

Evaluating Model Performance and Fairness without Labeled Representative Data

EVELIEN VAN DER LAAN, Utrecht University, The Netherlands and Netherlands Police, The Netherlands
NIKOLAJ TOLLENAAR, Utrecht University, The Netherlands and Research and Documentation Centre (WODC) Ministry of Justice and Security, The Netherlands
ANA ISABEL BARROS, Netherlands Police Academy, The Netherlands, Jheronimus Academy of Data Science, Tilburg University, The Netherlands, and TNO, The Netherlands
LUKAS SNOEK, Netherlands Police, The Netherlands

Evaluating machine learning models using data that reflect the deployment context is essential to draw reliable conclusions about both predictive performance and fairness. When models are evaluated on data drawn from a different population, time period, or domain, performance and fairness assessments can be misleading. In practice, however, evaluation data that is representative of the deployment setting is often unavailable, expensive to collect, or lacks the required annotations.

In this paper, we study how evaluation can be performed when only labeled data from a different, unrepresentative distribution is available. We propose an importance sampling approach that uses an unlabeled sample from the deployment context to reweight a labeled evaluation dataset, so that it better resembles the deployment population. A central challenge is deciding which features should be used to align the evaluation and deployment data. We show that using model-internal features—hidden-layer representations of the evaluated model itself—to reweight the evaluation data outperforms handcrafted features and general embeddings from other models when estimating deployment-time performance. Importantly, this approach does not require prior knowledge of how the evaluation and deployment data differ.

We further show that different evaluation and fairness metrics respond differently to changes in the outcome distribution between datasets. Metrics conditioned on the model's predictions (such as precision) require explicit alignment of outcome prevalence between the evaluation and deployment data. By contrast, metrics conditioned on the true outcomes (such as recall) are robust to shifts in the outcome distribution. Finally, we find that fairness evaluation is more challenging than performance evaluation: in most settings, subgroup disparities in the deployment context cannot be reliably inferred without access to sensitive attribute information.

Overall, this work clarifies how evaluating models on non-representative data can distort both performance and fairness assessments, and provides a practical method for approximating deployment-time evaluation when representative labeled data are unavailable.

CCS Concepts: • **General and reference** → **Evaluation**; • **Computing methodologies** → **Machine learning**.

Additional Key Words and Phrases: Model Evaluation, Algorithmic Fairness, Distribution Mismatch, Importance Sampling, Dataset Shift, Fairness Auditing, Representation Learning, Unlabeled Target Data

Authors' Contact Information: Evelien van der Laan, Utrecht University, Utrecht, The Netherlands, e.a.vanderlaan@uu.nl and Netherlands Police, Driebergen, The Netherlands; Nikolaj Tollenaar, Utrecht University, Utrecht, The Netherlands, n.tollenaar@uu.nl and Research and Documentation Centre (WODC) Ministry of Justice and Security, The Hague, The Netherlands; Ana Isabel Barros, Netherlands Police Academy, Appeldoorn, The Netherlands, ana.barros@politieacademy.nl and Jheronimus Academy of Data Science, Tilburg University, 's-Hertogenbosch, The Netherlands and TNO, The Hague, The Netherlands; Lukas Snoek, Netherlands Police, Amsterdam, The Netherlands, lukas.snoek@politie.nl.



This work is licensed under a Creative Commons Attribution 4.0 International License.

FACCT '26, Montreal, QC, Canada

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2596-8/2026/06

<https://doi.org/10.1145/3805689.3812303>

ACM Reference Format:

Evelien van der Laan, Nikolaj Tollenaar, Ana Isabel Barros, and Lukas Snoek. 2026. Evaluating Model Performance and Fairness without Labeled Representative Data. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26)*, June 25–28, 2026, Montreal, QC, Canada. ACM, New York, NY, USA, 24 pages. <https://doi.org/10.1145/3805689.3812303>

1 Introduction

In recent years, the rapid rise of large, pre-trained models has brought significant advances across a wide range of applications. Yet, it is well established that the performance of artificial intelligence (AI) models can differ substantially when deployed on data that differs from their training distribution—a phenomenon referred to as *distribution shift* [9, 38, 50]. Importantly, the consequences of distribution shift extend beyond predictive performance. Models that appear fair during development may exhibit algorithmic disparity (“bias”) when deployed in a different context, as fairness properties do not necessarily transfer across distributions [34, 40, 54]. As a result, evaluations based on data that do not reflect the deployment setting may lead to incorrect conclusions about both predictive performance and fairness, highlighting the importance of evaluating models in their intended deployment context.

In practice, however, conducting such evaluations is often challenging. A representative and fully annotated evaluation dataset that reflects the target population may be unavailable, expensive, or time-consuming to collect. As a result, model evaluations frequently rely on open-source benchmarks, data drawn from different populations, or historical datasets that may be affected by temporal drift [10, 61]. When development or evaluation is outsourced or based on pre-trained models released with performance or fairness claims computed on proprietary or undocumented datasets, the underlying evaluation distribution may be unknown. In these settings, reported metrics can be difficult to interpret and may fail to reflect a model’s behavior in real world deployment.

These concerns are increasingly reflected in the regulatory landscape. For example, the EU Artificial Intelligence Act (AI Act) requires training, validation, and test data to be “relevant, sufficiently representative and to the best extent possible, free of errors and complete in view of the intended purpose” (Article 10(3)) [19]. In practice, however, these requirements are often difficult to satisfy, motivating the need for evaluation methods that can provide reliable performance and fairness estimates even when representative labeled data from the target population are limited or unavailable.

Although reliable evaluation without representative labeled data is difficult to achieve across all scenarios, we show that deployment-time performance and fairness can be approximated using labeled data from a different distribution. To do so, we propose to use importance sampling based on the model’s internal features. Importantly, we outline the assumptions needed for this approach and highlight the limitations of the approach.

1.1 Relevant Literature

Distribution Shift and Evaluation. The distribution of datasets may differ in multiple ways. Formally, distribution shift occurs when the joint distributions of two datasets differ, i.e., $P_1(Y, X) \neq P_2(Y, X)$, where $P_n(Y, X)$ denotes the joint distribution of labels Y and covariates X in dataset n [25, 43]. The dataset may also include sensitive attributes A .

Moreno-Torres et al. [43] distinguish three common forms: covariate shift (changes in $P(X)$), label shift (changes in $P(Y)$), and concept shift (changes in $P(Y|X)$). More recent work extends this taxonomy to fairness-related changes. Demographic shift refers to changes in the distribution of sensitive groups (changes in $P(A)$) [23], while dependence shift captures changes in the relationship between sensitive attributes and outcomes (changes in $P(Y | A)$) [52]. Such shifts can directly affect fairness when models are deployed in new environments [57]. These distribution shifts can formally be specified as follows:

- Covariate shift: $P_1(X) \neq P_2(X)$, but $P_1(Y|X) = P_2(Y|X)$

- Label shift: $P_1(Y) \neq P_2(Y)$, but $P_1(Y|X) = P_2(Y|X)$
- Concept shift: $P_1(Y|X) \neq P_2(Y|X)$, but $P_1(X) = P_2(X)$
- Demographic shift: $P_1(A) \neq P_2(A)$
- Dependence shift: $P_1(Y | A) \neq P_2(Y | A)$

Distinguishing these shifts is important because they affect evaluation differently. When evaluation data do not reflect the deployment distribution, standard performance metric values may become deceptive [21].

Fairness evaluation introduces an additional challenge: disparity metrics compare group-conditioned quantities. Consequently, they are sensitive not only to global distribution changes (changes in $P(Y)$ and $P(X)$) but also to shifts in subgroup outcome distributions (changes in $P(A)$ and $P(Y | A)$). While prior work has mainly studied predictive performance under distribution shift, less attention has been given to how disparity metrics behave. Our paper contributes by systematically identifying which performance and fairness metrics remain reliable under different types of distribution shift.

A related line of work studies out-of-distribution (OOD) detection, which identifies when incoming data deviate from the training distribution and signals when predictions may become unreliable during inference [29, 70]. In contrast, our approach evaluates how a model is expected to perform on new data before deployment.

Fairness under Distribution Shift. If distribution shifts occur between training and deployment data fairness properties of a model do not necessarily transfer across distributions [34, 40, 54]. Recent work of Deho et al. [15] shows that algorithmic fairness properties can degrade under naturally occurring dataset shifts even when models were trained or post-processed to satisfy fairness criteria on the original distribution. As noted by [15], one implication in practice is the need for continuous monitoring and evaluation of the fairness of algorithms, as “the tides of fairness can turn really fast” (p. 15). We complement this work by investigating how performance and fairness can be evaluated and monitored with little or no labeled data from the deployment context.

Several studies propose training methods designed to maintain fairness under distribution shift [3, 31, 55, 65]. In line with our problem setup, [27, 51] modify the learning objective to enforce fairness on the target distribution despite the absence of labeled target data. Our work, however, focuses on the evaluation problem, given an existing (pre-trained) model: how to reliably measure model performance and fairness in the deployment context when the available (labeled) evaluation datasets do not match the deployment distribution.

Importance Sampling. If evaluation data are not representative of the deployment distribution, methods are needed to correct for this mismatch. One such approach is importance sampling, a Monte Carlo technique that estimates properties of a target distribution using samples drawn from a different source distribution [37]. In machine learning, importance weighting is widely used to address distribution shift, particularly under the covariate shift assumption where $P(Y | X)$ remains stable but the marginal feature distribution changes [58]. In this setting, observations from the source distribution are reweighted by the density ratio $p_{\text{target}}(x)/p_{\text{source}}(x)$, adjusting the influence of each observation so that the weighted training data better reflects the target distribution [62]. These ideas are commonly applied during model training in, for example, domain adaptation and distribution shift correction, where importance weighting is used to learn models that generalize to a target distribution; see Kimura and Hino [36] for an overview.

In this paper, however, we apply importance sampling not for model training but for *model evaluation*. Our goal is to estimate performance and fairness metrics that reflect the deployment distribution without modifying or retraining the model. Following Chen et al. [10], we combine a labeled evaluation dataset drawn from a *source distribution* with an unlabeled dataset drawn from the *target distribution*, consisting of inputs observed in the deployment environment. By estimating the density ratio between these two distributions, source samples can be weighted or sampled so that evaluation better reflects the target distribution.

A key challenge in applying importance weighting is determining which features capture the relevant differences between the source and target distributions. If important drivers of the shift are omitted, the resulting weights may fail to recover the correct target statistics [10]. Handcrafted features may overlook latent or higher-order shifts in the data and typically require additional feature design or annotation. To address this challenge, we investigate whether model-internal representations, that is, activations in hidden layers of the (pre-trained) model itself, can serve as an effective basis for importance weighting.

A large body of work develops practical estimators and model-selection procedures for importance weighting and density ratio estimation, and analyzes failure modes such as high-variance weights under limited overlap [10, 13, 24, 63, 64]. While our work focuses on the interaction between distribution shift, evaluation metrics, and feature-space selection for importance weighting, these refinements can be incorporated directly into our framework without altering its core objective.

1.2 Contributions

This paper studies how to reliably evaluate predictive performance and fairness when evaluation data do not match the deployment distribution. Our contributions are threefold.

First, we analyze how commonly used performance and fairness metrics behave under different forms of distribution shift. We show that evaluation metrics differ in their sensitivity to shifts in the underlying data distribution. Particularly, metrics conditioned on the true outcome (e.g., recall, equalized odds) remain stable under label shift, whereas metrics conditioned on predicted outcomes (e.g., precision, predictive parity) depend on the outcome distribution and therefore require evaluation data to reflect the outcome distribution of the target population. We extend this analysis to fairness metrics, highlighting how subgroup outcome distributions affect disparity estimates.

Second, we focus on the variable selection problem in importance sampling, i.e., the challenge of choosing which features capture the relevant differences between the source and target distributions, and should therefore be used for constructing importance weights. In particular, we investigate whether model-internal representations (activations of hidden layers of the evaluated model) can serve as a feature space for density ratio estimation, avoiding manual feature engineering. Our results show that reweighting based on these representations can recover target performance under distribution shift, while the ability to recover subgroup statistics for fairness evaluation depends on the extent to which sensitive attributes are represented in these features.

Third, we examine task-agnostic embeddings as an alternative feature space when internal model representations are unavailable. Through empirical analysis, we show that such embeddings generally fail to recover target performance, suggesting that effective importance weighting requires features closely aligned with the prediction task.

To illustrate and evaluate the breadth of our proposed approach, we developed three experimental settings: a theoretical setting based on synthetic data, a computer vision setting involving face datasets, and a natural language processing setting using textual data. The experiments show the versatility of our approach: it integrates with existing density ratio estimation methods, supports both classification and continuous outputs, and extends to any deep learning based architecture that involves hidden layers (including most modern computer vision and natural language processing models). It is also practical, as it eliminates the need for handcrafted features or, under specific assumptions, detailed prior knowledge about the nature of the distribution shift.

2 Preliminaries

We consider a fixed model $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ denote the domains of features and labels, respectively. In addition, the dataset may include sensitive attributes $A \in \mathcal{A}$, which may or may not be included in the model input $\mathcal{X}$, but are often required for fairness evaluation. We use uppercase letters (e.g., X, Y, A) to denote random

variables and lowercase letters (e.g., x, y, a) for their realizations. We write P for probability distributions and p for their corresponding densities when they exist.

We are given two datasets drawn from different distributions: a labeled source dataset with n_s samples, $\mathcal{D}_s = \{(y_i^s, x_i^s, a_i^s)\}_{i=1}^{n_s} \sim P_s(Y, X, A)$, and an unlabeled target dataset with n_t samples, $\mathcal{D}_t = \{x_i^t\}_{i=1}^{n_t} \sim P_t(X, A)$. In some cases, sensitive attributes A may also be partially or fully available in $\mathcal{D}_t$.

Throughout this paper, we focus on the effects of *covariate shift* and *label shift*, which may arise independently or simultaneously.

Assumptions. (i) Support overlap. We assume that the target support is contained in the source support, i.e. $\text{supp}(P_t) \subseteq \text{supp}(P_s)$, where $\text{supp}(P) = \{x \in \mathcal{X} : p(x) > 0\}$. This ensures that every point that can occur under the target distribution also occurs with non-zero probability under the source distribution. (ii) Conditional invariance. We assume that the conditional distribution of labels given features and sensitive attributes remains unchanged: $P_t(Y | X, A) = P_s(Y | X, A)$.

Our objective is to estimate the model’s predictive performance and fairness metrics in the target context, using the labeled source data $\mathcal{D}_s$ and the unlabeled target data $\mathcal{D}_t$.

2.1 Accuracy Metrics

2.1.1 Performance Metrics: Classifiers. We consider a binary classifier f_θ that maps covariates X to predictions $\hat{Y} \in \{0, 1\}$, with true labels $Y \in \{0, 1\}$. Performance metrics for such models differ in whether they are conditioned on the true outcome or on the model’s predicted outcome [42].

Metrics conditioned on the true label Y quantify how well the model captures the actual classes. A common example is recall, defined as $P(\hat{Y} = 1 | Y = 1)$, which measures the fraction of positive cases that are correctly identified. Such metrics are independent of the prevalence of the positive class and therefore remain comparable across populations with different outcome distributions.

In contrast, metrics conditioned on the prediction $\hat{Y}$ reflect the reliability of the model’s outputs. Precision, for example, is defined as $P(Y = 1 | \hat{Y} = 1)$, and depends on both the model’s error rates and the outcome prevalence.

This distinction can be formalized using Bayes’ theorem. Metrics conditioned on the true label can be written as

$$P(\hat{Y} | Y) = \frac{P(Y | \hat{Y})P(\hat{Y})}{P(Y)} = \frac{P(Y | \hat{Y})P(\hat{Y})}{P(Y | \hat{Y} = 1)P(\hat{Y} = 1) + P(Y | \hat{Y} = 0)P(\hat{Y} = 0)} \quad (1)$$

and are invariant to changes in $P(Y)$ [49, 56]. In contrast, prediction-conditioned metrics satisfy

$$P(Y | \hat{Y}) = \frac{P(\hat{Y} | Y)P(Y)}{P(\hat{Y})} = \frac{P(\hat{Y} | Y)P(Y)}{P(\hat{Y} | Y = 1)P(Y = 1) + P(\hat{Y} | Y = 0)P(Y = 0)}. \quad (2)$$

and therefore vary with the outcome distribution $P(Y)$.

As a result, label-conditioned and prediction-conditioned metrics respond differently when evaluation and deployment data differ in outcome prevalence. This distinction is central to understanding which performance and fairness metrics remain valid under reweighting.

2.1.2 Performance Metrics: Regression. For regression tasks, models predict a continuous outcome $Y \in \mathbb{R}$, and performance is assessed by measuring the discrepancy between predictions $\hat{Y} = f_\theta(X)$ and true outcomes. Common error metrics include the mean absolute error (MAE) and the mean squared error (MSE) [7].

We focus primarily on the mean absolute error, defined as $\text{MAE} = \mathbb{E}[|Y - \hat{Y}|]$, which measures the average magnitude of prediction errors. Compared to mean squared error, defined as $\text{MSE} = \mathbb{E}[(Y - \hat{Y})^2]$, MAE is less sensitive to outliers and easier to interpret [68].

2.2 Fairness Metrics

We assess model fairness by comparing performance metrics across subpopulations defined by sensitive attributes A . As with performance metrics, fairness notions differ in what they condition on, which determines whether observed disparities are sensitive to differences in outcome prevalence across groups.

Equalized odds is a label-conditioned fairness criterion. It requires that error rates be equal across groups for individuals with the same true outcome, i.e.,

$$P(\hat{Y} = 1 \mid Y = y, A = a) = P(\hat{Y} = 1 \mid Y = y, A = a'), \quad y \in \{0, 1\},$$

and is therefore invariant to differences in outcome prevalence $P(Y \mid A)$ [42]. This notion is also referred to as *separation*.

In contrast, *predictive parity* is conditioned on the model's prediction and requires equal precision across groups,

$$P(Y = 1 \mid \hat{Y} = 1, A = a) = P(Y = 1 \mid \hat{Y} = 1, A = a'),$$

also known as *sufficiency* [11]. Because it conditions on $\hat{Y}$, predictive parity is sensitive to differences in outcome prevalence between groups $P(Y \mid A)$, and observed disparities may reflect base-rate differences rather than differences in model behavior. Appendix D provides a Bayes' decomposition of predictive parity that makes this dependence explicit.

Finally, when outcomes are unavailable or contested, fairness is sometimes defined in terms of equal decision rates across groups, known as *statistical parity* or *independence* [42]: $P(\hat{Y} \mid A = a) = P(\hat{Y} \mid A = a')$.

For regression tasks, formal fairness criteria are less established [26]. We therefore assess fairness by comparing prediction errors across groups, using differences in mean absolute error (MAE) as a measure of disparity [5].

3 Methodology

Our proposed method to assess model performance under distribution shift estimates how well a fixed model would perform in a new context using labeled data from a different distribution. Since we focus on cases where the target dataset lacks labels, we apply reweighting techniques to approximate performance and fairness of the model in the target context. We use two related approaches: importance sampling and importance weighting. Both require estimates of the ratio of the probability density functions of the source and target distributions.

3.1 Importance Sampling: a Monte Carlo Method

Importance sampling is a Monte Carlo method for evaluating expectations under a target distribution using samples drawn from a different (source) distribution [64]. Suppose we are interested in estimating the expectation of a function $f(X)$ under a target distribution $p_t(x)$, but we only have access to samples from a source distribution $p_s(x)$. The expectation can be rewritten as:

$$\mathbb{E}_{p_t}[f(x)] = \int p_t(x)f(x) dx = \int \frac{p_s(x)}{p_s(x)} p_t(x)f(x) dx = \mathbb{E}_{p_s} \left[\frac{p_t(x)}{p_s(x)} f(x) \right] \approx \frac{1}{N} \sum_{i=1}^N \frac{p_t(x_i)}{p_s(x_i)} f(x_i) \quad x_i \sim p_s(x) \quad (3)$$

Each term is scaled by the importance weight $w_i = \frac{p_t(x_i)}{p_s(x_i)}$, which adjusts for the mismatch between the source and target distributions [64]. In importance sampling, these weights are used as sampling weights to draw a representative subsample from the source dataset: samples that better resemble the target distribution are more likely to be included. In contrast, importance weighting retains all source data points and reweights their contribution during metric computations.

These techniques allow us to estimate how the model would behave under the target distribution, despite not having access to the target outcomes.

3.1.1 Density Estimation. Density estimation is known to be difficult, especially in high-dimensional spaces, due to the curse of dimensionality, which slows the convergence of nonparametric estimators as dimensionality increases [45]. Because of this, density ratio estimation (i.e. estimating the ratio between two distributions directly) has become a widely studied alternative. Several methods have been proposed, such as moment matching, kernel mean matching, Kullback-Leibler Importance Estimation Procedure (KLIEP) and Least-Squares Importance Fitting (LSIF) (see Sugiyama et al. [62] for further details on density ratio estimation).

In this paper, we use a random forest classifier to distinguish between samples from the source dataset and the target dataset. Specifically, we construct a binary domain label $D \in \{0, 1\}$, where $D = 0$ indicates source samples and $D = 1$ indicates target samples. For each source data point, the classifier predicts the probability that it originates from the target distribution, which is then used to compute an importance weight. This provides a simple and effective approximation of the density ratio between the two distributions, while avoiding explicit density estimation. The approach has been shown to perform well in high-dimensional settings [6]. In our experiments, the classifier was implemented as a random forest with 100 trees, unrestricted tree depth, and balanced class weights. To obtain unbiased probability estimates and avoid overfitting, we used stratified 5-fold cross-validation with shuffled folds and computed out-of-fold predicted probabilities. All other parameters were left at their default values in the scikit-learn implementation (version 1.7.2) [48]. While we used this implementation in our experiments, the proposed framework is flexible and can accommodate alternative classifiers or density ratio estimation techniques.

3.2 Feature Space Importance Sampling

To estimate importance weights, we compare the distribution of observations from the source and target datasets in a shared feature space. Observations from the source dataset that are more similar to the target data receive higher weights, while those that are less similar receive lower weights. The quality of this feature space is therefore crucial: it must capture the relevant properties that differ between the two distributions.

We construct this space using two types of features: (i) *explicit, interpretable features*, and (ii) *model-internal features* derived from the last (non-output) layer of deep neural networks.

3.2.1 Explicit Feature Generation. Explicit features refer to interpretable characteristics of the data that are derived independently of the model whose performance we aim to evaluate. They may include metadata, manual annotations, or automatically generated attributes obtained through pre-processing pipelines. Such features capture domain-specific properties that help characterize source and target dataset differences.

In our computer vision experiments, we obtain explicit (facial) features using current state-of-the-art pre-trained models. These features include measures of image quality and sharpness [22], head pose [28], skin color estimates [20], and contrast-based appearance descriptors [69]. While noisy and imperfect, these explicit features provide structured, human-interpretable information.

3.2.2 Model-Internal Features. Neural networks are typically structured into three parts: an input layer, multiple hidden layers, and an output layer. As model input flows through the neural network, each hidden layer applies a nonlinear transformation that encodes the input data into progressively higher-level and more abstract feature representations. Early layers capture relatively simple patterns (such as edges in images or word-level context in text), while deeper layers encode more complex relations. The final hidden layer provides a lower dimensional representation of the input that contains the information considered most relevant by the model for making accurate predictions [2, 46].

We refer to these final hidden representations as model-internal features. They offer a number of advantages for our setting. First, they are general: every neural network architecture produces such internal activations/features, whether it is a feedforward network, a convolutional neural network (CNN), a recurrent neural network (RNN),

or a transformer. Second, they are domain-agnostic: the same principle applies regardless of whether the input is text, images, speech, or another modality. Finally, they are task-aligned: since they are derived directly from the trained model, they reflect the features that the model has learned to be salient for its predictive task.

The exact form of these features depends on the underlying architecture. In feedforward networks, they correspond to the activations of the final hidden layer. In CNNs, global pooling is typically applied to the final feature maps to obtain a fixed-length representation of the input. In transformer architectures, such as BERT or Vision Transformers, the final hidden state of the [CLS] token provides a compact embedding that summarizes the input sequence or image [2].

Unlike explicit features, which rely on predefined domain knowledge, model-internal features capture distributional differences as perceived by the model itself, providing a task-aligned basis for reweighting that directly reflects the model’s predictive behavior.

3.2.3 Dimension Reduction. Model-internal features are often high-dimensional. When the number of data points is limited, using such representations can lead to overfitting during density ratio estimation, and even with abundant data they can be computationally expensive [30]. Therefore, we apply dimensionality reduction as a preprocessing step.

Specifically, we use Principal Component Analysis (PCA) to project model-internal features into a lower-dimensional space while preserving most of their variance [1]. In practice, we retain the smallest number of components that jointly explain 90% of the total variance, a commonly used heuristic for balancing dimensionality reduction with information preservation [32]. We found that the reduced representation can serve as an effective substitute for the full feature set, balancing efficiency with preservation of task-relevant information.

3.2.4 Label Distribution. Under label shift, importance weighting must account not only for changes in the feature distribution but also for differences in outcome prevalence. In practice, however, the true target labels Y are typically unavailable at evaluation time. Instead, model-internal features encode information about the model’s predictions $\hat{Y}$, and therefore reflect the distribution $P(\hat{Y})$ rather than the true label distribution $P(Y)$. This subsection examines when such model-implied outcome information is sufficient for reweighting, and when explicit adjustment for the target outcome distribution is required.

A key case in which model-implied outcome information is sufficient, arises when the model is well-calibrated, i.e., when $\mathbb{E}[Y | \hat{Y} = s, A] = s$ [4, 14]. In the special case of binary outcomes, calibration not only guarantees this conditional relationship but also ensures that the marginal distributions $P(\hat{Y})$ and $P(Y)$ (as well as $P(\hat{Y} | A)$ and $P(Y | A)$) are closely aligned. In such settings, model-internal features can serve as a sufficient proxy for the outcome variable, avoiding the need for additional adjustment. However, this reliance on calibration is fragile in practice, as model calibration does not reliably transfer across domains [47].

For continuous outcomes, calibration ensures equality of conditional means but does not imply equality of the full distributions. As a result, $P(\hat{Y})$ and $P(Y)$ (and likewise $P(\hat{Y} | A)$ and $P(Y | A)$) may still diverge and relying solely on internal features is insufficient for accurately capturing outcome prevalence.

To address this more general case, outcome information must be explicitly incorporated during importance weight estimation. One approach is to include estimates of $P(Y)$ or $P(Y | A)$ alongside model features. Alternatively, importance weights can be rescaled to match the observed outcome distribution. Concretely, let $P_v(Y)$ denote the outcome distribution induced by the current weights v_i , and let $P_t(Y)$ be the empirical distribution of the observed labels in the target domain. For each outcome value y , define a rescaling factor $c_y = \frac{P_t(Y=y)}{P_v(Y=y)}$. Corrected weights are then obtained by setting $w_i = c_{y_i} v_i$, where c_{y_i} denotes the correction factor corresponding to the label value of observation i . By construction, the reweighted distribution satisfies $P_w(Y) = P_t(Y)$, ensuring that the weights reflect the observed outcomes rather than the model-implied ones. The same correction can be applied conditionally on sensitive attributes by defining $c_{a,y} = \frac{P_t(Y=y|A=a)}{P_v(Y=y|A=a)}$ and using $w_i = c_{a_i, y_i} v_i$.

Importantly, this procedure only requires knowledge of the marginal (or conditional) outcome distribution in the target domain, not full annotations. Such marginals can be obtained by labeling a small random subsample of target data or, in some settings, by relying on domain knowledge or prior assumptions about outcome prevalence. When the classifier is well calibrated, $P_t(\hat{Y})$ (and $P_t(\hat{Y} \mid A)$) provides a good proxy for the true outcome distribution, eliminating the need for additional estimation of $P_t(Y)$.

In summary, whether one considers $P(Y)$ for performance metrics or $P(Y \mid A)$ for fairness metrics, the same principle applies: in the binary calibrated case, model-internal features suffice as proxies for labels, while in all other cases explicit incorporation of outcome distributions remains necessary.

4 Results

We evaluate how importance sampling and importance weighting approximate target-domain performance and fairness across multiple settings. Our experiments span three domains: synthetic data (Appendix B), computer vision, and natural language processing. Table 2 in Appendix A provides an overview of all conducted experiments, including the tasks, datasets, models, feature spaces and feature dimensionalities. To facilitate reproducibility, code and data are publicly available.¹

4.0.1 Evaluation Methods. Across all experiments, we compare four evaluation strategies:

- **Source (naive):** Evaluation on the labeled source dataset.
- **Target (ground truth):** Evaluation on labeled target data (oracle; used only for reference).
- **Importance Sampling:** Evaluation on a weighted subsample of the source data.
- **Importance Weighting:** Evaluation on the full source data with importance-weighted metrics.

Unless stated otherwise, results are averaged over 1,000 bootstrap resamples, with confidence intervals computed from the same distribution. Note that the goal is to approximate the target-domain evaluation; thus, closer agreement with the target-domain estimate indicates better recovery via weighting or sampling.

4.1 Computer Vision Experiments

In this section, we examine how importance sampling and importance weighting can be used to evaluate model performance under distribution shift in computer vision, using age and gender prediction as case studies.

We study age and gender prediction using two widely used face datasets: UTKFace [71] and FairFace [35]. Although both contain facial images, there is a distribution shift between them due to different image sources and demographic composition. We use MiVOLO as a predictive model [39], a transformer-based face analysis model. To estimate importance weights, we compare three feature spaces: explicit (facial) attributes (Section 3.2.1), model-internal features, and task-agnostic face embeddings derived from ArcFace [16] using InsightFace.² We consider controlled intra-dataset shifts, cross-dataset shifts, image-quality-related shifts, and subgroup shifts relevant for fairness evaluation.

4.1.1 Sanity Checks under Controlled Distribution Shifts. We begin by validating our approach in settings where the distribution shift is fully known. In the binary classification case (gender), we constructed a synthetic target distribution consisting of individuals younger than 20 from the UTKFace dataset. The remaining portion of UTKFace served as the source dataset. Due to the high proportion of babies in UTKFace and the difficulty of gender classification for babies, the performance of the model on the target population is lower. When importance weights were computed from explicit meta-data (as defined in Section 3.2.1), both reweighted and subsampled estimates of precision and recall closely matched the performance measured directly on the target population (Experiment (a) Figure 1).

¹<https://github.com/EvelienvdLaan/importance-sampling>

²<https://www.insightface.ai>

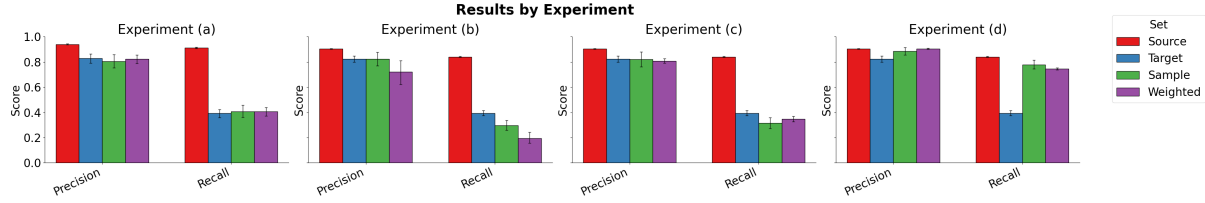


Fig. 1. Precision and recall for four computer vision experiments: (a) known intra-dataset shift with explicit feature weighting; (b) cross-dataset shift with explicit feature weighting; (c) cross-dataset shift with model-internal feature weighting; (d) cross-dataset shift with task-agnostic ArcFace embedding weighting, comparing naive source evaluation, importance sampling, and importance weighting with the ground-truth target evaluation.

We repeated the procedure in a regression setting (age estimation), where the target population was restricted to elderly individuals (>50 years). The MAE increased in the target population relative to the source. Nevertheless, using explicit features including ground-truth age in the weighting step yielded accurate estimates of target performance (Figure 10 in Appendix C).

Takeaway: In line with results from Chen et al. [10], the experiments show that when the axes of distribution shift are fully observed and known, importance weighting can reliably reconstruct performance on the target distribution.

4.1.2 Cross-Dataset Distribution Shifts. Next, we consider cross-dataset settings, where source and target samples are drawn from different benchmarks. We use FairFace as the source data and a subset of UTKFace as target data. Reweighting based on explicit features (as described in Section 3.2.1), aligns the marginal and joint feature distributions between the target and the weighted subsample of the source data (see Figure 11 in Appendix C). Although subsampling does approximate target performance (see Experiment (b) Figure 1), the estimates remain noticeably weaker than in the synthetic setting, where all axes of distribution shift were known and fully observable. This suggests that, in cross-dataset cases, part of the shift is driven by latent factors not captured by the explicit features.

Takeaway: When the true axes underlying distribution shift are unknown, explicit features do not suffice to capture all factors of cross-dataset shifts.

4.1.3 Model-Internal Features as Solution. We then evaluate whether the learned representations by the model provide a more comprehensive feature space for importance weighting. Using the final hidden layer of the gender classification model substantially improved estimates of target performance: both precision and recall approach the ground-truth target values (Experiment (c) Figure 1). Figure 2 provides a two-dimensional visualization (PCA) of the source, target, and subsampled source distributions, illustrating how importance sampling reshapes the source data to better match the target domain.

Beyond shifts in predicted attributes such as gender or age, the model-internal features also captured shifts along other axes, such as image quality, without requiring explicit knowledge of the type of shift (see Figure 12 in Appendix C). This highlights that model-internal features provide a versatile basis for reweighting even when the relevant distributional differences are not directly observable.

In order to use the final hidden layer of the model in the feature space for importance sampling, one needs full access to the weights of the evaluated model. As an alternative, we tested task-agnostic representations. Task-agnostic embeddings are feature representations obtained from models trained for a different objective (e.g., identity recognition via ArcFace [16] rather than age/gender prediction), making them broadly accessible but not aligned with the evaluated model’s decision boundaries. However, these embeddings failed to recover target

performance (see Experiment (d) from Figure 1). Unlike task-specific model-internal features, which encode information directly relevant to gender or age prediction, task-agnostic embeddings capture a broader set of facial attributes, making it harder to obtain a subsample that resembles the target population.

Takeaway: The experiments show that importance weights based on model-internal features enable accurate recovery of target performance, whereas general embeddings do not.

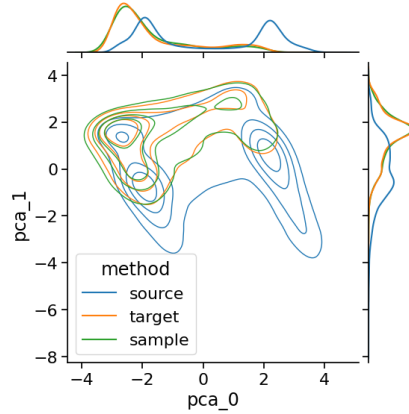


Fig. 2. Two-dimensional PCA projection of model-internal features for the source domain, target domain, and a weighted subsample of the source data. Importance sampling shifts the source distribution toward the target distribution in representation space.

4.1.4 Fairness Evaluation under Dataset Shift. We next evaluate whether importance weighting based on model-internal features can recover subgroup-specific distribution shifts required for fairness analysis. To isolate this effect, we conducted a synthetic shift experiment in which the age distribution for black and white individuals was perturbed between source and target datasets (both based on the UTKFace dataset). As shown in Table 1, importance sampling successfully aligned the distribution of model-predicted features (e.g., age and gender), yet it does not recover the marginal subgroup distribution $P(A)$ nor the conditional subgroup outcome distribution $P(Y|A)$. This divergence arises because the model-internal features capture task-relevant (age and gender estimation) variation but do not encode subgroup distributions in a way that makes them recoverable via reweighting. As a result, fairness metrics computed on the reweighted source data deviated substantially from those measured directly on the target domain (see Figure 3).

In principle, if model-internal features happen to encode sensitive attributes with sufficient fidelity, reweighting based on these features might recover subgroup distributions. However, it is generally unclear when such attributes are encoded and when they are not, and without explicit labels this cannot be verified, which limits the practical applicability of reweighting based on model-internal features for fairness assessment.

Takeaway: The experiments show that importance weighting based on model-internal features was insufficient for reliable fairness evaluation, as sensitive attributes and subgroup outcome distributions were not recoverable from these features.

	Source	Target	Sample
Mean Age	33.28	43.62	41.05
P(Gender)	0.52	0.58	0.58
P(Black)	0.19	0.50	0.20
Mean Age (Black)	33.14	49.50	37.46
Mean Age (White)	38.11	37.75	45.37

Table 1. Summary statistics by sample. Importance sampling aligns model-predicted features but fails to recover subgroup and conditional outcome distributions.

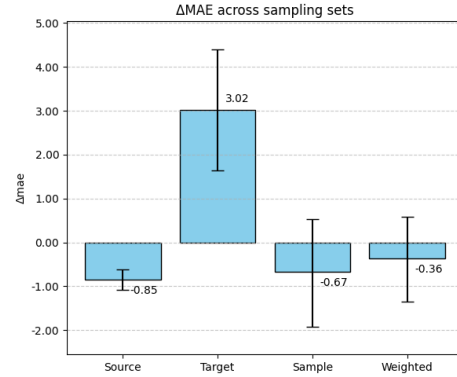


Fig. 3. Group differences in mean absolute error (MAE) between Black and White individuals under $P(X | A)$ shift, comparing naive source evaluation, importance sampling, and importance weighting with the ground-truth target evaluation. Importance weighting based on model-internal features does not recover subgroup error differences.

4.2 Natural Language Processing Experiments

In this section, we examine how importance sampling and importance weighting can be used to evaluate model performance and fairness under distribution shift in natural language processing, using natural language inference, sentiment classification, and language detection as case studies.

4.2.1 Cross-Dataset Distribution Shifts. First, we evaluate our approach on natural language inference (NLI), where labels are "entailment", "neutral", and "contradiction". We use a bert-base-uncased model [17] from Hugging Face, fine-tuned on SNLI [8]. The SNLI validation and test splits serve as source data, while the mismatched MNLI validation set [67] is used as the target domain. Although both datasets contain labeled sentence pairs, they differ in genre, style, and length, leading to a drop in accuracy under distribution shift.

To estimate performance in the target domain, we use task-specific model-internal features (the pooled CLS token from the final hidden layer) as the feature space for importance weighting. Using these representations substantially improves estimates of target performance, with both precision and recall approaching the ground truth target values (Experiments (a) in Figure 13 in Appendix C).

In parallel to the computer vision experiments, we evaluate whether access to model-internal features is necessary by replacing these representations with a general sentence embedding (sentence-transformers/all-mpnet-base-v2 [60]). As in the vision setting, this task-agnostic representation fails to recover target performance (Experiment (b) in Figure 13 in Appendix C).

Takeaway: The experiments show that importance weights based on task-specific model-internal features enable recovery of performance under distribution shift in NLI tasks, whereas general sentence embeddings do not.

4.2.2 Label Shift. Next, we consider a setting in which the distributional shift affects the outcome rather than the covariates. We use a bert-base-uncased model [17], fine-tuned for sentiment analysis on Amazon reviews [18, 41, 44]. The Amazon test split serves as source data. To construct the target sample, we draw tweets from TweetEval [53] and modify the class balance such that the proportion of positive sentiment is $P(Y = 1) = 0.2$. Consistent with the theoretical expectations outlined in Section 2.1.1, altering the marginal outcome distribution

affected precision but left recall unchanged (Figure 4). Note that the change in recall observed in the figure reflects the underlying domain shift between Amazon reviews and tweets, not the label manipulation itself.

Figure 4 displays two experiments. In the first, we applied importance weighting based solely on model-internal features. Because these features approximate $P(Y)$ through $P(\hat{Y})$, reweighting partially aligned the reweighted sample with the target label distribution. However, since $P(\hat{Y})$ is not a perfect proxy for $P(Y)$, due to imperfect prediction accuracy, the precision still deviated from the true target value. In other words, the model does not encode enough information about the actual labels.

To address this, we conducted a second experiment in which the initial importance weights were rescaled to match the empirical target label distribution, as described in Section 3.2.4. Incorporating this outcome information substantially improved precision estimates, see Experiment (b) in Figure 4. Importantly, this requires only a small labeled subsample of the target population, rather than full annotation. In cases where such labeling is infeasible, prior domain knowledge about the expected outcome prevalence could be used to approximate $P(Y)$ in the target population.

Takeaway: Correcting for $P(Y)$ shifts requires outcome information in addition to model features, particularly in settings with lower prediction accuracy.

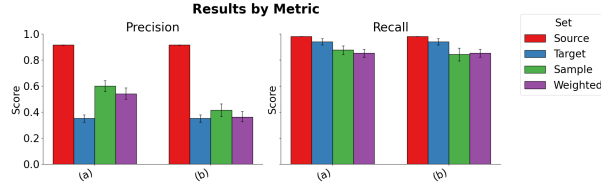


Fig. 4. Precision and recall for two sentiment classification experiments: (a) label shift with model-internal feature weighting; (b) label shift with model-internal feature weighting and rescaling to match the target outcome distribution $P(Y)$. Comparing naive source evaluation, importance sampling, and importance weighting with the ground-truth target evaluation.

4.2.3 Fairness Evaluation with Predicted Sensitive Attributes. We conclude with a fairness setting in which the predicted attribute itself defines the sensitive attribute. Using `fasttext-language-identification` [33], we compare Dutch and Afrikaans detection across two domains: CC100 (longer text segments) [12, 66] as the source and Tatoeba³ (short sentences) as the target. On CC100, the model achieved comparable precision and recall for both languages. When evaluated on Tatoeba, however, recall for Dutch exceeded recall for Afrikaans, indicating that reduced sentence length makes distinguishing between the two languages harder, with accuracy declining more for Afrikaans.

To determine whether this emerging disparity can be anticipated prior to deployment, we applied importance weighting based on model-internal features to subsample CC100 to resemble the target distribution. The weighted CC100 subset reproduced the recall asymmetry observed on Tatoeba, suggesting that reweighting can reveal emerging disparities when the predicted class functions as the sensitive attribute (Figure 5).

Takeaway: The experiment shows that importance weighting can reveal fairness disparities in the target domain when the predicted label is the sensitive attribute.

³<https://tatoeba.org/>

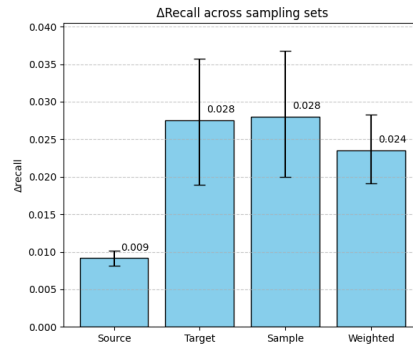


Fig. 5. Difference in recall between Dutch and Afrikaans under cross-domain distribution shift, Comparing naive source evaluation, importance sampling, and importance weighting with the ground-truth target evaluation. The target domain consists of shorter text segments.

5 Discussion and Conclusion

Our analyses highlight that performance and fairness metrics differ in their sensitivity to distribution shift. Metrics conditioned on the true outcome (e.g., recall or equalized odds) remain stable under label shift, whereas metrics conditioned on predicted outcomes (e.g., precision or predictive parity) depend on the outcome distribution. As a result, valid evaluation with these metrics requires evaluation data to match the outcome distribution of the target population. For fairness metrics, this additionally means matching the outcome distribution within each subgroup.

Our results also show that model-internal representations provide an effective feature space for reweighting non-representative evaluation data. Because these representations capture task-relevant structure learned by the model, they enable estimation of target-domain performance without requiring full annotation of deployment data or prior knowledge about how source and target distributions differ. This is particularly useful under cross-dataset shift, where the factors driving the distribution change are not fully observable.

Fairness evaluation is more demanding. Recovering subgroup disparities depends on the extent to which sensitive attributes are represented in the model-internal features. When these attributes are weakly encoded or absent, reweighting may recover overall performance while failing to recover group-level disparities. An important exception arises when the predicted label itself is the sensitive attribute, in which case model-internal features inherently encode group membership. Moreover, label shift requires separate treatment: when outcome prevalence differs between source and target populations, feature-based reweighting alone may not recover correct evaluation metrics. In practice, this can be addressed by adjusting the weights to match the target outcome distribution—either through a small labeled subsample, prior knowledge of expected prevalence, or by using $\hat{Y}$ as a proxy for Y when the model is well calibrated.

Across experiments, importance sampling and importance weighting produce comparable estimates of target performance. Importance weighting typically yields lower variance because it uses all available data, whereas importance sampling can be advantageous when additional annotation is required, since only a representative subset must be labeled.

We evaluate our approach on synthetic, computer vision, and natural language benchmarks to assess its behavior under different types of distribution shift and data modalities. These experiments suggest that the approach can generalize across domains, provided that suitable feature representations are available. However, the considered benchmarks do not cover all possible forms of distribution shift, and may not fully capture the

complexity of real-world deployment environments. Additionally, fairness analysis depends on the availability and quality of sensitive attribute annotations.

5.1 Assumptions and Limitations

Valid performance evaluation can still be supported without representative labeled data, provided that appropriate assumptions are made. First, the conditional relationship $P(Y|X)$ must remain stable; under concept shift, the model no longer represents the target mechanism and reweighting cannot recover valid evaluation metrics. Second, importance weighting requires sufficient support overlap between source and target feature distributions. If the target distribution lies outside the support of the source data, reliable reweighting is not possible. These assumptions are commonly satisfied or approximately hold in many real-world settings where distribution shift is driven by changes in population composition rather than changes in the underlying labeling mechanism.

Recovery of fairness estimates introduces additional constraints. If sensitive attributes are not explicitly available and are insufficiently represented in the feature space used for weighting, subgroup disparities cannot be reliably recovered. Before applying reweighting, it may be useful to assess the extent to which sensitive attributes are encoded in model-internal representations [59]. When sensitive attributes are strongly represented, importance sampling is more likely to recover subgroup distributions, and thus recover actual target model performance and fairness. In settings of weak representation, subgroup recovery would require incorporating explicit group information into the weighting scheme, for example via (automatically generated) annotations.

Finally, the approach inherits standard assumptions of density ratio estimation and importance weighting. We note that several refinements of importance sampling and density ratio estimation have been proposed to address common pitfalls such as high-variance weights and poor behavior under limited support overlap [10, 13, 24, 63, 64]. These techniques can be incorporated directly into our framework without changing its core objective.

5.2 Conclusions

This paper clarifies how accuracy and fairness metrics behave under distribution shift and proposes the use of model-internal features for reweighting as a practical method to approximate target-domain evaluation without fully labeled target data. Three conclusions follow.

First, distribution shift affects metrics unevenly: label-conditioned metrics remain stable under label shift, while prediction-conditioned metrics do not. Second, reweighting based on internal model representations successfully recovers target performance, it shines particularly when the axes of shift are unknown, as in cross-dataset settings. Third, fairness auditing may still require access to sensitive attribute information, as internal features alone may not reliably encode $P(A)$ and $P(Y | A)$.

Overall, importance reweighting based on model-internal features provides an accessible and practical alternative when target data annotations are limited or absent. The core message remains: reliable evaluation requires alignment with the distribution in which models are deployed.

Generative AI Usage Statement

ChatGPT (version 5.2) was used during manuscript preparation to help improve grammar, style, and overall readability of text written by the authors, as well as for limited code assistance such as generating docstrings. Generative AI was not used to create substantive manuscript text or research content. The authors remain fully responsible for the originality, accuracy, and integrity of the work.

Acknowledgments

This work was supported by the Netherlands Police. The authors thank Peter van der Heijden for his valuable feedback and helpful discussions during the development of this paper.

References

- [1] Hervé Abdi and Lynne J Williams. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics* 2, 4 (2010), 433–459.
- [2] Charu C Aggarwal et al. 2018. *Neural networks and deep learning*. Vol. 10. Springer.
- [3] Bang An, Zora Che, Mucong Ding, and Furong Huang. 2022. Transferring fairness under distribution shifts via fair consistency regularization. *Advances in Neural Information Processing Systems* 35 (2022), 32582–32597.
- [4] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.
- [5] Emma Beauxis-Aussalet. 2021. Error Variance, Fairness, and the Curse on Minorities. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 352–365.
- [6] Steffen Bickel, Michael Brückner, and Tobias Scheffer. 2009. Discriminative learning under covariate shift. *Journal of Machine Learning Research* 10, 9 (2009).
- [7] Alexei Botchkarev. 2018. Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. *arXiv preprint arXiv:1809.03006* (2018).
- [8] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326* (2015).
- [9] Tiffany Tianhui Cai, Hongseok Namkoong, and Steve Yadlowsky. 2023. Diagnosing model performance under distribution shift. *arXiv preprint arXiv:2303.02011* (2023).
- [10] Mayee Chen, Karan Goel, Nimit S Sohoni, Fait Poms, Kayvon Fatahalian, and Christopher Ré. 2021. Mandoline: Model evaluation under distribution shift. In *International conference on machine learning*. PMLR, 1617–1629.
- [11] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.
- [12] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 8440–8451. doi:10.18653/v1/2020.acl-main.747
- [13] Corinna Cortes, Yishay Mansour, and Mehryar Mohri. 2010. Learning bounds for importance weighting. *Advances in neural information processing systems* 23 (2010).
- [14] A Philip Dawid. 1982. The well-calibrated Bayesian. *Journal of the American statistical Association* 77, 379 (1982), 605–610.
- [15] Oscar Blessed Deho, Michael Bewong, Selasi Kwashie, Jiuyong Li, Jixue Liu, Lin Liu, and Srecko Joksimovic. 2025. Is it still fair? A comparative evaluation of fairness algorithms through the lens of covariate drift. *Machine Learning* 114, 1 (2025), 8.
- [16] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4690–4699.
- [17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR abs/1810.04805* (2018). arXiv:1810.04805 <http://arxiv.org/abs/1810.04805>
- [18] Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblani, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrom, Roman Solomatin, Ömer Çağatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, Rafał Poświata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Mariya Hendriksen, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Šuppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri Krishnakumar, Anna Maksimova, Silvan Wehrli, Maria Tikhonova, Henil Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide, Lester James Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Hongjin Su, Jimmy Lin, Howard Yen, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. 2025. MMTEB: Massive Multilingual Text Embedding Benchmark. *arXiv preprint arXiv:2502.13595* (2025). doi:10.48550/arXiv.2502.13595
- [19] European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).

- https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689. Published in the Official Journal of the European Union, L 1689, 13 June 2024.
- [20] Haiwen Feng, Timo Bolkart, Joachim Tesch, Michael J. Black, and Victoria Abrevaya. 2022. Towards Racially Unbiased Skin Tone Estimation via Scene Disambiguation. In *European Conference on Computer Vision*.
- [21] Giles M Foody. 2023. Challenges in the real world use of classification accuracy metrics: From recall and precision to the Matthews correlation coefficient. *Plos one* 18, 10 (2023), e0291908.
- [22] Federal Office for Information Security (BSI). 2025. Open Source Face Image Quality (OFIQ), version 1.1.0. <https://github.com/BSI-OFIQ/OFIQ-Project>. Accessed: 2025-11-27.
- [23] Stephen Giguere, Blossom Metevier, Yuriy Brun, Bruno Castro Da Silva, Philip S Thomas, and Scott Niekum. 2022. Fairness guarantees under demographic shift. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*.
- [24] Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, Bernhard Schölkopf, et al. 2009. Covariate shift by kernel mean matching. *Dataset shift in machine learning* 3, 4 (2009), 5.
- [25] Matthew Groh. 2022. Identifying the context shift between test benchmarks and production data. *arXiv preprint arXiv:2207.01059* (2022).
- [26] Furkan Gursoy and Ioannis A Kakadiaris. 2022. Error parity fairness: Testing for group fairness in regression tasks. *arXiv preprint arXiv:2208.08279* (2022).
- [27] Shreyas Havaladar, Jatin Chauhan, Karthikeyan Shanmugam, Jay Nandy, and Aravindan Raghuveer. 2024. Fairness under covariate shift: Improving fairness-accuracy tradeoff with few unlabeled test samples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 12331–12339.
- [28] Thorsten Hempel, Ahmed A. Abdelrahman, and Ayoub Al-Hamadi. 2024. Toward Robust and Unconstrained Full Range of Rotation Head Pose Estimation. *IEEE Transactions on Image Processing* 33 (2024), 2377–2387. doi:10.1109/TIP.2024.3378180
- [29] Dan Hendrycks and Kevin Gimpel. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136* (2016).
- [30] Shohei Hido, Yuta Tsuboi, Hisashi Kashima, Masashi Sugiyama, and Takafumi Kanamori. 2011. Statistical outlier detection using direct density ratio estimation. *Knowledge and information systems* 26, 2 (2011), 309–336.
- [31] Zhimeng Stephen Jiang, Xiaotian Han, Hongye Jin, Guanchu Wang, Rui Chen, Na Zou, and Xia Hu. 2023. Chasing fairness under distribution shift: A model weight perturbation approach. *Advances in Neural Information Processing Systems* 36 (2023), 63931–63944.
- [32] Ian T Jolliffe and Jorge Cadima. 2016. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences* 374, 2065 (2016), 20150202.
- [33] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of Tricks for Efficient Text Classification. *arXiv preprint arXiv:1607.01759* (2016).
- [34] Nathan Kallus and Angela Zhou. 2018. Residual unfairness in fair machine learning from prejudiced data. In *International Conference on Machine Learning*. PMLR, 2439–2448.
- [35] Kimmo Karkkainen and Jungseock Joo. 2021. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1548–1558.
- [36] Masanari Kimura and Hideitsu Hino. 2024. A short survey on importance weighting for machine learning. *arXiv preprint arXiv:2403.10175* (2024).
- [37] Teun Kloek and Herman K Van Dijk. 1978. Bayesian estimates of equation system parameters: an application of integration by Monte Carlo. *Econometrica: Journal of the Econometric Society* (1978), 1–19.
- [38] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*. PMLR, 5637–5664.
- [39] Maksim Kuprashevich and Irina Tolstykh. 2023. Mivolo: Multi-input transformer for age and gender estimation. In *International conference on analysis of images, social networks and texts*. Springer, 212–226.
- [40] Chao Lan and Jun Huan. 2017. Discriminatory transfer. *arXiv preprint arXiv:1707.00780* (2017).
- [41] Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. *Proceedings of the 7th ACM conference on Recommender systems* (2013). <https://api.semanticscholar.org/CorpusID:6440341>
- [42] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D’Amour, and Kristian Lum. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual review of statistics and its application* 8, 1 (2021), 141–163.
- [43] Jose G Moreno-Torres, Troy Raeder, Rocio Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. 2012. A unifying view on dataset shift in classification. *Pattern recognition* 45, 1 (2012), 521–530.
- [44] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. MTEB: Massive Text Embedding Benchmark. *arXiv preprint arXiv:2210.07316* (2022). doi:10.48550/ARXIV.2210.07316
- [45] Thomas Nagler and Claudia Czado. 2016. Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas. *Journal of Multivariate Analysis* 151 (2016), 69–89.
- [46] Michael A Nielsen. 2015. *Neural networks and deep learning*. Vol. 25. Determination press San Francisco, CA, USA.

- [47] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems* 32 (2019).
- [48] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.
- [49] David MW Powers. 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061* (2020).
- [50] Joaquin Quiñero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. 2022. *Dataset shift in machine learning*. MIT Press.
- [51] Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian D Ziebart. 2021. Robust fairness under covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 9419–9427.
- [52] Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. 2023. Improving fair training under correlation shifts. In *International Conference on Machine Learning*. PMLR, 29179–29209.
- [53] Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*. 502–518.
- [54] Jessica Schrouff, Natalie Harris, Sanmi Koyejo, Ibrahim M Alabdulmohsin, Eva Schnider, Krista Opsahl-Ong, Alexander Brown, Subhrajit Roy, Diana Mincu, Christina Chen, et al. 2022. Diagnosing failures of fairness transfer across distribution shift in real-world medical settings. *Advances in Neural Information Processing Systems* 35 (2022), 19304–19318.
- [55] Candice Schumann, Xuezhi Wang, Alex Beutel, Jilin Chen, Hai Qian, and Ed H Chi. 2019. Transfer of machine learning fairness across domains. *arXiv preprint arXiv:1906.09688* (2019).
- [56] Hongwei Shang, Jean-Marc Langlois, Kostas Tsioutsoulis, and Changsung Kang. 2023. Precision/recall on imbalanced test data. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 9879–9891.
- [57] Minglai Shao, Dong Li, Chen Zhao, Xintao Wu, Yujie Lin, and Qin Tian. 2024. Supervised algorithmic fairness in distribution shifts: A survey. *arXiv preprint arXiv:2402.01327* (2024).
- [58] Hidetoshi Shimodaira. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference* 90, 2 (2000), 227–244.
- [59] Congzheng Song and Vitaly Shmatikov. 2019. Overlearning reveals sensitive attributes. *arXiv preprint arXiv:1905.11742* (2019).
- [60] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MpNet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems* 33 (2020), 16857–16867.
- [61] Adarsh Subbaswamy, Roy Adams, and Suchi Saria. 2021. Evaluating model robustness and stability to dataset shift. In *International conference on artificial intelligence and statistics*. PMLR, 2611–2619.
- [62] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. 2012. *Density ratio estimation in machine learning*. Cambridge University Press.
- [63] Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul Von Büna, and Motoaki Kawanabe. 2008. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics* 60, 4 (2008), 699–746.
- [64] Surya T Tokdar and Robert E Kass. 2010. Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics* 2, 1 (2010), 54–60.
- [65] Haotao Wang, Junyuan Hong, Jiayu Zhou, and Zhangyang Wang. 2023. How robust is your fairness? evaluating and sustaining fairness under unseen distribution shifts. *Transactions on machine learning research* 2023 (2023), <https://openreview>.
- [66] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 4003–4012. <https://aclanthology.org/2020.lrec-1.494/>
- [67] Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426* (2017).
- [68] Cort J Willmott and Kenji Matsuura. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research* 30, 1 (2005), 79–82.
- [69] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems* 34 (2021), 12077–12090.
- [70] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. 2024. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision* 132, 12 (2024), 5635–5662.
- [71] Song Yang Zhang, Zhifei and Hairong Qi. 2017. Age Progression/Regression by Conditional Adversarial Autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

A Overview of Conducted Experiments

Task	Shift	Source	Target	Model	Weighting Features
Synthetic (Cls.)	Covariate, Label, Fairness	Simulated	Simulated	Logistic	$X(3), A(1), \hat{Y}(1)$
Age / Gender (CV)	Intra dataset	UTKFace	Subset UTKFace	MiVOLO	Explicit (10)
Gender (CV)	Cross dataset	Fairface	Subset UTKFace	MiVOLO	Explicit (10), Internal (384), Arcface (512)
Gender (CV)	Quality	FairFace	Subset UTKFace	MiVOLO	Internal (384)
Age (CV)	Fairness	UTKFace	Subset UTKFace	MiVOLO	Internal (384)
NLI	Cross-dataset	SNLI	MNLI (mm)	BERT	Internal (768), MPNet (768)
Sentiment	Cross-domain, Label	Amazon	Twitter	BERT	Internal (768)
Language Detection	Cross-domain	CC100	Tatoeba	fastText	Internal (256)

Table 2. Overview of conducted experiments, datasets, and tasks. Feature dimensionalities (before PCA) are shown in parentheses.

B Synthetic Experiments

In this appendix, we present controlled synthetic experiments that illustrate how importance sampling and importance weighting behave under different types of distributional mismatch. These experiments allow us to isolate specific mechanisms—such as covariate shift, label shift, and fairness-related shifts—and to verify the theoretical insights discussed in the main text.

B.1 Data Generating Process

We simulate binary outcome data $Y \in \{0, 1\}$ based on covariates $\mathbf{X} \in \mathbb{R}$ and a binary sensitive attribute $A \in \{0, 1\}$. The probability of the positive class is governed by a logistic model:

$$\begin{aligned} \text{logit}(P(Y = 1 \mid \mathbf{X}, A)) &= \beta_0 + \mathbf{X}^\top \boldsymbol{\beta}_1 + A\beta_2, \\ Y &\sim \text{Bernoulli}(\text{expit}(\text{logit})) \end{aligned} \quad (4)$$

Unless stated otherwise, we set $\beta_0 = -2$, $\boldsymbol{\beta}_1 = \mathbf{1}$, and $\beta_2 = 2$. For experiments involving extreme shifts in $P(A)$, we use $\beta_2 = 5$, while for fairness experiments we set $\beta_2 = 0$. Exact parameter configurations are provided on GitHub.

In the source domain, covariates are drawn as $\mathbf{X} \sim \mathcal{N}(0, 2)$ and $A \sim \text{Bernoulli}(0.5)$, yielding a class prevalence of $P(Y = 1) = 0.41$. Target domains are generated by introducing distributional shifts relative to the source, including changes to $P(A)$, shifts in the (marginal or conditional) distribution of $\mathbf{X}$, and adjustments to $P(Y)$ to simulate label shift.

We generate a fixed **source dataset** of size $n = 250,000$, multiple **target datasets** of size $n = 10,000$, and **source test subsets** of size $n = 125,000$.

B.2 Predictive Model and Density Ratio Estimation

The model whose performance we aim to estimate is fitted on the training subset of the source dataset. We use logistic regression to model $P(Y \mid \mathbf{X}, A)$.

Density ratio estimation is performed using a random forest classifier trained to distinguish between samples from the target dataset and the source test set. The predicted probabilities are then used to compute importance weights

$$w(\mathbf{x}) = \frac{p_{\text{target}}(\mathbf{x})}{p_{\text{source}}(\mathbf{x})}.$$

Our approach is model-agnostic and does not rely on assumptions about the predictive model. While we present results using logistic regression, we observe similar qualitative patterns when using more flexible models such as random forests to generate predictions $\hat{Y}$.

B.3 Experimental Scenarios

Using the data-generating process described above, we construct synthetic experiments that isolate different types of distributional mismatch. Specifically, we consider:

- **Covariate shifts**, induced by changes in $P(X)$ or subgroup proportions $P(A)$ while keeping $P(Y | X, A)$ fixed;
- **Label shift**, induced by changes in the marginal outcome distribution $P(Y)$;
- **Fairness-related shifts**, induced by changes in $P(X | A)$ that affect group-level performance without altering the conditional outcome model.

These controlled settings allow us to evaluate when importance sampling and weighting recover target-domain performance and fairness, and when they fail.

B.4 Results

B.4.1 Covariate Distribution Shifts. We first evaluate our method under covariate shift, induced either by translating the mean of X or by altering subgroup proportions $P(A)$, while keeping $P(Y | X, A)$ fixed. Figure 6 shows that as the mean of X in the target population drifts further from the source distribution, naive source evaluation increasingly deviates from true target performance, whereas both importance sampling and importance weighting closely approximate target-domain performance. Experiment (a) in Figure 7 illustrates the same pattern under an extreme shift in $P(A)$.

Takeaway: Importance weighting reliably corrects for covariate shifts, provided there is sufficient overlap in support.

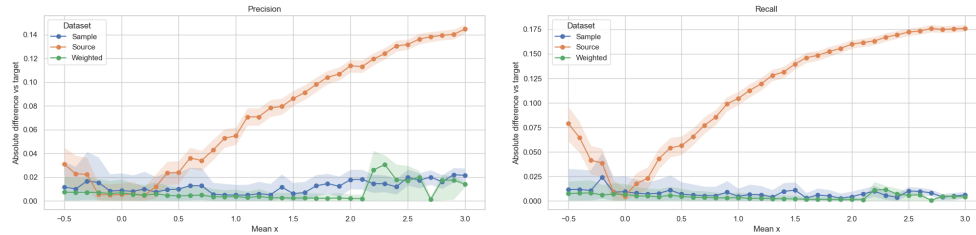


Fig. 6. Estimation error of target-domain precision and recall under increasing covariate shift in $P(X)$, comparing naive source evaluation, importance sampling, and importance weighting relative to the ground-truth target evaluation. The mean of the target covariate distribution is progressively shifted away from the source distribution.

B.4.2 Label Shift. Next, we simulate label shift by changing the prevalence of Y in the target distribution. As expected, recall remains stable while precision decreases as $P(Y = 1)$ declines (Experiment (b), Figure 7), confirming the theoretical result from Section 2.1.1.

Rescaling importance weights to match the target outcome distribution $P(Y)$, as described in Section 3.2.4, restores precision even under large prevalence shifts (Figure 8).

Takeaway: Label shifts affect prevalence-dependent metrics and can be corrected using outcome information.

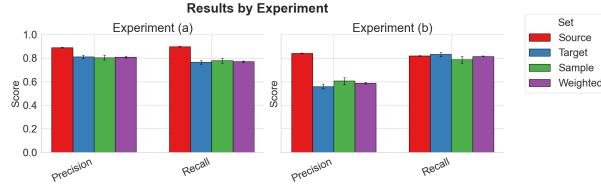


Fig. 7. Estimated precision and recall under synthetic covariate and label shift, comparing naive source evaluation, importance sampling, and importance weighting with the ground-truth target evaluation. Experiment (a) applies an extreme subgroup proportion shift in $P(A)$; Experiment (b) applies label shift with importance weights constructed using outcome information. Error bars denote standard errors across simulation runs.

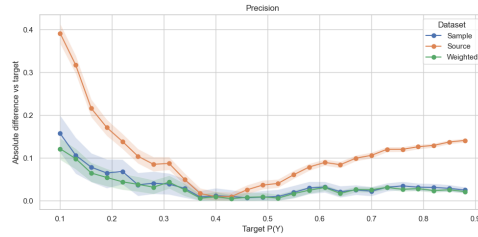


Fig. 8. Estimation error of target-domain precision and recall under varying outcome prevalence $P(Y)$, comparing naive source evaluation, importance sampling, and importance weighting relative to the ground-truth target evaluation. Precision varies with $P(Y)$.

B.4.3 Fairness Evaluation. We introduce a conditional covariate shift by sampling X from group-specific distributions, altering $P(X | A)$ while keeping $P(Y | X, A)$ unchanged. Although source evaluation shows no disparity, target evaluation reveals group-level differences in recall (Figure 9). Importance weighting recovers these disparities.

Takeaway: Importance weighting can reveal fairness disparities that remain hidden when evaluation is performed on non-representative source data.

B.4.4 Importance Sampling versus Importance Weighting. Finally, we compare importance sampling and importance weighting. While both methods yield similar point estimates, importance weighting produces narrower confidence intervals due to its use of the full source dataset (Figures 6, 7, and 8). Note that when manual annotation of evaluation data is required, subsampling may be preferable, since the estimates are nearly identical, yet the annotation effort is greatly reduced.

Takeaway: Importance sampling and weighting perform similarly; weighting improves statistical efficiency, while sampling reduces annotation effort.

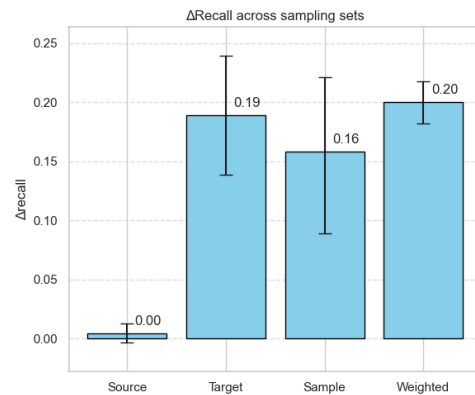


Fig. 9. Group differences in recall under a conditional covariate shift in $P(X | A)$, comparing source evaluation and importance-weighted estimates with the ground-truth target evaluation. The shift is induced by changing the mean of X for one sensitive group.

C Supplementary Results

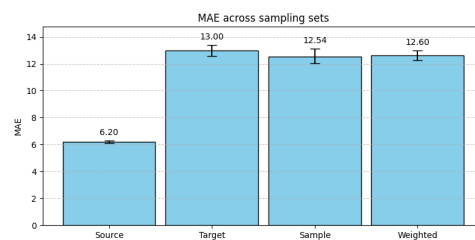


Fig. 10. Estimated mean absolute error (MAE) for age prediction under a controlled covariate shift to older individuals, comparing naive source evaluation, importance sampling, and importance weighting with the ground-truth target evaluation. Importance weights are computed using explicit features, including ground-truth age.

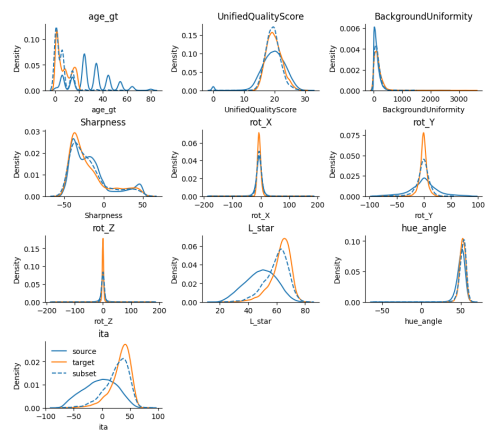


Fig. 11. Marginal distributions of explicit features for the source dataset (FairFace) and the target dataset (UTKFace subset). Reweighting the source samples using explicit features aligns the weighted sample of the source distributions with those of the target.

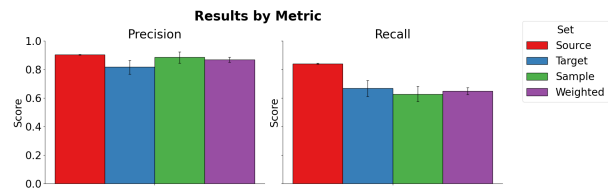


Fig. 12. Precision and recall for cross-dataset shift driven by image quality with model-internal feature weighting comparing naive source evaluation, importance sampling, and importance weighting with the ground-truth target evaluation.

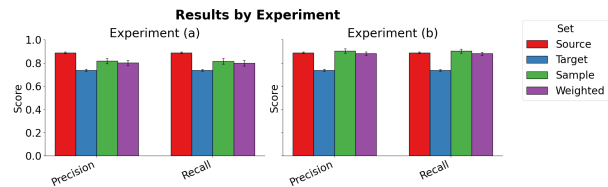


Fig. 13. Precision and recall for two natural language inference experiments: (a) cross-dataset shift with task-specific model-internal feature weighting; (b) cross-dataset shift with task-agnostic sentence embedding weighting, comparing naive source evaluation, importance sampling, and importance weighting with the ground-truth target evaluation.

D Bayes' Decomposition Predictive Parity

$$\begin{aligned}
P(Y|\hat{Y}, A) &= \frac{P(Y, \hat{Y}, A)}{P(\hat{Y}, A)} \\
&= \frac{P(\hat{Y} | Y, A)P(Y, A)}{P(\hat{Y}, A)} \\
&= \frac{P(\hat{Y}|Y, A)P(Y, A)P(A)}{P(\hat{Y}|A)P(A)} \\
&= \frac{P(\hat{Y}|Y, A)P(Y|A)}{P(\hat{Y}|Y = 1, A)P(Y = 1|A) + P(\hat{Y}|Y = 0, A)P(Y = 0|A)}.
\end{aligned} \tag{5}$$

Received 13 January 2026; revised 25 March 2026; accepted 16 April 2026