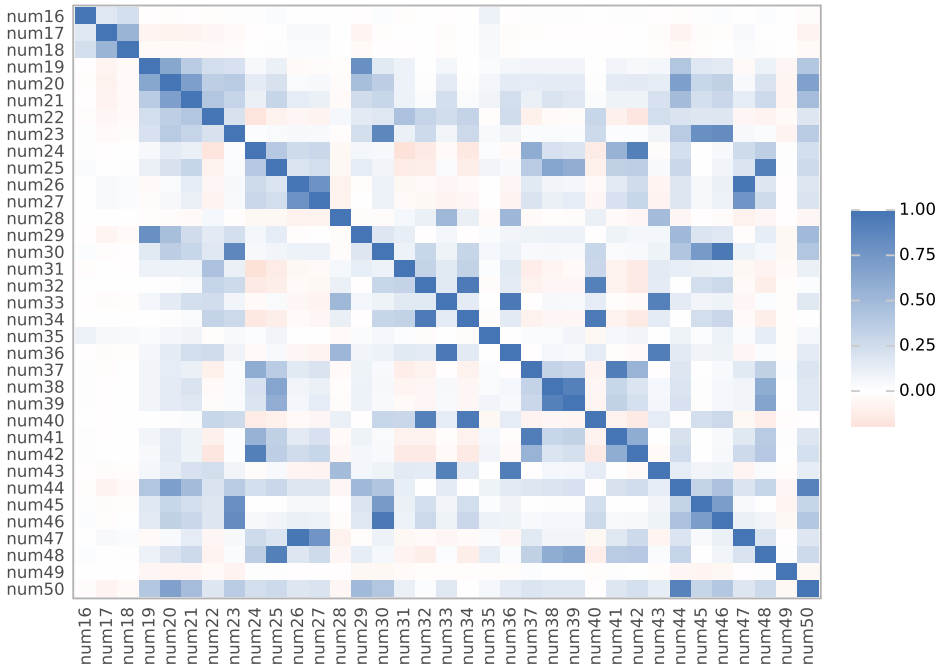


Figure 11: Correlations between numerical features in the **original** dataset.

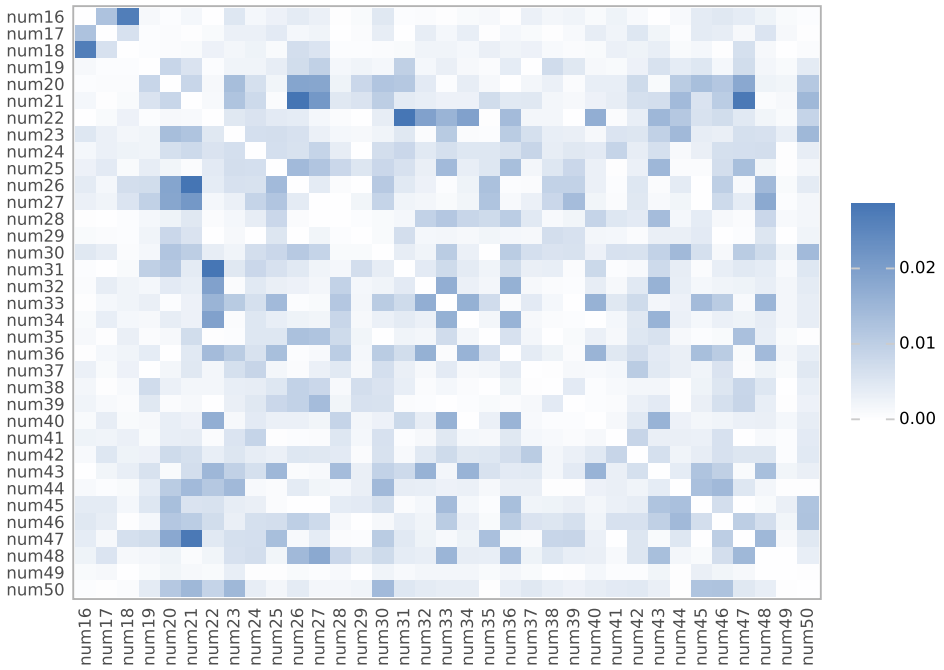


Figure 12: **Difference** in correlations between numerical features in the **original** and **FairJob** dataset.

B Experiments

B.1 Metrics

Fairness metrics. To measure fairness of the model, we consider *demographic parity* on senior job opportunities that computes the average difference of predictions given the protected attribute and

Table 5: Statistics for job campaign selection bias: we see that while female (internet)-population is slightly larger than male, the female population in the job campaign is much smaller, leading to selection bias of the job campaign – male users are almost twice more picked than female users.

population	advertising	job campaign	ratio
female	53.6%	39.2%	0.73
male	46.4%	60.7%	1.31

Table 6: Observed frequencies for clicks, senior job ads, and protected attribute

	not clicked		clicked		
	non-senior	senior	non-senior	senior	all
female	189982	342221	1274	2636	536113
male	166394	366140	917	2662	536113
all	356376	708361	2191	5298	1072226

Table 7: Observed frenquencies for user's protected attribute given the seniority of the shown job ad.

	non-senior	senior
female	0.533390	0.483224
male	0.466610	0.516776

Table 8: Observed joint frenquencies for clicks, senior job ads, conditional on the protected attribute.

	not clicked		clicked	
	non-senior	senior	non-senior	senior
female	0.354369	0.638337	0.002376	0.004917
male	0.310371	0.682953	0.001710	0.004965

the job seniority:

$$\mathrm{DP}(\hat{Y}|A) = \mathbb{E}\left(\hat{Y}|A=1,S=1\right) - \mathbb{E}\left(\hat{Y}|A=0,S=1\right). \tag{6}$$

Demographic parity requires equal proportion of positive predictions in each group ("No Disparate Impact") which in our case can be translated as same proportion of shown senior job ads in each group of the protected attribute. Demographic parity can be thought of as a stronger version of the US Equal Employment Opportunity Commission's "four-fifths rule", which requires that the "selection rate for any race, sex, or ethnic group [must be at least] four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate"[14].

Equalized odds [Zafar et al., 2017] ensures that a machine learning model works equally well for different groups. In case of job ads, equalized odds would force the prediction of both positive and negative outcomes to be the same which might generate more false positive predictions for one group versus others, resulting in worse outcome. *Equal of opportunity* [Hardt et al., 2016] ensures that a machine learning model works equally well only on positive outcomes for different groups which results in capturing the costs of misclassification disparities.

Performance metrics. The loss function is log-loss and report it as `NLLH` (negative log-likelihood). We also report `AUC` (Area under the ROC Curve) as a description of prediction power on strongly imbalanced data in binary classification problems. Additionally, we consider click rank utility U and product-rank utility for biased data $\tilde{U}$, proposed in the main text Section 3.4.

[14] See the Uniform Guidelines on Employment Selection Procedures, 29 C.F.R. §1607.4(D) (2015).

B.2 Models.

Training regimes. We train models in the following ways:

- **unfair** – the model uses protected attribute as a feature in the data,
- **unaware** – the model does not use the protected attribute in the data, corresponds to fairness through unawareness,
- **fair** – the model does not use the protected attribute in the data and trained with an additional fairness-enforcing penalty in the loss.

Algorithms. As baseline results, we perform three methods in different regimes:

- `Dummy` – classifier based on a single threshold for positive class probability (unaware),
- `XGB` – XGBoost algorithm (unfair and unaware),
- `LR` – Logistic Regression (unfair, unaware, and fair).

B.3 Possible improvements.

Base methods. We considered XGBoost and Logistic Regression models, but in some cases deep learning models such as Multi-Layer Perceptron, ResNets or Transformers might perform better, especially when focusing on improving mixed embeddings of categorical and numerical features [Gorishniy et al., 2022]. We also stress the challenge of strongly imbalanced classification, which can be further investigated.

Fairness methods. Another way to enforce fairness during training is to use an adversarial algorithm, e.g. Lahoti et al. [2020]. Alternatively, the pre-processing corrections are related to the direct modifications of the training set before training. The examples of the methods include separation of observatives X into two subsets $X_{\text{desc}(A)}$ and $X_{\text{non-desc}(A)}$ [Zemel et al., 2013]; transformation of X to some $\tilde{X}$ so that its factual and counterfactual distributions becomes the same, i.e. $P_{\tilde{X}}|do(A = a) = P_{\tilde{X}}|do(A = a')$ for all a, a'; learning "disentangled" representations [Locatello et al., 2019]. The pre-processing methods might be hard to apply to large and complex data as FairJob and might lead to losing a lot of data and, therefore, performance. However, the pre-processing and in-processing techniques might be combined together to achieve better results. We refer to Hort et al. [2023] for a recent survey on fairness inducing methods.

C Reproducibility

Source code for the experiments is hosted at https://github.com/criteo-research/fairjob-dataset/.

Resources. Experiments were conducted on Criteo internal cluster on instances with a RAM of 500Go and 46 CPUs available and 6 GPUs V100.

Tuning. We tune each model's hyperparameters following the procedure and the hyperparameters search spaces from Gorishniy et al. [2022]. Namely, the best hyperparameters are the ones that perform best on the validation set, so the test set is never used for tuning. For most algorithms, we use the Optuna library [Akiba et al., 2019] to run Bayesian optimization (the Tree-Structured Parzen Estimator algorithm) [Turner et al., 2021]. We set the number of tuning trails to 50, we use pruning and upper bound the number of data examples for hyperparameter search to 50000 as a compromise between optimality and training time constraint, however, we train the models with the optimal parameters on the whole training set. The amount of data and number of trails for tuning can be increased for the best performance.

C.1 Logistic regression details

Feature embeddings. We use a PyTorch [Paszke et al., 2019] module `Embedding` to compute embeddings for categorical features as a part of model training. Then, we created a mixed embedding layer where we concatenate categorical features embeddings with numerical features. The resulted

mixed embedding layer is used as a model input. The embeddings for both categorical and numerical features can be improved, for example, by following the benchmark paper of Gorishniy et al. [2022].

Fixed hyperparameters. We tested different batch sizes from [1024, 4000, 10000] and fixed batch=1024 as the best performing. Due to strong class imbalance, we oversample positive class examples for batch generation with sampling weights inversely proportional to the observed class frequencies in the data, using PyTorch [Paszke et al., 2019] utility `WeightedRandomSampler`. We fix number of epochs to `n_epochs` = 50.

Tuned hyperparameters. We tune the following hyperparameters:

- `embedding_size = UniformInt[4, 8]`,
- `weight_decay = LogUniform[1e-6, 1e-4]`,
- `scheduler_step_size = UniformInt[20, n_epochs]`,
- `scheduler_gamma = LogUniform[1e-2,1]`

Fairness parameters. We report the results for

- `fairness_multiplier` $\lambda \in [0.0, 0.1, 0.316, 1.0, 3.0, 5.0]$.

For each configuration we performed hyperparameter tuning described above.

Evaluation. For each tuned configuration, we run 10 experiments with different random seeds and report the average performance on the test set.

Experiment for density plots. The density plots illustrated in the main paper (Figure 5) correspond to the positive predictions densities for logistic regression model trained in *unfair*, *unaware* and *fair* ways. For easy and fast reproducibility, we upper-bound the number of training examples for hyperparameters tuning to 100000, `batch_size = 10000` and `number_of_epochs = 15`. The results are outputs of one simulation of logistic regression models with the following parameters and tuned on the following hyperspaces:

- `embedding_size = UniformInt[4, 5]`,
- `learning_rate = LogUniform[1e-4, 1e-2]`,
- `weight_decay = LogUniform[1e-6, 1e-4]`

For the model with fairness penalty, we use penalty coefficient equal to 3.0. We want to stress that the aim of this simulation was to provide a descriptive example rather than the optimal model for click classification. In the repository, we report the best parameters and predictions, as well as the code for the plot generation.

C.2 XGBoost details

Feature embeddings. We use the Python package for the XGBoost library [Chen and Guestrin, 2016], through the Scikit-Learn API [Pedregosa et al., 2011]. We encode the categorical features using the `TargetEncoder` in Scikit-Learn, learning the encoding on the training set [Micci-Barreca, 2001].

Fixed hyperparameters. We use the histogram (`hist`) tree contruction algorithm in order to speed up the model fit. In order to deal with the imbalance of the classes, we set the `scale_pos_weight` parameter to the ratio between negative and positive class occurrences in the training data, as suggested in the library documentation. We fix number of trials for tuning to `n_trials` = 100.

Tuned hyperparameters. We tune the following hyperparameters:

- `max_depth = UniformInt[3,10]`,
- `min_child_weight = LogUniform[0.0001,100]`,

- `subsample = Uniform[0.5,1]`,
- `learning_rate = LogUniform[0.001,1]`,
- `colsample_bytree = Uniform[0.5,1]`,
- `reg_lambda = LogUniform[0.1,10]`,
- `gamma = LogUniform[0.001,100]`.

D Comparison to other tabular datasets

D.1 Click-prediction datasets

We argue that the FairJob dataset is representative enough to evaluate click prediction algorithms, based on three main points:

- **real-world data** – the dataset is based on real-world data and contains all the necessary information. We have provided detailed documentation to support this.
- **comparative size** – the dataset is comparable in size to other widely-used specialized recommendation datasets.
- **baseline experiments** – we performed baseline experiments that produced reasonable results with respect to both performance and the utility-fairness trade-off.

The dataset contains pseudonymized users' context and publisher features collected from a job-targeting campaign run over five months. Therefore, the dataset contains *slices of real-world data*.

To properly compare the FairJob dataset to other click prediction datasets, we searched for the most used datasets for click prediction. We identified the following widely-used datasets for click prediction based on benchmark papers [Zhu et al., 2021, Mao et al., 2023]:

- Criteo 2014 Dataset. It consists of ad click data over a week and comprises 26 categorical feature fields and 13 numerical feature fields.
- Avazu 2015 Dataset. It contains 10 days of click logs and has a total of 23 fields with categorical features including app id, app category, device id, etc.

Both datasets are sampled from real click logs in production and contain tens of millions of samples. However, both lack explicit user and item field information. In addition to click prediction datasets, other recommendation system problems of similar complexity, such as MovieLens for predicting engagements and Frappe related to app usage, have datasets comparable in size to FairJob [Zhu et al., 2021, Mao et al., 2023]. See the detailed comparison in Table 9.

Table 9: Comparison of FairJob to other most frequently used tabular datasets for click prediction and recommendation system problems of similar complexity.

dataset	# rows	# features
Criteo	45,840,617	39
Avazu	40,428,967	22
MovieLens	2,006,859	3
FairJob	1,072,226	60
Frappe	288,609	10

We acknowledge that the FairJob dataset does not contain all features typically used for training click prediction models in the industry, due to business confidentiality. However, despite hand-picking the most relevant features, our dataset still contains 60 features, which is the largest number of real features in a dataset for click prediction.

The FairJob dataset does focus on a subproblem of click prediction within the context of job advertising campaigns, where users are pre-selected as having potential job-seeking profiles. Consequently, the dataset size is represented by a smaller number of records specific to this campaign. The click rate of 0.007% is consistent with the click rates observed in larger-scale datasets.

Furthermore, 1 million records for an unbalanced classification problem seems a reasonable size compared to other tabular datasets, as also stated in the next subsection. Our experiments in Appendix B demonstrate that baseline classification algorithms can successfully predict clicks on the test set.

D.2 Fairness-aware datasets

The shortcomings of existing fairness-aware data sets include: age of the data itself, measurement bias, missing values, label leakage, use of data sets for a purpose they were not intended for initially [Le Quy et al., 2022, Hort et al., 2023]. FairJob dataset is collected in 2024, does not have missing values and represents the real-world application for Fair AI methods.

From benchmark studies on encoding numerical and categorical features [Gorishniy et al., 2021, 2022, Grinsztajn et al., 2022, Matteucci et al., 2023] and surveys on fairness-aware tabular datasets [Le Quy et al., 2022, Hort et al., 2023], we extract the most used dataset and compare to FairJob in Table 10.

Table 10: Comparison of FairJob to other most frequently used tabular datasets and most frequently used tabular fairness-aware datasets.

name	# rows	# num	# cat	task type	protected attribute
COMPAS	7,214	14	37	binclass	sex, race
Gesture Phase	9,873	32	0	multiclass	-
Churn Modelling	10,000	10	1	binclass	-
California Housing	20,640	8	0	regression	income
House 16H	22,784	16	0	regression	-
Adult	48,842	6	8	binclass	sex, race, gender
Otto Group Products	61,878	93	0	multiclass	-
Higgs Small	98,049	28	0	binclass	-
Diabetes	101,766	10	40	binclass	gender
Facebook Comments Volume	197,080	50	1	regression	-
Santander Customer Transactions	200,000	200	0	binclass	-
KDD Cencus-Income	299,285	34	7	binclass	sex, race
Covertype	581,012	54	0	multiclass	-
FairJob	1,072,226	36	18	binclass	gender
MSLR-WEB10K (Fold 1)	1,200,192	136	0	regression	-

E Broader impact

Any machine learning system that learns from data runs the risk of introducing unfairness in decision making. Recent research [Speicher et al., 2018, Lambrecht and Tucker, 2019, Andreou et al., 2019, Ali et al., 2019] has identified fairness concerns in several AI systems, especially toward protected groups that are under-represented in the data. Thus, alongside the technical advancements in improving AI systems it crucial that we also focus on ensuring that they work for everyone.

Click prediction is a fundamental task in online advertising and recommendation systems, as noted in recent benchmarking papers [Zhu et al., 2021, Mao et al., 2023]. This is an active research field where improvements in click prediction algorithms lead to improvements in recommendation systems in general [Mao et al., 2023]. However, improvements in general models do not always translate to specialized models, thus, there is a need to verify algorithms directly with available resources. The FairJob dataset represents a particular case of the click prediction problem, focusing on job-seeking user profiles and job advertising product features. However, many features (user features and categorical product features) are not campaign-specific and can be found in other applications outside job advertising. By open-sourcing the FairJob dataset, we provide access to a real-world problem of finding a trade-off between utility and fairness in job offer advertising. We argue that advances in fair machine learning methods should be validated in real-world scenarios, as the distribution of features and their impact on outputs and fairness can vary significantly across different applications.

As described in Section 3, advertising, like almost any complex problem, is subject to various biases [Hort et al., 2023]. We have clearly stated the potential bias sources, analyzed their impact, and discussed possible corrections. While we cannot definitively prove that the properties of the FairJob dataset can be extended to every click prediction problem, it serves as a representation of a specific real-world scenario where the fairness-utility trade-off is encountered. If the FairJob dataset helps to find better methods for improving fairness in job advertising, these methods might also be applicable to other advertising campaigns, as similar user signals may be present in the features used to predict clicks or to associate with the utility-fairness trade-off.

One of the key practical challenges in addressing unfairness in AI systems is that most methods require access to protected demographic features, placing fairness and privacy in tension. We work toward addressing these important challenges by proposing a new benchmarking dataset to improve worst-case performance of protected groups, in the absence of protected group information but with a proxy variable.

One limitation of methods in this space is the difficulty of evaluating their effectiveness when we do not have demographics in a real application. Therefore, while we think developing better debiasing methods is crucial, there remains further challenges in evaluating them.