Model Selection for User-Level Targeting Models based on Heterogeneous Treatment Effects

Alex Wood-Doughty and Tianqi Wang

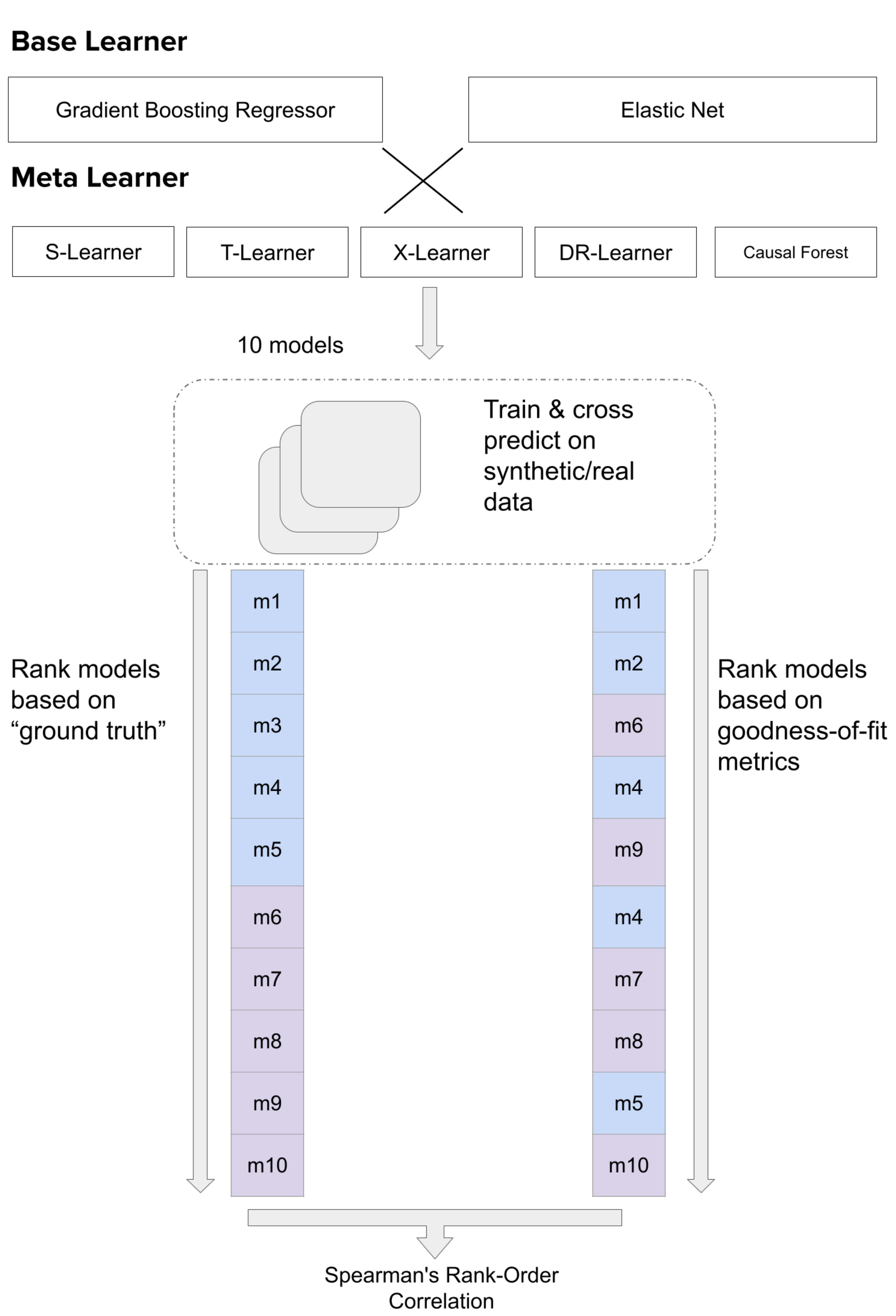
tl;dr

- User-level targeting is a common use-case for HTE models
- We define a new goodness-of-fit metric based on Off-Policy Evaluation (OPE)
- We show with synthetic and real data that this new metric outperforms existing methods on targeting problems

Motivation

- HTE goodness-of-fit is a challenging problem
  - Most literature focuses on the Precision of Estimating Heterogeneous Effects (PEHE)
    - $\mathbb{E}[(\tau - \hat{\tau})^2]$
  - We don't observe ground truth, so need to define metric that approximates PEHE
- A common use-case for HTE models is user-level targeting
  - e.g. marketing, personalized medicine, etc.
  - For targeting models, we care more about users who are near the decision boundary
  - PEHE equally weights all users, so potential to do better for targeting applications
- Gaps exist between practitioners and literature
  - Recent literature on improvements over R-Loss, but complicated
  - Most open-source libraries for HTE models use AUUC or R-Loss

Methodology



Outcomes (ground truth)

- Precision of Estimating Heterogeneous Effects (PEHE)
  - $\mathbb{E}[(\tau - \hat{\tau})^2]$
- Targeting (with known τ)
  - Sum up τ for top 50% of users
- Off-Policy Evaluation (OPE)
  - For a hypothetical policy a (e.g. give treatment to top 50% of users)

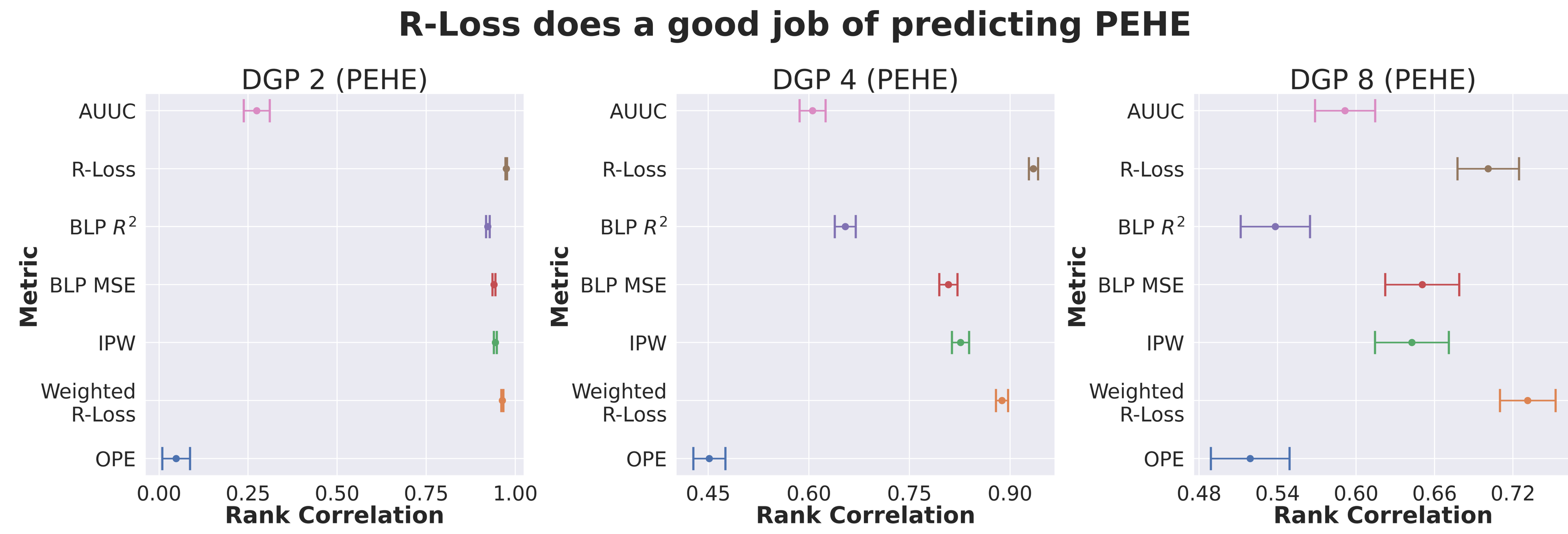
Goodness-of-Fit Metrics

- Area Under the Uplyft Curve (AUUC)
- R-Loss
- Best Linear Predictor R^2 and MSE
- Inverse Propensity Weighted Transformed Outcome (IPW)
- Weighted R-Loss
  - Upweight users who are near the decision boundary
- OPE (Doubly Robust)
  - For a hypothetical policy a (e.g. give treatment to top 50% of users)
  - $\mathbb{E}[(Y - \hat{Y}(a))\frac{\mathbb{1}(a=a')}{\hat{\pi}} + \hat{Y}(a')]$

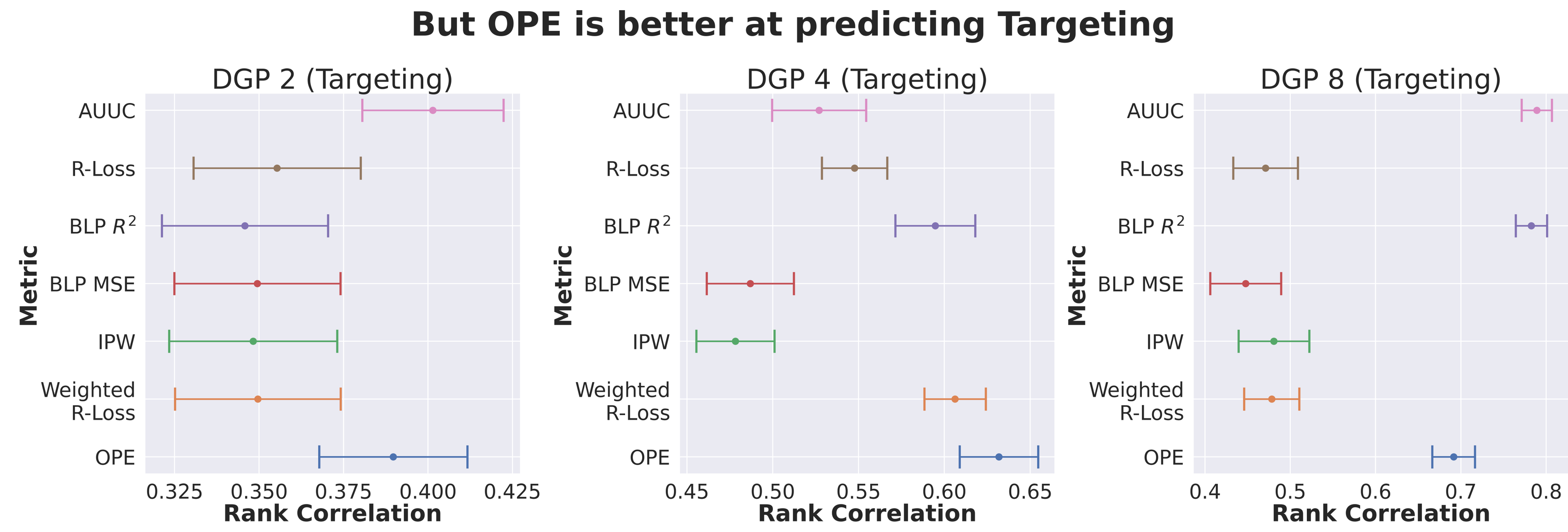
Synthetic Data

- Following Powers et al
  - N = 3000, split evenly into Train, Val, Test
  - 10 features, half standard normal, half Bernoulli(0.5)
  - 8 different DGPs
    - 8 different combinations of functions, one for tau (treatment effect), one for mu (baseline response).
    - All have random assignment
  - 100 bootstraps

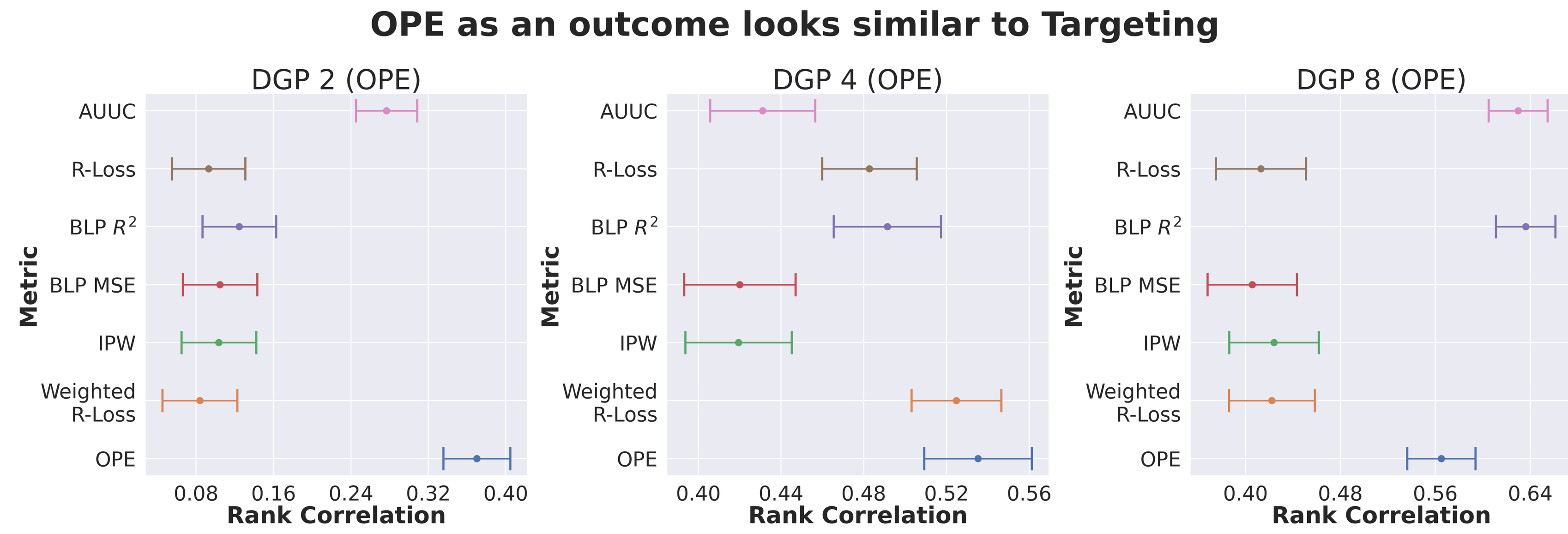
R-Loss does a good job of predicting PEHE



But OPE is better at predicting Targeting



OPE as an outcome looks similar to Targeting



Real Data

- Data from a Lyft incentive experiment
  - 174K observations, split evenly between treatment and control
  - Two outcomes (gain and cost), 49 features
  - 100 bootstraps
- Define "ground truth" as OPE estimate of profit-max allocation
  - multiplier $*$ gain $-$ cost > 0

OPE is much better on real data