Heterogeneous Treatment Effects and Targeting Decisions

Data Science for Marketing Decision Making
Günter J. Hitsch
Chicago Booth

Winter 2017

Overview

1. The classical CRM approach and predictive modeling
2. CATE (conditional average treatment effect): CRM based on heterogeneous incremental effects
3. Linear treatment-interactions model
4. Causal forest
5. Model validation with heterogeneous treatment effects
6. Profit prediction based on randomized samples

CRM and predictive modeling: The classical approach

Goal: Focus targeting efforts on customers to improve profitability and ensure positive ROI's.

In *churn managemen*t we predict

$$\Pr(\text{churn}_i|\boldsymbol{x}_i)$$

Strategy: Do not target customers with low churn rates. E.g., if $\Pr(\text{churn}_i|\boldsymbol{x}_i) = 0.002$, the churn probability can be reduced by at most 0.2 percentage points, and the cost to achieve such a small reduction may be larger than the incremental customer value.

In *customer development*, we predict the expected profit when targeting the customer:

$$\mathbb{E}(\text{profit}_i|\boldsymbol{x}_i) = m_i\mathbb{E}(Y_i|\boldsymbol{x}_i) - c_i$$

(Y_i is spending, m_i is the margin and c_i is the cost of targeting)

Strategy: Target customers only if $\mathbb{E}(\text{profit}_i|\boldsymbol{x}_i)$ is sufficiently high.

The classical CRM approach and targeting decisions

Targeting decision:

$$\mathbb{E}(\text{profit}_i|\boldsymbol{x}_i) > 0 \;\Leftrightarrow\; m_i\mathbb{E}(Y_i|\boldsymbol{x}_i) > c_i$$

Key question: What is the alternative to targeting the customer?

The targeting decision above implicitly assumes a profit of 0 when not targeting a customer.

Is this assumption realistic?

Alternative: Assume the profit when not targeting a customer is a fraction $\alpha < 1$ of the targeting profit. Then the modified targeting rule says that we should target a customer if and only if:

$$\begin{aligned} & m_i\mathbb{E}(Y_i|\boldsymbol{x}_i) - c_i > \alpha\left(m_i\mathbb{E}(Y_i|\boldsymbol{x}_i)\right) \\ \Leftrightarrow\; & (1-\alpha)m_i\mathbb{E}(Y_i|\boldsymbol{x}_i) > c_i \end{aligned}$$

Introduce some notation: W_i is the *targeting indicator*.

$W_i = 0$ if we do not target the customer, $W_i = 1$ if we target the customer.

Now express expected profits conditional on the targeting status $w = 0, 1$:

$$\mathbb{E}(\text{profit}_i | \boldsymbol{x}_i, W_i = w)$$

The targeting logic on the previous slide can then be expressed as follows: Target if and only if

$$\begin{aligned} & \mathbb{E}(\text{profit}_i | \boldsymbol{x}_i, W_i = 1) > \mathbb{E}(\text{profit}_i | \boldsymbol{x}_i, W_i = 0) \\ \Leftrightarrow \; & m_i \mathbb{E}(Y_i | \boldsymbol{x}_i, W_i = 1) - c_i > m_i \mathbb{E}(Y_i | \boldsymbol{x}_i, W_i = 0) \\ \Leftrightarrow \; & m_i \left(\mathbb{E}(Y_i | \boldsymbol{x}_i, W_i = 1) - \mathbb{E}(Y_i | \boldsymbol{x}_i, W_i = 0) \right) > c_i \end{aligned}$$

Potential outcomes model

To make a targeting decision, we need to predict

$$\mathbb{E}[Y_i | \boldsymbol{x}_i, W_i = 1] - \mathbb{E}[Y_i | \boldsymbol{x}_i, W_i = 0]$$

If we use the notation $Y_i(0)$ for the sales outcome when customer i is not targeted and $Y_i(1)$ for the sales outcome when the customer is targeted, we can write the expected sales differences as

$$\mathbb{E}[Y_i(1) | \boldsymbol{x}_i] - \mathbb{E}[Y_i(0) | \boldsymbol{x}_i] = \mathbb{E}[Y_i(1) - Y_i(0) | \boldsymbol{x}_i]$$

A correct targeting decision requires a prediction of $Y_i(1) - Y_i(0)$, the **causal** (**incremental**) effect of targeting customer i vs. not targeting customer i.

Conditional average treatment effect

We want to make customer-level predictions of the causal effect of a treatment, $Y_i(1) - Y_i(0)$.

$\boldsymbol{x}_i$ is a vector of variables (attributes) that we observe for customer i.

CATE (conditional average treatment effect):

$$\tau_i = \mathbb{E}[Y_i(1) - Y_i(0)|\boldsymbol{x}_i]$$

The CATE is closely related to the ATE (average treatment effect), but specific to the sub-population of units (customers) characterized by the attributes $\boldsymbol{x}_i$. τ_i is a **heterogeneous** (across units) **treatment effect**.

Examples:

- Incremental purchase when a customer is targeted vs. not targeted conditional on past transaction data (time elapsed since the last purchase, order frequency during the last two years, ...)
- Incremental click-through or conversion if a customer (cookie) is exposed to a display ad vs. not being exposed to a display ad, conditional on a customer profile obtained from past browsing data

Estimation of the conditional average treatment effect

Ideally we would use data, $\mathcal{D} = ((\tau_1, \boldsymbol{x}_1), \ldots, (\tau_n, \boldsymbol{x}_n))$, to estimate the CATE using regression analysis.

For example, if the relationship between $\boldsymbol{x}_i$ and τ_i is approximately linear, then

$$\tau_i = \delta_0 + \sum_{k=1}^{p} \delta_k x_{ik} + \epsilon_i$$

Is this approach feasible?

Estimation

Notation:

$$Y_i \equiv Y_i^{\text{obs}} = \begin{cases} Y_i(0) & \text{if } W_i = 0 \\ Y_i(1) & \text{if } W_i = 0 \end{cases}$$

The data set actually observed is $\mathcal{D} = ((Y_1, W_1, \boldsymbol{x}_1), \ldots, (Y_n, W_n, \boldsymbol{x}_n))$. Hence we observe $Y_i(0)$ if $W_i = 0$, and $Y_i(1)$ if $W_i = 1$.

The **fundamental problem in causal inference** is that we observe either $Y_i(0)$ or $Y_i(1)$. We never observe both parallel worlds and hence we never observe both $Y_i(0)$ and $Y_i(1)$.

Linear model with treatment-interactions

Proposed regression model:

$$Y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + \delta_0 W_i + \sum_{k=1}^{p} \delta_k (x_{ik} W_i) + \epsilon_i$$

- Interactions $x_{ik} \cdot W_i$

Allows us to estimate the conditional expectation function

$$\mathbb{E}[Y_i | \boldsymbol{x}_i, W_i] = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + \delta_0 W_i + \sum_{k=1}^{p} \delta_k (x_{ik} W_i)$$

Goal: Estimate the CATE,

$$\begin{aligned} \tau_i &= \mathbb{E}[Y_i(1) - Y_i(0) | \boldsymbol{x}_i] \\ &= \mathbb{E}[Y_i | \boldsymbol{x}_i, W_i = 1] - \mathbb{E}[Y_i | \boldsymbol{x}_i W_i = 0] \end{aligned}$$

We can predict the CATE from the regression function:

$$\begin{aligned}\tau_i &= \mathbb{E}[Y_i|\boldsymbol{x}_i, W_i = 1] - \mathbb{E}[Y_i|\boldsymbol{x}_i, W_i = 0] \\ &= \left(\beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + \delta_0 + \sum_{k=1}^{p} \delta_k x_{ik}\right) - \left(\beta_0 + \sum_{k=1}^{p} \beta_k x_{ik}\right) \\ &= \delta_0 + \sum_{k=1}^{p} \delta_k x_{ik}\end{aligned}$$

When will we obtain consistent estimates of the parameters δ_k and hence consistently estimate the CATE?

Consistent estimation of CATE

1. **Unconfoundedness**: The potential outcomes, $Y_i(0)$ and $Y_i(1)$ are statistically independent of the treatment W_i *conditional* on $\boldsymbol{x}_i$.
2. **Overlap**: The probability of receiving the treatment is neither 0 nor 1 for all $\boldsymbol{x}_i$:
$$0 < \Pr\{W_i = 1|\boldsymbol{x}_i\} < 1$$
3. **SUTVA** (**stable unit treatment value assumption**): The treatment assignment for one unit i does not affect the outcome for a different unit, $k \neq i$

Note that violation of SUTVA is very unlikely in a CRM situation: Equilibrium effects are of no concern as CRM efforts are small interventions. Word-of-mouth effects *might* arise, although we may want to incorporate WOM effects in the estimated treatment effect.

Violation of overlap

Terminology: The conditional probability of receiving the treatment $W_i = 1$,

$$e(\boldsymbol{x}_i) = \Pr\{W_i = 1 | \boldsymbol{x}_i\},$$

is called the **propensity score**.

Situation where overlap is violated in a target marketing situation?

Unconfoundedness

Unconfoundedness can always be assured if W_i is randomized in a RCT.

Indeed, all that is required is that W_i is randomized in each sub-population characterized by identical attribute vectors $\boldsymbol{x}_i$.

Through randomization we can also ensure that overlap is satisfied.

Warning: If W_i is randomized in each sub-population, but not in the overall population, we need to be careful when characterizing the overall, average treatment effect. See the example on the next slides.

Correctly using randomized data

Example: Two customer segments, $\mathcal{A}$ and $\mathcal{B}$

Segment	Population percentage	$\Pr\{W_i = 1\}$	$Y_i(0)$	$Y_i(1)$	τ_i
$\mathcal{A}$	0.5	0.05	20	40	20
$\mathcal{B}$	0.5	0.9	100	100	0

Within each of the two segments the treatment assignment is random.

However, note that in the whole population (segments $\mathcal{A}$ and $\mathcal{B}$ together) the treatment W_i is not independent of the two potential outcomes.

Implications:

1. We can consistently estimate the CATE, τ_i, in each segment
2. We cannot estimate the ATE (average treatment effect in the population) based on the pooled data using all segments
3. We can estimate the ATE by taking a population-weighted average of the CATE in each segment

First, what is the probability that a customer is in segment $\mathcal{A}$ or $\mathcal{B}$ conditional on the treatment status, $W_i = 0, 1$? We can figure this out using Bayes' law:

$$\begin{aligned}\Pr\{i \in \mathcal{A}|W_i = 1\} &= \frac{\Pr\{W_i = 1|i \in \mathcal{A}\} \cdot \Pr\{i \in \mathcal{A}\}}{\Pr\{W_i = 1|i \in \mathcal{A}\} \cdot \Pr\{i \in \mathcal{A}\} + \Pr\{W_i = 1|i \in \mathcal{B}\} \cdot \Pr\{i \in \mathcal{B}\}} \\ &= \frac{0.05 \cdot 0.5}{0.05 \cdot 0.5 + 0.9 \cdot 0.5} \\ &= 0.0526\end{aligned}$$

$$\begin{aligned}\Pr\{i \in \mathcal{B}|W_i = 1\} &= \frac{\Pr\{W_i = 1|i \in \mathcal{B}\} \cdot \Pr\{i \in \mathcal{B}\}}{\Pr\{W_i = 1|i \in \mathcal{A}\} \cdot \Pr\{i \in \mathcal{A}\} + \Pr\{W_i = 1|i \in \mathcal{B}\} \cdot \Pr\{i \in \mathcal{B}\}} \\ &= \frac{0.9 \cdot 0.5}{0.05 \cdot 0.5 + 0.9 \cdot 0.5} \\ &= 0.9474\end{aligned}$$

Second, we can then predict the expected outcome conditional on the treatment status, pooled over both segments

$$\begin{aligned}\mathbb{E}[Y_i|W_i=0] &= \mathbb{E}[Y_i(0)|i\in\mathcal{A}]\cdot\Pr\{i\in\mathcal{A}|W_i=0\}+\mathbb{E}[Y_i(0)|i\in\mathcal{B}]\cdot\Pr\{i\in\mathcal{B}|W_i=0\}\\ &= 20\cdot 0.9048+100\cdot 0.0952\\ &= 27.62\end{aligned}$$

$$\begin{aligned}\mathbb{E}[Y_i|W_i=1] &= \mathbb{E}[Y_i(1)|i\in\mathcal{A}]\cdot\Pr\{i\in\mathcal{A}|W_i=1\}+\mathbb{E}[Y_i(1)|i\in\mathcal{B}]\cdot\Pr\{i\in\mathcal{B}|W_i=1\}\\ &= 40\cdot 0.0526+100\cdot 0.9474\\ &= 96.84\end{aligned}$$

The inferred population ATE (average treatment effect) using the pooled data is:

$$\mathbb{E}[Y_i|W_i=1]-\mathbb{E}[Y_i|W_i=1]=96.84-27.62=69.22$$

The true ATE is:

$$\begin{aligned}\tau &= \Pr\{i\in\mathcal{A}\}\cdot\tau_{\mathcal{A}}+\Pr\{i\in\mathcal{B}\}\cdot\tau_{\mathcal{B}}\\ &= 0.5\cdot 20+0.5\cdot 0\\ &= 10\end{aligned}$$

Note: If the treatment probability (propensity score) is constant across all segments, then we can estimate the true ATE using the pooled data. — Confirm this!

Causal forest

Recently introduced method in the machine learning/treatment effects literature

- Wager and Athey (2016): "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests" (manuscript)

The causal forest algorithm is an extension of a random forest algorithm to **directly predict heterogeneous treatment effects**

Potential benefits over regression approach with treatment-interactions:

- Non-parametric method, designed to automatically detect non-linear relationships between the inputs and the treatment effect and interactions between the inputs
- The causal forest is directly trained to predict the CATE, τ_i, not the outcome level, Y_i

Causal forest algorithm

Like a random forest, a causal forest is based on an ensemble of trees.

Ideally, the leaves of the trees include (almost) homogenous units i, and the treatment assignment in each leaf is random or as good as random, such that the unconfoundedness assumption is satisfied.

In a standard regression tree, the predicted output for observation i in leaf l is given by

$$\hat{Y}_i = \hat{Y}_{\mathcal{R}_l} = \frac{1}{N_l} \sum_{j \in \mathcal{R}_l} Y_j$$

Splitting criterion in the causal forest algorithm

In a causal tree the predicted CATE for observation i in leaf l is

$$\hat{\tau}_i = \hat{\tau}_{\mathcal{R}_l} = \frac{1}{N_{1l}} \sum_{j \in \mathcal{R}_{1l}} Y_j - \frac{1}{N_{0l}} \sum_{j \in \mathcal{R}_{0l}} Y_j$$

Here, $\mathcal{R}_{1l}$ includes all observations j in leaf l with treatment status $W_j = 1$, and $\mathcal{R}_{0l}$ includes all observations j with treatment status $W_j = 0$. N_{1l} and N_{0l} are the corresponding observation numbers.

We see that the predicted CATE is based on the difference in the mean outcome of the treated and untreated units. If the leaves are sufficiently small such that all units are (almost) homogenous, and under random assignment of W_i, the mean-difference in outcomes is an *unbiased estimator of the true CATE*, τ_i.

Trees are grown by successively adding splits. Unlike in a standard regression tree algorithm, growing a tree based on a split that reduces

$$RSS(\mathcal{R}) = \sum_{i \in \mathcal{R}} (Y_i - \hat{Y}_{\mathcal{R}})^2$$

by the largest amount is not directly feasible.

Why? — Because a causal tree predicts conditional average treatment effects. Ideally we would use the following criterion function to find the best split:

$$RSS(\mathcal{R}) = \sum_{i \in \mathcal{R}} (\tau_i - \hat{\tau}_{\mathcal{R}})^2$$

But this criterion is infeasible, because $\tau_i = Y_i(1) - Y_i(0)$ is not observed, unlike Y_i.

Instead, the algorithm finds a split that maximizes the variance of the predicted treatment effects, $\hat{\tau}_i$, that result from a new split. Intuition: Maximizing the variance across treatment effects adds a large degree of additional predictive power = ability to predict differences in outcomes across units i.

Example — `Causal-Forest.Rmd`

Simulated data with two inputs: `recency` and `web_buyer`

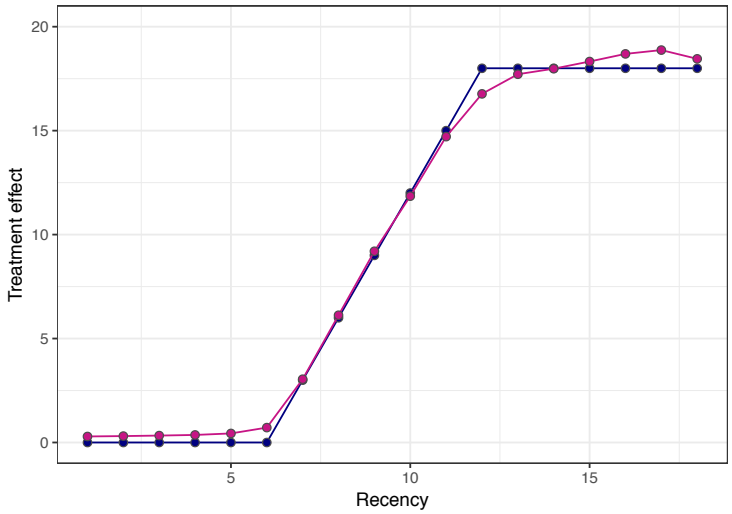
Output: Dollar spending

Customers are randomly targeted

CATE, τ:

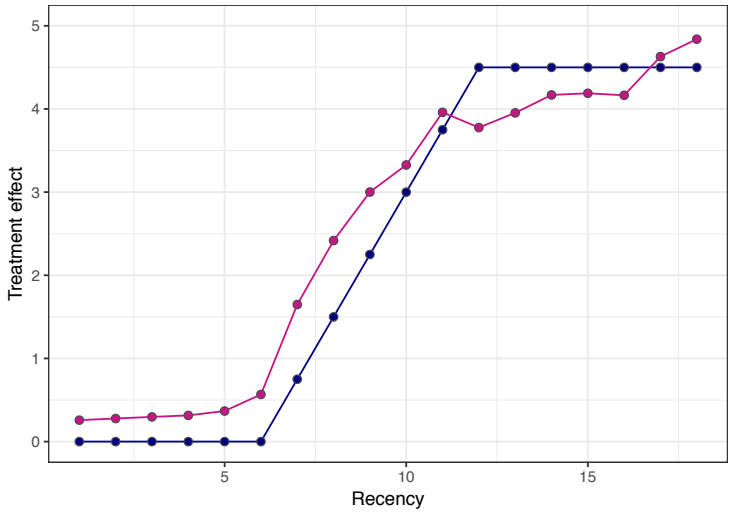
- If `recency` ≤ 6 customers are not responsive to a targeting effort
- For $6 <$ `recency` ≤ 12 τ increases and then remains constant for all values `recency` ≥ 12
- τ is four times larger for customers who are not web-buyers compared to web-buyers

Actual and estimated τ for customers who are not web-buyers (segment with large CATE):



Note: The red curve is the average prediction from the causal forest algorithm (`causalForest`)

Actual and estimated τ for web-buyers:



Note that the treatment effect is on a different scale compared to the previous graph

Heterogeneous treatment effects: Model validation

As in the standard predictive modeling approach, we can examine the predictive power of a model using a lift table/chart

Standard approach:

1. Estimate model in training sample
2. Classify units i in the validation sample into groups (scores) based on the predicted output
3. For each group (score) $\mathcal{S}_k$ predict the average output in the validation sample:

$$\frac{1}{N_k} \sum_{i \in \mathcal{S}_k} Y_i$$

Modification when predicting heterogeneous treatment effects:

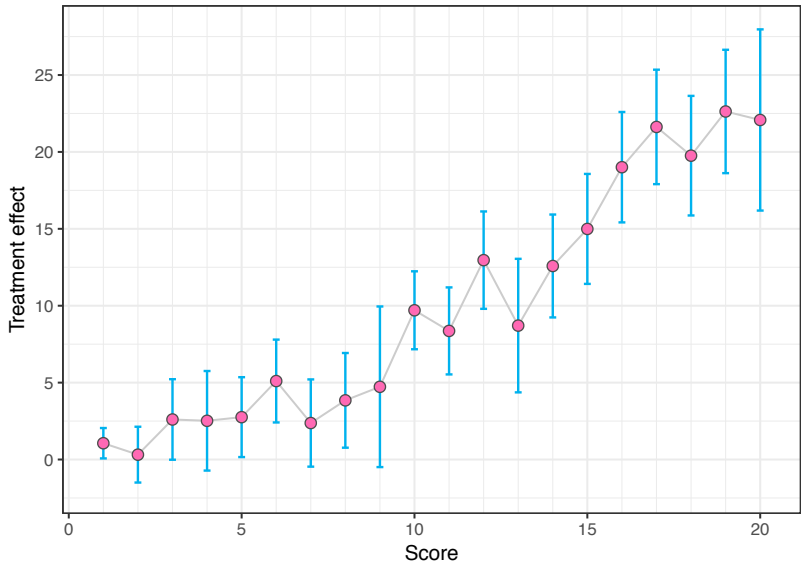
2'. Create scores based on *predicted treatment effects*, $\hat{\tau}_i$

3'. Predict the average treatment effect in group (score) $\mathcal{S}_k$:

$$\frac{1}{N_{1k}} \sum_{i \in \mathcal{S}_{1k}} Y_i - \frac{1}{N_{0k}} \sum_{i \in \mathcal{S}_{0k}} Y_i$$

A sufficient condition for this to work: The treatment is randomly assigned and the treatment probability, $e = \Pr\{W_i = 1\}$, is constant across units i.

Example



Confidence intervals: Difference in means from two independent samples

Model validation using predicted profits

Lifts are a good method to quickly assess if there is some predictive power in the model. However, lifts do not provide a measure of the **degree of the predictive power** of the model.

In any CRM application, the goal of predictive modeling is to increase profits. Hence, ideally we would compare:

(i) Profits based on a targeting strategy from predictive model $\mathcal{M}^*$

(i) Profits based on a targeting strategy from an alternative model, $\mathcal{M}$, or based on a "heuristic"

Examples of *heuristic targeting strategies* — strategies that might be reasonable but not derived from solid principles:

(a) Target all customers

(b) Target all customers who have made at least one purchase in the last 18 months but have not made a purchase in the last 6 months

Predicted profit associated with targeting strategy T

Step 1: Estimate model $\mathcal{M}^*$, for example a causal forest, in a training sample

Step 2: Predict incremental sales for each customer i in a new sample (validation sample or sample used for the actual implementation of the strategy):

$$\tau_i = \mathbb{E}(Y_i|\boldsymbol{x}_i, W_i = 1) - \mathbb{E}(Y_i|\boldsymbol{x}_i, W_i = 0)$$

Step 3: Propose a targeting strategy T:

$$T_i = T(\boldsymbol{x}_i) = \begin{cases} 0 & \text{if } m\tau_i \leq c_i, \\ 1 & \text{if } m\tau_i > c_i. \end{cases}$$

(m is the margin, c is the targeting cost c)

The profit for this targeting stategy is

$$\Pi(T) = \sum_{i=1}^{n} \left[(1 - T_i) \cdot mY_i(0) + T_i \cdot (mY_i(1) - c)\right]$$

Estimation of $\Pi(T)$ from a randomized sample (Hitsch and Misra 2017)

$\Pi(T)$ is the profit that will be realized if we implement the proposed targeting strategy T.

Can we predict this profit using past data, even though customers were not targeted using the strategy T?

The answer is yes, provided the *actual* targeting W_i in the data satisfies:

1. W_i is randomly assigned
2. The propensity score $e = \Pr\{W_i = 1\}$ satisfies $0 < e < 1$ (overlap)

Then we can construct an unbiased profit estimator $\hat{\Pi}(T)$ such that

$$\mathbb{E}[\hat{\Pi}(T)] = \mathbb{E}\left[\Pi(T)\right]$$

Profit estimator

Estimate of profit from the targeting strategy T:

$$\begin{aligned}\hat{\Pi}(T) &= \sum_{i=1}^{n}\left[\left(\frac{1-W_i}{1-e}\right)(1-T_i)\cdot mY_i(0) + \left(\frac{W_i}{e}\right)T_i\cdot(mY_i(1)-c)\right] \\ &= \sum_{i=1}^{n}\left[\frac{\mathbb{I}\{W_i,T_i=0\}}{1-e}\cdot mY_i(0) + \frac{\mathbb{I}\{W_i,T_i=1\}}{e}\cdot(mY_i(1)-c)\right]\end{aligned}$$

Notes:

- $\mathbb{I}\{\cdot\}$ is an indicator $=1$ if the condition in brackets is true and $=0$ otherwise
- W_i is the observed, randomized treatment based on the randomized targeting strategy in the data

In $\hat{\Pi}(T)$ we sum only over the realized profits when our proposed targeting strategy **coincides** with the observed treatment, $T_i = W_i$.

Note that $\hat{\Pi}(T)$ can be calculated from the data, because

$$Y_i \equiv Y_i^{\text{obs}} = \begin{cases} Y_i(0) & \text{if } W_i = 0 \\ Y_i(1) & \text{if } W_i = 0 \end{cases}$$

Hence, we can calculate $\hat{\Pi}(T)$:

$$\hat{\Pi}(T) = \sum_{i=1}^{n}\left[\left(\frac{1-W_i}{1-e}\right)(1-T_i)\cdot mY_i + \left(\frac{W_i}{e}\right)T_i\cdot(mY_i-c)\right]$$

Why does $\hat{\Pi}(T)$ predict $\Pi(T)$?

Focus on one observation, i, and suppose $T_i = 0$ — customer i is not targeted using the targeting strategy T.

Claim:

$$\mathbb{E}\left[\frac{\mathbb{I}\{W_i, T_i = 0\}}{1-e} \cdot mY_i(0)\right] = \mathbb{E}\left[\mathbb{I}\{T_i = 0\} \cdot mY_i(0)\right]$$

Because:

$$\mathbb{E}\left[\frac{\mathbb{I}\{W_i, T_i = 0\}}{1-e} \cdot mY_i(0)\right] = \mathbb{I}\{T_i = 0\} \cdot \frac{\Pr\{W_i = 0\}}{1-e} \cdot \mathbb{E}\left[mY_i(0)\right]$$

Why can we factor the original term into the three components?

- T_i and hence $\mathbb{I}\{T_i = 0\}$ is not random — we are focusing on observation i with $T_i = 0$, hence $\mathbb{I}\{T_i = 0\} = 1$
- $\mathbb{E}[\mathbb{I}\{W_i = 0\}] = \Pr\{W_i = 0\}$
- W_i is randomized and hence independent from $mY_i(0)$

Now remember that $e = \Pr\{W_i = 1\}$, and hence $\Pr\{W_i = 0\} = 1 - e$.

Therefore:

$$\begin{aligned}\mathbb{I}\{T_i = 0\} \cdot \frac{\Pr\{W_i = 0\}}{1-e} \cdot \mathbb{E}\left[mY_i(0)\right] &= \mathbb{I}\{T_i = 0\} \cdot \frac{1-e}{1-e} \cdot \mathbb{E}\left[mY_i(0)\right] \\ &= \mathbb{I}\{T_i = 0\} \cdot \mathbb{E}\left[mY_i(0)\right] \\ &= \mathbb{E}\left[\mathbb{I}\{T_i = 0\} \cdot mY_i(0)\right]\end{aligned}$$

In exactly the same manner we can show that

$$\mathbb{E}\left[\frac{\mathbb{I}\{W_i, T_i = 1\}}{1-e} \cdot (mY_i(1) - c)\right] = \mathbb{E}\left[\mathbb{I}\{T_i = 1\} \cdot (mY_i(1) - c)\right]$$

Intuition

1. In $\hat{\Pi}(T)$ we sum only over observations i when our proposed targeting strategy **coincides** with the observed treatment, $T_i = W_i$
2. If $T_i = 0$ for observation i, the probability that the observed treatment is $W_i = 0$ is $1 - e$, hence only $1 - e$ percent of all these observations can be used to predict the profit from the proposed strategy T. We therefore weight all these observations by

$$\frac{1}{1-e}$$

   (This is called inverse probability weighting)
3. If $T_i = 1$ for observation i, the probability that the observed treatment is $W_i = 1$ is e, hence only e percent of all these observations are usable. We therefore weight all these observations by

$$\frac{1}{e}$$

Using $\hat{\Pi}(T)$ in practice

The key insight is that we can evaluate the profit consequence of **any** targeting strategy T using a randomized sample (with overlap).

- Usage example I: Train different models (e.g. LASSO with treatment-interactions, causal forest, etc.), then predict the profit levels and differences in profit predictions across models in the validation sample.
- Usage example II: Train different models in year (quarter, ...) t data, evaluate models and pick one of them for the actual targeting in year (quarter, ...) $t + 1$. However, we do not target all customers but set aside a *holdout sample*, in which the actual targeting is randomized. Then we can evaluate the profit implications of the models trained using year t data in the $t + 1$ data (**external validity**).

This approach will be substantially cheaper than running A/B tests to evaluate all possible targeting strategies, and furthermore we can evaluate targeting strategies that we hadn't even thought of when we collected the data.

Applications

Unlimited:

- Customer development (e-mail, catalog, ...)
- Treatment effect of churn management program
- Incremental response to acquisition effort
- Click-through and conversion from display ad exposure
- Effectiveness of re-targeting
- Target coupons

In all these situations we can estimate heterogeneous treatment effects using ideas from machine learning and causal inference.

With randomized samples we can directly compare the profitability consequences of any targeting strategy.

Summary

- Correct targeting decisions should be based on heterogeneous treatment effects — CATE
- Conditional average treatment effects can be estimated from randomized samples
- Possible estimation methods for CATE:
  1. Linear treatment-interactions model
  2. Causal forest
- Model validation:
  - Lift charts/tables to compare predicted and "observed" CATE's
  - Unbiased estimator of profits under a proposed targeting strategy