Student: Pavlos Markesinis (P.Markesinis@student.tudelft.nl) | Professor: Anna Lukina

Analysing Data Features on Algorithmic Fairness in Machine Learning

Comparing the sensitivity of data features under fairness properties between different sectors

Introduction

Fairness in Machine Learning: Not a solved topic. Many definitions, some of which are contradicting [1].
Dynamic Monitoring of Fairness: An active area of research [2]. Continuously examining fairness during runtime of a machine learning algorithm [3].

Research Question

Which data features are the most sensitive when monitoring fairness properties on criminal data, and how do these features perform when monitoring fairness properties on data from different sectors?
Definitions:

- **Feature:** Attribute of a dataset (age, gender, race, education level).
- **Fairness property:** A rule used to evaluate the presence of fairness.
- **Sensitivity:** The extend of a violation in a fairness property.
- **Sectors:** Criminal Justice, Healthcare, Education, Finance.

Methodology

Fairness Properties:

- **Demographic Parity:** Rate of minority group over majority group. Threshold value: **0.8** $\frac{P[Y=1 \,|\, feature = minority]}{P[Y=1 \,|\, feature = majority]} \geqslant c$
- **Equal Opportunity:** True positive outcome is the same no matter the feature . Threshold value: **0.05**

$$|P[Y_P = 1 | Y_A = 1, feature = A] - P[Y_P = 1 \,|\, Y_A = 1, feature = B]\,| \leqslant \varepsilon$$

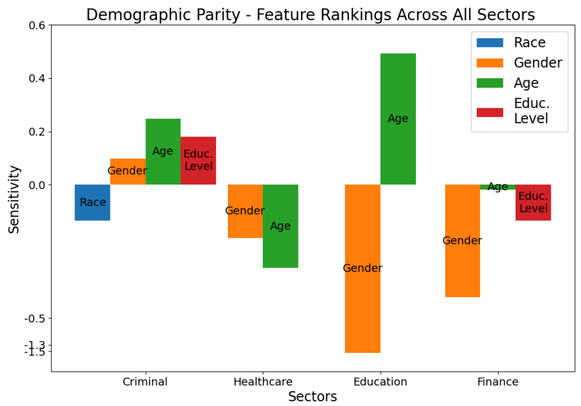
Algorithm Verification:

- Uses Logistic Regression to predict the target value of each dataset.
- Monitors both fairness properties during runtime of the algorithm and calculates sensitivity values for each feature.

Results

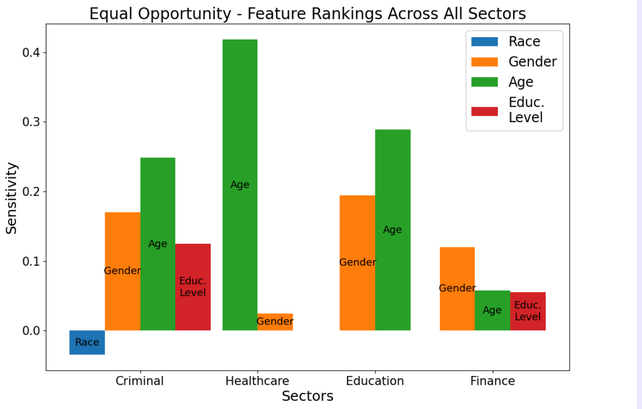
Demographic Parity:

- **Age was the most sensitive feature across all sectors with one exception in Healthcare.**



Equal Opportunity:

- **All features except race violated the fairness property.**
- **Age was the most sensitive feature across all sectors with one exception in Finance.**



Evaluation & Limitations

Statistical Analysis with Mann-Whitney U Test:

- Non-parametric statistical test: Does not assume normal distribution of the data.
- Can be used with ranked data.
- Suitable for small sample sizes.

No statistical significance was revealed despite the pattern for both Demographic Parity and Equal Opportunity

Biggest Limitation: Lack of datasets.

Conclusion

- **Key findings:** Analysis showed age as the most sensitive feature affecting fairness in machine learning across sectors, with gender and education also notable but less impactful.
- **Statistical Insight:** Differences in sensitivity across sectors were not statistically significant, highlighting the need for further investigation to confirm observed patterns.

References

1. Ninareh Mehrabi, Fred Morstatter, Nripsuta Sax- ena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. CoRR, abs/1908.09635, 2019.
2. Thomas A. Henzinger, Mahyar Karimi, Kon- stantin Kueffner, and Kaushik Mallik. Runtime Monitoring of Dynamic Fairness Properties. In 2023 ACM Conference on Fairness, Accountabil- ity, and Transparency, pages 604–614, June 2023. arXiv:2305.04699 [cs].
3. A. D'Amour et al., "Fairness Is Not Static: Deeper Understanding of Long Term Fairness via Simulation Studies," in Conference on Fairness, Accountability, and Transparency (FAT), Barcelona, Spain, ACM, New York, NY, USA, pp. 12, 2020 .