Université Paris Cité

DriftFair: A socio-technical approach to maintaining fairness under covariate shift

Marc Hulcelle, Sijie Dong, Soror Sahri, Themis Palpanas

LIPADE-TR-N° 13

Avril 2026

Technical Report

LIPADE

Laboratoire d'Informatique PAris DEscartes

DriftFair: A socio-technical approach to maintaining fairness under covariate shift

Marc Hulcelle
Université Paris Cité
Paris, France
marc.hulcelle@u-paris.fr

Sijie Dong
Université Paris Cité
Paris, France
sijie.dong@etu.u-paris.fr

Soror Sahri
Université Paris Cité
Paris, France
soror.sahri@parisdescartes.fr

Themis Palpanas
Université Paris Cité
Paris, France
themis@mi.parisdescartes.fr

Abstract

Ensuring fairness in machine learning is particularly challenging when data distributions shift between the training and serving phases. Such *data drift* can degrade model performance and worsen biases against underrepresented populations. We present DriftFair, a socio-technical pipeline that detects harmful drift, interprets where fairness is at risk, and mitigates biased outcomes through targeted data augmentation. First, DriftFair locates low-accuracy regions correlated with fairness-sensitive attributes, labeling them as "unfairness-prone." It then monitors changes in these regions under new data; if fairness metrics deteriorate, we classify the drift as harmful and selectively retrain the model with synthetic samples to rebalance the representation. Throughout, social science insights guide threshold definition, identify sensitive attributes, and contextualize distributional changes. Experiments on the U.S. Census Bureau's American Community Survey (ACS) and the COMPAS dataset show that our pipeline preserves fairness under realistic shifts, underscoring the importance of integrating technical detection strategies with social-contextual understanding.

1 Introduction

Machine learning (ML) is increasingly used to make decisions with high social impact like hiring, medical diagnosis, education, and criminal justice. These applications directly impact people's lives and can harm our society if not designed and engineered with fairness and ethics considerations. Previous work studied the existence and effects of disparities in ML applications and suggest they can be observed in society at various levels and perpetuated in ML applications [31].

Potential sources of fairness issues in machine learning predictions can arise from biases in the data causing unfair decisions [11, 29]. One form of bias is data drift [2], which refers to changes in the distribution of features such as gender, race, sex, and age among others. It may be harmful as it can negatively impact the performance of machine learning models in a significant way, leading the models to inaccurate predictions and unfair decision-making. Data drift occurs in two ways [19, 21]: covariate shift, which refers to changes in the distribution of input features between training and serving data; and concept drift which occurs when the relationship between input features and the target variable changes over time. If the drift disproportionately affects certain features' subgroups, this can lead to an accuracy drop in the model's performance, as well as fairness, thus perpetuating and possibly worsening pre-existing societal biases. Addressing data drift is then crucial to mitigate unfair decisions.

From a social science perspective, under-representation can translate into systematic model biases that disproportionately affect marginalized groups, as illustrated in Figure 1. However, not all shifts require immediate intervention: if there is a slight decrease in an already small subpopulation and a corresponding increase in a majority group, fairness metrics may remain stable, making costly retraining unnecessary. Thus, the central motivation is to distinguish between harmful and non-harmful drift in a manner that robustly preserves fairness without continuously updating the model for benign distributional changes. To address this, we introduce DriftFair, a socio-technical framework designed to detect and mitigate unfairness issues under covariate shifts.

A key challenge is that generic drift detection methods often fail to assess the social implications of model errors. They may signal a distributional shift without indicating whether it disproportionately harms sensitive groups. Conversely, existing fairness-aware methods typically assume a stable data distribution. Designing a pipeline that pinpoints exactly where feature space bias intensifies due to data drift while avoiding unwarranted retraining, requires an integrated approach. Social science expertise is essential for determining which attributes are sensitive in a given societal context, setting domain-relevant thresholds for harmful drift, and interpreting fairness beyond raw metrics. Moreover, balancing accuracy and fairness becomes more complex when the serving data do not mirror training distributions. In high-stakes scenarios, unrecognized biases may reinforce existing disparities, while unnecessary retraining on every small shift imposes resource and labor costs. Addressing these issues demands a careful multi-disciplinary approach that acknowledges the interplay between statistical drift and social facts. Therefore, we formulate the following research questions: (**RQ1**) What is the impact of data drift on fairness? (**RQ2**) Can we detect fairness issues during the detection and mitigation of data drift? (**RQ3**) Is there a trade-off between accuracy and fairness? (**RQ4**) How does the choice of a fairness framework impact different populations?

The remainder of this paper is organized as follows. In Section 2, we discuss related work on fairness in ML and existing approaches to data drift detection. Section 3 provides background on data drift and introduces the formal definitions of fairness metrics used in our

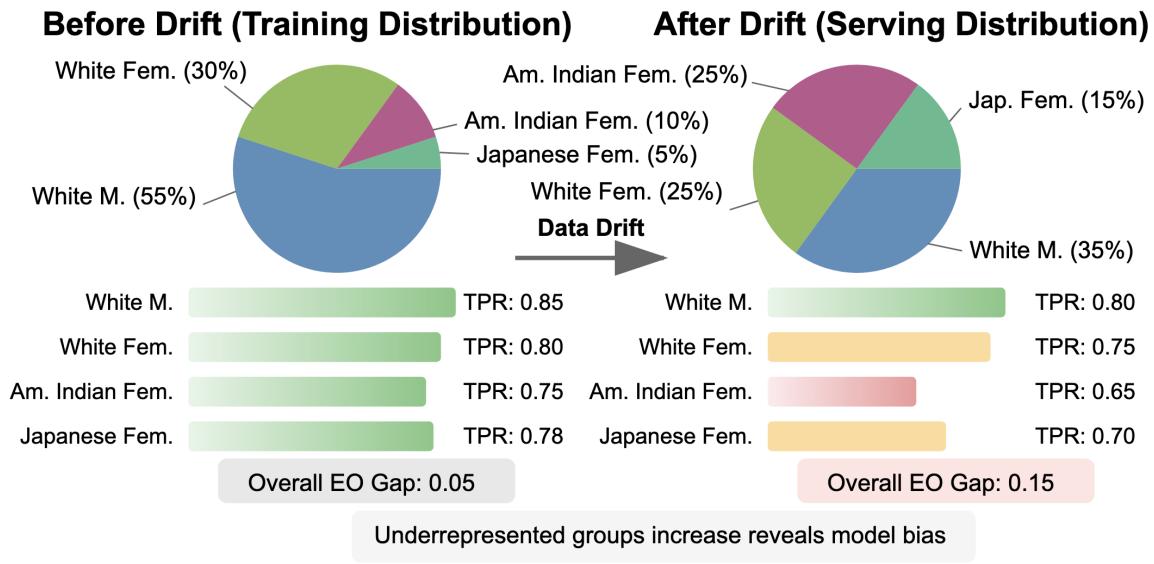

Figure 1: Example of covariate shift. Here, the proportion of initially under-represented groups significantly rises to the point of being the majority group. An ML model that faces such data drift might reveal its inherent biases towards minority groups. ***TPR:*** **True Positive Rate.** ***EO:*** **Equal Opportunity metric, gap being represented by the difference between TPR.**

study. Section 4 presents DriftFair, detailing its steps from fairness-aware drift detection to mitigation strategies, addressing RQ1 and RQ2. In Section 5, we present experimental results on the ACS and COMPAS datasets, illustrating how our method identifies and addresses harmful drift, thus answering RQ3. Section 6 concludes with a discussion of the broader implications of our socio-technical perspective and directions for future research. Answers to RQ4 are provided throughout the paper, as the choice of a fairness framework has implications over the entirety of the pipeline, from the theoretical conception to its application on real-world data and the analysis of results.

2 Related Work

2.1 Data Drift

Data drift presents a significant challenge in maintaining fairness in ML, as shifts in feature distributions can exacerbate biases and degrade model performance [8]. Several approaches aim to mitigate fairness degradation under such conditions. A model-agnostic framework has been proposed to balance accuracy and multiple fairness constraints while ensuring theoretical guarantees [43]. Another line of work introduces a meta-algorithm that reformulates fairness constraints into convex classification problems with provable guarantees [12]. To address trade-offs between accuracy and fairness, a multi-objective optimization framework has been developed to select Pareto-optimal models, an approach we adopt due to its flexibility [33].

Beyond static fairness-aware models, recent research explores fairness under data drift. One approach integrates drift detection into an entropy-based fairness-aware objective to ensure stability over time [23]. Decision tree classifiers have also been modified to maintain statistical parity when deployed in non-stationary environments [44]. Another strategy employs Conformance Constraints to detect drift and adjust models dynamically, either by selecting the best-matching model (DIFFAIR) or reweighting training samples (CONFAIR) to align with serving data without requiring sensitive attributes at inference [41]. However, a comparative study evaluating multiple fairness-aware algorithms under covariate shift reveals that no existing method remains fully robust when subgroup-specific drift is large [15], emphasizing the need for more adaptive solutions.

2.2 A Socio-Technical Approach to Fairness

Maintaining fairness in ML applications is not only a technical challenge but is the object of increasingly multidisciplinary research, to which legal specialists and social scientists contribute [6, 25]. We argue that fairness has mostly been studied from an ML point of view, which reinforces a techno-solutionism approach to societal issues. Such an approach can lead to a decontextualization of issues that should be countered [7, 8]. The social perspective links fairness metrics to social practices to better understand the roots of fairness issues. For example, in hiring applications, it is demonstrated that job applicants are affected by and relate to gender stereotypes, e.g. men tend to apply for more technical and ambitious jobs when compared to women [24]. According to the Gender Equality Division of UNESCO, ML systems used in the recruitment software are biased against women [36]. Despite the recent interest of the research community on fairness in machine learning related to gender biases, more research work and studies are needed to address gender disparities and make ML applications fairer to all users.

3 Preliminaries

In this section, we present the concept of data drift and define the fairness metrics used to assess model performance under evolving distributions. We then discuss how existing drift detection frameworks inform our approach to mitigating fairness concerns.

3.1 Data Drift

Data drift, also referred to as data shift, occurs when the distribution of input features changes between training and serving data. It can be broadly categorized into *covariate shift*, where the marginal distribution of input features changes while the conditional distribution of the target remains the same, and *concept drift*, where the relationship between input features and the target variable itself evolves over time. In this work, we primarily focus on covariate shift, as it directly impacts model predictions while keeping the underlying decision boundary relatively stable.

Several methods have been proposed to detect and localize data drift [1, 5, 27, 39]. Typically, these approaches compare the training distribution of features $\mathcal{D}_{\text{train}}$ with the model's predictions distribution $\mathcal{D}_{\text{serve}}$. Recent frameworks go further by identifying where in the feature space the drift occurs, enabling more targeted responses. Among these, Detection of Drift in *Low-Accuracy Areas* (DDLA) [16] pinpoints sub-regions where the model's predictive accuracy declines the most when moving from training to serving data. In this work, we adopt a region-based perspective inspired by DDLA to precisely capture the interplay between drift and fairness: we seek not just any drift, but drift that leads to a disproportionate increase in errors for specific subgroups-e.g. determined by sensitive attributes.

Drift can be harmful from a fairness perspective if it significantly increases model inaccuracies or disparities for specific subgroups. We therefore monitor how fairness metrics change when data shift and localize the drift to low-accuracy regions correlated with sensitive attributes. This provides a more robust identification of biased outcomes than aggregate measures alone.

3.2 Fairness Metrics

We focus on group-based fairness notions that compare performance across demographic subgroups of a sensitive attribute A, such as gender, or race. Let $\hat{Y}$ be the model's prediction and Y the ground truth. We use two representative metrics in our experiments [10, 20]:

Demographic Parity (DP), also called Statistical Parity, requires:

$$P(\hat{Y} = 1 \mid A = 0) \;=\; P(\hat{Y} = 1 \mid A = 1),$$

indicating that a protected group ($A = 1$) receives positive outcomes (here, $\hat{Y} = 1$) at the same rate as the non-protected group ($A = 0$). This metric is used to measure disparities in how different groups receive beneficial treatment.

Equal Opportunity (EO) focuses on true positive rates:

$$P(\hat{Y} = 1 \mid A = 0, Y = 1) \;=\; P(\hat{Y} = 1 \mid A = 1, Y = 1).$$

Here, the emphasis is on whether each group have similar recall (or TPR), which is crucial when false negatives can produce serious adverse consequences. Other fairness definitions exist, such as Equalized Odds, Conditional Statistical Parity, but DP and EO capture two complementary perspectives: overall rate of positive predictions and error rate disparities on true positives. In the rest of the paper, we use DP and EO, but our methodology is flexible to other fairness metrics.

In the next section, we detail our pipeline, which builds on the above notions of data drift and fairness metrics. Our method classifies drift as harmful only if it magnifies unfairness, and selectively updates the model with targeted data augmentation rather than retraining for every observed distributional shift.

4 Proposed Approach

4.1 Formalization

Let $\mathcal{D}_{\text{train}}$ be the training distribution of (X, A, Y), where $X \in \mathbb{R}^d$ denotes non-sensitive features, A a sensitive attribute, and Y the binary label. A model $M : \mathbb{R}^d \times \Omega_A \to \{0, 1\}$ is trained on samples from $\mathcal{D}_{\text{train}}$. At serving time, samples arrive from a possibly shifted distribution $\mathcal{D}_{\text{serve}}$. We note $\mathcal{R}_{\text{low}} \subseteq \mathcal{X} \times \Omega_A$ as the region where M has low accuracy with respect to a threshold τ. Let $\mathcal{R}_{\text{unfair}} \subseteq \mathcal{R}_{\text{low}}$ be the subset in which performance disparities across subgroups $(A = 0)$ and $(A = 1)$ are high under a fairness metric $\mathcal{F}$. We note $\pi_{\text{unfair}}(\mathcal{D}_{\text{serve}})$ the proportion of $\mathcal{D}_{\text{serve}}$ lying in $\mathcal{R}_{\text{unfair}}$.

A drift is *non-harmful* if $\pi_{\text{unfair}}(\mathcal{D}_{\text{serve}}) \leq \pi_{\text{unfair}}(\mathcal{D}_{\text{train}}) + \delta$ and $\mathcal{F}(\mathcal{D}_{\text{serve}}) + \delta \geq \mathcal{F}(\mathcal{D}_{\text{train}})$ for a small $\delta > 0$. Otherwise, if either $\pi_{\text{unfair}}(\mathcal{D}_{\text{serve}})$ exceeds or $\mathcal{F}(\mathcal{D}_{\text{serve}})$ falls behind its training counterpart by more than some $\eta > \delta$, the drift is *harmful* and triggers retraining. Specifically, additional samples are generated to cover $\mathcal{R}^*_{\text{unfair}} \subseteq \mathcal{R}_{\text{unfair}}$, where the model shows biased or inaccurate outcomes under $\mathcal{D}_{\text{serve}}$. Training on these augmented samples is aimed at improving fairness metrics while accommodating the new data distribution.

4.2 DriftFair Architecture

The main steps of our proposed DriftFair approach are depicted in Figure 2. We present each step in the following subsections.

4.2.1 Fairness model assessment. Our proposed approach starts with an initial assessment of the model's performance and fairness on the test set from the training data. In the rest of the paper, we will be using the following metrics: accuracy, equal opportunity (EO), and demographic parity (DP). Although we use accuracy as a metric for all predictions, EO and DP are used to measure differences in predictions for both genders. As explained in Section 3.2, EO measures the difference in the true positive rate (TPR) between genders, while DP measures the distribution difference between genders. We chose these metrics because we want to make sure that our models reduce their amount of false negatives equally for both genders, given that false negatives had strong implications for the use-case of our datasets.

4.2.2 Fairness detection through data drift. Using the DDLAs framework for data drift detection and mitigation proposed in [16] and presented above, we detect low-accuracy regions. We then leverage these identified regions to assess the presence of sensitive features. For example, if low-accuracy regions correlate with a sensitive feature such as gender, we label these regions as "unfairness-prone regions," as they are more likely to exhibit fairness issues under data drift. This will result in analyzing the fairness under data

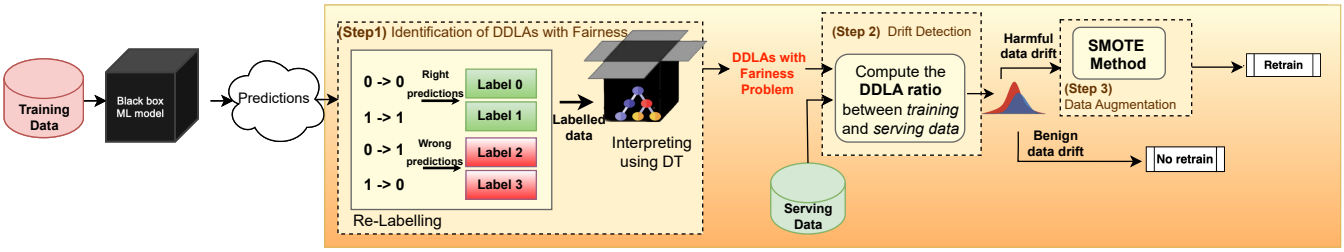

Figure 2: Pipeline of our framework to detect fairness issues during data drift. We train a model and compute fairness metrics for an initial assessment. Once the model is trained, we relabel its predictions to determine which type of error it makes. We train a Decision Tree on the relabeled data and use it to detect Data Distribution with Low Accuracy (DDLA). We look for fairness issues among training data by examining if these DDLAs include a sensitive attribute. We compare the DDLA ratio between training and serving data to determine whether harmful drift is occurring, thus potentially unveiling fairness issues in the serving data. We retrain the model with data augmentation to mitigate previously detected unfairness.

drift according to two scenarios: (i) occurs when DDLAs include sensitive attributes, thus revealing fairness issues linked to these attributes; (ii) arises when DDLAs don't allow to detect fairness-related to data drift, i.e. DDLAs don't involve sensitive attributes. This scenario could imply either no fairness issues exist, or fairness issues are embedded within the model itself due to learned biases from the training data. Following this, we compare the identified low-accuracy regions, with the serving data's to assess whether the data drift is harmful or not.

4.2.3 Fairness mitigation in data drift. We monitor data drift by comparing distributions between the training data and the serving data by examining whether data drift has caused an increase in the ratio of these unfairness-prone regions within the serving data compared to the training data.

When harmful drift is detected, we first apply a data augmentation technique inspired by SMOTE to balance the representation of fair and unfair regions. We first identify samples from the training data that are deemed fair according to fairness metrics. From there, we generate synthetic samples that extend in the feature space direction of samples considered unfair, while aiming to preserve the ground truth as closely as possible. This helps create a balanced distribution aligned with the new serving data, reducing bias in unfairness-prone regions. We then retrain the model with the additional augmented data to improve fairness metrics specifically within the newly identified unfair regions, addressing disparities that arise from the detected harmful data drift. We re-evaluate the model's fairness metrics on the serving data to assess improvement after retraining. At this final step, social scientists can conduct an analysis to ensure that new biases are not introduced and existing unfairness is not amplified.

4.2.4 Social Science feedback for Maintaining fairness . The integration of social science perspective into our approach enhances both the societal relevance and interdisciplinary robustness of our approach. This interdisciplinary feedback-driven process allows to enrich model fairness by embedding social context, technical fairness metrics, and anticipates modifications made to the ML models. In the fairness model assessment step, social scientists can play a crucial role in selecting and defining fairness metrics that best reflect real-world social justice issues. They also assist in determining which features are considered sensitive within specific socio-cultural contexts, such as gender, race, age, and socioeconomic status. In the fairness detection step, for scenarios where DDLAs allow to detect fairness issues due to drift by identifying sensitive attributes, social science expertise provides critical insights into how these drifts may impact protected groups. This input complements fairness metrics, such as demographic parity or equalized odds, which quantify fairness but lack contextual understanding. For black-box models, where interpretability challenges hide how decisions are made, interdisciplinary feedback clarifies fairness risks tied to data drift, guiding model adjustments to align outputs with both technical fairness metrics and social fairness standards.

5 Experimental Evaluation

All experiments were performed with Python 3.9. The source code for our experiments will be available on GitHub at the following repository, allowing reproducibility and further exploration of the results.

We apply our pipeline to conduct a between-states analysis of the US home ownership market and uncover fairness issues. For this, we first apply our pipeline to discover fairness issues that are related to data drift by training a random forest that only optimizes accuracy. We call this our baseline model. We compare this with another version of the pipeline where we replace the random forest with a neural network that optimizes for both accuracy and other fairness metrics such as EO and DP[33], presented in Section 3.2. Fairness issues can be detected by comparing performance between genders according to DP and EO.

Datasets: We collected data from the U.S. Census Bureau's American Community Survey (ACS) Public Use Microdata Sample (PUMS) files [1] from 2021. These files are a set of records from individual people that correspond to housing units, with disclosure protection enabled so that individuals or housing units cannot be identified. We downloaded data for each of the 52 states individually and proceeded to keep 165 variables out of all available ones. Variables

[1] https://www.census.gov/programs-surveys/acs/microdata.html

describe the respondent of the survey - e.g. ethnicity, monthly income -, the family it lives with - e.g. total family income, total number of people, their relationships -, and the housing unit - type of housing unit, total surface area, number of bedrooms, electricity included in rent. We cleaned and kept data related to houses that are not indicated as vacant, and houses that are rented or owned. We also eliminated data points where one of the variables was missing. This process deleted on average 40% of the data per state. The data contains 38 variables that are noncategorical, such as the year the house was built or the family income from the past 12 months for instance. We get $1,896,284$ data points across all states, with an average of $37,182 \pm 41,313$ points per state.

With the dataset, we want to explore the differences in homeownership between genders between states so that we can reveal potentially unexplored fairness issues. We thus define the "SEX" attribute as sensitive and also add the race variable "HHLDRRAC1P" as an attribute to watch according to intersectional studies [18]. The former encodes gender as a binary variable: "MALE" or "FEMALE". The latter encodes race as follows, as given by the original dataset: "White", "Black or African American", "American Indian", "Alaskan", "American Indian and/or Alaskan Native tribes", "Asian", "Native Hawaiian or Pacific Islander", "Some other race", "Two or more races". Except for "SEX", all categorical variables are one-hot encoded. Noncategorical variables are scaled to follow a normal distribution $\mathcal{N}(0,1)$. To detect said differences, we train a model on data from a single state and use data from another state as serving data. The model is trained to predict the variable "TEN" that encodes if a family owns or rents the house in which they live.

We also collected data from the COMPAS dataset, which is commonly used in the fairness community [4]. A private company developed the COMPAS - which stands for Correctional Offender Management Profiling for Alternative Sanctions - recidivism risk scores, and now multiple states of the USA use the COMPAS score as a risk screening system. However, numerous research showed that the recidivism scores were heavily biased against African-American populations [17]. Given its real-life implications, not only is it necessary to counterbalance these biases with fairness tools, but it is also important to understand the impact of models on populations that are discriminated against. We cleaned the dataset from any missing entries which yielded 1881 data points for 15 variables. We use this dataset to predict whether an individual will re-offend given his criminal history. The two most common sensitive attributes for this dataset are the "race" and "sex" ones. The dataset is composed of 49.8% African-Americans, 36% Caucasians, 10.2% Hispanics, 0.5% Asians, 0.5% Native Americans, and 2.9% other ethnicities. This data is split between 81.4% of men and 18.6% of women. We use the "sex" attribute as a sensitive feature for our experiments but also include "race" as an attribute to watch for according to prison studies [35]. We simulate data drift between the training and the serving data by splitting the dataset according to the "race" attribute as follows for the training set: 50% of African-American from the initial dataset, 30% of Hispanic, 50% of Caucasian, 80% of Asian, 80% of other ethnicities. This yields 924 data points for the training and 957 data points for the serving data.

5.1 Results

We report the results of training of both the baseline and fairness model on both datasets in Table 1. On the COMPAS dataset, we observe that the optimization of both DP and EO leads to a small increase of both metrics compared to the baseline that already has a high DP. However, optimizing only for EO leads to a small decrease in the score. Both changes are statistically non-significant according to a Kruskal-Wallis test, $p > .05$. This result shows that gender-only issues are not that important, indicating that the variable "race" should also be considered a fully sensitive attribute for which fairness should be optimized too, as suggested by feminist literature and intersectional studies [35]. As for the US Census dataset, we observe that the fairness model improves its accuracy, while maintaining a similar EO score, $p < 0.05$ according to a Kruskal-Wallis test. Similarly here, this shows that other variables should have been included for the fairness optimization in the choice of sensitive attributes, such as "race" or possibly "age" [26, 28]. Given these results, we analyze how extracted rules of detected unfairness-prone regions on the serving data evolve before and after retraining for each dataset. According to our formulization and previous results, fairness issues are entailed by data drift (RQ1).

US Census. For both the baseline and the fairness model, we first looked for US states that had a housing market that was significantly different from others by analyzing which state had significant harmful data drift with other states. Five states emerged from this cross-comparison: California, Texas, New York, Pennsylvania, and Illinois. For the following, we consider California as the training data and use all other states as individual serving datasets.

We observe that some rules that relate to "Correct" predictions and include "SEX" or "HHLDRRAC1P" are based on many samples. During testing, we counted $7,206 \pm 13,282$ samples, and $10,440 \pm 17,747$ rules after retraining. There is a great variability in the number of rules. However, in general, there are very few rules relating to false negatives or positives that include a large number of samples, which means that harmful data drift relates to small regions in the feature space. This indicates that fairness issues are very specific for each state. Most comparisons between California and other states lead to few rules that are based on a fairly small amount of samples approximately 10-, with a few exceptions. Compared to Florida, the harmful data drift detection yields a small region of unfairness that is based on only 3 samples. After retraining to mitigate this, the fairness detection creates more rules that total a little more than 600 samples. As for the comparison with Hawaii, there are 1000 samples per rule on average for false predictions before and after retraining. Here is an example of a rule for false positives based on 395 samples: African-American families who live in a mobile home and make more than 1000$ in agricultural sales a year, and who perceive more than the average public assistance income. This example is not found again after retraining. This extracted rule could be interpreted either as a positive or negative outcome, depending on whether the state wants to encourage home ownership for disadvantaged populations or if we think about their ability to reimburse their mortgage. Depending on the socio-economic framework of a state, the latter could have substantial repercussions, similar to what happened during the 2008 real estate crises [32]. If we look at data drift with Idaho, we can find a similar example of a rule

Dataset	Metric	Random Forest	Fairness at training	Fairness after retraining
US Census	Accuracy	0.798 ± 0.093	0.821 ± 0.092	0.821 ± 0.092
	Equal Opportunity	0.988 ± 0.091	0.989 ± 0.008	0.990 ± 0.008
COMPAS (Exp. 1)	Accuracy	0.65 ± 0.07	0.68 ± 0.04	0.67 ± 0.03
	Equal Opportunity	0.74 ± 0.22	0.70 ± 0.21	0.71 ± 0.20
COMPAS (Exp. 2)	Accuracy	0.65 ± 0.07	0.68 ± 0.07	0.68 ± 0.06
	Demographic Parity	0.90 ± 0.06	0.94 ± 0.05	0.92 ± 0.05
	Equal Opportunity	0.74 ± 0.22	0.76 ± 0.12	0.75 ± 0.11

Table 1: Comparison of accuracy and fairness metrics across different model configurations. Results are shown for both the US Census and COMPAS datasets. For the Fairness model, we report metrics on the test set before retraining ("step 1") and after retraining ("step 3"). Equal Opportunity and Demographic Parity are used as fairness metrics. Results are reported as mean ± standard deviation across multiple runs.

for false positives based on 249 samples: Retired white American households with a significantly below average retirement income as well as social security income who make more than 1000$ in yearly agricultural sales, that do not have to pay for water. This example is not found again after retraining however. We found a similar example when comparing with Iowa for Non-African-American families that rely on public assistance income to live. This again shows that such a model can have a significant impact on populations and that their deployment should be engineered and deployed with caution, and several multidisciplinary feedback loops.

There is greater variability in the number of samples that DDLA rules are based on after retraining for a quite stable amount of rules: there are 4.9 ± 3.5 rules for 13.3 ± 9.5 samples during testing on the serving data, and 4.4 ± 4.8 rules after retraining that are based on 608.1 ± 1491.6 samples. It is interesting to note that close to all false negative rules disappear after retraining which is consistent with the fact that our model is optimizing equal opportunity, i.e. reducing the amount of false negatives for both genders so they reach a similar rate. However, this comes with the cost of rules for false positives that are based on many more samples.

COMPAS. For the rules, we observe a small number of rules based on a few samples that are extracted by our pipeline containing the "sex" variable. At training, there are 10 ± 4 rules that total 22 ± 12 samples, while after retraining there are 7 ± 5 rules that are based on a total of 16±12 samples. This confirms the fact that "sex" alone does not present much unfairness in the dataset as previously analyzed, but the inclusion of both race and sex leads to a better understanding of how each population is specifically being discriminated against [35].

Accuracy-fairness trade-off. The fairness literature generally states that accuracy and fairness metrics are in tension and that a bit of one has to be compromised for the other to reach a higher value [40]. However, this is not necessarily mathematically true and also introduces a techno-solutionism narrative that accuracy is more desirable than fairness since this position posits that a model should sacrifice accuracy to be fairer [14][38]. We argue that the existence of this trade-off depends on the problem's framing of accuracy and fairness.

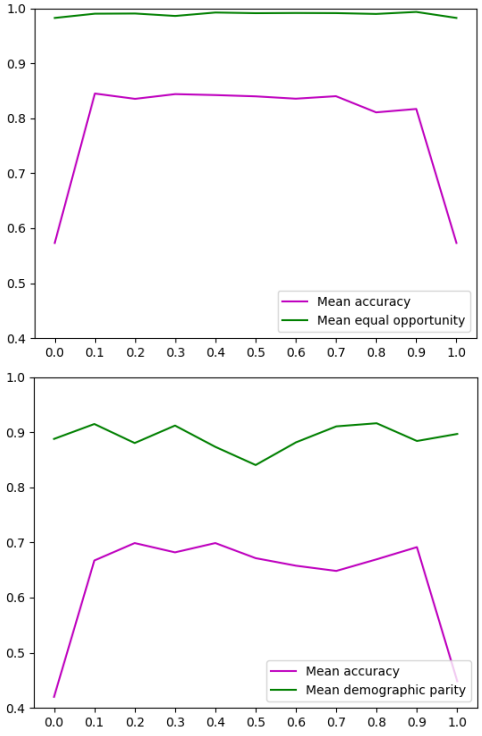

Figure 3: Mean accuracy and equal opportunity according to the weight given to accuracy during training computed on serving data before retraining. *Top:* Trade-off computed with the US Census dataset on the following seven states as serving data: Texas, Florida, New York, Pennsylvania, Illinois, Ohio, and Georgia. *Bottom:* Trade-off computed with the COMPAS dataset.

Here, we use a model that is trained with multiple objectives. We optimize accuracy as well as one or more fairness metrics. To better investigate whether our model operates a trade-off between all metrics or if optimizing another metric also improves the initial

objective, we analyze the variations of accuracy and equal opportunity when changing the weights assigned to each loss function. Thus we have :

$$\mathcal{L} = w * \mathcal{L}_{acc} + (1 - w) * \mathcal{L}_{eo} \quad (1)$$

where $\mathcal{L}$, $\mathcal{L}_{acc}$, $\mathcal{L}_{eo}$ respectively represent the total training loss of the model, the loss to optimize accuracy - i.e. binary cross-entropy -, the loss that optimizes equal opportunity, and w is the weight. We train a model on data from California and use the following seven states as serving data: Texas, Florida, New York, Pennsylvania, Illinois, Ohio, and Georgia. We compute accuracy and EO at step 1 before any retraining at step 3 to better capture the model's capacity to generalize. We run experiments with $w \in [0, 1]$ with incremental steps of 0.1. We report the results in Figure 3. We executed the same study with the COMPAS dataset, this time with DP instead of EO. The results are reported in Figure 3. We observe the same pattern previously seen with the US Census dataset. There is no apparent trade-off in any of the two cases between accuracy and the fairness metric when the sensitive attribute is sex.

5.2 Socio-technical analysis

Most studies solely focus on the computational aspect of fairness issues. Authors generally check the validity of their approach through ML-oriented metrics, which can reinforce a techno-solutionism narrative. We argue that a socio-technical analysis of results provides significant feedback for understanding, as well as checking for potential ethical issues. Here, our analysis showed that California had fairness issues that were significantly different from other states. We discuss this by examining California's real estate history to better understand how their market became specific, and how that may impact comparisons with other states.

California is currently known to face a housing crisis with more than 180,000 people who are not housed according to the federal government, and a fourth of renters who are severely rent-burdened, meaning that they spend more than half their earnings on rent [13]. This crisis can be explained by multiple political decisions. California favored the construction of single-family housing units, even in dense regions such as large cities [30]. In the 1970s, 55% of households owned their homes [9]. Buildings with multiple units also had to have enough parking lots for all residents, which made available space in dense areas even scarcer. However, these decisions faced significant challenges due to the intense waves of African-American immigration in the 1970s, which put the housing market under significant pressure [22]. House prices, including those of affordable ones, soared as they reached ten times their initial price from 1979 in 2019 [3]. Since then, the federal government passed many decrees to counteract these effects, for instance, by allowing housing units that were originally built for single families to be split into two or more units. However, due to the political stratification of California enforced by the Housing Accountability Act, these decrees did not yet produce the expected outcomes [37]. To this day, California still remains the US state with the highest housing cost. Due to this significant increase in housing prices, most current owners, who come from white American families, have houses that were passed on by their parents [34]. This creates an unequal market where minorities - specifically African Americans and Latinx Americans - are unable to afford a house or rent in a way that is financially sustainable for them. This state's history created a housing market that is very different from other states' markets, which needs to be taken into account when validating a tool such as ours.

6 Conclusion

We proposed a socio-technical pipeline that guarantees fairness under data drift by identifying low-accuracy regions tied to sensitive attributes such as race or gender and classifying shifts as harmful only when they exacerbate bias. Our framework retrains the model only when harmful drift occurs with a data augmentation approach to re-balance the representation in the newly identified sensitive regions, rather than retraining on every change. Experiments on the U.S. Census ACS and COMPAS datasets confirm our pipeline's effectiveness in maintaining equitable outcomes, underscoring the value of uniting technical and social perspectives to navigate evolving data distributions [42]. Throughout the pipeline, we integrated social science insights to ensure that every decision at each step is well-informed and aligns with real-world contexts. Such insight informs choices from the fairness metrics to the analysis of unfairness-prone regions to interpret latent biases that cannot be inferred solely from algorithmic metrics. There still remain a few research avenues worth investigating. We could broaden our approach by incorporating other fairness metrics simultaneously, while ensuring that the combinations are relevant, and extend data augmentation to other strategies for a more flexible mitigation step. We only addressed covariate shift in this paper, but the pipeline would benefit from also handling concept drift where the relationship between features and labels evolves. This would entail incorporating more regular social science feedback loops to ensure an alignment with shifting societal expectations and regulatory standards.

References

[1] Samuel Ackerman, Eitan Farchi, Orna Raz, Marcel Zalmanovici, and Parijat Dube. 2020. Detection of data drift and outliers affecting machine learning model performance over time. *arXiv preprint arXiv:2012.09258* (2020).

[2] Samuel Ackerman, Orna Raz, Marcel Zalmanovici, and Aviad Zlotnick. 2021. Automatically detecting data drift in machine learning classifiers. *arXiv preprint arXiv:2111.05672* (2021).

[3] U.S. Federal Housing Finance Agency. 2024. All-Transactions House Price Index for California [CASTHPI]. retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/CASTHPI, January 02, 2025.

[4] J. Angwin, J. Larson, S. Mattu, , and L. Kirchner. 2016. Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks.

[5] Lucas Baier, Josua Reimold, and Niklas Kühl. 2020. Handling Concept Drift for Predictions in Business Process Mining. *2020 IEEE 22nd Conference on Business Informatics (CBI)* 1 (2020), 76–83.

[6] C. Malik Boykin, Sophia T. Dasch, Vincent Rice Jr., Venkat R. Lakshminarayanan, Taiwo A. Togun, and Sarah M. Brown. 2021. Opportunities for a More Interdisciplinary Approach to Measuring Perceptions of Fairness in Machine Learning. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21)*. Association for Computing Machinery, Article 1.

[7] Philipp Brandt. 2022. Sociology's Stake in Data Science. *Sociologica* 16, 2 (Oct. 2022), 149–166. doi:10.6092/issn.1971-8853/13434

[8] Philipp Brandt. 2023. Machine Learning, Abduction, and Computational Ethnography. (2023).

[9] US Census Bureau. [n. d.]. Homeownership Rates. Retrieved December 12, 2024.

[10] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. 2022. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports* 12, 1 (2022), 4209.

[11] Simon Caton and Christian Haas. 2024. Fairness in Machine Learning: A Survey. *ACM Comput. Surv.* 56, 7, Article 166 (April 2024), 38 pages.

[12] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. 2019. Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 319–328. doi:10.1145/3287560.3287586

[13] Ben Christopher and Manuela Tobias. 2024. Californians: Here's why your housing costs are so high. https://calmatters.org/explainers/california-housing-costs-explainer/.

[14] A Feder Cooper, Ellen Abrams, and Na Na. 2021. Emergent unfairness in algorithmic fairness-accuracy trade-off research. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 46–54.

[15] Oscar Blessed Deho, Michael Bewong, Selasi Kwashie, Jiuyong Li, Jixue Liu, Lin Liu, and Srecko Joksimovic. 2024. Is it Still Fair? A Comparative Evaluation of Fairness Algorithms through the Lens of Covariate Drift. arXiv:2409.12428 [cs.LG] https://arxiv.org/abs/2409.12428

[16] Sijie Dong, Qitong Wang, Soror Sahri, Themis Palpanas, and Divesh Srivastava. 2024. Efficiently mitigating the impact of data drift on machine learning pipelines. *Proceedings of the VLDB Endowment* 17, 11 (2024), 3072–3081.

[17] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4, 1 (2018), eaao5580.

[18] Douglas N Evans, Kwan-Lamar Blount-Hill, and Michelle A Cubellis. 2019. Examining housing discrimination across race, gender and felony history. *Housing Studies* 34, 5 (2019), 761–778.

[19] João Gama, Indre Zliobaite, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. *ACM Comput. Surv.* 46, 4 (2014), 44:1–44:37.

[20] Pratyush Garg, John Villasenor, and Virginia Foggo. 2020. Fairness metrics: A comparative analysis. In *2020 IEEE international conference on big data (Big Data)*. IEEE, 3662–3666.

[21] 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.

[22] Elizabeth M. Grieco. 2014. The "Second Great Wave" of Immigration: Growth of the Foreign-Born Population Since 1970. https://www.census.gov/newsroom/blogs/random-samplings/2014/02/the-second-great-wave-of-immigration-growth-of-the-foreign-born-population-since-1970.html Retrieved January 02, 2025..

[23] Shreyas Havaldar, Jatin Chauhan, Karthikeyan Shanmugam, Jay Nandy, and Aravindan Raghuveer. 2024. Fairness under Covariate Shift: Improving Fairness-Accuracy Tradeoff with Few Unlabeled Test Samples. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 11 (Mar. 2024), 12331–12339. doi:10.1609/aaai.v38i11.29124

[24] Léo Hemamou, Arthur Guillon, Jean-Claude Martin, and Chloé Clavel. 2021. Don't judge me by my face: An indirect adversarial approach to remove sensitive information from multimodal neural representation in asynchronous job video interviews. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–8.

[25] Th Kirat, Olivia Tambou, Virginie Do, and Alexis Tsoukiás. 2023. Fairness and Explainability in Automatic Decision-Making Systems. A challenge for computer science and law. *EURO journal on decision processes* 11 (2023), 100036.

[26] Meghan Kuebler and Jacob S. Rugh. 2013. New evidence on racial and ethnic disparities in homeownership in the United States from 2001 to 2010. *Social Science Research* 42, 5 (2013), 1357–1374. doi:10.1016/j.ssresearch.2013.06.004

[27] Ankur Mallick, Kevin Hsieh, Behnaz Arzani, and Gauri Joshi. 2022. Matchmaker: Data Drift Mitigation in Machine Learning for Large-Scale Systems. *Proceedings of Machine Learning and Systems* 4 (2022), 77–94.

[28] Brian J McCabe. 2018. Why buy a home? Race, ethnicity, and homeownership preferences in the United States. *Sociology of Race and Ethnicity* 4, 4 (2018), 452–472.

[29] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6 (2021).

[30] Stephen Menendian, Shahan Shahid Nawaz, and Samir Gambhir. 2024. Single-Family Zoning in California: A Statewide Analysis. https://belonging.berkeley.edu/single-family-zoning-california-statewide-analysis.

[31] Ayesha Nadeem, Babak Abedin, and Olivera Marjanovic. 2020. Gender bias in AI: A review of contributing factors and mitigating strategies. (2020).

[32] David L. Olson. 2016. The real estate crash of 2007-8 as a systemic failure. *Human Systems Management* 35, 4 (2016), 267–277.

[33] Kirtan Padh, Diego Antognini, Emma Lejal-Glaude, Boi Faltings, and Claudiu Musat. 2021. Addressing fairness in classification with a model-agnostic multi-objective algorithm. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (Proceedings of Machine Learning Research, Vol. 161)*, Cassio de Campos and Marloes H. Maathuis (Eds.). PMLR, 600–609. https://proceedings.mlr.press/v161/padh21a.html

[34] Shane Phillips, Carolina Reid, and Dana Cuff. 2022. *Housing And Community Development In California: An In-Depth Analysis of the Facts, Origins and Trends of Housing and Community Development in California.* Technical Report. UCLA, Lewis Center for Regional Policy Studies.

[35] Gwenola Ricordeau. 2019. *Pour elles toutes.* Lux éditeur.

[36] UNESCO. 2020. *Artificial intelligence and gender equality: key findings of UNESCO's Global Dialogue.* Technical Report. UNESCO's Global Dialogue.

[37] Natalia Vega Varela and Nancy L. Cohen. 2024. The Origins of California's Housing Crisis: Fifty Years of Rising Rental Housing Costs and Their Unequal Impacts on Californians. *Gender Equity Policy Institute* (2024).

[38] Xiaomeng Wang, Yishi Zhang, and Ruilin Zhu. 2022. A brief review on algorithmic fairness. *Management System Engineering* 1, 1 (2022), 7.

[39] Geoffrey I Webb, Loong Kuan Lee, Bart Goethals, and François Petitjean. 2018. Analyzing concept drift and shift from sample data. *Data Mining and Knowledge Discovery* 32 (2018), 1179–1199.

[40] Michael Wick, Jean-Baptiste Tristan, et al. 2019. Unlocking fairness: a trade-off revisited. *Advances in neural information processing systems* 32 (2019).

[41] Ke Yang and Alexandra Meliou. 2023. Non-Invasive Fairness in Learning through the Lens of Data Drift. (3 2023). http://arxiv.org/abs/2303.17566

[42] Ke Yang and Alexandra Meliou. 2024. Non-invasive fairness in learning through the lens of data drift. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 2164–2178.

[43] Hantian Zhang, Xu Chu, Abolfazl Asudeh, and Shamkant B. Navathe. 2021. OmniFair: A Declarative System for Model-Agnostic Group Fairness in Machine Learning. In *Proceedings of the 2021 International Conference on Management of Data* (Virtual Event, China) *(SIGMOD '21)*. Association for Computing Machinery, New York, NY, USA, 2076–2088. doi:10.1145/3448016.3452787

[44] Wenbin Zhang and Albert Bifet. 2020. FEAT: A Fairness-Enhancing and Concept-Adapting Decision Tree Classifier. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12323 LNAI, 175–189. doi:10.1007/978-3-030-61527-7_12