Robustness of Fairness Metrics: A Casestudy for the German Credit Dataset

Sven Maurice Morlock

Social Data Science and AI Lab
LMU Munich

July 21, 2025

Motivation

- starting point: ML algorithms might exhibit fairness issues
    - due to bias in data itself
    - amplified by ML algorithm depending on how data is processed
- for a given ML model this can be used to vary fairness metrics arbitrarily
- and hence lead to different decisions in the end
- $\rightsquigarrow$ research question: how robust are fairness metrics?
    - how to assess robustness?
    - which metric?
    - what dataset?

Assessing Robustness

- idea from [SPK24]
- consider changes in final decision
- "more robust if small changes in input lead to same output"
- focus here on preprocessing steps, e.g.
  - include/exclude observation
  - binning, scaling
  - scaling
- $\Rightarrow$ would different changes in preprocessing lead to different fairness metrics and hence different decisions?

Fairness Metrics and Dataset

- [Bel+18] provide a toolkit for implementing several fairness metrics
  - Statistical Parity Difference, Disparate Impact, Equal Opportunity Difference
- German Credit dataset used to classify credit worthiness of customers
  - in total 20 covariates and binary target(good/bad)
  - [KCP12; Fel+15] indicate issues for covariate "gender"
  - [Zli17; BS16] indicate issues for covariate "age"
  - [GBM17] indicate issues for covariate "foreign worker"

Project Agenda

- create different possible datasets by altering the input
  - include/exclude gender
  - bin age into categories
  - include exclude foreign worker
- model and dataset available at OpenML
- trace changes in fairness metric(s)

Open Questions

- only three out of 20 covariates might be an issue but interaction important?
    - many possible combinations
    - assess interaction effect
- vary model/metrics?
- just preprocessing step ($\approx$ different input data)?

References I

[Bel+18] Rachel K. E. Bellamy et al. *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias*. Oct. 2018. URL: https://arxiv.org/abs/1810.01943.

[BS16] Solon Barocas and Andrew D Selbst. "Big data's disparate impact". In: *California Law Review* 104.3 (2016), pp. 671–732. URL: https://scholarship.law.berkeley.edu/californialawreview/vol104/iss3/2.

[Fel+15] Michael Feldman et al. "Certifying and removing disparate impact". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2015, pp. 259–268. DOI: 10.1145/2783258.2783311.

References II

[GBM17] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. "Fairness testing: Testing software for discrimination". In: *Proceedings of the 2017 26th ACM SIGSOFT International Symposium on Software Testing and Analysis*. ACM. 2017, pp. 498–501. DOI: 10.1145/3092703.3098236.

[KCP12] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. "Data preprocessing techniques for classification without discrimination". In: *Knowledge and Information Systems* 33.1 (2012), pp. 1–33. DOI: 10.1007/s10115-011-0463-8.

References III

[SPK24] Jan Simson, Florian Pfisterer, and Christoph Kern. "One Model Many Scores: Using Multiverse Analysis to Prevent Fairness Hacking and Evaluate the Influence of Model Design Decisions". In: *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '24. ACM, June 2024, pp. 1305–1320. DOI: 10.1145/3630106.3658974. URL: http://dx.doi.org/10.1145/3630106.3658974.

[Zli17] Indre Zliobaite. "Measuring discrimination in algorithmic decision making". In: *Data Mining and Knowledge Discovery* 31.4 (2017), pp. 1060–1089. DOI: 10.1007/s10618-017-0517-4.