Bias Correction for Supervised Learning in Email Marketing

Moumita Sinha, Yancheng Li, Wei Shung Chung, Paul Hsiung
{mousinha, yancli, wchung, hsiung}@adobe.com
Adobe Inc.

ABSTRACT

With the popularization of data driven decisions in recommendations, next best action predictions, loan approvals, decision on policies and myriad other applications, it becomes important to explore closely, whether decisions have inherent biases. When there are biases in past decisions, that persist in data and are propagated through machine learning (ML) models. When not corrected, the predicted outcome from a model can be biased with respect to a specific attribute, in two ways, *group bias* and *individual bias.*

In this work, to avoid learning biased decisions from historic data a Generative Adversarial Network is used to transform the training data that is used for predictive models. We introduce two new loss functions that constrain the transformation so that point-wise deviance as well as group bias is reduced, in addition to a loss function that matches the transformation to the original data in distribution. The proposed approach is evaluated on a proprietary email marketing data set, where the task is to determine which consumers should receive a targeting message for marketing purposes. The results show that our approach reduces both group bias and individual bias in the data with respect to sensitive attribute of loyalty card membership of the brand who sends the targeting email messages. Our approach compares favorably to other competing methods. Moreover, the proposed approach of transformation in the email marketing data identifies higher number of individuals among those who had greater engagement with similar emails in the past. These individuals are more likely to respond to future marketing emails, for targeting with email marketing messages, a desirable outcome from predictive modeling of the data. For comparison, we also used the publicly available Adult data set.

KEYWORDS

email marketing, bias correction, GAN, fairness in marketing target, group bias, individual bias

ACM Reference Format:
Moumita Sinha, Yancheng Li, Wei Shung Chung, Paul Hsiung. 2020. Bias Correction for Supervised Learning in Email Marketing. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 8 pages.

1 INTRODUCTION

While decisions based on human intuitions may be biased, even data driven decisions can be discriminatory, because of perpetuated information within the training data. The proposed approach learns a mapping from the original data to a new transformation with reduced bias, subjected to limited data distortion. This proposed approach is named 'DebiasGAN' and is aimed to provide bias corrected training data for input to fitting predictive models, as shown in Figure 1.

Then we show that, building a predictive model with the transformed data, reduces accuracy of the prediction of who should receive the next message. The reduction in the accuracy is desirable, as it shows that the proposed model is predicting away from the ground truth, which is based on 'biased' data.

Let us now consider the notations of the supervised learning task:

- Y is a binary random variable representing the classification decision for an individual.
- X is the entire data set for all individuals for different attributes. $x \in X$ is a vector for each of the attributes for all the individuals. These features can be either binary, multinomial or continuous variables.
- S is an additional binary random variable, whose values represent the different groups to which an individual can belong to and for which bias reduction is aimed to be achieved for the particular outcome Y. S is described as sensitive feature or the selected feature.

Group bias measures the discrimination with respect to the sensitive feature S and target treatment Y and when both Y and S are binary, the group bias score, is defined as

$$\phi_G = |P(Y = y|S = 0) - P(Y = y|S = 1)|$$

This score is desired to be as close as possible to 0.

Individual bias measures the consistency of treating similar individuals with similar activity, disregarding the membership status to different groups of the sensitive feature S. This is measured by first clustering all the individuals into K groups, using K-means clustering of X. Then from each cluster, the sum of the individuals in each of the classes of Y, which is a binary variable, is calculated. A ratio is calculated with the size of the smaller of the two classes of Y in each cluster as the numerator and the larger one as the denominator. A weighted sum of this ratio across clusters is the individual bias, the weight being the relative size of the cluster compared to the complete sample size. More formally, the individual bias is defined as:

$$\phi_I = \sum_{k=1}^{K} w_k \frac{\min[\sum_i (Y_{ik}), \sum_i (1 - Y_{ik})]}{\max[\sum_i (Y_{ik}), \sum_i (1 - Y_{ik})]},$$

where all the individuals in the data set are assigned to one of K clusters, Y_{ik} is the decision variable for the i^{th} individual (takes values $0/1$) in the k^{th} cluster. $w_k = \frac{n_k}{N}$ is the weights for each cluster in individual bias score, where, n_k is the number of individuals in the k^{th} cluster and N is the total number of individuals in the data

SIGIR eCom'20, July 30, 2020, Virtual Event, China

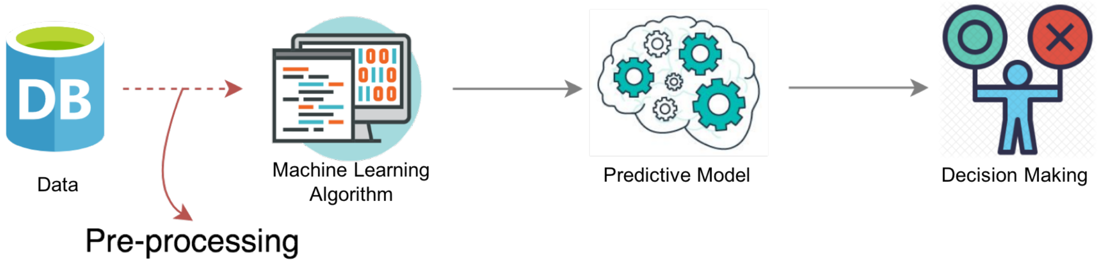


Figure 1: Bias Correction in Pre-Processing

set. In an ideal situation, each cluster is as similar as possible based on the individual's activity and in the absence of individual bias, all individuals in each cluster will be awarded similarly, that is have the same value of Y_{ik}, leading to the value of the numerator in ϕ_I equal to 0. Thus individual bias score is also desired to be as close as possible to 0.

2 METHODS

Our proposed architecture for DebiasGAN, as shown in Figure 2 is composed of one generator G and three loss functions denoted by $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_{MSE}$. The input to the generator is from a latent representation of the original data from a pre-trained autoencoder.

Autoencoder For the autoencoder, the original data used is non-sensitive feature vectors, X and response variable Y, concatenated as input. Let us consider the latent representation from this autoencoder be z, which is then used as the input to the DebiasGAN (Equation 1).

$$z = Enc(X, Y) \tag{1}$$

Decoder learns how to reconstruct the latent representation in a lower dimension back to the original sample space. The loss function for the autoencoder training is minimization of Mean Square Error (MSE) between the original and reconstructed data:

$$L_{AE} = min||Dec(Enc(X, Y)) - (X, Y)||_2^2 \tag{2}$$

DebiasGAN The aim of the adversarial network is to generate a transformation of data so that it stays as similar as possible to the original data, but with reduced dependency on the selected sensitive variable S in the relation between Y and X and in the process transform X, Y such that individuals with similar values of transformed X, have similar transformed values of Y.

$$\Delta((\hat{X}, \hat{Y}), (X, Y)) = ||[\hat{X}, \hat{y}] - [X, y]||_2^2 \tag{3}$$

The proposed approach uses a generator G that takes the learned latent feature z from the autoencoder as input, and after a three-layer neural network, outputs (X', Y'), under 3 constraints, namely *data distortion, data utility* and *dependency on the selected feature* for bias correction.

$$G(z) = (X', Y') \tag{4}$$

The three constraints used in our architecture are defined as follows.

Data Distortion (MSE): The data distortion controls for the transformation of $\{Y, X\} \rightarrow \{Y', X'\}$ constrained to reduce or remove altogether pointwise any large deviations between the original and the new generated data and is defined as

$$E[\Delta((X, Y), (X', Y'))|S, X, Y] \leq c_{s,x,y}$$

The pointwise constraint helps to maintain for every individual, the transformed data to be as close as possible to the original data. In this submission, we have considered the metric for data distortion as mean squared error, i.e.

$$\Delta((X, Y), (X', Y')) = ||G(z) - [X, Y]||_2^2 \tag{5}$$

Data Utility (D_1): Data utility aims to have the statistical distribution of $\{X', Y'\}$ close to that of $\{X, Y\}$, that is, the joint distribution of transformed data $p_{\{X', Y'\}}$ should be statistically close to the original distribution $p_{\{X,Y\}}$ and is achieved by the discriminator D_1. The objective function, maximized over D_1 and minimized over $G(.)$ is defined by

$$\underset{G}{min}\,\underset{D_1}{max}\, E_{\{X,Y\}}[log\, D_1(X, Y)] + E_{G(z)}[log\,(1 - (D_1(G(z)))]$$

Data Dependency (D_2): The main objective in reducing group bias is to obfuscate the sensitive information S from (X', Y') and hence remove the dependency or association between then. The discriminator D_2 distinguishes between samples from $P[G(z)|S = 1]$ and $P[G(z)|S = 0]$, while the generator G(.) aims to have samples from both as similar as possible. The objective function, maximized over D_2 and minimized over G(.) is defined as

$$\underset{G}{min}\,\underset{D_2}{max}\, E_{\{G|S=1\}}[log\, D_2(G(z))] + E_{\{G|S=0\}}[log\,(1 - (D_2(G(z)))]$$

$$\underset{G}{min}\,\underset{D_2}{max}\, E[log\, D_2(G(z)] + E_{G(z)}[log\,(1 - D_2(G(z))]$$

Within transformed data, prediction $\hat{Y} \perp S$ and non-sensitive features $\hat{X} \perp S$. D_2's objective is to precisely predict the sensitive attribute, S, given $(\hat{X}, \hat{Y})$, which has the same purpose as the pre-trained classifier. The only difference is that the input of the classifier are real samples (X, Y), while for D_2 inputs are the generated samples $(\hat{X}, \hat{Y})$. At the same time, G seeks to fool D_2, where the generated samples do not encode any information about S. Once the discriminator is not able to accurately predict S, the independence between transformed data and sensitive variable is satisfied and group fairness has been achieved.

Overall, DebiasGAN performs a minimax optimization between generator and discriminators. D_1 aims to accurately distinguish between real and generated samples and D_2 seeks to distinguish between samples with different group membership of the selected

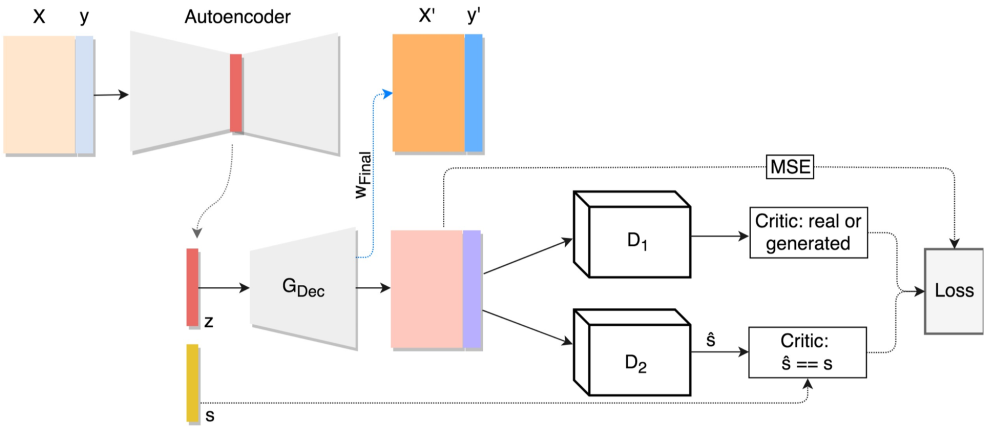


Figure 2: DebiasGAN architecture

feature. The DebiasGAN also aims to reduce the MSE between the original and the generated data, to control data distortion. The overall loss function is then defined as

$$\min_{G}\max_{D1,D2} \mathcal{L}(G,D_1,D_2) = \lambda_1\mathcal{L}_1(G,D_1) + \lambda_2\mathcal{L}_2(G,D_2) + \lambda_3\mathcal{L}_{MSE} \tag{6}$$

such that $\mathcal{L}_1(G,D_1) =$

$$\min_{G}\max_{D_1} E_{\{X,Y\}}[log\ D_1(X,Y)] + E_{G(z)}[log\ (1-(D_1(G(z)))]$$

$$\begin{aligned}\mathcal{L}_1(G,D_1) &= \mathcal{L}_{discriminator1}([x,y],G(z)) \\ &= \frac{1}{N}\sum_{i=1}^{N}[\log(D_1(x,y,s)) + \log(1-D_1(G(z),s))]\end{aligned} \tag{7}$$

$$\begin{aligned}\mathcal{L}_2(G,D_2) &= \mathcal{L}_{discriminator2}(s,G(z)) \\ &= \frac{1}{N}\sum_{i=1}^{N}[\log(D_2(G(z),s)) + \log(1-D_2(G(z),s))]\end{aligned} \tag{8}$$

$\mathcal{L}_2(G,D_2) =$

$$\min_{G}\max_{D_2} E_{\{G|S=1\}}[log\ D_2(G(z))] + E_{\{G|S=0\}}[log\ (1-(D_2(G(z)))]$$

and $\mathcal{L}_{MSE} = \Delta((X,Y),(X',Y'))$

Activation Functions Similarly, D_2 is initialized with weights of a pre-trained classifier.

Both discriminators D_1 and D_2 employ a fully-connected neural network with three hidden layers with Leaky-ReLU activation function after each layer and Sigmoid function as output layer. During training, all models adopt the Adam optimizer. For the GAN learning process, generator G is initialized with the pre-trained weights from the autoencoder.

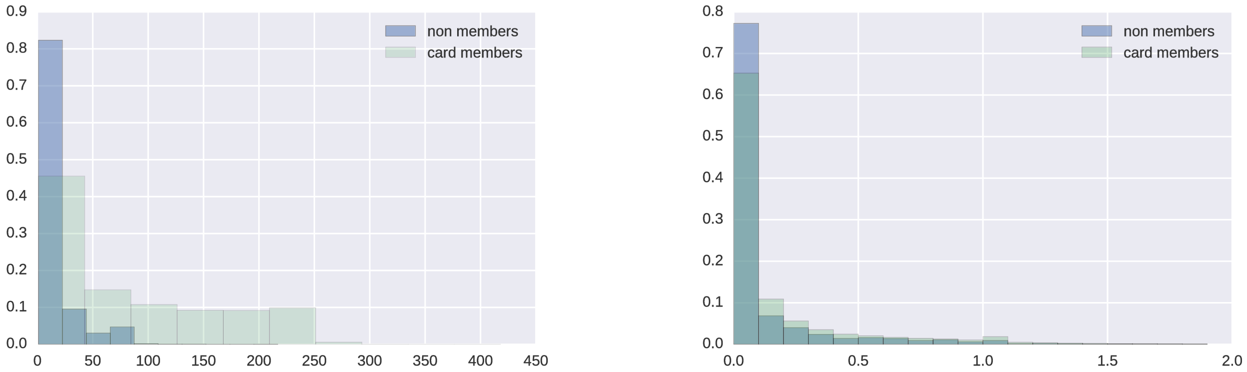
3 RELATED WORK

Discrimination prevention in machine learning has caught attention over the past few years and researchers have studied notions of fairness and methods for addressing the bias problem. In terms of notions of fairness, some have tried to achieve only increased group fairness in their work [1, 10, 12, 16], while [9] has introduced individual fairness. On the other hand [21] has proposed an approach to achieve increased fairness in both group and individual fairness simultaneously. For measurements, the mean difference score [6] and disparate impact factor [10] are widely used for measuring demographic parity, which is the difference or ratio of the probabilities of receiving positive treatment between different groups. Other methods of measuring bias have been proposed as disparate mistreatment [20] and inequality indices [18].

After quantifying the level of bias within data, fairness-aware approaches are adopted to prevent embedding bias into machine learning based decisions. These approaches are generally categorized into two main groups based on the time of bias correction intervention: (1) pre-processing approaches to perform transformations or modifications on the original training set before training or applying any machine learning models [7, 8, 13, 19, 21]; Calmon et. al.[7] formalizes a convex optimization to learn a transformation of the original data to achieve fairness and control for data distortion. Zemel et. al.[21] proposes a learning framework for probabilistic mapping of individuals to representations that achieve both group and individual fairness. (2) post-processing methods involve modifying the predictive models to take into account the data biases [2, 15, 17, 22]. Hardt et. al. [11] elaborates an optimized way of adjusting predicted decisions to remove discrimination.

Among the existing works that uses generative adversarial network, a recent study shows fairness modeling through adversarial learning [19]. It proposes a generative adversarial network, FairGAN, to generate new, group bias corrected data with controlled distortion from the original training data distribution, but does not consider individual bias. [22] presents an adversarial network that can achieve different fairness objectives during the classifier learning process and [5] explores the effects caused by different data distributions on fairness through an adversarial algorithm.

DebiasGAN focuses on generating bias corrected data for the pre-processing step of a predictive modeling workflow. It corrects for both group bias and individual bias, controlling for minimal distributional difference of the transformed data as well as minimal point-wise difference from the original data. In addition, there is a



(a) Messages

(b) Open Rate

Figure 3: Proportion of Consumers with (a) Number of Messages and (b) Open Rates in a Random Sample of 100K, with and without Loyalty Card Membership

pre-step, where our approach learns a transformation on original training data through a learned latent representation of an encoder.

4 EXPERIMENTS AND RESULTS

In this section, four sets of experiments are performed to better understand the contribution of the DebiasGAN in reducing the biases and how this transformation affects the final goal of prediction. First, in Section 4.3 the group and individual biases with respect to a specific sensitive attribute in the transformed data is compared to the original data and other baseline approaches of achieving similar reduction in biases. Second, in Section 4.4 numerical computations are performed to compare how the data distortion, data utility and data dependency are affected by the transformations. Third, in Section 4.5, the contribution of each of the three loss functions in DebiasGAN to group bias and individual bias are compared. Finally, in Section 4.6 the accuracy in the predictive modeling task is compared and explained.

4.1 Data

The proposed DebiasGAN approach is compared with the learning fair representation (LFR) approach as defined by Zemel et al. [21]. The experiments are performed using a combination of two data sets, the adult data set [4], and an email marketing decisions data. The email marketing data set contain a total of $740,608$ instances, characterized by 11 attributes (1 target attribute, 9 non-sensitive attributes and 1 sensitive attribute). The goal is to predict whether the marketer will continue to send out marketing emails to an individual (Y) based on his or her responding and purchasing pattern, described by open and click indicators and purchase records. The 9 non-sensitive features are: during the last 12 months, number of emails received, open rates, click rates, purchase count and total purchase cost for each. We also considered binary features based on the last email received, that is, whether it was opened, clicked, led to a purchase and reaction to last email. Within the data set, 84.28% customers are loyalty card members, while the rest 15.72% without loyalty membership are considered minorities. The loyalty card membership has been considered in this experiment as the sensitive attribute S, and we compare how decision to send marketing rewards are related to this membership versus the consumers' actual interactions with the e-commerce platform. We have 2 years of email interaction data. We use 1 year's data to generate the features for our model to predict for the second year. This entire data set is split into training and evaluation data in $80:20$ proportion.

We use an additional public data 'Adult Data Set' [4] which has a sample size of 32256, with the outcome variable Y being whether an individual makes above or below $50K$ dollars per year. The sensitive feature S is gender (male or female). Each individual is then described by other 11 attributes, denoted by a matrix X.

4.2 Motivation for Bias Correction

As observed in Figure 3a, the number of messages received by the loyalty card members during the period is much larger than those without the cards. However, the open rates for both groups (Figure 3b) are comparable. The proportion of non-card members who do not open messages are higher than those who are card members. But there are some non-card members, whose open rates are close to 2, which means they have opened each email multiple times. Thus the data shows that there are some non card holders who have high engagement and may be ready for more communications and relevant rewards from the brand. From this data, our hypothesis is that, some of the consumers who do not have card, may be ready for more communications from and engagement with the brand. However, the current data, when used to build a predictive model to determine who should receive the next wave of messages, it will input the past information where card members received more messages than those who did not. This information may bias the prediction, because several features in the predictive model may be associated with the fact that one set of people are loyalty card holders. This will have the potential to penalize those customers

Table 1: Comparison of Bias Scores

	Email Marketing Data			Adult Data		
	Original	LFR	DebiasGAN	Original	LFR	DebiasGAN
Group Bias	0.319	**0.084**	0.251	0.196	0.187	**0.004**
Individual Bias	0.248	0.139	**0.022**	0.267	0.057	**0.0007**

who do not have card membership but have active interactions with similar emails in the past.

In the Adult data set, the proportion of individuals in the higher income group is almost 20 % more in one gender group than other. The aim of the bias correction is to transform this original data so that the bias towards a particular gender group is reduced in the raw data set and in the process reduce the individual bias as well. Then this transformed data can be used for building a model to predict income groups in future.

4.3 Comparisons of Bias Score Reduction

To reduce the perpetuated bias in data, DebiasGAN has been proposed. Table 1 compares the bias reductions in transformed data for both the email marketing data and adult data set. The original score given in the table is the group bias and individual scores calculated on the raw data before transformations. The bias scores for LFR are based on transformation of data using [21] and the parameters were set as $A_z = 50$ (group fairness), $A_x = 0.01$ (individual fairness), and $A_y = 1$ (prediction accuracy). These are compared with the bias scores based on data transformed by DebiasGAN approach. The parameters for the DebiasGAN approach were set as $\lambda_1 = \lambda_2 = 0.4$ and $\lambda_3 = 0.8$. Changing the values of A_x, A_y, A_z and increasing the number of iterations for LFR did not change the bias scores significantly.

4.4 Measure of deviance in transformed data

The transformations are to reduce the associations of a specific selected attribute with the outcome variable and other attributes in the data, which leads to reduction in group and individual biases. At the same time, the aim is to have minimal deviance in the transformed data from the original. The deviance is measured by data distortion, data utility and data dependency as defined in Section 2.

Data distortion measures the pointwise deviance of the transformed data from the original. The loss function to control for data distortion is given in Equation 6. To numerically measure this pointwise deviance, mean squared error (MSE) was calculated between the transformed $\{X', Y'\}$ and original $\{X, Y\}$. Each variable was standardized with mean 0 and standard deviation 1 to measure this deviance. Table 2 provides the mean square error for each of the methods in the two data sets. The distortion is maintained similarly in all the scenarios with a higher distortion in the DebiasGAN approach in the email data.

Table 2: Data Distortion: MSE

	Email Marketing Data	Adult Data
LFR	0.38735	0.38185
DebiasGAN	0.6393	0.26314

Data utility measures the difference in the overall distribution of the variables and is constrained by the loss function $\mathcal{L}_1(G, D_1)$ in DebiasGAN. To compare the distributions of the original data with the transformed data, the density estimation of the standardized values were compared. Attributes with multi-classes were converted to binary attributes and the proportion of each class is compared between the original and the transformed. The distributions before and after the transformations look similar in both the LFR and DebiasGAN approach.

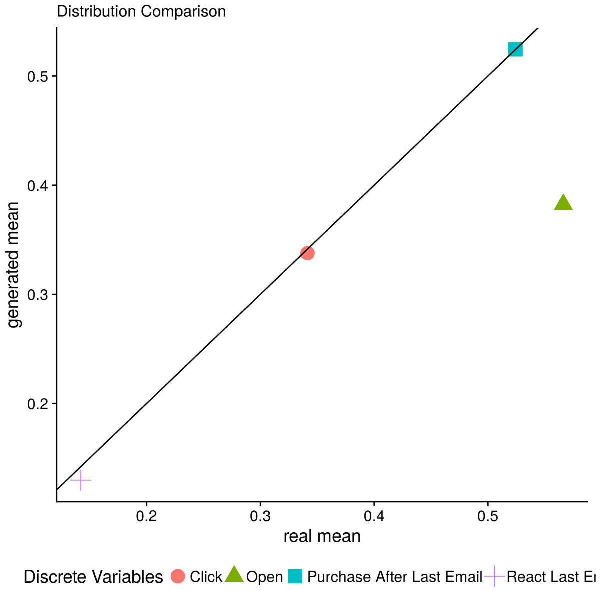


Figure 4: Discrete binary variables: Utility Evaluation

In Figure 4, we have the comparisons of the mean of the discrete predictors between the original (X-axis) and the transformed data (Y-axis) in the email marketing data. We can see that 3 out of 4 variables (click indicator, react to last email indicator, and purchase after last email indicator) lies along the line as expected, which indicates a highly preserved data utility for these three variables. The open indicator for the last email is below the line, which signals a shift in distribution for the transformed data. This shift is caused by the bias removal process. Opening an email is an important indicator, whether an individual is continued to be sent emails. It has a strong association with S, the loyalty card membership, with a large proportion of non-members not opening emails. Therefore, the constrained generation process transforms the open indicator of last email more, to mitigate bias. The number of messages received in the 12 months is a multinomial discrete variable. The utility is

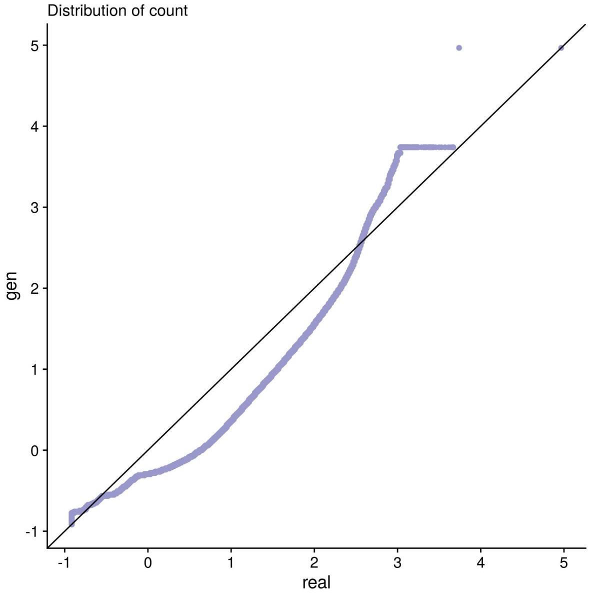


Figure 5: Discrete multinomial variable: Utility Evaluation

then evaluated, using Quantile-Quantile (Q-Q) plot (Figure 5) of the standardized variable of the count. The distributions of the original and the transformed data are close specially for the interval $[-1, 1]$ within which 80% of the standardized count lies in the data set.

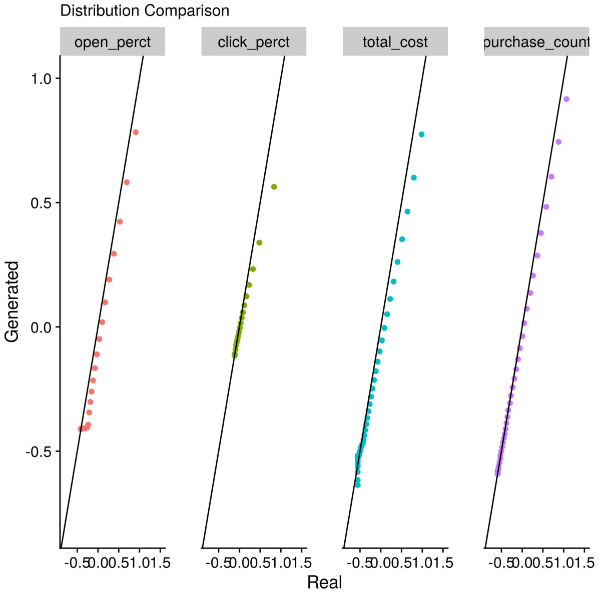


Figure 6: Continuous variables utility evaluation

One thing to notice is that for the open percentage variable, the small shift in lower percentiles indicates the transformation effects performed specifically on this variable. This shift can further lead to a change in the response variable, Y, where the DebiasGAN was trained to achieve group fairness.

Table 3: Data Dependency: AUC of ROC of prediction of S

	Email Marketing Data	Adult Data
Original	0.9016	0.8778
LFR	0.9823	0.7731
DebiasGAN	0.6651	0.5283

Data dependency measures the association between the sensitive variable S and the transformed variables (X', Y'). To assess the reduction in dependency between the selected feature S and the transformed data, a logistic regression has been fitted to predict $P[S|X', Y']$. The AUC of the ROC curve of the prediction provides a measure of accuracy of the ability of the model to predict S, given the transformed data. Table 3 compares this AUC with that of logistic regressions that predict S in the original data, generated data from LFR and DebiasGAN. Both email marketing data and adult data set results are shown. The AUC for the original data for both the email and the adult data is close to 0.9, showing a high degree of accuracy in predicting S from X, Y. Thus this shows that the X, Y contains information that are associated with S. Now the transformations X', Y' by both LFR and DebiasGAN are aimed to reduce the association between S and X', Y', which should result in a lower accuracy in predicting $P[S|X', Y']$. Table 3 shows that the AUC from the DebiasGAN transformation is close to 0.5 (meaning the desired no association is reached) in adult data set and it is 0.6651 in the email marketing data set. These reductions in AUC values is much more than that in LFR transformations, which are 0.7731 and 0.9823 respectively. This shows that DebiasGAN has been able to reduce the association between the selected feature and the rest of the information much more than the LFR and that this reduction in some scenarios reaches close to no association, as desired.

The comparable accuracy and better performance on AUC with respect to target transformed response variable indicates that DebiasGAN is learning the true data distribution and being able to mitigate the bias simultaneously. The significantly different accuracy comparing with true response variable Y indicates the necessity of fairness modeling.

4.5 Contribution of Loss Functions

The previous sections show how the data transformations control for point-wise and distributional deviance, reducing bias in the outcome variable for the selected attribute. The transformations in DebiasGAN are achieved by simultaneously optimizing three loss functions. Table 4 shows how the bias scores are affected if the transformations are performed by optimizing a subset of the loss functions instead of all three of them. The optimizations by dropping the $\mathcal{L}_{\mathcal{MSE}}$ and keeping just $\mathcal{L}_1(G, D_1)$ and $\mathcal{L}_1(G, D_2)$ did not converge after trying varying number of iterations. As seen in the table, the transformations by dropping one of the loss functions in both the data sets however, did not yield any satisfactory reductions in either the group or the individual bias scores. This shows that the probabilities of reaching optimal values without one of the loss functions is reduced considerably and the inclusion of

Table 4: Contribution of the Loss Functions

	Email Marketing Data		Adult Data	
	Group Bias	Individual Bias	Group Bias	Individual Bias
Original	0.319391	0.248	0.196099	0.267007
DebiasGAN: $\mathcal{L}_1(G, D_1)$, $\mathcal{L}_2(G, D_2)$, $\mathcal{L}_{MSE}$	0.251	0.022	0.004359	0.000749
$\mathcal{L}_1(G, D_1)$, $\mathcal{L}_{MSE}$	0.264792	0.405334	0.211477	0.192387
$\mathcal{L}_2(G, D_2)$, $\mathcal{L}_{MSE}$	0.319656	0.419354	0.013693	0.047581

all three loss functions are important to achieve reduction in the bias scores.

4.6 Predictions Using Transformed Data

The primary aim of the data transformations have been to reduce bias with respect to a selected attribute for an outcome prediction. In the previous sections it has been shown how DebiasGAN successfully reduces group and individual bias in the data. In this section, the email marketing decision data is used to show how the transformation in the data affects the prediction of the outcome variable. The DebiasGAN results are compared with that of the original data as well as the LFR transformations.

Table 5: Past Open Rates in the Labelled Classes Using Original and Transformed Data in Evaluation Data Set

	Labelled	Mean	Median
Original Data	Send	25.4	8.6
	Don't Send	10.2	0
LFR Data	Send	28.5	10.7
	Don't Send	4.4	0
DebiasGAN Data	Send	35.2	15.9
	Don't Send	1.6	0

When marketers send email messages, one of the metrics is to measure open rates of the sent emails. The future open rates of the predicted outcome classes based on LFR and DebiasGAN transformed data cannot be measured, because open rates exist for only those who were sent marketing messages. Thus, Table 5 compares the past open rates of individuals in these labelled outcome classes in the original and the transformed data, for the evaluation data set. For the original data, Y denotes whether an individual will be sent a marketing email and X is a set of nine attributes as described above. Bias with respect to selected feature loyalty card membership, was perpetuated from the fact that more targeting messages were sent to the members. This original data is then transformed to Y' and X' respectively using LFR and DebiasGAN. The transformation is aimed at reducing the bias of who are sent next set of messages, which changes the labels of 0/1 in Y'. The past open rates in the labelled send and do not send groups change after the transformations (Table 5). The set of individuals who were labelled as, to be sent marketing messages, based on original data had past open rates at 25.4%, which is much lower than that the past open rates of 35.2% in the same group in DebiasGAN group. The corresponding increase in the LFR transformed data is not as much. Consequently, the past open rates of those who were labelled as to be not sent marketing messages based on original data are higher by about three folds compared to those to be not sent messages by the DebiasGAN data. LFR transformation also yielded a reduction in the past open rates of individuals who are labelled as to be not sent targeted messages. Thus with the reduction in the bias scores and the shifting of the values in Y', resulted in changes in the past open rates in the two outcome groups. This overcomes choosing too many non-card members to be not sent emails and instead helps to choose individuals who had more interactions with the emails in the past.

Table 6: Logistic Regressions with Transformed Data

AUC of ROC	
Original	**0.8837**
LFR-Gen	0.9960
LFR-Orig	0.8810
DebiasGAN-Gen	0.9170
DebiasGAN-Orig	**0.6770**

*-Gen and -Orig are generated and original data respectively as ground truth

While the overall bias is reduced in the transformed data with respect to the selected feature, and has shifted the groups of individuals in the outcome groups with respect to their past interactions with the emails, it remains of interest on how this transformation affects the accuracy of the prediction of the outcome variable. These transformed data are then used to fit a logistic regression to predict which individuals will be sent targeting messages in the future. Table 6 shows the accuracy of predictions from the logistic regressions, using AUC of the ROC curve. The AUC for the prediction using the original data is shown as 0.8837. When this data is transformed using DebiasGAN, the AUC for predicting the transformed outcome, that is Y', is 0.9170. However, the AUC for predicting the original outcome values that is, Y, using the transformed data X' is 0.6770. This shows that the logistic regression as a model has a high accuracy in predicting the binary outcomes. However, as the aim of the transformation is to change the outcome, such that bias is controlled with respect to a selected feature, the accuracy of predicting the original outcome reduces. This is the desirable outcome and shows that the transformed data is predicting a new pattern than what marketers have been using to determine whom to send the next set of messages. The results from LFR transformation are also compared. The AUC for the LFR transformed data compared to original data is 0.8810. The LFR transformed data maintains a similar accuracy in predictions, as the original data. However, the reduction in the bias scores in most cases (Table 1) are lesser as compared to that in the transformed data from DebiasGAN.

Moreover, among the individuals predicted to not receive the next set of messages based on the original data, 30% of of them are non-card members. When the transformed DebiasGAN data is used for prediction, the percentage of non-card members reduce to 20% of the individuals predicted to not receive the next sent of messages. This reduced percentage of non-card members in the not-send group is close to the population percentage of 14% non-card members, thus confirming the reduction in group bias as a result of DebiasGAN.

5 DISCUSSIONS AND CONCLUSIONS

Bias reduction and fairness is of tremendous importance with the advent of more and more automatic data driven applications to help the society to take decisions. Sometimes these predictive models can inadvertently be biased based on information propagated from past activities or actions [3, 14]. Corrections of similar bias in recommendations and personalizations of digital marketing approaches have also become important.

In this submission, a novel application DebiasGAN, using Generative Adversarial Network framework has been proposed to correct for biases with respect to a selected attribute for an outcome variable. This application is employed during the pre-processing phase of ML workflow, so that the transformed data are used as input for the model building steps. This approach addresses both group and individual biases. Other similar approaches like FairGAN [19] uses a GAN framework to mitigate bias, but it only addresses group bias. In LFR [21], the authors formulated fairness as an optimization problem with the ability to search across different representations of original data, to identify that representation which reduces both group and individual bias, with minimal loss in deviations from original data. In the results presented in Section 4 it is clear that although the deviations are marginally higher in DebiasGAN, compared to LFR, the reductions in bias scores are much larger. The DebiasGAN transformations also relabels more relevant individuals with the appropriate outcome classes, with very little decrease in accuracy of the models, predicting the outcomes. It has also been showed that all three proposed loss functions for DebiasGAN are critical in decreasing the bias scores and dropping any one of then reduces the probability to reach any optimal values and the bias scores stay very similar without any reduction in them.

While the proposed approach of fairness is applicable to a myriad of applications, this work has shown how this can be utilized specifically in the context of email marketing. Without the bias correction, the marketing emails were missing out on an opportunity to reach out to consumers who were not necessarily loyalty card members, but were highly engaged with the brand.

As future work it would be of interest to try multiple hyper parameters and multiple combinations of these loss functions to identify the individual nature of these loss functions in their contribution to the bias scores. Additional interesting future extensions would be to see under what circumstances the bias correction during pre-processing step is more efficient than during the modeling step and how a combination of both may work.

REFERENCES

[1] Tameem Adel, Isabel Valera, Zoubin Ghahramani, and Adrian Weller. 2019. One-network adversarial fairness. In *Thirty-Third AAAI Conference on Artificial Intelligence.*

[2] Junaid Ali, Muhammad Bilal Zafar, Adish Singla, and Krishna P Gummadi. 2019. Loss-Aversively Fair Classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society.* ACM, 211–218.

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchne. 2016. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[4] Arthur Asuncion and David Newman. 2007. UCI machine learning repository. (2007).

[5] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075* (2017).

[6] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2 (2010), 277–292.

[7] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems.* 3992–4001.

[8] Silvia Chiappa. 2019. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7801–7808.

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

[10] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 259–268.

[11] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems.* 3315–3323.

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

[13] Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. 2019. ifair: Learning individually fair data representations for algorithmic decision making. In *2019 IEEE 35th International Conference on Data Engineering (ICDE).* IEEE, 1334–1345.

[14] Sam Levin. 2016. A beauty contest was judged by AI and the robots didn't like dark skin. https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people.

[15] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed impact of fair machine learning. *arXiv preprint arXiv:1803.04383* (2018).

[16] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. 2011. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 502–510.

[17] Luca Oneto, Michele Donini, and Massimiliano Pontil. 2019. General fair empirical risk minimization. *arXiv preprint arXiv:1901.10080* (2019).

[18] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual &Group Unfairness via Inequality Indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* ACM, 2239–2248.

[19] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2018. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data).* IEEE, 570–575.

[20] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web.* International World Wide Web Conferences Steering Committee, 1171–1180.

[21] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International Conference on Machine Learning.* 325–333.

[22] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society.* 335–340.