|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Paper | Key concepts | Use of SQF | Main message of the paper | Methods |
| 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) | * 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 | Model: Logistic regression  Target: criminal possession of weapon  Train pop: SQF  Target pop: whole NYC pop  PA: race | Fairness interventions on the training data (Equal Opportunity or Equalized Odds), do not guarantee fairness in the population the algortihm should be used on. This is the case when the training sample is generated via a biased process (censoring) 🡪 basically: group fairness interventions are not enough if your sample is not representative | 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)  (authors: zhoua@usc.edu, [kallus@cornell.edu](mailto:kallus@cornell.edu)) |
| 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 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)? | Simulation study: simulate varying degrees of bias (no exact details on the process used) |
| 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. |  |
| The selective labels problem | Read if still time |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

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?

What is my problem?

I am still unsure whether I am handling the data correctly, unsure what to do with the papers I found on SQF and what I have done so far. The options I currently see i) try to replicate the papers ii) use the data like I currently do, get the results, try to tie back the results to what they did in the paper

Either way i) I need to describe what exactly they did in the paper ii) as a next step I can still work with my data and see whether it paints a similar picture 🡪 if not, look for explanations why

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