# VFAEs for Endogenously Determined Recidivism Prediction

**Paul Lou**
CS Dept
UCLA
UID: 805352210
pslou@cs.ucla.edu

**Pratyush Garg**
ECE Dept
UCLA
UID: 305430437
pratyushg@g.ucla.edu

## Abstract

In recent years, fair ML literature [1] has attempted to formalize different philosophical standpoints into theoretical notions of fairness and provide a discussion on their compatibility [2][3]. Recently, the overwhelming consensus is that fairness must be carefully defined based on domain specific knowledge in which statistical and machine learning models are applied. We consider the domain of recidivism prediction, where the task is to predict repeat arrests. Recently Jung et al. show that assuming endogenously determined actions of crime, the only fairness metric consistent with minimizing crime rates is equalized odds [4, 5]. One pre-processing technique on the input feature vectors, to achieve a notion of fairness with respect to a sensitive attribute is the use of variational fair autoencoders (VFAEs) [6]. We build a VFAE to produce invariant data representations with respect to race to predict recidivism. We measure precision and equalized odds in a binary prediction setting among African Americans and Caucasians in four models: COMPAS' score, baseline logistic regression, baseline with unawareness, and baseline with a VFAE pre-processing step. Our analysis shows that all four models fail to achieve equalized odds and instead meet approximate precision parity concluding that in the endogenous setting, none of these models achieve a notion of fairness consistent with minimizing crime rates.

## 1 Background

### 1.1 Variational Fair Autoencoders

**Producing "fair" data representations** Variational fair autoencoders (VFAEs), introduced by Louizos et al. [6], are models which take a feature vector $X$ and specified sensitive attribute $S$ and produces a representation $X'$ such that $X'$ is uninformative about sensitive attribute $S$. We call such a representation $X'$ a latent representation. Moreover, $X'$ retains maximal information about non-sensitive attributes with respect to the objective. If one views the sensitive attribute as contributing unfairness relative to some concrete fairness metric, then the sensitive attribute is justifiably removed and the resultant data should improve fairness of subsequent learning algorithms performed on the representation $X'$. The VFAE comes with a semi-supervised flavor that is aimed to address the issue of correlation between the sensitive attribute $S$ and the target variable, which is in our case the event of recidivism. Addressing the correlation prevents crippling the data of its predictive utility for use in subsequent prediction tasks.

### 1.2 Statistical notions of group fairness

**Overview of notions** There has been a general literature of results in statistical notions of fairness. Some concrete examples and definitions of statistical notions of group fairness include

1. *Statistical Parity*: if the population consists of 30 percent blue aliens and 70 percent green aliens, then a binary classifier $C$ satisfies statistical parity if 30 percent of the positive predictions, (individuals $x$ on which $C(x) = 1$), are blue aliens and 70 percent of the positive outcomes are green aliens. More generally, statistical parity says that the demographics of those individuals classified positively are equivalent to the demographics of the original population.

2. *Equalized Odds*: Equalized odds is satisfied by a binary classifier when the true positive rates (recall) and false positive rates are equivalent among the sensitive groups. We note that Kleinberg et. al's notion of *balance for the positive class* and *balance for the negative class* generalize the notion of equalized odds [4]. Hardt et al. contribute a post-processing algorithm for both binary predictors and real-valued score functions that produces a model that satisfies equalized odds and preserves near optimality. The procedure is model agnostic in

the sense that the post-processing procedure is independent from the input model training algorithm.

3. *Calibration [2]*: Kleinberg et al. introduced calibration within groups as a form of fairness. Calibration within groups stipulates simply that a score function accurately scores within each group.

**Conflicting notions of fairness for recidivism** The whole issue of fairness in recidivism came to light when ProPublica and the creator of the COMPAS risk prediction algorithm, Northpointe, disagreed on notions of fairness. ProPublica showed that equalized odds was violated by COMPAS scores among the African American and Caucasian groups. However, Northpointe responded that predictive parity was met; where predictive parity means that the accuracy within each groups is approximately equivalent. In other words COMPAS was well-calibrated among the African American and Caucasian groups. This problem got further confused when it was shown that Equalized odds and well-calibration within groups were inherently incompatible notions of fairness by Kleinberg et. al [2]. This incompatibility gives us a choice between the two notions.

**Correct notion of fairness for recidivism** Roth et al. recently show that equalized odds is consistent with the goal of minimizing crime rate when assuming that the choice to commit crimes is determined by societal policies. [5]. These societal policies include legal standards for conviction and enforcement decisions. The nomenclature for such an assumption is that, if the choice of an agent to commit a crime is determined by these societal policies and social behaviour, it is said to be endogenously determined. Endogenously determined choices are in contrast to exogenously determined choices. To be clear, an exogenously determined choice means the agent's choice of committing a crime is independent of societal policies. In summary, the view of exogenous and endogenous determination affects which notions of fairness match the natural objectives of crime reduction and in particular, assuming endogenous determination equalized odds was shown to be the correct metric. *Hence, we take equalized odds as our fairness metric.*

## 2 Project Overview

Having discussed the various different parts of the problem, we now define what our objective is for the project. We focus on comparing and contrasting the predictive precision, true positive rate (recall), and false positive rate segmented by race among four different models that are chosen to fulfil the objectives below.

We believe that this analysis should a) shed light on how good/bad the COMPAS algorithm is when compared to some pretty basic models, b) check VFAE compatibility within the recidivism domain as opposed to the image domain it was initially designed to be used on, c) confirm Kleinberg's incompatibility result, d) give information about "proxy" variables and their influence.

The four models we compare are:

1. **COMPAS' score as a binary predictor**: we threshold COMPAS' decile scores by $[0, 3]$ as predicting no recidivism and $[4, 10]$ as predicting recidivism. This is consistent with their interpretation of "Medium-High Risk" individuals as likely to recidivate.

2. **Baseline logistic model**: a logistic regression on the available dataset. The performance of the baseline model is highly sensitive to various score thresholds. Once again, we choose a threshold of 0.35 to be consistent with COMPAS' decile scoring system in which scores in [0, 3] are considered low risk and scores [4, 10] are considered at risk.

3. **Unaware logistic model**: the baseline logistic regression model trained without the race column. This models the naive idea that if we blind the model to the explicit race values, the model should become subsequently fair.

4. **VFAE-encoded input to baseline model**: encoding the original feature vectors through a VFAE and feeding the encoded representation as training data into the baseline model. Comparing this to the unaware model will give us information about both the correlation of the output with the sensitive attribute and the race proxy information in the other feature columns.

We focus on a binary race variable, the two categories being African American and Caucasian. This choice is due to previous attention on differences in these two race groups in existing academic and journalistic work.

### 2.1 Dataset

The dataset used in this study was freely publicised by ProPublica [7]. In the associated link in the citation, ProPublica's methods of legally obtaining the data are fully described. In particular we focus on the *cox-parsed.csv* file. The dataset contains 11364 rows representing data on individuals from the Broward County Jail in Florida. Among the 52 features, we filter by case identification number, sex, age, age category, race (African American and Caucasian categories), juvenile felony count, juvenile misdemeanor count, juvenile other crimes

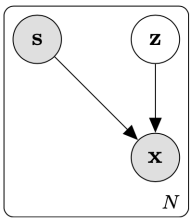



Figure 1: Unsupervised model

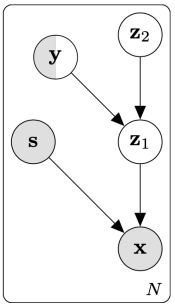



Figure 2: Semi-supervised model

Figure 1: The VFAE Model Structure

count, priors count, days spent in jail, and charge degree for a total of 11 features. The true target variable is given by the column labeled as `is_recid` representing whether or not the convict committed another crime within two years. See our Github repository for the data cleaning mechanism.

## 3 Model Building with VFAE

To tackle the problem of possible degeneracy with respect to the objective when using vanilla VAEs, the authors of [6] use a semi-supervised architecture. The structure is given in Figure 1.

As explained before, the idea is to build a latent representation that encodes maximum information from $X$ wrt to the objective $y$, but does not include information wrt $S$, the protected attribute. A double stacked variational autoencoder is used to this effect and the solution is attempted using the mean-field VI strategy. More details about the VFAE model can be found in the paper [6]. Basically, the following generative story is used.

$$
\begin{aligned}
q_\phi\left(\mathbf{z}_{1n}|\mathbf{x}_n, \mathbf{s}_n\right) &= \mathcal{N}\left(\mathbf{z}_{1n}|(\boldsymbol{\mu}_n, \boldsymbol{\sigma}_n) = f_\phi\left(\mathbf{x}_n, \mathbf{s}_n\right)\right) \\
q_\phi\left(\mathbf{y}_n|\mathbf{z}_{1n}\right) &= \mathrm{Cat}\left(\mathbf{y}_n|\boldsymbol{\pi}_n = \mathrm{softmax}\left(f_\phi\left(\mathbf{z}_{1n}\right)\right)\right) \\
q_\phi\left(\mathbf{z}_{2n}|\mathbf{z}_{1n}, \mathbf{y}_n\right) &= \mathcal{N}\left(\mathbf{z}_{2n}|(\boldsymbol{\mu}_n, \boldsymbol{\sigma}_n) = f_\phi\left(\mathbf{z}_{1n}, \mathbf{y}_n\right)\right) \\
p_\theta\left(\mathbf{z}_{1n}|\mathbf{z}_{2n}, \mathbf{y}_n\right) &= \mathcal{N}\left(\mathbf{z}_{1n}|(\boldsymbol{\mu}_n, \boldsymbol{\sigma}_n) = f_\theta\left(\mathbf{z}_{2n}, \mathbf{y}_n\right)\right) \\
p_\theta\left(\mathbf{x}_n|\mathbf{z}_{1n}, \mathbf{s}_n\right) &= f_\theta\left(\mathbf{z}_{1n}, \mathbf{s}_n\right)
\end{aligned}
$$

All functions $f_\theta$ and $f_\phi$ used in the generative story are modelled by 2 layer fully connected networks. Sampling for the gaussian pdf is done using the reparameterization trick to make the step part of the computation graph. The loss functions are then set up appropriately to include the reconstruction losses for both $x$ and $z_1$, invariance properties between $z_1$ and $s$ that the authors call Maximum Mean Discrepancy and classification loss for $y$.

An interesting thing to note here is that the model behaves differently during training and testing. Since, the computation graph includes an operation on the prediction $y$, which is unavailable during testing, we use the psuedo-output $\hat{y}$ that is trained along with the latent representation for this very purpose. However, these pseduo-outputs do not represent the final predictions since we run a logistic regression model on the latent representations to determine those.

Finally, we train the VFAE for 30 epochs (when the cross-validation loss curves saturated), using a batch size of 50 and taking 40 as the hidden dimension for all the feed forward networks (these parameters were chosen as a result of a basic hyperparamter search). The trained model was then used to generate latent representations for the test dataset and then fed into the baseline logistic regression model.

## 4 Results

For the following commentary, please refer to table 1 and table 2.

**Baseline model** Despite using only 11 features instead of COMPAS' private questionnaire results encompassing 137 features, our baseline logistic regression model achieved comparable precision to that of COMPAS' equivalent binary prediction. The main deficiency in the baseline model is a marked drop in recall. In terms of equalized odds, the baseline model is far more unfair than COMPAS' equivalent binary prediction model: note the large differences in TPR and FPR among African Americans and Caucasians in table 2.

**Blinded baseline model** As we expected, removing the race feature reduced the differences in TPR and FPR while also reducing the magnitude of difference in precision. Interestingly, the overall accuracy for the blinded model wasn't too compromised and hence, the approach seems better than the baseline.

**VFAE-encoded input to baseline model** The VFAE encoded baseline model performance,

| Model | Precision (AA) | Precision(C) | TPR (AA) | TPR (C) | FPR (AA) | FPR (C) |
|---|---|---|---|---|---|---|
| COMPAS' Binary Prediction | 0.424 | 0.374 | 0.811 | 0.65 | 0.648 | 0.421 |
| Baseline | 0.491 | 0.519 | 0.679 | 0.343 | 0.416 | 0.136 |
| Unaware Baseline | 0.503 | 0.5 | 0.613 | 0.407 | 0.357 | 0.174 |
| VFAE Baseline | 0.415 | 0.401 | 0.711 | 0.385 | 0.564 | 0.244 |

Table 1: Precision, TPR, and FPR for each model. AA = African American, C = Caucasian.

| Model | Precision Diff. | TPR Diff. | FPR Diff. |
|---|---|---|---|
| COMPAS' Binary Prediction | 0.05 | 0.161 | 0.227 |
| Baseline | -0.028 | 0.336 | 0.28 |
| Unaware Baseline | 0.003 | 0.206 | 0.183 |
| VFAE Baseline | 0.014 | 0.326 | 0.32 |

Table 2: Differences in precision, TPR, and FPR [African American's minus Caucasian's]

however, was much worse than what we had initially expected. The model keeps almost the same precision as the baseline model and the overall accuracy is also similar. The disappointing result is the equalized odds disparity. The model is as worse as the baseline model in this regard. We will discuss some possible reasons for this in the next section.

## 5 Discussion

**Unfairness of COMPAS and simple recidivism prediction models** This study focuses on binary prediction rather than producing a score function that aims to assign probabilities of recidivism. Equalized odds is straightforward to define and measure in the binary prediction following Hardt et al [4]. In particular, we emphasize that these results, especially under the assumption of endogenous determination of crime, reinforce that COMPAS is not fair in the sense of equalized odds. Moreover, our results show that, despite the theoretical disadvantages of fairness through unawareness, removing the race feature empirically improved fairness.

**Kleinberg's incompatibility** The VFAE structure focuses on removing information about the protected attribute from the latent representation. However, it does not in any way model the equalized odds condition. Specifically, the loss function used for disentanglement may be implicitly encouraging precision parity which, as we know by Kleinberg's condition is incompatible with Equalized Odds. The authors of the paper do not discuss which metric they will measure fairness by and instead just use a custom discrimination ratio to justify the model.

**Proxy variables? Not really...** Another issue with our dataset, besides the lack of training points, is the dearth of features. We have just 11 features that are utilizable for training. In these, though, we confirmed using pearson coefficients, that there is very little correlation between race and other columns since there are no real "proxy" variables that model the race identity implicitly. This helps to show that the unaware model is a decent idea for this dataset and hence performs better than the VFAE which is bogged down by training concerns.

**VFAE performing closer to random** All these factors combine to make our result, a negative result on the effectiveness of VFAEs as a pre-processing step for obtaining equalized odds in the COMPAS recidivism problem. The Kleinberg incompatibility, lack of proxy variables, small training data, all contribute to making VFAEs not very useful without any modifications. In fact, we find (from the ROC) that the VFAE encoded model is closer to the random classifier.

**Future Directions** A final point on the project is aimed at addressing the Kleinberg incompatibility issue. Specifically, if instead of using VFAE's to work on some notion of disentanglement which maybe encouraging precision parity, what if we use the VFAE architecture (which is very suitable for this task) to explicitly encourage Equalized Odds in the loss. We believe this will yield interesting results, and will try this in our research work in the near future.

## 6 Acknowledgement

We thank Prof. Chang for the opportunity to pursue such an interesting project. We also thank ProPublica for their insightful journalism in raising this issue and in making their dataset public and free for use.

## References

[1] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY,

USA, 2012. Association for Computing Machinery.

[2] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores, 2016.

[3] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, 2017.

[4] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning, 2016.

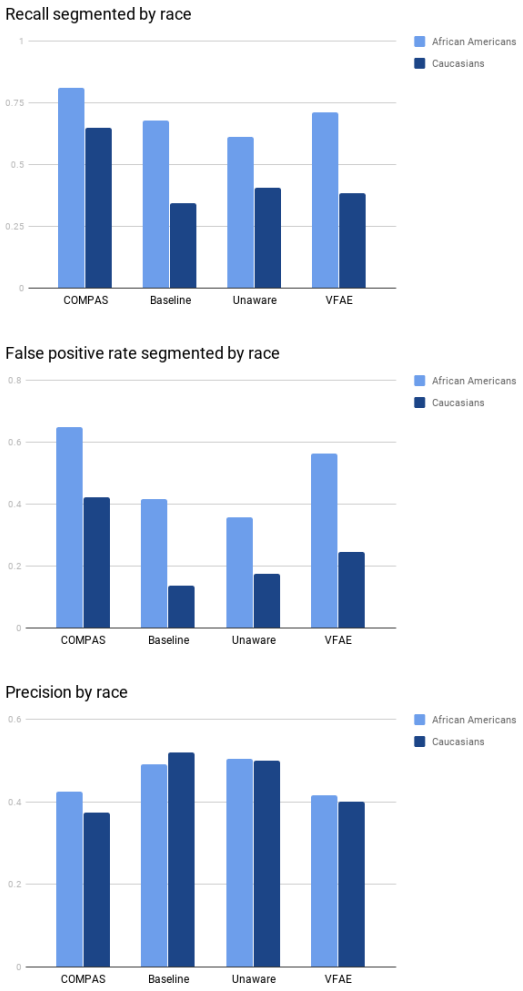
[5] Christopher Jung, Sampath Kannan, Changhwa Lee, Mallesh M. Pai, Aaron Roth, and Rakesh Vohra. Fair prediction with endogenous behavior, 2020.

[6] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder, 2017.

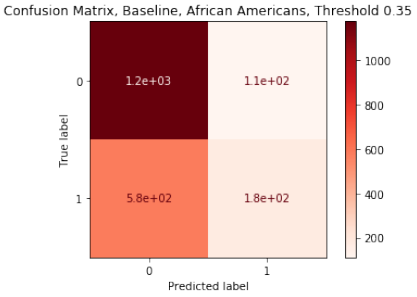
[7] ProPublica. Compas recidivism risk score data and analysis, 2016.

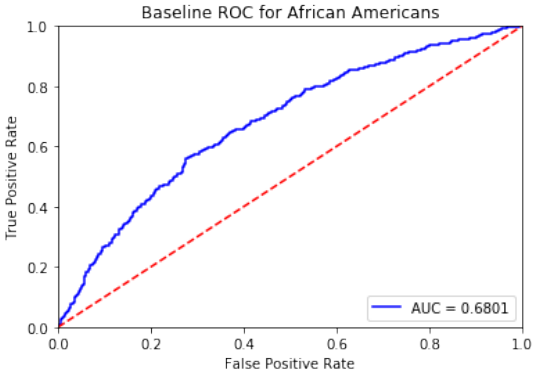
# 7 Appendix A: Charts

## 7.1 Overview of model performance and fairness by race

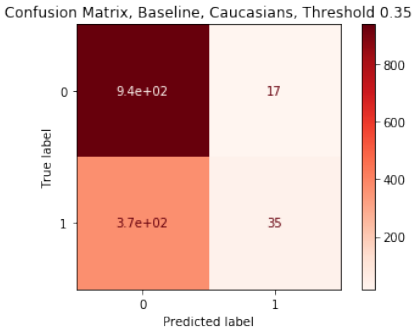


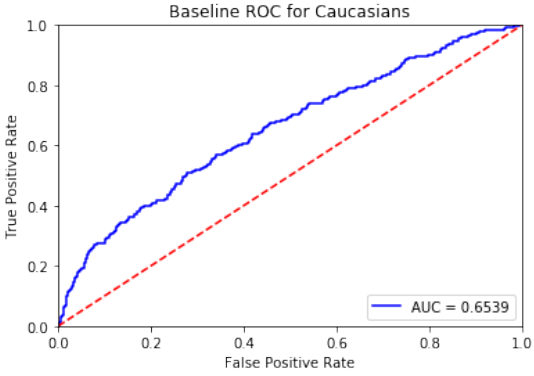

## 7.2 Baseline Model ROC and Confusion Matrix (African American)



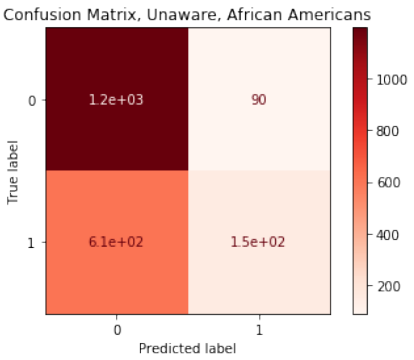



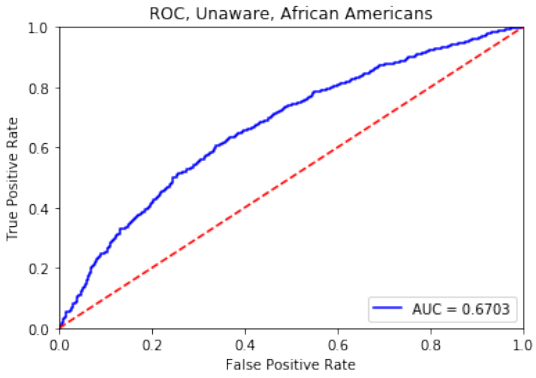

## 7.3 Baseline Model ROC and Confusion Matrix (Caucasian)



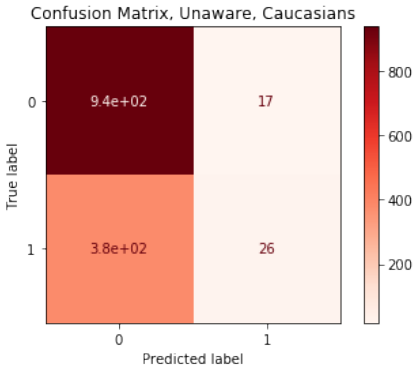



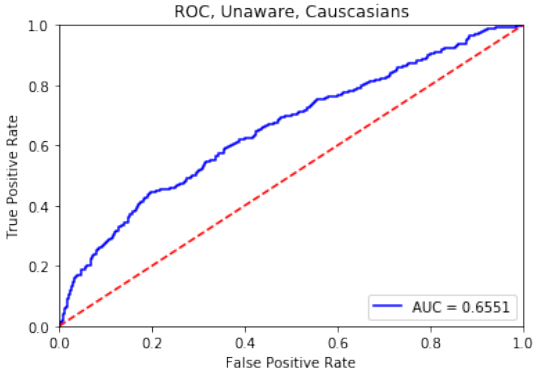
### 7.4 Unaware Model ROC and Confusion Matrix (African American)



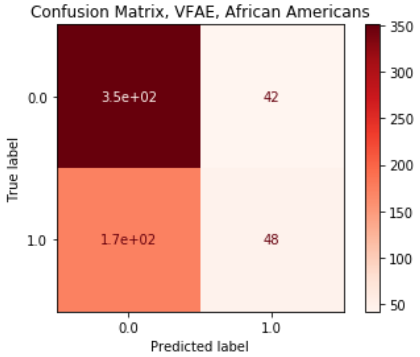



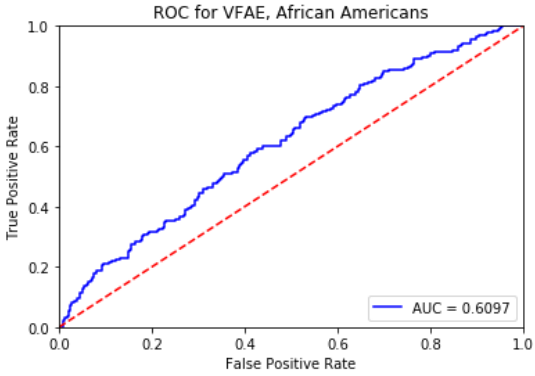

### 7.5 Unaware Model ROC and Confusion Matrix (Caucasian)




### 7.6 VFAE Model ROC and Confusion Matrix (African American)




### 7.7 VFAE Model ROC and Confusion Matrix (Caucasian)