Predictive Power at What Cost? Economic and Racial Justice of Data-driven Algorithms by Ranae Jabri

Discussant: Gianmarco Daniele (University of Milan – CLEAN Bocconi)

the paper

- Very relevant question
- Increase role of algorithms in policymaking and the justice system
- Machine learning as a «Black box»: fairness concerns

- Recidivism risk score: COMPAS
- Inclusion of neighborhood variables (where the defendent lives) marginally increases prediction but it widen disparities in defendant risk scores and false positive rates for black defendants

comments

- Ideal experiment: same data and same algorithm as COMPAS
- However, COMPAS algorithm is proprietary, so…the paper tries to replicate COMPAS

- We don’t know how different are the data used by COMPAS
- We don’t know how different is the algorithm used by COMPAS
- The model explains 64–65 percent of the total variation (R-squared) in the COMPAS raw risk scores
- Overall concern: limited knowledge about the importance of those omitted variables (included in COMPAS but not in the paper)
- Their interaction with neighborhood variables might or might not increase the relevance of census-track variables for the prediction

comments

- How do input variables predict recidivism outcomes out-of sample?

- Why do you use only logistic models and not more complex models (e.g. Lasso, Ridge, XGBoost, etc.)?
- More details on the prediction: how do you split the sample in training and test?
- Why not looking also at ROC as measure of performance?
- What is the object function of COMPAS? Maximise prediction (recall)? Or maximize precision (reduce false positive)?

comments

- COMPAS score might affect pre-trial detention which might affect recidivism, this complicates the interpretation as you might observe recidivism because they get a high score
- Ideal complementary exercise: predict recidivism with/without neighborhood data on a sample without COMPAS or any other score

comments

- neighborhood variables increase false positive for black defendants (3p.p.) but they reduce false positive for white people (by a larger magnitude 6p.p.) worth discussing? Any guess why is this the case? It suggests removing those variables is not a Pareto improvement.

- P. 16: "The neighborhood fixed effects account for 6.7 percent of the total variation explained by the linear model.": is this really a negligibly improvement? What would it be a substantial improvement?