Metalearners for uplift modeling

ECML/PKDD'22 Uplift Modeling Tutorial

Wouter Verbeke & Szymon Jaroszewicz

Metalearner

- Modeling **strategy** or **framework** to estimate the **conditional average treatment effect (CATE)** that can be implemented with **any ML method**
  - Base learner
  - Cfr. ensemble methods

- Different metalearners:
  - **T-learner**
  - **S-learner**
  - **X-learner**
  - R-learner
  - DR-learner
  - …

- Appropriate learner? Depends on the data generating process!

T – learner

1. Estimate **two separate models** for the **two groups (C & T)** separately, to estimate the average outcomes $\mu_0(x)$ and $\mu_1(x)$:

$$\mu_0(x) = \mathbb{E}(Y(0)|X = x) \text{ using } \{X_i, Y_i\}_{T_i=0}$$
$$\mu_1(x) = \mathbb{E}(Y(1)|X = x) \text{ using } \{X_i, Y_i\}_{T_i=1}$$

For binary treatment variable!
For *any* type of outcome variable

- Two models can have different base learners as well as variables X_i

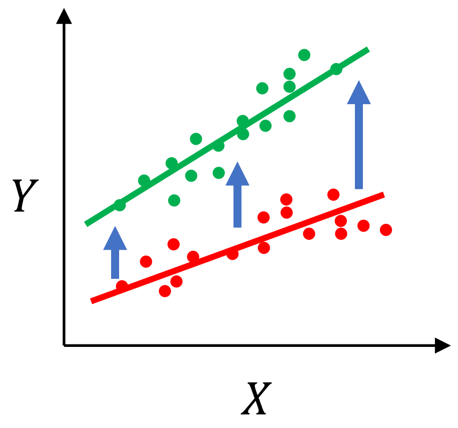
2. Obtain CATE estimate as follows:

$$\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$$

- Appropriate when ***response surfaces*** are different
- UM literature: Two-model approach

Based on: https://causalml.readthedocs.io/en/latest/methodology.html

T – learner



Control group

Treatment group

$$Y^{T=1} = \beta_0^{T=1} + \beta_1^{T=1}X$$

$$Y^{T=0} = \beta_0^{T=0} + \beta_1^{T=0}X$$

$$\hat{\tau}(x) = \beta_0^{T=1} + \beta_1^{T=1}X - \beta_0^{T=0} + \beta_1^{T=0}X$$

S – learner

1. Estimate the average outcomes $\mu(x,t)$ for both control and treatment groups with a **single** model:

$$\mu(x,t) = \mathbb{E}(Y|X = x, T = t)$$

2. Obtain CATE estimate by imputing $T = 1$ and $T = 0$:

$$\hat{\tau}(x) = \hat{\mu}(x,1) - \hat{\mu}(x,0)$$

- Appropriate when response surfaces are ***similar***
- **Propensity scoring** can be applied to reduce treatment assignment bias
- UM literature: treatment dummy approach

For *any* type of treatment variable
For *any* type of outcome variable
(continuous, multiple, time-dep., …)
BUT: effect → baseline!

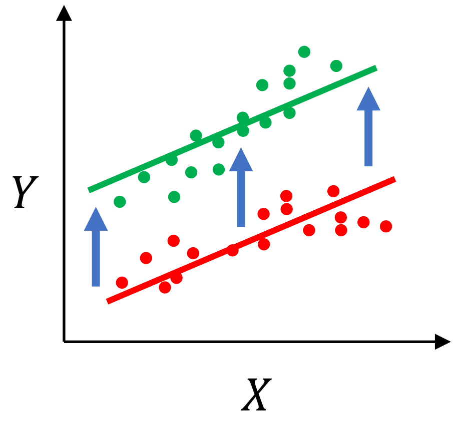
Based on: https://causalml.readthedocs.io/en/latest/methodology.html

S – learner

- Limited flexibility?
- Add interaction terms to extend model flexibility (ATE > CATE)

Control group

Treatment group



$$\hat{\tau}(x) = \beta_0 + \beta_1 X + \beta_2 T + \beta_3 XT$$

β_2 is the estimated average treatment effect (ATE)

X – learner

1. Estimate the **average outcomes** $\mu_0(x)$ and $\mu_1(x)$ using machine learning models:

$$\mu_0(x) = \mathbb{E}(Y(0)|X = x)$$

$$\mu_1(x) = \mathbb{E}(Y(1)|X = x)$$

2. Impute the **treatment effects** based on the **observed** and **estimated outcome**:

Control group $\quad D_i^0 \coloneqq \hat{\mu}_1\left(X_i^1\right) - Y_i^0 \quad \rightarrow \quad \tau_0(x) = E[D^0|X = x]$

Treatment group $\quad D_i^1 \coloneqq Y_i^1 - \hat{\mu}_0\left(X_i^0\right) \quad \rightarrow \quad \tau_1(x) = E[D^1|X = x]$

then estimate $\tau_0(x) = E[D^0|X = x]$ and $\tau_1(x) = E[D^1|X = x]$ with machine learning models

3. Obtain CATE estimate as weighted average of both estimates:

$$\hat{\tau}(x) = g(x)\,\hat{\tau}_0(x) + \big(1 - g(x)\big)\hat{\tau}_1(x)$$

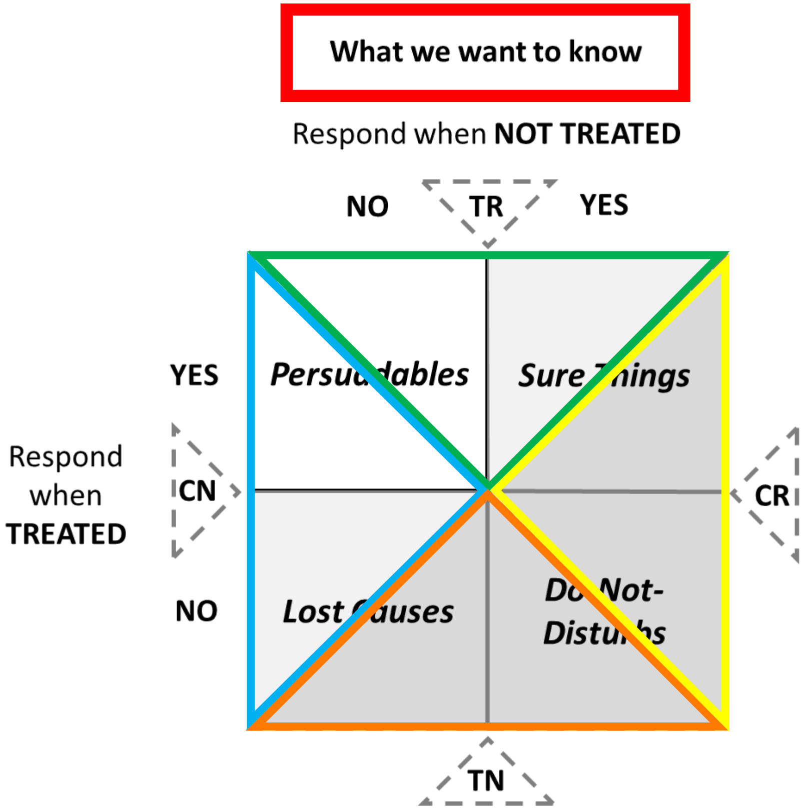
- with $g(x) \in [0,1]$, e.g., a ***propensity score*** to reduce treatment assignment bias

→ Appropriate when C & T samples are imbalanced

Based on: https://causalml.readthedocs.io/en/latest/methodology.html

Transformed outcome method

Depending on the **two potential outcomes**



Depending on the **applied treatment** and the **observed outcome**

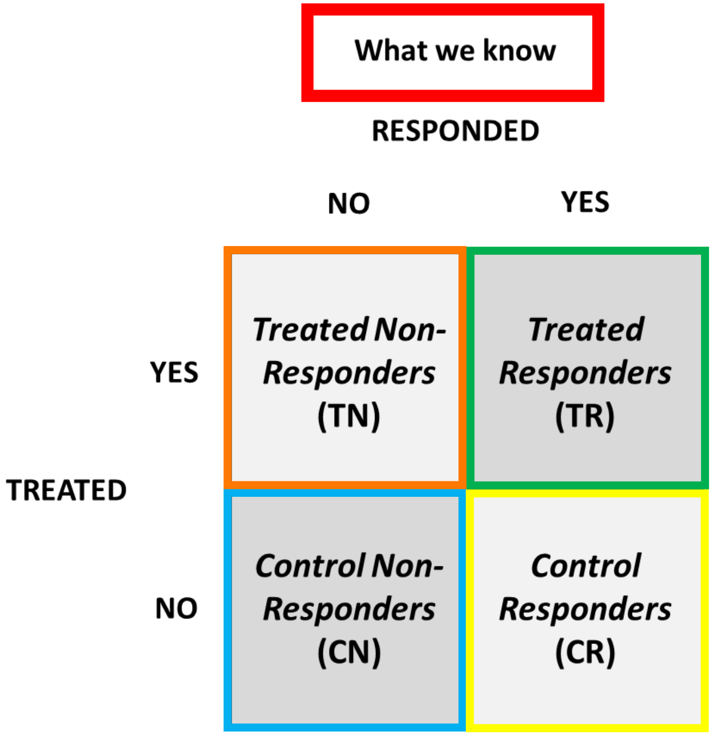


Figure adopted from Kane, 2014

Transformed outcome method

x_1	x_2	...	t	y	
32	1,530€	...	1	1	TR
48	2,680€	...	1	0	TN
...	...	...	...	...	
23	1,720€	...	0	0	CN
39	2,390€	...	0	1	CR
...	...	...	...	...	

transformed outcome

y'
1
0
...
1
0
...

- Estimate the **transformed outcome** $\boldsymbol{y'}$ using machine learning models

References

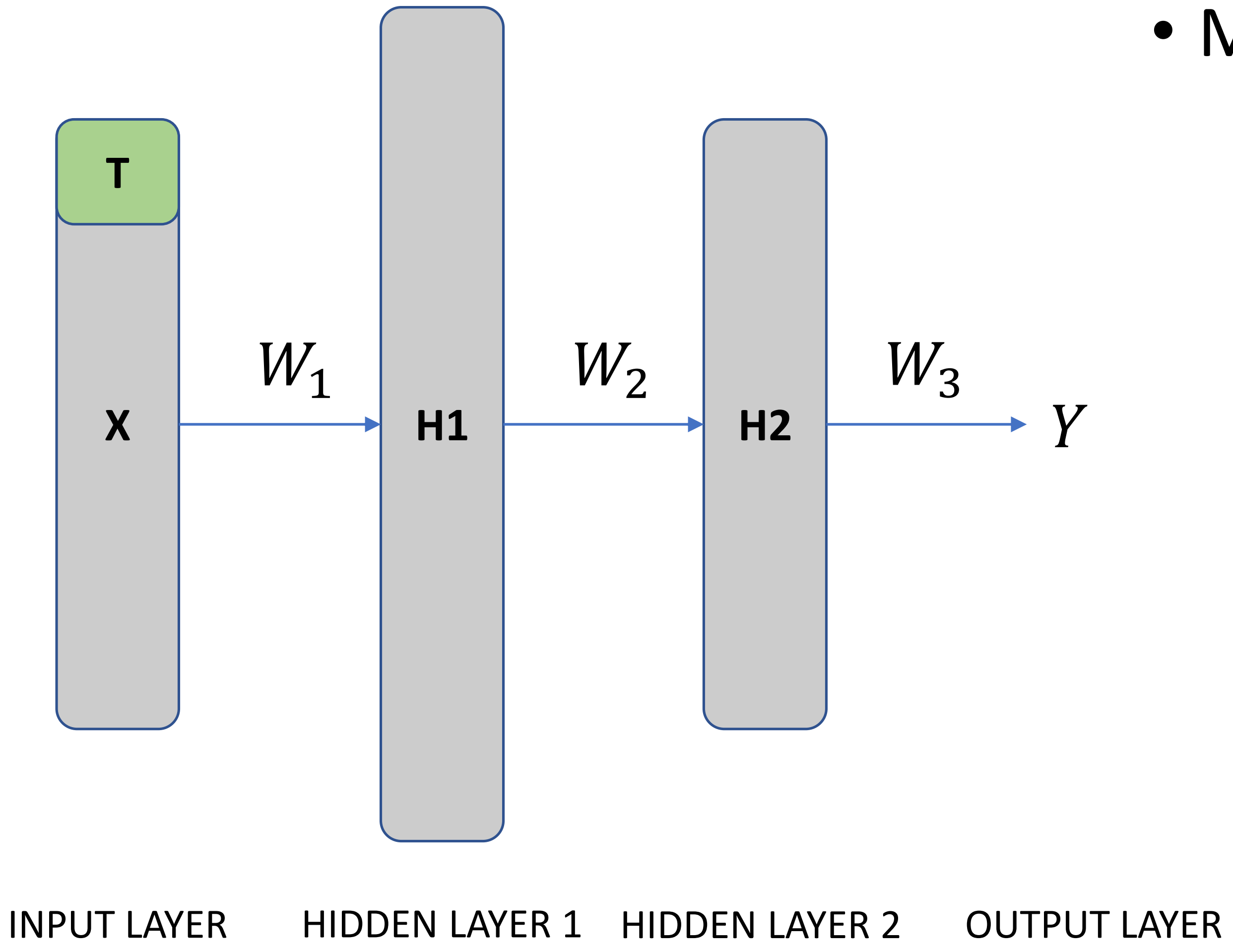
- Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, *116*(10), 4156-4165.
- Curth, A., & van der Schaar, M. (2021, March). Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *International Conference on Artificial Intelligence and Statistics* (pp. 1810-1818). PMLR.
- Causal ML package documentation: https://causalml.readthedocs.io/

Deep learning for uplift modeling

ECML/PKDD’22 Uplift Modeling Tutorial

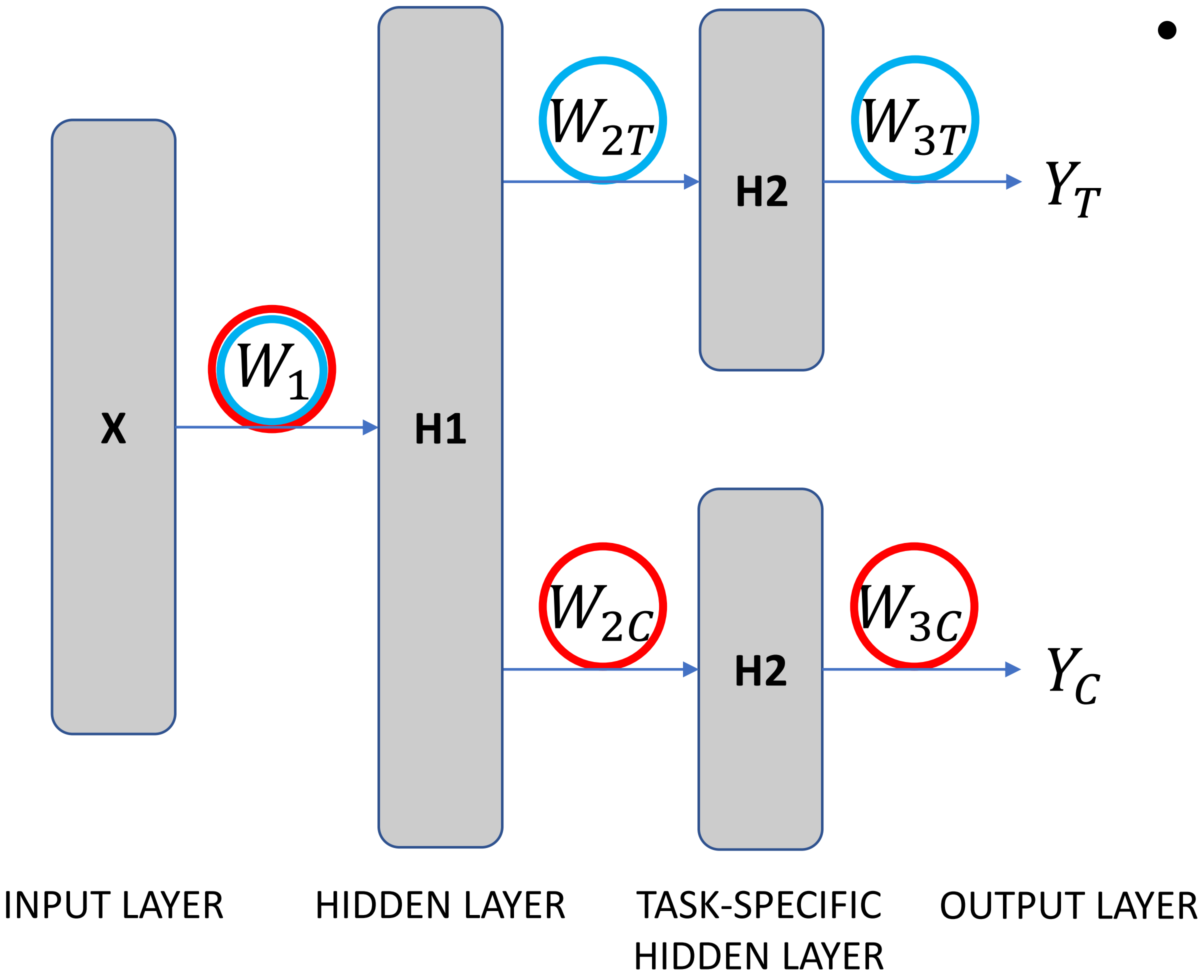
Wouter Verbeke & Szymon Jaroszewicz

Deep learning



- MLP as base learner in metalearner
  - E.g., S-Learner (treatment dummy)

Deep learning



- MLP-specific approach: **Y-net**
  - Multi-task learning
  - Hybrid **two-model architecture**
  - For binary treatment

  - Observations of the treatment group for learning (partial updates) W_1 and W_{2T} and W_{3T}
  - Observations of the control group for learning (partial updates) W_1 and W_{2C} and W_{3C}

Balancing

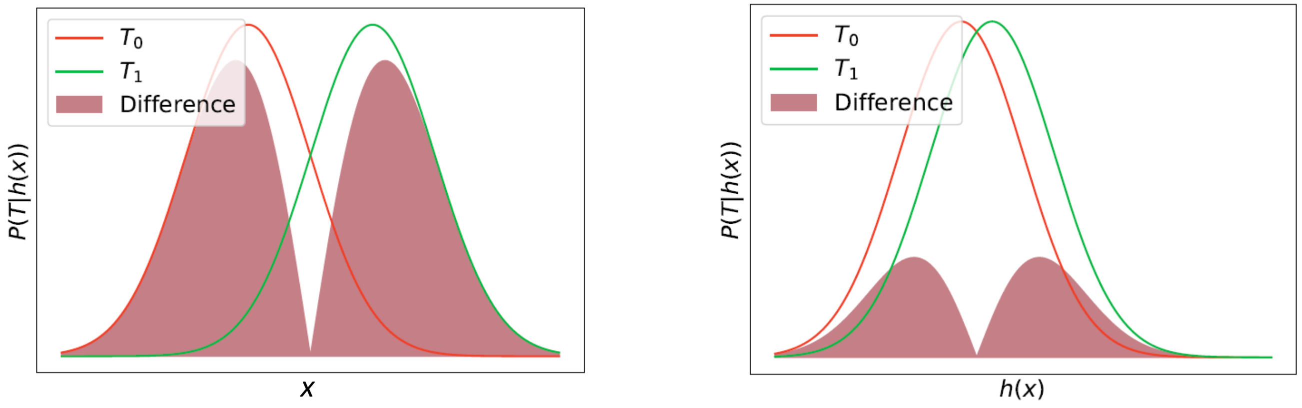
How to learn from ***observational data****?*

➔ Treatment assignment bias

➔ Learn ***similar*** **representations** of **treatment** and **control** **group**

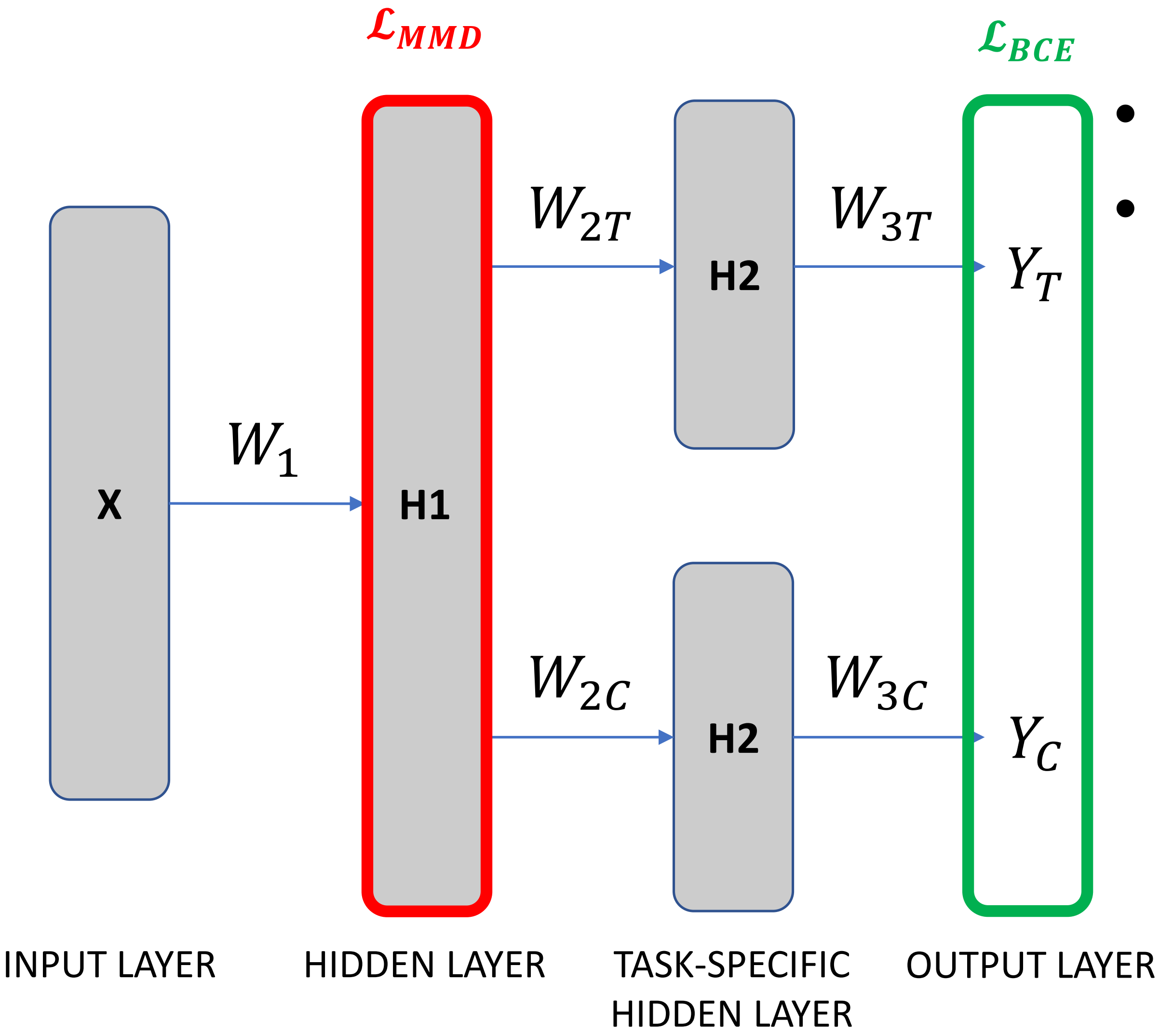
- Minimize distributional distance between both groups
- Integral probability metrics as regularization
  - E.g., Wasserstein distance, maximum mean discrepancy, …

x: *original data*
h(x): *representation*



(a) Large MMD (b) Small MMD

Deep learning with balancing



- Learn a **bias-free representation of X**
- **How?**
  - Measure amount of bias in Hidden layer H1, e.g., using **MMD**
  - Extended loss function, e.g.: **Binary cross-entropy loss + MMD**

$$Loss = \mathcal{L}_{BCE} + \alpha \mathcal{L}_{MMD}$$

α : **hyperparameter**

$\mathcal{L}_{BCE}$: **Binary cross-entropy loss**

$\mathcal{L}_{MMD}$: **Maximum mean discrepancy**

$$= \sum_{i=1}^{d} \left| \frac{1}{N_T} \sum_{i_T=1}^{N_T} h(x_{i_T}^i (t=1)) - \frac{1}{N_C} \sum_{i=1}^{N_C} h(x_{i_C}^i (t=0)) \right|$$

References

- Johansson, F., Shalit, U., & Sontag, D. (2016, June). Learning representations for counterfactual inference. In *International conference on machine learning* (pp. 3020-3029). PMLR.
- Shalit, U., Johansson, F. D., & Sontag, D. (2017, July). Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning* (pp. 3076-3085). PMLR.

Learning to rank for uplift modeling

ECML/PKDD'22 Uplift Modeling Tutorial

Wouter Verbeke & Szymon Jaroszewicz

Learning to rank for uplift modeling

- Learning to rank (L2R) techniques:
  - Stem from the information retrieval community,
  - Comprise techniques specifically designed to optimize the quality of predicted rankings directly,
  - Rather than the quality of predicted values that serve to rank instances

- Aim in uplift modeling: ranking!
- L2R for UM:
  - Requires **appropriate metric** for evaluating quality of ranking (objective)
  - Cfr. Supra: evaluation measures (e.g., CROC measure)

Devriendt et al., 2020, Learning to Rank for uplift modeling, IEEE Transactions on Knowledge and Data Engineering, https://doi.org/10.1109/TKDE.2020.3048510

Evaluating uplift models

ECML/PKDD'22 Uplift Modeling Tutorial

Wouter Verbeke & Szymon Jaroszewicz

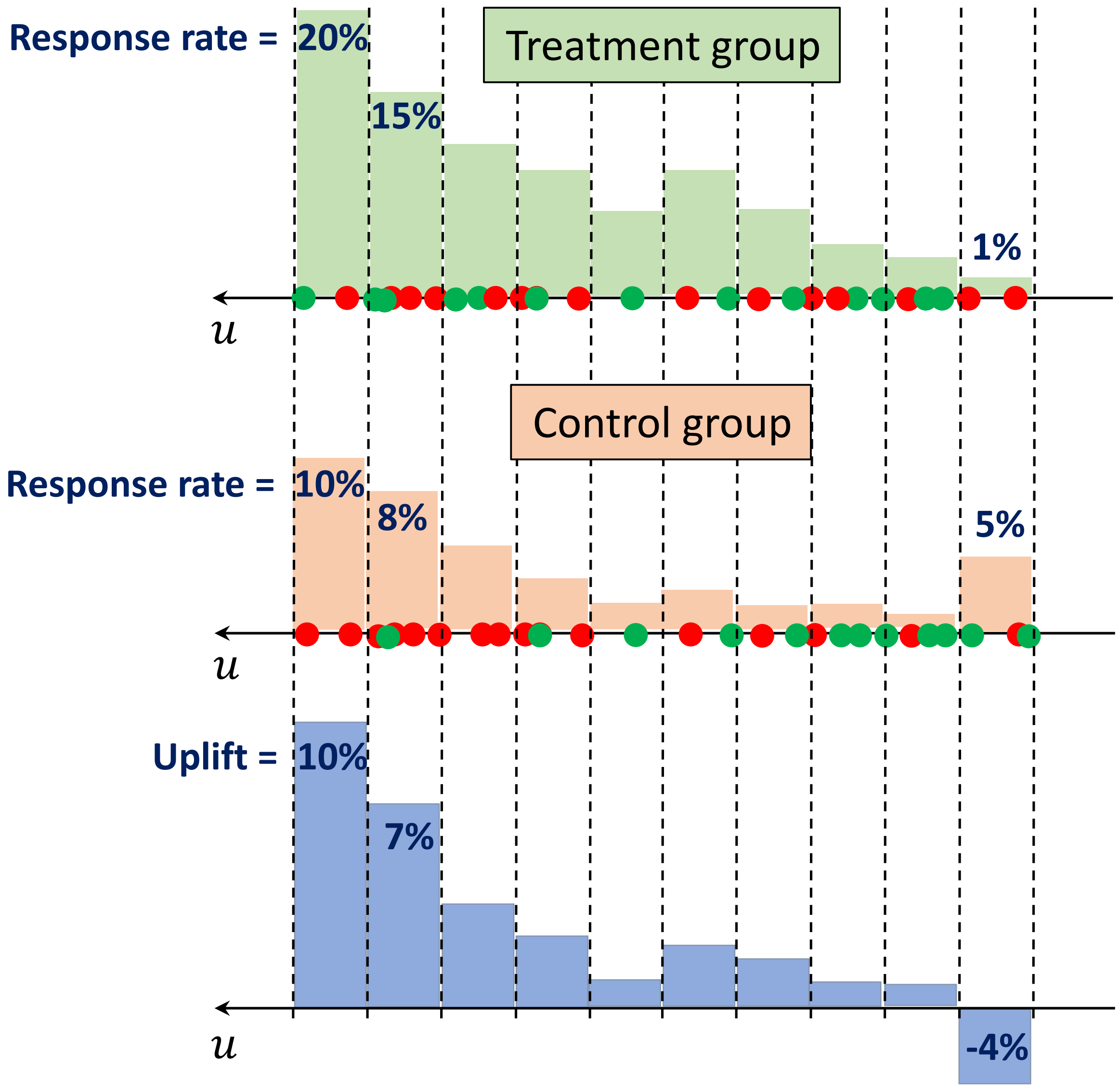
Evaluation

- **PEHE** = RMSE (root mean squared error)
- Synthetic or semi-synthetic data
  - Research <> Business decision-making (e.g., marketing)

$$\epsilon_{PEHE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\underbrace{[Y_1^{(i)} - Y_0^{(i)}]}_{\text{ITE}} - \underbrace{[\hat{Y}_1^{(i)} - \hat{Y}_0^{(i)}]}_{\text{ITE estimate}})^2}$$

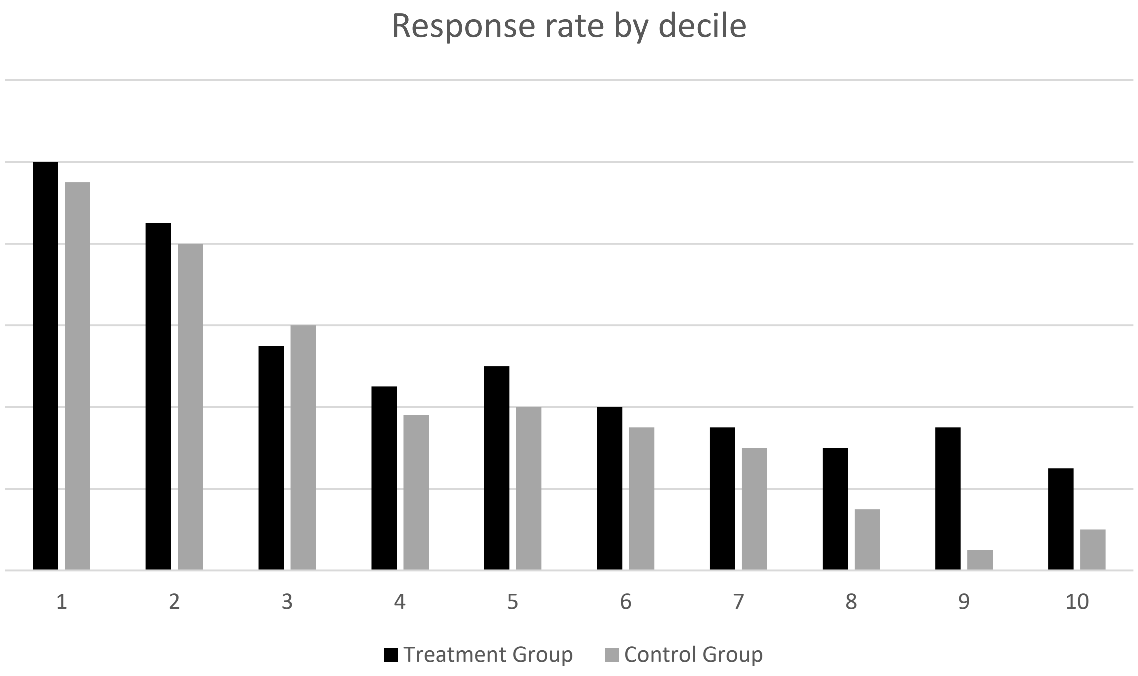
Evaluation

- Uplift curve
  - For binary target
  - Evaluation by comparing outcomes for ***similar*** groups
  - Uplift model allows to score and rank all instances
  - **Uplift-curve:** increase in positive outcome rate
    - E.g., per decile
- Note: observational data …



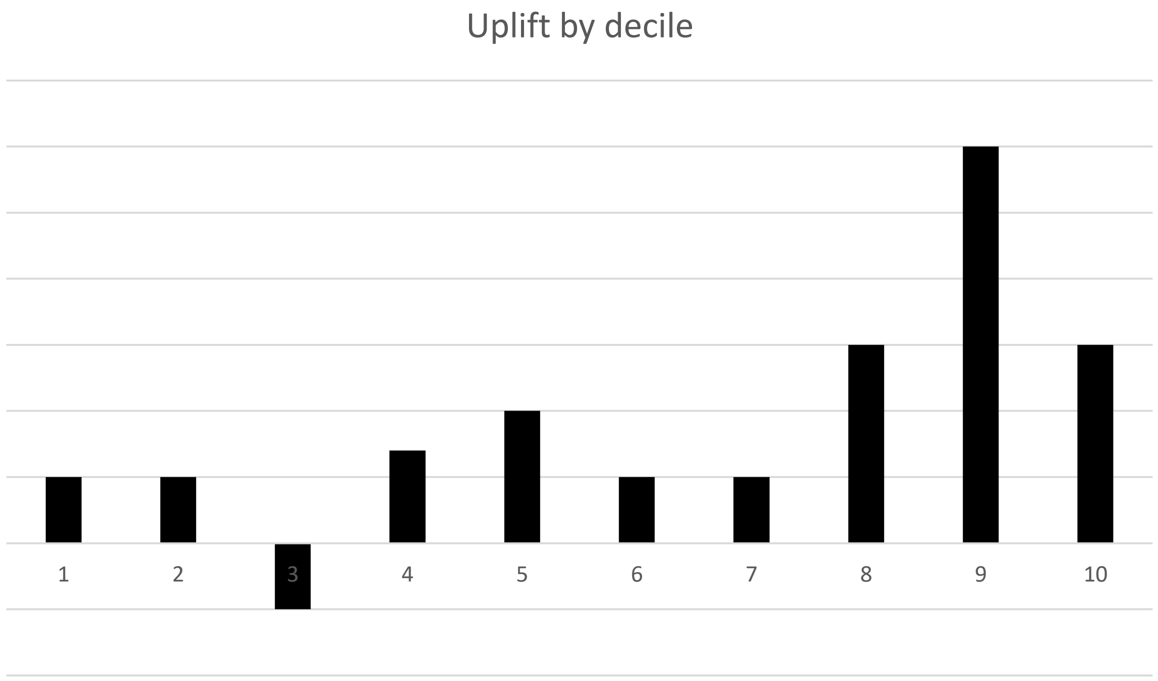
Evaluation

- Response rate by decile



Evaluation

- Uplift by decile



Good model?

Evaluation

First, we should rank persuadables (ITE = 1)
Then, we should rank lost causes and sure things (ITE = 0)
Finally, we should have the sleeping dogs (ITE = -1)

Cfr. infra: transformed outcome method

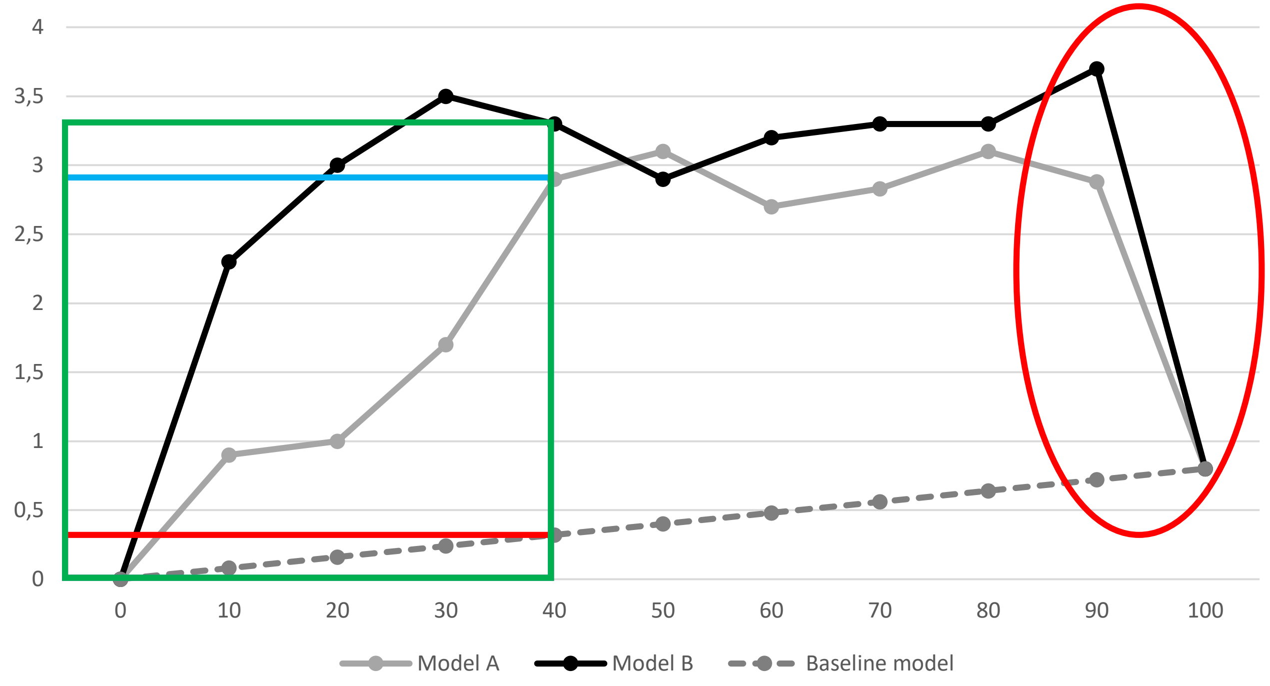
- Uplift by decile **… for a *perfect* model?**

Evaluation

Link with ATE?

Cumulative incremental gains or Qini curve (cfr. Gini curve)

Y: **Increase in response rate (%)**

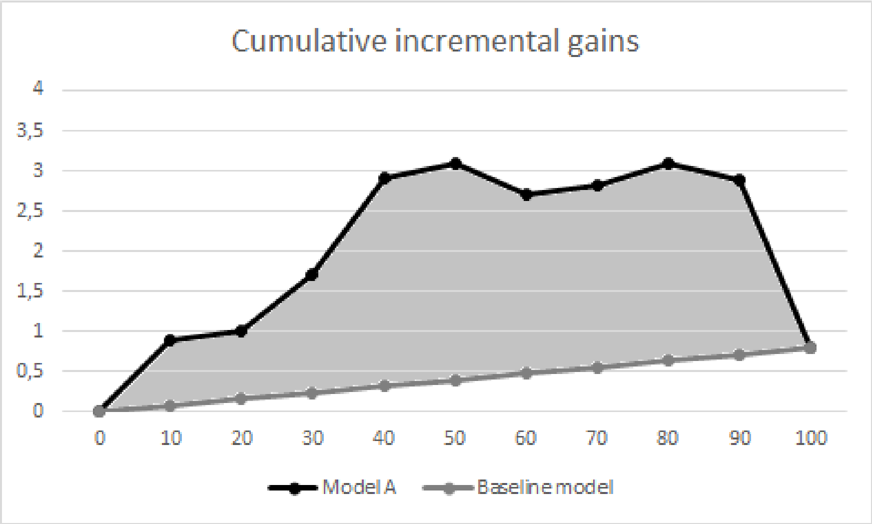


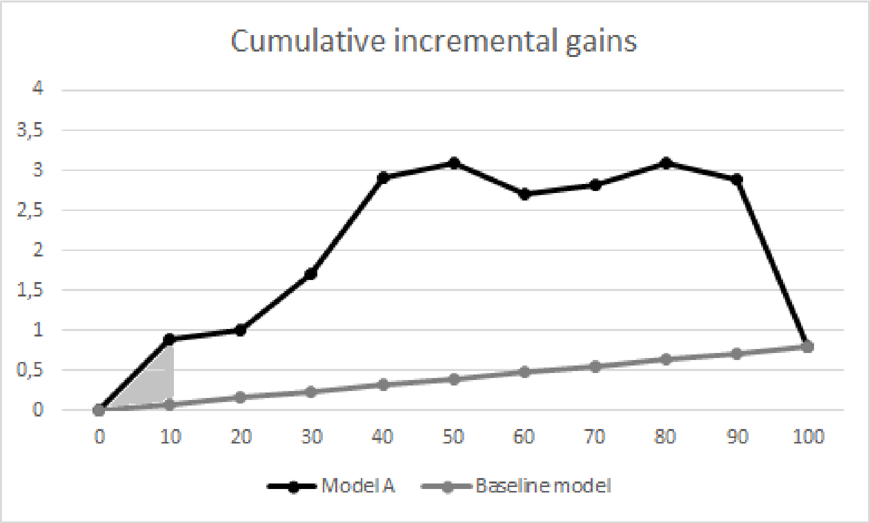
X: **Treatment rate (%)** of test sample ranked with model from large to small estimated uplift

Baseline (random) model:	40% treated	→	0.7% increase
Model A:	40% treated	→	2.9% increase
Model B:	40% treated	→	3.3% increase

Evaluation

- **Qini measure** = Area Under the Uplift Curve (AUUC <> AUC)
- **Quantile uplift**: how much uplift achieved at specified targeting depth?
  - Similar: top-decile qini





Evaluation: CROC

- Cfr. infra: transformed outcome method
  - Apply to transform evaluation in binary classification evaluation

x_1	x_2	…	t	y	
32	1,530€	…	1	1	TR
48	2,680€	…	1	0	TN
…	…	…	…	…	
23	1,720€	…	0	0	CN
39	2,390€	…	0	1	CR
…	…	…	…	…	

transformed outcome

y'
1
0
…
1
0
…

- Then, apply, e.g., ROC analysis:

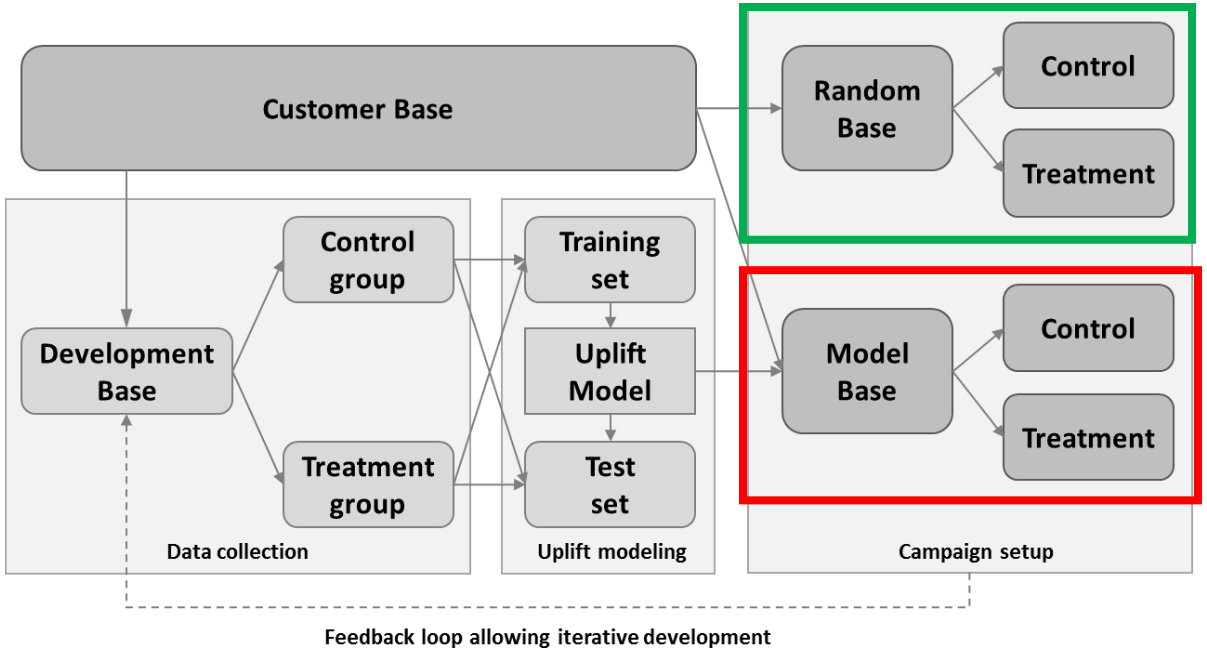
  → Causal ROC curve (CROC curve)

  → Area under the CROC curve (AUCROC measure)

Verbeke et al., 2022, To do or not to do: Cost-sensitive causal decision-making, European Journal of Operational Research

Evaluation

- Monitoring model performance
- Iterative ***learning*** and improving or optimizing



RCT

Exploration
vs.
Exploitation

Active learning
Bandits, Reinforcement learning

Evaluation

- Uplift modeling: ranking per CATE to optimize targeting

- How many to target?
  - I.e., where to set the threshold?

- Bringing in costs and benefits to optimize decision-making!
  - Cost of a treatment
    - May depend on the outcome
    - E.g., discount in case of a positive outcome only
  - Benefit of **causing** a positive outcome
  - Cost of **causing** a negative outcome

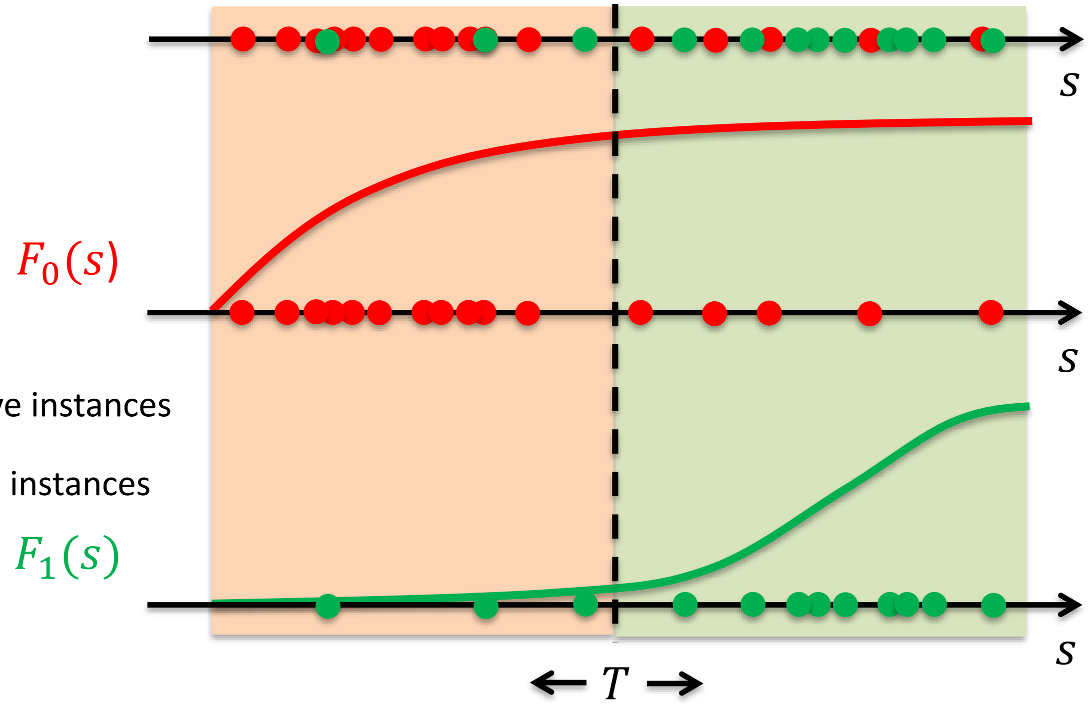
		Buy after campaign	
		No	Yes
Buy without campaign	No	Lost causes	Persuadables
	Yes	Sleeping dogs	Sure things

Classification model allows to score and rank all instances

N: number of instances

π_0: proportion of Negative instances

π_1: proportion of Positive instances



Confusion matrix	Predicted Negative	Predicted Positive
Actual Negative	$F_0(\boldsymbol{T})\pi_0 N$	$(1 - F_0(\boldsymbol{T}))\pi_0 N$
Actual Positive	$F_1(\boldsymbol{T})\pi_1 N$	$(1 - F_1(\boldsymbol{T}))\pi_1 N$

Arbitrary threshold!

Cost-benefit matrix	Predicted Negative	Predicted Positive
Actual Negative	b_0	c_0
Actual Positive	c_1	b_1

Maximum Profit measure

Confusion matrix	Predicted Negative	Predicted Positive
Actual Negative	$F_0(\boldsymbol{T})\pi_0 N$	$(1 - F_0(\boldsymbol{T}))\pi_0 N$
Actual Positive	$F_1(\boldsymbol{T})\pi_1 N$	$(1 - F_1(\boldsymbol{T}))\pi_1 N$

$\circ$

Cost-benefit matrix	Predicted Negative	Predicted Positive
Actual Negative	b_0	c_0
Actual Positive	c_1	b_1

(with $\circ$ the Hadamard product)

- **Average Profit** (P) per instance:

$$P(\boldsymbol{T}; b_0, c_0, b_1, c_1) = b_0\, F_0(\boldsymbol{T})\pi_0 + b_1\, F_1(\boldsymbol{T})\pi_1 - c_0(1 - F_0(\boldsymbol{T}))\pi_0 - c_1(1 - F_1(\boldsymbol{T}))\pi_1$$

$$P = P(T) = \sum\sum(C \circ CB)$$

- **Maximum Profit** (MP) measure:

$$MP = \max_{\forall T} P(T; b_0, c_0, b_1, c_1) = P(\boldsymbol{T}^*; b_0, c_0, b_1, c_1)$$

  - with $\boldsymbol{T}^*$ **the optimal threshold** under the given cost-benefit distribution:

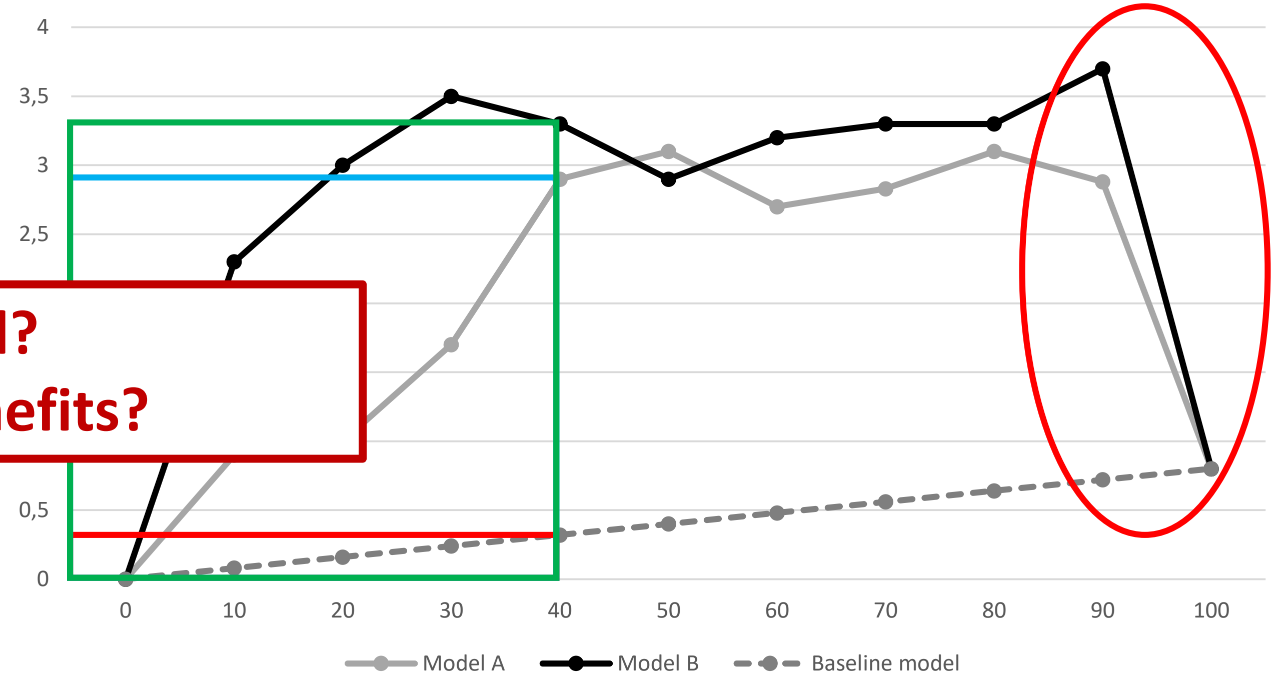
$$T^* = \arg\max_{\forall T} P(T; b_0, c_0, b_1, c_1)$$

Evaluation

Link with ATE?

Cumulative incremental gains or Qini curve (cfr. Gini curve)

Y: **Increase in response rate (%)**



Threshold?
Costs and benefits?

X: **Treatment rate (%)** of test sample ranked with model from large to small estimated uplift

Baseline (random) model:	40% treated	→	0.7% increase
Model A:	40% treated	→	2.9% increase
Model B:	40% treated	→	3.3% increase

Uplift model allows to score and rank all instances

Control group

Applied treatment: W=0

$F_0^C(u)$

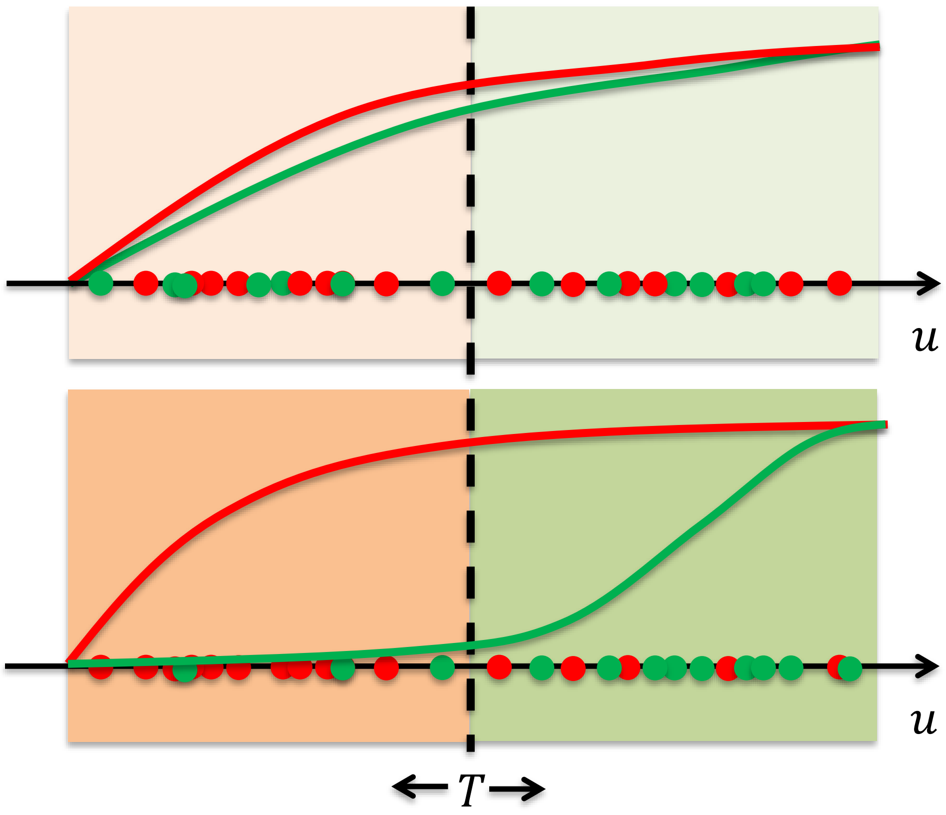
$F_1^C(u)$

Treatment group

Applied treament: W = 1

$F_0^T(u)$

$F_1^T(u)$



Treatment - Outcome matrix		**W = 0**	**W = 1**
Control group	**Outcome Negative**	$F_0^c(T)\pi_0$	$(1 - F_0^c(T))\pi_0$
	Outcome Positive	$F_1^c(T)\pi_1$	$(1 - F_1^c(T))\pi_1$
Treatment group	**Outcome Negative**	$F_0^T(T)\pi_0$	$(1 - F_0^T(T))\pi_0$
	Outcome Positive	$F_1^T(T)\pi_1$	$(1 - F_1^T(T))\pi_1$

Treatment - outcome matrix (TO):
Simulated outcome distributions for ***some threshold*** τ

Treatment - outcome matrix	W = 0	W = 1
Outcome Y = 0	$\pi_0^c F_0^c(\tau)$	$\pi_0^T(1 - F_0^T(\tau))$
Outcome Y = 1	$\pi_1^c F_1^c(\tau)$	$\pi_1^T(1 - F_1^T(\tau))$

Observed in control group

Observed in treatment group

Net - effect matrix (NE):
Change in outcome distributions compared to ***baseline treatment W=0***

Net - effect matrix	W = 0	W = 1
Outcome Y = 0	0	$\pi_0^T\left(1 - F_0^T(\tau)\right) - \pi_0^c(1 - F_0^C(\tau))$
Outcome Y = 1	0	$\pi_1^T\left(1 - F_1^T(\tau)\right) - \pi_1^c(1 - F_1^C(\tau))$

Cost (C) and Benefit (B) matrices:
Costs & benefits depend on treatments & outcomes

Cost matrix	W = 0	W = 1
Outcome Y = 0	$c_{(0,0)}$	$c_{(1,0)}$
Outcome Y = 1	$c_{(1,0)}$	$c_{(1,1)}$

Benefit matrix	W = 0	W = 1
Outcome Y = 0	$b_{(0,0)}$	$b_{(1,0)}$
Outcome Y = 1	$b_{(1,0)}$	$b_{(1,1)}$

Average Profit (P) per instance:

$$P(T) = \sum\sum(NE \circ B - TO \circ C)$$

(with $\circ$ the Hadamard product)

Maximum Profit Uplift (MPU) measure:

$$MPU = \max_{\forall T} P(T)$$

- with T^* **the optimal threshold** under the given cost-benefit distribution: $T^* = \arg\max_{\forall T} P(T; b_0, c_0, b_1, c_1)$

References

- Devriendt et al., 2020, Learning 2 Rank for uplift modeling, IEEE Transactions on Knowledge and Data Engineering, https://doi.org/10.1109/TKDE.2020.3048510
- Verbeke et al., 2021, The foundations of cost-sensitive causal classification, ArXiv
- Verbeke et al., 2022, To do or not to do: Cost-sensitive causal decision-making, European Journal of Operational Research, https://doi.org/10.1016/j.ejor.2022.03.049

Research agenda

ECML/PKDD'22 Uplift Modeling Tutorial

Wouter Verbeke & Szymon Jaroszewicz

Beyond double binary causal classification

- Continuous treatments?
  - Discount, price, production parameters, …
  - Treatment dummy approach

- Multiple treatments?
  - Also for continuous treatments with binning
    - E.g.: discount, price, …
  - T-Learner or multi-model approach: one model per treatment vs. control
  - S-Learner or treatment dummy approach: multiple treatment dummies

Beyond double binary causal classification

- Continuous outcome?
  - Revenue, yield, quality, time-until-churn/failure/…
  - Two-model approach
  - Treatment dummy approach

With observational data?

Inverse propensity score weighting?

- Multiclass outcome?
  - Also for continuous outcomes with binning
  - Multi-model approach
  - Treatment dummies approach
  - Multi-task learning approach

Beyond double binary causal classification

- High-dimensional treatments:
  - E.g., organITE
- Interpretability or explainability?
- Taking into account cost of treatment and benefit of outcome
  - Objective: maximize profits
  - E.g., customer retention:
    - To retain as much customers as possible?
    - To retain as much value as possible!

Cost-sensitive learning
Profit-driven analytics

Beyond double binary causal classification

- Time-dependent treatments and outcomes?
  - Survival analysis: personalized medicine
  - Forecasting: demand steering

- Concept drift?
  - How much data needed for (re-)training?
    - ‘Representative’ sample?
    - Still use for ‘old’ data?
    - E.g., change in retention offer, market conditions, …

Bandits
Reinforcement learning

Cases

ECML/PKDD’22 Uplift Modeling Tutorial

Wouter Verbeke & Szymon Jaroszewicz

Beyond: Cases

Case: Machine maintenance

- Predictive maintenance vs. prescriptive maintenance
- Take into account costs and benefits?
- Note: close link with optimization

Case: Waste oven process

- Process instances are variable

Questions:

- **Double binary causal classification?**
- **RCT data?**

Beyond: Cases

Case: Pricing – **ITE model for customer price elasticity?**

- Pricing **grid** – segmentation based on:
  - Demand characteristics (when, #, …)
  - Customer characteristics?
- Note:
  - Infeasible to price at the individual level?
  - ITE estimates still allow to optimize segmentation
  - Fences
  - Close link with optimization

Questions:

- **Double binary causal classification?**
- **RCT data?**
- **Ethical concerns?**

Beyond: Cases

Case: Credit risk management

- Active credit risk management: measures to prevent default?
  - E.g.: Practice of active credit risk management in economic downturn periods
- Minimizing losses due to default: recovery process optimization

Questions:

- **Double binary causal classification?**
- **RCT data?**

Beyond: Cases

Case: Human resources management

- Effect of benefits, policies, … with respect to turnover, illness, …
- E.g.: Compensation and benefits
  - Modeling compensation and benefit impact on employee retention, satisfaction and performance?
  - Note: ‘slow’ vs. immediate effects (e.g., in marketing)
    - See also health: effect of exposure to …

Questions:

- **Double binary causal classification?**
- **RCT data?**

Beyond: Cases

Case: Learning analytics
https://www.sciencedirect.com/science/article/pii/S0167923620300750

Case: Health: personalized medicine – Van der Schaar lab @Cambridge
https://www.youtube.com/watch?v=YQ8HX4T5OuE

Case: Fraud risk management?

- Preventive fraud measures?