Fairness Measures for Machine Learning in Finance

Sanjiv Das [1 2] **Michele Donini** [1] **Jason Gelman** [1] **Kevin Haas** [1] **Mila Hardt** [1] **Jared Katzman** [1]
Krishnaram Kenthapadi [1] **Pedro Larroy** [1] **Pinar Yilmaz** [1] **Bilal Zafar** [1]

Abstract

We present a machine learning pipeline for fairness-aware machine learning (FAML) in finance that encompasses metrics for fairness (and accuracy). Whereas accuracy metrics are well understood and the principal ones used frequently, there is no consensus as to which of several available measures for fairness should be used in a generic manner in the financial services industry. We explore these measures and discuss which ones to focus on, at various stages in the ML pipeline, pre-training and post-training, and we also examine simple bias mitigation approaches. Using a standard dataset we show that the sequencing in our FAML pipeline offers a cogent approach to arriving at a fair and accurate ML model. We discuss the intersection of bias metrics with legal considerations in the US, and the entanglement of explainability and fairness is exemplified in the case study. We discuss possible approaches for training ML models while satisfying constraints imposed from various fairness metrics, and the role of causality in assessing fairness.

1. Introduction

Fairness-aware machine learning (FAML) is a critical need in several areas, such as finance, hiring, criminality assessment, medicine, and college admissions, as shown by several recent high profile algorithmic bias incidents O'Neil (2016). However, establishing fair ML models is not a natural outcome of a prediction framework, as noted in Fazelpour and Lipton (2020). ML models that are trained on data affected by societal biases may learn to exhibit bias. Analyzing bias and explainability of ML models is growing in importance for ML-driven products and services, driven by customer needs, regulatory/legal requirements, and societal expectations. Bias and discrimination have been part of several regulations and legal frameworks including the US Civil Rights Act of 1964, and the European Union's General Data Protection Regulation (GDPR). In addition, the finance industry has witnessed additional regulations such as the Fair Credit Reporting Act (FCRA), Equal Credit Opportunity Act (ECOA), SR-11, and Reg B. Due to these regulations, there are concerns in the finance industry around deploying and maintaining more advanced models into production, because fairness becomes harder to establish with more complex models.

We propose a collection of techniques for measuring bias and mitigating bias on protected characteristics, with a focus on the finance sector. We present a case study of a FAML system pipeline applied to a dataset of loans, and show how to apply bias measurement and mitigation at different stages in the ML pipeline, namely, pre-training and post-training. We discuss practical challenges in applying fairness techniques in the financial services industry, pertaining to the intersection of bias metrics with certain legal considerations, the tension between different fairness notions, and the assessment and choice of dataset for bias measurement.[1]

The rest of this paper proceeds as follows. Section 2 provides the background on algorithmic bias and the need for fairness-aware machine learning techniques in finance. Section 3 outlines various approaches to measuring bias in ML, before training the model and after training as well (in Section 4 we discuss the generalization to cases where the protected characteristic and/or the labels may be non-binary). Note that we will use the terminology "protected characteristic" for the variable on which bias may occur, e.g., gender, and we may also call this the "attribute of interest." The latter terminology is more general in that it suggests that bias is bi-directional, e.g., we do not want a gender imbalance, irrespective of gender, whereas the former terminology tends to be uni-directional, where one class is the protected or disadvantaged one, and the other is advantaged. Alternatively, we may not need to specify that one class is the protected one, as it is perfectly plausible that bias in either direction is undesirable. In Section 5,

[1]Amazon Web Services [2]Santa Clara University, CA, USA. Correspondence to: Sanjiv Das <sanjivda@amazon.com>, Michele Donini <donini@amazon.com>, Krishnaram Kenthapadi <kenthk@amazon.com>, Bilal Zafar <zafamuh@amazon.com>.

Working paper, Palo Alto, California.

[1]None of the metrics and methodologies herein are intended as assurances of legal compliance, but mere algorithms to guide fairness in machine learning. In the end, fairness is an ethical and/or legal question, not an algorithmic one.

we will examine various approaches to mitigating bias in the machine learning pipeline (with an application in Section 6 concerning a dataset of loans). Section 7 discusses why fairness measures are often in opposition to each other, and whether measurement should be undertaken on training datasets or test datasets. These are unresolved issues as of now in the financial services industry. Finally, Section 8 offers closing discussion and possible future directions.

2. Background: Algorithmic Bias and Finance

Fairness in Broader Context: This collection of metrics can be viewed as a component of a broader ecosystem of machine learning dimensions, including Fairness, Accuracy, Causality, Explainability, and Trust. Trust includes Security and Privacy.[2] Measuring bias in data sets and models is gaining attention now. On March 28, 2020, a federal court ruled that research assessing whether online algorithms result in racial, gender, or other discrimination does not violate the Computer Fraud and Abuse Act (CFAA).[3] The topics of fairness and bias in datasets and models are gaining significant attention from industry, academia, and government - for example, fairness and bias is a major topic in the Ethics Guidelines for Trustworthy AI published by the EU Commission's High Level Expert Group on Artificial Intelligence, as well as the OECD Principles on AI.

Algorithmic Bias and Fairness: There is an antecedent literature from the 1950s that relates to statistical discrimination, see Hutchinson and Mitchell (2019) for a survey. This is connected to the more recent literature on FAML (fairness aware machine learning), and several ethical black box models and datasets are available. One approach, as in FairML, determines if restricted characteristics have high feature importance. There are several open source repositories that offer various bias management approaches, such as: fairness, Aequitas, Themis, responsibly, IBM's AI Fairness 360, which have more than fifty metrics, and themis-ML. Other references are the UC Berkeley course CS294 titled "Fairness in Machine Learning" and this NIPS17 tutorial, which is based on the excellent book titled "Fairness and Machine Learning" by Barocas et al. (2019). In short, unfairness (bias) can be measured in different ways. The particular choice of bias metric depends on social and legal considerations and various stakeholders including representatives of the disadvantaged group can weigh in.

Algorithmic Bias in Finance: Mitigation of unforeseen effects of bias can help stave off adverse treatment of protected groups and protect financial companies from the unforeseen effects of ML models. For example, redlining in mortgage lending has been a persistent race issue in the US.[4] Gender bias may be present in small business lending decisions, given that fewer women in the historical record who have taken small business loans leads to models trained on an unbalanced data set, see Alesina et al. (2013); Chen et al. (2017); Brock and Haas (2019). Hiring in the finance industry has also been male-dominated for decades (Iris, 2016) and as hiring is increasingly driven by ML, care will be needed to make sure that any potential biases are mitigated.[5]

Fairness in ML Pipeline: There are three broad approaches to FAML: (i) Methods that try to manage biases in the data used for training, (ii) methods that impose fairness during training, and (iii) methods that mitigate bias post-training. This paper offers a large set of these fairness measures, suggests various approaches to resolve the trade off across different measures, and offers recommendations for end users on the appropriate selection of fairness metrics during the machine learning process. We pay special attention to the implementation of FAML in the finance industry, as there are several areas in which fairness is both an ethical and legal requirement. Our goal in this paper is to narrow the focus to one industry and demonstrate how implementations of fairness might be undertaken as an ongoing ML approach in the financial services industry.

Practical Challenges in FAML: The trouble with implementing FAML is that there are too many notions of fairness and a lack of clarity on the prioritization of these definitions. We hope to catalog most of these measures for the finance domain, connect them to industry terminology and regulation, and resolve some conflicting definitions, but not all. At first glance, it would be comforting to assume that the more measures of fairness we use, the better, so that all types of fairness are imposed in, say, a lending algorithm. However, as we make an algorithm fair on one measure, it may become unfair on another, since the commonly used metrics for fairness often conflict with each other. Berk et al. (2017) present a set of six comprehensive measures of fairness. This paper shows that it is mathematically difficult to mitigate bias all measures. The intuition is that since there are just four numbers in the confusion matrix, and many more metrics, it is very hard to change these four numbers in a way in which all metrics change in the same direction. This behooves the modeler and other stakeholders to choose just a small set of relevant fairness metrics. Many more metrics are presented in the survey by Mehrabi et al. (2019), while a scathing critique of fairness measures is pre-

[2] There is widespread interest in this area, and Amazon is supporting the NSF in providing grants for FAML: https://beta.nsf.gov/science-matters/supporting-foundation-fairness-ai.

[3] See https://www.aclu.org/press-releases/federal-court-rules-big-data-discrimination-studies-do-not-violate-federal-anti.

[4] As are models that measure the likelihood of criminal recidivism, e.g., the COMPAS algorithm. More recently, fairness in the use of AI for job interviews is being regulated.

[5] There are positive signs in this area, cf. Barron's, "100 Most Influential Women in Finance", Leslie Norton, March 9, 2020.

sented in Corbett-Davies and Goel (2018). However, despite these critiques, striving for FAML is still an important goal, especially in the realm of financial services.

Bias may be measured against a "golden truth", i.e., some notion of fairness in the population on the chosen restricted characteristic. For example, bias appears in small business lending datasets because the historical record reflects the fact that women in several preceding decades (and even nowadays) applied for far fewer business loans. As a consequence, ML algorithms may pick up on this attribute as a decider in who should be granted a loan if gender remains as a feature in the dataset. It is important that such distortions are detected and corrected, because class imbalance in the historical record may favor the majority class. In this case, the dataset may be set closer to a golden truth by rebalancing classes on gender, while leaving out gender as a feature during ML training. However, such corrective actions do not always rule out algorithmic bias, which may still creep in to the data set for various reasons such as the use of features that are correlated with gender. Therefore, achieving a "fair" ML algorithm on a given dimension requires both ex ante and ex post model corrections using features and labels as inputs. Even with this, the model algorithm itself may still exacerbate whatever bias remains after various adjustments are made to the features and labels, and additional bias mitigation through model adjustments may be necessary. As these adjustments are undertaken, the accuracy of the model may suffer, tasking the data scientist with a difficult trade-off between accuracy and bias, a very important part of FAML.

3. Bias Metrics

3.1. Sources of Bias

Bias in a model arises in many ways. We provide a six-category taxonomy of sources of bias.

1. Biased labels. This arises from human biases and accumulates in datasets. It is particularly prevalent in public datasets with multiple labelers, like police data, public opinion datasets, etc., see Wauthier and Jordan (2011).

2. Biased features, also known as "curation" bias. Here, bias arises from selecting some features and dropping others and can occur directly or indirectly. For example, in lending, a modeler may choose features that are more likely to disadvantage one group and leave out features that would favor that group. While this may be deliberate, it is also possible to have these be done as part of an unconscious process. O'Neil (2016) gives a great example where her model for why children love eating their vegetables was an outcome of culinary curation, where they seem to eat all their vegetables given no servings of pizza, potatoes, meat, etc.

3. Objective function bias, noted by Menestrel and Wassenhove (2016). One case in which this occurs is when the loss function may be overly focused on outliers and if outliers are of specific types in the dataset, the modeler may inject bias.

4. Homogenization bias, where machines generate the data to train later models, perpetuating bias. In these settings, future outcomes are biased, which create a feedback loop through the models, making future models, decisions, and data even more biased.[6]

5. Active bias. Here the data is simply made up and results in biases in people's inferences, opinions, and decisions. Fake news is the prime example. Such bias may be managed by considering the source, by reading beyond the headline, checking authorship, running fact checks[7], verifying dates, making sure that it is really a fact and not a joke, satire, or comedy. More importantly, carefully consider expert credentials, and carefully check for confirmation bias.[8]

6. Unanticipated machine decisions. Untrammeled machine-learning often arrives at optimal solutions that lack context, which cannot be injected into the model objective or constraints. For example, a ML model that takes in vast amounts of macroeconomic data and aims to minimize deficits may well come up with unintended solutions like super-normal tariffs leading to trade wars. This inadmissible solution arises because the solution is not excluded in any of the model constraints. The model generates untenable answers because it does not have context.

No matter how bias originates, it behooves us to measure it and mitigate it. We consider measurement next.

3.2. Measuring Bias

Measuring bias in ML models is a first step to mitigating bias and regulatory compliance. Each measure of bias corresponds to a different notion of fairness. Even considering simple notions of fairness leads to many different measures applicable in various contexts. Consider fairness with respect to gender or race/ethnicity groups, for example, and, for simplicity, that there are two relevant classes, an advantaged class and a disadvantaged one. In the case of an ML model for lending, we may want small business loans to be issued to equal numbers of both classes. Or, when processing job applicants, we may want to see equal numbers of members of each class hired. However, this approach may assume that equal numbers of both classes apply to these jobs, so we may want to condition on the number that apply.

[6] See Wired.com - 12/2011.
[7] See Factcheck.org.
[8] Facebook suggestions on fake news.

Further, we may want to consider not whether equal numbers apply, but whether we have equal numbers of qualified applicants. Or, we may consider fairness to be an equal acceptance rate of qualified applicants across classes, or, an equal rejection rate of applicants, or both. We may use datasets with different proportions of data on the attributes of interest. This imbalance can conflate the bias measure we pick. Our models may be more accurate in classifying one class than in the other. Thus, we need to choose bias notions and metrics that are valid for the application and the situation under investigation.

We detail several bias metrics in this section. Bias may be measured before training and after training, as well as at inference. Our exposition presents the specific case of binary classification models for simplicity, but we generalize the metrics to multi-category classification models and models where the labels are continuous. We specify the following simple notation to make presentation of the metrics easier.

The goal of binary classification is to predict an unknown binary outcome, y given an observed set of features. For example, in a loan application setting, the outcome could correspond to whether an applicant will pay back the loan or not and the observed set of features could include income, net worth, credit history, and loan amount. Suppose we have labeled training data where each example consists of the values of the observed features and the corresponding label, y (= 0 or 1). The classifier maps the observed features to a predicted label, $\hat{y}$ (= 0 or 1) that we hope agrees with the observed label, y. Suppose that there is a restricted feature (which we also call the demographic group, attribute of interest, or facet) associated with each example (e.g., gender or race/ethnicity groups), and based on the value of this feature, we designate the example as part of either the advantaged group (marked/subscripted by a) or the disadvantaged group (marked/subscripted by d). Bias may be measured from a comparison of the original labels ($y \in \{0, 1\}$) of the data sample with the predicted labels ($\hat{y} \in \{0, 1\}$). Assume that $\hat{y} = 1$ is the accepted case and $\hat{y} = 0$ is the rejected case (for example, in the loan application use case). The ML classifier sets $\hat{y}_i = 1$ for observation i if $p(X_i) > H$, where $p \in (0, 1)$ is the probability generated by the classifier operating on feature set X, and H is a threshold cutoff level of probability, taken to be $H = 1/2$ by default.

In the training dataset we may count the number of labels of values 0 and 1, grouped by the restricted feature (denoted X_r). Denote the number of observed labels of value 0, 1 as $n^{(0)}, n^{(1)}$, respectively, and the number of labels of each class as n_a, n_d. These comprise labels of the advantaged and disadvantaged class, i.e., $n_a^{(0)}, n_a^{(1)}$ and $n_d^{(0)}, n_d^{(1)}$, respectively. We also have that $n_a^{(0)} + n_a^{(1)} = n_a$ and $n_d^{(0)} + n_d^{(1)} = n_d$. Corresponding to this notation for observed labels, we have a parallel notation for predicted labels $\hat{y}$, with counts $\hat{n}^{(0)}, \hat{n}^{(1)}$, etc. This minimal notation provides several measures of bias (the list is far from exhaustive).

Therefore, bias measurement is implemented using inputs $\{y, X, X_r\}$ and outputs $\{\hat{y}, p(X)\}$ generated by the machine learning model. These quantities are then used to generate several measures of bias. Therefore, bias measurement is model agnostic, because these quantities are not tied to a specific model. We note, however, that bias mitigation may not always be model agnostic, though the mitigation approaches we develop in this paper will be, i.e., we can treat the ML model as a black box for the purposes of mitigation as long as we amend only the five quantities above and not the structure of the model itself. We treat this is as our definition of model agnosticity throughout the paper, both for measurement and mitigation of bias. We note here that the protected characteristic X_r may not always be part of the model feature set and might be separately required. It may be available but in the event it is not, the dataset may be combined with other data to create a protected class variable for the purpose of assessing bias, see for example Kallus et al. (2020).

3.3. Pre-Training Metrics

We want to develop metrics that can be computed on the raw dataset before training as it is important to identify bias before expending time/money on training, which may also exacerbate pre-existing bias in the training data. One may wish to use a survey approach to determine the "golden truth" and compare it to the dataset to make sure the data is not too contaminated with bias to be useful. The golden truth is the joint statistical distribution of model inputs we would like to have or have deemed to be fair before we train any model. These distributions may not always be available, so pre-training bias metrics will be measures for comparison to a golden truth, were it to be available. If not, modelers and others reviewing the metrics will at least be able to assess whether the pre-training bias metrics are in violation of a judgment threshold level. The following pre-training metrics are of course model-independent.

1. Class imbalance (CI): Bias is often generated from an under-representation of the disadvantaged group in the dataset, especially if the desired "golden truth" is equality across groups. As an example, algorithms for granting small business loans may be biased against women because the historical record of loan approvals contains very few women, because women did not usually apply for loans to start small businesses. This imbalance can carry over into model predictions.

   We will report all measures in differences and normalized differences. Since the measures are often probabilities or proportions, we want the differences to lie in $(-1, +1)$. We define $CI = \frac{n_a - n_d}{n} \in (-1, +1)$ in normalized form. We see that CI can also be negative, denoting reverse

bias.

Mostly, the proportion difference is what is needed, but sometimes we may need the normalized difference, as in the case of the 80% Rule, that may be used to measure certain types of employment discrimination, see the 80% rule. In this case, it is the *ratio* that is important, so the normalized probabilities are able to capture this.

2. Difference in positive proportions in observed labels (DPL): Let $q_a = \frac{n_a^{(1)}}{n_a}$ be the ratio of type 1 for the advantaged class and $q_d = \frac{n_d^{(1)}}{n_d}$ be the same for the disadvantaged class. $DPL = q_a - q_d$, and $DPL = \frac{q_a - q_d}{q_a + q_d} \in (-1, +1)$. Clearly, DPL measures bias resident in the dataset at the outset. If $DPL \approx 0$, then we say that "demographic parity" has been achieved ex-post (i.e., already in the historical record). DPL is necessary, but not sufficient, to claim that the labels in the dataset are uncorrelated with the protected characteristic. Whereas here we are considering demographic parity pre-training, we note that demographic parity is a concept that also applies to post-training predictions from a model.

3. Kullback and Leibler (1951) Divergence (KL): Vasudevan and Kenthapadi (2020) propose a few pre-training measures of bias. We adapt the definition in the paper to our purposes here. We compare the probability distribution of the advantaged class (P_a) with that of the disadvantaged class (P_d), using KL divergence, i.e., relative entropy (Kullback in fact preferred the term "discrimination information"). The following formula measures how much information is lost when we move from P_a to P_d, i.e., the divergence of P_d from P_a:

$$KL(P_a, P_d) = \sum_y P_a(y) \log \left[\frac{P_a(y)}{P_d(y)} \right] \geq 0.$$

For the binary class data we have here, we may compute the KL divergence for all features one by one, and for the labels, i.e., the variable y refers to both types of quantities, whether they are binary, multi-category, or continuous. In financial services, greater attention would be paid to KL measurements on the labels, which are in effect, alternate forms of DPL.

4. Jensen-Shannon divergence (JS): if the distribution of the combined classes is P, then

$$JS(P_a, P_d, P) = \frac{1}{2} \left[KL(P_a, P) + KL(P_d, P) \right] \geq 0.$$

After computing the divergences for all features, we may re-order them to highlight the features that are most different across the two classes.

5. L_p norm (LP): Instead of the entropy differences in KL and JS, we may consider norm differences. For $p \geq 1$, we have

$$L_p(P_a, P_d) = \left[\sum_y |P_a(y) - P_d(y)|^p \right]^{1/p} \geq 0.$$

We note that this metric cannot be negative, hence, we cannot distinguish between bias and reverse bias.

6. Total variation distance (TVD): this is half the L_1 distance:

$$TVD = \frac{1}{2} L_1(P_a, P_d) \geq 0.$$

As with LP, this measure is also non-negative.

7. Kolmogorov-Smirnov (KS), two-sample approximated version:

$$KS = \max(|P_a - P_d|) \geq 0.$$

It is possible to evaluate the KS statistical test from the following distance measure, where the null hypothesis is rejected at level α:

$$KS > c(\alpha) \sqrt{\frac{n_a + n_d}{n_a \cdot n_d}}.$$

The value of $c(\alpha)$ is given by $c(\alpha) = \sqrt{-\ln(\frac{\alpha}{2}) \cdot \frac{1}{2}}$.

8. Conditional Demographic Disparity in Labels ($CDDL$): Wachter et al. (2020) developed this measure, which can be applied pre-training and also post-training. The metric asks the following question: Is the disadvantaged class a bigger proportion of the rejected outcomes than the proportion of accepted outcomes for the same class? We note that just this question alone would lead to an answer to whether demographic disparity exists (DD), not $CDDL$. Conditioning on an attribute is needed to rule out Simpson's paradox.[9] The example arises in the classic case of Berkeley admissions where men were accepted at a 44% rate and women at a 36% rate overall, but when this metric was examined department by department, the women on average were admitted in higher proportions. This is because more women applied to departments with lower acceptance rates than men did, though even in these departments, women were accepted at a higher rate than men.

We define

$$D = \frac{\text{No of rejected applicants from the protected facet}}{\text{Total rejected applicants}} = \frac{n_d^{(0)}}{n^{(0)}}$$

and

$$A = \frac{\text{No of accepted applicants from the protected facet}}{\text{Total accepted applicants}} = \frac{n_d^{(1)}}{n^{(1)}}$$

If $D > A$, then demographic disparity (DD) exists. $CDDL$ arises when demographic disparity exists on average across all strata of the sample on a user-supplied attribute. We will subgroup the sample and compute DD for each subgroup, and then compute the count-weighted average of DD. The function is as follows:

$$CDDL = \frac{1}{n} \sum_i n_i \cdot DD_i$$

[9] https://www.britannica.com/topic/Simpsons-paradox .

where i subscripts each subgroup and n_i is the number of observations in each subgroup, such that $\sum_i n_i = n$.

Therefore, of the eight pre-training bias metrics, the first two can detect negative bias, whereas the next five are agnostic to which class is advantaged or disadvantaged. The last one is positive. As an additional note, we stipulate that each measure might be compared to a golden truth or judged for its absolute level, but there is no metric under which we may undertake a cross-sectional comparison of these metrics, as they measure different attributes of the distribution of labels and attributes of interest. If we find that there is a class imbalance, we may wish to rebalance the sample before proceeding to train the model (see Appendix 6 where we considered this intervention in greater detail). We now move on to consider post-training bias metrics.

3.4. Post-Training Metrics

At this stage we have computed the pre-training metrics and we may also have rebalanced the sample to address any class imbalances that may exist. After training the ML model, we then compute the following bias metrics.

1. Difference in positive proportions in predicted labels ($DPPL$): Let $\hat{q}_a = \frac{\hat{n}_a^{(1)}}{n_a}$ be the ratio of type 1 for the advantaged class and $\hat{q}_d = \frac{\hat{n}_d^{(1)}}{n_d}$ be the same for the disadvantaged class. $DPPL = \hat{q}_a - \hat{q}_d \in (-1, +1)$. A comparison of DPL with $DPPL$ assesses if bias initially resident in the dataset increases or decreases after training.

   This is also similar to a metric known as Mean Difference (MD) in the Themis-ML package. There is an alternate computation for this. Compute $\bar{y}_a = E[\hat{y}_a]$ and $\bar{y}_d = E[\hat{y}_d]$, i.e., take the mean of the predicted values of both classes. Then we define $MD = \bar{y}_a - \bar{y}_d$. A positive mean difference implies bias against the disadvantaged class.

   This is also known as the Calders and Verwer (2010) (CV score). The only difference in our measure here is the additional normalization. Galhotra et al. (2017) call this the "group discrimination score." This is also called "statistical parity" in Berk et al. (2017), Corbett-Davies and Goel (2018). $DPPL$ has the obvious flaw that it assumes that the advantaged and disadvantaged classes are equally qualified (for a loan, job, etc.). Imposing $DPPL$ legally amounts to affirmative action if one of the classes is less qualified, however, it may still be required to correct unfairness. Therefore, it is definitely a useful measure from a societal point of view, even if it violates some notions of statistical fairness.

2. Disparate Impact (DI): The ratio version of the $DPPL$ measure is known as "disparate impact" and is formulated as follows:
$$DI = \frac{\hat{q}_d}{\hat{q}_a}.$$
   In an employment context in the US, regulators will use 80% as a "rule of thumb" for measuring disparate impact. This is an ex-ante measurement of demographic parity. The next measure takes a more nuanced view of this situation.

3. Difference in conditional outcomes (DCO): This metric examines the difference in proportions of acceptance or rejection between the two classes of qualified observations. Then we define two measures:

   - Difference in Conditional Acceptance (DCA): Define $c_a = \frac{n_a^{(1)}}{\hat{n}_a^{(1)}}$ and $c_d = \frac{n_d^{(1)}}{\hat{n}_d^{(1)}}$, as ratios of observed labels to predicted labels. $DCA = c_a - c_d$. This metric comes close to mimicking human bias. For example, when loan applicants from class d are approved for loans by the model, but the human overrides these approvals (as in redlining), then $n_d^{(1)} < \hat{n}_d^{(1)}$, leading to $c_d < 1$, and likewise, we might have $c_a > 1$ (positive bias). This would generate $DCA > 0$. DCA is related to equality of opportunity. It is a way of measuring "active" bias by a loan officer, where equally qualified applicants in both classes are treated differently to the detriment of class d. Interestingly, our nomenclature here may relate $DCA > 0$ as "bias" and $DCA < 0$ as "affirmative action."
   - Difference in conditional rejection (DCR): This metric examines the difference in proportions of rejection between the two classes of qualified observations. Define $r_a = \frac{n_a^{(0)}}{\hat{n}_a^{(0)}}$ and $r_d = \frac{n_d^{(0)}}{\hat{n}_d^{(0)}}$. Then we define $DCR = r_d - r_a$. From a terminology point of view, when both DCA and DCR are related to equalized odds (ex-post, i.e., in practice). When both DCA and DCR are very close to 0, we can conclude that the proportion of qualified (as suggested by observed labels) applicants accepted by the model and the proportion of unqualified applicants rejected are nearly equal across both classes.

4. Recall difference (RD): Here, higher recall for the advantaged class suggests that the ML model is better at distinguishing true positives from false negatives than for the disadvantaged class, i.e., it finds more of the actual true positives for the advantaged class than the disadvantaged class, which is a form of bias. We define $RD = \text{Recall}_a - \text{Recall}_d$. Berk et al. (2017) call this "conditional procedure accuracy equality."

5. Difference in Label rates (DLR): The metrics in this category come in three flavors:

   - Difference in acceptance rates (DAR): This metric measures whether qualified applicants from the

advantaged and disadvantaged class are accepted at the same rates. It is the difference in the ratio of true positives divided by the predicted positives for each class.

$$DAR = \frac{TP_a}{\hat{n}_a^{(1)}} - \frac{TP_d}{\hat{n}_d^{(1)}}.$$

Satisfying $DAR \approx 0$ implies members of both groups are accepted at the same rates. Note that DAR and DCA are somewhat similar, because we also see that $DAR = \frac{TP_a}{TP_a+FP_a} - \frac{TP_d}{TP_d+FP_d}$, whereas $DCA = \frac{TP_a+FN_a}{TP_a+FP_a} - \frac{TP_d+FN_d}{TP_d+FP_d}$, which is also therefore, related to the measure for Treatment Equality (TE) in adjusted form, below.

DAR may also be interpreted as Precision difference (PD = Precision$_a$ − Precision$_d$). It implies that more of the predicted true positives are valid compared to false positives for the advantaged group. These separate precision values for each class are computed by dividing the confusion matrix into two parts, grouped on class. Berk et al. (2017) call this "conditional use accuracy equality". It is also denoted as "predictive parity." This metric is the same as DAR above.

- Difference in rejection rates (DRR): This metric measures whether qualified applicants from the advantaged and disadvantaged class are rejected at the same rates. It is the difference in the ratio of true negatives divided by the predicted negatives for each class.

$$DRR = \frac{TN_d}{\hat{n}_d^{(0)}} - \frac{TN_a}{\hat{n}_a^{(0)}}.$$

Also, just as DAR is related to DCA, we see that DRR is related to DCR.

6. Accuracy Difference (AD): We apply the usual definition for accuracy, i.e.,

$$\text{Accuracy}_k = \frac{TP_k+TN_k}{TP_k+TN_k+FP_k+FN_k}, \quad k = \{a, d\}.$$

We measure the accuracy of the classifier for each of the classes (i.e., we split the confusion matrix into two matrices, one for each class) and then assess which one is measured more accurately. Here bias is the difference in classifier accuracy across the classes, i.e., AD = Accuracy$_a$ − Accuracy$_d$.

7. Treatment Equality (TE): Berk et al. (2017) defined this as the ratio of false negatives to false positives (or vice-versa). Compute this for both classes, $\tau_a = FN_a/FP_a$, $\tau_d = FN_d/FP_d$. Then we define $TE = \tau_d - \tau_a$.

8. Conditional Demographic Disparity of Predicted Labels ($CDDPL$): this is the same metric as in the pre-training case, except applied to predicted labels rather than observed labels.

9. Counterfactual difference (CD_n): This measure is counterfactual analysis in economist-speak. The idea is that if there is no bias on a characteristic, then "flipping the bit" on the characteristic should make no difference to the model prediction. We describe three versions of this metric, CD_1, CD_2, CD_3. An extensive discussion of the hidden assumptions and usefulness of counterfactual analysis is provided in Barocas et al. (2020).

It is easy to measure in a binary class situation, simply flip the dummy variable identifying the advantaged (a) and disadvantaged (d) classes and see what proportion of predictions change from 0 to 1 and vice-versa for each of the two classes. Let n'_a be the number of flipped outcomes for class a and n'_d for class d. The measure is $CD_1 = \frac{n'_a+n'_d}{n} \in (0, 1)$, where $n = n_a + n_d$. Galhotra et al. (2017) call this the "causal discrimination score." Counterfactuals and causality are linked and a good discussion is in Pearl (2009). For a general description of causal theory based on counterfactuals, see Menzies and Beebee (2019).

We add an additional score to discriminate further, i.e., we check whether class a is more robust to the protected characteristic than class d. Let $\delta_a = \frac{n'_a}{n_a}$ be the proportion of predictions of class a that flip and $\delta_d = \frac{n'_d}{n_d}$ be the flipped proportion for class d. Then the bias measure will be $CD_2 = \delta_d - \delta_a$ and $CD_3 = \frac{\delta_d-\delta_a}{\delta_a+\delta_d} \in (-1, +1)$. If $CD_2, CD_3 > 0$ then class d is more sensitive to characteristic than class a, which is also a form of bias. This new measure extends the GBM score to a nuanced one. We call it "counterfactual sensitivity".

There is no guarantee that flipping the bit ensures counterfactual fairness, because it may still allow the ML model to be biased against the disadvantaged class by using features that are correlated with the protected characteristic. Therefore, it may require the modeler to ensure that the model is fair under the counterfactual when using a matched sample, i.e., for each member of the disadvantaged class we find a matching member of the advantaged class that is very close on all variables except the protected characteristic. Such approaches, known by different terminology such as propensity score matching (in econometrics) and optimal transport (in computer science) are effective and provide a computationally inexpensive approach to ensuring fairness in ML, see for example the following approach, denoted FlipTest.

We may also use causal influence quantification to answer fairness questions, see Janzing et al. (2019). For example, hypothetically speaking, if the rejection rates for women applying for loans is higher than that for men, it may be because women apply for loans to set up more risky businesses, where the rejection rates are higher than for loans where business risk is lower, to which more men apply. In such settings, the counterfactual bias measure looks like it is keying off gender, but if a

causality-based approach is taken, then we will find that rejection is not gender based as there will be no causal link from gender to rejection, but from business type instead. By requiring that a causal link be established from the protected characteristic to the decision, we are able to carefully assess unfairness at a deeper level. However, the information requirements to implement such tests are more onerous. On the other hand, the FlipTest does not impose additional information requirements.

10. FlipTest (FT_n): Black et al. (2020) introduced this method as a black-box technique to detect bias in classifiers by leveraging an optimal transport framework. FT evaluates the cardinality of two different sets of examples: $F^+(h,G) = \{x \in D | h(x) > h(G(x))\}$, and $F^-(h,G) = \{x \in D | h(x) < h(G(x))\}$, where h is our model, D is the discriminative neural network of the GAN (i.e., generative adversarial network, see Goodfellow et al. (2014)) and G, the generator neural network part, is a transformation from the space of the disadvantaged examples to the space of the advantaged ones. The function G can be learned in different ways. In the original work, G is a Wasserstein-GAN model from Arjovsky et al. (2017), with a modified loss function that imposes on the generator to produce artificial examples from (the distribution of) the advantage class using an input example from the disadvantage one, and keeping a small distance (in the feature space) between them. In our case, we approximate directly $h(G(x))$ with a kNN approach on top of the prediction of the model h. In fact, we infer the label of a disadvantage example by using the kNN prediction (i.e., the most represented label among the k closest examples from the advantage class). We use this prediction as our $h(G(x))$ in order to populate the Flipsets. From the two Flipsets, we generate the final metrics as follow:

    - $FT_1 = \frac{|F^+(h,G)|+|F^-(h,G)|}{n_d} \in [0,1]$,
    - $FT_2 = \frac{|F^+(h,G)|-|F^-(h,G)|}{n_d} \in [-1,+1]$.

    Our approach is simplified and defines the metric as

$$FT = \frac{F^+ - F^-}{n_d} \in [-1,+1]$$

    where F^+ is the number of disadvantaged group members with an unfavorable outcome whose nearest neighbors in the advantaged group received a favorable outcome, F^- is the number of disadvantaged group members with a favorable outcome whose nearest neighbors in the advantaged group received an unfavorable outcome.

We summarized the various metrics and cross-reference the metrics to the various terminology used in the literature, as many metrics have multiple nomenclature (see Table 1). We note that the notion of "total fairness" is achieving all these metrics, which is not possible. Why? Most of the metrics are permutations of the four numbers in a binary classification confusion matrix and it is impossible to satisfy all fairness metrics with just these 4 numbers. For example, Canetti et al. (2019) show that there is no general way to post-process a classifier to equalize positive predictive value (DCA) or negative predictive value (DCR) for the disadvantaged class. Therefore, customers will need to choose one or two metrics that they need to comply with.

Some bias measures are legal artifacts and may be evidenced by the existence of one or many of the metrics below, such as "disparate treatment"—a US construct (of intentional discrimination) where someone is treated differently based on their membership in a protected class, often more strongly evidenced by counterfactual unfairness, see Kusner et al. (2018). The measures we consider may not capture all legal notions of unfairness, because they may be hard to detect, such as the US notion of "disparate impact" (also known as unintentional discrimination), which occurs when the disadvantaged class experiences bias even when the approach taken is apparently fair, e.g., when a HR test is required by all candidates but just happens to be harder for one gender because of social conditioning. Here this bias does not come from the ML model, but may still be detected by it. From a ML point of view, we do not make a distinction between disparate impact and treatment. In Table 1 below, we specifically identify disparate impact with the definition used in practice, i.e., metric #10, whereas all metrics that treat people of different groups differently are generally described as examples of disparate treatment.

Table 1 shows a summary of the bias metrics, their definitions, and related nomenclatures.

4. Non-Binary Attributes and Labels

The metrics we examined in Section 3 apply to binary attributes of interest (also known as protected attributes or classes, e.g., gender male vs female) and binary target variables (e.g., a loan was approved or not). In this section, we discuss extending these situations to non-binary settings. Whereas most cases of fairness assessment relate to binary cases, we recognize that there may arise a need for non-binary settings as well.

4.1. Non-Binary Attributes of Interest

Attributes of interest (protected characteristics) need not be binary. For example, ethnicity may be divided into multiple categories, such as White, Black, Asian, Latinx, Indigenous Populations, etc. In this case we may need to aggregate multiple categories into binary ones for the purposes of attribution of fairness. Here, for instance, we may collect Black, Latinx, and Indigenous Populations into a group called "minority" and the others into a group called "non-minority." We then proceed as before to compute metrics

Table 1. FAML Metrics. Notation: Original labels ($y = \{0, 1\}$); predicted labels ($\hat{y} = \{0, 1\}$). The ML classifier sets $\hat{y}_i = 1$ for observation i if $p(X_i) > H$, where $p \in (0, 1)$ is the probability generated by the classifier operating on feature set X, and H is a threshold cutoff level of probability, taken to be $H = 1/2$ by default; restricted feature (X_r); advantaged group (a); disadvantaged group (d); number of observed labels of type 0, 1 are $n^{(0)}, n^{(1)}$, respectively; number of labels of each class are n_a, n_d; labels of the advantaged and disadvantaged class, i.e., $n_a^{(0)}, n_a^{(1)}$ and $n_d^{(0)}, n_d^{(1)}$, respectively; predicted labels $\hat{y}$, with counts $\hat{n}^{(0)}, \hat{n}^{(1)}$; ratio of type 1 for the advantaged class ($q_a = n_a^{(1)}/n_a$) and $q_d = n_d^{(1)}/n_d$ for the disadvantaged class; probability distribution of the advantaged (disadvantaged) class $P_a, (P_d)$, with average distribution P; $\hat{q}_a = \hat{n}_a^{(1)}/n_a$ be the ratio of predicted type 1 for the advantaged class and $\hat{q}_d = \hat{n}_d^{(1)}/n_d$ be the same for the disadvantaged class; $c_a = n_a^{(1)}/\hat{n}_a^{(1)}$ and $c_d = n_d^{(1)}/\hat{n}_d^{(1)}$, as ratios of observed labels to predicted labels; $r_a = n_a^{(0)}/\hat{n}_a^{(0)}$ and $r_d = n_d^{(0)}/\hat{n}_d^{(0)}$; TP: true positives; FP: false positives; TN: true negatives; FN: false negatives; ratio of false negatives to false positives (or vice-versa). Compute this for both classes, $\tau_a = FN_a/FP_a$, $\tau_d = FN_d/FP_d$; n_a' be the number of flipped outcomes for class a and n_d' for class d; $\delta_a = \frac{n_a'}{n_a}$ be the proportion of predictions of class a that flip and $\delta_d = \frac{n_d'}{n_d}$ be the flipped proportion for class d. The alternate nomenclature refers to both, exact same measures or related ones.

Metric	Indicative Formula	Related Nomenclature
1. Class Imbalance (CI)	$n_a - n_d$	-
2. Difference in positive proportions in observed labels (DPL)	$q_a - q_d$	pre-training demographic parity
3. KL Divergence (KL)	$\sum_y P_a(y) \log\left[\frac{P_a(y)}{P_d(y)}\right]$	-
4. JS Divergence (JS)	$\frac{1}{2}\left[KL(P_a, P) + KL(P_d, P)\right]$	-
5. L_p norm (LP)	$\left[\sum_y \lvert P_a(y) - P_d(y)\rvert^q\right]^{1/q}$	-
6. Total Variation Distance (TVD)	$\frac{1}{2} L_1(P_a, P_d)$	-
7. Kolmogorov-Smirnov (KS)	$\max(\lvert P_a - P_d\rvert)$	-
8. Conditional Demographic Disparity ($CDDL$)	$CDDL = \frac{1}{n}\sum_i n_i \cdot DD_i$	-
9. Difference in positive proportions in predicted labels ($DPPL$)	$\hat{q}_a - \hat{q}_d$	mean difference, demographic parity, statistical parity, disparate treatment, group discrimination score
10. Disparate Impact (DI)	$DI = \frac{\hat{q}_d}{\hat{q}_a}$	disparate impact
11. Difference in conditional acceptance (DCA)	$c_a - c_d$	equality of opportunity, individual fairness, disparate treatment
12. Difference in conditional rejection (DCR)	$r_a - r_d$	equalized odds (includes DCA), disparate treatment
13. Recall difference (RD)	$\frac{TP_a}{TP_a+FN_a} - \frac{TP_d}{TP_d+FN_d}$	sufficiency, conditional procedure accuracy, false positive rate, success prediction error, disparate treatment
14. Difference in acceptance rates (DAR)	$\frac{TP_a}{\hat{n}_a^{(1)}} - \frac{TP_d}{\hat{n}_d^{(1)}}$	equality of opportunity, individual fairness, disparate treatment
15. Difference in rejection rates (DRR)	$\frac{TN_d}{\hat{n}_d^{(0)}} - \frac{TN_a}{\hat{n}_a^{(0)}}$	equalized odds (includes DCA), disparate treatment
16. Precision difference (PD)	$\frac{TP_a}{TP_a+FP_a} - \frac{TP_d}{TP_d+FP_d}$	false negative rate, failure prediction error, conditional use accuracy, disparate treatment, predictive parity
17. Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	-
18. Accuracy difference (AD)	$\text{Accuracy}_a - \text{Accuracy}_d$	disparate treatment
19. Treatment Equality (TE)	$\tau_d - \tau_a$	-
20. Counterfactual difference	$CD_1 = \frac{n_a'+n_d'}{n}$	-
	$CD_2 = \delta_d - \delta_a$	counterfactual fairness, disparate treatment
21. Fliptest (FT)	$FT_1 = \frac{\lvert F^+(h,G)\rvert + \lvert F^-(h,G)\rvert}{n_d}$	counterfactual analysis
	$FT_2 = \frac{\lvert F^+(h,G)\rvert - \lvert F^-(h,G)\rvert}{n_d}$	counterfactual analysis
	$(F^+ - F^-)/(n_d)$	counterfactual analysis

based on this new variable. Of course, we might need to create more than one protected characteristic variable for various permutations of the multi-category class variable.

Another case we encounter is that of cross-categories. For example, we may want to treat under-represented classes who are senior citizens as a disadvantaged class. In this case the attribute of interest would be a new feature variable that is the product of the under-represented feature and the senior citizen feature. Again, this allows the computation of bias metrics to proceed seamlessly using the metrics above. In short, we convert multi-category or continuous class variables into binary ones for the purpose of bias measurement.

4.2. Non-Binary Labels (target variables)

Again, we may encounter multi-category, and continuous target variables, in addition to the standard case of a binary target variable. An example of a multi-category target variable is a credit rating of an individual who borrows money on a website such as Lending Club. The site assigns a credit rating that is a discrete one from a multitude of rating levels. An example of a continuous target variable is the interest rate offered on a loan or the salary offered in a hiring situation. In all these cases, we are interested in assessing whether the disadvantaged class is treated differently than the advantaged class.

In Table 2, we present how the non-binary case will apply to each of our bias metrics. For a multi-category target variable (label) requires that for each label we take the difference between facets (classes, e.g., male vs female) for each label and report the maximum difference. For a continuous variable we report the mean difference in label across facets. As we see from Table 2, it is feasible to handle non-binary labels in FAML.

5. Bias Mitigation

Bias corrections can take many forms and may lead to different tradeoffs between fairness and accuracy for each ML model. Some common corrections that may be applied are as follows:

1. Removal of the class variable from the feature set. For example, restricted characteristics such as gender, race/ethnicity, and age may be part of the feature set and removal of these will mitigate some or all of the bias metrics mentioned above. However, as is to be expected, this will also impact accuracy. Moreover, the real problem often lies elsewhere, given that protected attributes are almost always eliminated from feature sets, but not all features that are correlated with the attribute.

2. Rebalance the training sample pre-training. This corrects unfairness from differences in base rates. Synthetically increase the number of observations $n_d^{(1)}$ if $n_a^{(1)} > n_d^{(1)}$. Synthetic oversampling is undertaken using standard algorithms such as SMOTE, available in SkLearn. Likewise, decrease $n_d^{(0)}$ if $n_a^{(0)} < n_d^{(0)}$. Both these corrections are in the spirit of affirmative action. Both these adjustments are intended to result in equal sized classes across $(0, 1)$ labels. If the class variable truly matters then rebalancing usually results in a loss in accuracy. Random perturbation of class labels is also possible instead of using oversampling. But, this approach results in different results every time. One can also transform the features such that their joint distribution without the class variables remains more or less the same, but the correlation with the class variable is reduced, as close to zero as possible.

3. Adjust labels on the training dataset and re-train. For the advantaged class, adjust the ground truth such that $y_a = 0$ if $\hat{y}_a = 1$ and $p_a(X) < H + \eta$ for some well-defined hyperparameter η. That is, downgrade some of the borderline positive labels for the advantaged class. Likewise, set $y_d = 1$ if $\hat{y}_d = 0$ and $p_d(X) > H - \eta$, i.e., upgrade some of the borderline negative labels for the disadvantaged class. Then re-run the ML model fit. Recompute the various bias and accuracy metrics.

4. Adjust cutoffs post-modeling. The cutoff probability is usually set at $H = 1/2$. If bias is present, then the cutoff for the advantaged class can be adjusted to $H + \delta$, and the cutoff for the disadvantaged class will be reduced to $H - \delta$. This will change the predicted counts $\hat{n}_a^{(0)}, \hat{n}_a^{(1)}, \hat{n}_d^{(0)}, \hat{n}_d^{(1)}$, and change many of the bias measures as well as the accuracy of the model. Hyperparameter δ can be tuned appropriately, until a desired level of fairness and/or accuracy is achieved. The legal milieu may not accommodate direct alteration of the predictions, so the availability of this mitigation is subject to the domain of application.

These bias corrections will result in changes in fairness and accuracy for all of the ML models that are applied to the training dataset. Since we have several fairness metrics, and there is a tension amongst them, we train our models with these fairness metrics as constraints, either applied ex-post or at training time. With multiple constraints, we need to either (i) choose one constraint (which is limiting), (ii) weight the constraints to consolidate them into a single constraint, or (iii) apply a min-max criterion, i.e., minimize the maximum bias metric under all the different constraints we choose to include while training the model.

6. Sequencing FAML in Practice: A Case Study

Our FAML process has several stages and offers a series of steps for the ML modeler to decide whether the final model

Table 2. Treatment of bias metrics for non-binary targets. We note that the outcome variable is denoted "label" and the class variable is called a "facet". We will collapse all multicategory and continuous facets to binary and then apply the rules in this table to non-binary cases of labels.

Metric	Multi-category	Continuous
DPL	diff in proportions of each observed label between facets, and report the max diff	diff in mean actual label value between facets for each label
KL, JS	max divergence across facets for each label	max divergence for continuous distributions across facets
L_p, TVD, KS	max metric of the absolute diff in distribution across facets for each label	metric of the absolute diff in continuous distribution across Facets
$CDDL$	max diff in proportion of actual rejects and proportion of actual accepts across facets, conditional on feature subgroup, for each label	not defined
$DPPL$	diff in predicted proportions for each label between Facets and report the max diff	diff in mean predicted label value between facets
DI	ratio of predicted proportions for each label across facets should be within (0.80-1.20), report the max	ratio of mean predicted label for Facets should be in the range (0.80-1.20)
DCA, DCR	diff in ratio of actual labels to predicted labels across facets for each label; report max	diff in ratio of mean actual values to mean predicted values across facets for each label
RD	diff in recall for each label across facets	not defined
DAR, DRR, PD	diff in rates of correctly predicted labels for each facet	diff in ratio of mean predicted values to actual mean values across facets for each label; Note: cannot be computed for PD in this case
AD	(1) diff in accuracy scores from the confusion matrix; (2) diff in AUC from the ROC analysis	diff in AUC from the ROC analysis
TE	diff in ratio of false positives to false negatives across facets	not defined
FT	avg diff in labels for kNNs across facets	avg diff in means of kNNs across facets
$CDDPL$	max diff in proportion of predicted rejects and proportion of predicted accepts across facets, conditional on feature subgroup, for each label	not defined

in the training process has achieved desired fairness levels. The schematic for the model is shown in Figure 1.

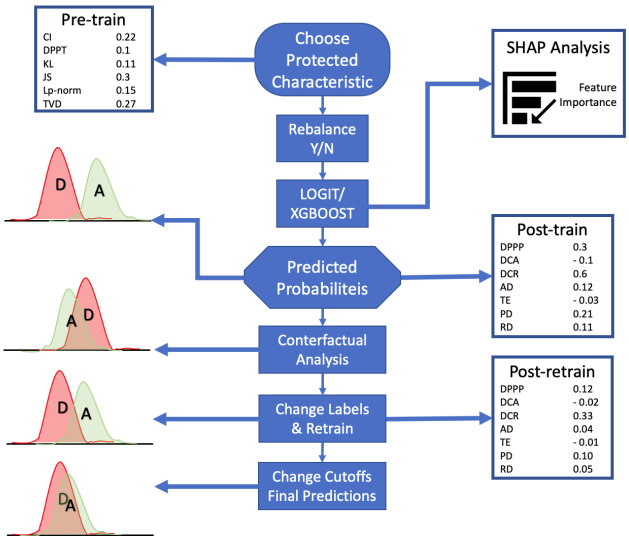

Figure 1. Fairness aware machine learning pipeline. There are 10 steps in the pipeline show here. (1) Read in data, (2) Choose protected characteristic, (3) Pre-training bias metrics, (4) Rebalance the sample if needed, (5) Train with logit or xgboost, and (6) (a) Show the distributions of predicted probabilities for each class and (b) the post-training metrics, (7) Counterfactuals using a flipped class variable, (8) Show distributions with flipped class variable as counterfactuals, (9) Adjust labels and re-train the model and then (a) plot probability distributions and (b) metrics, (10) Change cutoffs for final predictions and plot distributions and metrics.

We implement the model pipeline on a canonical sample, i.e., the well-known German credit dataset.[10] This small dataset comprises 1000 loan applicants with 20 features, such as checking account status, loan maturity, credit history, loan purpose, loan amount, whether the borrower has a savings account and how much balance they carry, employment status, instalments as a percent of income, gender, other debt, years of residence, property ownership, age, other instalment plans, rent or own, number of existing loans, type of job, number of dependents, telephone or not, and whether the person is a foreign worker. We one-hot encoded categorical variables as needed and clubbed some others in a process of standard data engineering. Eventually, we arrived at 59 features. The label is binary, taking a value 1 if the borrower was classified as a good credit, else it was given a value of 0. The class variable we chose to examine was gender, and we coded the gender feature as equal to 1 if female, and 0 for male. We assess bias metrics on both training and/or testing datasets. In our example here, we use all the data for training, so bias is measured on the training data. But we might be better off using a separate test data set to measure bias on as well. Which one to use is discussed in greater detail in Section 7.7.

Our first step comprises examining the data with the pre-training bias measures from Section 3.3. We computed these and the visualization of these is shown in Figure 2. For example, we see that class imbalance is $CI = 0.38$, which is because 31% of the sample comprises women and therefore difference in the percentage of men and women is $0.69 - 0.31 = 0.38$. The other metrics of pre-training bias are milder. DPL shows that imbalance in labels disfavors women (the bias is positive). The other measures are also positive but that is because they lie between 0 and 1, so may not necessarily reflect bias against women.

The next step in the pipeline is to decide whether to rebalance the sample or not. We decide not to do so (we will consider this later). Moving on, we train the model using logistic regression, achieving an accuracy level of 0.78 and an AUC (area under the curve) of 0.83. The model inputs and outputs are used to compute the various post-training metrics, and we also plot the distributions of predicted probabilities for mean (type 0) and women (type 1) to see how these differ. Results are shown in Figure 3. We see that some bias metrics are positive and some negative, exemplifying the point made earlier that bias metrics do not all track in the same direction.

We see from the $DPPL \approx 0$ metric that demographic parity is achieved. However, $DCA = 0.089$, i.e., statistical parity is not well maintained, as this indicates that men are given 9% preference relative to women in acceptances. But on the other hand, women are 25% disfavored ($DCR = 0.25$) in rejections. A combination of DCA and DCR suggests that the trained model is unfair. The other four metrics clearly do not seem to be adverse to the disadvantaged class. The difference in distributions of predicted probabilities is mild, though the distribution for men is shifted more to the right versus that for women, suggesting some small bias. But the bias metrics suggest that bias mitigation is predicated.

The next step is counterfactual analysis, where we flip the gender bit to see if the distributions of predicted probabilities changes, i.e., if the protected characteristic influences the model, then the difference in distributions should become smaller. The results are shown in Figure 4, but this does not show much change, suggesting that the protected characteristic itself does not matter, and that the bias is arising from other features in the data.

Moving down the system pipeline, we next adjust the labels and retrain the model. Labels were adjusted using a cut-off shift of $\delta = 0.05$, i.e., we relabeled the training data such that for women, any labels that were 0, but had predicted probabilities greater than 0.45 ($H - \delta$) were labeled as 1 instead. Likewise, for men, any labels that were 1 were rela-

[10] See the German data set in the UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets.php.

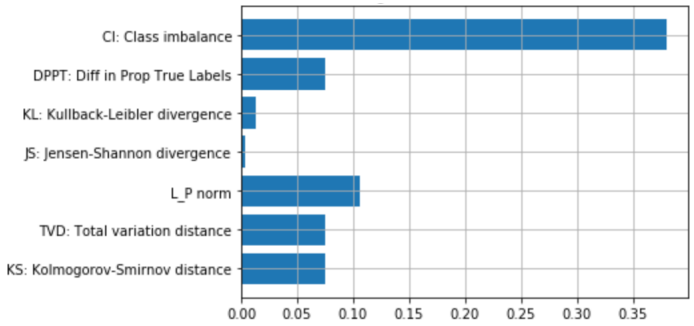

Figure 2. Pre-training metrics for the German credit dataset.

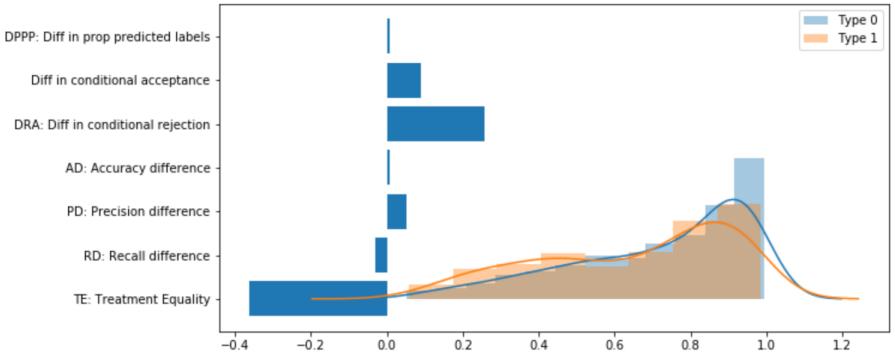

Figure 3. Post-training metrics for the German credit dataset. The blue distribution is the one for men (type 0) and the orange one is women (type 1).

beled as 0, if the predicted probabilities were less than 0.55 ($H - \delta$). We chose δ to bring down $DPPL, DCA, DCR$ as much as possible. The results are shown in Figure 5. This is essentially retraining the model on the features in the training dataset but with new labels.

We can see that we have minimized the bias quite dramatically across all measures, which is interesting. However, it exemplifies how the ML modeler may be able to use our framework to create a model that is less biased. Note also that the difference in distributions is also less than before bias mitigation.

We now apply the last mitigation approach, though it is hard to see that we can improve on what we have much further. Here, we do not adjust the labels and retrain the model, but instead, we directly adjust the predictions. We do this by choosing optimally a prediction shift parameter $\eta = 0.045$ and changing all predictions for women between $(H-\eta, H)$ from 0 to 1, and for men between $(H, H + \eta)$ to 0. The resulting metrics are shown in Figure 6.

We see that the bias has reduced from the original post-training metrics but not as much as before, though the outcomes are very close. It's clear that the best way here is to adjust labels and retrain the model. We see that bias mitigation works extremely well in the sequential approach here, without any rebalancing.

It is useful to ask: Can we do better using a different ML model? Instead of the logistic regression, we use an xgboost model and see if we can achieve higher levels of accuracy, while maintaining satisfactory levels of fairness. When we retrained the original model using xgboost, we achieved much higher accuracy levels with just two epochs of training, i.e., accuracy increased to 0.86 (from 0.78 in the logistic model) and AUC increased to 0.93 (versus 0.83 in logistic). The post-training metrics before any bias mitigations are shown in Figure 7. We also ran counterfactual differences and noted that the difference in distributions was not reduced (not shown for parsimony).

We can see that the bias is less than with logistic regression at the same stage in the pipeline. But we not apply both bias mitigations optimally: (i) relabel and retain, and (ii) adjust predictions. Figure 8 shows the results of both mitigations.

In both mitigation approaches, the best values of δ and η are unable to reduce the DCA and DCR a lot when we keep $DPPL$ as low as possible. This exemplifies the tension between the different bias measures. It also offers an example of the accuracy-bias tradeoff. We get higher accuracy from xgboost, but it comes at a cost of fairness. Ultimately, the user has to choose between the models. In

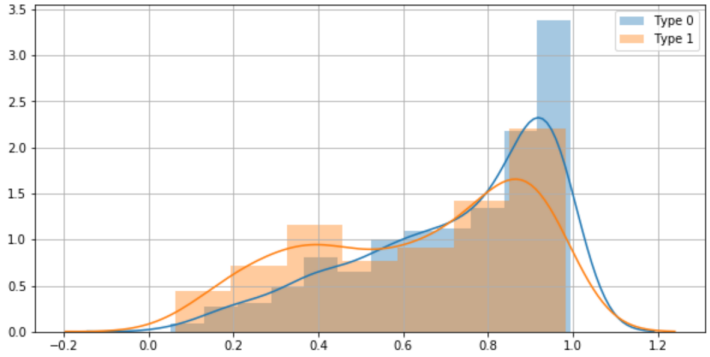

Figure 4. Counterfactual analysis. The blue distribution is the one for men (type 0) and the orange one is women (type 1).

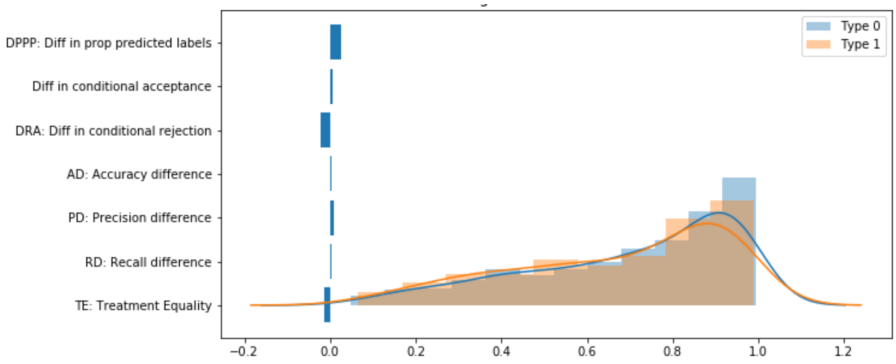

Figure 5. Metrics for the German credit dataset after adjusting labels and retraining the model. The blue distribution is the one for men (type 0) and the orange one is women (type 1).

the next section, we use this example as a backdrop to discuss some important issues to be taken into account when deciding on a final tuned FAML model.

7. Issues in Financial Fairness

7.1. Contradictions amongst Fairness Metrics

As we have seen above, there is a tradeoff between fairness and accuracy. There are also tradeoffs between the various fairness metrics. These arise from the limited degrees of freedom amongst the inputs into the fairness metrics. These issues are best illustrated with an example.

Assume a dataset of loan applications from 150 customers, of which 100 are men and 50 are women. Therefore, $CI = 0.33$. Of the men (class a), 70 were given loans and of the women (class d), 27 were granted loans, based on the actual labels in the data. Therefore, $DPL = 0.70 - 0.54 = 0.16$. The confusion matrix with a probability cut off of $H = 0.5$, for each class, is shown in the top part of Table 3. The top third of the table shows the confusion matrices with a probability cutoff of 0.5. The middle third gives the results when we move the cutoff to 0.55. The bottom of the table shows accuracy and fairness measures for both cutoffs. The combined confusion matrix for the model is the sum of the two confusion matrices. For illustration, we compute some of the accuracy and bias metrics off these confusion matrices, shown in Table 3.

We see that when the cutoff is moved to 0.55, the number of positive predictions decreases as it should, and the number of rejections increases for both the advantaged and disadvantaged class. The accuracy of the model is attenuated, falling from 0.82 to 0.79. However, demographic parity $DPPL$ improves as the bias falls from 0.25 to 0.18. Therefore, we see that an inverse tradeoff between fairness and accuracy in terms of this metric. This is also the case for some other fairness metrics that also improve, i.e., DCA, DAR, DRR, and AD. But, DCR, which was negative (reverse bias) is now positive. Hence, not all fairness metrics improve. In financial settings, the fact that disadvantaged class loan applicants are being rejected in practice more than that suggested by the algorithm would suggest malicious rejections by the loan officer. This example illustrates that it is not possible to improve all fairness metrics together. This comes from the fact that a small number of values in confusion matrices are unlikely to span all metrics in a monotonic manner. Therefore, this suggests that modelers generate multiple measures across a few different models

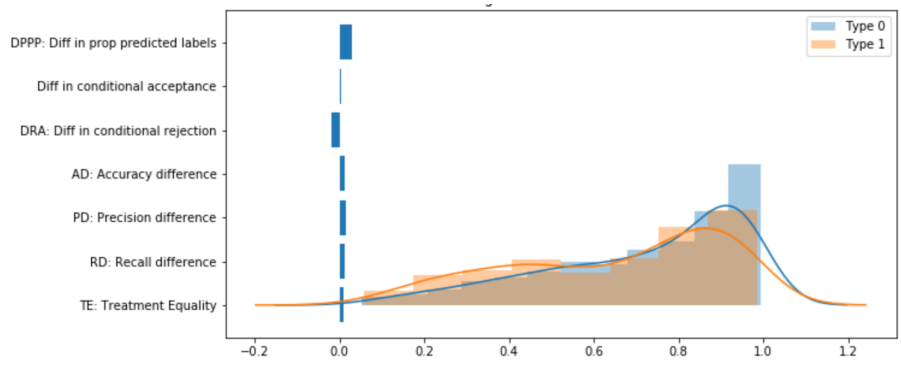

Figure 6. Metrics for the German credit dataset after changing the predictions. The blue distribution is the one for men (type 0) and the orange one is women (type 1).

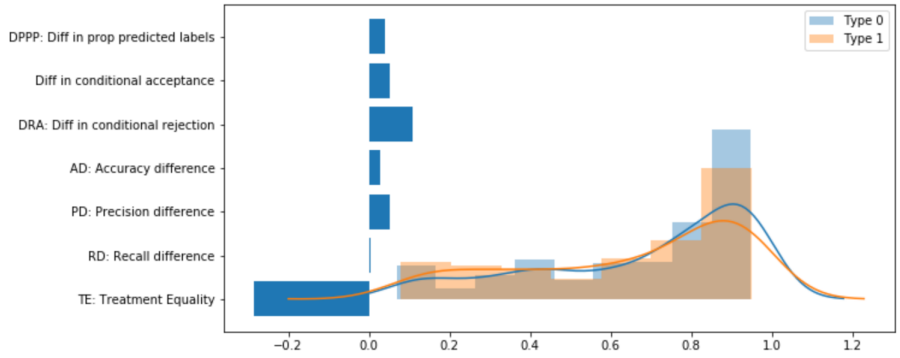

Figure 7. Post-training metrics for the German credit dataset using xgboost. The blue distribution is the one for men (type 0) and the orange one is women (type 1).

and minimize the bias metric that matches their situation most appropriately.

7.2. Regulations

The financial services industry has a history of regulation that imposes fairness. For example, the notion of a level playing field is embodied in the Truth in Lending Act (TILA, 1968), which protects borrowers against inaccurate and unfair credit billing and credit card practices. It requires lenders to provide loan cost information so that borrowers can comparison shop for loans. The Rumford Fair Housing Act of 1963 and the Civil Rights Act of 1964 were initial attempts to address discrimination, followed by the Fair Housing Act of 1968. According to the Department of Housing and Urban Development (HUD), examples of discriminatory practices include: different prices for the sale or rental of a dwelling, delaying or failing to maintain or repair homes for certain renters, or limiting privileges, services, or facilities of a dwelling based on a person's gender, nationality, or racial characteristics. These laws formed the basis for fairness in lending.

The Consumer Credit Protection Act (CCPA, 1968), Fair Credit Reporting Act (FCRA, 1970), and the Equal Credit Opportunity Act (ECOA, 1974) prohibits discrimination in credit transactions based on race or color, national origin, religion, sex, marital status, age, whether an applicant receives income from a public assistance program, and an applicant's exercise, in good faith, of any right under the Consumer Credit Protection Act. These are a set of some but not all protected characteristics financial services firms need to pay attention to in FAML. These are anti-discrimination laws and therefore impose a legal fairness requirement. Some of these laws also impose an explainability requirement: the ECOA in conjunction with Reg B require an adverse action notice (AAN) to be sent to any customer who is denied credit within 30 days. Reg B implements the ECOA and is enforced by the Consumer Finance Protection Bureau (CFPB). Reg B requires no more than 4 reasons—known as "principal reason explanations"—be provided for rejecting a loan, and Appendix C of Reg B provides a non-exhaustive list of reasons that may be provided for a credit denial. Principal reason explanations are legal constructs and not necessarily those supported by mathematical notions of feature importance.

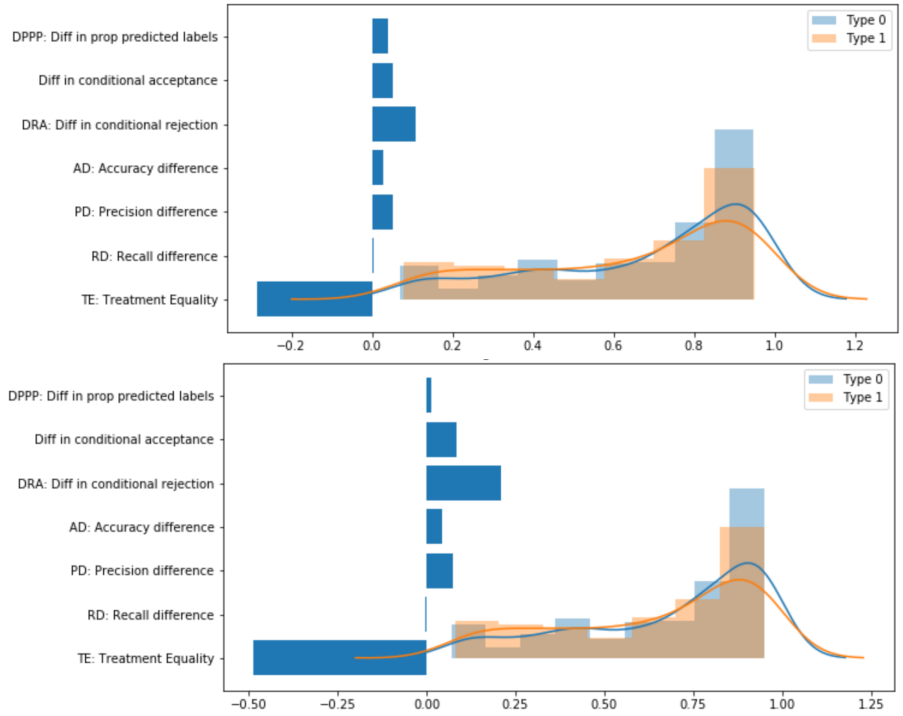

Figure 8. Post-mitigation metrics for the German credit dataset using xgboost. The blue distribution is the one for men (type 0) and the orange one is women (type 1). The upper plot is for mitigation via adjust labels ($\delta = -0.15$) and retrain, and the lower plot shows the outcome from adjusting the predictions directly ($\eta = 0.04$).

7.3. Questions to be asked by a financial user

The various metrics examine different nuances and ways in which bias may arise. However, because there is a multitude of metrics, it is not always easy to know which ones apply in a particular situation or domain. Therefore, Table 4 is prepared as a checklist of questions that may be helpful to a financial practitioner in order to gain guidance on which metric is most applicable to the model.

7.4. Simpson's paradox

As noted in Wachter et al. (2020), unfairness is an outcome of aggregation in a protected class and when a stratified view will show no bias. In the well-known Berkeley admissions study, see Bickel et al. (1975), the overall admissions rate for women was much lower than that for men, but when the data was stratified by department, the results showed on average, women were admitted at higher rates. The aggregate bias result emanates from the fact that more women applied to departments with lower admissions rates, but stratification clarifies that admissions were fair. This is known as Simpson (1951)'s paradox.

The metric conditional demographic disparity ($CDDL$) in Section 3.3 is designed to reflect bias after accounting for the possible presence of the paradox. We implemented this metric as an example of both pre-training and post-training of the model on the German data set. The class variable is gender. The group variable is Housing which has three categories: rent, own, free.

First, pre-training, we find that demographic disparity (DD) is 0.076, biased against women. When stratified into the three housing groups, we see that DD is 0.043, 0.057, 0.125, respectively, suggesting that most of the bias against women comes from the free housing sector. So, we compute $CDDL$, i.e., the group wise weighted average of DD, and this works out to 0.062, therefore, $CDDL < DD$.

Second, we trained the loan prediction model using logistic regression and xgboost, and then calculated the DD and $CDDL$ values post-training. For logistic regression, $DD = 0.119$ and $CDDPL = 0.094$, where the three group values of DD are 0.171, 0.084, 0.0382. We see that most of the bias in the model changes to emanating from the rent housing category. For xgboost, $DD = 0.107$ and $CDDPL = 0.079$, where the three group values of DD are 0.088, 0.072, 0.111. We see that most of the bias in this model comes from the free housing category. $CDDPL$ therefore, not only offers a check for the existence of Simpson's paradox, but it also drills down into the subgroups in which bias may disproportionately occur, in the pre-training dataset and in the model as seen in post-training outcomes.

Table 3. Accuracy and Bias Metrics. The table shows the bias metrics generated from the data in the confusion matrices for the two classes. The metrics are: difference in positive prediction proportions $DPPL$; difference in conditional acceptance DCA; difference in conditional rejection DCR; difference in acceptance rates DAR; difference in rejection rates DRR; accuracy difference AD. The top third of the table shows the confusion matrices with a probability cutoff of 0.5. The middle third gives the results when we move the cutoff to 0.55. The bottom of the table shows accuracy and fairness measures for both cutoffs.

$H = 0.50$:

Class a	ACTUAL		
PREDICTED	0	1	Total
0	20	5	25
1	10	65	75
Total	30	70	100

Class d	ACTUAL		
PREDICTED	0	1	Total
0	18	7	25
1	5	20	25
Total	23	27	50

$H = 0.55$:

Class a	ACTUAL		
PREDICTED	0	1	Total
0	25	15	40
1	5	55	60
Total	30	70	100

Class d	ACTUAL	
PREDICTED 0	1	Total
0	20	9	29
1	3	18	21
Total	23	27	50

Metric	$H = 0.50$	$H = 0.55$
Accuracy	0.8200	0.7867
$DPPL$	0.2500	0.1800
DCA	-0.1467	-0.1190
DCR	-0.2800	0.0431
DAR	0.0667	0.0200
DRR	0.0800	0.0595
AD	0.0900	-0.0647

7.5. Explainability

While this paper focuses on fairness, we should not ignore the interplay between fairness and model explainability. Using these features related to protected characteristics in a lending model may be considered discrimination. We may detect these features using explainability models, such as the popular SHAP model, see Lundberg and Lee (2017), Lundberg et al. (2018). Using the German dataset, we obtain the top ten features by Shapley importance, shown in Figure 9. The most important determinants of whether the borrower would be granted credit are whether or not she has a checking account, the loan maturity, and loan amount. In order to check for fairness, we correlate the top features with the gender variable, shown on the right side of Figure 9. As we see, the correlations are low, except for the Age variable, which is negatively correlated with the gender, i.e., women borrowers are younger than men borrowers, on average. However, this may even suggest that women are being discriminated against for being younger rather than on gender. There may thus be a direct effect of Age on loan approval and only an indirect effect of gender. This is where the causality-based framework of Janzing et al. (2019) comes in handy. In our example here, though, we retained the gender variable in our feature set, and since it does not appear in the list of top Shapley features, it is unlikely that unfairness on gender exists, though this is not sufficient condition to exclude gender bias.

7.6. Disparate Impact and Treatment

Disparate impact occurs when policies are equally applied but still impact the disadvantaged group more than the advantaged one. For example, a policy of lending to people who exceed an income threshold has a disparate impact on minorities, whose income levels tend to be lower than that of the average borrower in the US. This would be detected in our $DPPL = \hat{q}_a - \hat{q}_d$ metric. Legally, this is checked by computing the measure $DI = \hat{q}_d/\hat{q}_a \geq 0.80$, i.e., the 80% rule. Because of the existence of disparate impact laws, it is important that all financial FAML be cognizant of the metrics in Section 3.

Discrimination via disparate treatment is illegal and occurs when the lender discriminates based on a protected characteristic. This may be done through the model, but also by the human in the loop post-model decision. It may also occur when a proxy for the protected characteristic is used, e.g., zip code for race, as the two are highly correlated, which forms the basis for redlining. Both DCA and DCR are metrics that may indicate disparate treatment.

The outputs from the pipeline may also be used to determine whether predatory lending exists. We may find that reverse bias exists, i.e., the disadvantaged class was given proportionately more loans, i.e., $DPL < 0$ and $DI > 1.0$, but at a higher interest rate. So the model may appear to be fair in terms of loan proportions, but one group is being taken advantage of via usurious loan rates. In our example, there is no evidence of predatory lending on women, because $DPL > 0$, $DPPL > 0$, $DI < 1$, and the correlation of the proxy for interest rates (i.e., Instalment percent of income) is negatively correlated with gender, i.e., lower for women. There are several other ways in which one group may be taken advantage of such as hidden fees, unclear disclosures, encouraging borrowers to over-leverage, requiring additional legal processes, etc. It is hard to detect these other forms of unfairness in the lending process, but many of our metrics may provide warning signs for these practices.

7.7. Training vs Testing?

Accuracy of ML models is not assessed on the training set; we are interested in the performance "in the wild," not on training data, and we should be more interested in the behavior of the model on unseen data (test/validation set). Does this also apply to fairness metrics? Whereas the metrics we propose here apply equally well to any data set (training, validation, or testing), it is important to stipulate at which point in the pipeline we will impose fairness computation.

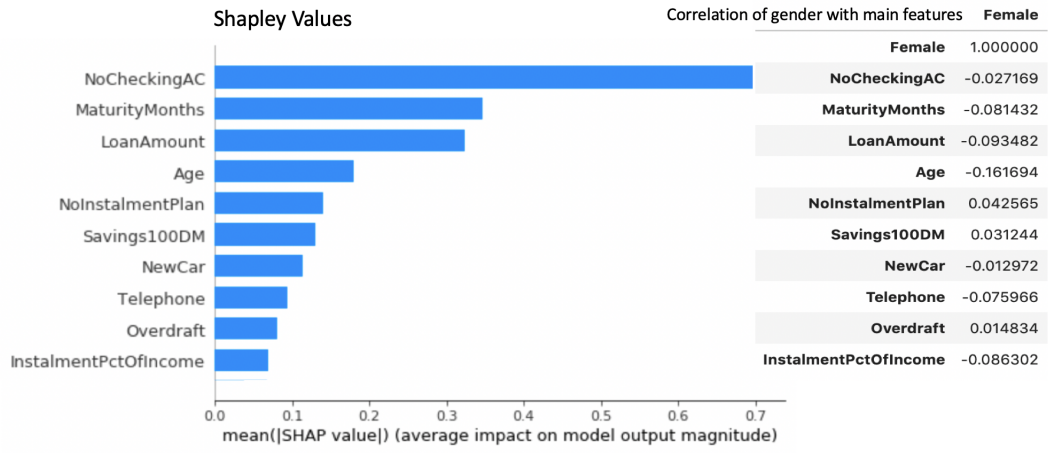

Figure 9. Shapley values for the xgboost model. We show the top ten features and also the correlation of these features in the data with the gender variable.

This is a practical issue that all financial services companies will need to address in their model development and deployment.

Fortunately, the logical timing of fairness computation is easily determined. For pre-training bias measurement ("data analysis") we believe it is best to look at the entire labeled dataset (train + validation + test data combined). The reasons are: (1) we would like to measure the representativeness of the available labeled dataset to determine if we need to collect more data or do reweighting/resampling; (2) we would like to measure any biases with labeling itself, by measuring the differences in the label distribution across different classes. As our goal is to detect these two issues on the dataset, we can measure them before the train/validation/test split. If the train/validation/test split was done randomly, the observed characteristics are likely to hold on each of these as well. If at all, we may flag when the above measures deviate significantly on the test data, which may suggest that the test data distribution is different from that of the overall dataset.

For post-training bias measurement ("model analysis"), we recommend focusing on the test dataset. The reason is that the test data has not been used as part of the training/tuning stages, and hence it is better to compute bias metrics and the effect of bias mitigation over this dataset. In particular, we can use the bias measures computed over the validation data as part of bias mitigation algorithms.

Turning to explainability, for global explanations of the model, we recommend applying explainability algorithms like SHAP over the test data. This is better from the computational angle as well since the test data is usually smaller in size.

8. Concluding Discussion

This paper discusses a family of fairness functions that may be applied to machine learning models for the financial services industry. There is a growing base of model-driven lending and credit card activity, implemented by established firms and startups. These firms are all required to adhere to legal regulations imposing fairness (see Appendix 7.2). Are automated machine learning lending models more or less fair than traditional ones? Bartlett et al. (2017) find that FinTech algorithms statistically discriminate just as face-to-face lenders do, but to an extent that is 40% less.

We provide a large number of fairness metrics, applicable before model training and after. We link these measures to the legal environment and attempt to provide a taxonomy using standard nomenclature. Simple bias mitigation approaches are also offered and a ML pipeline is described, illustrating how a lender may arrive at a model that is both accurate and fair. This can be done in an automated way or with a human in the loop.

ML models entail an accuracy-fairness tradeoff. We provide an example of this tradeoff and also show how there may be a tension amongst the fairness metrics, where an attempt to reduce bias on one measure may lead to exacerbation of bias on another. At the simplest level we can optimize a model for accuracy and check fairness measures and then adjust the model for mitigation and retrain, through an iterative process. In general, given a machine learning model, we can apply fairness metrics as constraints (i) after training and/or (ii) during training. We note that the former is an iterated approach to implementing the latter. This is because in a classification scheme, we train the model, compute bias metrics, and then use this information to decide the next step to improve the fairness of our model, until we are satisfied that the bias levels are acceptable.

However, we may want to train our models for the highest level of accuracy (optimization) while simultaneously ensuring a high level of fairness (regularization). Therefore, as future work, we will extend our thinking of FAML to regularization during or after optimization (where the third option, i.e., before optimization, is part of the pre-processing of the data, without considering the predictive model itself).

An interesting application of this idea is the following: using machine learning methods requires us to choose certain configuration parameters, which have to be selected before the optimization of our model. Usually these parameters are called hyper-parameters and they are picked following different validation procedures, i.e., sequentially repeating the training of our model with different configurations, and selecting the one that is more promising with respect to our goal (in our case, a model that is accurate and fair). The validation procedure can be made more efficient by applying hyper-parameter optimization (HPO) techniques (Feurer and Hutter (2019)) such as those based on Bayesian Optimization (BO) methods.

BO is a class of global optimization algorithms for minimizing expensive-to-evaluate objective functions. Sequential BO is performed by building a surrogate model of the objective function and then sampling a subsequent point determined from an acquisition function criterion. This is repeated until either the given budget is exhausted or the user terminates the loop. Interestingly, it is possible to enforce a fairness constraint on the BO framework in order to optimize black-box models for performance subject to specified fairness metrics. In this sense, the constrained BO technique is able to explore areas of the space of the configurations with higher probability of having accurate and fair models simultaneously. It has recently been demonstrated that accurate and fair models can be obtained by tuning a machine learning model's hyperparameters using this approach (Perrone et al. (2020)).

Although equalizing error rates is an intuitive and well-studied group fairness notion, it may be desirable in some application settings to minimize the largest group error rate. As an example, for a lending application scenario wherein most or even all of the targeted population is disadvantaged, it may be desirable to ensure that the group with the largest error has as less error as possible. Such approaches have been explored as part of the recently introduced minimax group fairness framework (Martinez et al. (2020); Diana et al. (2020)).

Further, it is important to mention that we may also consider causal models for determining bias, as noted in Janzing et al. (2019). Causal models are much harder to implement, for they often require additional information to be brought into the analysis and also require stronger statistical conditions to be met. Causal models may also tease out the difference between disparate impact and disparate treatment.

Finally, we recognize that the notions of bias and fairness are highly application dependent and that the choice of the attribute(s) for which bias is to be measured, as well as the choice of the bias metrics, may need to be guided by social, legal, and other non-technical considerations. Building consensus and achieving collaboration across key stakeholders (such as product, policy, legal, engineering, and AI/ML teams, as well as end users and communities) is a prerequisite for the successful adoption of fairness-aware ML approaches in practice (Bird et al. (2019)).

References

Alesina, A. F., F. Lotti, and P. E. Mistrulli (2013). Do Women Pay More for Credit? Evidence from Italy. *Journal of the European Economic Association 11*(s1), 45–66.

Arjovsky, M., S. Chintala, and L. Bottou (2017, December). Wasserstein GAN. *arXiv:1701.07875 [cs, stat]*. arXiv: 1701.07875.

Barocas, S., M. Hardt, and A. Narayanan (2019). *Fairness and Machine Learning*. fairmlbook.org.

Barocas, S., A. D. Selbst, and M. Raghavan (2020, January). The Hidden Assumptions Behind Counterfactual Explanations and Principal Reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, Barcelona, Spain, pp. 80–89. Association for Computing Machinery.

Bartlett, R. P., A. Morse, R. Stanton, and N. Wallace (2017, December). Consumer Lending Discrimination in the FinTech Era. SSRN Scholarly Paper ID 3063448, Social Science Research Network, Rochester, NY.

Berk, R., H. Heidari, S. Jabbari, M. Kearns, and A. Roth (2017, May). Fairness in Criminal Justice Risk Assessments: The State of the Art. *arXiv:1703.09207 [stat]*. arXiv: 1703.09207 version: 2.

Bickel, P. J., E. A. Hammel, and J. W. O'Connell (1975, February). Sex Bias in Graduate Admissions: Data from Berkeley. *Science 187*(4175), 398–404. Publisher: American Association for the Advancement of Science Section: Articles.

Bird, S., B. Hutchinson, K. Kenthapadi, E. Kiciman, and M. Mitchell (2019). Fairness-Aware Machine Learning: Practical Challenges and Lessons Learned. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3205–3206.

Black, E., S. Yeom, and M. Fredrikson (2020, January). FlipTest: Fairness Testing via Optimal Transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAccT '20, Barcelona, Spain, pp. 111–121. Association for Computing Machinery.

Brock, J. M. and R. D. Haas (2019, October). Gender discrimination in small business lending: Evidence from a lab-in-the-field experiment in Turkey.

Calders, T. and S. Verwer (2010, September). Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery 21*(2), 277–292.

Canetti, R., A. Cohen, N. Dikkala, G. Ramnarayan, S. Scheffler, and A. Smith (2019, January). From Soft Classifiers to Hard Decisions: How fair can we be? *arXiv:1810.02003 [cs, stat]*. arXiv: 1810.02003.

Chen, D., X. Li, and F. Lai (2017, December). Gender discrimination in online peer-to-peer credit lending: evidence from a lending platform in China. *Electronic Commerce Research 17*(4), 553–583.

Corbett-Davies, S. and S. Goel (2018, July). The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *arXiv:1808.00023 [cs]*. arXiv: 1808.00023.

Diana, E., W. Gill, M. Kearns, K. Kenthapadi, and A. Roth (2020). Convergent Algorithms for (Relaxed) Minimax Fairness. *arXiv preprint arXiv:2011.03108*.

Fazelpour, S. and Z. C. Lipton (2020, February). Algorithmic Fairness from a Non-ideal Perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, New York, NY, USA, pp. 57–63. Association for Computing Machinery.

Feurer, M. and F. Hutter (2019). Hyperparameter Optimization. In F. Hutter, L. Kotthoff, and J. Vanschoren (Eds.), *Automated Machine Learning: Methods, Systems, Challenges*, The Springer Series on Challenges in Machine Learning, pp. 3–33. Cham: Springer International Publishing.

Galhotra, S., Y. Brun, and A. Meliou (2017). Fairness Testing: Testing Software for Discrimination. *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering - ESEC/FSE 2017*, 498–510. arXiv: 1709.03221.

Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio (2014, December). Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, Montreal, Canada, pp. 2672–2680. MIT Press.

Hutchinson, B. and M. Mitchell (2019). 50 Years of Test (Un)fairness: Lessons for Machine Learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, 49–58. arXiv: 1811.10104.

Iris, B. (2016). *What Works: Gender Equality by Design*. Harvard University Press.

Janzing, D., K. Budhathoki, L. Minorics, and P. Blöbaum (2019, December). Causal structure based root cause analysis of outliers. *arXiv:1912.02724 [cs, math, stat]*. arXiv: 1912.02724.

Kallus, N., X. Mao, and A. Zhou (2020, January). Assessing Algorithmic Fairness with Unobserved Protected Class using Data Combination. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAccT '20, Barcelona, Spain, pp. 110. Association for Computing Machinery.

Kullback, S. and R. A. Leibler (1951, March). On Information and Sufficiency. *The Annals of Mathematical Statistics 22*(1), 79–86.

Kusner, M. J., J. R. Loftus, C. Russell, and R. Silva (2018, March). Counterfactual Fairness. *arXiv:1703.06856 [cs, stat]*. arXiv: 1703.06856.

Lundberg, S. and S.-I. Lee (2017, November). A Unified Approach to Interpreting Model Predictions. *arXiv:1705.07874 [cs, stat]*. arXiv: 1705.07874.

Lundberg, S. M., B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim, and S.-I. Lee (2018, October). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering 2*(10), 749–760.

Martinez, N., M. Bertran, and G. Sapiro (2020). Minimax Pareto Fairness: A Multi Objective Perspective. In *Proceedings of the 37th International Conference on Machine Learning*. Vienna, Austria, PMLR 119.

Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan (2019, September). A Survey on Bias and Fairness in Machine Learning. *arXiv:1908.09635 [cs]*. arXiv: 1908.09635.

Menestrel, M. L. and L. N. Wassenhove (2016). Subjectively biased objective functions. *EURO Journal on Decision Processes 4*(1), 73–83. Publisher: Springer.

Menzies, P. and H. Beebee (2019). Counterfactual Theories of Causation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2019 ed.). Metaphysics Research Lab, Stanford University.

O'Neil, C. (2016, September). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Reprint edition ed.). Broadway Books.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys 3*, 96–146.

Perrone, V., M. Donini, K. Kenthapadi, and C. Archambeau (2020). Fair Bayesian Optimization. In *ICML AutoML Workshop*.

Simpson, E. H. (1951). The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society. Series B (Methodological) 13*(2), 238–241. Publisher: [Royal Statistical Society, Wiley].

Vasudevan, S. and K. Kenthapadi (2020). LiFT: A scalable framework for measuring fairness in ml applications. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, CIKM '20.

Wachter, S., B. Mittelstadt, and C. Russell (2020, March). Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI. SSRN Scholarly Paper ID 3547922, Social Science Research Network, Rochester, NY.

Wauthier, F. L. and M. I. Jordan (2011). Bayesian Bias Mitigation for Crowdsourcing. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 24*, pp. 1800–1808. Curran Associates, Inc.

Table 4. Fairness questions that may arise in a financial ML model. This table provides some example questions that arise while checking for bias. This list is indicative and not exhaustive. It is a checklist of questions that may be helpful to a financial practitioner in order to gain guidance on which metric is most applicable to the model.

Example Questions	Metrics	Related to
Is the data for each class balanced? Signals that bias may occur in the trained ML model if there is not enough data from a Facet. Example: in small business lending, datasets may contain very few loans to one group, resulting in a trained model that may disfavor that group.	CI	-
Is there a big disparity in outcomes in the dataset across the classes? Indicates possible bias in the dataset if there is a big imbalance in labels across classes. Example: the rates at which various races are granted parole is very different.	DPL, KL, JS, LP, TVD, KS	pre-training demographic parity
Is a certain facet a bigger proportion of the rejected outcomes than the proportion of accepted outcomes? Note: just this question would lead to an answer to whether demographic disparity exists. The conditioning on an attribute is needed to rule out Simpson's paradox. Conditional Demographic Disparity (CDDL or CDDPL in this paper) is explained in the paper by Wachter et al. (2020). The example arises in the classic case of Berkeley admissions where men were accepted at a 44% rate and women at a 36% rate overall, but then this metric was examined department by department, the women on average were admitted in higher proportions. This is because more women applied to departments with lower acceptance rates than men did, though even in these departments, women were accepted at a higher rate than men.	$CDDL$, $CDDPL$	demographic disparity (not to be confused with demographic parity); relates to direct discrimination where a person is treated unequally based on a protected characteristic, and also to indirect discrimination where a policy that appears to be fair, impacts one class more adversely than others.
Does the model give proportionately more loans to one class than another? Note: (1) this ignores completely that one class may be more qualified than the other (as suggested by the observed labels). This metric demands parity irrespective of the fitness of each class. (2) Interestingly, a model that is purely random, i.e., has accuracy=0.5 in a binary label setting will automatically be fair by DPPL.	$DPPL$	mean difference, post-training demographic parity, statistical parity, disparate impact, group discrimination score
Is the ratio of proportions of loans given to each class the same? Does this satisfy the 80% rule, i.e., the ratio of the minimum proportion to the maximum one should be $>= 0.8$? Note: this is nothing but $DPPL$ in ratio form.	DI	disparate impact, 80% rule.
The model may grant more or less loans to a certain class than what the observed labels suggested (e.g., observed labels contain 50 loans for class 1 and 30 for class 2, predicted labels contain 60 loans for class 1 and 20 for class 2). Is the observed vs. predicted ratio the same for different classes? In the current example, there is a bias against class 2.	DCO, DCA, DCR	disparate treatment
Is the model different across classes in picking up the truly deserving borrowers? This assumes that the observed labels are unbiased. Given this assumption, here is an example: if the number of truly deserving borrowers from class 2 is 60 and the model only recommends that 50 get a loan, and the truly deserving borrowers in class 1 are 50 and the model recommends that 45 get a loan, it is biased against both classes, but it is more biased against class 2. This difference in bias across the classes is measured by RD.	RD	sufficiency, conditional procedure accuracy, false positive rate, success prediction error
Is the model accepting equal proportions of the qualified members of each class? Is the model also rejecting equal proportions of the unqualified members of each class? These are related to ideas of equal opportunity and equalized odds.	DAR, DRR, PD	equality of opportunity, equalized odds, individual fairness, predictive parity
Does the model predict the labels for one class more accurately than for others? Example: If the process for granting loans to under-represented populations is much more noisy than for everyone else, we may be mistreating deserving members of under-represented populations, even when there is no bias according to many of the other measures. This may be indicative of a more insidious form of discrimination. It is related to individual fairness.	AD	disparate treatment, individual fairness
Even if the accuracy across classes is the same, is it the case that errors are more harmful to one class than another? TE measures whether errors are compensating in the same way across classes. Example: 100 people from class 1 and 50 people from class 2 apply for a loan. 8 people from class 1 were wrongly denied a loan and another 6 were wrongly approved. For class 2, 5 were wrongly denied and 2 were wrongly approved. $TE = 0.75$ for class 1 and $TE = 0.40$ for class 2, even though $accuracy = 0.86$ for both groups. (This measure was promoted by Berk et al. (2017), but has issues when FN=0.)	TE	-
Are a small group of people from class 1, matched closely on all features with a person from class 2, paid on average more than the latter? Note: when this measure is applied irrespective of class, we can check if individual fairness is achieved.	FT	counterfactual fairness, individual fairness