Learnings from scaling Causal ML at Netflix

Adrien Alexandre
October 2024, NABE TEC

Applications of Causal ML at Netflix

Proxy Metric Research

Given a North Star Metric, e.g. Revenue or Retention, can we derive a more sensitive short-term metric that is predictive of that North Star?

Trade-Off evaluation

Given two conflicting A/B test Metric (Good Signal + Bad Signal), which rollout decision should we opt for?

Valuation of endogenous user actions

Given a set of user actions (e.g. “Thumbing”), can we value and rank each action?

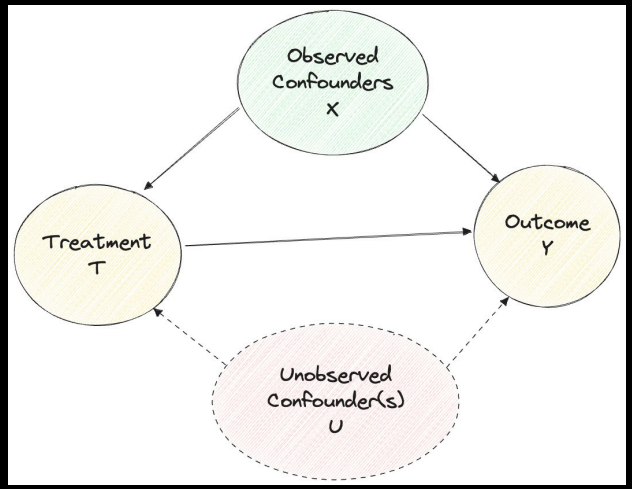
Reasons to prefer CausalML over Experimentation.

1. Experiments are costly ($$)
2. It’s often difficult to induce variation in our metric of interest.
3. What about Meta-Analysis of past experiments? Possible: but noisy, and correlated measurement errors makes causal interpretation tricky.

In contrast, CausalML (with observational data) provides a cheap and easy way to compute causal* estimates without such constraints.

CausalML also has challenges ...

Confounding Bias



Model misspecification can result in improper causal identification.

$$R_i = \beta A_i + f(X_i) + \varepsilon_i$$

Example: Partially Linear Model (PLM) is unable to recover the true treatment effect of A -> R when they are heterogeneous.

Netflix has many, many examples of non-constant treatment effects. For example, **the effect of playing a game will be very different for first-time gamers than for frequency gamers.**

Design principles of scaling CausalML at Netflix

Accessible… ish.

CausalML should feel “hard”. We want to enable broad access, but also prompt the right questions from the users.

Foreground diagnostics - not results.

We make it easy for users to get their analysis results, but we print out robustness checks before the results themselves.

Centralized support

We want to scale our solution and enable users, while providing access to a centralized group that can help answer questions and think deeply about the intricacies of specific use-cases.

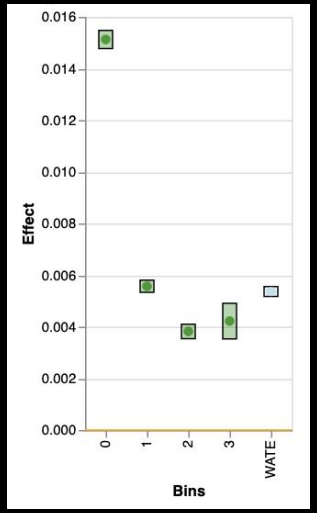
Our solution in practice

1. Internal package with “opinionated” models that scale well across Netflix use cases.

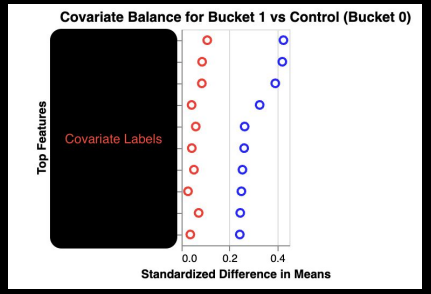
2. Config-driven notebook that produces estimates quickly with proper diagnostics

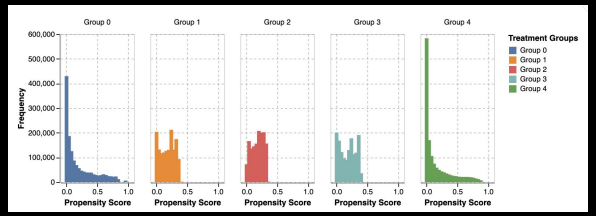
3. A data component to help with assembling required datasets.

Results



Diagnostics





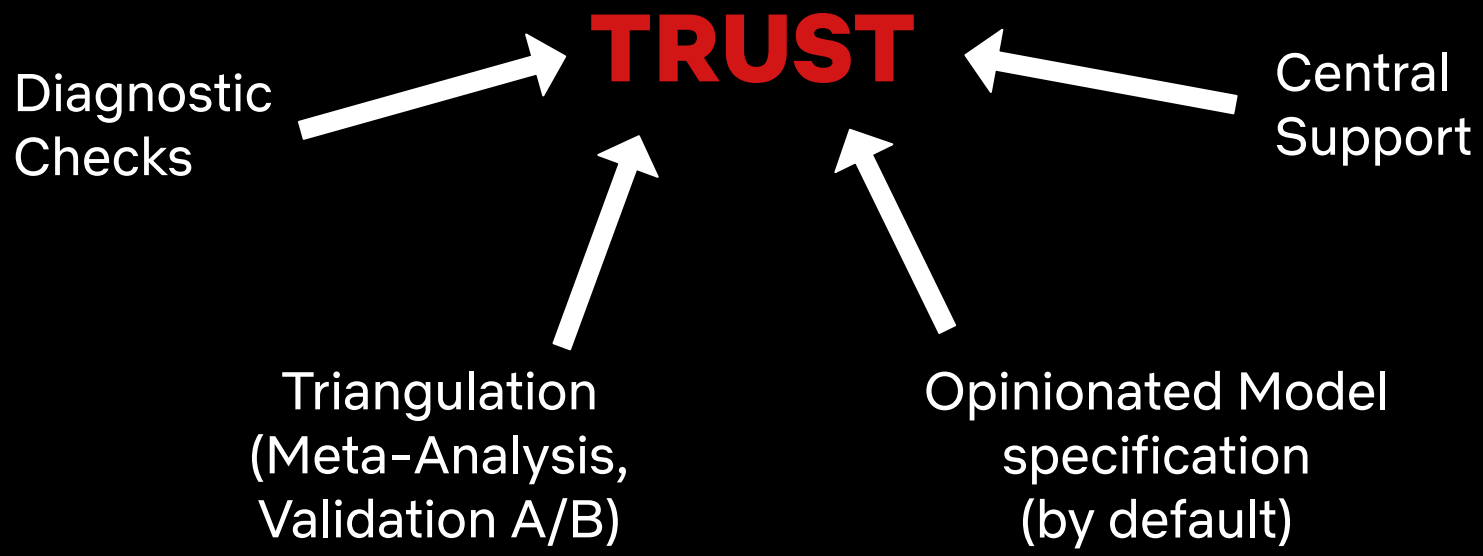
SENSITIVITY ANALYSIS

Treatment robustness value (RV%) = 1.702%

This means that omitted variables that explain more than RV% of the residual variation both of the treatment (`cf.d`) and of the outcome (`cf.y`) are sufficiently strong to make the estimated effect interval include 0.

For context, **benchmark covariates** explain 1.08% of the outcome variation and 1.73% percent of the treatment variation above the base set of confounders included in the model.

Learnings from producing and consuming Causal ML at Netflix



THANKS