Understanding Unfairness in Fraud Detection through Model and Data Bias Interactions
José Pombal 1,2,3, André F. Cruz 1, João Bravo 1, Pedro Saleiro 1, Mário A.T. Figueiredo 2,3, Pedro Bizarro 1
firstname.lastname@feedzai.com
1 Feedzai, 2 Instituto Superior Técnico, 3 Instituto de Telecomunicações

ABSTRACT
In recent years, machine learning algorithms have become ubiquitous in a multitude of high-stakes decision-making applications. The unparalleled ability of machine learning algorithms to learn patterns from data also enables them to incorporate biases embedded within. A biased model can then make decisions that disproportionately harm certain groups in society — limiting their access to financial services, for example. The awareness of this problem has given rise to the field of Fair ML, which focuses on studying, measuring, and mitigating unfairness in algorithmic prediction, with respect to a set of protected groups (e.g., race or gender). However, the underlying causes for algorithmic unfairness still remain elusive, with researchers divided between blaming either the ML algorithms or the data they are trained on. In this work, we maintain that algorithmic unfairness stems from interactions between models and biases in the data, rather than from isolated contributions of either of them. To this end, we propose a taxonomy to characterize data bias and we study a set of hypotheses regarding the fairness-accuracy trade-offs that fairness-blind ML algorithms exhibit under different data bias settings. On our real-world account-opening fraud use case, we find that each setting entails specific trade-offs, affecting fairness in expected value and variance — the latter often going unnoticed. Moreover, we show how algorithms compare differently in terms of accuracy and fairness, depending on the biases affecting the data. Finally, we note that under specific data bias conditions, simple pre-processing interventions can successfully balance group-wise error rates, while the same techniques fail in more complex settings.

KEYWORDS
Algorithmic Fairness, Data Bias, Machine Learning

1 INTRODUCTION
With the increasing prominence of machine learning in high-stakes decision-making processes, its potential to exacerbate existing social inequities has been a reason of growing concern [3, 26, 37]. Financial services have been no exception, with multiple works in the field warning against potential discrimination [5, 6, 8, 32]. By leveraging complex information from data to make decisions, these models can also learn biases that are encoded within. Using biased patterns to learn to make predictions without accounting for possible underlying prejudices can lead to decisions that disproportionately harm certain social groups. The goal of building systems that incorporate these concerns has given rise to the field of Fair ML, which has grown rapidly in recent years [35].

Fair ML research has focused primarily on devising ways to measure unfairness [4], and to mitigate it in algorithmic prediction tasks [11, 35]. Mitigation is broadly divided in three approaches: pre-processing, in-processing, and post-processing [34], which map respectively to interventions on the training data, on the model optimization, and on the model output. Another focal point of discussion revolves around the underlying sources of algorithmic unfairness. The aforementioned mitigation methods reflect different beliefs with respect to the origins of unfair predictions. Pre-processing assumes that the cause is bias in the data, while in- and post-processing shift the onus to modeling choices and criteria.

Research seems to be divided along the same lines in what concerns uncovering the source of bias in the ML pipeline. On the one hand, there is work defending that bias in the data is at the root of downstream unfairness in predictions [12, 13, 42, 44]. On the other hand, some researchers have adverted to the crucial role that model choices have in algorithmic unfairness [18, 25]. However, the consequences of different sources of bias on unfairness produced by ML algorithms remains unclear. Little attention has been paid to the interaction between biases in the data and model choices. At best, the relationship between the two is recognized but, save a few studies, mostly left unexplored. At worst, one of them is outright disregarded.

We maintain that the two views are complementary, not mutually exclusive. In fact, we aim to add to this discussion by showing that the landscape of algorithmic bias and fairness does change dramatically with the specific bias patterns present in a dataset. Conversely, under the same data bias conditions, different models incur in distinct fairness-accuracy trade-offs. We show this empirically by devising a series of controlled experiments with fairness-blind ML algorithms that map such trade-offs to types of bias present in the data. Each experiment is motivated by a hypothesis about these trade-offs, and some are built to reflect biases that naturally arise in fraud detection, such as the selective labels problem [33], or the fact that certain agents are actively trying to evade fraud. To this end, we propose a taxonomy of different conditions under which a dataset may be considered biased with respect to a protected group.

The experiments are conducted on a large, real-world, bank-account-opening fraud dataset, into which bias is injected through additional synthetic features. The synthetic nature of these additional features does not limit our analysis; rather, by allowing full control of the sources of bias, it lets us draw clear links between generic dataset characteristics and subsequent fairness-accuracy
Figure 1: Fairness-accuracy trade-offs attained by a variety of ML algorithms under distinct group prevalences: group A has 4x the prevalence of group B. Disparities for FPR and FNR move in opposite directions. As suggested by Equation 1, the group with highest prevalence is disproportionately affected by false positives.

trade-offs. Moreover, these synthetic features have a clear grounding in real-world data bias patterns (e.g., different group-wise prevalences, under-represented minorities).

This work has two overarching goals. First, to provide empirical evidence that predictive unfairness stems from the relationship between data bias and model choices, rather than from isolated contributions of either of them. Second, to steer the discussion towards relating algorithmic unfairness to concrete patterns in the data, allowing for more informed, data-driven choices of models and unfairness mitigation methods.

Summing up, we make the following contributions:

- A formal taxonomy to characterize data bias between a protected attribute, other features, and the target variable.
- Experimental results for a comprehensive suite of hypotheses regarding fairness-accuracy trade-offs ML models make under distinct types of data bias, pertinent, but not restricted to fraud detection.
- Showing how, by changing data bias settings, the picture of algorithmic fairness changes, and how comparisons among algorithms differ.
- Raising awareness to the issue of variance in fairness measurements, underlining the importance of employing robust models and metrics.
- Evaluation of the utility of simple unfairness mitigation methods under distinct data bias conditions.

2 RELATED WORK

2.1 Fairness-Fairness Trade-offs

Fairness is often at conflict with itself. It has been shown that when a classifier score is calibrated and group-wise prevalences are different, it is impossible to achieve balance between false positive (FPR) and false negative (FNR) error rates [15, 21, 31]. Corbett-Davies and Goel [16] further discuss these metrics’ trade-offs and point out their statistical limitations. Speicher et al. [41] compare between-group and in-group fairness metrics, showing that solely optimizing for one may harm the other.

It is clear that no single fairness metric is ideal, and that its choice is highly dependent on assumptions made and the problem domain [39]. With this in mind, as motivated in Section 4.3, we will use FPR parity to measure fairness. Decomposing this metric as per Equation 1 allows for a better understanding of the aforementioned trade-offs and of how they result from an interaction between the data and classifier. For two protected groups A and B, let \( p_i \) be the prevalence of group \( i \in \{A, B\} \), and \( PPV_i \), \( FNR_i \) be the precision and false negative rate, respectively, of a classifier on group \( i \). Then, as shown by Chouldechova [15],

\[
\frac{FPR_A}{FPR_B} = \frac{p_A \cdot 1 - PPV_A}{p_B \cdot 1 - PPV_B} \cdot \frac{1 - FNR_A}{1 - FNR_B}. \tag{1}
\]

Notice how FNR parity must be sacrificed in order to guarantee FPR parity, if prevalence \( p_i \) differs between groups but \( PPV_i \) are balanced. Prevalence is only related to the data itself, while the other metrics result from an interaction between the classifier and the dataset. Indeed, for any classifier under different group-wise prevalences, we must sacrifice at least one of: FPR parity, FNR parity, or calibration\(^1\) (PPV parity). Figure 1 illustrates this relation under different group-wise prevalences. Although different algorithms achieve different fairness-accuracy trade-offs, they all follow the same trend: the group with higher prevalence is disproportionately affected by false positives, and subsequently less affected by false negatives.

2.2 Relating Trade-offs and Data Bias

Label Bias. In the criminal justice context, Fogliato et al. [22] assume that their target labels (arrest) are a noisy version of the true outcome they wish to predict (re-offense). They then propose a

\(^1\)A score is deemed calibrated if it reflects the likelihood of the input sample being positively labeled, regardless of the group an individual belongs to.
framework to analyze the impact of this imperfection on protected groups (e.g., race). Wang et al. [44] propose a method to mimic label bias that is particularly interesting to us: they corrupt a portion of the target labels in their training data, where the amount of corrupted labels depends both on the protected group and on the target. Afterwards, they assess the impact of this on downstream unfairness mitigation methods. Most prove to be less effective under this type of bias. In case of account opening fraud, label bias can arise in the form of the selective label problem [33], which will be explained in Section 4.2.5.

**Group-size disparity, prevalence disparity, and relations between protected attribute and other dataset features.** As part of a larger suite of experiments, Blanziesky and Cunningham [7] study the impact of prevalence disparities on underestimation, which is the ratio between a group’s probability of being predicted positive, and the probability of being labelled positive. They test several fairness-blind algorithms on a fully synthetic dataset. The main finding is that the smaller the number of minority group observations, the stronger underestimation becomes. We build on this work by using a larger dataset, experimenting with more bias conditions and models, and evaluating them with popular metrics in the Fair ML community.

Akpinar et al. [2] study the effects on observational unfairness metrics (e.g.: demographic parity, TPR parity, etc...) of training models on several types of data bias. They propose a sandbox tool to allow practitioners to inject bias in datasets, so as to run controlled experiments and evaluate the robustness of their systems. Our work is similar in the bias injection process, but its overarching goal is somewhat different. We focus on formalizing the data bias conditions, and conducting a thorough analysis of the fairness-accuracy trade-offs different algorithms exhibit when exposed to bias.

Finally, we draw inspiration from Reddy et al. [38], who study the impact of several data bias conditions on a large suite of deep learning unfairness mitigation methods. The authors create a synthetic variant of the MNIST dataset [19] (CI-MNIST), where they emulate and test the impact of group-size disparities, correlations between the target and the sensitive attribute (essentially prevalence disparity), and correlations between non-sensitive features and the target. The UCI Adult dataset, a real dataset, is also experimented on. However, the authors outline the importance of synthetic data, by stating that it is not possible to fully emulate some bias conditions on real data. While we make use of a real dataset, we augment it synthetically for this reason. The key takeaway is that the landscape of algorithmic fairness changes drastically under more extreme bias scenarios.

### 3 BIAS TAXONOMY

Throughout this work, we refer to a dataset’s feature set as $X$, the class label as $Y$ and the protected attribute as $Z$. A generic value taken by any of these is represented as its lowercase letter. It is important to stress that the following definitions use the inequality sign ($\neq$) to mean a statistically significant difference.

Despite the multitude of definitions, there is still little consensus on how to measure data bias or its impact on the predictive performance and fairness of algorithms [35]. In this paper, we propose a broad definition: there is bias in the data with respect to the protected attribute, whenever the random variables $Y$ and $X$ are sufficiently statistically dependent from $Z$.

**Bias Condition 0** (Protected attribute bias).

$$P[X,Y] \neq P[X,Y|Z]$$

(2)

For Condition 0 to be satisfied, the distribution of $Z$ must be statistically related to either $X$, $Y$, or both. If $Y$ is directly dependent on $Z$ or indirectly through $X$, algorithms may use $Z$ to predict $Y$.

We will study the effect of three specific bias conditions (or types). The following conditions all imply Condition 0.

**Bias Condition 1** (Prevalence disparity).

$$P[Y] \neq P[Y|Z]$$

(3)

i.e., the class probability depends on the protected group. For example, if we consider residence as $Z$ and crime rate as $Y$, certain parts of a city have higher crime rates than others.

**Bias Condition 2** (Group-wise distinct class-conditional distribution).

$$P[X|Y] \neq P[X|Y,Z]$$

(4)

Note that this condition allows for $P[Y] = P[Y|Z]$ (no prevalence disparity). Consider an example in account opening fraud in online banking. Assume that the fraud detection algorithm receives a feature which represents how likely the submitted e-mail is to be fake ($X$) and the client’s reported age ($Z$) as inputs. In account opening fraud, fraudsters tend to impersonate older people, as these have a larger line of credit to max out, but use fake e-mail addresses to create accounts. Therefore, the e-mail address feature will be better to identify fraud instances for reportedly older people, potentially generating a disparity in group-wise error-rates, even if age groups have an equal likelihood of committing fraud in general. Figure 2 provides a visual example using generic features.

**Bias Condition 3** (Noisy Labels). The noisy labels condition is

$$P^*[Y|X,Z] \neq P[Y|X,Z]$$

(5)

where $P^*$ is the observed distribution and $P$ is the true distribution. This is, some observations belonging to a protected group have been incorrectly labeled.

Inaccurate labeling is a problem for supervised learning in general. It is common for one protected group to suffer more from this
ailment, if the labeling process is somehow biased. For example, women and lower-income individuals tend to receive less accurate cancer diagnoses than men, due to sampling differences in medical trials [20]. In account-opening fraud, label bias may arise due to the selective label problem. Work on the impact of this bias on downstream prediction tasks is discussed in Section 2.

We will also study the effect of the following bias extensions. An extension is a property that does not imply Condition 0, but has consequences on algorithmic fairness.

**Bias Extension 1** (Group size disparities). Let $z$ be a particular group from a given protected attribute $Z$, and $N$ the number of possible groups. Under group size disparities, we have

$$P \{Z = z \} \neq \frac{1}{N}. \quad (6)$$

Intuitively, this represents different group-wise frequencies. A typical example is religion: in many countries, there tends to be a dominant religious group and a few smaller ones.

**Bias Extension 2** (Train-test disparities). Let $BC_i$ be a set of bias conditions $BC$ on a dataset $i$. Then, under train-test disparities:

$$BC_{\text{train}} \neq BC_{\text{test}}. \quad (7)$$

In supervised learning, it is assumed that the train and test data are independent and identically distributed (i.i.d.). It is crucial that the training set follows a representative distribution of the real world, so that models generalize well to unseen data. The test set is the practitioner’s proxy for unseen data, and concept drift may greatly affect subsequent model performance and fairness. In fraud detection this can be particularly important, if we consider that fraudsters are constantly adapting to avoid being caught. As such, a trend of fraud learned during training can easily become obsolete when models are ran in production.

## 4 METHODOLOGY

### 4.1 Dataset

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

Fraud rate (positive label prevalence) is about 1% in both sets. This means that a naïve classifier that labels all observations as *not fraud* achieves a test set accuracy of almost (99%). Such large class imbalance entails certain additional challenges for learning [24], and calls for a specific evaluation framework that is presented in Section 4.3.

### 4.2 Experimental Setup

#### 4.2.1 Overview

Each experiment in this paper is based on injecting a unique set of bias conditions $BC$ (defined in Section 7?) into a base dataset $D$ and analyzing subsequent fairness-performance trade-offs made by supervised learning models.

We append a set $S$ of synthetically generated columns to the data, in such a way that each condition $BC_i \in BC$ is satisfied. In all cases, the protected attribute $Z$ under analysis is part of $S$, allowing us to control how the data is biased with respect to $Z$. This way, we further our understanding of how a given bias type affects downstream fairness and performance. $Z$ can take values A or B.

For any given set of bias conditions, we repeat the above process 10 times, yielding 10 distinct datasets, which differ in the synthetic columns. This makes our conclusions more robust to the variance in the column generation process.

Models are trained on all 10 seeds for each $BC$ set. Results are obtained for both fairness-aware and unaware models. In the former, models have access to $Z$ in the training process, in the latter, they do not. Section 4.3 will provide further details on this.

#### 4.2.2 Hypothesis H1: Group size disparities alone do not threaten fairness

We would like to assess whether a protected attribute that is uncorrelated with the rest of the data can lead to downstream algorithmic unfairness. In particular, the goal is to compare the case where the two groups are of the same size, with that in which there is a majority and a minority protected group.

We append a single column to the base dataset: a protected attribute that takes value A or B (groups) for each sample. This feature is generated such that $P \{Z = A\} = s_A \neq s_B$, but is independent of features $X$ and target $Y$. This can be achieved by having each row of the new column take the value of a (biased, if $s_A \neq s_B$) coin flip.

Note how, according to our taxonomy in Section 3, this generative process for $Z$ satisfies Bias Extension 1. However, since it is a simple coin flip, it does not satisfy Bias Condition 0. As such, $Z$ remains unequivocally unbiased towards both $X$ and $Y$, on average.

Our Baseline will be for $P \{Z = A\} = 0.5$, when group sizes are equal, and so no Bias Condition or Bias Extension is satisfied — a
We expect this to facilitate detecting fraud for group B, thereby we randomly flip the training labels of negative examples belonging without group size disparities (Bias Extension 1). As such, the above $P[Y|Z=A] = 2*P[Y|Z=B]$. Afterwards, we randomly flip the training labels of group A positive observations, until there are no longer disparities in prevalence: $P^*|Y|Z=A = P|Y|Z=B$.

There are several works that propose more complex label massaging procedures to mitigate unfairness in a dataset [27, 28]. In this context, our method may be seen as a naive approach to achieve balanced prevalence via label flipping.

### 4.3 Evaluation

#### 4.3.1 Fairness metrics. The real-world setting in which these models would be employed — online bank account-opening fraud detection — motivates the choice of fairness and performance evaluation metrics adopted in this work.

In account-opening fraud, a malicious actor attempts to open a new bank account using a stolen or synthetic identity (or both), in order to quickly max out its line of credit [1]. A false positive (FP) is a legitimate individual who was wrongly flagged as fraudulent, and wrongly blocked from opening a bank account. Conversely, a false negative (FN) is a fraudulent individual that was able to successfully open a bank account by impersonating someone else, leading to financial losses for the bank.

We must ensure that automated customer screening systems do not disproportionately affect certain protected sub-groups of the population, directly or indirectly. Fairness w.r.t. the label positives is measured as the ratio between group-wise false negative rates (FNR). Equalizing FNR is equivalent to the well-known equality of opportunity metric [23], which dictates equal true positive rates (TPR), $TPR = 1 - FNR$. In our setting, this ensures that a proportionately equal amount of fraud is being caught for each sub-group. On the other hand, fairness w.r.t. the label negatives is measured as the ratio between group-wise false positive rates (FPR). Within our case-study, equalizing FPR (also known as predictive equality [17]) ensures no sub-group is being disproportionately denied access to banking services.

#### 4.3.2 Performance metrics. Bank account providers are not willing to go above a certain level of FPR, because each false positive may lead to customer attrition (unhappy clients who may wish to leave the bank). At an enterprise-wide level, this may represent losses that outweigh the gains of detecting fraud. The goal is then to maximize the detection of fraudulent applicants (high global true positive rate, TPR), while maintaining low customer attrition (low global false positive rate). As such, we evaluate the model’s TPR at a fixed FPR, imposed as a business requirement in our case-study. We assess the FPR ceiling of 5%. A more typical metric such as accuracy would not be informative, since it is trivial to obtain 99% accuracy by classifying all observations as not fraud (recall that fraud rate is around 1%).

### 4.4 Algorithms and models

We test 6 different ML algorithms: XGBoost (XGB) [14], LightGBM (LGBM) [29], Logistic Regression (LR) [43], Decision Tree (DT) [10], Random Forest (RF) [9], and Feed Forward Neural Network (MLP) trained with the Adam optimizer [30]. The first two are gradient
boosted tree methods, which have stood out as top performers for tabular data in recent years [40]. The other four are popular supervised learning algorithms, used in a variety of applications. All the above algorithms are fairness-blind, in the sense that they do not consider fairness constraints in their optimization process. This choice is intentional: we wish to analyze fairness-accuracy tradeoffs under different kinds of bias in the data, before fairness is taken into consideration. Still, we evaluate the models’ predictions when they have access to the protected attribute at training time (awareness), and when they do not (unawareness). The idea is to assess which types of data bias still lead to predictive unfairness, even when the algorithm is oblivious of the sensitive attribute.

Lastly, hyperparameter choice greatly influences performance and fairness [18]. As such, for each algorithm, we randomly sample 50 hyperparameter configurations from a grid space to be used in all experiments.

5 RESULTS

We summarize our findings in the following sections. In each, we discuss the key takeaways of an hypothesis, and detail the interplay of fairness metrics. We also present a series of plots, highlighting relevant phenomena.

Figure 4 shows results for H1, outlining how sample variance can harm algorithmic fairness, even when models are expected to be fair. Figure 6 shows how different algorithms fared in terms of performance and both fairness metrics, on each hypothesis. Figure 5 deep dives on the LGBM algorithm, to show how Precision plays a part in error-rate disparities, depending on the bias afflicting the data.

On all Figures, the y-axis represents a ratio of group error rates $\frac{FPR_A}{FPR_B}$ or $\frac{FNR_A}{FNR_B}$. As such, it will be in a log$_2$ scale, which allows points to be laid out symmetrically$^2$. The two red dashed lines are at the log$_2$ of 0.8 and 1.25, following the ”80% rule”, used by the US Equal Employment Opportunity Commission [36]. That is, a group’s error rate should be at least 80% of the other groups’ rates to be considered fair.

The plots exhibit the top performing model configuration, in terms of TPR, for each of the 10 dataset seeds. This information is summarized in error-bars, whose center is the median performance and both fairness metrics, on each hypothesis. The error bars may be coloured by algorithm, or by hypothesis, depending on the context. The idea is to focus on models which would be chosen for production in the ‘world’ of each dataset seed – that is, the top performers. Thus, their fairness, or lack thereof, is particularly relevant to the practitioner.

5.1 H1: Group size disparities do not threaten fairness.

5.1.1 Key Takeaways. Models are fair in expectation. On average, if there are no differences in each group’s data distribution, models will not necessarily discriminate the minority. In fact, large group size disparities lead to high fairness variance, possibly resulting in unfair models for either group (see Figure 4).

5.1.2 Fairness Metrics Interplay. Both predictive equality and equality of opportunity are achieved on average since the target variable $Y$ does not depend on the protected attribute $Z$ in any way.

5.2 H2: Groups with higher fraud rate have higher error rates.

5.2.1 Key Takeaways. The group with higher positive prevalence (in our case, fraud) has higher FPR and lower FNR, if group-wise precision is balanced. Interventions such as unawareness or equalizing prevalence are sufficient to balance error rates.

5.2.2 Fairness Metrics Interplay. In practice, group-wise FPR and FNR move in opposite directions, indicating that the classifier is uncalibrated for this group. Different fairness metrics point to different disadvantaged groups: practitioners must carefully weigh the real-world consequences of a FP and a FN.

5.3 H3: Correlations between fraud, other features, and the sensitive attribute influence fairness.

5.3.1 Key Takeaways. Contrary to H2, FPR and FNR are skewed in the same direction: on the group that has more adept fraudsters, innocent people are systematically flagged as fraudulent more often (higher FPR), and fraudulent individuals evade detection more (higher FNR). Equalizing prevalence is no longer useful (it is equal), and unawareness actually aggravates predictive equality disparities. Random Forests, the most robust algorithm in terms of fairness on other hypotheses, was the most unfair and volatile algorithm on this scenario.

5.3.2 Fairness Metrics Interplay. Since prevalence is constant across groups, it cannot be the source of unfairness. Instead, with models better classifying observations from one group, error rate disparities stem from precision divergences. Models have higher precision on one group, leading to relatively higher error rates for the other. Innocent individuals belonging to the group that is “better” at committing fraud are flagged as fraudulent more often than the other group (higher FPR).

5.4 H4: The selective label problem may have mixed effects on algorithmic fairness.

5.4.1 Key Takeaways. Inaccurate labelling leads to harmful effects if the disadvantaged group’s prevalence is further increased (similar to H2). Inaccurate labelling leads to beneficial effects if the disadvantaged group’s prevalence is decreased (label massaging [28]).

5.4.2 Fairness Metrics Interplay. When group A’s prevalence is artificially increased, together with its reduced precision due to noisy labelling, predictive equality is skewed against group A. On the other hand, when inaccurate labeling is used to artificially equalize group-wise prevalence, models tend to fulfill fairness in both predictive equality and equality of opportunity.
| Algorithm | XGB | LGBM | LR | DT | RF | MLP |
|-----------|-----|------|----|----|----|-----|
| Hypothesis | Baseline H1 H2.1 H3 H4.1 |

| Algorithm | Global TPR | Groupwise PRECISION Ratio (A/B) (log2) |
|-----------|------------|-----------------------------------|
| XGB       | 0.5 0.6 0.7 0.8 0.9 1.0 | 2 1 0 1 2 |
| LGBM      | 0.5 0.6 0.7 0.8 0.9 1.0 | 2 1 0 1 2 |
| LR        | 0.5 0.6 0.7 0.8 0.9 1.0 | 2 1 0 1 2 |
| DT        | 0.5 0.6 0.7 0.8 0.9 1.0 | 2 1 0 1 2 |
| RF        | 0.5 0.6 0.7 0.8 0.9 1.0 | 2 1 0 1 2 |
| MLP       | 0.5 0.6 0.7 0.8 0.9 1.0 | 2 1 0 1 2 |

Figure 4: Results for Hypothesis H1: group size disparities alone do not threaten fairness. Left plot: 50% group A, 50% group B. Middle plot: 90% group A, 10% group B. Right plot: 99% group A, 1% group B. Results obtained for a global threshold of 5% FPR. The center of the cross is the median of each metric, and each bar represents the minimum and maximum in each dimension. Slightly different samples from an unbiased data generation process may still lead to algorithmic unfairness in downstream prediction tasks.

![Deep dive into LGBM group-wise precision ratios.](image)

Figure 5: Deep dive into LGBM group-wise precision ratios. In contrast to H2.1, the median precision ratios for H3 and H4.1 are significantly skewed in favour of group B, meaning that models are better at classifying fraud in this group. In H3, this happens because group B fraud is easier to detect given $x_1$ and $x_2$. In H4.1, some of group A's fraud labels are false, giving models more accurate information to classify observations that belong to group B. Furthermore, in H4.1, A is apparently more fraudulent than B (double the prevalence), contributing to a steeper FPR disparity than in H2.1 (see Figure 6).

6 CONCLUSION

The underlying causes for algorithmic unfairness in prediction tasks remain elusive. With researchers divided between blaming the data or blaming the algorithms, little attention has been heeded to interactions between the two.

Our main contribution to this discussion is a comprehensive analysis on different hypotheses regarding fairness-accuracy trade-offs exhibited by ML algorithms, when subject to different types of data bias with respect to a protected group. The use case of this work is fraud detection, but its conclusions are extensible within and outside the scope of the Financial domain.

We can confidently state that the landscape of algorithmic fairness is a puzzle where both algorithm and data are vital, intertwined pieces, essential for its completion. Our results show how an algorithm that was fair under certain biases in the data, may become unfair in other circumstances. For example, Random Forests were the fairest models when the protected group was directly linked to the target, but became quite unfair once dependencies through other features were introduced. Further exploring these interactions is a relevant avenue for future research on the causes of unfairness.

Crucially, we have brought to the fore the often overlooked dangers of variance, by experimenting on several samples of the same underlying bias settings. This showed how algorithmic fairness is subject to the idiosyncrasies of a dataset, especially when groups have significantly different sizes. A model may be fair on one sample, and drastically unfair on another, even though the generative process for both samples was the same (with differences merely stemming from sampling variance). Research is usually focused on whether a model is fair on average, which understates the importance of building systems that are robust to sample changes.

A useful side product of our study was finding that simple unfairness mitigation methods are enough to balance error rates, under certain bias conditions. We also reinforced the relevance of choosing the appropriate fairness metric by exposing the shortcomings of ratios, and showing how error rate ratios move in opposite directions under group-wise prevalence disparities — a fact that is well-grounded mathematically.

We have proposed a data bias taxonomy, and studied several biases by injecting them synthetically into real data. An interesting avenue for further research would be to develop methods to detect and characterize these bias patterns without prior knowledge.

All in all, by relating data bias to fairness-accuracy trade-offs in downstream prediction tasks, one can make more informed, data-driven decisions with regards to the unfairness mitigation methods employed, and other choices along the Fair ML pipeline. We firmly believe that this path holds the key to a better understanding of algorithmic unfairness, that generalizes well to any domain and application.

In the fraud detection domain, and the financial services industry in general, gaining a better understanding of algorithmic unfairness should be a top priority. This will lead to more effective mitigation, which is a core step towards guaranteeing that all groups in society have equal access to financial services, and thus equality in general.
Figure 6: Median, minimum, and maximum performance and fairness levels for top LGBM, RF, MLP, and LR models on each dataset seed (all at 5% global FPR). Group sizes are always balanced except for H1. At a higher level, this shows how different types of bias yield distinct fairness-accuracy trade-offs. At a lower level, each algorithm exhibits particular trade-offs. For example, contrary to its counterparts, LR shows more balanced FPR rates on H2.1 than on H3. XGB is omitted because results were identical to LGBM. DT is omitted because performance was too low.

Hypotheses Recap - Baseline: Unbiased setting — both group sizes are equal, no bias conditions nor extensions satisfied. H1: group size disparities alone do not threaten fairness (case shown is for group A representing 90% of the dataset). H2.x: Groups with higher fraud prevalence have higher error rates (in H2.1 both train and test sets are biased, and H2.x bars represent the case where A has double the fraud rate of B). H3: Algorithmic unfairness arises when groups leverage features unequally to avoid fraud detection. H4.x: The selective label problem may have mixed effects on algorithmic fairness. (H4.1 studies harmful effects on fairness, and H4.2 beneficial ones).
