Exploring the case for parametrised resampling for fairness optimisation driven by nuanced metrics for measuring ML bias

Atif Khan

Newcastle University , UK
a.khan21@ncl.ac.uk

Abstract. The paper by Zelaya et al. [1] introduced a parametrised data sampling technique that can be used to optimise classifier’s prediction performance towards any subgroup within the data in accordance with any preferred fairness definition. This prediction performance is measured using three metrics i.e. DPR, EOR and CFR, which are probability ratios informing classifier’s level of prediction disparity for different subgroups. But Borkan et al. [2] introduced a more comprehensive ML bias measuring metrics they claim will provide a nuanced view of ML bias using classifier’s probability score distributions. In this paper we apply ML bias mitigation technique by [1] on public benchmark dataset (Census Income) and measure classifier’s prediction performance in metrics introduce by [2] and report the results for different values of fairness optimisation parameter ‘d’

Keywords: ML bias measurement.

1 Introduction

Zelaya et al. [1] introduced novel machine learning bias mitigation technique, based on resampling training data instances guided by fairness optimisation parameter ‘d’. The technique is both classifier and fairness definition agnostic, so that “fairness” improvement is as prescribed by user and implemented by setting optimisation parameter ‘d’ to appropriate value which in effect resamples the training dataset accordingly.

Borkan et al. [2] introduced five nuanced metrics for measuring machine learning bias in a classifier. All these metrics are based on prediction probability score distributions of classifier for different subgroups. The metrics are quantified using Area under the Receiver Operating Characteristic Curve (ROC-AUC) and Average Equality Gap (AEG) expressed in Mann-Whitney U metrics. These metrics gives detailed picture of most forms of machine learning bias in any given classifier/model [2].

We used a public benchmark dataset (Census Income) i.e. one of datasets used by [1] and we deliberately choose it for its relative larger size than others. We used “Preferential Sampling” to resample training data instances for different values ($\in [-1, 1]$) of ‘d’ the disparity correction parameter and measure the five nuanced metrics from classifier outputs for each d value.

Rest of the paper is set out as follows, section 2 explains parametrised ML bias mitigation technique by [1] , section 3 explains the five nuanced metrics by [2] , section 4 reports the experiment and results and paper concludes in section 5.

2 Parametrised Training Data Resampling for Fairness Optimisation

[1] The general idea of the technique is to resample training data such that model predictions are optimised to fair outcome i.e. fairness defined by user/application. The process starts by identifying necessary attributes in the data and some fairness definition

- Protected attribute (PA) refers to predictor variable in the dataset that identify the subjects/individuals that can be object of discrimination e.g. gender, race or belief etc. Using PA value, the number of subgroups in the data are determine. The paper [1] discuss application of fairness optimisation only in context of single PA
- Positive Ratio (PR) & Negative Ratio (NR): For binary label/response variable, this refers to ratio of positive instances i.e. label = 1 to total instances and respectively negative instances i.e. label =0 to total instances. This can be calculated for whole dataset or individual subgroups
- Favoured & Unfavoured subgroups (F & U): In context where positive label (=1) refer to positive outcomes like approval of credit card. The favoured group(F) refer to subgroup with highest PR and unfavoured (U) refer to one with lowest PR
- Fairness worldviews: the paper mention three philosophical fairness worldviews “what you see is what you get” (WYSIWUG) i.e. leaving the data as is, “We are all equal” WAAE i.e. all subgroups should be

treated equally and Affirmative correction i.e. reversing the bias amongst the subgroups

The paper proposes parametrised PR correction to impose desired fairness optimisation or a certain worldview. The parametrisation work as below

Let parameter "d" ($\in [-1, 1]$), $f^+(d)$ is function dependent on PR such that

$f^+(1) = PR(F)$, $f^+(0) = PR(D)$, $f^+(-1) = PR(U)$

smoothing function that satisfy above conditions is a quadratic polynomial

$f^+(d) = ad^2+bd+c$; $u^+(d) = ad^2-bd+c$

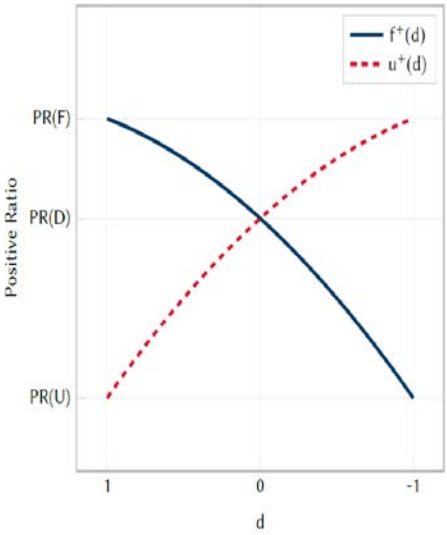
The coefficients of about equation can be expressed in terms of PR

$a= (PR(F)+PR(U))/2 - PR(D)$

$b= (PR(F)-PR(U))/2$, $c= PR(D)$

similarly, $f^-(d)= 1-f^+(d)$; $u^-(d)=1-u^-(d)$

[1]



As expressed above, disparity correction parameter d in terms of PR of F, U & D where F= favoured group, U = unfavoured group & D= whole dataset. These $f^-(d)$, $f^+(d)$, $u^-(d)$ & $u^-(d)$ corrected ratios are used to produce a d-resamples training set to change predictions of classifier as per desired fairness definition

To change or balance the PR & NR of various subgroups the paper [1] discuss four resampling techniques

1. Random Undersampling: random favoured positive & unfavoured negative instances are removed from training set

2. <u>Random Oversampling</u>: random favoured negative & unfavoured positive instances are added/duplicated to the training set

3. <u>SMOTE Oversampling:</u> This too is oversampling technique, but it uses synthetic minority oversampling technique (SMOTE)

4. <u>Preferential Sampling (PS):</u> In this simultaneous oversampling of favoured negatives & unfavoured positives and undersampling of favoured positives & unfavoured negative instances take place

The outcome of this resampling (using any of above sampling techniques) for different d values on prediction performance of classifier is measured in terms of three fairness definitions-based metrics

1. **Demographic Parity**: The probability of being classified as positive should be same across subgroups

   $P\ (Ypred =1\ |subgroup=U) = P\ (Ypred = 1|subgroup=F)$
   This can be converted into ratio
   $DPR = P\ (Ypred =1\ |subgroup=U)\ /\ P\ (Ypred = 1|subgroup=F)$ *this ideally be '1'*

2. **Equality of Opportunity**: the probability of being classified as positive for true positive should be equal across subgroups

   $P\ (Ypred =1\ |subgroup=U, Y=1) = P\ (Ypred = 1|subgroup=F, Y=1)$
   This can be converted into ratio
   $EOR = P\ (Ypred =1\ |subgroup=U, Y=1)\ /\ P\ (Ypred = 1|subgroup=F, Y=1)$ *this ideally be '1'*

3. **Counterfactual Fairness**: A classifier Ypred is counterfactually fair if under all condition predictors X=x & subgroups A=a

   $P\ (Ypred_{A<-U} =1|\ X=x, A=a) = P\ (Ypred_{A<-F} =1|\ X=x, A=a)$

   This can be simplified and defined as follows

$CFR = P\ (Ypred_{A<-\ U} =1|\ X=x,\ A=a)/\ P\ (Ypred_{A<-\ F} =1|\ X=x,\ A=a)$ *this ideally be '1'*

This project is about exploring alternate ML bias measurement metrics to above three. For this an interesting work by Jigsaw team at google promises comprehensive nuanced ML bias measurement metrics in their paper [2]

3 Nuanced Machine Learning Bias Metrics

Borkan et al. [2] whiles working to measure unintended bias of machine learning models/classifiers in incorrectly predicting negatively (or positive for toxicity) text comments about some subgroups (i.e. unfavoured groups). For example, A text comment that just had a word like "gay" or "Muslim" without any negative connotation was still predicted to be toxic. This is because more of the negative instances (or positive for toxicity) in the training set consist of subgroups "gay" or "Muslim" [2]. And ML models usually pick the simplest relationship between predictors and response i.e. instead of relying on swear or abusive words for toxicity it incorrectly uses these subgroups for prediction

To see detailed picture of bias in any ML model they propose five threshold & classifier agnostic metrics. These metrics are based on measuring classifier's prediction probability scores distribution for different subgroups compare to background. Three of these metrics are measured using ROC-AUC i.e. sensitive to ranking of probability scores and other two using Average Equality Gap (AEG) i.e. sensitive to gap between probability score distributions

3.1 Area Under the Receiver Operating Characteristic Curve (ROC-AUC or simply AUC):

ROC: It is the curve drawn between False Positive Rate (FPR- as known as specificity) and True Positive Rate (TPR- as known as sensitivity) for all possible thresholds.

AUC: Is ratio of the area under this curve to total area

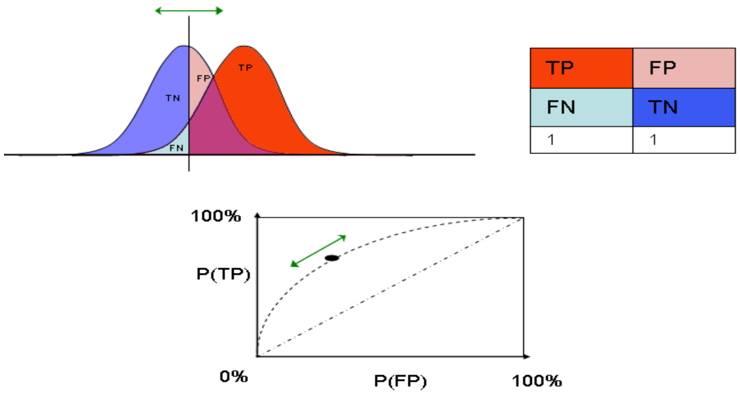


Fig. 1. ROC-AUC illustration [3]

As illustrated above for a classifier with well separated predictive probability distributions for negative & positive classes, the area under the curve will be maximum and vis versa. AUC (∈ [0, 1])

Unlike misclassification metrics like TPR etc. which are threshold dependent AUC is independent of thresholds. Also, AUC is immune to imbalance of populations/sizes of negative & positive class distributions e.g. there might be 1000 samples predicted as positive and just 10 as negative but this will not skew AUC score. But as mention earlier AUC is only sensitive to rank ordering i.e. predicted probability score distribution of negative class should rank lower to positive class for a higher AUC score

In other words, for any model, AUC measure the probability that a randomly chosen negative sample will receive a lower predictive probability score than a randomly chosen positive sample. And a perfect AUC score of 1 refers to case where all negative samples received lower score than all positive samples [2]

The three AUC based nuanced ML bias measurement metrics are 1. Subgroup AUC 2. Background Positive and Subgroup Negative (BPSN) AUC and 3. Background Negative and Subgroup Positive (BNSP) AUC

To formally defines these metrics, let’s consider D to be predicted test dataset, D^+ be positive samples in the background (favoured group i.e. all data minus subgroup or unfavoured group), D^- be negative samples in the background, D_g^+ be positive samples in the subgroup (i.e. unfavoured group) and D_g^- be negative samples in the subgroup[2]

1. **Subgroup AUC**: This will calculate AUC only on test samples from subgroup. This show negative and positive class rank ordering within the subgroup

   $$\text{Subgroup AUC} = \text{AUC}\ (D_g^+ + D_g^-)$$

2. **BPSN AUC**: This will calculate AUC on positive samples in the background and negative samples in the subgroup. A lower BPSN AUC value shows lesser separability between positive background samples and negative subgroup samples i.e. more false positives for subgroup samples

   $$\text{BPSN AUC} = \text{AUC}\ (D^+ + D_g^-)$$

3. **BNSP AUC**: This will calculate AUC on negative samples in the background and positive samples in the subgroup. A lower BNSP AUC value shows lesser separability between negative background samples and positive subgroup samples i.e. more false negatives for subgroup samples

   $$\text{BNSP AUC} = \text{AUC}\ (D^- + D_g^+)$$

As mentioned earlier AUC score is only sensitive to rank ordering but to capture issues related to gap between predictive probability of subgroups [2] introduced two additional metrics based on AEG

3.2 Average Equality Gaps

Equality gap is different between true positive rate (TPR (D_g)) of subgroup and TPR (D) of background at a certain threshold. When this is calculated for all possible values of threshold this is termed as positive Average Equality Gap (Positive AEG). Similarly, if true negative rates are considered instead the counterpart metric is Negative AEG

More formally this can be defined as, *For each threshold t , if you plot the true positive rate of the subgroup as x(t) and the true positive rate of the background as y(t) then the Positive Average Equality Gap is the area between the curve (x(t),y(t)) and the line y = x* [2]

Positive AEG = $\int_0^1 (y(t) - x(t)\, dx(t)$

This can be more generally described in Mann-Whitney U metrics, given two data points (test samples from subgroups) are selected at random one from D_g^+ and other from D^+ given both are positive samples the probability that either predictive probability score is higher than the other should be same

$P\ (Y^{\wedge}_D > Y^{\wedge}_g \mid Y_D \in D^+ , Y_g \in D_g^+) \quad = ½$

Positive AEG $=$ ½ - $P(Y^{\wedge}_D > Y^{\wedge}_g / Y_D \in D^+ , Y_g \in D_g^+)$

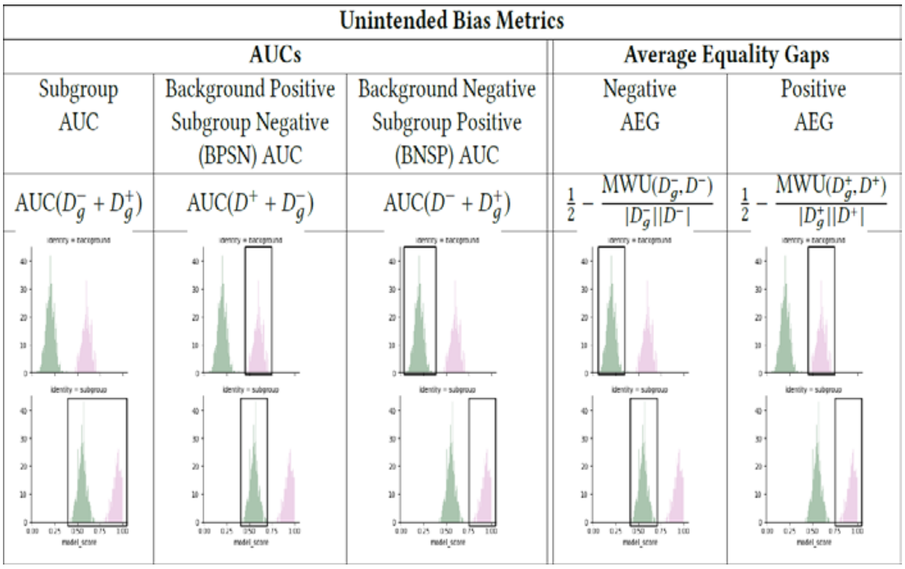
4. **Positive AEG** = ½ - (MWU (D^+, D_g^+)/ $|D_g^+| \, | D^+|$)

the range of Positive AEG [-0.5, 0.5]

5. **Similarly, Negative AEG** = ½ - (MWU (D^-, D_g^-)/ $|D_g^-| \, | D^-|$)

the range of Negative AEG [-0.5, 0.5]

Below figure 2 illustrate the kind of ML biases each of these nuanced metrics can identify

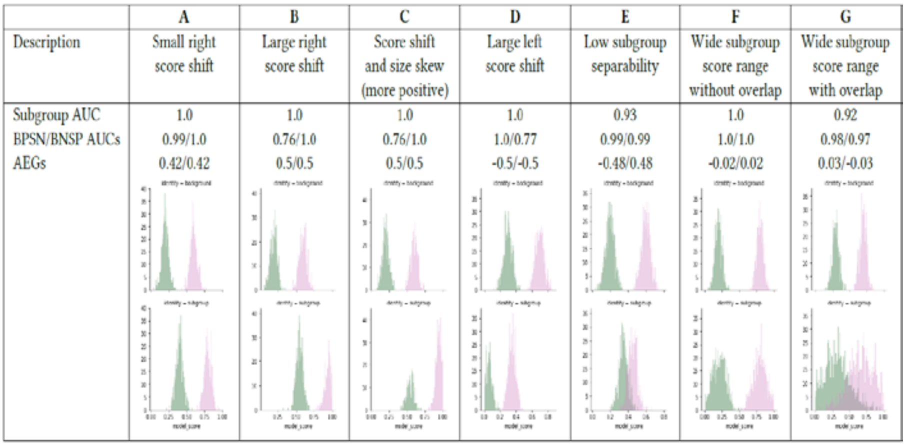


Unintended Bias Metrics				
AUCs			Average Equality Gaps	
Subgroup AUC	Background Positive Subgroup Negative (BPSN) AUC	Background Negative Subgroup Positive (BNSP) AUC	Negative AEG	Positive AEG
$\mathrm{AUC}(D_g^- + D_g^+)$	$\mathrm{AUC}(D^+ + D_g^-)$	$\mathrm{AUC}(D^- + D_g^+)$	$\frac{1}{2} - \frac{\mathrm{MWU}(D_g^-, D^-)}{\lvert D_g^-\rvert\lvert D^-\rvert}$	$\frac{1}{2} - \frac{\mathrm{MWU}(D_g^+, D^+)}{\lvert D_g^+\rvert\lvert D^+\rvert}$

identity = background
identity = subgroup
model_score

Fig. 2. Typical ML biases each metric can identify [2]

[2] also discuss the strengths and weakness of these metrics in identifying some common ML biases. As illustrated, these metrics are good at identify most biases except the case F “Wide subgroup score range without overlap”



	A	B	C	D	E	F	G
Description	Small right score shift	Large right score shift	Score shift and size skew (more positive)	Large left score shift	Low subgroup separability	Wide subgroup score range without overlap	Wide subgroup score range with overlap
Subgroup AUC BPSN/BNSP AUCs AEGs	1.0 0.99/1.0 0.42/0.42	1.0 0.76/1.0 0.5/0.5	1.0 0.76/1.0 0.5/0.5	1.0 1.0/0.77 -0.5/-0.5	0.93 0.99/0.99 -0.48/0.48	1.0 1.0/1.0 -0.02/0.02	0.92 0.98/0.97 0.03/-0.03

Fig. 3. Typical scores for some common ML biases [2]

4 Experiment and Results

Aim: Explore the case for parametrised resampling for fairness optimisation driven by nuanced metrics for measuring ML bias

To attain this aim I have set following objectives

- Select a dataset and divide it into training/test sets
- Ascertain subgroup (unfavoured) & background (favoured) in the data
- Resample the training set for different values of disparity correction parameter ‘d’ using anyone one of sampling techniques mentioned by [1]
- Measure the outcome of each of these d-parametrised resampled set using five nuanced ML bias metrics
- Draw inference from results

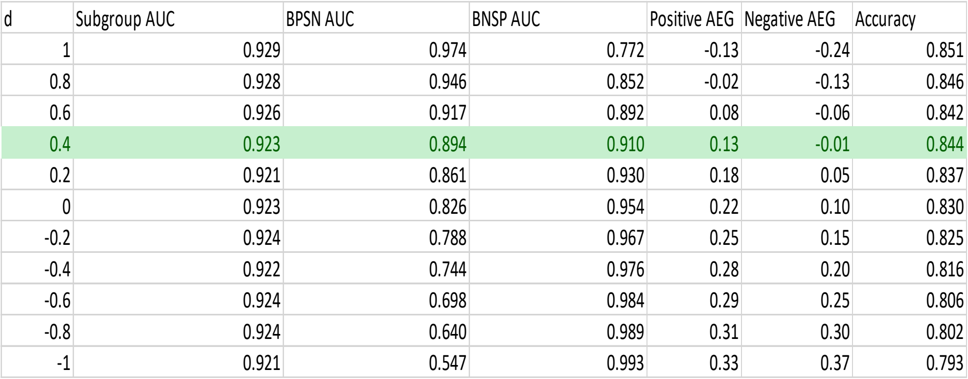
Dataset. I selected “income” dataset for our experiment as used by [1] because of its relatively large size which will be helpful because of possibility of getting more test samples for all of our combinations e.g. BPSN, BNSP etc.

Resampling technique. I selected preferential sampling as paper [1] suggest its faster attainment to optimisation with respect to ‘d’. And following d values where used for resampling [1, 0.8, 0.6, 0.4, 0.2, 0, -0.2, -0.4, -0.6, -0.8, -1]

Model. I selected logistic regression as classifier same as [1]

4.1 Results

Following are the metrics scores and accuracy I recorded for resampled training dataset for corresponding d values



d	Subgroup AUC	BPSN AUC	BNSP AUC	Positive AEG	Negative AEG	Accuracy
1	0.929	0.974	0.772	-0.13	-0.24	0.851
0.8	0.928	0.946	0.852	-0.02	-0.13	0.846
0.6	0.926	0.917	0.892	0.08	-0.06	0.842
0.4	0.923	0.894	0.910	0.13	-0.01	0.844
0.2	0.921	0.861	0.930	0.18	0.05	0.837
0	0.923	0.826	0.954	0.22	0.10	0.830
-0.2	0.924	0.788	0.967	0.25	0.15	0.825
-0.4	0.922	0.744	0.976	0.28	0.20	0.816
-0.6	0.924	0.698	0.984	0.29	0.25	0.806
-0.8	0.924	0.640	0.989	0.31	0.30	0.802
-1	0.921	0.547	0.993	0.33	0.37	0.793

Table 1. Metrics Scores for different ‘d’ values

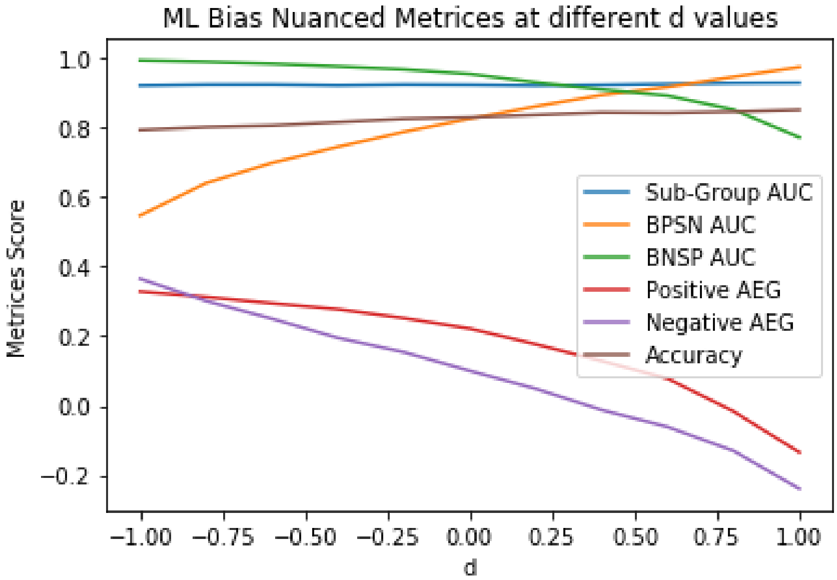


Fig. 4. Plot illustrating metrics core for different ‘d’ values

4.2 Inference (Score interpretation)

- Subgroup AUC is concisely close to ideal value 1 (i.e. at around 0.92) irrespective of resampling disparity correction parameter d. This is understandable and reaffirm that ranking of predictive probabilities for positive & negative classes are very well separated

- BPSN AUC for original dataset was close to ideal value 1 (i.e. at around 0.97) but as the d value is reduced, the BPSN AUC score also goes down. This is because the score signifies ranking order or separation of positive predictive score distribution of background (favoured) and negative predictive score distribution of subgroup (unfavoured). With decreasing d value there will be fewer positive background instances also fewer negative subgroup instances hence reducing the separation

- BNSP AUC for original dataset is 0.77 but as the d value is reduced, the BNSP AUC score increases. This is because the score signifies ranking order or separation of negative predictive score distribution of background (favoured) and positive predictive score distribution of subgroup (unfavoured). With decreasing d value there will be more positive subgroup instances also more negative background instances hence increasing the separation between these two distributions

- Positive AEG oscillates between its extreme values i.e. [-0.5 to 0.5] with respect to resampling disparity correction parameter d. This is because, if we recall

  $$Positive\ AEG = \frac{1}{2} - P(\hat{Y}_D > \hat{Y}_g \mid Y_D \in D^+ , Y_g \in D_g^+)$$

  This suggest at first (original data) probability of positive background sample getting more score than a positive subgroup is greater, and this reduces with d

- Negative AEG follows similar pattern as Positive AEG but in reverse direction

5 Conclusion and Future Work

[2] argue that predictive probability scores-based metrics give more detail picture of bias than class predictions-based metrics and we find this to be true for this project. While metrics by [1] i.e. DPR, EOR and CFR give some information of bias, but they lack granular details of structures of predictive probability score distributions for various subgroups. Also reconciling these three metrics to get optimum d value is difficult. On the other hand, reconciling the nuanced metrics to get optimum d value is much easier and visualising the bias is made easier with various predictive probability distributions

Future work.

Taking implications from this experiment. Theoretically predictive probability score-based resampling can be more effective in removing the bias whiles retaining accuracy of the model. That is

- Instead of oversampling random positives in unfavoured group and random negatives in favoured group, oversample unfavoured samples that have top 5% predictive probability score and favoured samples that have bottom 5% predictive probability score
- It is highly like that these extreme samples are more influenced by other predictors than the PA. hence retaining or improving the accuracy

References

1. Zelaya, Vladimiro & Missier, Paolo & Prangle, Dennis. (2019). Parametrised Data Sampling for Fairness Optimisation.
2. Borkan, Daniel & Dixon, Lucas & Sorensen, Jeffrey & Thain, Nithum & Vasserman, Lucy. (2019). Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification.
3. http://www.biosoft.hacettepe.edu.tr/easyROC/, last accessed 20/02/2020
4. Jigsaw. 2017. Perspective API. https://www.perspectiveapi.com/, last accessed 20/02/2020
5. Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In Proceedings of AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society.
6. Daniel Borkan, Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Limitations of Pinned AUC for Measuring Unintended Bias. CoRR abs/1903.02088, 1903.02088v2 (2019). arXiv:1903.02088v2
7. Nithum Thain, Lucas Dixon, and Ellery Wulczyn. 2017. Wikipedia Talk Labels: Toxicity. (2 2017). https://doi.org/10.6084/m9.figshare.4563973.v2