A study of influence of inputs on automated recidivism decisions

Karina Huang, Lipika Ramaswamy
Institute for Applied Computational Science, Harvard University

INTRODUCTION

- Automated decision making has gained popularity in hiring and sentencing procedures. Fairness of automated processes, however, remains an open question due to limited interpretability of the black box models.
- The Quantitative Input Influence (QII) framework (Datta et al., 2016) offers effective and straight-forward explanations of algorithms by quantifying the influence of input features on individual and group observations.
- Influence is computed as a difference between the actual feature distribution and a random feature distribution.
- We applied this framework to classification algorithms, using criminal records of 7,214 individuals in Broward County, FL, to investigate fairness of COMPAS assignment of decile scores.
- We modeled the assignment of violence decile scores using individuals' *Age*, *Sex*, *Race*, *Priors*, and *Misdemeanor Counts*. We trained three classifiers and report on the Decision Tree.

METHOD I: UNARY QII

- **Unary QII on individual outcomes** communicates the average influence of a feature on all individuals in the dataset. Intuitively, this influence represents how important a feature is to predicting the outcome.

$$\mathbb{E}(i \text{ is pivotal for } c(X)) = \sum_{\mathbf{x} \in X} Pr(X = x) \cdot \mathbb{E}(i \text{ is pivotal for } c(X) | X = \mathbf{x})$$

- **Unary QII on group outcomes/group disparity** denotes the influence of a feature on a group of individuals.

$$\iota^{\mathcal{Y}}_{disp}(i) = Q^{\mathcal{Y}}_{disp}(X) - Q^{\mathcal{Y}}_{disp}(X_{-i}U_i)$$

- The quantity of interest for unary QII is the probability of being predicted a given decile score category.

RESULTS

Figure 1. Unary QIIs on Individual Outcomes

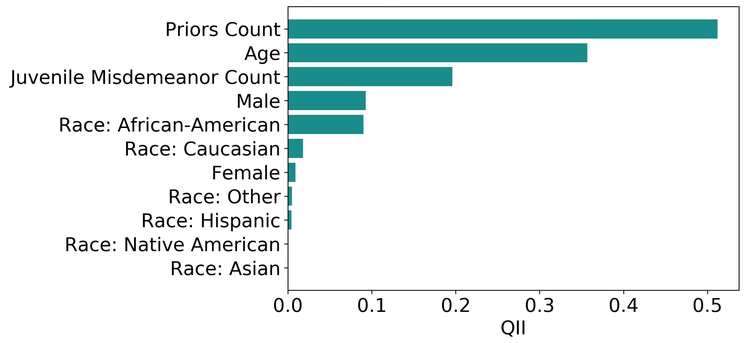


Figure 2. Unary QII on Group Disparity – Caucasian vs. Other

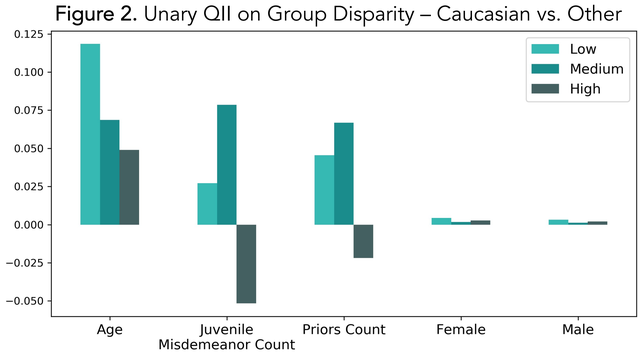
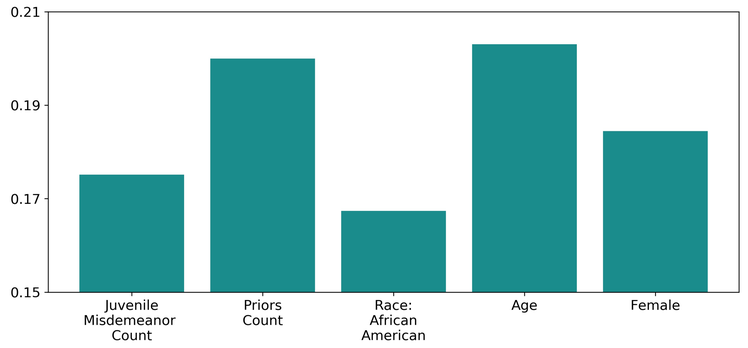


Figure 3. Transparency Report for Ms. X

Ms. X's Information	
Age	26
Juvenile Misdemeanor Count	0
Juvenile Felony Count	0
Juvenile Other Count	0
Priors Count	2
Violence Decile Score (Category)	4 (Low)
Race	African-American
Two-Year Recid	Yes



METHOD II: MARGINAL QII

- Transparency reports can provide individuals with detailed information on (un)expected classification outcomes.
- Marginal QII, reflected in transparency reports, represents the influence of a single feature within a subset of features.

$$\iota^{\mathcal{Q}}(i, S) = Q(X_{-S}U_S) - Q(X_{-S\cup\{i\}}U_{S\cup\{i\}})$$

- One influence measure for marginal QII is the Shapley value. This metric is grounded in cooperative game theory.

DISCUSSION

Unary QII

- Figure 1 shows that the feature with the highest influence is *Priors Count*, followed by *Age* and *Juvenile Misdemeanor Count*.
- Group disparity reports that for predicted low and medium decile scores, these three features appear to bias for or against Caucasians.
- Note that the direction of bias is not captured by this measure. Further, the magnitude of bias is arguably minor given the scale.

Marginal QII

- A transparency report is presented for Ms. X in Figure 3. The Decision Tree classifier predicted a high violence decile score for Ms. X, when her assigned violence decile score was low.
- The result might be surprising at first glance. However, the transparency report shows that the most influential features that led to this classification were *Age* and *Priors Count*.
- This alerts us of the problematic behaviors of the underlying classifier, which heavily weights *Age*.