

Impact Evaluation in Firms and Organizations

With Applications in R and Python

Chapter 5: Causal Machine Learning

5.1 Motivating Causal Machine Learning

5.2 Elements of Causal Machine Learning

5.3 A Brief Introduction to Several Machine Learning Algorithms

5.4 Effect Heterogeneity and Optimal Policy Learning

Aim to compare similar groups

- Treatment and control groups must be comparable in observed characteristics X
- Analysts decide ex ante which characteristics (e.g., age or education) to include
- Selection of characteristics typically relies on domain-specific knowledge

Selection-on-observables assumption

- Included characteristics must capture all factors affecting both intervention and outcome
- Covariate choice is guided by the aim of satisfying this assumption
- Often based on intuition, theory, or findings from previous studies

Example

Evaluating the impact of discounts on sales

- The company observes customer characteristics such as demographics and purchase history
- These characteristics may jointly affect discount receipt and sales per customer
- It is assumed that no unobserved factors jointly affect discount receipt and sales given these covariates
- Internal rules may reveal which characteristics influence the likelihood of receiving a discount
- Such domain knowledge helps identify relevant covariates for making treatment and control groups comparable

Challenges in Selecting Characteristics

Characteristic selection is not straightforward

- It is often unclear which characteristics are sufficient for selection on observables
- Challenge intensifies when many characteristics are observed

Big data contexts

- Digitization generates increasingly rich sets of consumer characteristics
- Many characteristics may be irrelevant for causal identification
- Relevant subset is often unknown prior to analysis

Learning relevant subsets of covariates

- Data-driven methods can learn relevant characteristics
- Objective is to identify subsets that make treatment and control groups comparable

Ad hoc characteristic selection

- Analysts often lack clear guidance on which characteristics to include
- Common practice is to try different covariate sets and choose the most favorable one
- Conventional impact evaluation methods are not robust to such covariate snooping
- Conclusions may be unreliable when based on ad hoc selection

Take away

- Pre-testing problem: Due to the analyst's preselection of characteristics, effect estimates and inference may be biased
- Methods that learn characteristics without invalidating inference are desirable

Causal Machine Learning (CML) approach

- CML is distinct from purely predictive machine learning
- Selects relevant characteristics in a data-driven way
- Aims to ensure comparability between treatment and control groups

Fundamental assumptions still required

- CML cannot overcome basic identification requirements such as selection on observables

Potential limitations

- Assumption fails if important confounders remain unobserved
- For example, personality traits affecting both discounts and purchasing behavior

Applicability of CML

- Useful when relevant characteristics are not known in advance
- Learns the subset of observed characteristics needed to ensure comparability

Requirements for CML

- Number of truly relevant characteristics must be limited
- Observed characteristics must adequately capture key drivers of treatment and outcome

Performance of CML

- Does not compromise performance when the relevant covariates are sparse
- Remains competitive despite learning characteristics from the data

Example

Use CML to evaluate train ticket discounts in Switzerland

- Outcome of interest is demand shifts, such as rescheduling trips
- Many factors may jointly affect discounts and travel behavior

Motivation for CML

- Large number of potentially relevant demand-related characteristics
- CML selects important characteristics in a data-driven manner

Heterogeneous effects

- Interventions may affect population segments differently, with some groups responding more strongly than others
- CML allows analysis of differential impacts across subpopulations
- Subpopulations are defined by observed characteristics X

Use of heterogeneous effects

- Insights support improved targeting of interventions
- Enables tailoring strategies to specific population segments

Example: Heterogeneity Analysis of Coupons

Example

Analyze heterogeneous effects of a coupon campaign

- Study effect heterogeneity across customer segments using CML
- Segments defined by age, income, and pre-campaign expenditures

Findings

- Coupon effects vary with past purchasing behavior
- Strongest effects among customers with very low or very high prior purchases (see Figure 5.1)

Effect Heterogeneity Across Past Purchases

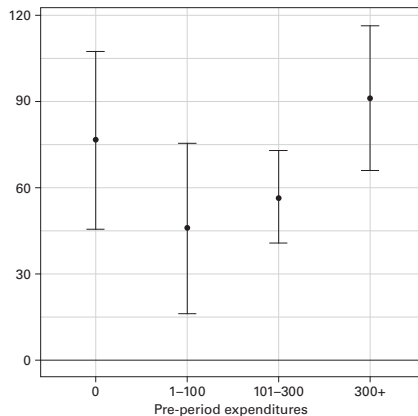


Figure 5.1: Effect heterogeneity across past purchases

Heterogeneous effects for optimal targeting

- CML can be used to derive optimal targeting strategies
- Determines which segments should receive an intervention
- Can incorporate costs and benefits to maximize net gains
- Enables more efficient allocation of limited resources

Example

Apply optimal policy learning to coupon campaigns

- Objective is to maximize overall customer purchases
- Optimal targeting varies with income, pre-campaign expenditures, and family size

5.1 Motivating Causal Machine Learning

5.2 Elements of Causal Machine Learning

5.3 A Brief Introduction to Several Machine Learning Algorithms

5.4 Effect Heterogeneity and Optimal Policy Learning

Machine learning (ML) components

- CML uses ML to estimate statistical models
- Models include the propensity score $\Pr(D = 1 | X)$ and conditional mean outcomes $E[Y | D = 1, X]$, $E[Y | D = 0, X]$

Interpretation of models

- Propensity score: likelihood of receiving the intervention given X
- Conditional means: expected outcomes for given X and treatment status

Learning important characteristics

- ML identifies important characteristics in X that strongly influence both treatment assignment and the outcome

Double machine learning

- ML models for treatment and outcome enable impact estimation
- Effects can be estimated for all units or treated units only
- Estimation relies on doubly robust (DR) expressions

Error tolerance of DR expressions

- Allow limited errors in treatment and outcome models
- ML models need not be perfectly specified to perform well

Conditions for good performance

- Number of relevant characteristics must be limited
- Sample size must be large enough for reliable learning

Comparison to other methods

- Matching, IPW, and regression are more sensitive to model errors
- CML therefore typically relies on doubly robust estimators

Dividing the data

- CML performs different tasks on different parts of the data
- Dataset is randomly split into two subsamples

Estimation procedure

- In the first subsample, machine learning is used to learn the treatment and outcome models
- In the second subsample, the learned models (from the first subsample) are used in doubly robust expressions to estimate the impact
- Tasks are then swapped across subsamples, and treatment effects averaged over the two subsamples

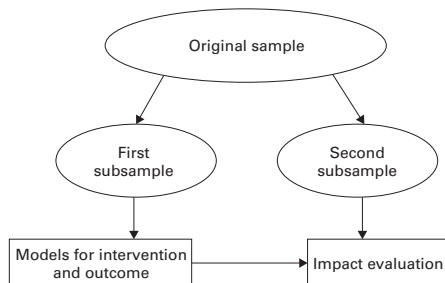
Overfitting

- Overfitting may occur when the same data are used both to learn important characteristics and to estimate the intervention effect
- This can increase variability and reduce stability of impact estimates across datasets
- Sample splitting helps mitigate overfitting

Cross-fitting procedure

- Perform the two tasks in different subsamples
- Swap the tasks across subsamples
- Final estimate averages the impacts across subsamples

Cross-fitting for Impact Evaluation



- In a second step, the roles of the subsamples are swapped: intervention/outcome models obtained in second subsample and impact evaluated in first subsample.
- In a third step, the impacts obtained in both subsamples are averaged.

Figure 5.2: Cross-fitting for impact evaluation

5.1 Motivating Causal Machine Learning

5.2 Elements of Causal Machine Learning

5.3 A Brief Introduction to Several Machine Learning Algorithms

5.4 Effect Heterogeneity and Optimal Policy Learning

Available ML techniques

- Many ML methods can estimate intervention and outcome models, e.g., decision trees
- Identify covariates influencing the intervention or outcome
- These methods serve as ingredients in CML procedures

Decision trees

- Recursively split the dataset based on values of characteristics X
- Splits aim to create subsets with similar intervention or outcome values

Splitting by characteristics

- Algorithm splits data when intervention values differ across groups, e.g., older customers often receive discounts, younger ones rarely do
- Such splits reveal characteristics influencing treatment receipt

Recursive refinement

- Splits are applied repeatedly to form finer subsets
- Uses characteristics such as age, education, or gender
- Subsets become increasingly homogeneous in intervention values

Decision tree

- Splitting rules are displayed as a tree structure
- Tree structure highlights important characteristics

Tree growth and stopping rules

- Data is repeatedly split to grow a decision tree
- Stopping rules may limit tree depth or minimum leaf size
- Optimal tree size can be chosen in a data-driven way via validation

Avoiding overfitting

- Validation separates tree growing and performance assessment
- Closely related to cross-fitting discussed earlier
- Helps avoid overfitting when modeling the intervention

Using tree leaves

- Final subsets formed by the tree are called leaves
- Observed intervention values within a leaf inform predictions
- Subjects with similar characteristics receive similar predictions

Decision Tree

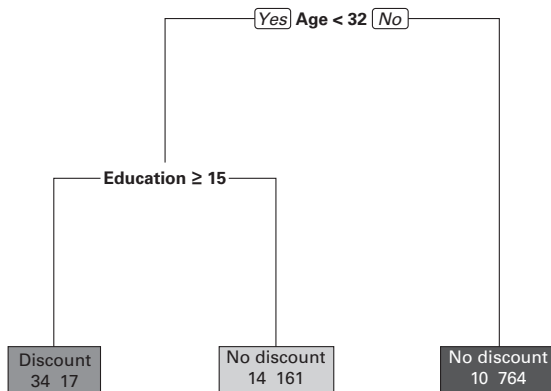


Figure 5.3: Decision tree

Structure of the tree in Figure 5.3

- Tree predicts discount receipt based on age and education
- First split at age 32: younger customers are more likely to receive discounts
- Second split among younger customers at 15 years of education

Leaf predictions

- Each leaf contains a predicted intervention outcome
- Leaves also report numbers of recipients and non-recipients
- Majority rule determines the predicted outcome

Limitations of single decision trees

- Individual decision trees tend to have high variance
- Small changes in the data can lead to very different trees
- As a result, predictive accuracy of a single tree is often limited

Averaging multiple trees

- Trees are grown on many random subsamples of the data
- Process is repeated many times, yielding many different trees
- Predictions are aggregated by averaging across trees

Bagging and random forests

- Bagging averages trees grown on random subsamples of observations
- Random forests additionally randomize the set of characteristics used for splitting
- Both reduce variance and improve predictive accuracy

Partial inclusion of characteristics

- Some methods can shrink the influence of characteristics instead of treating inclusion as an all-or-nothing decision
- Often improves predictive accuracy

Lasso and ridge regression

- Penalize model complexity
- Allow characteristics to have varying influence

Choosing optimal influence

- Cross-validation selects tuning parameters
- Aims to maximize predictive accuracy

Neural network structure

- Neural networks consist of interconnected regression models
- Covariates serve as inputs to nodes
- More nodes increase model flexibility and complexity

Model complexity

- Number of layers and nodes affects prediction accuracy
- Deep learning refers to networks with many layers
- Widely used in image and language applications

Choosing network size

- Validation helps determine optimal architecture
- Goal is accurate prediction of intervention or outcome

Neural Network for Treatment Prediction

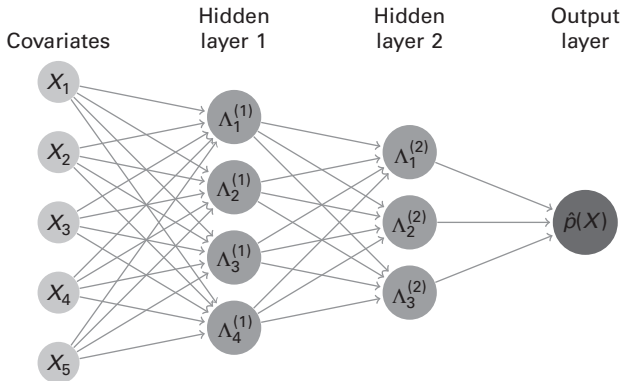


Figure 5.4: A neural network for treatment prediction

Combining algorithms

- Ensemble methods average predictions from several machine learning algorithms
- Aim to improve predictive accuracy over individual algorithms

Performance gains

- Averaging can outperform each single algorithm
- Combines strengths of multiple predictive models
- Provides a more robust overall prediction

Choosing weights

- Cross-validation determines optimal weights for each algorithm
- Weights chosen to maximize predictive accuracy
- Ensures best-performing ensemble for intervention prediction

5.1 Motivating Causal Machine Learning

5.2 Elements of Causal Machine Learning

5.3 A Brief Introduction to Several Machine Learning Algorithms

5.4 Effect Heterogeneity and Optimal Policy Learning

Heterogeneity in responses

- Individuals may respond differently to the same intervention
- Treatment effects can therefore vary across customers
- The ATE averages over these heterogeneous individual effects

Example

Heterogeneous responses to discount promotion

- Discounts may increase spending for some customers
- Cannibalization can reduce spending for others
- A third group may remain unaffected

Average and Heterogeneous Effects

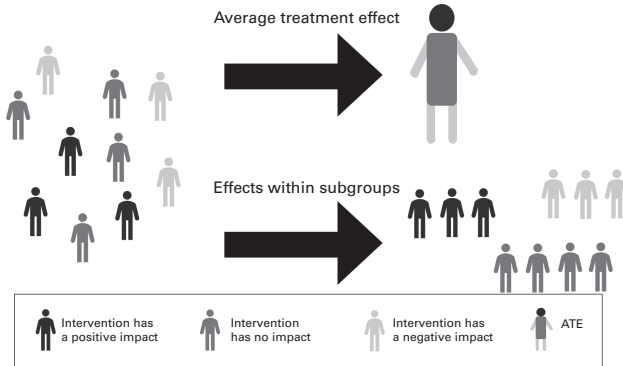


Figure 5.5: Average and heterogeneous effects

Concept of heterogeneity

- CML allows assessment of how effects vary across subgroups
- Subgroups are defined by observed characteristics X

Relevance in practice

- Treatment effects often differ across customer segments
- Segments may be defined by age, gender, education, or income
- Identifying groups with particularly large effects is valuable

Implications for targeting

- Enables more effective and cost-efficient interventions
- Supports targeting customers with high expected impact
- Helps maximize overall gains from the intervention

Causal tree approach

- Applicable to A/B tests and under the selection-on-observables assumption
- Uses tree structure to detect heterogeneous causal effects

Data-splitting procedure

- One part derives splitting structure maximizing effect heterogeneity
- Another part estimates treatment effects within subsets
- Performing different tasks in different subsamples prevents overfitting bias and provides more honest assessments of the treatment effect

Tree-based heterogeneity detection

- Modified decision trees predict causal effects, not outcomes
- Dataset is recursively split using characteristics X
- Splits aim to form subgroups with homogeneous effects
- Enables data-driven identification of responsive customer segments

Tree structure

- Tree defines subgroups with maximally different effects
- Number of splits is predefined
- Structure reveals where interventions are more or less effective and provides interpretable subgroup definitions

Causal Tree

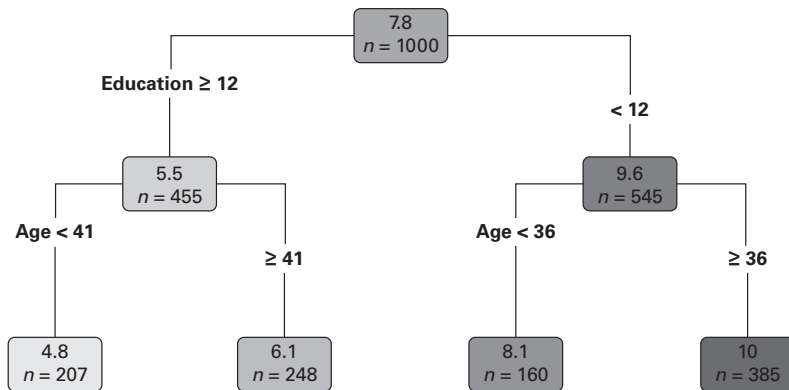


Figure 5.6: Causal tree

Structure of the causal tree

- Root node shows ATE: \$7.8 increase in daily sales
- Total sample consists of 1,000 customers
- First split based on education (12 years threshold)

Segment-specific impacts

- Impact is \$5.5 for higher educated customers
- Impact is \$9.6 for less educated customers
- Both groups are further split by age

Leaf-level heterogeneity

- Four final leaves represent customer segments
- Impacts range from \$4.8 to \$10

Further methods for assessing effect heterogeneity

- Causal forest: Averages over many causal trees (applied to many random subsamples of the data)
- Other ML methods such as lasso regression can be adapted so that they detect effect heterogeneity

Conditional average treatment effects (CATEs)

- Represent the average effect for subjects with specific values of characteristics X , rather than for the entire population
- Allow investigation of how effects change across values of X
- CATEs are defined as

$$\Delta_x = E[Y(1) - Y(0) \mid X = x] \quad (5.1)$$

Conditional Average Treatment Effects (CATEs)

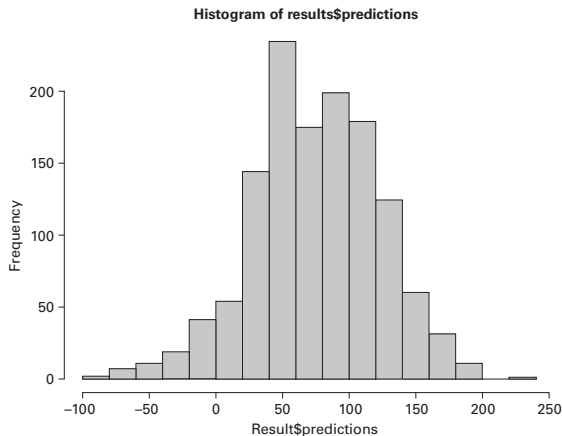


Figure 5.8: CATEs

Tailoring interventions

- Individualized effects enable targeted interventions
- Goal is to allocate costly interventions efficiently

Decision-making considerations

- Targeting decisions account for intervention costs
- Costs and benefits may vary across subgroups

Defining customer segments

- Segments defined using observed characteristics X
- Enables fine-tuned pricing or marketing strategies

Policy trees

- Policy trees learn both segmentation and intervention assignment
- Only the number of segments must be preset
- Yields optimal intervention rules for different subgroups
- Can be described by the tree structure of how the data are split

Algorithm

- Splits data into subgroups based on observed characteristics
- Assigns interventions across subgroups to maximize overall gains

Policy Tree for a Marketing Intervention

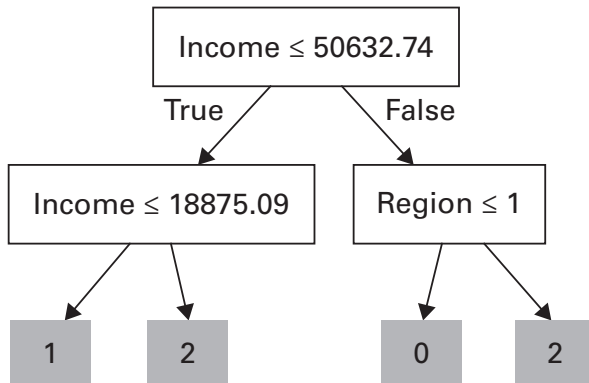


Figure 5.7: Policy tree for a marketing intervention

Set-up

- Intervention levels: no discount (0), small discount (1), large discount (2)
- Objective is to maximize net benefits per customer
- Four customer segments are learned from the data

Optimal allocation according to the policy tree

- Annual income and region are key determinants
- Lower-income customers should receive a small discount, while middle-income customers should receive a large discount
- Among high-income customers, those in region 1 should receive no discount, while those in other regions should receive a large discount

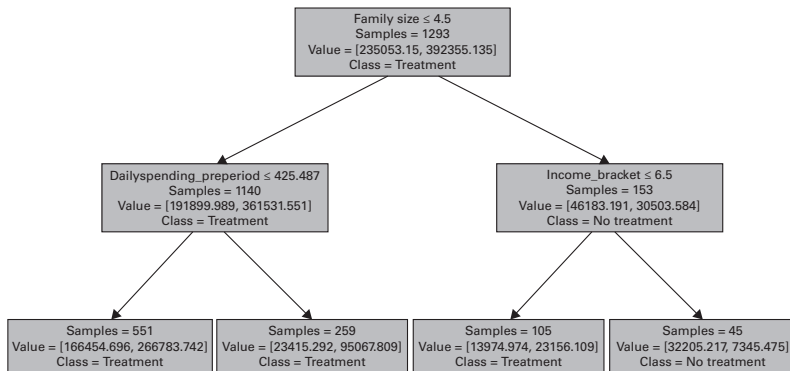


Figure 5.10: Policy tree in Python