Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements

Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu,
Pierre Kreitmann, Jonathan Bischof, Ed H. Chi
{alexbeutel, jilinc, tulsee, hqian, woodruff, cmluu, kreitmann, bischof, edchi}@google.com
Google

Abstract

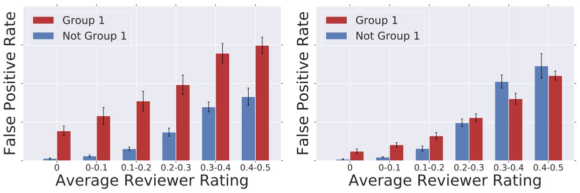
As more researchers have become aware of and passionate about algorithmic fairness, there has been an explosion in papers laying out new metrics, suggesting algorithms to address issues, and calling attention to issues in existing applications of machine learning. This research has greatly expanded our understanding of the concerns and challenges in deploying machine learning, but there has been much less work in seeing how the rubber meets the road.

In this paper we provide a case-study on the application of fairness in machine learning research to a production classification system, and offer new insights in how to measure and address algorithmic fairness issues. We discuss open questions in implementing equality of opportunity and describe our fairness metric, *conditional equality*, that takes into account distributional differences. Further, we provide a new approach to improve on the fairness metric during model training and demonstrate its efficacy in improving performance for a real-world product.

Introduction

By almost every measure, there has been an explosion in attention and research on machine learning fairness: there is a quickly growing amount of research on how to define, measure, and address machine learning fairness, and products are evaluated with these concerns in mind. Despite this significant attention, there has been much less published work detailing how fairness concerns are measured and addressed by product teams in industry. In this paper, we hope to shed light on the challenges in following these principles and learnings in an applied production setting, and to offer metrics and methods developed in the process.

We focus on a classification system where adverse actions are taken against examples predicted to be in the positive class. This is similar to not giving a person a mortgage if a model predicts they will default on it (Hardt et al. 2016), using recidivism prediction for setting bail (Chouldechova 2017), or removing comments on the web if they are predicted to be abusive (Dixon et al. 2017). In all of these cases, each item is associated with a user, and if the classifier makes a mistake and the adverse action is taken against their example, that is bad for the user. More generally, if examples from certain groups of users more often have adverse actions taken against them, it could effect the health of the service.



(a) Before

(b) After Improvements

Figure 1: We observe a significant improvement in the gap in false positive rate by training the model with absolute correlation regularization.

As a result, improving group fairness (Hardt et al. 2016) is both the right thing to do and important to the health of the product.

We focus on equality of opportunity (Hardt et al. 2016), in particular comparing false positive rate (FPR) between groups. While the model being calibrated (Crowson, Atkinson, and Therneau 2016) is an important mathematical property, it does not reflect the experience of users and the implications of representation on the service. However, while (Hardt et al. 2016) provides great intuition and philosophical guidance, we find that in practice it leaves significant wiggle room in how the metric is calculated based on how the evaluation data is sampled or generated. Further, as shown by Corbett-Davies et al. (2017), distributional differences can result in unintended side-effects and costs when implementing fairness changes. We address these issues through a generalized form of the metric, *conditional equality*, that makes these decisions more explicit, and we describe how we navigated these challenges in our use case. Figure 1 presents a summary of our results in an applied production setting[1].

Given this metric, we consider how to improve it under the practical constraints of a product. For example, we are unable to reliably observe the sensitive attribute at inference time, preventing approaches like using different thresholds (Lipton, Chouldechova, and McAuley 2017). Further, as with many engineering systems, simplicity and maintain-

[1] Due to the sensitive nature of these tests, we must omit the numerical values on the y-axis of all plots. In all cases, plots that are juxtaposed keep the same range on the y-axis such that results can be compared.

ability are core requirements. We begin with exploring the use of adversarial training techniques (Beutel et al. 2017a; Zhang, Lemoine, and Mitchell 2018; Madras et al. 2018), which have been shown to be effective. However, as with many adversarial training approaches, we find these are sometimes unstable and difficult to reliably train well. As a result, we offer a new approach, *absolute correlation regularization*, which, while not provably optimal at convergence, we find empirically can stably improve our algorithmic fairness metrics.

Finally, we test these approaches on the production model to improve the metrics for items from two sensitive groups. We find that adversarial training and absolute correlation regularization both improve these metrics significantly.

The open questions of how to design a practical algorithmic fairness metric and how to improve that metric under the constraints of a production system are crucial to putting academic learning into practice in industry. While our approaches are tailored to the application and constraints at hand, we believe they can offer guidance to ML practitioners and call attention to gaps in the current literature that researchers can work to address and practitioners should be mindful of. We list below our contributions:

1. **Metrics:** We demonstrate the challenge in "correctly" measuring equality of opportunity, and describe our metric, conditional equality, that makes practitioner decisions explicit and takes into account varying difficulty of examples across groups.
2. **Optimization:** We offer a new regularization technique, called absolute correlation regularization, to encourage equality of opportunity during training.
3. **Improvement:** We demonstrate improvements to our algorithmic fairness metrics. In particular, we find that traditional modelling, such as through larger models, can improve algorithmic fairness. Second, we find that absolute correlation regularization stably and significantly improves algorithmic fairness metrics.

Background and Related Work

We begin with some background material on algorithmic fairness metrics and relevant related work.

Metrics Many different metrics have been proposed to measure machine learning fairness, particularly for binary classification. One line of work called individual fairness rests on the view that similar examples should receive similar predictions (Dwork et al. 2012b); but this leaves open the question of similarity. Another line of work focuses on group fairness, where examples are grouped by a particular sensitive attribute and statistics about the model predictions are aggregated within the group and compared between groups. Multiple group fairness metrics have been proposed. Demographic parity (Calders and Verwer 2010) asserts that the average prediction for each group should be equal:

$$P(\hat{y} = 1|s = 0) = P(\hat{y} = 1|s = 1), \qquad (1)$$

where the model prediction is $\hat{y}$ and the sensitive attribute (group identity) is given by s. One issue with this view is that different groups could have very different labels y (often called different base rates). Equality of opportunity (Hardt et al. 2016) addresses this by analyzing the accuracy and asserting that the model should not mistake $y = 0$ examples for $y = 1$ examples at a higher rate for one group than another:

$$P(\hat{y} = 0|s = 0, y = 0) = P(\hat{y} = 0|s = 1, y = 0) \qquad (2)$$

Empirically this means that we are comparing the false positive rate (FPR) for examples from each group, which makes sense if a false positives result in a high cost to the group. A symmetric statement can be made for the false negative rate, and putting these together is defined as equality of odds. A third popular group fairness metric has been calibration (Crowson, Atkinson, and Therneau 2016):

$$E[y|s = 0, \hat{y} = p] = E[y|s = 1, \hat{y} = p] \;\; \forall p \in [0, 1] \qquad (3)$$

Significant work has analyzed these metrics and their gaps. A number of results have shown that achieving all of them (or even pairs of them) is only possible in limited cases (Kleinberg, Mullainathan, and Raghavan 2016; Pleiss et al. 2017). Other research has considered expanding them to more complex combinations of multiple sensitive attributes, which we refer to as intersectional testing (Kearns et al. 2017; Hébert-Johnson et al. 2017). Another different line of work has explored using the language of causality to define fairness (Kilbertus et al. 2017), but this has had limited traction (Garg et al. 2018) due to the difficulty of knowing the causal graph.

Modeling With this wide variety of measures of fairness, another line of research has explored how to address algorithmic fairness issues in models. One line of work has built on adversarial training. This approach began for domain adaptation (Ajakan et al. 2014; Ganin et al. 2016; Bousmalis et al. 2016) and was quickly applied to fairness (Zemel et al. 2013; Louizos et al. 2015; Edwards and Storkey 2015). More recent work has modified this to align with different ML fairness metrics (Beutel et al. 2017a; Zhang, Lemoine, and Mitchell 2018; Madras et al. 2018). Others have focused on constrained optimization (Agarwal et al. 2018; Goh et al. 2016), or using a variety of regularization techniques (Kamishima, Akaho, and Sakuma 2011; Zafar et al. 2015; Bechavod and Ligett 2017). Our regularization approach draws from all of these bodies of work.

Other approaches have been advocated for such as using different thresholds for each group during binary classification (Lipton, Chouldechova, and McAuley 2017), but this is not feasible without observing the sensitive attribute at inference. Another approach has been data augmentation (Dixon et al. 2017), but it is often unclear how to do this in applications with more complex, arbitrary feature sets, e.g., loan default prediction over individuals or recommender systems where domain adaptation is difficult.

Application: Much of the published work on addressing fairness concerns focuses on public policy applications like recidivism prediction (Chouldechova 2017; Lum 2017), predictive policing (Lum and Isaac 2016), and child services (Chouldechova et al. 2018). These works explore some related practical difficulties, e.g., (Chouldechova 2017) discusses how metrics could be calculated after conditioning on other covariates like prior convictions. Recently, Holstein et al. (2018) surveyed practitioners on the challenges to improving fairness in industry.

Application Setting

We begin with an overview of our application, and the properties of it that are key to how we define and address any fairness concerns. We focus on a binary classification model that predicts if each example follows or breaks a pre-determined product policy. Examples that break the policy have an adverse action taken directly against them; examples that fall within policy have no action taken against them. This is similar to abuse classification literature (Dixon et al. 2017), the common loan-default prediction problem (Hardt et al. 2016), or recidivism prediction (Chouldechova 2017).

Because of the quantity of data, reviewers cannot rate all examples on the service. Rather, we use human raters to score a subsample of the examples. Human raters give a score $y \in [0, 1]$ and K raters score each example, producing an average ground truth score $y_i = \sum_k \frac{y_{i,k}}{K}$ for example i. We can choose which examples to get rated, but each rating is relatively expensive such that we can only have a small fraction of the data rated. This is particularly exacerbated by the fact that only a small percentage of the examples break policy and as such, random sampling of examples produces relatively little data with high y.

Because most of the examples are unrated, we use a model f to predict the ground truth score: $f(\mathbf{x}_i) = \hat{y}_i \approx y_i$ where $\mathbf{x} \in \mathbb{R}^d$ are features of the example. We take $\mathbf{x}$ to be a general feature vector; in practice, it contains a wide range of features including direct features as well as embedding and signals produced from other models and systems. The model is trained as a regression model with squared error: $\min_f \sum_i (f(\mathbf{x}_i) - y_i)^2$. At inference time, adverse action is taken against all examples with prediction over some threshold: $f(\mathbf{x}_i) = \hat{y}_i > \tau$.

For evaluating algorithmic fairness, we consider group fairness (Hardt et al. 2016) over groups of examples or users. Each example is associated with a user, but features about the users are not reliably observed, i.e., we consider the case where users are not associated with demographic information. Rather, a small number of users are willing to share their demographics, which we can use during training and for offline evaluation of algorithmic fairness metrics. We refer to the group membership by feature s. Because most users do not share their demographic information, we cannot use those features as input to the model or in any way at inference time. Further, because the number of examples with demographic information is relatively small, a more expansive intersectional evaluation is not feasible (Kearns et al. 2017; Hébert-Johnson et al. 2017).

Assumptions We take the product policy as ground truth. Further, we make the simplifying assumption that the human raters provide an unbiased estimate of the ground truth score. At the present, this is difficult to evaluate and further research is needed on how to detect and evaluate rater bias. We offer an expanded discussion of the assumptions and limitations of our analysis at the end of the paper.

Baseline Model

We consider now how our baseline model performs, in particular the FPR. This original model is a linear model over diverse feature set $\mathbf{x}$. We consider the FPR for two important sensitive groups, which we will refer to as Group 1 and Group 2. In each case, we compare the FPR to that of examples not from that respective user group, i.e., Not-Group 1 and Not-Group 2. Due to the sensitive nature of the measurements, we present all results in relativistic terms. For example, the FPR Ratio is defined as:

$$\text{FPR Ratio}_{\text{Group1}} = \frac{\text{FPR}_{\text{Group1}}}{\text{FPR}_{\text{Not}-\text{Group1}}} \tag{4}$$

We present these measurements of the original system in Figure 3a and 4a. As observed in the plots, we find that for both Group 1 and Group 2 the FPR Ratio > 1, with FPR Ratio$_{\text{Group1}} = 5.3\times$ and FPR Ratio$_{\text{Group2}} = 2.18\times$. Both of these numbers indicate a gap in the equality of opportunity and that examples from these groups are more frequently incorrectly having adverse action taken against them. Although hard to measure due to limited data, we generally observe lower FNR for subgroups than the majority, and because adverse actions are taken for predicted positives, we focus on the FPR in the rest of our analysis.

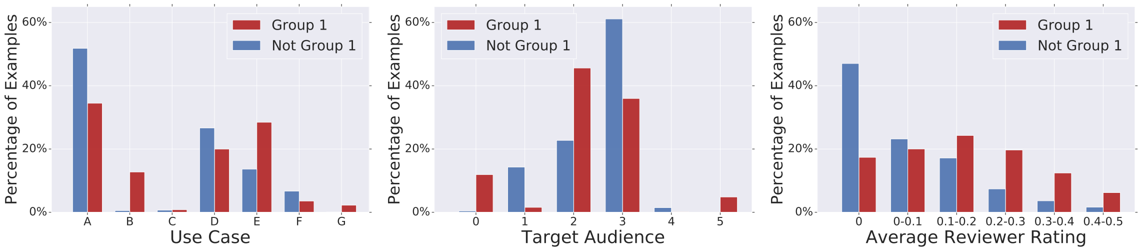
ML Fairness Metric

While equality of opportunity (Hardt et al. 2016) provides insight into measuring cost to users of mistakes across groups, it leaves many open questions that in practice need to be addressed. In particular, how should the data that the metric is calculated over be sampled? What if the distribution of data differs? How do we address these differences?

Data Distribution

One immediate open question in the analysis and practical implementation is: *how should the data be sampled?* FPR and FNR are *only* meaningful with respect to a given distribution of evaluation data, and we find that different ways of generating that evaluation data give significantly different results. Further, as pointed out by Corbett-Davies et al. (2017), examples from different groups may have different distributions of risk, and directly applying equality of opportunity over these different risks can raise its own issues.

The analysis above is based on a dataset built by sampling examples proportional to their usage, but ignore many other differences in the data. Unsurprisingly, different groups of users are associated with different types of examples, such as with different use cases or target audiences. For example, in Figures 2a and 2b we find that for one particular demographic property, the distribution of both use cases of the examples as well as the target audiences are quite different between the groups (considered over negative examples from each group).



(a) Distribution of Use Cases (b) Distribution of Audience (c) Distribution of Ratings y

Figure 2: Different groups have different distribution of examples.

While the data can be stratified by many of these dimensions, no principled way has been given for how or when to do so. Here, we take inspiration from Corbett-Davies et al. (2017), which suggests the importance of addressing different risk distributions. However, Corbett-Davies et al. (2017) analyze risk through the model's predictions rather than through some externally observable property. Here, we deviate in that we observe a real-valued policy y averaged over multiple human raters. As can be seen in Figure 2c, we find that even within examples that would be considered negatives ($y < \tau$), there is a notably different distribution between groups. In particular, the sensitive subgroup has relatively more examples close to the policy threshold τ, suggesting uncertainty by human raters about how the examples align with the policy.

Distribution-Dependent Metrics

Understanding and addressing these differences in distribution is critical to interpreting the results. Therefore, we begin with laying out a formalization for conditional group fairness metrics and then discuss the reasoning and implications of our choice of what to condition on.

First we build on (Ritov, Sun, and Zhao 2017) and define a conditional group fairness for our case:

Definition 1. *Conditional Equality of Opportunity is defined for conditioning on feature A that takes values $\mathcal{A}$:*

$$\begin{aligned} P(\hat{y} \geq \tau | s = 1, y < \tau, A = a) \\ = P(\hat{y} \geq \tau | s = 0, y < \tau, A = a) \; \forall a \in \mathcal{A} \end{aligned} \tag{5}$$

This definition does not give a concrete metric and leaves the question of how to prioritize different $a \in \mathcal{A}$. We can make this precise by defining the conditional equality of opportunity gap:

Definition 2. *Conditional equality of opportunity gap is defined conditioning over feature A taking values in $\mathcal{A}$, with each gap weighted by a probability p_a for $a \in \mathcal{A}$:*

$$\begin{aligned} EOGap = \sum_a p_a [P(\hat{y} \geq \tau | s = 1, y < \tau, A = a) \\ - P(\hat{y} \geq \tau | s = 0, y < \tau, A = a)] \end{aligned} \tag{6}$$

By setting $p_a = \frac{1}{|\mathcal{A}|}$, this metric equally weights each possible value $a \in \mathcal{A}$. This is generally a good prior absent a strong reason to deviate. However, if one group has a skew in A then the uniform prior may not represent the experience for a user in that group. In this case, another option is to set $p_a = P(A = a | s = 0)$, which aligns with importance weighting the data from the background distribution to match the distribution of the focused subgroup[2].

Crucially, with this definition we still must choose a feature A on which to condition. Conditioning on a particular property, for example the example use case, would have the implication that error rates can be different across use cases, as long as they are the same across groups for the same use case; but if different groups prefer different use cases, then this metric would not necessarily support those group preferences. As discussed above, we believe the averaged human rating addresses a balance of desired properties for our metric. The averaged human rating does not prioritize different use cases, target audiences, etc., but rather we can interpret it as giving us a way of observing the inherent difficulty (or risk as in (Corbett-Davies et al. 2017)) of an example. If humans are uncertain if an example meets a policy, it is understandable for the model to make a mistake as well. As we see in Figure 1, we observe that the FPR does increase with the averaged human rater score, and while we still observe a gap in FPR between the background data and the subgroup, it is partially explained by the difference in the distribution of examples.

Note, these proposals have significant connections to related work. Most related is (Ritov, Sun, and Zhao 2017), which first proposed this generalized fairness metric and included conditioning on arbitrary variables. The similarity function in individual fairness can be thought of as conditioning on different features, and that work grapples with many of these same issues (Dwork et al. 2012a). Chouldechova (2017) and Corbett-Davies et al. (2017) both mention that metrics can be calculated condition on other variables, and for recidivism prediction they condition on the number of prior convictions, but neither give a general framework for when and how to condition. In examining fairness through a causal lens, (Kilbertus et al. 2017) explores the question of what is a "resolving variable," but ultimately this is left as a philosophical choice. Further, our conditional equality metric can be viewed as a special case of the intersectional fairness analysis (Kearns et al. 2017;

[2] When reporting ratios, we compute the ratio of average FPRs to align with this view of data sampling.

Hébert-Johnson et al. 2017), which conditions on any and all combination of covariates, but as we discussed, this requires a significant amount of data. Ultimately, all of these approaches leave the question of how to interpret conditioning on different terms.

We hope that by forcing the definition of the fairness metric to specify the conditioned variables, practitioners consider the distribution of their evaluation data, how it was sampled, and the implications. Unfortunately, this still does not diminish the importance of considering how the dataset is sampled or generated, and requires careful consideration by practitioners in determining this procedure.

Correlation Loss

Although previous work has laid out multiple techniques for training models with fairness metrics as part of the objective, these approaches generally come with notable engineering concerns. Here, we present a new, lightweight approach for improving the desired fairness metrics more effectively than previous approaches.

One perspective on group fairness metrics is that the output distribution should match across groups, possibly after reweighting or resampling the data (Beutel et al. 2017a; Madras et al. 2018). Notably, (Zhang, Lemoine, and Mitchell 2018) focuses on the prediction $\hat{y}$ rather than an intermediate layer. One way we could conceptualize this goal is to compare the distributions of $\hat{y}$ between groups and encourage low mutual information or KL divergence. Here, we take a simplified view of that by minimizing the absolute correlation between $\hat{y}$ and the group membership s:

$$\min_f \left[\sum_{(\mathbf{x}_i, y_i) \in \mathcal{X}} L(y_i, f(\mathbf{x}_i)) \right] + \lambda |\mathrm{Corr}_{\mathcal{X}^-}| \tag{7}$$

where

$$Corr_{\mathcal{X}^-} = \frac{(\sum_{\mathbf{x}_i \in \mathcal{X}^-} f(\mathbf{x}_i) - \mu_{\hat{y}})(\sum_{s_i \in \mathcal{X}^-} s_i - \mu_s)}{\sigma_{\hat{y}} \sigma_s}$$

$$\mu_{\hat{y}} = \frac{1}{|\mathcal{X}^-|} \sum_{\mathbf{x}_i \in \mathcal{X}^-} f(\mathbf{x}_i) \qquad \mu_s = \frac{1}{|\mathcal{X}^-|} \sum_{s_i \in \mathcal{X}^-} s_i$$

$$\sigma_{\hat{y}}^2 = \frac{1}{|\mathcal{X}^-|} \sum_{\mathbf{x}_i \in \mathcal{X}^-} (f(\mathbf{x}_i) - \mu_{\hat{y}})^2$$

$$\sigma_s^2 = \frac{1}{|\mathcal{X}^-|} \sum_{s_i \in \mathcal{X}^-} (s_i - \mu_s)^2$$

Following (Beutel et al. 2017a; Madras et al. 2018), we use $\mathcal{X}^- = \{x_i \in \mathcal{X} | y_i < \tau\}$ in order to optimize for equality of opportunity; this could be extended to other reweighting or resampling schemes to address other metrics as in (Madras et al. 2018). In practice, we use $\tilde{\mathcal{X}}^- \subset \mathcal{X}^-$, which is a minibatch of examples sampled from $\mathcal{X}^-$. This follows a similar pattern as previous adversarial approaches in adding a penalty based on the distribution of the output, but unlike all of the adversarial approaches, no training of an adversary is required, which we find to greatly improve stability in practice. While minimizing this term does not provably minimize the fairness metric, we find we get good results in practice, as we will show below.

Improvements in Practice

In practice, we seek to improve both the general equality of opportunity metric as well as the conditional equality metric. All results here build on the description of the application setting, metrics and methods above. In particular, we explore the incremental process by which we worked to improve this classifier, and how each change effects the end metrics. Results are summarized in Figures 3 and 4 for Group 1 and Group 2 respectively, but we will walk through each of the results below. Error bars are based on training the model 10 times and averaging results. We find there are a number of trade-offs and directions for future work.

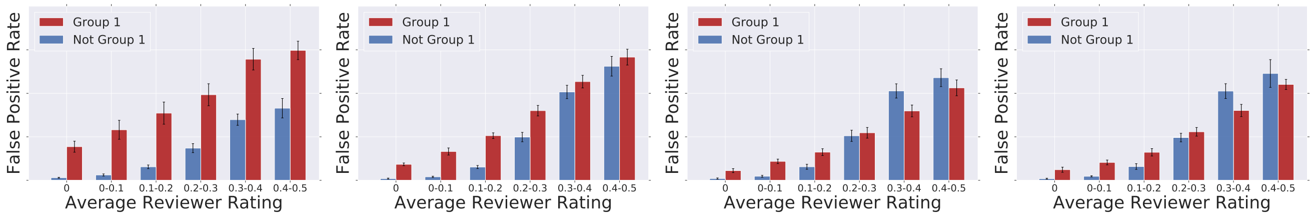
Model Capacity As discussed previously, the original model was a linear model, and we observe a significant gap in FPR between the sensitive subgroup and the rest of the data for both Group 1 and Group 2, as seen in Figures 3a and 4a. The first step we took was to change from a linear classifier to a DNN with a single hidden layer. Previous work (Chen, Johansson, and Sontag 2018) has suggested that theoretically model capacity can be a driver of disparities in accuracy. As can be seen in Figure 3b, the move to DNN decreases the FPR for Group 1, as well as decreasing the FPR gap from $2.62\times$ to $1.44\times$. While this is good to see, this does not always hold true: we observe the FPR *increases* for Group 2 and Not-Group 2 users[3]; the FPR gap decreases from $1.76\times$ to $1.25\times$. As such, while increasing model capacity may help in some cases, it does not necessarily improve accuracy everywhere.

Adversarial training Building on the DNN model, we next considered how well adversarial training can improve the FPR gap. We follow an approach similar to Beutel et al. (2017a) of training an additional head taking as an input the last hidden layer of the model and trying to predict the sensitive attribute s while the model tries to learn a representation that is independent of s; we use only data for which $y < \tau$, as we are concerned with the FPR. As we see in Figure 3c, this significantly decreases the FPR gap from $1.44\times$ to $1.04\times$ and simultaneously decreases the FPR for Group 1, despite that not being an explicit objective.

Correlation Loss As mentioned previously, adversarial training has been well studied and we see has strong performance, but from an engineering perspective is challenging due to its instability during training. As a result, we pursued absolute correlation regularization to stabilize training. We observe in Figure 3d that using absolute correlation regularization keeps the FPR gap approximately the same $(1.05\times)$. The practical benefits of keeping a low FPR gap while improving stability is highly valuable in practice.

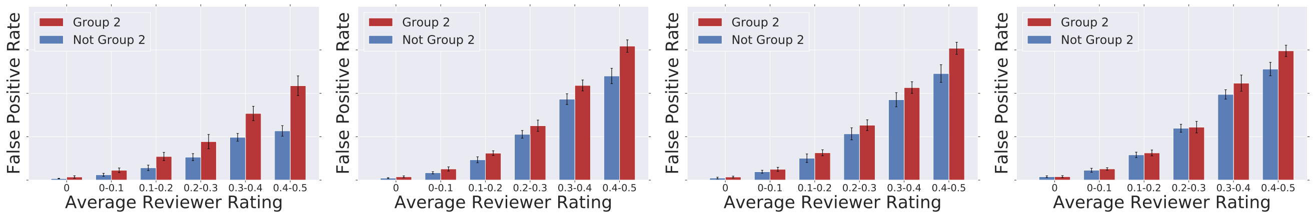
Transfer across Groups One idea that has been debated in the literature (Madras et al. 2018) is if and when there is transfer learning of fairness across groups. We consider here

[3] Note, Not-Group 1 and Not-Group 2 are different in that they are based on different methods for getting demographic data s.



(a) Linear (b) DNN (c) DNN with Adversary (d) DNN with corr. reg.

Figure 3: Improvements for Group 1 users.



(a) Linear (b) DNN (c) DNN + Group 1 corr. reg. (d) DNN + Both corr. reg.

Figure 4: Improvements for Group 2 users.

how the application of absolute correlation regularization to Group 1 effects the FPR gap for Group 2 users. In Figure 4c we observe a very slight improvement in the FPR gap, bringing it down from 1.37× to 1.31×.

Improving for Multiple Groups Finally, we look at if we can simultaneously improve the FPR gap for both Group 1 and Group 2 users simultaneously. To do this, we add two different absolute correlation regularization terms to the DNN training, one for each group. As we see in Figure 4d, we are able to improve the FPR gap for Group 2 to 1.11×. Unfortunately, we do not find that this decreases the FPR for Group 2; we believe focusing on making the model more inclusive by improving the accuracy not just decreasing the gap is an important step for future work (Beutel et al. 2017b; Chen, Johansson, and Sontag 2018).

Future Directions

We have focused on how to improve an individual model that directly effects the user experience, but by no means does every use of machine learning fit into these settings. To expand the applicability, we believe there are a number of areas that deserve more research attention.

Human raters: This work, like most of the algorithmic fairness literature, assumes that the labels are unbiased. We believe more attention is needed to understand if and when there is bias in crowd-sourced ratings, and how to remove it.

Binary actions: We only consider the case where we are taking binary actions directly against the examples (at the known threshold). When the prediction is treated as a continuous score or used in conjunction with other signals, it becomes harder to evaluate the effect on the user experience (Dwork and Ilvento 2018) and further research is needed in this direction.

Observed Examples: We evaluate our system against the examples currently in the system. However, that distribution is of course affected by the system's previous performance. Use cases that were previously not well supported may be underrepresented in our sample. Unfortunately, we do not know of a way to infer the distribution of examples that would exist under a different previous system. As a result, for the time being, we focus on evaluating and improving metrics for the current state, with the belief that it will improve the performance of the system for the sensitive subgroups and we can continue to evaluate the performance as the subgroup evolves. Further research, along the lines of (Liu et al. 2018), would be valuable to better understand the long term evolution of the system.

Discussion

In this work we have offered details on how applying algorithmic fairness principles to a production classifier fares. In particular, we have explored how equality of opportunity depends on how the data is sampled, and how different groups can have notably different distributions of data. To address issues with these distributional differences, we offered a general evaluation approach that takes into account example difficulty. Further, we have offered a new mechanism, absolute correlation regularization, for improving algorithmic fairness metrics that we find to be more stable than adversarial training. We demonstrate the ability of these approaches to improve the FPR gap for two different groups in a production classifier and offer analysis of how the model performance is effected by these different training procedures.

Acknowledgements: We would like to thank Margaret Mitchell for her valuable feedback on this paper.

References

[Agarwal et al. 2018] Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. M. 2018. A reductions approach to fair classification. *CoRR* abs/1803.02453.

[Ajakan et al. 2014] Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; and Marchand, M. 2014. Domain-adversarial neural networks. *CoRR* abs/1412.4446.

[Bechavod and Ligett 2017] Bechavod, Y., and Ligett, K. 2017. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*.

[Beutel et al. 2017a] Beutel, A.; Chen, J.; Zhao, Z.; and Chi, E. H. 2017a. Data decisions and theoretical implications when adversarially learning fair representations. *CoRR* abs/1707.00075.

[Beutel et al. 2017b] Beutel, A.; Chi, E. H.; Cheng, Z.; Pham, H.; and Anderson, J. 2017b. Beyond globally optimal: Focused learning for improved recommendations. In *Proceedings of the 26th International Conference on World Wide Web*, 203–212. International World Wide Web Conferences Steering Committee.

[Bousmalis et al. 2016] Bousmalis, K.; Trigeorgis, G.; Silberman, N.; Krishnan, D.; and Erhan, D. 2016. Domain separation networks. In *Advances in Neural Information Processing Systems*, 343–351.

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

[Chen, Johansson, and Sontag 2018] Chen, I.; Johansson, F. D.; and Sontag, D. 2018. Why is my classifier discriminatory? *arXiv preprint arXiv:1805.12002*.

[Chouldechova et al. 2018] Chouldechova, A.; Benavides-Prado, D.; Fialko, O.; and Vaithianathan, R. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, 134–148.

[Chouldechova 2017] Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5(2):153–163.

[Corbett-Davies et al. 2017] Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 797–806. ACM.

[Crowson, Atkinson, and Therneau 2016] Crowson, C. S.; Atkinson, E. J.; and Therneau, T. M. 2016. Assessing calibration of prognostic risk scores. *Statistical methods in medical research* 25(4):1692–1706.

[Dixon et al. 2017] Dixon, L.; Li, J.; Sorensen, J.; Thain, N.; and Vasserman, L. 2017. Measuring and mitigating unintended bias in text classification.

[Dwork and Ilvento 2018] Dwork, C., and Ilvento, C. 2018. Fairness under composition. *CoRR* abs/1806.06122.

[Dwork et al. 2012a] Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012a. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226. ACM.

[Dwork et al. 2012b] Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. S. 2012b. Fairness through awareness. In *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, 214–226.

[Edwards and Storkey 2015] Edwards, H., and Storkey, A. 2015. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*.

[Ganin et al. 2016] Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research* 17(59):1–35.

[Garg et al. 2018] Garg, S.; Perot, V.; Limtiaco, N.; Taly, A.; Chi, E. H.; and Beutel, A. 2018. Counterfactual fairness in text classification through robustness. *arXiv preprint arXiv:1809.10610*.

[Goh et al. 2016] Goh, G.; Cotter, A.; Gupta, M. R.; and Friedlander, M. P. 2016. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems*, 2415–2423.

[Hardt et al. 2016] Hardt, M.; Price, E.; Srebro, N.; et al. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 3315–3323.

[Hébert-Johnson et al. 2017] Hébert-Johnson, Ú.; Kim, M. P.; Reingold, O.; and Rothblum, G. N. 2017. Calibration for the (computationally-identifiable) masses. *CoRR* abs/1711.08513.

[Holstein et al. 2018] Holstein, K.; Vaughan, J. W.; Daumé III, H.; Dudík, M.; and Wallach, H. 2018. Improving fairness in machine learning systems: What do industry practitioners need? *arXiv preprint arXiv:1812.05239*.

[Kamishima, Akaho, and Sakuma 2011] Kamishima, T.; Akaho, S.; and Sakuma, J. 2011. Fairness-aware learning through regularization approach. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, 643–650. IEEE.

[Kearns et al. 2017] Kearns, M. J.; Neel, S.; Roth, A.; and Wu, Z. S. 2017. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *CoRR* abs/1711.05144.

[Kilbertus et al. 2017] Kilbertus, N.; Rojas-Carulla, M.; Parascandolo, G.; Hardt, M.; Janzing, D.; and Schölkopf, B. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, 656–666.

[Kleinberg, Mullainathan, and Raghavan 2016] Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

[Lipton, Chouldechova, and McAuley 2017] Lipton, Z. C.; Chouldechova, A.; and McAuley, J. 2017. Does mitigating ml's disparate impact require disparate treatment? *arXiv preprint arXiv:1711.07076*.

[Liu et al. 2018] Liu, L. T.; Dean, S.; Rolf, E.; Simchowitz, M.; and Hardt, M. 2018. Delayed impact of fair machine learning. *CoRR* abs/1803.04383.

[Louizos et al. 2015] Louizos, C.; Swersky, K.; Li, Y.; Welling, M.; and Zemel, R. 2015. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*.

[Lum and Isaac 2016] Lum, K., and Isaac, W. 2016. To predict and serve? *Significance* 13(5):14–19.

[Lum 2017] Lum, K. 2017. Limitations of mitigating judicial bias with machine learning. *Nature Human Behaviour* 1.

[Madras et al. 2018] Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. S. 2018. Learning adversarially fair and transferable representations. *CoRR* abs/1802.06309.

[Pleiss et al. 2017] Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J. M.; and Weinberger, K. Q. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems*, 5684–5693.

[Ritov, Sun, and Zhao 2017] Ritov, Y.; Sun, Y.; and Zhao, R. 2017. On conditional parity as a notion of non-discrimination in machine learning. *arXiv preprint arXiv:1706.08519*.

[Zafar et al. 2015] Zafar, M. B.; Valera, I.; Rodriguez, M. G.; and

Gummadi, K. P. 2015. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*.

[Zemel et al. 2013] Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 325–333.

[Zhang, Lemoine, and Mitchell 2018] Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating unwanted biases with adversarial learning. *CoRR* abs/1801.07593.