Automated Feature Engineering for Algorithmic Fairness

Ricardo Salazar
TU Berlin
ricardo.salazar@alumni.tu-berlin.de

Felix Neutatz
TU Berlin
f.neutatz@tu-berlin.de

Ziawasch Abedjan
Leibniz Universität Hannover
L3S Research Center
abedjan@dbs.uni-hannover.de

➢ Motivation

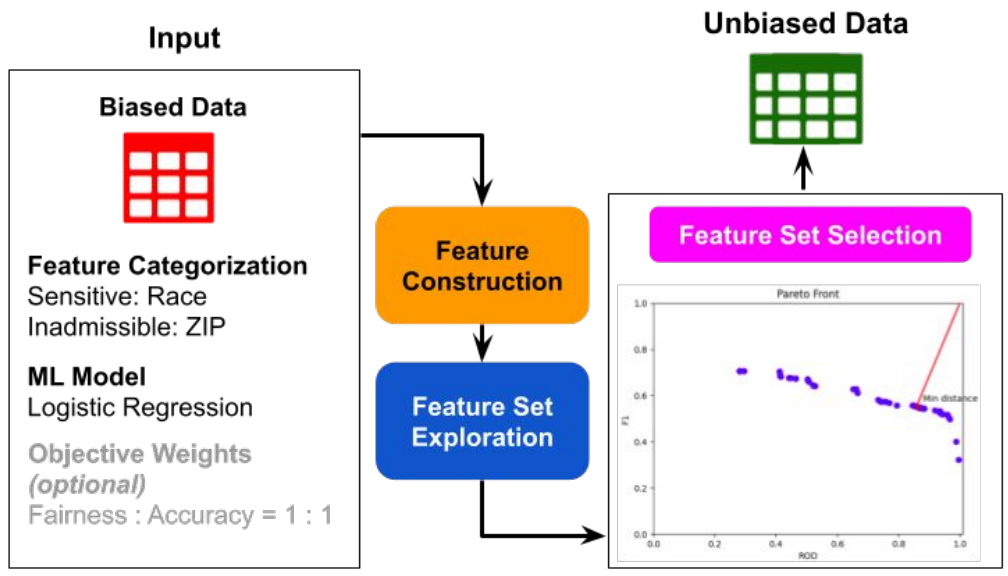
- Machine learning applications might reinforce bias against certain groups of people with a discrimination history [1, 2].
- State-of-the-art pre-processing approaches [3, 4, 5] remove bias from the training set using a horizontal strategy, i.e., adding and removing tuples.
- This horizontal approach can cause a decrease in accuracy due to information loss and might also lead to fairness overfitting.

➢ Research Questions

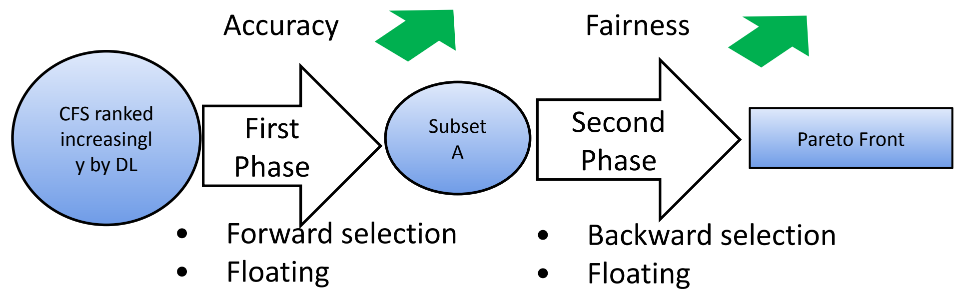
- How can we leverage feature engineering to find a viable alternative to horizontal approaches?
  - How can we generate features that replace existing inadmissible features?
  - How can we efficiently traverse the exponential space of feature transformations?

FairExp achieves competitive results compared to state-of-the-art pre-processing horizontal strategies.

➢ FairExp Architecture



1. **Feature Construction**
   - We apply recursively feature construction operators proposed by ExploreKit [6].
2. **Feature Set Exploration**



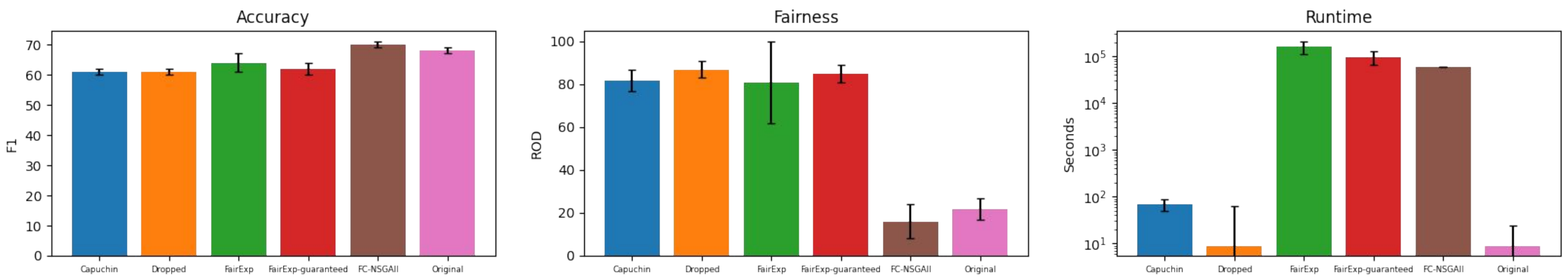
CFS: Constructed Feature Set
DL : Description Length

3. **Feature Set Selection**

$$z = \underset{x \in \mathcal{P}}{\operatorname{argmax}}\, w_{\text{fair}} * ROD + (1 - w_{\text{fair}}) * \text{F1 score}$$

➢ Experimental Results

Results for the Adult Dataset. Target: >50k usd/year; Sensitive: Sex; Inadmissible: Marital-status



References

[1] Julia Stoyanovich, et.al. 2020. Responsible Data Management. PVLDB.
[2] Julia Stoyanovich, et.al.. 2018. Panel: A Debate on Data and Algorithmic Ethics. PVLDB.
[3] Babak Salimi, et.al. 2019. Interventional Fairness: Causal Database Repair for Algorithmic Fairness. SIGMOD.
[4] Flávio du Pin Calmon, et.al. 2017. Optimized Pre-Processing for Discrimination Prevention. NeurIPS.
[5] Michael Feldman et al. 2015. Certifying and Removing Disparate Impact. KDD.
[6] Gilad Katz, et.al. 2016. ExploreKit: Automatic Feature Generation and Selection. ICDM.

Acknowledgment

The contribution of Felix Neutatz was funded by the German Ministry for Education and Research as BIFOLD - Berlin Institute for the Foundations of Learning and Data (ref. 01IS18025A and ref. 01IS18037A).