Using Causal ML to select articles for promotion on digital channels

Joel Persson, Research Associate @ ETH Zürich (Switzerland)
Dr. Cristina Kadar, Data Science & Machine Learning PO @ NZZ (Switzerland)

WAN-IFRA Data Science Expert Group
Webinar, November 2023

Why do we need causal modelling?

Digitalization has enabled data-driven decision making in the media industry

In the past decades we have witnessed vast improvements in data collection, storage, and processing.

This enabled common decision making use cases to be data-driven:

- **Which** content should be promoted on the frontpage, social media channels, etc.?
- **When** should a paywall be shown?
- **How much** should a certain subscription be priced?

Machine learning is increasingly used, yet it is not designed for decision making

Recent access to more data and computational power has enabled algorithms based on machine learning (ML)

Problem:

- ML learns a model from many past examples (fits a function) to predict outcomes from correlations. In standard application, it makes predictions given the status quo (e.g., forecasting).
- Decision-making requires predicting outcomes caused by different actions ("treatment" vs. "control"). Thus, we are interested in predicting changes in outcomes (i.e., the causal treatment effect or uplift).

Ex: In targeting users to reduce churn, we should not target those with highest predicted likelihood of churn, but those with the largest reduction in likelihood to churn if they receive an offer vs. if not.

=> For data-driven decision-making, ML needs causal inference (the study of designs and estimation methods needed for drawing causal conclusions from data).

Designs for getting data: RCTs and Observational Studies

Randomized controlled trials (A/B tests) are the “gold standard” for estimating treatment effects.

Pros:

- Randomization ensures causality
- No modelling required
- Easy to implement

Cons:

- Provides an average treatment effect
- “Costly” (wrt profits, logistics, ethics, etc.)
- Delayed learning (must wait until the end)

Observational studies (historical data with non-random treatment) sometimes the feasible approach.

Pros:

- No “cost” from randomization
- Often much larger sample size
- No delay; can use existing data

Cons:

- Req’s controlling for all confounders
- Req’s specifying a regression model
- => Prone to bias from misspecification, omitted variables, and estimation errors

Statistical methods for estimating treatment effects

Historically, methods for estimating treatment effects relied on parametric statistical models.

Ex: The **average treatment effect** is □ in a linear regression model

- **RCT:** Outcome = intercept + □ × treatment + controls + error
- **Obs study:** Outcome = intercept + □ × treatment + confounders + controls + error

Ex: The **heterogeneous treatment effect** function is □ **× moderators** in a linear regression model

- **RCT:** Outcome = intercept + □ × moderators × treatment + controls + error
- **Obs study:** Outcome = intercept + □ × moderators × treatment + confounders + controls + error

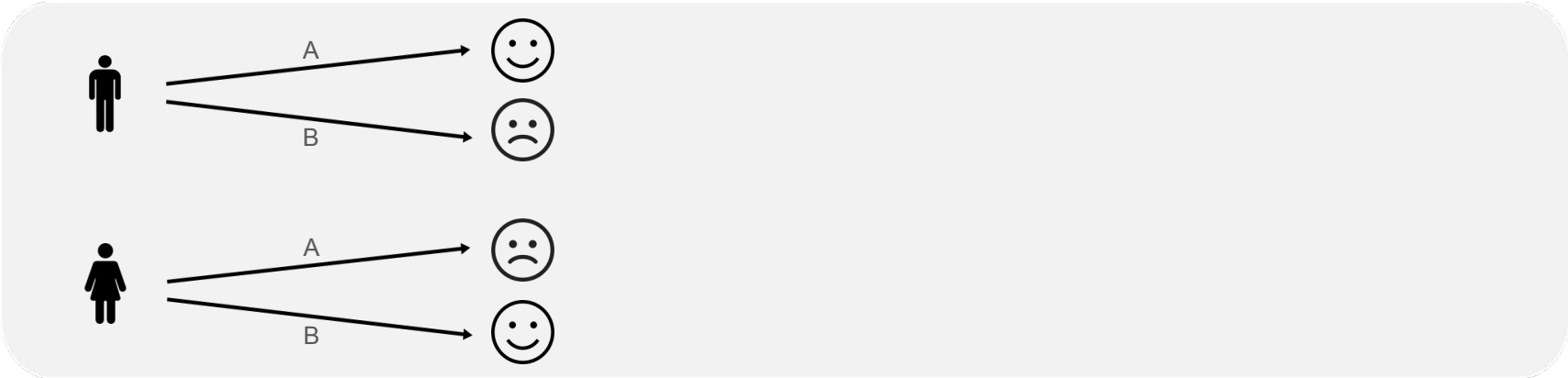
Challenge:

- Unbiased estimates (i.e., get true estimate on average) req's the specified model to be correct
- Valid inference (e.g., p-values, conf intervals) req's specifying the model before seeing the data

Latest estimation methods: ML

Lately, ML methods have been adapted to causal inference to address these challenges.

Causal ML improves estimation of heterogeneous treatment effects by framing it as a prediction problem.



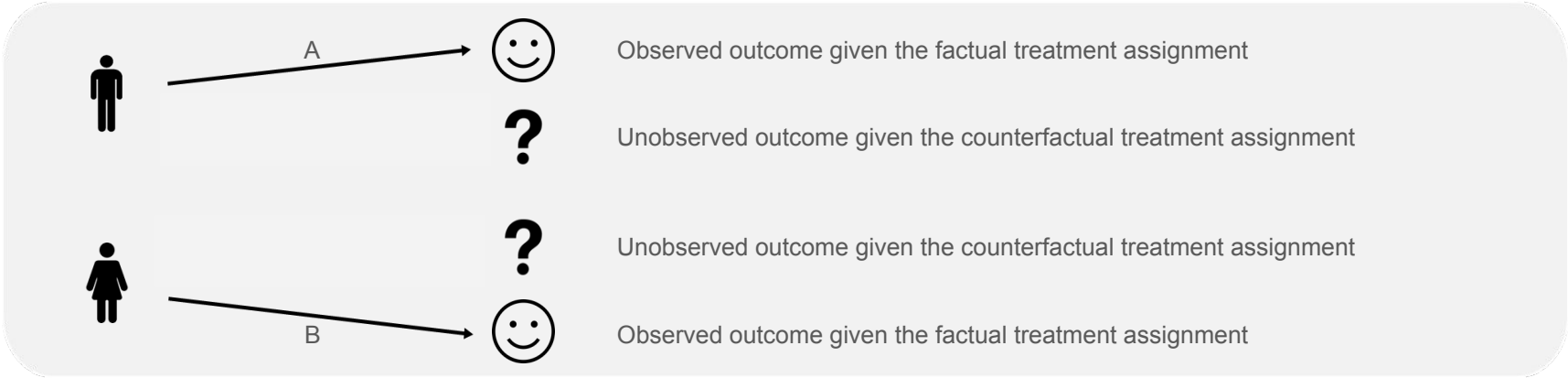
Benefits of Causal ML

- Data-driven: Learns the model that best fit the data => less prone to bias from incorrect model
- Regularization: Automatically selects which variables are moderators, controls, and confounders
- Prediction performance: Often better at out-of-sample prediction under treatment or control

Latest estimation methods: ML

Lately, ML methods have been adapted to causal inference to address these challenges.

Causal ML improves estimation of heterogeneous treatment effects by framing it as a prediction problem.



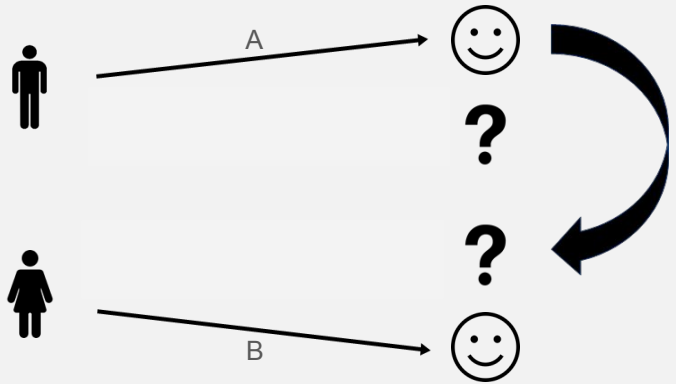
Benefits of Causal ML

- Data-driven: Learns the model that best fit the data => less prone to bias from incorrect model
- Regularization: Automatically selects which variables are moderators, controls, and confounders
- Prediction performance: Often better at out-of-sample prediction under treatment or control

Latest estimation methods: ML

Lately, ML methods have been adapted to causal inference to address these challenges.

Causal ML improves estimation of heterogeneous treatment effects by framing it as a prediction problem.



Predict the unobserved outcome using data on similar users with the counterfactual treatment assignment

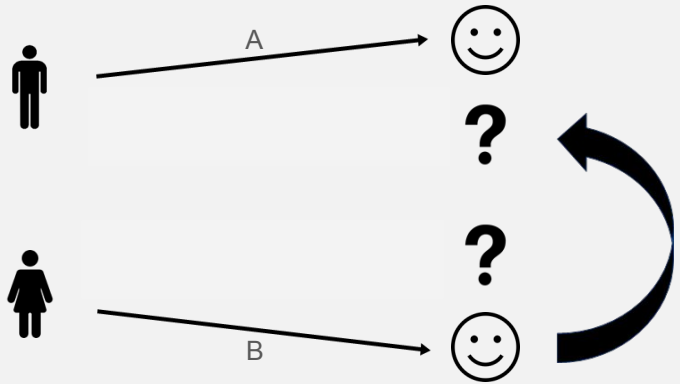
Benefits of Causal ML

- Data-driven: Learns the model that best fit the data => less prone to bias from incorrect model
- Regularization: Automatically selects which variables are moderators, controls, and confounders
- Prediction performance: Often better at out-of-sample prediction under treatment or control

Latest estimation methods: ML

Lately, ML methods have been adapted to causal inference to address these challenges.

Causal ML improves estimation of heterogeneous treatment effects by framing it as a prediction problem.



Predict the unobserved outcome using data on similar users with the counterfactual treatment assignment

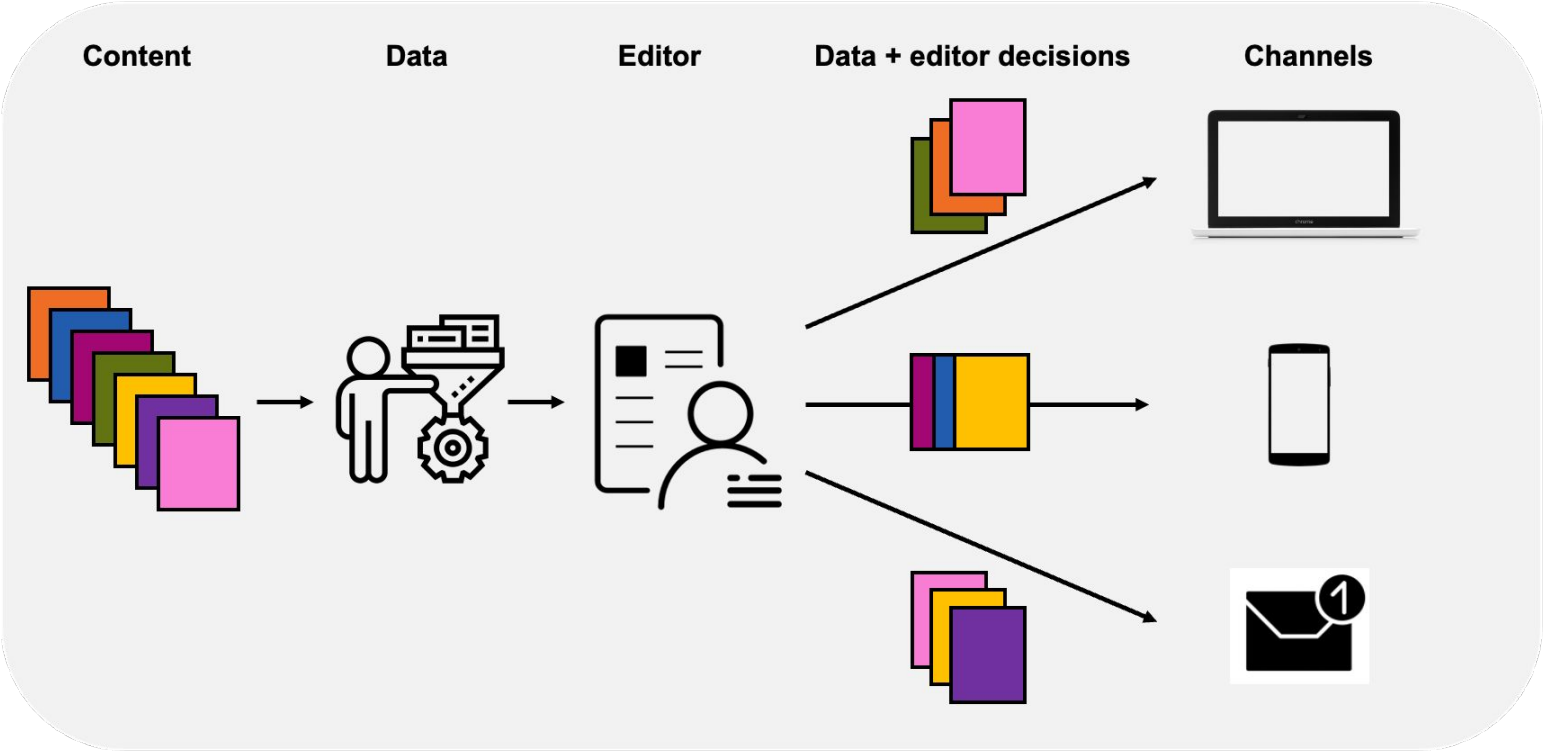
Benefits of Causal ML

- Data-driven: Learns the model that best fit the data => less prone to bias from incorrect model
- Regularization: Automatically selects which variables are moderators, controls, and confounders
- Prediction performance: Often better at out-of-sample prediction under treatment or control

NZZ’s use case:

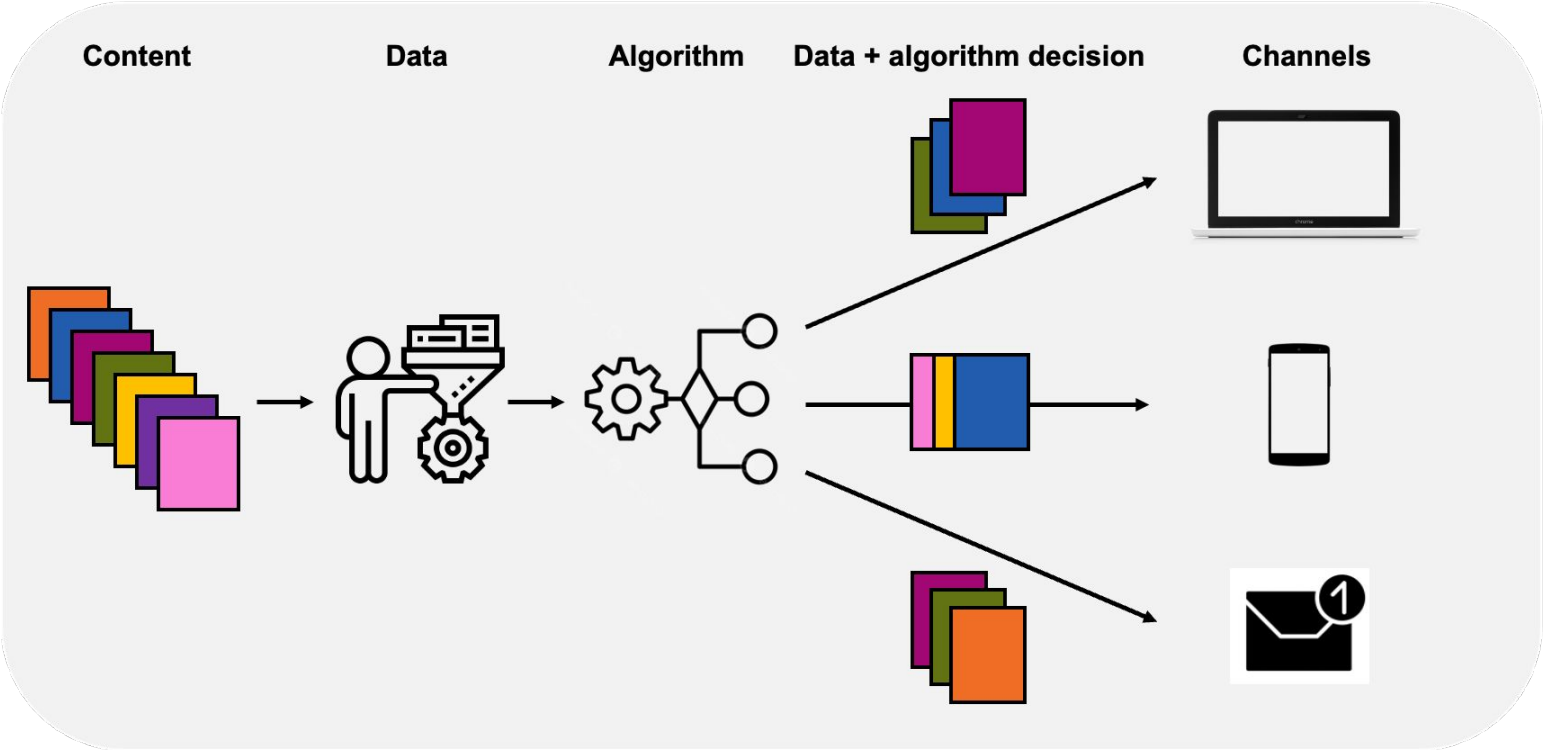
Editorial content promotion on the digital channels, such as the CH and DE frontpages

Current setup

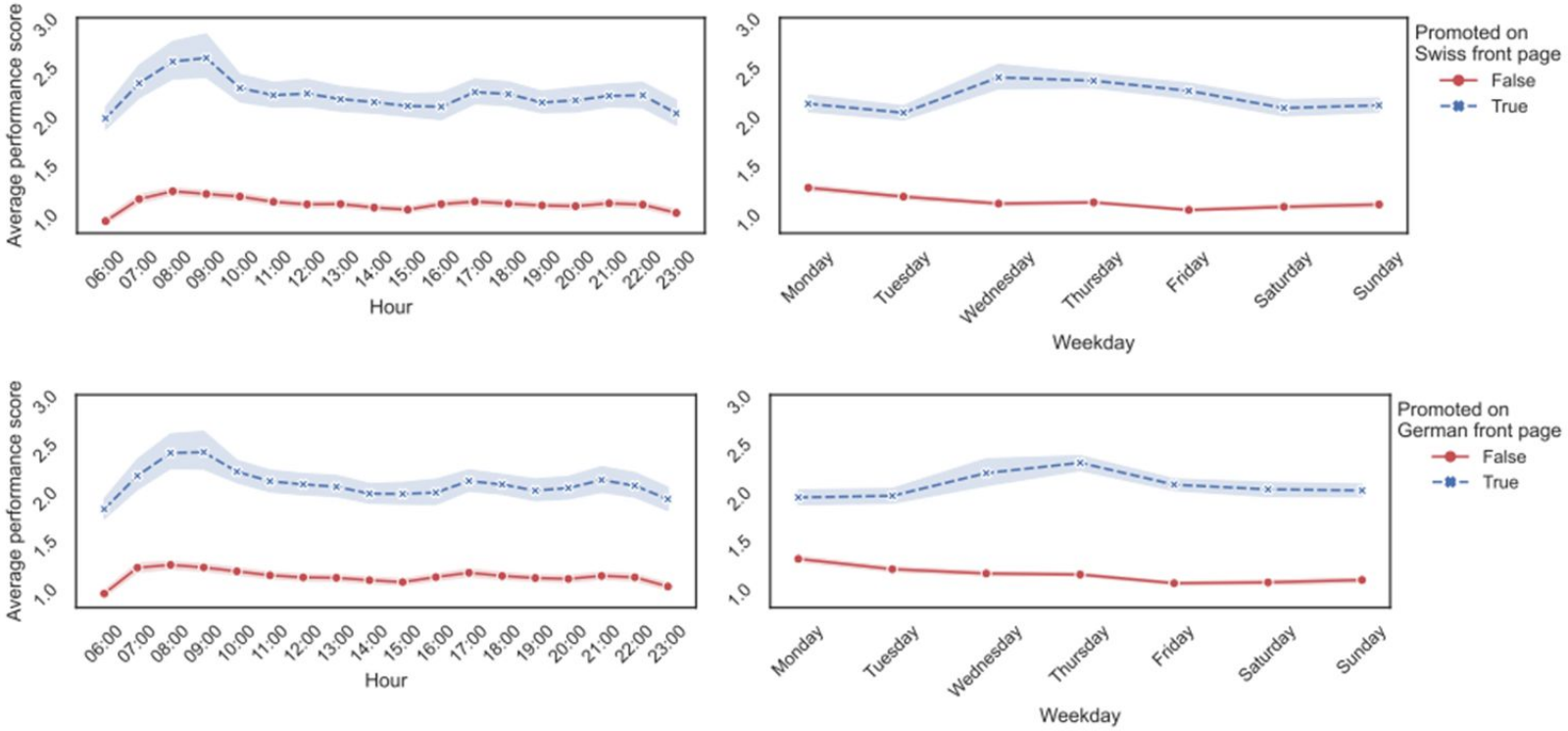


Current system used by the editors to decide what to promote when and where

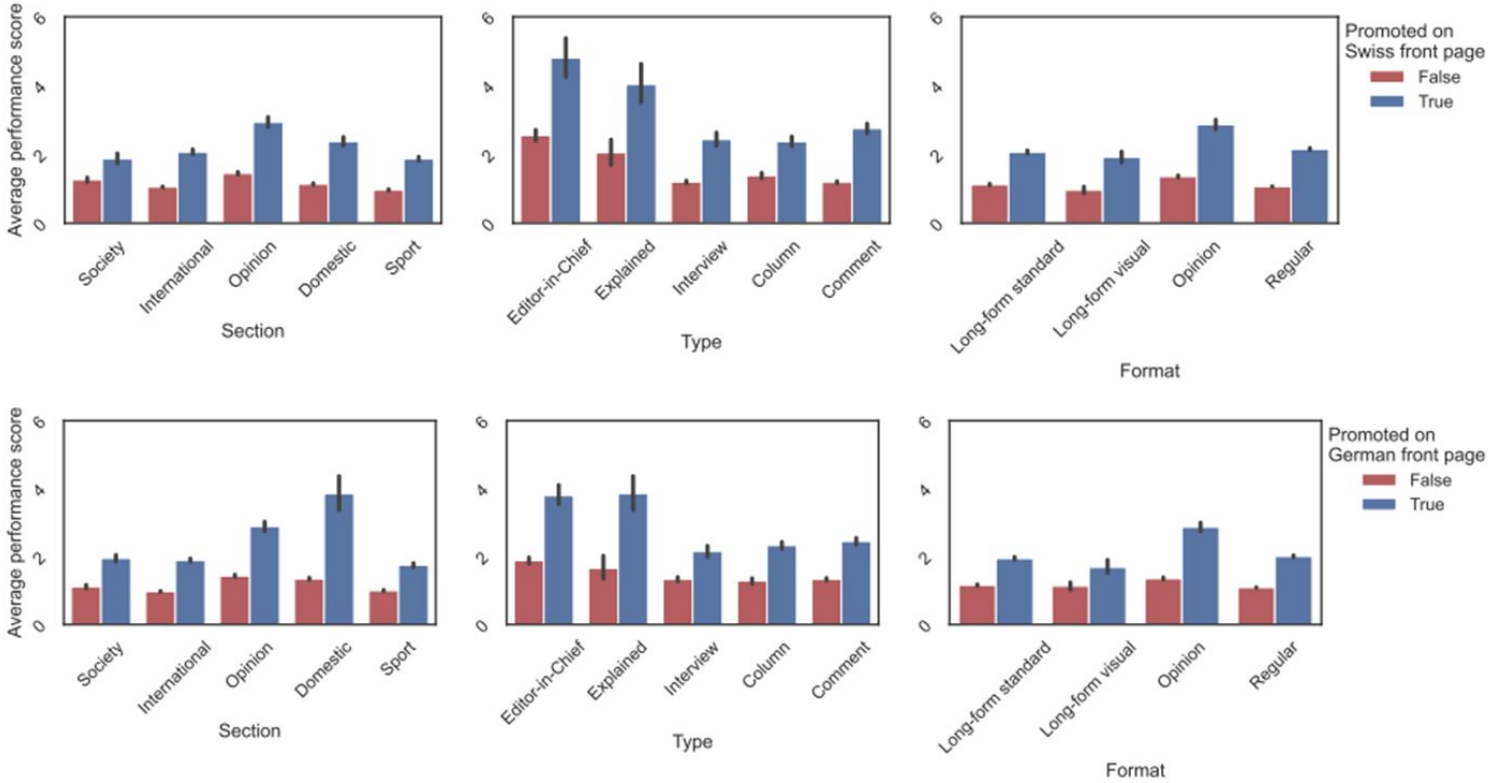
Envisioned setup (e.g. with human in the loop)



Descriptive statistics: promoted articles have higher performance score than non-promoted articles on the CH and DE frontpages



Descriptive statistics: promoted articles have higher performance score than non-promoted articles across sections, types, and formats



Causal inference approach

To learn what content is best to promote where and when, we must learn the heterogeneous uplift of promoting content. This implicitly requires thinking about counterfactuals.

*What are the potential **outcomes** (performance scores) for the different **choices** (promoted Y/N)?*

*Given the potential outcomes from the different choices, take the action with the expected greatest **uplift** and promote those articles that would receive the greatest benefit from being featured on the frontpage.*

Step 1: Estimation of the nuisance models on training data

Outcome model:

Gradient Boosting Regressor

$$\textbf{performance\_hat}_{a,t} = \textbf{f}(\textbf{promotion}_{a,t}, \textbf{metrics}_{a,t-1}, \textbf{article\_covariates}_{a}, \textbf{time\_covariates}_{t})$$

Propensity score model:

Gradient Boosting Classifier

$$\textbf{promotion\_hat}_{a,t} = \textbf{g}(\textbf{metrics}_{a,t-1}, \textbf{article\_covariates}_{a}, \textbf{time\_covariates}_{t})$$

Step 2: Estimation of the uplift model on the training data

We combine the predictions from the outcome and propensity score models to a “doubly robust (DR) score” of the predicted performance over the coming hour:

$$DR\_score_{a,t} = performance\_hat_{a,t} + (1\{\text{promotion decision in the data} = a\}_{.t} / promotion\_hat_{a,t}) \times (performance\_hat_{a,t} - performance_{a,t})$$

Main prediction term

Bias correction for prediction errors by the nuisance models

We get the predicted uplift by taking the difference in DR_score given promotion or not:

$$\text{Predicted uplift}_t = DR\_score_{a = \text{promoted},t} - DR\_score_{a = \text{not promoted},t}$$

Finally, we fit an uplift model so that we can predict uplifts for content in the test data:

$$\text{Predicted uplift}_t = h(metrics_{a,t-1}, article\_covariates_a, time\_covariates_t)$$

ForestDRLearner

Step 3: Estimate optimal policy via simulation on the test data

Recommend top-N articles with the highest predicted uplift as per the model in step 2 for promotion on the frontpage in the next hour.

Optimal promotion policy$_{a,t}$ =

- **IF** predicted **uplift rank$_{a,t}$** > N-1, **Promote**
- **ELSE**, **Do not promote**

Compare this optimal policy to the behavior policy (current practice of the editors).

In the simulation, our method makes promotion decisions that outperform the editors' historical decisions at all points in time

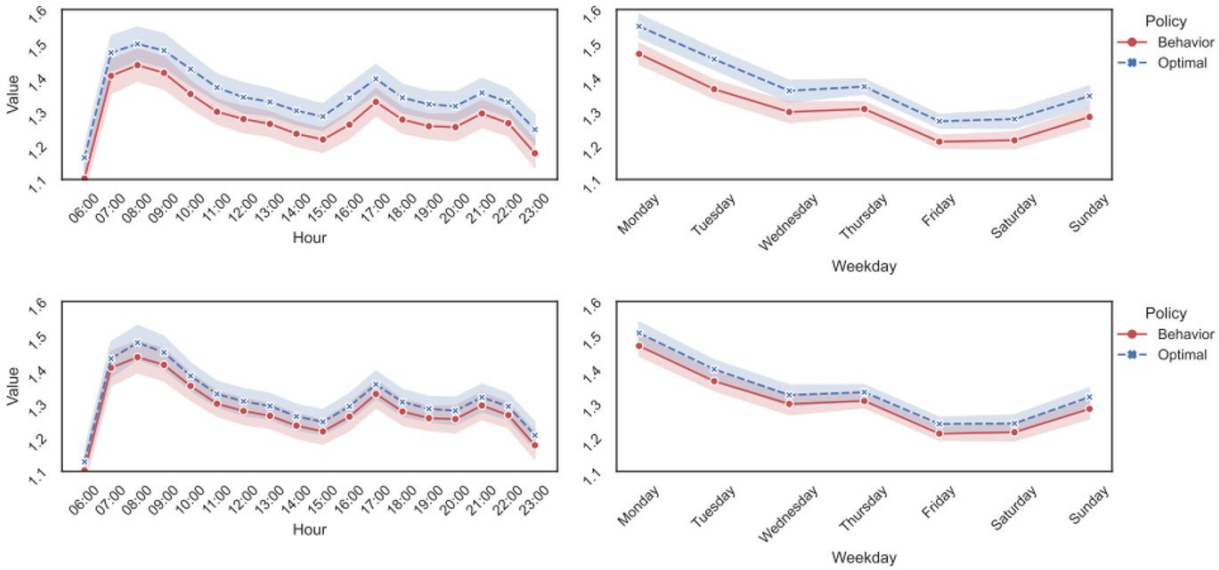
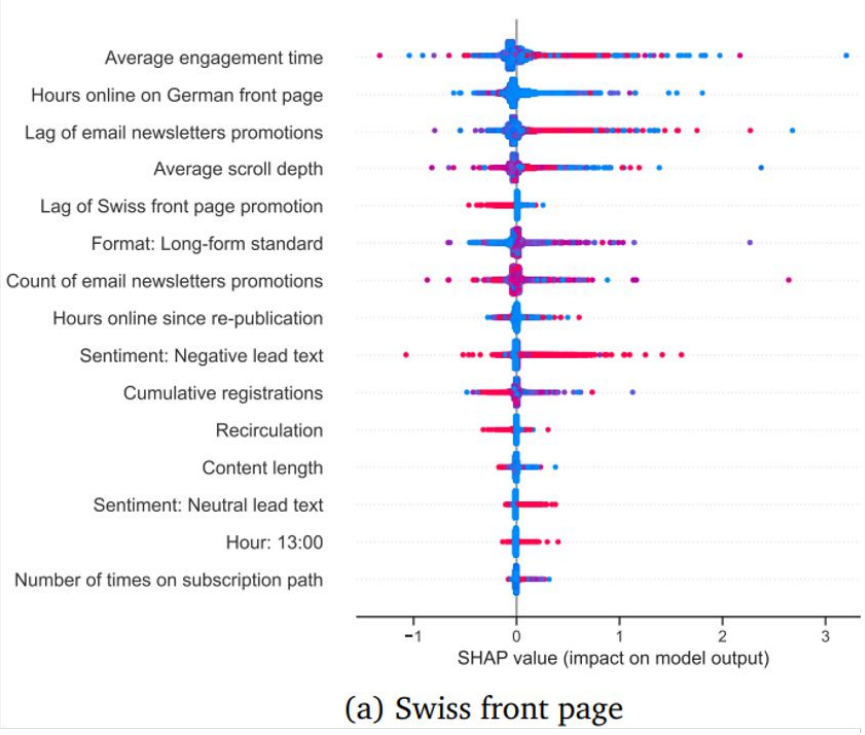
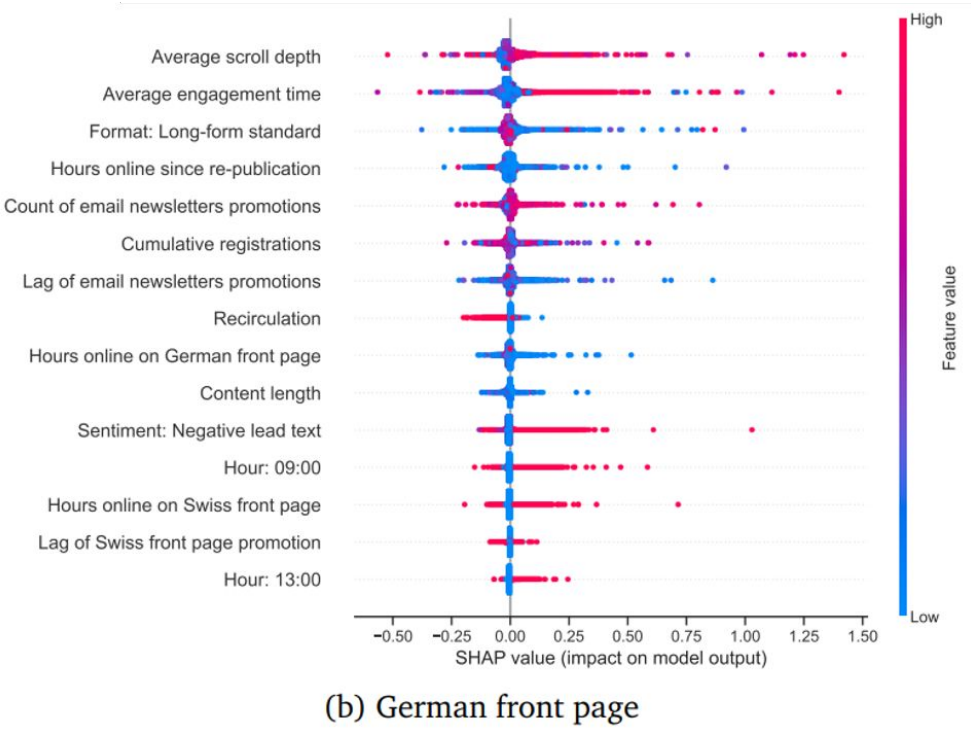


Figure 3.4: Value of the behavior policy (current practice) and our optimal policy by hour and day of the week. Shaded regions are 95% confidence intervals across 1000 bootstrap runs. Top: Swiss front page. Bottom: German front page.

Which content features explain the heterogeneous treatment effect of the promotion and thus the optimal promotion policy?



(a) Swiss front page



(b) German front page

Thank you for your attention!