|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Paper | 1 |  | Main message of the paper | Methods and concepts |
| Selective labels | Non-representative sample, selectively labelled train data  *Contraction* as solution | They don’t use SQF data. They just study a problem the SQF data probably has. |  |  |
| Residual Unfairness[[1]](#footnote-1) | Non-representative sample, the difference between train and target pop | Model: Logistic regression  Target: criminal possession of weapon  Train pop: SQF  Target pop: whole NYC pop  PA: race | The paper argues that even when machine learning models are adjusted to meet fairness criteria like Equal Opportunity or Equalized Odds, residual unfairness can persist due to inherent biases in the training data. This residual unfairness leads to inequity of opportunity, where certain groups continue to be disadvantaged. Concepts like first-order stochastic dominance and disparate benefit of the doubt illustrate formalize systematic discrimination of certain groups into mathematical propositions. The paper emphasizes that addressing fairness at the algorithmic level is insufficient if the underlying data is prejudiced; systemic biases in the data must also be tackled to achieve true fairness. | Fairness methods: post-processing strategy that outputs group specific thresholds that guarantee equal opportunity or equalized odds between groups (in the training population)  Simulation method: they need to simulate target population that resembles whole NYC population and do this by using their own weighing technique (inspired by propensity score matching)  I could text the authors and ask for code or data (zhoua@usc.edu, [kallus@cornell.edu](mailto:kallus@cornell.edu))  Key concepts:   * Disparate benefit of the doubt (one group gets an advantage over the other by historically repeated better treatment of the group) * Inequity of opportunity: quantifies the unequality between two groups * Stochastic dominance: one group has consistently higher probability scores than the other |
| Bias in Bias out | Bias reversal and bias inheritance | Synthetic data from SQF  Task 1  Model: log reg  Target: possession of contraband (given searched)  Task 2  Model: log reg  Target: being searched  Task 3  Model: log reg  Target: searched and contraband | Bias in does actually not necessarily mean bias out but it depends on a) the data (is it representative?) and b) your task (what is the target, what do you condition on)?  They basically say that predicting weapon is affected by bias reversal (performing better on poc) while predicting stop is affected by bias inheritance (performing better on white) | Bias reversal  Bias inheritance  Three different prediction scenarios:  Predict possession of a weapon, given you were searched. Predict search. Predict weapon when you were searched. |
| Precinct or Prejudice |  | Model:  Target: possession of weapon  Train data: SQF limited on the stops due to CPW | Investigating three million stops over five years, the researchers found that in more than 40% of cases where individuals were suspected of criminal possession of a weapon, the likelihood of finding a weapon was less than 1%. They also noted that Black and Hispanic individuals were disproportionately stopped in these low hit rate contexts. |  |
| Data Transparency |  | Model: six different models (log reg, RF, …)  Target: Arrest  PA: Race, Sex | No racial discrimination in arrestment rates |  |
| Provable Detection of Propagating Sampling bias (preprint) |  | Not directly typical setting, they simulate varying degrees of differential sampling bias and use race as a target for varying degrees of bias and quantify discrimination with a Bias Scan. | Unsure to what conclusion they come for the dataset |  |
| Through Lens of Causality |  | Target: arrestment  PA: race (but they compare black Hispanic men vs white men) | They introduce two causal group fairness metrics (FACE, FACT) and don’t come to the same conclusion. FACE says there’s group level causal discrimination and FACT says there is none. |  |

1. What are the main theoretical concepts and statement the paper makes and how are they illustrated with the SQF dataset?
2. What questions with respect to the dataset are they trying to answer?/ How are they using the dataset? What is the target? What is the feature? What is the model?

In der Präsentation:

Dataset “naiv” verwenden um fairness notions klar zu machen

In der Arbeit:

Zeigen, welche challenges der Datensatz mit sich bringt, wie der Datensatz bis jetzt genutzt wurde, den Datensatz in den größeren Kontext bringen und anhanddessen argumentieren welche fairness Metriken sich hier wie auswirken (und warum sie deshalb passender oder weniger passend sind)

What can I take aways fron the two papers?

What happens if we train the logistic classifier (to predict weapon yes no) on the SQF as is (Kallus), don’t do a post processing fairness intervention (NO Hardt et. al) and test the classifier on the target population (that is created via the weighing method of Kallus and Zhou)? I think according to \cite{RambachanBBOEFW} we should observe bias reversal.

Take data from before 2013 (as example with racial bias) and data I already have and compare the predictions rates between groups.

Find out a more general data pipeline (mlr3??) so I don’t have to do the whole processing again.

1. Widersprechen sich diese zwei paper? [↑](#footnote-ref-1)