FairCanary: Rapid Continuous Explainable Fairness

Avijit Ghosh*
ghosh.a@northeastern.edu
Northeastern University
Boston, MA, USA

Aalok Shanbhag*
aalokshanbhag@gmail.com
Snap Inc.
Mountain View, CA, USA

Christo Wilson
cbw@ccs.neu.edu
Northeastern University
Boston, MA, USA

ABSTRACT

Systems that offer *continuous model monitoring* have emerged in response to (1) well-documented failures of deployed Machine Learning (ML) and Artificial Intelligence (AI) models and (2) new regulatory requirements impacting these models. Existing monitoring systems continuously track the performance of deployed ML models and compute feature importance (a.k.a. *explanations*) for each prediction to help developers identify the root causes of emergent model performance problems.

We present Quantile Demographic Drift (QDD), a novel model bias quantification metric that uses quantile binning to measure differences in the overall prediction distributions over subgroups. QDD is ideal for continuous monitoring scenarios, does not suffer from the statistical limitations of conventional threshold-based bias metrics, and does not require outcome labels (which may not be available at runtime). We incorporate QDD into a continuous model monitoring system, called FairCanary, that reuses existing explanations computed for each individual prediction to quickly compute explanations for the QDD bias metrics. This optimization makes FairCanary an order of magnitude faster than previous work that has tried to generate feature-level bias explanations.

CCS CONCEPTS

• **Information systems** → **Data analytics**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

fairness, drift, model explanation, continuous measurement

1 INTRODUCTION

Machine Learning (ML) and Artificial Intelligence (AI) models that are deployed into the field cannot guarantee consistent performance

over time [54]. One of the reasons for this might be that the underlying data has changed stochastically. This phenomenon, called *drift*, has been well-studied in the literature, from sudden [48] to gradual drifts [60]. Drifts may also be caused by true shifts in the relationship between the underlying variables (e.g., due to changes in the population over time), sampling issues [53], or even bugs that impact downstream data collection.

In scenarios where a deployed ML model is making sensitive decisions, we argue that analyzing the impact of drift on the *fairness* of the model is equally, if not more, important than assessing the impact of drift on traditional performance metrics like accuracy and recall. Regulators are also concerned about this issue: for example, the European Commission's recently proposed Artificial Intelligence Act states *"[ML] providers should be able to process…special categories of personal data, as a matter of substantial public interest, in order to ensure the bias monitoring, detection and correction in relation to high-risk AI systems"* as part of a *"robust post-market monitoring system."* [11] Similar regulations have been proposed in New Zealand [34], Canada [49], the US [1], and the UK [20].

The recognition that drift can negatively impact model performance, coupled with looming regulations, has spurred the creation of many commercial systems that offer *continuous model monitoring* [14]. In general, these systems track live model predictions over time, alert the operator if performance metrics change substantively, and compute feature importance (a.k.a. *explanations*) for each prediction using methods like LIME [52] or the Shapley Value [16, 40, 43].[1] Some of these monitoring systems incorporate fairness metrics in addition to traditional performance metrics [47].

In this paper, we present a novel model bias quantification metric called Quantile Demographic Disparity (QDD) that uses quantile binning to measure differences in the overall prediction distributions over subgroups. Because QDD is measured over continuous distributions it does not require developers to choose specific (and often ad hoc) thresholds for measuring fairness, unlike most conventional fairness metrics (see Table 1).[2] Additionally, QDD does not require outcome labels, which may not be available at runtime. We incorporate QDD into FairCanary, a continuous model monitoring system that offers significant advantages versus state-of-the-art commercial systems that help ensure model fairness over time. In particular, FairCanary reuses explanations computed for each individual prediction to quickly compute explanations for the QDD bias metrics. This optimization makes FairCanary an order

---

---

[1]At a high-level, LIME and SHAP are *surrogate models* that are trained alongside a target model and predict how changes to feature values will impact the target model's output. High-impact features are likely to be very important to the target model.
[2]We provide evidence for why ad hoc bias detection thresholds may miss unfairness in § 2.3 and § 4.

Metric/Framework	Related Terms	CO?	E?
Demographic parity [18]	mean difference, demographic parity, disparate treatment	✗	✗
Conditional statistical parity [13]	statistical parity, conditional procedure accuracy, disparate treatment	✗	✗
Equalized odds [27]	equalized odds, false positive/negative parity, disparate treatment	✗	✗
Equal opportunity [27]	equality of opportunity, individual fairness, disparate treatment	✗	✗
Counterfactual fairness [8, 37]	counterfactual fairness, disparate treatment, fliptest	✗	✗
Statistical independence [26]	HGR coefficient, independence	✓	✗
Distributional difference [44]	KL divergence, JS Divergence, Wasserstein distance	✓	✓

Table 1: Summary showing whether conventional classes of fairness metrics support Continuous Output (CO) and feature-level Explanations (E). Metric families are inspired by Mehrabi et al. [42] and the related terminology is from Das et al. [15].

of magnitude faster than previous work that has tried to generate feature-level bias explanations [44].

The rest of the paper is structured as follows. In § 2 we present an overview of *concept drift* in ML, existing work on continuous fairness monitoring, and ML fairness approaches. Next, in § 3, we introduce our system, FairCanary, present an overview of its operation and capabilities, formally define our QDD metric, and discuss how to obtain explanations for it by reusing existing prediction attributions. In § 4 we present a synthetic case study that highlights FairCanary's capabilities, and conclude in § 5.

2 BACKGROUND

We begin by reviewing the concept of drift and the problems it causes for ML models. We highlight how continuous model monitoring can be used to identify and mitigate the problems caused by drift, but also shortcomings of existing monitoring systems. Finally, we present an overview of ML fairness metrics from the literature and discuss how their shortcomings motivated us to develop novel metrics for this project.

2.1 Drift in Machine Learning

Even though a ML model may pass quality control in terms of performance during training, once deployed on live data the model may encounter issues over time that degrade or destroy its performance [55]. One of the main issues that can arise in deployment is *drift*, which is caused by divergence between the data and context under which the model was trained, and the real-world context into which it is deployed. *Data drift* occurs when the runtime data is significantly different from the training data, by virtue of the constant changing of real world data [10]. *Concept drift*, in contrast, occurs when the relationship between the model output and the feature variables change [48, 53, 60].

Scholars have noted that model performance issues caused by drift extend to questions of algorithmic fairness [2], i.e., the removal of unfair and unjustified biases from ML and AI systems. For example, a temporal analysis by Liu et al. [39] showed how the changing of fairness metrics over time, due to data drift, concept drift, or otherwise, could actually harm sensitive groups.

The most popular methods for detecting concept drift [6, 22] assume that the labels for the predicted variable are immediately available. This may not be feasible in practice, however, especially if the labels correspond to sensitive features of human beings.

Furthermore, even if labels are immediately available, concept drift may have rendered them unreliable, thus defeating the purpose of using them to detect concept drift. Given these issues, prior work [17, 24, 50, 65] has measured the drift of prediction distributions as a proxy for concept drift.

FairCanary also uses prediction distribution drift to measure temporal unfairness. Instead of measuring the drift of the production prediction distribution against training prediction distribution, like in [17, 24, 50, 65], we measure the shift in the prediction distributions between different protected groups. If the prediction distributions for two protected groups start diverging over time, that is an indication of unfairness.

The primary mitigation against drift is retraining models on fresher data. Retraining may be expensive, however, so determining when to retrain models is crucial: retraining too frequently wastes (potentially substantial) resources [4], while waiting too long runs the risk of performance degradation.

2.2 Model Monitoring and Explanations

Continuous model monitoring systems are designed to help developers ensure that deployed models perform as expected over time in the face of problems like drift. A number of commercial tools are available that offer model monitoring [14]. In general, these systems offer the following features:

- Continuously record model inputs and model predictions.
- Measure and report traditional performance metrics over time, like precision, recall, and accuracy. Some systems also measure bias/fairness metrics.
- Calculate and record feature-level explanations using techniques like LIME [52] or SHAP [16, 40, 43], which are useful for post-mortem analysis if problems are observed.
- Generate alarms if particular metrics fall below an operator-specified threshold.

Continuous model monitoring systems are useful for uncovering a variety of issues with models at deployment time, including issues caused by drift. Once the developer has identified an issue they can apply mitigations, such as model retraining.

While virtually all of the commercially available monitoring systems explain predictions in terms of constituent features, none of them (to the best of our knowledge) offer explanations for measures of unfairness. We argue that it is equally important to understand which particular input features are responsible

(a) Probability Density Plot   (b) Cumulative Distribution Plot

Figure 1: Probability distribution plots for two hypothetical demographic groups. As demonstrated by the CDF plot on the right, at a threshold of $x = 10$ the positive prediction probability for both groups is about 0.95, thereby satisfying Demographic Parity $[P(Y^*)|D_1 = P(Y^*)|D_2]$, but this is misleading: the Wasserstein distance is nonzero since the two distributions have markedly different shapes. In contrast, moving the threshold to $x = 8$ immediately disadvantages one group, since the positive prediction probability for group 1 falls to 0.6 while for group 2 it only falls to 0.9, thereby violating Demographic Parity.

for causing unfairness to the model over time, especially given the "right to explanation" that is increasingly being enshrined in regulation [56].

Unfortunately, the interpretation of fairness metrics in terms of the input features to the model has not been studied extensively so far. Explaining conventional fairness metrics (see Table 1) that rely on ground truth labels using Shapley values is possible by making the assumption that the perturbed values retain the original output label. This approach can be misleading, however, because the perturbations change the nature of the instance, and can even create Out-of-Distribution (OOD) points [35]. Another approach, proposed by Shanbhag et al. [58], explains differentiable distance metrics using integrated gradients, but this technique only applies to differentiable models, which limits its practical applications. Finally, Miroshnikov et al. [44] developed methods to explain the Wasserstein-1 distance using a Shapley value formulation. However, this approach also suffers from practical challenges: (1) it requires that explanations be computed for every possible pair of protected groups, and (2) it is computationally challenging to compute Shapley values over large samples.

FairCanary is closely related to the work by Miroshnikov et al. [44], with a couple of key differences. While Miroshnikov et al. [44] calculated fairness explanations from scratch using Shapley-based methods, for FairCanary we assume that a system that continuously generates prediction explanations, like the systems in [14], are already available. FairCanary sits on top of such a system and reuses these existing prediction explanations to generate fairness explanations in linear time (§ 3.3).

2.3 Shortcomings of Conventional Fairness Metrics

Several conceptual definitions of fairness have been discussed in the literature that, according to Corbett-Davies and Goel [12], fall into three general classes: (1) *anti-classification*, where protected features and their proxies are not used to make decisions, (2) *classification parity*, where measures of model predictive performance are equal across protected groups, and (3) *calibration*, where the outcomes, conditional on priors, are independent of protected features. Corbett-Davies and Goel dissect fairness metrics that implement these definitions, claiming that they have "deep statistical limitations" [12], with several metrics at odds with one another.

Table 1 shows an overview of the terminology and limitations of different classes of fairness metrics in the literature. We refer to the first five frameworks (demographic parity, conditional statistical parity, equalized odds, equal opportunity, and counterfactual fairness) as "conventional" fairness metrics because of their prevalence in algorithmic fairness literature [42] and in the industry [3]. The last two classes, statistical independence and distributional difference, are relatively niche and new to the discussion.

Conventional fairness metrics have impossibility results [46]. Prior work [12, 33, 44] points out that it is impossible to satisfy both *classification parity* and *calibration* metrics at the same time in general, and therefore context becomes key when picking a metric [2, 57].

These statistical limitations extend to group membership limitations. Conventional fairness metrics require groups and subgroups to be discrete variables and cannot work with continuous

Figure 2: A diagram illustrating how FairCanary monitors the inputs and outputs of a trained model over time, identifies bias, alerts the developer, and assists in mitigation. See § 3.1 for further details.

variables [25]. Similarly, "confusion matrix based-metrics" [46] do not support continuous outputs (which is often the case in problems like regression and recommendation). This limitation necessitates that practitioners choose thresholds for determining if the given fairness metric has been violated, but the process for choosing these thresholds is ad hoc and may lead to wildly different conclusions about the fairness of a model. We show an example of this phenomenon in Figure 1.

The fairness metric we developed for FairCanary is called Quantile Demographic Drift (QDD). It is a quantile-based optimized version of the Wasserstein-1 distance metric [62]. It falls under the distributional difference family (see Table 1) and thus lends itself to continuous measurement and explainability. We describe in detail the advantages of QDD over other metrics in § 3.2.

3 FAIRCANARY

We now describe FairCanary, our system for performing continuous model monitoring. First, we present the context in which FairCanary is designed to operate and describe its operations at a highlevel. Next, we discuss how FairCanary measures bias and introduce our novel Quantile Demographic Disparity (QDD) metric. Finally, we describe how FairCanary provides explanations that attribute observed biases to specific features, and the bias mitigation options provided by FairCanary.

3.1 Overview

FairCanary is a system for performing continuous model monitoring. It is designed to be deployed into production environments alongside a trained ML model to help the developers monitor the model's performance over time in terms of traditional and fairness performance metrics. In this paper, our focus will be on the latter, fairness metrics.

The developer of the model must configure FairCanary, a priori, by defining the (intersectional) groups for which unfairness will be monitored, identifying the feature(s) in the dataset that encode group membership, establishing base rate statistics for these groups

(i.e., as ascertained from the model's underlying training data), and setting thresholds to trigger bias alerts.

Figure 2 illustrates FairCanary's mode of operation and some of its key capabilities. (**a**) As new data arrives it is fed into the trained model, which (**b**) produces predictions that are stored by FairCanary. Over time, FairCanary maintains a record of the predictions for each group at an operator-specified time granularity.

(**c**) Periodically, FairCanary computes the fairness metric (QDD, see § 3.2) for the model and alerts the developers if any group performs below the preconfigured threshold. FairCanary provides explanations along with alerts that inform the developer which feature(s) are attributable to the issue (see § 3.3). (**d**) Subsequently, the developer may mitigate the emergent unfairness using tools provided by FairCanary (see § 3.4), which (**e**) should return the model to a state where predictions are fair across groups.

3.2 Quantile Demographic Disparity

In this section, we describe a new metric to measure bias in the predictions of a ML model, at both the group and individual level. The prediction tasks covered by our metric include any single dimensional output, such as regression output, or the output of any particular class in a multi-class classification model.

Our metric, Quantile Demographc Disparity (QDD), falls within the distributional difference family of fairness metrics (see Table 1). We argue that there are two reasons for assessing the fairness of an ML model by comparing its prediction distributions over the groups of interest, versus focusing on post-threshold outcomes. The *first* reason is to ensure that we measure bias across the whole spectrum of classified individuals, as opposed to focusing solely on the individuals that are above the threshold of selection, or on group-level approximations. *Second*, as groups of interest get smaller, they reveal more information about intra-group disparities that would have otherwise been lost due to aggregation [25], all the way down to groups of one, i.e., individuals. This helps remove aggregation bias from the bias measurement itself.

3.2.1 **Desired Properties of a Bias Metric**. We now discuss desirable properties of a distributional fairness metric that fit our stated objectives:

(1) The metric should be in the units of the model's prediction scores. The utility of this is especially evident when dealing with continuous output models. This is desirable because it provides insight into the extent of the problem, before human intervention is applied, such as deciding and applying a threshold.

(2) The metric should take the value zero only if the prediction distributions being compared are exactly the same. The benefit of this is that, when taken along with the first property, it gives the ML practitioner a mental scale to understand the extent of the bias.

(3) The metric should be continuous with respect to changes in the geometry of the distribution [44]. This ensures that any distributional change is captured.

(4) The metric should be non-invariant with respect to monotone transformations of the distributions [44]. For example, given two samples of points S_1 and S_2, if we multiply the value of each point in the samples by a constant k, the distance between the modified samples should now depend on k. Jensen-Shannon Divergence (JSD) [38], for example, does not satisfy this property.

(5) The metric should be bias-transforming as described in [63], i.e., the metric should not be satisfied by a model that preserves the biases present in the data.

QDD satisfies all of these properties when the number of bins is equal to the number of samples. The choice of number of bins can be adjusted to satisfy these properties.

3.2.2 **Formalization**. We now describe our QDD metric, which is a function of the quantile bin that a prediction event lies in. QDD is a novel formulation of the Wasserstein-1 distance metric [62], and thus it is designed to work for continuous outputs and can be customized to provide sliced views down to the individual-level.

For two groups G_1 and G_2, let the two distributional samples of model scores be S_1 and S_2. We divide the samples into B bins of equal size N_1 and N_2, respectively. This is equivalent to segmenting by quantiles. For example, if there are 10 bins, we are essentially bucketing individuals between the 0th–10th percentile, 10th–20th percentile, and so on.

We define QDD for bin b as

$$QDD_b = \mathbb{E}_{G_{1,b}}[S_1] - \mathbb{E}_{G_{2,b}}[S_2]. \tag{1}$$

This can be approximated as

$$QDD_b = \frac{1}{N_1} \sum_{n=1}^{N_1} S_{1,n} - \frac{1}{N_2} \sum_{n=1}^{N_2} S_{2,n}. \tag{2}$$

The QDD, when conditioned on certain attributes C, becomes the Conditional Quantile Demographic Disparity.

To demonstrate the flexibility of QDD, we demonstrate how it can be used to measure three different conceptualizations of bias.

(1) Intra-Group Bias is defined as the maximum QDD across the b bins of a given group of individuals. This quantity is useful to combat aggregation bias within groups.

(2) Disparity with Base Rate is defined as the difference between the QDD calculated over the production data and the QDD of the training data. This quantity is most relevant when the training data is representative of the population the model is expected to encounter during deployment.

(3) Individual Fairness via Alignment. QDD is defined between two groups over a given number of bins, which determines the resolution of the metric. If the number of bins is equal to the number of instances in the sample, QDD becomes a comparison between individuals at the same rank or percentile. This is equivalent to the concept of *alignment* proposed by Shanbhag et al. [58].

Computing QDD over individual instances gives us a clean way to obtain individual fairness insights, with the counterfactual example being the same ranked counterpart in the opposite group. This method does not require us to compute complex counterfactuals, which could have their own biases and errors [32]. The principle we use to justify this insight is as follows: if there is no bias between two groups, and we have a large enough sample of both, then the distance between individuals of the same rank in the prediction space should be zero.

3.3 Explanation

Explainability of ML systems that are deployed in production is a very important part of the practice of responsible AI [5, 41]. This especially applies to models that are contributing to decisions that can impact peoples' lives. Such decisions cannot be inscrutable, and thus the internal workings employed by the ML models must be human-verified to be logical and normatively justifiable.

FairCanary incorporates two state-of-the-art methods for explaining the output of predictions in terms of specific features: Shapley value-based methods [40] and (if the model being monitored is differentiable) Integrated Gradients (IG) [61]. We adopted these methods because they satisfy the desirable axiom of *efficiency* [40], which helps provide a precise accounting of bias.[3]

Just like explanations for individual predictions, we argue that it is vital to be able to explain measures of fairness or bias, so that the features that are responsible for the bias can be identified. FairCanary incorporates a novel method for explaining the feature importance contributing to QDD that we call Local Quantile Demographic Disparity attribution.

The Local QDD attribution for feature f, for prediction sample S_1 over S_2 in bin b, $QDDA_{b,A,f}$ is a measure of the change in QDD in bin b that can be attributed to (a.k.a. explained by) feature f using attribution method A that satisfies the efficiency axiom. r denotes the reference and t denotes the target distribution. We define

$$QDDA_{b,A,f} = \frac{1}{N_t} \sum_{n=1}^{N_t} \text{attr}_{n,A,S_1,f} - \frac{1}{N_r} \sum_{n=1}^{N_r} \text{attr}_{n,A,S_2,f} \tag{3}$$

where $\text{attr}_{n,A,S_i,f}$ refers to the attribution of the nth data point to feature f for a prediction from bin b of distribution S_i using attribution method A. Given that the attribution method A satisfies the efficiency axiom, $QDD_b = \sum_{f=1}^{F} QDDA_{b,A,f}$.

---

[3]Although there are other explanation methods that satisfy efficiency [7, 59] we do not explore them in this work.

Proof: Since the attribution method A satisfies efficiency, for each instance in the sample S_1 and S_2, $\sum_{f=1}^{F}$ attributions$_f$ = prediction − baseline prediction.

For the same baseline,

$$\frac{1}{N_t} \sum_{n=1}^{N_t} \sum_{f=1}^{F} \text{attr}_{n,A,S_1,f} - \frac{1}{N_r} \sum_{n=1}^{N_r} \sum_{f=1}^{F} \text{attr}_{n,A,S_2,f} =$$

$$\frac{1}{N_1} \sum_{n=1}^{N_2} S_{1,n} - \frac{1}{N_2} \sum_{n=1}^{N_2} S_{2,n}$$

$$\therefore \text{QDD}_b = \sum_{f=1}^{F} \text{QDDA}_{b,A,f}.$$

Explaining bias in this manner enables a single attribution to be used for multiple explanations across groups. In contrast, Shapley values over a particular metric must be re-calculated for every grouping. Our explanation technique therefore is much more computationally efficient than previous techniques [44] since it requires the calculation of attributions only once. To elaborate, Shapley values without approximation are exponential in the number of model features. While there exist approximation techniques, the complexity is worse than linear time. Hence, Wasserstein Shapley computation (n points times d features) is worse than calculating Shapley values for n points separately for d features. Additionally, we can re-use the Shapley values computed for a data point when calculating QDD between any combination of protected features, whereas the whole computation needs to be repeated for each combination in the case of Wasserstein Shapley.

3.4 Mitigation

Mitigation is a key outcome of monitoring bias, enabling corrective action to be taken. FairCanary provides an option for developers to automatically mitigate bias revealed by our QDD metric using a quantile norming approach. In essence, this approach replaces the score of the disadvantaged group with the score of the corresponding rank in the advantaged group, similar to the mitigations proposed in [29, 45]. The justification for quantile norming is that if (1) bias is known to exist, (2) bias is the only rational explanation for disparity, and (3) bias is assumed to be equal within the disadvantaged group, then normalizing across ranks is normatively justifiable.

In essence, quantile norming is a post-processing mitigation. The advantages of post-processing mitigations as opposed to pre-training debiasing are discussed by Geyik et al. [23]. Additionally, quantile norming is a relatively computationally inexpensive approach to bias elimination.

We note that all bias mitigation approaches, including quantile norming, should only be adopted in practice after conducting a thorough examining of their consequences on the outputs of a model. Corbett-Davies et al. [13] demonstrate several cases where mitigation may cause additional harm to individuals or to particular groups. Developers that adopt FairCanary are under no obligation to use quantile norming for mitigation, and are free to adopt other, perhaps more thorough and computationally expensive, approaches (e.g., model retraining [54], data preprocessing [30], etc.) that better suit their needs.

Feature	Values	Distribution
Location	{'Springfield', 'Centerville'}	70:30
Education	{'GRAD', 'POST_GRAD'}	80:20
Engineer Type	{'Software', 'Hardware'}	85:15
Experience (Years)	(0, 50)	Normal Distribution
Relevant Experience (Years)	(0, 50)	Normal Distribution
Gender	{'MAN', 'WOMAN'}	50:50

Table 2: Features, values, and their distributions used in our synthetic case study. Note that the gender feature is only used for measuring and mitigating bias, it is not used for model training or prediction.

4 CASE STUDY

In this section, we present an example of FairCanary in action via a case study on a synthetic dataset. This allows us to inject controlled drifts into the data stream to demonstrate how FairCanary, via QDD, can detect and explain the resulting bias. Additionally, we present comparisons to conventional fairness metrics.

4.1 Scenario

In this case study, we posit a scenario where a developer has trained a model to predict the starting salary of job seekers based on relevant features of their resume, such as education level and years of experience (see Table 2). Note that the output of this model is continuous. Additionally, the developer designed the model to be fair with respect to the binary gender of job seekers, i.e., the distribution of salaries predicted for men and women should be nearly identical. We assume that the model was audited and found to be fair relative to the data that was available at training time.

Let us assume that the model has learned the following relationship to predict an individual's salary from the features in Table 2:

$$\text{Salary} = 50,000 + (20,000 \times \text{location}) + (20,000$$
$$\times \text{education}) + (5,000 \times \text{relevant\_experience})$$
$$+ (100 \times \text{experience}) + (10,000 \times \text{engineer\_type})$$

In our scenario, the developer deploys this model into production along with FairCanary to continuously monitor its output. We generate 20,000 synthetic job seekers' data per day for three days that are fed into the model, using feature values drawn from the distributions given in Table 2 (with the added constraint that experience ≥ relevant experience). On Day One and Day Three we generate all of the candidate data correctly, but crucially, on Day Two, we simulate a data engineering bug that erroneously labels all women as 'GRAD' instead of 'POST_GRAD' regardless of their true educational attainment. This reduces the estimated salary for all women post-graduates by $20,000 on Day Two.

We argue that the scenario we have outlined here is realistic. ML-based resume screening and analysis tools are widely available, and given that they gate access to employment opportunities, it is crucial that these systems be fair [9, 51]. The bug we intentionally simulate on Day Two could easily occur in practice, e.g., due to the temporary malfunctioning of a resume parser that prepares data for the salary prediction model.

(a) Prediction distribution on Day One   (b) Prediction distribution on Day Two   (c) Prediction distribution on Day Three

(d) Continuous plot of the QDD metric over time. There is a clear dip on the second day.

(e) QDD Explanations for Day One   (f) QDD Explanations for Day Two   (g) QDD Explanations for Day Three

Figure 3: Plots for our case study showing how FairCanary would detect and explain the bias against women on Day Two on a continuously running salary prediction model. The explanations for Day Two clearly indicate Education as the feature responsible for the bias, which enables the practitioner to correct the data integrity issue and fix the biased predictions.

4.2 Analysis

Figure 3 shows how FairCanary would detect and explain the fairness problem that occurs on Day Two. The model outputs on Day One show that the prediction distributions for men and women are mostly aligned (Figure 3a), thereby being fair. On the second day (Figure 3b), due to the data integrity error discussed above, the prediction distributions differ. When we examine the running plot for QDD[4] (Figure 3d), we notice a sharp dip on Day Two—QDD goes from an average value of $156 on Day One to

-$8677 on Day Two—indicating a bias against women.[5] Note that the absolute value of QDD goes up, indicating an increase in bias, and would trigger the alarm system like in Figure 2. Similarly, the feature explanations (generated here using Integrated Gradients) go from being distributed among the different features on Day One (Figure 3e) to assigning the majority of blame to the education feature on Day Two (Figure 3f).

FairCanary would alert the model developer of the problem on Day Two, and its explanations could help the developer perform

---

[4]For simplicity, we set the number of quantile bins as 1 for the case study. Thus, the explanations are for the entire distribution and not any particular quantile bin.

[5]Recall in § 3.2.1 we say that one useful feature of QDD is that the metric has the same units as the predicted output. Having the QDD value in dollars clearly helps users to understand the extent of bias and thereby aids usability.

Threshold	Day One		Day Two	
	SPD	DI	SPD	DI
$50000	0.00009	1.00009	-0.00556	0.99439
$100000	0.00911	1.01749	-0.08290	0.84569
$200000	0.00088	1.02876	-0.01049	0.65544

Table 3: The performance of two conventional fairness metrics, Statistical Parity Difference (SPD) and Disparate Impact (DI), against different salary thresholds for the case study. The predictions on Day One were fair, while they were unfair to women on Day Two. Only one metric catches the bias, and only at one threshold (highlighted in red).

root-cause analysis of the bias issue. Based on this information, the developer could then identify and correct the underlying data engineering bug. Once corrected, we observe that the model's predictions are again aligned for men and women on Day Three (Figure 3c), and the QDD values have returned to their expected range (Figure 3d).

To further illustrate the utility of QDD we compare it with two conventional fairness metrics—Statistical Parity Difference (SPD)[6], and Disparate Impact (DI)[7]—to see if a monitoring system using these metrics would have caught the bias against women on the second day.

Table 3 shows the values of the two conventional metrics for different salary thresholds (i.e., for the positive outcome) on Day One and Day Two. We configure the alert threshold for both metrics[8] in accordance with the US UGESP 4/5[ths] rule [19] that is commonly used in disparate impact analysis [64]. Alarmingly, we observe that, as configured here, SPD would not catch the bias on the second day at all, and DI would only catch it at one threshold level.

5 DISCUSSION

In this work we present a novel metric called QDD that improves on conventional fairness metrics by not requiring prediction labels or threshold values (§ 3.2). We utilize this metric in FairCanary, a system for performing continuous monitoring of deployed ML models. FairCanary includes all of the typical capabilities of ML monitoring systems [14]: it records inputs to and outputs from the model over time, calculates traditional measures of model performance (e.g., accuracy), allows operators to set configurable alerts if model performance changes dramatically, and calculates explanations for individual predictions using existing techniques [40, 61].

Additionally, FairCanary is able to provide explanations for QDD by reusing the explanations for individual predictions, which is (1) a capability not offered by conventional fairness metrics and (2) less

---

[6]Statistical or Demographic Parity Difference is the difference in the positive outcome rate between the privileged and unprivileged group. SPD = $\Pr(\hat{y} = 1|p = 1)$ - $\Pr(\hat{y} = 1|p = 0)$.
[7]Disparate Impact is the ratio of the passing rate of the the privileged and unprivileged group. DI = $\frac{Pr(\hat{y}=1|p=1)}{Pr(\hat{y}=1|p=0)}$.
[8]For Statistical Parity Difference, since there is no conventionally accepted value, we set the threshold to 20% to be consistent with Disparate Impact.

computationally demanding than similar approaches from prior work [44] (§ 3.3).

Through examples (Figure 1) and a synthetic case study (§ 4), we demonstrate the functionality of FairCanary and the useful properties afforded by our QDD metric. We publicly release the code [9] used to generate the plots in our case study.

5.1 Limitations and Future Work

While threshold independence is one of the strengths of QDD, it is also a potential weakness: without ground truth labels, calculated disparities are, at the end of the day, best-case approximations of the discrimination that actually takes place in society. We therefore do not advocate for the elimination of conventional fairness metrics that require ground truth labels and thresholds, but instead propose using them in conjunction with QDD to obtain a fuller picture of real life harms in a context dependent manner [21].

We used Integrated Gradients as the explanation method for our case study. However, the choice of explanation method is potentially important, as recent research [28, 36] shows that different explanation methods often do not produce the same results, and ensembling them is superior than using any one of them in isolation.

Finally, FairCanary/QDD is not completely automated: there are still manual parameters that need to be set, like number of bins and alert sensitivity. Providing FairCanary users with guidance on how to tune the system for their use case and context will be crucial for real use cases. Additionally, all fairness monitoring systems should consider providing actionable recourse tips [31] through explanations to end-users via a carefully designed, accessible interface.

5.2 Broader Impact

Regardless of whether ML models are regulated to mandate audits and continuous monitoring, we argue that ML practitioners have a professional and moral obligation to ensure that the systems they deploy do not misbehave. Given that issues like drift are known to occur, and that these issues may cause unfairness and bias, we argue that monitoring systems should become a standard component of most, if not all, deployed ML-based systems.

We hope that FairCanary (or other monitoring systems that incorporate its capabilities) will equip companies and institutions with improved tools to monitor, understand, and mitigate problems in their deployed ML systems, in real time. In turn, we hope that these capabilities will bring more equity and justice to the individual stakeholders impacted by deployed models.

---

[9]https://github.com/fiddler-labs/faircanary

REFERENCES

[1] 116th Congress (2019-2020). [n. d.]. H.R.2231 - Algorithmic Accountability Act of 2019. https://www.congress.gov/bill/116th-congress/house-bill/2231.

[2] Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, Duncan Wadsworth, and Hanna Wallach. 2021. Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs. *arXiv preprint arXiv:2103.06076* (2021).

[3] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. 2018. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943* (2018).

[4] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.*

[5] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency.* 648–657.

[6] Albert Bifet and Ricard Gavalda. 2007. Learning from time-changing data with adaptive windowing. In *Proceedings of the 2007 SIAM international conference on data mining.* SIAM, 443–448.

[7] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. 2016. Layer-wise relevance propagation for neural networks with local renormalization layers. In *International Conference on Artificial Neural Networks.* Springer, 63–71.

[8] Emily Black, Samuel Yeom, and Matt Fredrikson. 2020. Fliptest: fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* 111–121.

[9] Miranda Bogen and Aaron Rieke. 2018. Help wanted: An examination of hiring algorithms, equity, and bias. (2018).

[10] Eric Breck, Neoklis Polyzotis, Sudip Roy, Steven Whang, and Martin Zinkevich. 2019. Data Validation for Machine Learning.. In *MLSys.*

[11] European Commission. [n. d.]. Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence.

[12] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[13] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining.* 797–806.

[14] Jakub Czakon. 2022. Best Tools to Do ML Model Monitoring. (2022). https://neptune.ai/blog/ml-model-monitoring-best-tools

[15] Sanjiv Das, Michele Donini, Jason Gelman, Kevin Haas, Mila Hardt, Jared Katzman, Krishnaram Kenthapadi, Pedro Larroy, Pinar Yilmaz, and Muhammad Bilal Zafar. 2021. Fairness Measures for Machine Learning in Finance. *The Journal of Financial Data Science* 3, 4 (2021), 33–64.

[16] Anupam Datta, Shayak Sen, and Yair Zick. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP).* IEEE, 598–617.

[17] Denis Moreira dos Reis, Peter Flach, Stan Matwin, and Gustavo Batista. 2016. Fast unsupervised online drift detection using incremental kolmogorov-smirnov test. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 1545–1554.

[18] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference.* 214–226.

[19] Equal Employment Opportunity Commission, Civil Service Commission, et al. 1978. Uniform guidelines on employee selection procedures. *Federal Register* 43, 166 (1978), 38290–38315.

[20] UK Office for Artificial Intelligence. [n. d.]. Ethics, Transparency and Accountability Framework for Automated Decision-Making. https://www.gov.uk/government/publications/ethics-transparency-and-accountability-framework-for-automated-decision-making.

[21] Center for Data Science and Public Policy. [n. d.]. Aequitas: Fairness Tree. http://www.datasciencepublicpolicy.org/projects/aequitas/.

[22] Joao Gama, Raquel Sebastiao, and Pedro Pereira Rodrigues. 2013. On evaluating stream learning algorithms. *Machine learning* 90, 3 (2013), 317–346.

[23] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* 2221–2231.

[24] Sindhu Ghanta, Sriram Subramanian, Lior Khermosh, Swaminathan Sundararaman, Harshil Shah, Yakov Goldberg, Drew S. Roselli, and Nisha

Talagala. 2019. ML Health: Fitness Tracking for Production Models. *CoRR* abs/1902.02808 (2019). arXiv:1902.02808 http://arxiv.org/abs/1902.02808

[25] Avijit Ghosh, Lea Genuit, and Mary Reagan. 2021. Characterizing Intersectional Group Fairness with Worst-Case Comparisons. *arXiv preprint arXiv:2101.01673* (2021).

[26] Vincent Grari, Boris Ruf, Sylvain Lamprier, and Marcin Detyniecki. 2019. Fairness-Aware Neural Réyni Minimization for Continuous Features. *arXiv preprint arXiv:1911.04929* (2019).

[27] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413* (2016).

[28] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems* 32 (2019).

[29] Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. 2020. Wasserstein fair classification. In *Uncertainty in Artificial Intelligence.* PMLR, 862–872.

[30] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems* 33, 1 (2012), 1–33.

[31] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2020. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv preprint arXiv:2010.04050* (2020).

[32] Niki Kilbertus, Philip J Ball, Matt J Kusner, Adrian Weller, and Ricardo Silva. 2020. The sensitivity of counterfactual fairness to unmeasured confounding. In *Uncertainty in artificial intelligence.* PMLR, 616–626.

[33] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[34] Alistair Knott. [n. d.]. Moving Towards Responsible Government Use of AI in New Zealand). https://digitaltechitp.nz/2021/03/22/moving-towards-responsible-government-use-of-ai-in-new-zealand/.

[35] I Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. 2020. Problems with Shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning.* PMLR, 5491–5500.

[36] I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. 2020. Problems with Shapley-value-based explanations as feature importance measures. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119),* Hal Daumé III and Aarti Singh (Eds.). PMLR, 5491–5500. https://proceedings.mlr.press/v119/kumar20e.html

[37] Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *arXiv preprint arXiv:1703.06856* (2017).

[38] Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory* 37, 1 (1991), 145–151.

[39] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed impact of fair machine learning. In *International Conference on Machine Learning.* PMLR, 3150–3158.

[40] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30,* I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[41] Aniek F Markus, Jan A Kors, and Peter R Rijnbeek. 2021. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics* 113 (2021), 103655.

[42] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).

[43] Luke Merrick and Ankur Taly. 2020. The Explanation Game: Explaining Machine Learning Models Using Shapley Values. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction.* Springer, 17–38.

[44] Alexey Miroshnikov, Konstandinos Kotsiopoulos, Ryan Franks, and Arjun Ravi Kannan. 2020. Wasserstein-based fairness interpretability framework for machine learning models. *arXiv preprint arXiv:2011.03156* (2020).

[45] Preetam Nandy, Cyrus Diciccio, Divya Venugopalan, Heloise Logan, Kinjal Basu, and Noureddine El Karoui. 2021. Achieving Fairness via Post-Processing in Web-Scale Recommender Systems. *arXiv preprint arXiv:2006.11350* [stat.ML]

[46] Arvind Narayanan. [n. d.]. 21 fairness definitions and their politics. https://fairmlbook.org/tutorial2.html.

[47] David Nigenda, Zohar Karnin, Muhammad Bilal Zafar, Raghu Ramesha, Alan Tan, Michele Donini, and Krishnaram Kenthapadi. 2021. Amazon SageMaker Model Monitor: A System for Real-Time Insights into Deployed Machine Learning Models. *arXiv preprint arXiv:2111.13657* (2021).

[48] K. Nishida, S. Shimada, S. Ishikawa, and K. Yamauchi. 2008. Detecting sudden concept drift with knowledge of human behavior. In *2008 IEEE International Conference on Systems, Man and Cybernetics.* 3261–3267. https://doi.org/10.1109/

ICSMC.2008.4811799

[49] Government of Canada. [n. d.]. Responsible use of artificial intelligence (AI). https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai.html.

[50] Fábio Pinto, Marco OP Sampaio, and Pedro Bizarro. 2019. Automatic model monitoring for data streams. *arXiv preprint arXiv:1908.04240* (2019).

[51] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAT*)*.

[52] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[53] Marcos Salganicoff. 1997. Tolerating concept and sampling shift in lazy learning using prediction error context switching. In *Lazy learning*. Springer, 133–155.

[54] Sebastian Schelter, Felix Biessmann, Tim Januschowski, David Salinas, Stephan Seufert, and Gyuri Szarvas. 2018. On challenges in machine learning model management. (2018).

[55] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. 2015. Hidden technical debt in machine learning systems. *Advances in neural information processing systems* 28 (2015).

[56] Andrew Selbst and Julia Powles. 2018. "Meaningful Information" and the Right to Explanation. In *Conference on Fairness, Accountability and Transparency*. PMLR, 48–48.

[57] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*. 59–68.

[58] Aalok Shanbhag, Avijit Ghosh, and Josh Rubin. 2021. Unified Shapley Framework to Explain Prediction Drift. *arXiv preprint arXiv:2102.07862* (2021).

[59] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *International conference on machine learning*. PMLR, 3145–3153.

[60] Kenneth O Stanley. 2003. Learning concept drift with a committee of decision trees. *Informe técnico: UT-AI-TR-03-302, Department of Computer Sciences, University of Texas at Austin, USA* (2003).

[61] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*. PMLR, 3319–3328.

[62] Cédric Villani. 2009. The wasserstein distances. In *Optimal transport*. Springer, 93–111.

[63] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2020. Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law. *W. Va. L. Rev.* 123 (2020), 735.

[64] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. 2021. Building and auditing fair algorithms: A case study in candidate screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 666–677.

[65] Indre Žliobaite. 2010. Change with delayed labeling: When is it detectable?. In *2010 IEEE International Conference on Data Mining Workshops*. IEEE, 843–850.