Individual and Group Fairness under Uncertainty

Victor Kaberuka Shyaka

[1] *Computer Science, Stanford University*

Introduction

In the realm of predictive modeling, ensuring fairness alongside accuracy is paramount, particularly in contexts with significant societal impacts, such as criminal justice. This project focuses on reproducing the results from the paper "Individual Fairness under Uncertainty"[1] using the COMPAS dataset, which is widely used for predicting recidivism. The objective is to develop a Fair Individual Cox Model (FairIndvCox) that integrates fairness considerations into the Cox proportional hazards model, ensuring both accurate and fair predictions under conditions of uncertainty.

The paper in question argues that the individual fairness algorithm surpasses other models at their Fair Normalized Discounted Cumulative Gain score (FNDCG). They offer no explicit work with group fairness criteria, or a mix of the two. This work explores the possibility of improved performance in a blend of individual and group fairness criteria.

Method

In our experimental model, we create a combination of Individual and Group fairness criteria, the fairIndGrpCox algorithm. For our group fairness metric we chose disparate impact, which measures the positive outcomes between different demographic groups, and we chose race. We chose this for its wide recognizability and this metric ensures that the rates of positive outcomes (such as non-recidivism) are similar across different demographic groups, promoting fairness across these groups.

$$\mathcal{L}(\beta) = -\text{partial likelihood} + \text{Norm}\left(\gamma(1 - \text{FNDCG}) + \delta(1 - \text{group fairness})\right)$$

Background

Traditional fairness approaches based on the Lipschitz condition require that **differences in treatment or predictions** be **proportional to differences in features**, which is challenging to calibrate, especially with censored data. To address this, the paper introduces FNDCG, a measure that evaluates the consistency between similarity rankings in the feature space and the risk prediction space. FNDCG promotes preserving the relative order of similar individuals in both spaces, computed using the similarity matrices *SimD'* (feature space) and *SimD* (prediction space), formulated as:

$$\text{SimD}_{ij} = \exp(-\|r_i - r_j\|) \qquad FNDCG = \frac{DCG_k(SimD)}{DCG_k(SimD')}$$

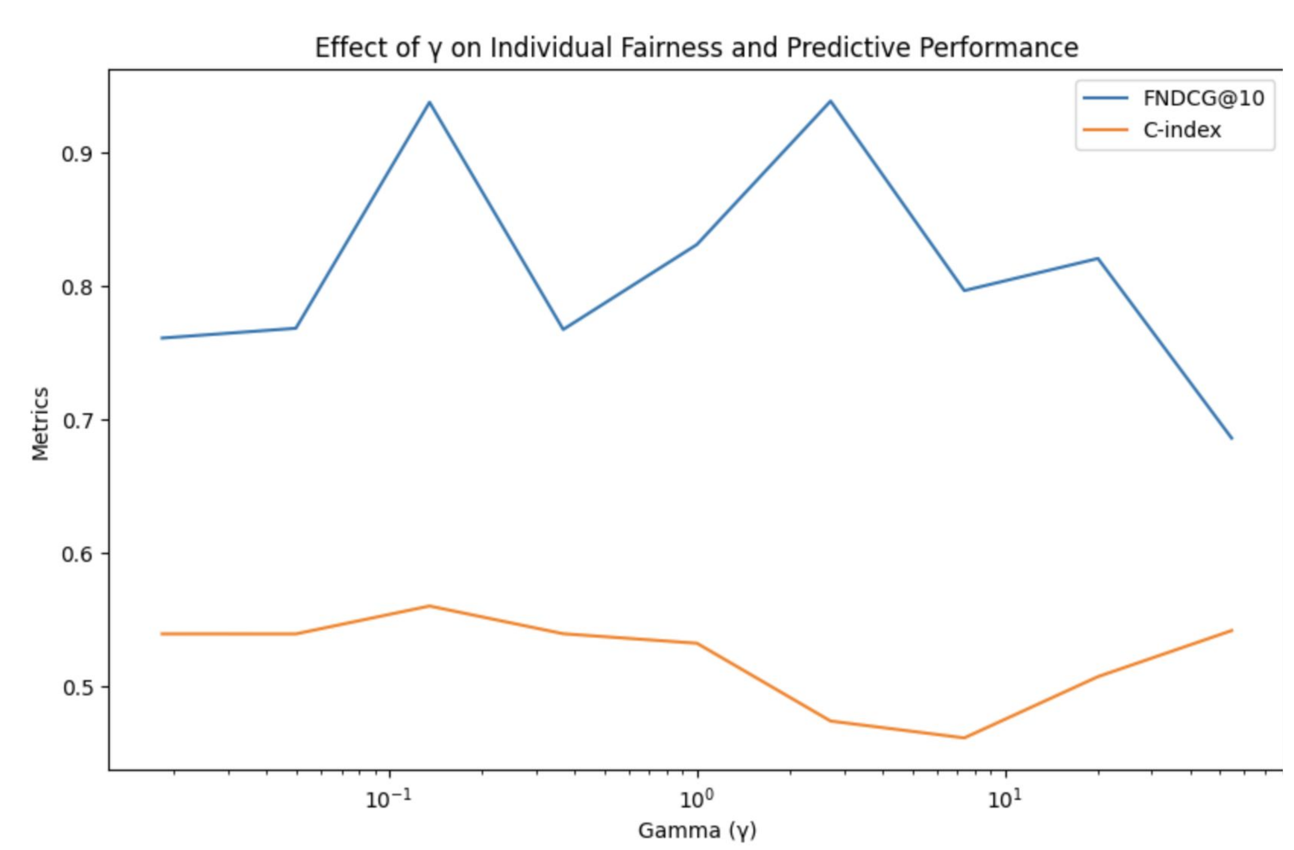
$$DCG_k = \sum_{i=1}^{k} \frac{SimD(i)}{\log_2(i+1)} \qquad \text{FNDCG} = \frac{1}{N}\sum_{i=1}^{N} \text{FNDCG}_k$$

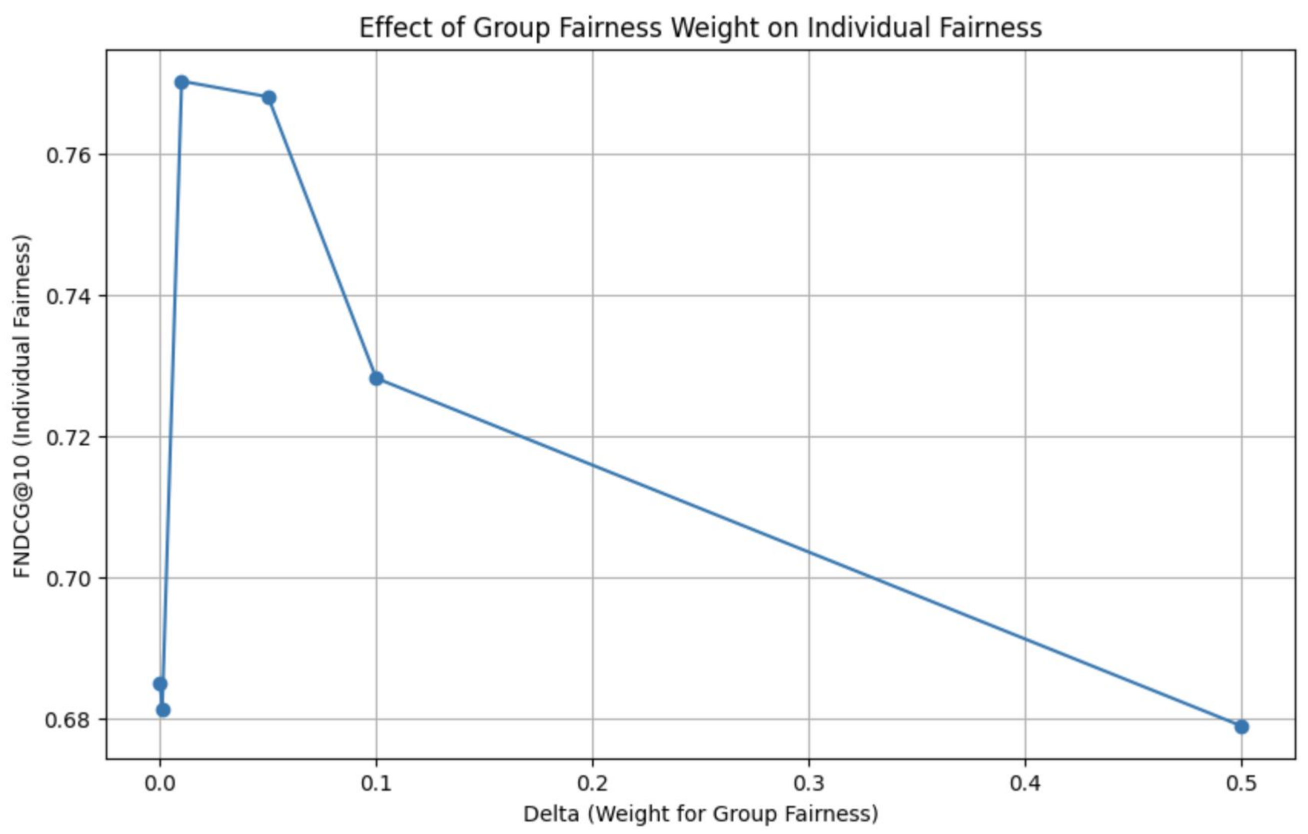
The FairIndvCox model combines the partial likelihood from the Cox model with a fairness term based on FNDCG, balanced by a hyperparameter γ:

$$\mathcal{L}(\beta) = -\text{partial likelihood} + \gamma(1 - \text{FNDCG}) \qquad \mathcal{L}_{\text{partial}}(\beta) = \sum_{i:E_i=1}\left(X_i\beta - \log\sum_{j:T_j \geq T_i} \exp(X_j\beta)\right)$$

This combined loss function ensures both accurate and fair predictions, making it more suitable for applications requiring nuanced and personalized fairness compared to traditional group fairness methods.

Results





Conclusions

The results do not show an improvement from the mixed criteria. Despite there being a processing of parameters between the two criteria, the ceiling of the baseline FNDCG values are not surpassed.

That may be for various reasons, with the most pertinent being **Group choice**. For this experiment, the group of choice was race for its more obvious quality, however an analysis on whether it may collide with the individual metric should be performed to answer that question. Another may be **Metric choice**. Disparate impact, although very convenient, may not be complex enough to account for the intricate space where group criteria fill in the gaps in individual fairness criteria.

It has been shown before that group and individual fairness criteria can foster better results [2], so although the results do not show an improvement there are reasons to think there might still be ways to improve this method further with more complex metrics.

Future work entails trying more complex group choices and metrics, which would require more computing power to process efficiently.

References

[1] Hebert-Johnson, U., Kim, M. P., Reingold, O., & Rothblum, G. N. (2018). *Individual Fairness under Uncertainty*. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2018), pp. 13996-14006.

[2] Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org.