Personalized Ads Using ML

Anthony Strittmatter

Lecturer

Prof. Anthony Strittmatter, Ph.D.

Research Interests: Business, Labour, and Health Economics, Program Evaluation, Computational Data Analytics

Positions:

Since 2023	Full Professor for Applied Econometrics at UniDistance Suisse in Valais/Switzerland
2022-2024	Senior Economist at Amazon in London/UK
2020-2023	Institut Polytechnique in Paris/France
2014-2020	University of St.Gallen/Switzerland, with research visits at UC Berkeley/US, Stanford University/US, and LMU Munich/Germany
2009-2016	Albert-Ludwig University of Freiburg/Germany

- Email: anthony.strittmatter@unidistance.ch
- Webpage: www.anthonystrittmatter.com

Lecture Outline

1. **Economics of Advertising**
   - Why firms advertise: information, persuasion, heterogeneity
   - Limits of mass and rule-based targeting
2. **Prediction with Machine Learning for Ads**
   - ML as a tool for predicting individual outcomes
   - Decision trees and random forests
   - Using prediction models to personalize advertising
3. **From Prediction to Causal Effects (Uplift)**
   - Prediction vs. treatment effects
   - Causal forests for personalized advertising
   - Policy learning for targeting

References

- Ascarza (2018): "Retention Futility: Targeting High-Risk Customers Might be Ineffective" Journal of Marketing Research, 55(1), 80-98, download.
- Strittmatter (2025): "Machine Learning for Causal Inference in Economics", IZA World of Labor, No. 516, download.
- Mullainathan and Spiess (2017): "Machine Learning: An Applied Econometric Approach", Journal of Economic Perspectives, 31 (2), pp. 87-106, download.
- Athey (2017): "Beyond Prediction: Using Big Data for Policy Problems", Science, 355 (6324), pp. 483-485, download.
- Cagala, Rincke, Glogowsky, Strittmatter (2021): "Optimal Targeting in Fundraising: A Machine Learning Approach", download.
- James, Witten, Hastie, Tibshirani (2023): "An Introduction to Statistical Learning", 2nd edition, Springer, download.

PC Lab Exercises for Self-Study

- I provide code for three key use cases covered in the lecture
- The exercises let you replicate, explore, and extend these use cases independently

- All course materials (slides, data, and code) are available on GitHub: github.com/AStrittmatter/Machine-Learning-Course
- PC labs are based on interactive Jupyter notebooks that run directly in your browser: mybinder.org

General

- Feel free to interrupt me at any time when you have questions.
- Tell me when I'm too slow or too fast. Ask me to repeat material in case something was not clear.
- You can also send me an email with questions: anthony.strittmatter@unidistance.ch
- Proposals to improve the lecture are also welcome.

Economics of Advertising

From demand shifts to the economic rationale for personalized advertising

What Is Advertising?

- Advertising is a firm's instrument to influence consumer demand
- It operates by affecting:
  - information (what consumers know)
  - beliefs (how consumers evaluate products)
  - attention (which options are considered)
- It differs from pricing and product design

Economic Definition

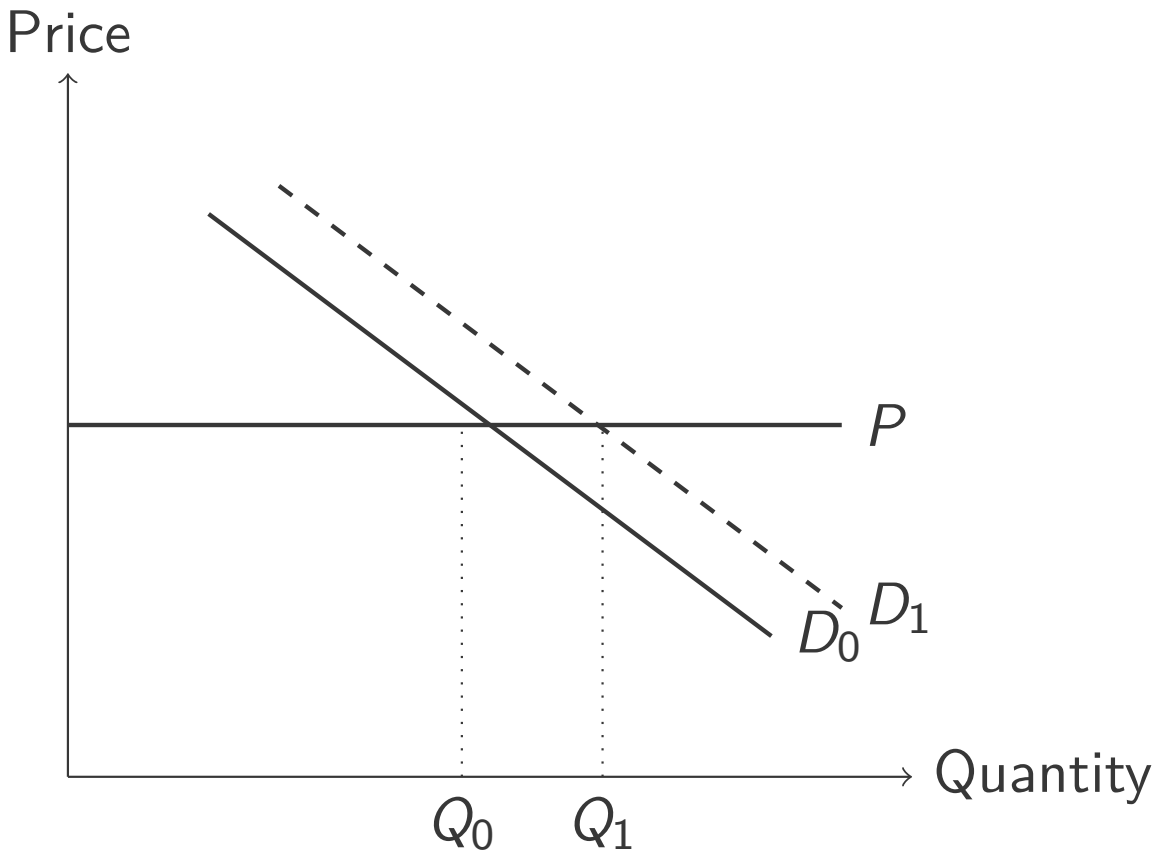
Advertising shifts demand by changing information or salience, without directly changing prices or product characteristics.

Why Do Firms Advertise?

1. **Reduce information frictions**
   - Consumers may not know that a product exists
   - They may be uncertain about price, quality, or fit
2. **Create awareness among potential buyers**
   - A product cannot be purchased if consumers are unaware of it
   - Advertising moves consumers into the "consideration set"
3. **Influence marginal purchase decisions**
   - Advertising can tip decisions when consumers are almost indifferent
   - Small effects on many consumers can generate large revenues

Economic implication: Advertising shifts or rotates demand by increasing willingness to buy at a given price.

Advertising as a Demand Shift



- Advertising shifts demand from D_0 to D_1
- Prices P are fixed in the short run
- Sales increase from Q_0 to Q_1

Example: Amazon Prime

- Some Prime subscribers consider canceling their membership
- They underestimate how often they use Prime benefits (delivery, video, music)
- Amazon sends personalized emails and in-app reminders about recent usage

Economic mechanisms:

- Consumers have imperfect recall of past usage and benefits (information frictions)
- Personalized summaries make benefits more visible at decision time
- Better information raises perceived surplus from staying subscribed

Why this matters economically:

- Reducing information frictions stabilizes demand without lowering prices
- Small reductions in churn have large effects on lifetime customer value

Example: Google Search Ads

- A consumer searches for *"running shoes"* on Google
- Sponsored ads appear above organic search results

Economic mechanisms:

- Ads make consumers aware of brands and products they did not know
- Only visible products enter the consideration set of consumers
- Higher placement in search results increases click probability
- Ads reduce the effort required to find relevant offers

Why this matters economically:

- Products not seen are effectively not demanded
- Advertising shifts demand by expanding and reshaping choice sets
- Even without changing prices, firms can reallocate market shares

Example: Netflix

- Netflix advertises a new flagship series (e.g. *Stranger Things*) on Instagram and YouTube
- Subscription price and overall catalog remain unchanged
- A modest increase in new subscriptions is observed after the campaign

Economic mechanism:

- Many consumers are *almost indifferent* between subscribing and not
- Advertising increases awareness and salience of specific content
- This slightly raises perceived value and willingness to pay
- For some consumers, willingness to pay crosses the subscription price

Why this matters economically:

- Individual effects are small, but applied to millions of users, they generate substantial revenue

Heterogeneous Consumers

- Consumers differ systematically in:
  - **preferences** (needs, tastes, brands)
  - **price sensitivity** (budget constraints, urgency)
  - **attention and responsiveness** to advertising

Example: Same Ad, Different Effects

- Customer already planning to buy running shoes $\rightarrow$ ad accelerates purchase
- Casual browser $\rightarrow$ ignores the ad
- Loyal customer of another brand $\rightarrow$ ad has no effect

Key insight: The same ad can increase demand for some consumers and have no (or even negative) effect for others

Different Responses to the Same Ad

Scenario: Online ad for a luxury watch shown to all users

High-income consumer

- Already interested in luxury brands
- High willingness to pay
- Ad increases purchase probability

Low-income consumer

- Budget constrained
- Low relevance of the product
- Ad has no effect on behavior

Economic implication:
Showing the same ad to everyone wastes budget and reduces effectiveness.

Traditional (Mass) Advertising

How advertising used to work

- One ad message shown to everyone
- No individual-level information used
- Common channels:
  - TV commercials
  - newspapers
  - billboards

Problem:
Most impressions go to consumers with low or zero responsiveness.

Inefficiency of Mass Advertising

Scenario: TV advertising for the PlayStation 5

Relevant audience

- households without a current-generation console
- active video game players
- parents considering a console purchase for children

Irrelevant audience

- households that already own a PlayStation 5
- viewers with no interest in video games
- individuals with budget or age constraints

Economic problem:
Mass advertising reaches many consumers with zero probability of purchase, leading to wasted advertising expenditure and low average returns.

Why Do Firms Use Targeting?

Advertising as an investment decision

- Each ad impression has a cost
- Consumers differ in their probability of responding
- Showing ads to low-response users wastes budget

Economic objective

Maximize expected returns by allocating ads to consumers with higher expected response.

Rule-Based Targeting

Early approach to personalized advertising

- Advertising decisions based on simple if–then rules
- Rules designed manually by marketers
- Consumers grouped into a small number of predefined segments

Typical decision logic

IF user satisfies the rule $\rightarrow$ show ad
ELSE $\rightarrow$ do not show ad

Key feature: The same rule is applied to all users within a segment.

Example: Gym Membership Advertising

Example: Planet Fitness / Anytime Fitness

- Facebook ad campaign to acquire new gym members
- Rule-based targeting:
  - Age between 20 and 40
  - Lives within 10 km of a gym
- Users satisfying the rule differ strongly in:
  - interest in fitness
  - past exercise behavior
  - likelihood of signing up

Economic insight: Rules improve targeting relative to mass advertising, but still treat unequal users equally.

Common Types of Rule-Based Targeting

- **Demographic targeting**
  - age, gender, income, location
  - e.g. ads for student loans shown to young adults
- **Behavioral targeting**
  - past purchases, browsing history
  - e.g. retargeting users who visited a product page
- **Interest-based targeting**
  - broad categories inferred from activity
  - e.g. “fitness enthusiasts”, “travel lovers”

Common feature: Rules rely on coarse groupings rather than individual-level predictions.

Why Rule-Based Targeting Is Limited

Economic problem of coarse segmentation

- Large heterogeneity within segments
- Rules ignore interactions and nonlinearities
- Manual rules do not scale to rich user data

Economic consequence

Ads are shown to many low-response users and withheld from some high-response users.

Implication: There are large gains from individual-level, data-driven targeting.

From Rules to Data-Driven Targeting

What has changed in advertising markets?

- Platforms observe rich, high-dimensional user data
  - browsing behavior, search queries, app usage, location, time
- Simple if–then rules cannot exploit this information
- Manually designing rules does not scale

Economic shift

Targeting moves from hand-crafted rules to data-driven decision problems.

Next step: Prediction and causal inference with machine learning

Key Takeaways: Economics of Advertising

- Advertising shifts demand by changing information and attention
- Consumers respond heterogeneously to ads
- Uniform advertising is inefficient in heterogeneous markets
- Targeting improves efficiency by matching ads to responsive consumers
- The complexity of targeting decisions motivates ML-based personalization

Prediction with Machine Learning for Ads

Using data to predict individual advertising outcomes

The Core Decision Problem in Digital Advertising

Think like an advertiser:

- You can potentially show ads to **many** users, but:
  - ads cost money
  - user attention is limited
  - users differ in how likely they are to respond

Example: Online shoe retailer

The firm wants to decide:

- who should see an ad,
- how much to bid for an impression,
- which message to show.

Where Machine Learning Appears in Practice

- **Google Search Ads**
  - *Who should see the ad:* predict which queries and users are likely to click
  - *How much to bid:* estimate expected value of an impression to guide bids
  - *Which message to show:* select ad text matching the search intent
- **Uber**
  - *Who should see the message:* predict which riders are likely to stop using Uber
  - *How much to spend:* choose the minimum incentive needed to prevent churn
  - *Which message to show:* decide between a price discount, ride reminder, or feature highlight

What Exactly Do Firms Predict in Advertising?

- Firms do not predict actions directly – they predict **probabilities and expected values**
- Common prediction targets:
  - **Click through rate (CTR):** $Pr(\text{click} \mid X)$
  - **Conversion rate:** $Pr(\text{purchase} \mid \text{click}, X)$
  - **Expected revenue:** $E(\text{value} \mid X)$
  - **Churn risk:** $Pr(\text{churn} \mid X)$

Interpretation

Each user receives a *score* summarizing how valuable or responsive they are expected to be.

Machine Learning for Advertising: What Does It Do?

- **Machine learning (ML)** builds predictive models from data

$$\widehat{Y} = \hat{f}(X)$$

- In digital advertising:
  - Y: clicks, purchases, revenue, or churn
  - X: user characteristics, past behavior, context
- The output is a **score for each user**:
  - how likely they are to click
  - how likely they are to buy
  - how much revenue they are expected to generate

Interpretation: ML turns rich user data into predictions that can be used to target users.

Why Use Machine Learning in Advertising?

Why ML works well in practice:

- handles many variables at once (high-dimensional data)
- captures nonlinearities and interactions automatically
- often delivers strong **out-of-sample** prediction accuracy

Important limitations:

- many ML models are hard to interpret ("black box")
- good prediction does *not* imply causal relationship

Key takeaway: ML is excellent for predicting users behaviour, but prediction alone does not answer whether ads *change* behavior.

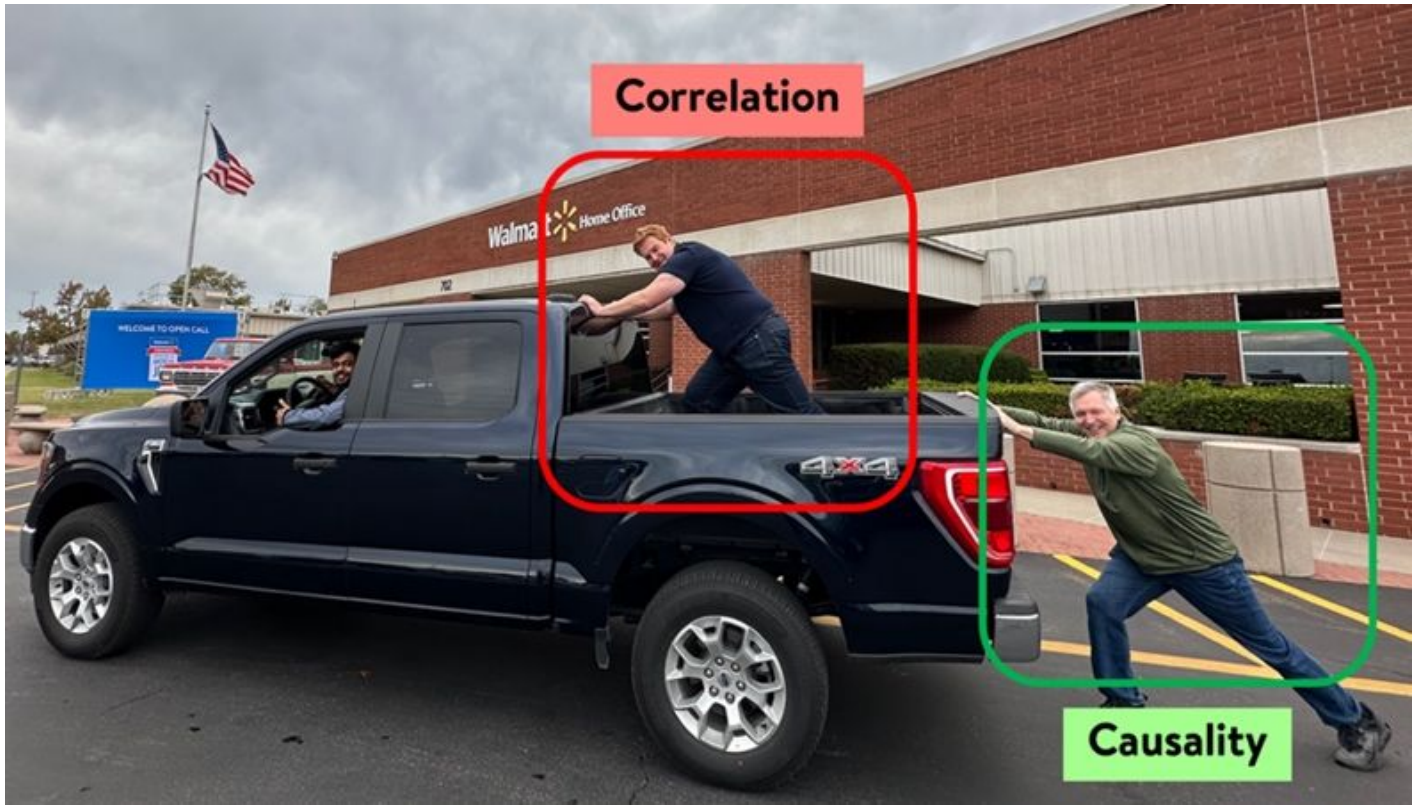
Prediction vs. Causality in Advertising

- **Prediction answers:**
  - Who is likely to buy?
  - Who has high expected spending?
- **Causality answers:**
  - Who buys *because* of the ad?
  - For whom does the ad change behavior?

Why this matters

A user with a high predicted purchase probability may have bought anyway. Prediction success does not imply that advertising caused the purchase.

Prediction vs. Causality in Advertising



John List and Jeffrey Lachman

When Prediction Is Still Useful for Advertising

- Prediction alone does not identify causal effects
- But real-world advertising decisions rely on **economic assumptions**

Common assumptions used in practice

- Users with very low churn risk do not need advertising
- Only high-spending users are worth persuading
- Users close to a purchase decision are more responsive to ads

Key message: Prediction becomes decision-relevant once combined with explicit behavioral assumptions.

How Do Trees Make Predictions?

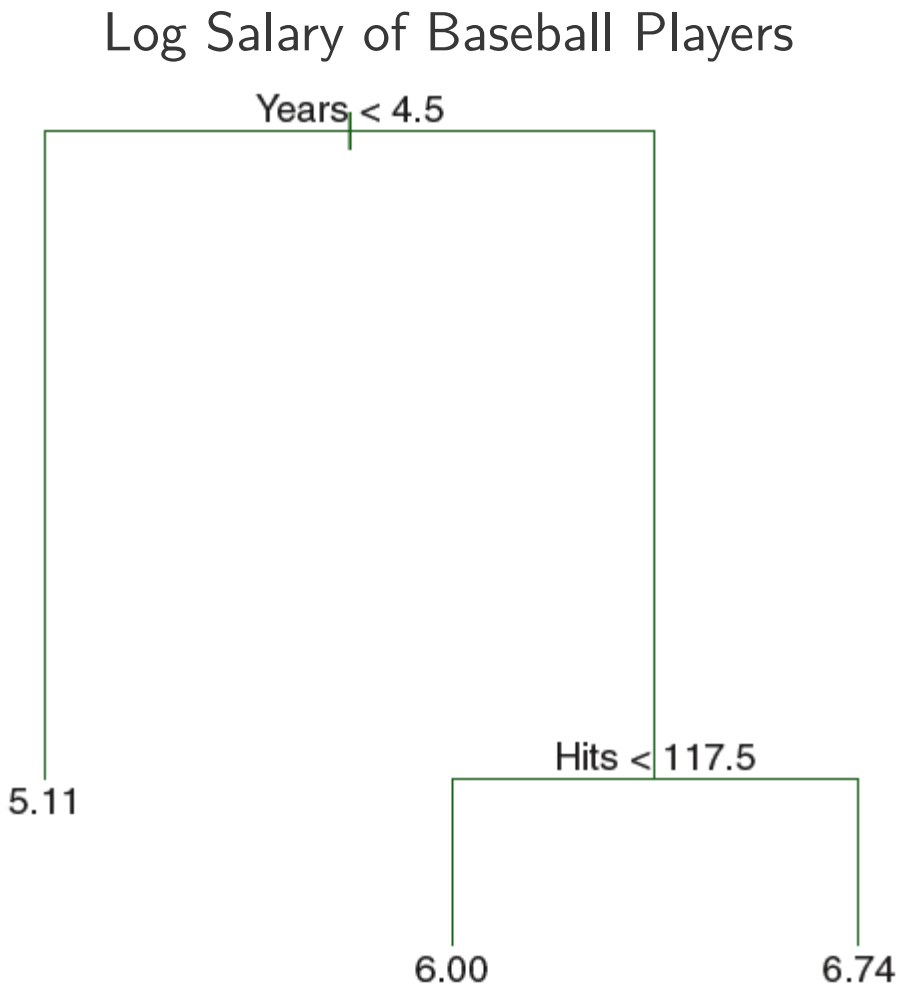
Idea: partition the market into "if–then" segments

1. Start with all consumers in one group
2. Pick a split (e.g., *visited sports sites?*) that best reduces prediction error
3. Repeat inside each subgroup until the tree is "good enough"

What the tree outputs at the end

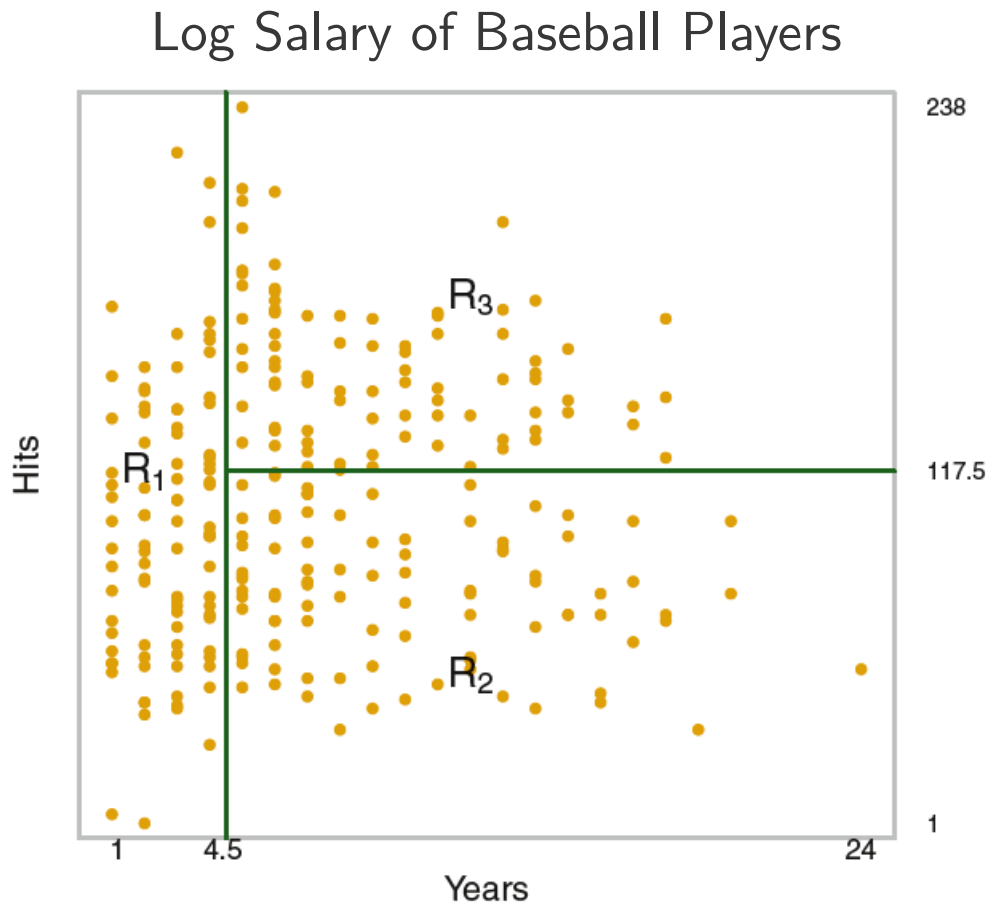
Each terminal node (leaf) stores a **predicted outcome**, e.g.
$\widehat{E}[Y \mid X]$ = predicted spending / purchase probability.

Example: Shallow Tree



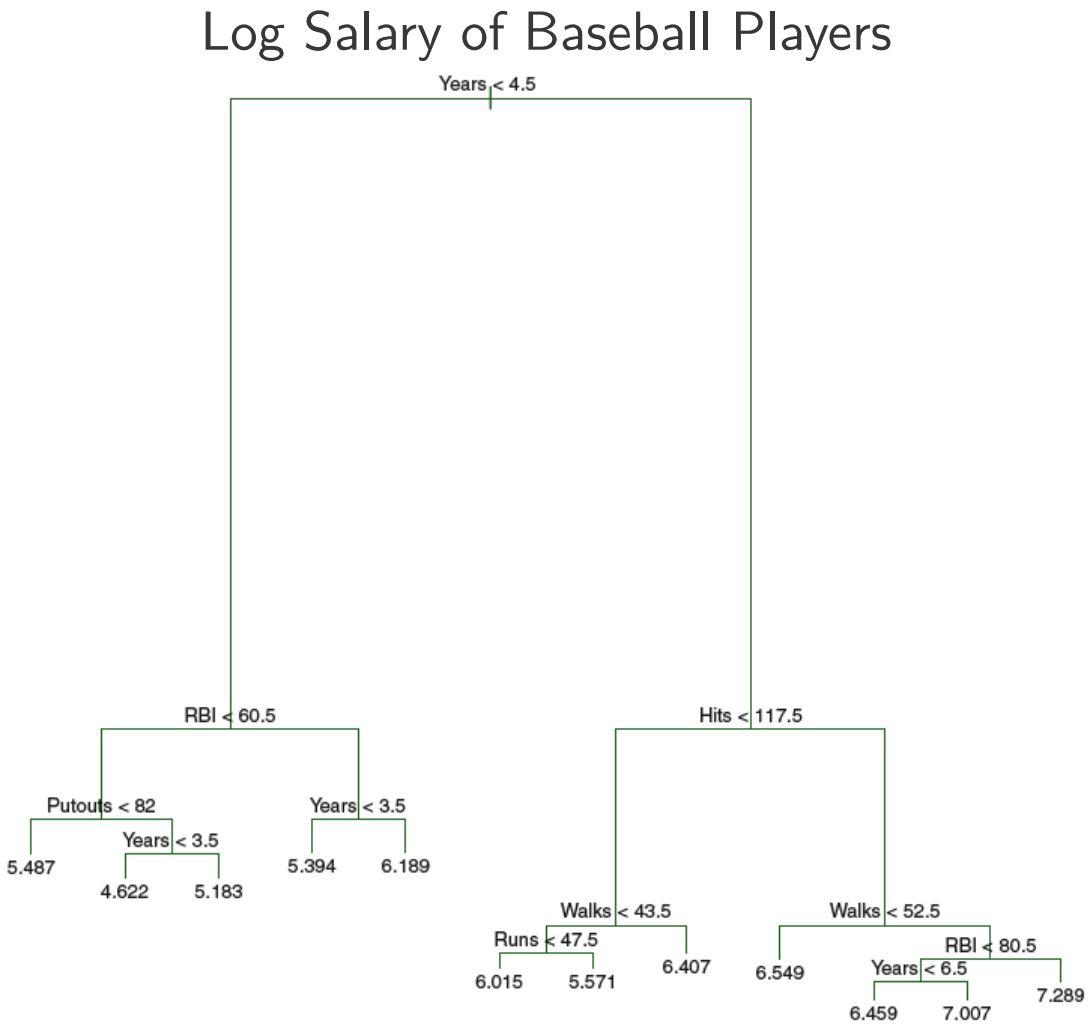
Source: James, Witten, Hastie, Tibshirani (2013)
Interpretation: Each terminal node defines a group of players with similar characteristics; the model predicts their log salary as the average log salary observed in that segment.

Example: Shallow Tree (cont.)



Source: James, Witten, Hastie, Tibshirani (2013)

Example: Deep Tree



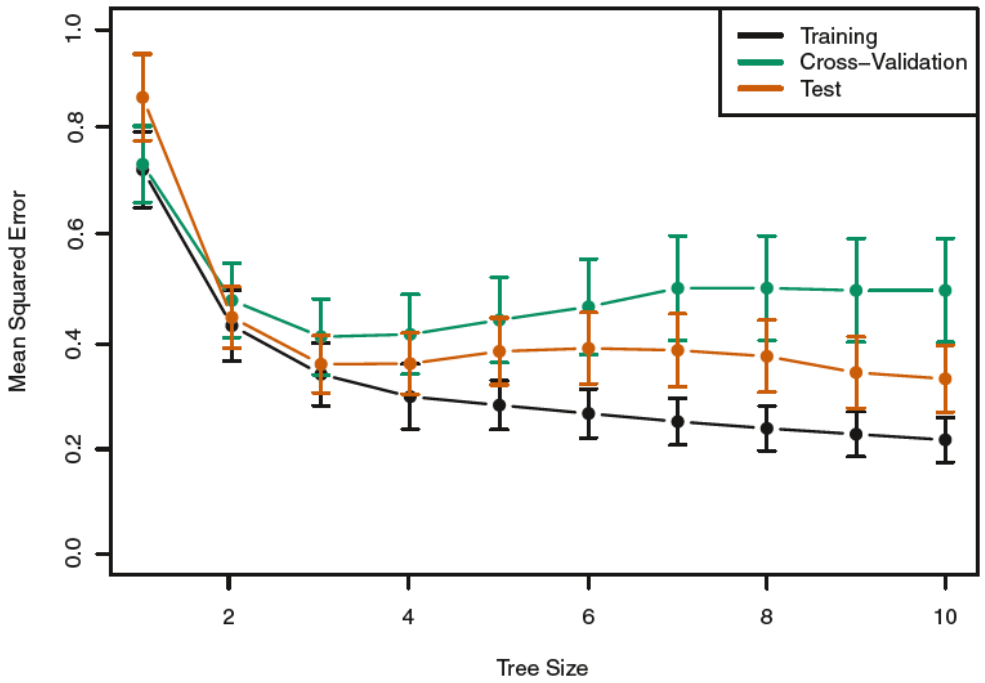
Source: James, Witten, Hastie, Tibshirani (2013)

Shallow vs. Deep Trees

- Shallow trees:
  - easy to explain and communicate
  - may miss important patterns (underfitting)
- Deep trees:
  - very flexible
  - can overfit noise (poor out-of-sample performance)

Trade-off: interpretability vs. predictive accuracy

Selecting Optimal Tree Size



→ Very large trees can fit past data extremely well.

→ But they may perform poorly on new, unseen data.

→ Choose the tree size that predicts *future outcomes* most accurately.

Source: James, Witten, Hastie, Tibshirani (2013)

Limitations of Single Trees

- Small data changes can lead to very different trees
- Predictions may be unstable
- This may limits out-of-sample performance

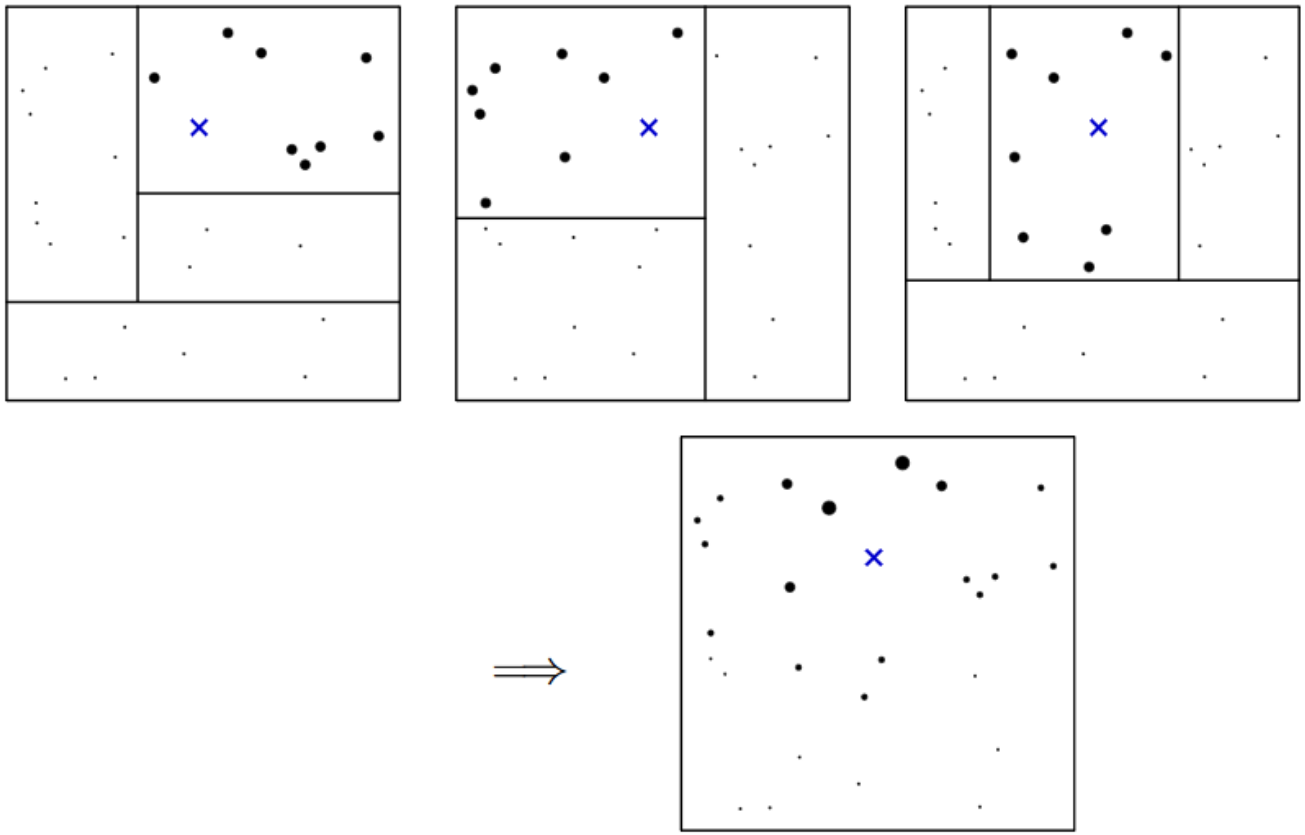
Idea: Combine many trees to improve prediction

Random Forests: Core Idea

- Build many trees on different subsamples of the data (and different candidate predictors)
- Each tree makes a prediction
- Final prediction is the average across trees
- Lower variance than single trees

Economic intuition: Aggregation reduces noise and improves predictive accuracy

Random Forest: Weighted Representation



Source: Athey, Tibshirani, Wager (2018)

Use Case: Comscore Data for Digital Advertising

- We use **real household-level data** collected by **Comscore**
- Comscore is a global analytics company measuring:
  - online purchases,
  - website traffic,
  - digital advertising exposure
- Its data are widely used by:
  - advertisers,
  - media companies,
  - large digital platforms

Why this matters

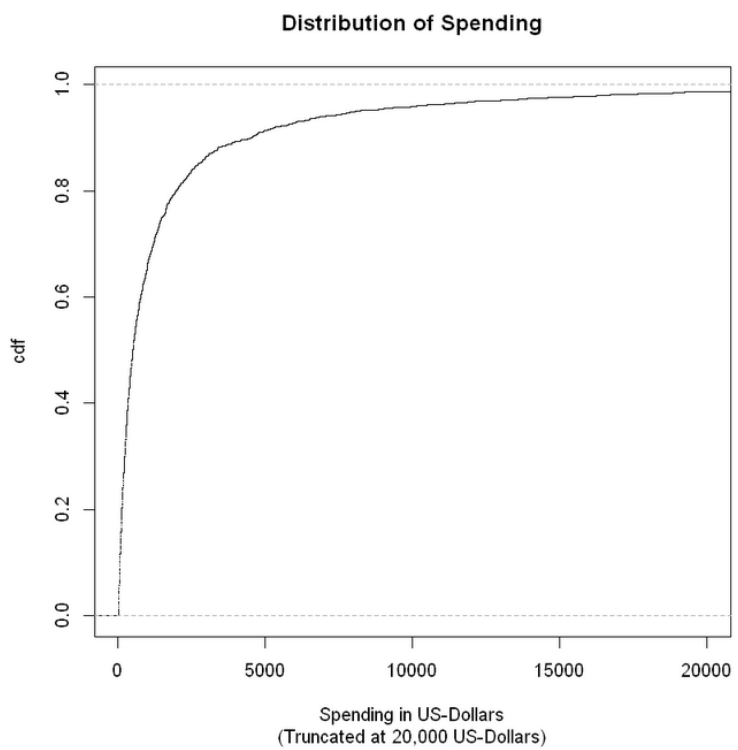
This is the type of data firms actually use to make targeting, bidding, and personalization decisions in digital advertising.

What Do We Observe in the Comscore Data?

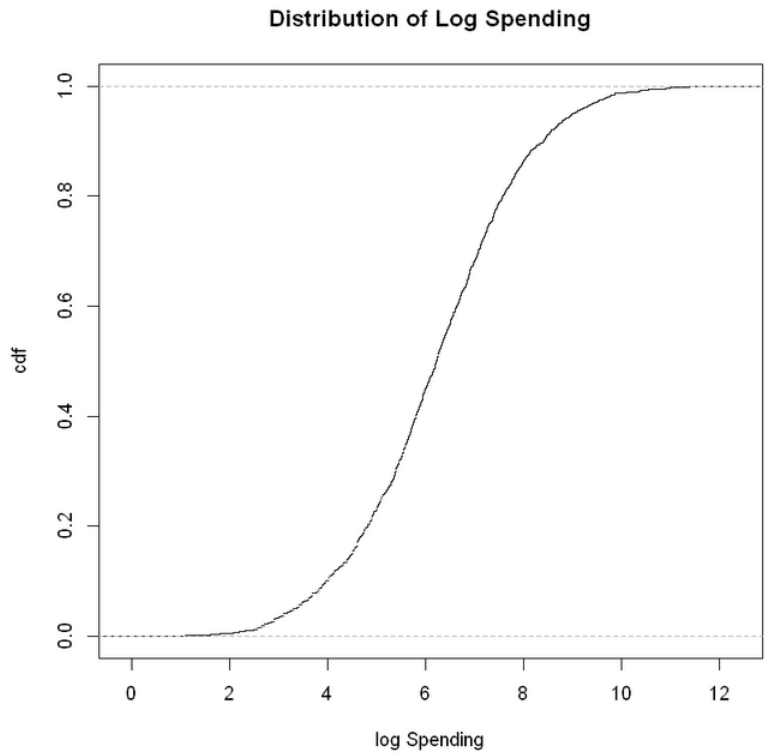
- For each household, we observe:
  - **Annual online spending** (observed in historical data)
  - **Browsing behavior** across the **1,000 most visited websites**
- Browsing data are measured as:
  - share of total online time spent on each website
- Two groups of households:
  - **Historical users:** browsing + spending observed
  - **New users:** browsing observed, spending not yet observed

Goal: Use historical browsing behavior to predict the value of new users for advertising.

Outcome Distribution: Spending Is Highly Skewed



Spending (USD)



Log spending (more symmetric)

- Many households spend little; a few spend a lot

Decision Trees with Comscore Data

- Consider an advertiser selling products online
- The advertiser wants to identify high-value households for advertising

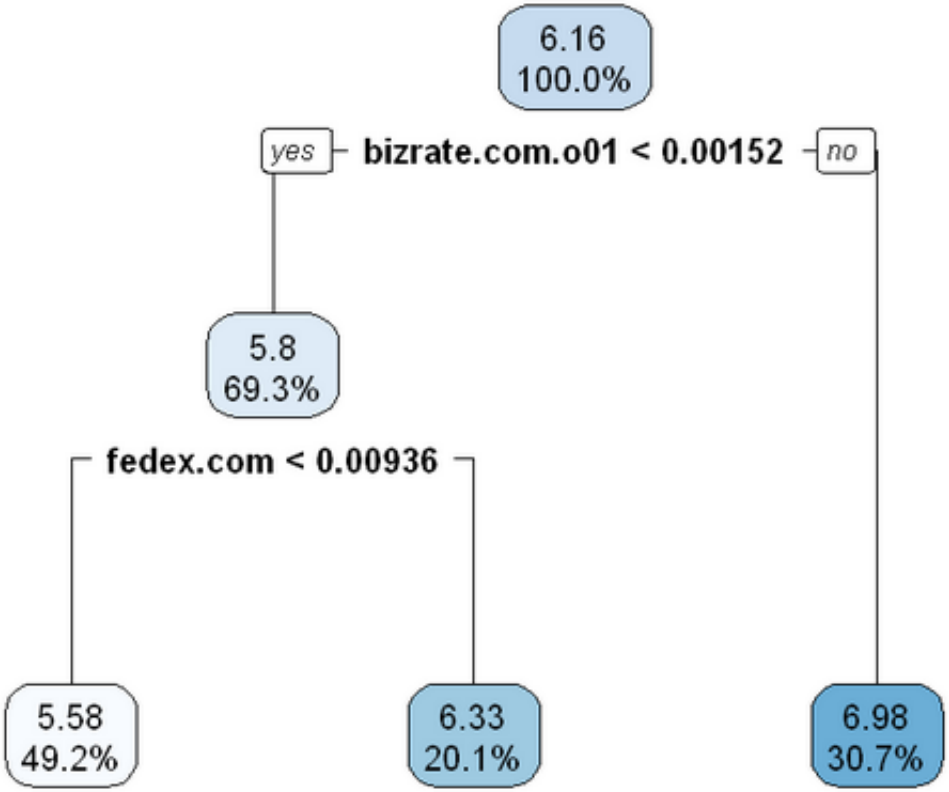
What the decision tree learns

- The tree uses detailed web browsing behavior across many websites
- It splits households into segments with similar browsing patterns
- Each terminal node corresponds to a group with similar *predicted online spending*

Key idea

Households with similar browsing behavior tend to have similar online spending patterns.

Example Tree: Browsing Predicts Spending



- Splits are based on browsing shares of particular websites
- Each leaf corresponds to a segment with different predicted spending

From Prediction to Advertising Decisions

- For new households, only browsing behavior is observed
- A tree predicts expected online spending

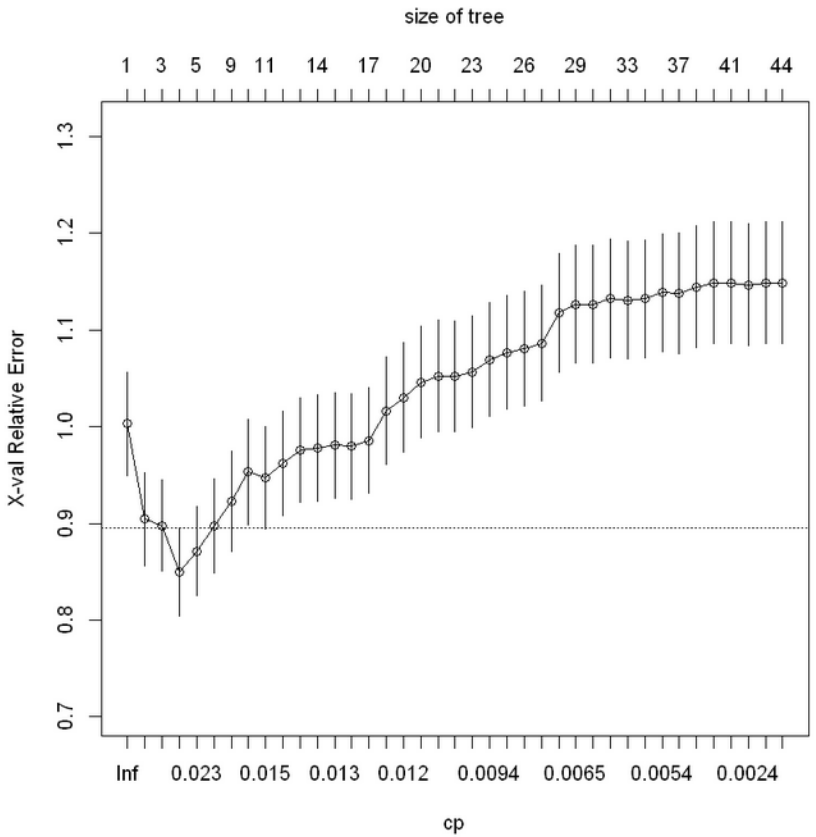
How advertisers act on predictions

- **High predicted spending:** bid more, show premium ads
- **Medium predicted spending:** show standard ads or promotions
- **Low predicted spending:** limit exposure or show no ads

Behavioral assumption

Households with higher online spending tend to be more responsive to advertising.

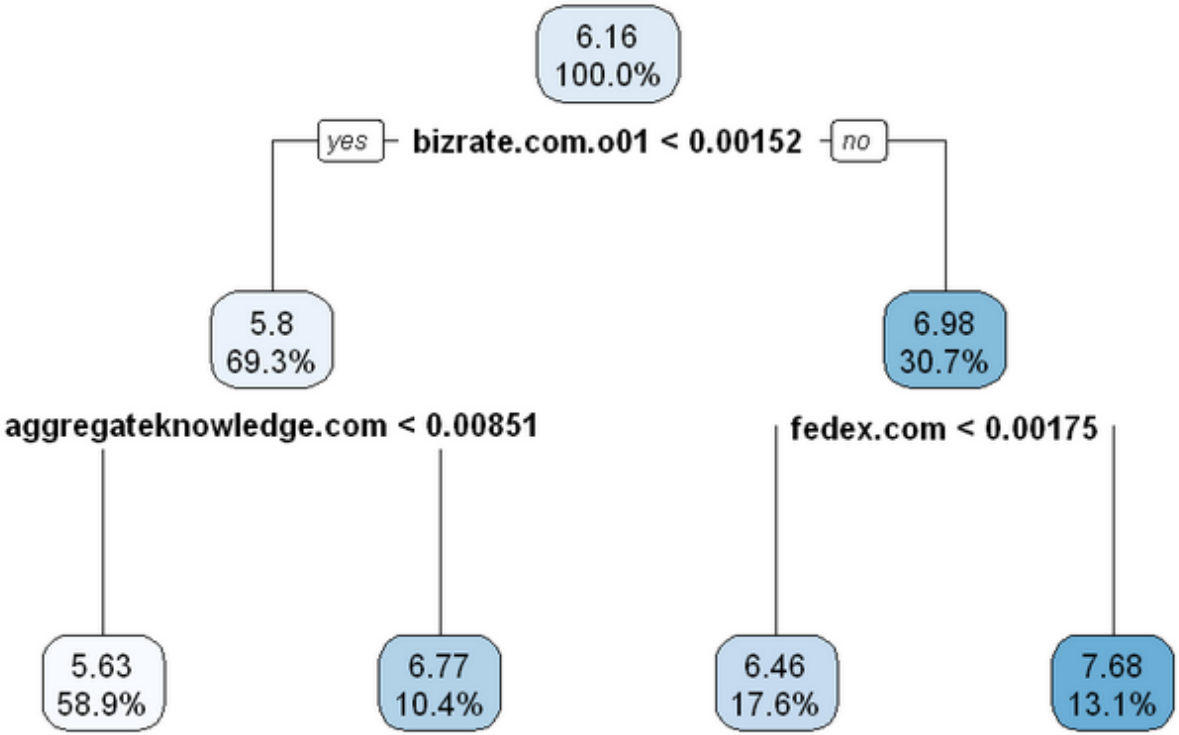
Avoiding Overfitting: Pruning with Cross-Validation



Cross-validation error vs. tree complexity

- Deep trees can fit noise (training accuracy looks great)
- Pruning improves stability and generalization
- Cross-validation chooses complexity that predicts well on new users

Optimized Tree Size



Why Stop at One Tree?

- A single tree can be unstable: small data changes $\Rightarrow$ different splits
- This instability can change the ranking of users (bad for targeting/bidding)

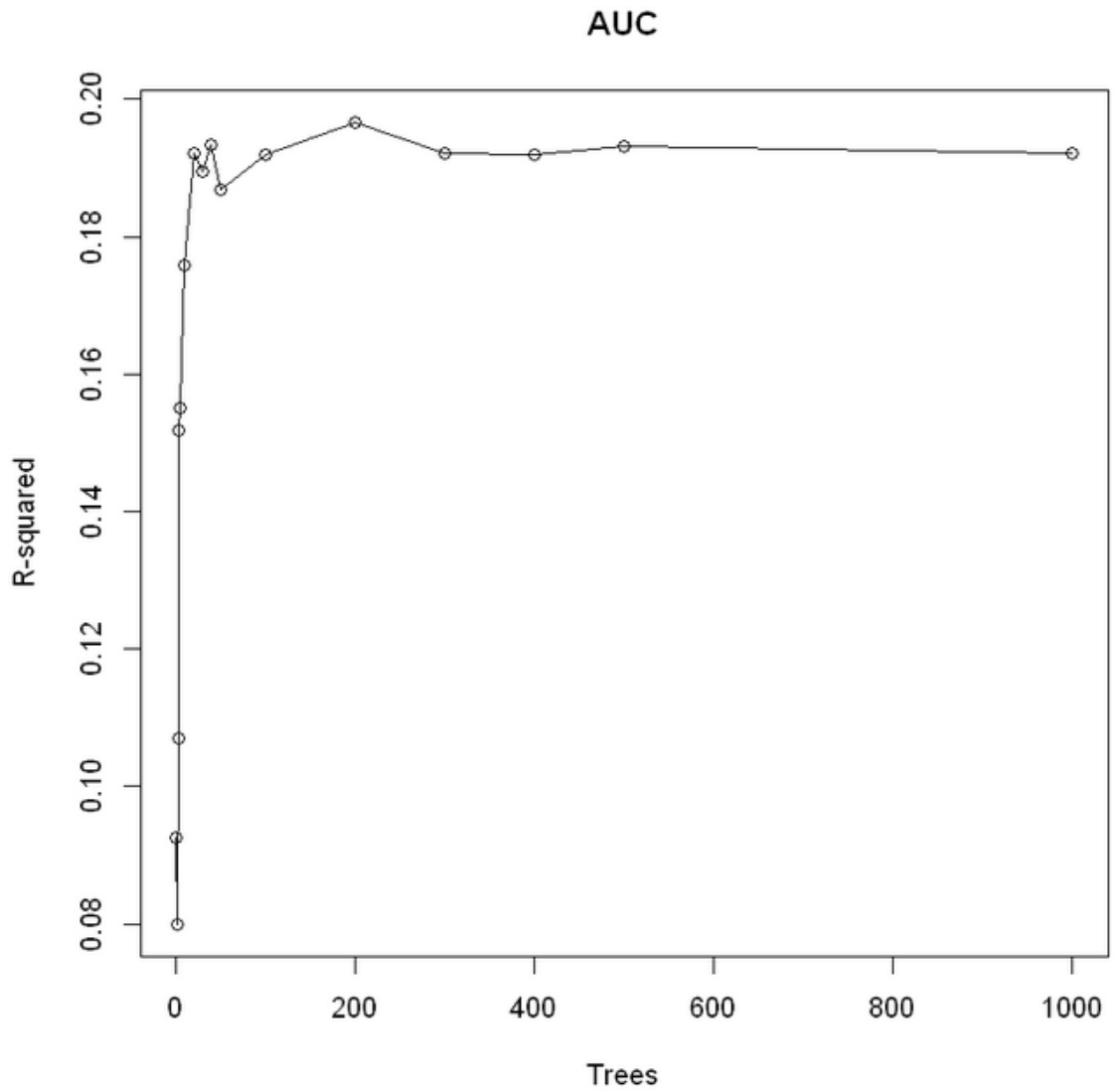
Solution: build many trees and average them (random forests).

Random Forests: Many Trees, Better Prediction

- Build many trees on different subsamples (and different candidate predictors)
- Each tree is noisy; averaging reduces variance
- Compared to a single tree:
  - higher predictive accuracy
  - more stable rankings of consumers
  - less interpretability

Economic intuition: better prediction $\Rightarrow$ better budget allocation.

Predictive Performance Improves with More Trees



- Performance improves quickly as we add trees

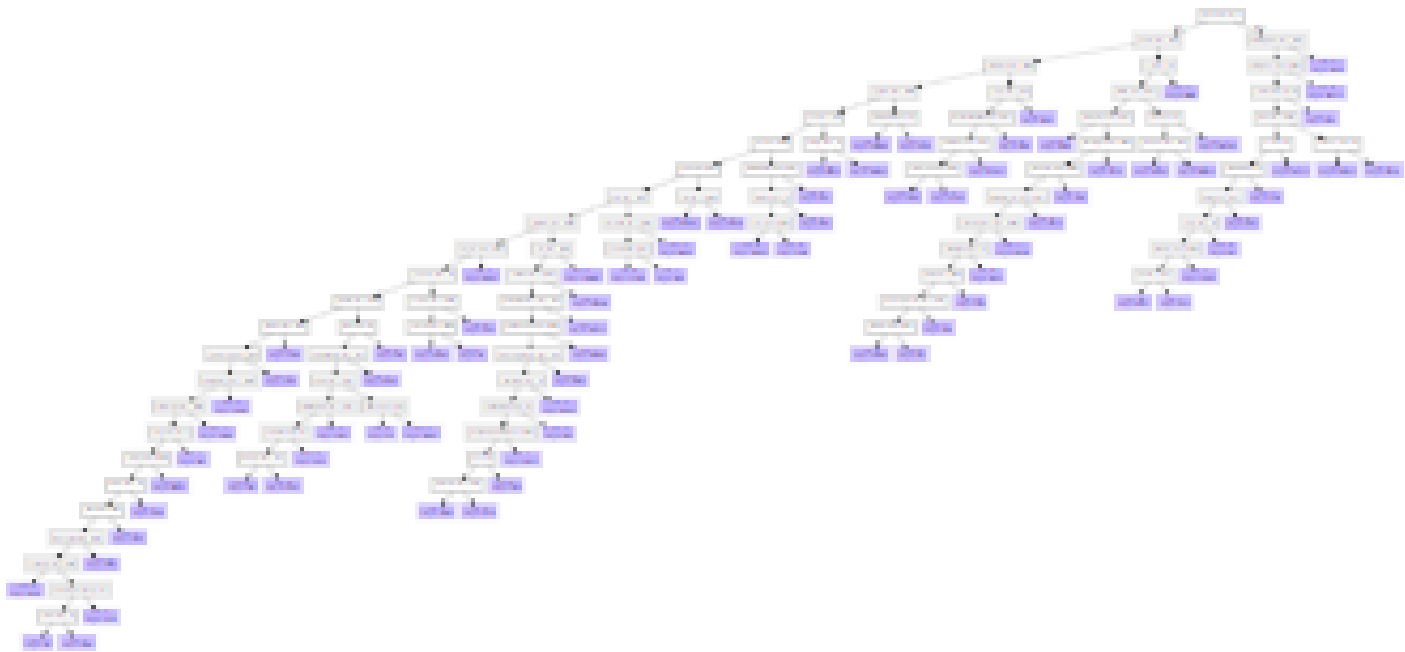
Why Random Forests Are Hard to Interpret

- A forest aggregates predictions from many deep trees
- Each tree uses different subsamples and different candidate splits
- There is no single, simple set of rules explaining the final score

Black-box intuition

A single tree gives one transparent segmentation rule. A forest gives a highly accurate score, but the "rule" is the average of many rules.

A Single Tree Inside a Random Forest (Illustration)



Variable Importance: Which Browsing Variables Matter?

- Even if the model is complex, we can summarize it
- Variable importance: which predictors frequently help create informative splits
- Two versions (as in our code):
  - importance from early splits (max.depth $= 1$)
  - importance from first few levels (discounted, max.depth $= 4$)

Variable Importance (Top Predictors)

Website	Importance
bizrate.com.o01	0.264
aggregateknowledge.com	0.180
fedex.com	0.142
ups.com	0.087
liveperson.net	0.039
marriott.com	0.033
jcpenney.com	0.025
searchmarketing.com	0.023

Note: Importance based on early splits (max.depth = 1)

Variable Importance (First Four Splits)

Website	Importance
bizrate.com.o01	0.220
aggregateknowledge.com	0.140
fedex.com	0.113
ups.com	0.069
liveperson.net	0.036
marriott.com	0.029
jcpenney.com	0.022

Note: Discounted importance from early levels (max.depth = 4)

Variable Importance (Why Be Cautious?)

- **Not causal:** importance $\neq$ "this site causes spending"
- **Correlation:** correlated variables can substitute for each other
- **Measurement:** variables with more variation/split opportunities can look more important

Takeaway: variable importance is useful for summarizing prediction, not for identifying mechanisms.

Deployment: Personalized Ads for New Households

1. Train the model on historical data (spending observed)
2. Predict spending for new households (only browsing observed)
3. Rank households by predicted value (model score)
4. Implement an ad policy:
   - target top $K\%$ (budget constraint)
   - bid more for high-value users in the auction
   - match premium vs. discount messaging

Key point: Prediction enables personalization *before* observing purchases.

Prediction and Personalized Advertising

- Trees and random forests predict **individual outcomes**
  - expected spending,
  - purchase probability,
  - user value

- Firms use these predictions to:
  - rank users,
  - allocate ads,
  - reduce waste compared to broad targeting rules

Key limitation

A user with high predicted spending may have purchased even without seeing the ad. Prediction alone does not measure *ad effectiveness*. To optimize advertising, we need to measure **incremental effects**.

Key Takeaways: Prediction for Personalized Ads

- Prediction is central to modern digital advertising
- Machine learning turns rich user data into **individual-level scores**
- These scores guide:
  - targeting,
  - bidding,
  - ad selection
- Prediction improves efficiency, but:
  - it does not answer whether ads *change* behavior

Looking ahead

To decide *who should be shown an ad*, we must move from prediction to causal effects (uplift and policy learning).

From Prediction to Causal Effects (Uplift)

Why predicting outcomes is not enough for personalized advertising

Roadmap of This Section

1. Why causal modeling is better than prediction for targeting
2. Potential outcomes framework (and identification)
3. Heterogeneous effects: CATE and causal forests
4. From effects to actions: why policy learning
5. Fundraising field experiment: policy learning for targeting

Big picture

Prediction ranks users by expected outcomes.
Causal ML estimates *incremental* effects.
Policy learning chooses actions to maximize an objective (profit/welfare) under constraints.

Prediction Is Not the Same as Advertising Impact

- In prediction, we learn $E[Y \mid X]$ (e.g., spending, purchase probability)
- In advertising, the key object is the **incremental effect** of showing the ad

Key distinction

- **Prediction:** Who is likely to buy?
- **Causality/Uplift:** Who buys *because of the ad*?

Implication: High predicted buyers can be *wasted budget* if they would buy anyway.

Why Causal Modeling Is Better for Targeting

- If you target based only on prediction, you need extra assumptions like:
  - "High predicted buyers are also the most persuadable"
  - "Low churn risk $\Rightarrow$ no need to advertise"
- Causal modeling is designed to answer the question directly:

What changes if we show the ad?

Takeaway

Causal estimates reduce the reliance on ad-hoc behavioral assumptions when choosing who to target.

Simple Example: Three Types of Consumers

Three stylized types

- **Sure buyers:** buy with or without the ad (uplift ≈ 0)
- **Persuadables:** buy only if they see the ad (uplift > 0)
- **Never buyers:** do not buy even with the ad (uplift ≈ 0)

Targeting goal: spend budget on **persuadables**.
Prediction failure mode: sure buyers often look "best" in predicted conversion.

Advertising as a Treatment: Potential Outcomes

- Let $D_i \in \{0, 1\}$ indicate whether user i is shown an ad
- Potential outcomes:
  - $Y_i(1)$: outcome if shown the ad (purchase, spending, churn, etc.)
  - $Y_i(0)$: outcome if not shown the ad

Individual ad effect (uplift)

$$\tau_i = Y_i(1) - Y_i(0)$$

Fundamental problem: we never observe both $Y_i(1)$ and $Y_i(0)$ for the same user.

Observed Outcomes and SUTVA

- Observed outcome:

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$$

- **SUTVA** (stable unit treatment value assumption):
  - no interference between users (one user's ad exposure doesn't change others' outcomes)
  - well-defined treatment (what exactly counts as "seeing the ad"?)

In advertising

SUTVA can be strained by spillovers (word-of-mouth), market-level effects, or auction dynamics.

Treatment Effects: ATE and CATE

- **ATE (average treatment effect):**

$$\tau = E[Y(1) - Y(0)]$$

- **CATE (conditional average treatment effect):**

$$\tau(x) = E[Y(1) - Y(0) \mid X = x]$$

- **Targeting** is about heterogeneity: who has large $\tau(x)$?

Why CATE is the natural object for personalization

It formalizes "uplift varies with user features" (purchase stage, price sensitivity, prior usage, etc.).

Identification: Randomized Experiments

If D is randomized (A/B test), then:

$$(Y(1), Y(0)) \perp D$$

and we can identify:

$$E[Y(1) - Y(0)] = E[Y \mid D = 1] - E[Y \mid D = 0]$$

In ads

Platforms routinely run experiments:

- holdout groups (no ads)
- randomized ad exposure / randomized bids / randomized creatives

Identification: Observational Data (Harder)

Without randomization we need assumptions, e.g. **unconfoundedness**:

$$(Y(1), Y(0)) \perp D \mid X$$

plus overlap:

$$0 < P(D = 1 \mid X) < 1$$

Why ads are tricky observationally

Users who get ads often differ systematically (intent, browsing, retargeting exposure), so confounding can be severe.

Prediction Metrics vs. Uplift Metrics

- Prediction cares about:
  - AUC / log-loss for click/purchase prediction
  - MSE for spending prediction
- Uplift cares about:
  - incremental conversions: $E[Y(1) - Y(0)]$
  - incremental profit / iROAS: incremental value minus ad cost

Key point

A model can be excellent at predicting Y and still be poor for targeting if it fails to predict $\tau(x)$.

Uplift Targeting as an Economic Decision

Let v_i be the value of a conversion (or expected margin) and c_i the cost of showing an ad.

Decision rule

Show the ad to i if

$$v_i \cdot E[Y(1) - Y(0) \mid X_i] > c_i$$

Interpretation: target when expected *incremental profit* is positive.

Where Uplift Matters in Practice

- **Retargeting:** high baseline purchase $\neq$ high incremental effect
- **Discount ads:** uplift must exceed margin loss
- **Ad fatigue:** some users can have *negative* uplift
- **Creative choice:** different messages produce different treatment effects
- **Bidding:** impressions are worth more if incremental value is higher

Transition

To personalize, we need methods for estimating heterogeneous treatment effects: CATE and causal forests.

CATE as "Personalized Uplift"

$$\tau(x) = E[Y(1) - Y(0) \mid X = x]$$

- X can include: prior purchases, time since last visit, device, location, browsing intensity
- CATE answers: **which user types are persuadable?**

Interpretation for targeting

Rank users by $\hat{\tau}(X_i)$, then target the top group (under a budget constraint).

Methods to Estimate Heterogeneous Effects

Classic idea: estimate $E[Y \mid D = 1, X]$ and $E[Y \mid D = 0, X]$ then difference.

- Meta-learners (T-learner, S-learner, X-learner)
- Doubly robust / orthogonalized learners (R-learner / AIPW-based)
- **Causal trees and causal forests** (tree-based CATE estimation)

Why trees/forests?

They capture nonlinearities and interactions and produce flexible heterogeneity patterns without manual specification.

Why Not Use Off-the-Shelf Prediction Trees?

- Prediction trees split to improve fit of $E[Y \mid X]$
- For treatment effects we need splits that maximize differences in *uplift*

Causal splitting intuition

Choose splits that create groups with different estimates of

$$E[Y(1) - Y(0) \mid X \in \text{leaf}]$$

not simply different levels of $E[Y \mid X]$.

Causal Forest: Intuition

- Like random forests, but targeted at treatment effects
- Build many "causal trees" on subsamples
- Each tree creates local neighborhoods; estimate treatment effect locally
- Average across trees for stability and accuracy

Output

A prediction $\hat{\tau}(X_i)$ for each user i (estimated CATE / uplift).

Honesty and Overfitting in Causal Trees

- In treatment-effect estimation, adaptive splitting can bias effect estimates
- **Honest trees** use sample splitting:
  - one subsample chooses splits
  - another subsample estimates effects within leaves

Why honesty matters

It reduces "finding heterogeneity that is just noise," improving out-of-sample performance for uplift.

Orthogonalization: Deconfounding + Better Learning

In observational (or stratified) settings we often estimate nuisance functions:

$$\mu(x) = E[Y \mid X = x], \qquad e(x) = P(D = 1 \mid X = x).$$

Orthogonalized signals (idea)

Use transformed outcomes that remove dependence on nuisance estimation error, so we can learn $\tau(x)$ more robustly (double machine learning logic).

Practical takeaway: cross-fitting + orthogonalization stabilizes causal ML in high dimensions.

Inference for CATEs (High Level)

- Modern causal forests can produce:
  - point estimates $\hat{\tau}(x)$
  - standard errors / confidence intervals (via forest-based variance estimates)
- But: interpreting many CATEs is hard and multiple testing is a concern

Econometric message

Causal ML is useful, but we still need disciplined evaluation: validation, uncertainty, and decision-based criteria.

Use Case: Causal Forest for Uplift Targeting

- Goal: estimate **who responds because of the treatment** (uplift), not who responds anyway
- We use a **causal forest** to estimate **heterogeneous treatment effects** (CATEs)
- Interpretation for ads:
  - **treatment** = showing an ad / sending a coupon / delivering a promo message
  - **outcome** = purchase / revenue / donation amount

Why this matters in personalized advertising

High baseline buyers are not always the best targets. The best targets are the **persuadables** (high uplift).

Setting: A Real Randomized Campaign (Ad Analogy)

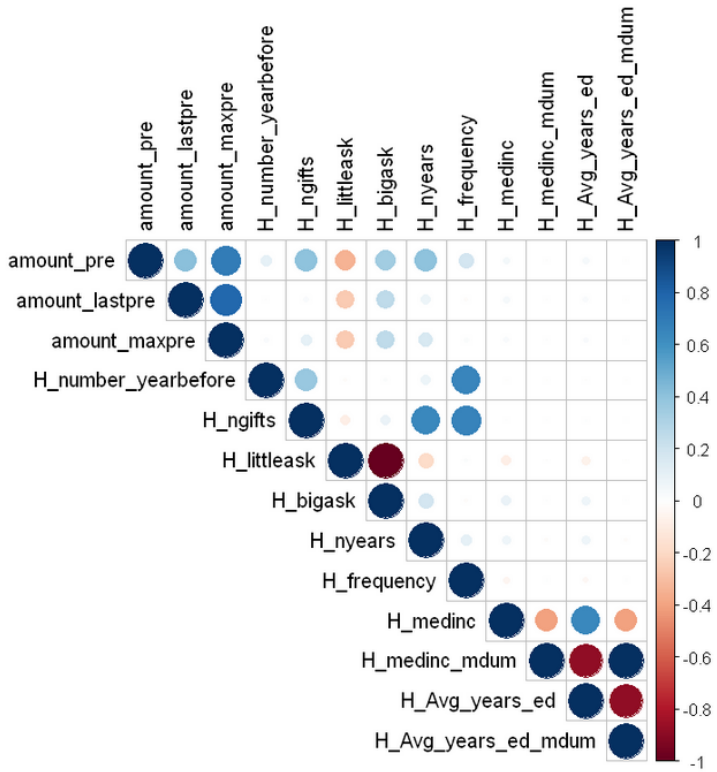
- Think of a campaign where individuals are randomly assigned to:
  - **Treatment** ($D = 1$): receive a "nudge" (e.g., gift/coupon/ad exposure)
  - **Control** ($D = 0$): no nudge
- Outcome Y: monetary response (e.g., donation amount / purchase value)
- Features X: rich user characteristics (history, demographics, context)

Personalized ads translation

A/B testing provides the variation in exposure needed to estimate **incremental effects** by user type.

Rich Covariates: High-Dimensional User Profiles

- We observe many predictors capturing:
  - prior behavior (past donations / past purchases)
  - recency and frequency measures
  - intensity/value measures (e.g., previous amounts)
- Many variables are correlated:
  - similar "behavioral signals" overlap
  - interactions can matter for uplift



Correlation structure among key features

Why ML helps: causal forests can handle many predictors and complex interactions when estimating heterogeneity.

Causal Forest: What It Estimates

- Potential outcomes:

$$Y_i(1),\ Y_i(0)$$

- Individual treatment effect (uplift):

$$\tau_i = Y_i(1) - Y_i(0)$$

- Conditional average treatment effect (CATE):

$$\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$$

Causal forest output

For each individual (or segment), we get an estimate $\widehat{\tau}(X_i)$: predicted **incremental impact** of the treatment.

What Drives Heterogeneity? (Split Frequencies)

- Causal forests search for splits that maximize **treatment effect heterogeneity**
- A simple summary is how often variables are used for splits

Feature (examples)	Split count
`amount_maxpre`	1986
`amount_pre`	1736
`amount_lastpre`	1557
`H_number_yearbefore`	1292
`H_frequency`	714

Ad interpretation: uplift depends strongly on **past value and past engagement** (common in ad response models too).

Average Effect (ATE): Often Not the Main Story

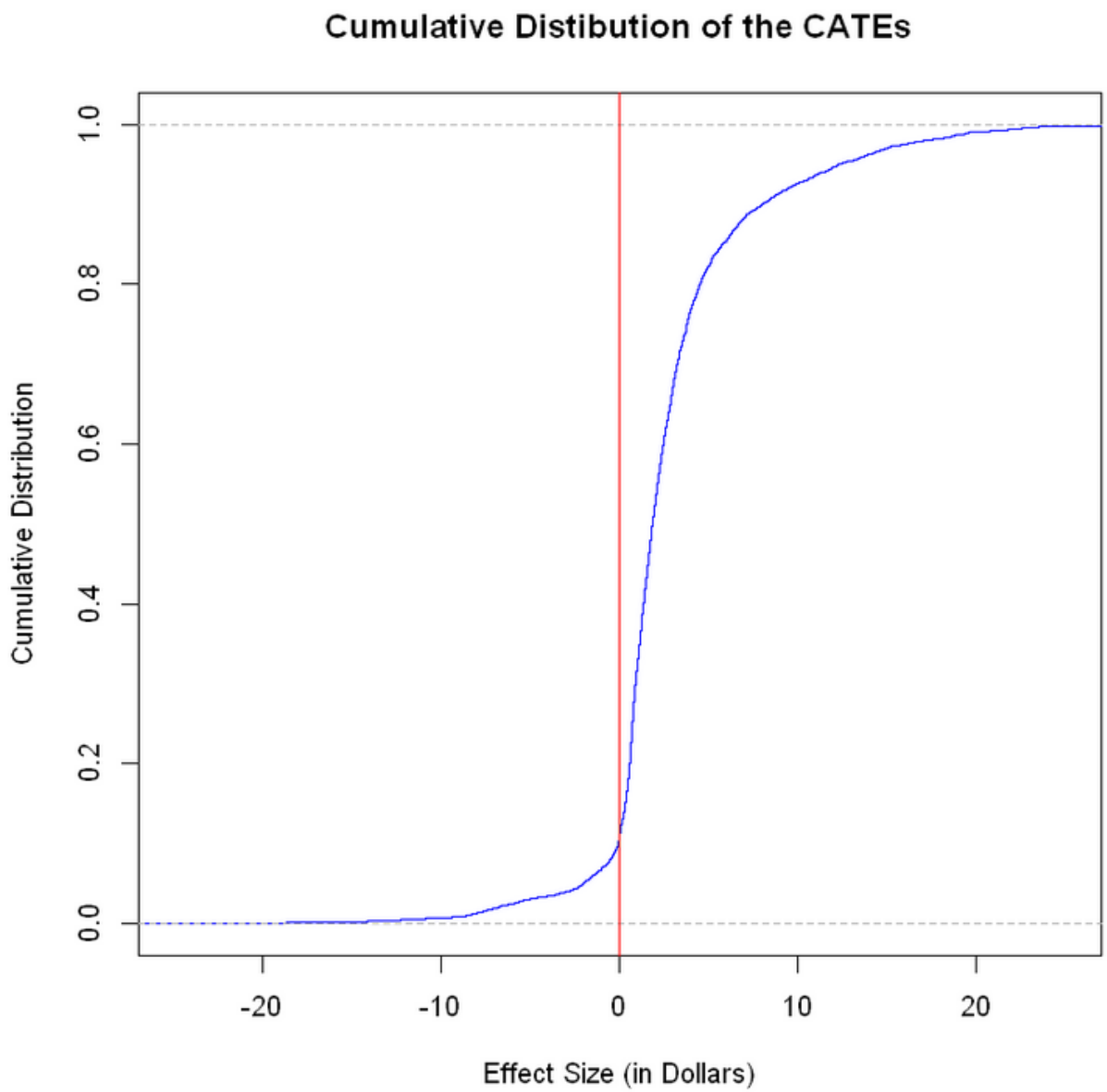
- Even with strong heterogeneity, the **average** effect can be small or imprecise
- In the causal forest output we obtain an ATE estimate:

Average treatment effect (from the estimation output)

$$\widehat{\text{ATE}} = 3.54 \quad \text{(s.e. } 4.32\text{)}, \quad p = 0.413$$

Key message: for targeting, we care less about the average and more about **who gains** (CATEs).

Distribution of Estimated CATEs (Uplift)



Estimated CATE distribution (uplift per person or individual)

Positive vs. Negative Uplift (Practical Implication)

- The estimation output reports:
  - **Number with positive effects:** 4099
  - **Number with negative effects:** 526
  - **Share positive:** 88.6%

Interpretation for ads

Some users can be **hurt** by exposure (ad fatigue, annoyance, reactance). A good policy avoids spending on them.

Caution: sign and magnitude are estimated with noise; policy evaluation should be out-of-sample / cross-fitted.

From CATEs to Targeting: A Profit Rule

- CATEs answer: *"What is the incremental effect for this user?"*
- A firm still needs a **decision rule** under costs and budgets

Simple profit-based targeting rule

Treat (show the ad / send the coupon) if:

$$\widehat{\tau}(X_i) \times \text{Value per response} > \text{Cost of treatment}$$

Ad example: show the ad if (uplift in purchase probability $\times$ margin) exceeds CPM/CPC cost.

How This Maps to Personalized Ads (Concrete)

- In digital ads, the platform often chooses among:
  - **who** gets an impression (targeting)
  - **how much** to bid (auction)
  - **which** message to show (creative)
- Uplift modeling improves all three by focusing on **incremental value**

Key takeaway

Prediction ranks users by expected outcomes.
Causal forests rank users by **incremental outcomes** (uplift) — closer to profit-maximizing ad allocation.

CATEs Are Effects — They Don't Directly Tell Us the Action

- $\hat{\tau}(x)$ estimates the effect of treatment vs no treatment
- A firm still needs a **policy**:
  - treat or not treat (binary)
  - which creative (multi-action)
  - how often (continuous)
  - how much to bid (continuous)

Decision layer

To choose actions, we must combine effects with constraints and costs (budgets, margins, fatigue).

Policy Learning: Learning Actions Directly

- A **policy** $\pi(x)$ maps features to an action

$$\pi(x) \in \mathcal{A}$$

- Examples in ads:
  - $\mathcal{A} = \{0, 1\}$: show ad vs no ad
  - $\mathcal{A} = \{\text{brand}, \text{discount}\}$: choose creative
  - $\mathcal{A} = \{0, 1, 2, 3\}$: frequency caps

Goal

Choose π to maximize expected objective (profit, donations, welfare), not just to estimate $\tau(x)$.

The Policy Value (Objective)

For binary treatment $D \in \{0, 1\}$ and policy $\pi(x) \in \{0, 1\}$:

$$V(\pi) = E\big[Y(\pi(X))\big]$$

With costs (ad cost, coupon cost), we often maximize:

$$V(\pi) = E\big[Y(\pi(X)) - c(X) \cdot \pi(X)\big]$$

Economic interpretation

This is the "choose who to treat" problem under costs and constraints.

Why Learn Policies Directly (Rather Than CATE Then Threshold)?

- Converting CATEs into actions requires extra steps:
  - choose a threshold
  - incorporate costs correctly
  - handle constraints (budget, limited capacity)
- A policy learner can be **targeted** to the objective:
  - maximize net donations
  - maximize profit
  - maximize iROAS under a budget

Takeaway

Good effect estimation does not automatically imply a good decision rule.

Evaluating a Policy Out-of-Sample (Off-Policy Evaluation)

- A policy is useful only if it performs well on new data
- We evaluate $V(\pi)$ using:
  - sample splitting / cross-fitting
  - doubly robust estimators (AIPW)

Bridge to application

Next: a real field experiment where the decision is whether to include a fundraising gift.

Application: Fundraising as a Targeting Problem

- Charities spend substantial shares on fundraising
- Fundraising instruments (gifts, matching, reminders) have heterogeneous effects
- The decision is a policy problem:
  - whom to contact
  - which incentive to include

Analogy to ads

A gift in fundraising is like an ad incentive: it can increase response but has a cost.

Fundraising Expenditures

- Charities spend between 5% and 25% on fundraising (Andreoni and Payne, 2011).
- Money spend on fundraising cannot be used to finance the actual charitable activity.
- Donors are averse to charities with high overhead costs (Tinkelman and Mankaney, 2005, Gneezy et al., 2014).

$\Rightarrow$ Efficient fundraising is crucial for charities!

Optimal Targeting

- Many different fundraising instruments have been proposed (e.g. matching grants, gifts).
- Due to heterogeneity in donors preferences (e.g., altruism, warm-glow), the effects of any fundraising instrument are likely to be heterogeneous across individuals.
- Optimal targeting exploits this effect heterogeneity with the purpose to maximise the net donations (= donations - costs).
- Feasible allocation rules for fundraising instruments are based on observable characteristics that proxy heterogeneous preferences.

Field Experiment with Gifts

- Field experiment with small unconditional gifts (Dürrer's flower postcards) accompanied by a solicitation letter ($N \approx 20'000$).

- Individuals in the randomly selected treatment group received a mailer with the gift and solicitation letter.
- Individuals in the randomly selected control group received the solicitation letter, but not the gift.

Potential Effects of Gifts

- In theory, gifts work through social preferences by triggering a reciprocal reaction (Benabou and Tirole, 2006, Dufenberg and Kirchsteiger, 2004).
- In line with this theory, Falk (2007) finds positive effects of gifts on donations.
- In contrast, Landry et al. (2010) and Yin et al. (2020) find that gifts can backfire and lower donations.
- Alpizar et al. (2008) find that gifts do not raise donations sufficiently high to justify the additional costs.
- Survey evidence suggests that 2/3 of donors do not want to receive gifts (Cygnus Applied Research, 2011).

$\Rightarrow$ Gifts appear to be a context with interesting heterogeneity!

Notation

- Treatment variable D_i (for $i = 1, ..., N$):

$$D_i = \begin{cases} 1 & \text{when a gift was sent, and} \\ -1 & \text{otherwise.} \end{cases}$$

- Potential outcomes:
  - $Y_i(1)$: Potential donations in response to the mailer with a fundraising gift.
  - $Y_i(-1)$: Potential donations in response to the mailer without a fundraising gift.
- Stable unit treatment value assumption (SUTVA):

$$Y_i = Y_i(-1) + \frac{1 + D_i}{2} \left(Y_i(1) - Y_i(-1) \right).$$

Treatment Effects

- Individual causal effects:

$$\delta_i = Y_i(1) - Y_i(-1)$$

- Average treatment effect (ATE):

$$\delta = E[\delta_i] = E[Y_i(1) - Y_i(-1)]$$

- Conditional average treatment effect (CATE):

$$\delta(x) = E[\delta_i | X_i = x] = E[Y_i(1) - Y_i(-1) | X_i = x]$$

Identifying Assumptions

- SUTVA
- Stratified randomisation with regard to observable characteristics Z_i:
  - CIA: $(Y_i(1), Y_i(-1)) \perp\!\!\!\perp D_i | Z_i = z$
  - Propensity score: $p(z, x) = Pr(D_i = 1 | Z_i = z, X_i = x) = Pr(D_i = 1 | Z_i = z) = p(z)$
  - Common support: $0 < p(z) < 1$

- Z_i are confounders that are relevant for identification.
- X_i are potentially relevant for effect heterogeneity.
- Z_i and X_i are not necessarily equivalent, but they may overlap.

Experimental Data

- Field experiment in cooperation with a fundraiser operating within the structure of the Catholic church in an urban area in Germany in 2014.
- All experimental participants received a letter with information about the fundraiser's cause (maintaining clergy houses, parish centers, and churches) and a donation request.
- A randomly selected treatment group additionally received a small unconditional gift.
- Attached to the letter is a bank transfer form pre-filled with the fundraiser's bank account information and the recipient's name.
- Donations are made exclusively via bank transfer, and the fundraiser does not provide any information about individual donations to the church parishes.

Heterogeneity Variables

- **Socio-economic characteristics:**
  Gender, age, marital status, years residency.
- **Donation history:**
  Number of previous donation, total previous donations, maximum previous donations, yearly donations of the previous 5 years.
- **Geo-spatial information of home address:**
  Number of restaurants, supermarkets, medical facilities, cultural facilities, and churches in the proximity (300 meters radius), distance to city hall, main station, main church, and airport, travel distance to main station, elevation.

Descriptive Statistics

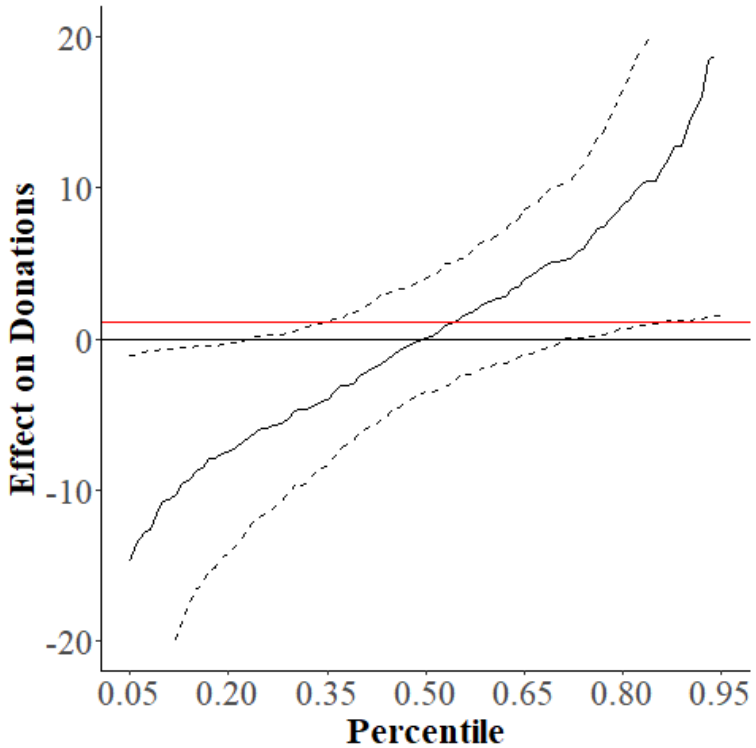
	Warm-list		Cold-list	
	Mean	Std. Dev.	Mean	Std. Dev.
	(1)	(2)	(3)	(4)
Socio-economic characteristics				
Female dummy	0.53		0.50	
Single dummy	0.50		0.64	
Widowed dummy	0.05		0.02	
Age (in years)	68.51	18.30	48.40	19.32
Duration residency in urban area (in years)	7.43	1.67	5.97	2.82
Donation history before the experiment				
Number of donations previous 8 years	3.97	2.83	0	
Max. donations previous 8 years (in Euro)	36.02	42.90	0	
Total donations previous 8 years (in Euro)	125.9	176.0	0	
Observations	2,354		17,425	

Average Effects

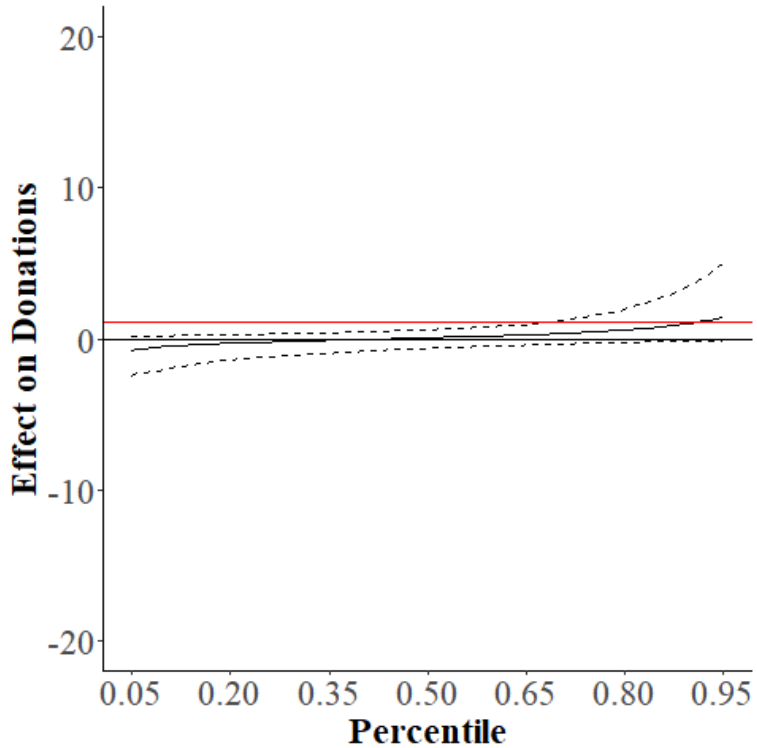
	Warm list	Cold list
	(1)	(4)
ATE	1.22	0.19*
	(1.15)	(0.10)
ATE - costs	0.06	-0.97***
	(1.15)	(0.10)
Strata controls	Yes	Yes
Observations	2'354	17'425

Notes: Outcome is donation amount (Euro) during the first year after the gift was sent.

Effect Heterogeneity



Warm list



Cold list

Notes: Figure based on the sorted effects model of Chernozhukov, Fernández-Val, and Luo (2018).

Targeting Rule

- A targeting rule $\pi(x) \in \{-1, 1\}$ allocates the gift based on observable X_i:
  - Individuals with $\pi(X_i) = 1$ receive the mailer with the gift
  - Individuals with $\pi(X_i) = -1$ receive the mailer without the gift

- Expected net donations (= expected donations - costs of gift),

$$P(\pi(X_i)) = E\left[Y_i(\pi(X_i)) - \frac{1 + \pi(X_i)}{2}c\right],$$

where c are the variable costs of the gift.

Benchmarks

- Everybody receives the gift:

$$P(\pi_1) = E[Y_i(1)] - c$$

- Nobody receives the gift:

$$P(\pi_{-1}) = E[Y_i(-1)]$$

- 50/50 randomization:

$$P(\pi_R) = \tfrac{1}{2}\left(E[Y_i(1) + Y_i(-1)] - c\right)$$

Value Added of Targeting Rule

- Compared to everybody receives the gift:

$$Q_1(\pi) = P(\pi) - P(\pi_1) = E\left[\frac{\pi(X_i) - 1}{2}(\delta_i - c)\right].$$

- Compared to nobody receives the gift:

$$Q_{-1}(\pi) = P(\pi) - P(\pi_{-1}) = E\left[\frac{1 + \pi(X_i)}{2}(\delta_i - c)\right].$$

- Compared to 50/50 randomization:

$$Q_R(\pi) = P(\pi) - P(\pi_R) = \frac{1}{2}E\left[\pi(X_i)(\delta_i - c)\right].$$

Augmented Inverse Probability Weighting (AIPW)

- δ_i is crucial for targeting but unobservable.
- Replace δ_i with an approximation score Γ_i.
- AIPW score (Robins et al., 1994; Chernozhukov et al., 2018):

$$\hat{\Gamma}_i = \hat{\Gamma}_i(1) - \hat{\Gamma}_i(-1)$$

with

$$\hat{\Gamma}_i(1) = \hat{\mu}_1(Z_i) + \frac{1 + D_i}{2} \cdot \frac{Y_i - \hat{\mu}_1(Z_i)}{\hat{p}(Z_i)},$$
$$\hat{\Gamma}_i(-1) = \hat{\mu}_{-1}(Z_i) - \frac{D_i - 1}{2} \cdot \frac{Y_i - \hat{\mu}_{-1}(Z_i)}{1 - \hat{p}(Z_i)}.$$

Augmented Inverse Probability Weighting (AIPW)

- AIPW identifies ATEs $\delta = E[\Gamma_i]$ and CATEs $\delta(x) = E[\Gamma_i | X_i = x]$.
- Chernozhukov et al. (2018): the ATE estimator

$$\hat{\delta} = \frac{1}{N} \sum_{i=1}^{N} \hat{\Gamma}_i$$

  is $\sqrt{N}$-consistent and semiparametrically efficient (with sufficiently fast nuisance estimation).
- Key practical advantage: allows fully flexible heterogeneity.

Estimation of the Optimal Targeting Rule

- Athey and Wager (2019): maximize sample analog of $Q_R(\pi)$:

$$\pi^* = \arg\max_{\pi} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \pi(X_i)(\hat{\Gamma}_i - c) \right\}.$$

- Equivalent weighted classification problem:

$$\pi^* = \arg\max_{\pi} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \pi(X_i) \operatorname{sign}(\hat{\Gamma}_i - c) \, |\hat{\Gamma}_i - c| \right\}.$$

Value Added of Machine Learning

- We could estimate π^* with a weighted logit (or any weighted classifier).
- But then we must choose X manually.
- Bias-variance trade-off:
  - too few features $\Rightarrow$ miss relevant heterogeneity
  - too many features $\Rightarrow$ overfit, poor out-of-sample policy value
- ML can balance this in a data-driven way.
- In the main specs we use optimal policy trees (Zhou et al., 2019).

Out-of-Sample Off-Policy Evaluation

- Once we have obtained π^*, estimate:

$$\hat{P}(\pi^*) = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{\Gamma}_i(\pi^*(X_i)) - \frac{1 + \pi^*(X_i)}{2} c \right),$$

$$\hat{Q}_R(\pi^*) = \frac{1}{2N} \sum_{i=1}^{N} \pi^*(X_i) \left(\hat{\Gamma}_i - c \right).$$

- Estimators are consistent and semiparametrically efficient (Chernozhukov et al., 2018).
- We apply cross-fitting to assess targeting rules out-of-sample.

Out-of-Sample Results for the Warm List

	Share Treated	Net Donations	Optimal Targeting Rule vs. Everybody	Nobody	Random
	(1)	(2)	(3)	(4)	(5)
Panel A: Results for Target Variable					
Net Donation Amount (1st year)	0.334	17.61*** (0.971)	2.141*** (0.817)	2.199*** (0.813)	2.170*** (0.575)
Panel B: Second Order Effects					
Net Donation Amount (1st and 2nd year)		32.94*** (1.661)	2.328* (1.405)	3.753*** (1.412)	3.040*** (0.995)
Donation Probability (1st year)		0.503*** (0.013)	0.007 (0.013)	0.025** (0.010)	0.016* (0.008)
Donation Probability (1st and 2nd year)		0.582*** (0.013)	0.001 (0.013)	0.017* (0.009)	0.009 (0.008)

Notes: Donation amounts in Euro. Standard errors in parentheses.

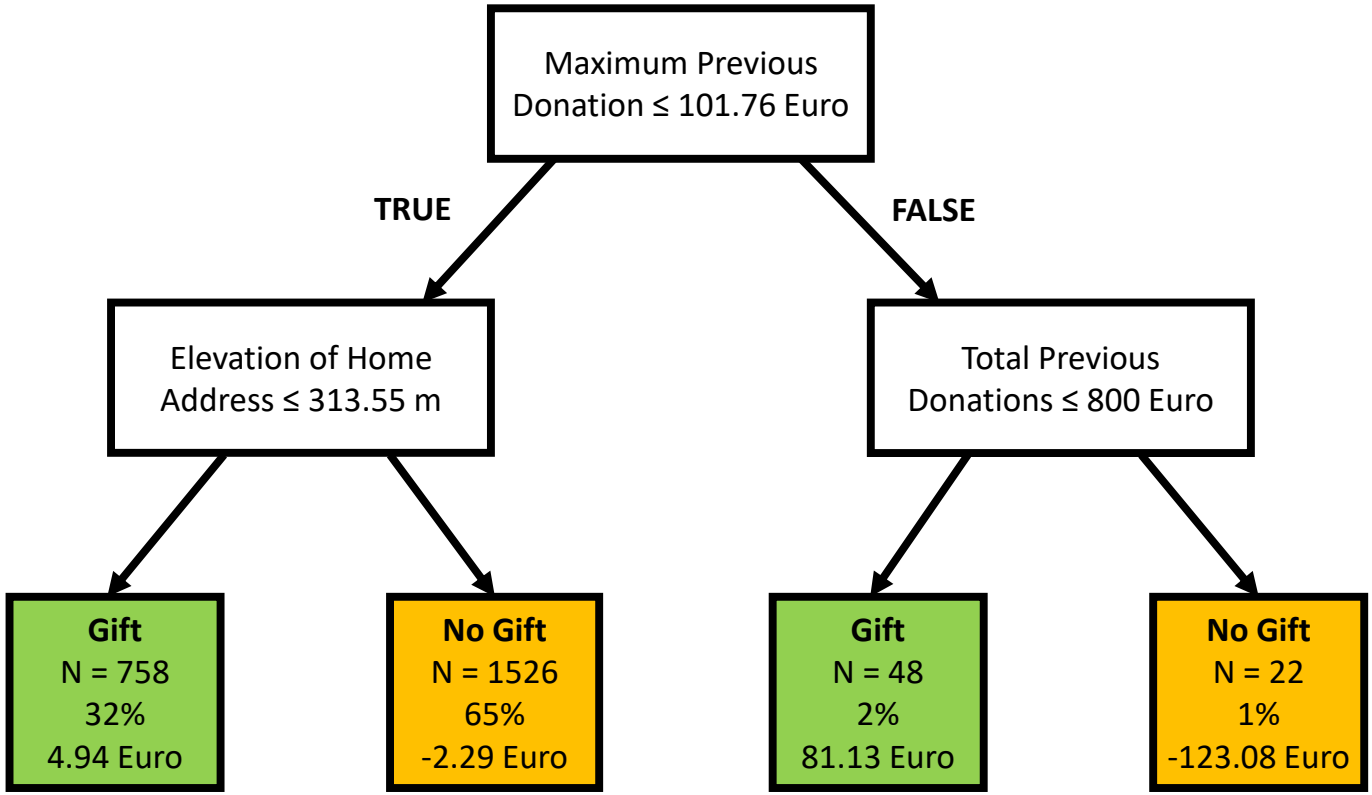
$\Rightarrow$ 14% increase in donations during 1st year.

Out-of-Sample Results for the Cold List

	Share Treated	Net Donations	Optimal Targeting vs. Everybody	Optimal Targeting vs. Nobody	Optimal Targeting vs. Random
	(1)	(2)	(3)	(4)	(5)
Panel A: Results for Target Variable					
Net Donation Amount (1st year)	0.014	0.15***	0.97***	-0.005	0.48***
		(0.02)	(0.10)	(0.012)	(0.05)
Panel B: Second Order Effects					
Net Donation Amount (1st and 2nd year)		0.44***	0.96***	0.04	0.50***
		(0.07)	(0.13)	(0.06)	(0.07)
Donation Probability (1st year)		0.009***	-0.007***	0.001	-0.003**
		(0.001)	(0.003)	(0.001)	(0.001)
Donation Probability (1st and 2nd year)		0.017***	-0.006*	0.001	-0.003
		(0.001)	(0.003)	(0.001)	(0.002)

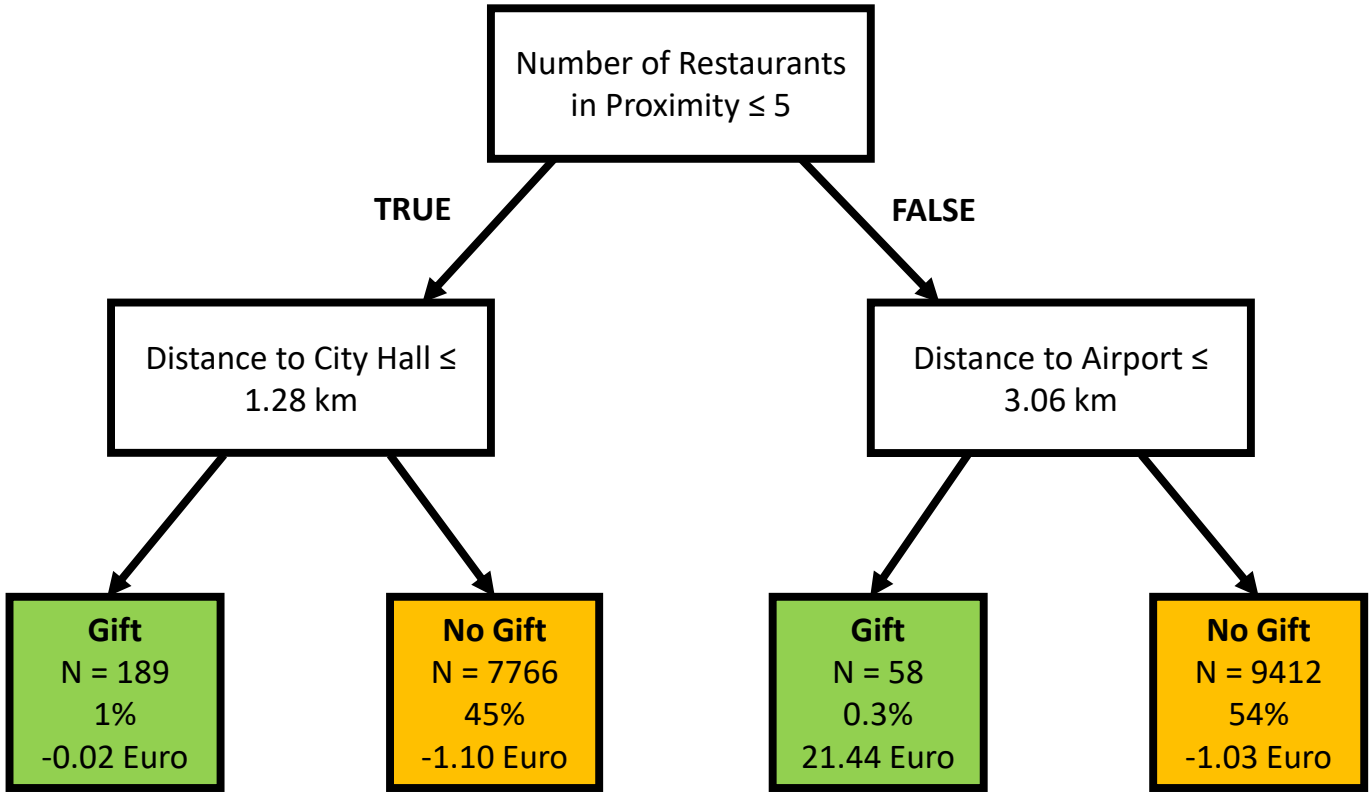
Notes: Donation amounts in Euro. Standard errors in parentheses.

Exact Policy Tree Warm List



Note: Figure based on the optimal policy tree of Zhou et al. (2019) and Sverdrup et al. (2020).

Exact Policy Tree Cold List



Note: Figure based on the optimal policy tree of Zhou et al. (2019) and Sverdrup et al. (2020).

Characteristics of Net Donors in the Warm List

	Individuals targeted by the algorithm				Std.
	Yes		No		Diff.
	Mean	Std. Dev.	Mean	Std. Dev.	
	(1)	(2)	(3)	(4)	(5)
Socio-economic characteristics					
Female dummy	0.507		0.539		6.459
Age (in years)	68.08	18.23	68.72	18.34	3.488
Donation history before the experiment					
Num. donations prev. 8 years	4.097	2.827	3.900	2.829	6.934
Max. don. prev. 8 years (in Euro)	39.94	44.61	34.05	41.89	13.63
Geospatial information (selected)					
Elevation (in meters)	308.66	6.266	321.38	9.524	157.80
Distance to main station (in km)	3.247	2.521	3.245	1.867	0.053
Observations	787		1'567		

(Slide shortened; keep full table in appendix if needed.)

Characteristics of Net Donors in the Cold List

	Individuals targeted by the algorithm				Std.
	Yes		No		Diff.
	Mean	Std. Dev.	Mean	Std. Dev.	
	(1)	(2)	(3)	(4)	(5)
Socio-economic characteristics					
Female dummy	0.558		0.503		11.04
Age (in years)	47.58	21.00	48.41	19.30	4.132
Geospatial information (selected)					
Distance to main station (in km)	1.995	0.890	2.874	2.028	56.14
Distance to airport (in km)	4.143	1.038	5.567	1.642	103.68
Observations	251		17'174		

(Slide shortened; keep full table in appendix if needed.)

Relevant Data Sources Warm List

Share Treated	Net Donations	Targeting Rule vs. Everybody	Nobody	Random	All Data Sources
(1)	(2)	(3)	(4)	(5)	(6)
Socio-Economic Characteristics					
0.55	15.71***	0.24	0.29	0.27	-1.91**
	(0.79)	(0.86)	(0.77)	(0.58)	(0.89)
Donation History					
0.12	17.20***	1.73**	1.79**	1.76***	-0.41
	(0.97)	(0.82)	(0.81)	(0.58)	(0.55)
Geo-Spatial Information					
0.49	17.40***	1.93**	1.98**	1.95***	-0.22
	(0.91)	(0.84)	(0.79)	(0.58)	(0.61)
Donation History and Geo-Spatial Information					
0.33	17.61***	2.14***	2.20***	2.17***	0
	(0.97)	(0.82)	(0.81)	(0.58)	

Notes: Donation amounts in Euro. Standard errors in parentheses.

Relevant Data Sources Cold List

Share Treated	Net Donations	Targeting Rule vs. Everybody	Nobody	Random	All Data Sources
(1)	(2)	(3)	(4)	(5)	(6)
Socio-Economic Characteristics					
0.015	0.14***	0.96***	-0.014**	0.47***	-0.01
	(0.02)	(0.10)	(0.007)	(0.05)	(0.014)
Geo-Spatial Information					
0.014	0.15***	0.97***	-0.005	0.48***	0
	(0.02)	(0.10)	(0.012)	(0.05)	

Notes: Donation amounts in Euro. Standard errors in parentheses.

Alternative Estimators Warm List

	Share Treated	Net Donations	Optimal Targeting vs. Everybody	Optimal Targeting vs. Nobody
	(1)	(2)	(3)	(4)
Logit (baseline)	0.47	16.19***	0.72	0.78
		(0.90)	(0.81)	(0.82)
Logit-Lasso	0.83	15.86***	0.39	0.45
		(0.91)	(0.62)	(0.98)
CART (CV depth)	0.33	17.40***	1.93**	1.98**
		(0.96)	(0.83)	(0.80)

Notes: Donation amounts in Euro. Standard errors in parentheses.

(Slide shortened; keep full table in appendix if needed.)

Alternative Estimators Cold List

	Share Treated	Net Donations	Optimal Targeting vs. Everybody	Optimal Targeting vs. Nobody
	(1)	(2)	(3)	(4)
Logit (baseline)	0.047	0.15***	0.96***	-0.01
		(0.04)	(0.10)	(0.03)
CART (CV depth)	0.0006	0.16***	0.97***	-0.001**
		(0.02)	(0.10)	(0.0003)

Notes: Donation amounts in Euro. Standard errors in parentheses.

(Slide shortened; keep full table in appendix if needed.)

Interpreting the Fundraising Results

- Warm list: the optimal targeting rule treats about one-third of donors and increases net donations
- Cold list: optimal rule treats a very small share (because gifts are costly and baseline response is tiny)
- Policy learning naturally incorporates:
  - treatment effect heterogeneity
  - costs of treatment
  - capacity/budget constraints (implicitly via objective)

Connection to advertising

This is structurally the same as: "Who should see the ad (or coupon) given heterogeneous uplift and per-impression cost?"

Ethical Concerns

- Statistical discrimination even if we omit critical variables (e.g., gender, migration, etc.)
- Examples: hiring decisions, flight prices, program assignments
- More or less discrimination than humans?
- Targeting rules can also reduce discrimination, but must be used appropriately
- Current scandals: Cambridge Analytica, Amazons' unethical hiring algorithm

Key Takeaways: Uplift and Policy Learning

- **Prediction** ranks users by expected outcomes $E[Y \mid X]$
- **Causal ML** estimates incremental effects $\tau(x) = E[Y(1) - Y(0) \mid X = x]$
- **Policy learning** chooses actions $\pi(x)$ to maximize an objective with costs/constraints
- In practice, policy learning can outperform "estimate CATE then threshold"
- Fundraising example: targeting gifts increases net donations by treating the right subset

Bridge

Next: how to implement these ideas for advertising targeting and auction bidding in real platforms.

Lecture Conclusions I: From Prediction to Decisions

- Modern advertising is fundamentally a **decision problem** under uncertainty:
  - limited budgets,
  - heterogeneous users,
  - costly interventions.
- **Prediction** answers:
  - Who is likely to click, buy, or spend?
- **Causal modeling (uplift)** answers:
  - Who changes behavior *because of* the ad?
- **Policy learning** goes one step further:
  - It directly learns *which action to take* for each user,
  - taking costs, constraints, and objectives seriously.

Core lesson

Good prediction is not enough. Effective personalization requires **causal**

Lecture Conclusions II: Economic Perspective

- Personalized advertising fits naturally into the economist's toolkit:
  - treatment effects,
  - heterogeneous responses,
  - welfare and profit maximization.
- Machine learning is most powerful when:
  - combined with economic structure,
  - grounded in causal identification,
  - evaluated by decision outcomes, not just predictive accuracy.
- The fundraising application illustrates a general insight:
  - optimal targeting can substantially improve outcomes,
  - even when average treatment effects are small or negative.

Final takeaway

Causal ML and policy learning transform rich data into **economically meaningful decisions**. This perspective extends far beyond advertising to public policy, health, labor markets, and beyond.