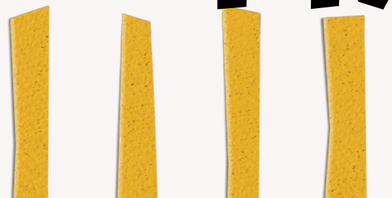
AUGENBLICK 2025 - CAUSAL INFERENCE CHALLENGE

Understanding Treatment Effects in Healthcare

-PIXELPAIR



CAUSAL INFERENCE TECHNIQUES

Propensity Score Matching (PSM)

Estimates treatment effects by matching treated and control units with similar propensity scores, ensuring comparability.

Inverse Probability Weighting (IPW):

Adjusts for confounding by assigning weights to observations based on their probability of receiving treatment.

Double Machine Learning (DML):

Uses machine learning models to flexibly control for confounders while estimating treatment effects.

RandomForestRegressor:

Predicts factual and counterfactual outcomes by training separate models for treated and control groups.

ESTIMATING TREATMENT EFFECTS

Average Treatment Effect (ATE):
Measures the overall impact of treatment by comparing the mean outcomes of treated and control groups. It provides a population-level estimate of effectiveness.

$$ATE = E[Y(1)] - E[Y(0)]$$

Individual Treatment Effect (ITE):
Estimates the treatment effect for each individual by comparing their predicted outcomes with and without treatment.

$$ITE_i = Y_i(1) - Y_i(0)$$

Conditional Average Treatment Effect (CATE):
Evaluates how treatment effects vary across different subgroups based on specific characteristics. It helps identify which groups benefit more or less from the intervention.

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

OBJECTIVES

Objective:

This project analyzes treatment effects in healthcare using causal inference to estimate true intervention impacts while addressing confounding biases.

Dataset:

The Infant Health and Development Program (IHDP) dataset is commonly used in causal inference research. It simulates real-world treatment effects and provides counterfactual outcomes, which are usually unknown in real-world data.

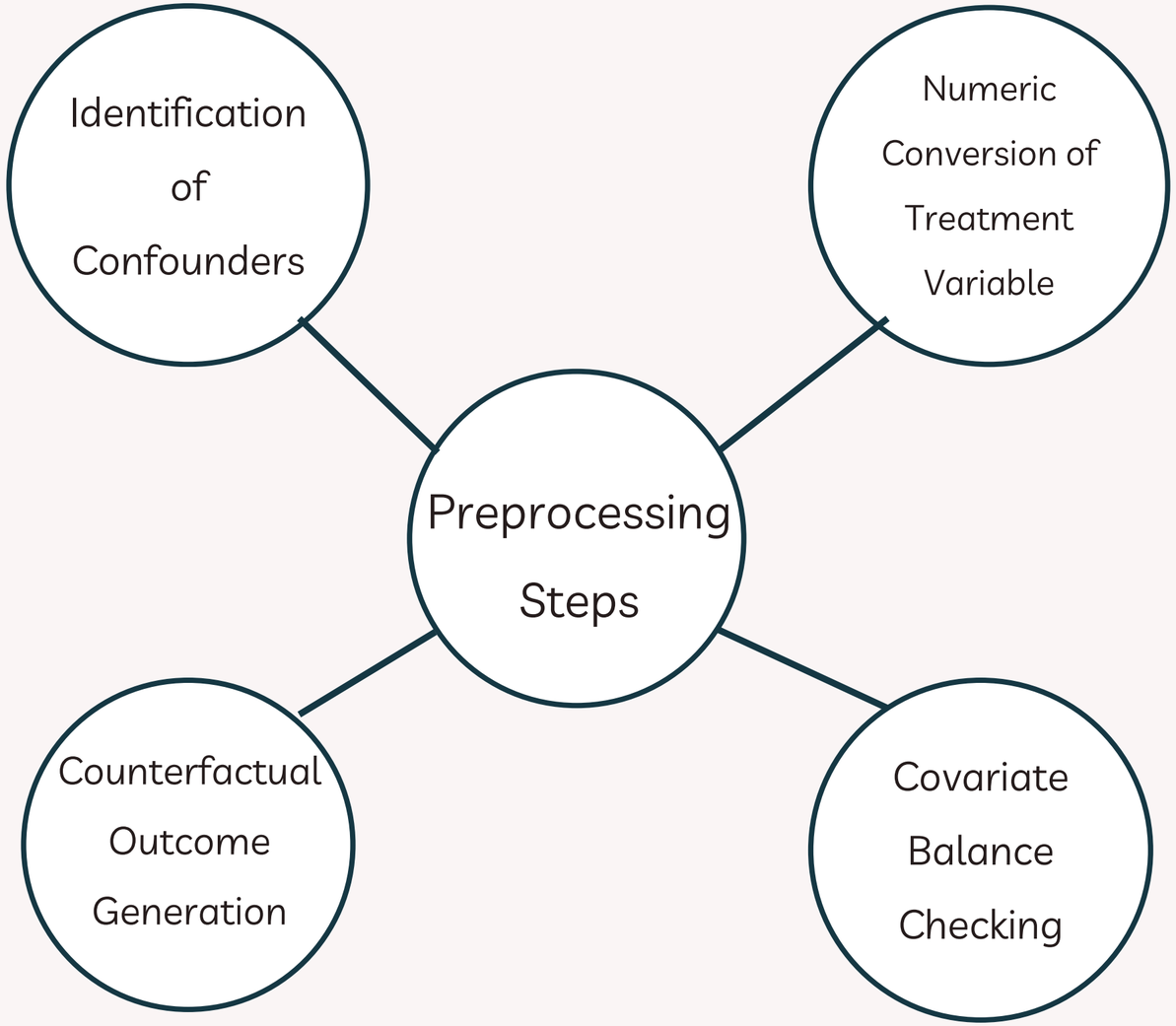
treatment	1 = received childcare, 0 = did not
y_factual	Observed cognitive score at age 3
X1 - X25	25 covariates (infant & mother details: birth conditions, socio-economic status, etc.)
y_cf	Counterfactual score (what would have happened with different treatment)
mu0, mu1	True potential outcomes (hidden ground truth for benchmarking)

METHODOLOGY :

The study utilizes the Infant Health and Development Program (IHDP) dataset, a widely used benchmark for causal inference. In cases of missing data, synthetic data generation techniques serve as a fallback. The dataset includes treatment assignment, factual and counterfactual outcomes, covariates such as socioeconomic indicators, and health-related features.

Treated Group	608
Controlled Group	608

After Random Sampling



MODEL AND VALIDATION

- **Feature Selection:**

The model leverages key socioeconomic and health-related features such as income, birth weight, education, health index, housing quality, and neighborhood safety to predict treatment effects.

- **Training Pipeline:**

Normalization: StandardScaler is applied to ensure numerical stability.

Separate Models: Two Random Forest Regressor models are trained independently for the treated (1) and control (0) groups to estimate causal effects.

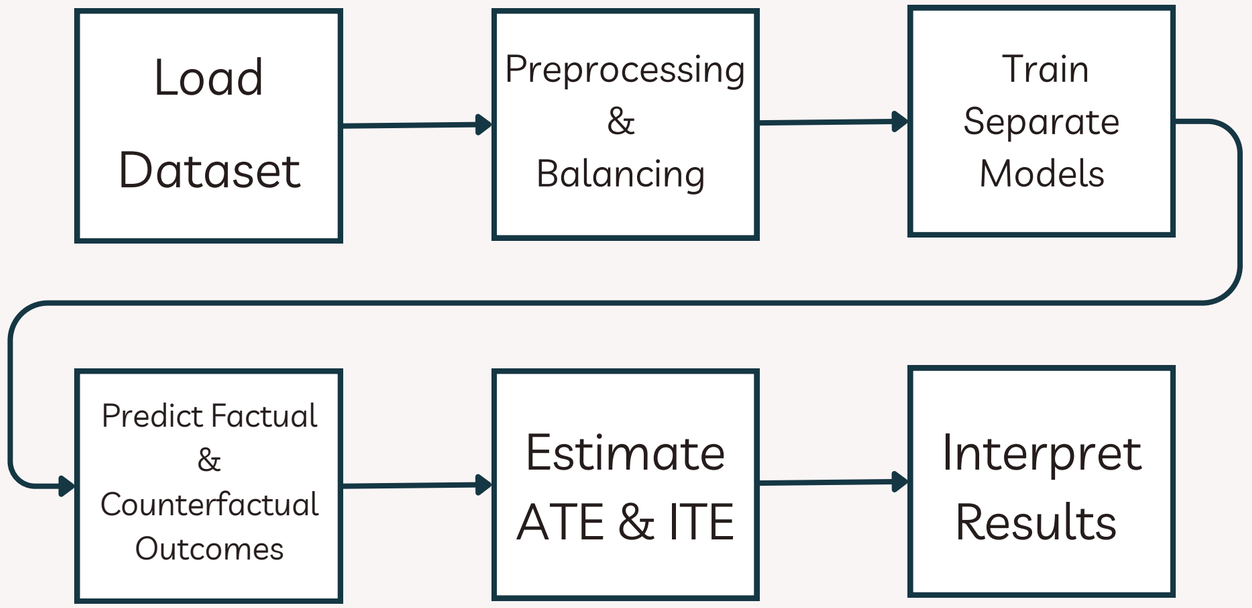
Model Performance: Evaluated using key metrics:

  - R^2 Score (Goodness of fit)
  - Mean Squared Error (MSE) (Prediction accuracy)

- **Validation Approach:**

The model's predictions are validated against synthetic counterfactual outcomes to assess reliability.

Ensures that the estimated treatment effects align with the expected causal relationships.



KEY FINDINGS & INSIGHTS

Treatment Effects: Specialized childcare positively impacts cognitive test scores, with varying effects across subgroups like income and education.

Confounder Balancing: Some confounders (e.g., parental education, income) show imbalance, requiring advanced adjustments to reduce bias.

Model Performance: Random Forest models show strong predictive accuracy with high R^2 and low MSE, aligning with expected causal relationships.

Heterogeneous Effects: Treatment benefits vary across individuals, highlighting the need for personalized intervention strategies in healthcare.

CONCLUSION

In this project, we developed a causal inference analysis tool using Streamlit to estimate treatment effects from observational data. Our approach integrates Random Forest regression models to predict outcomes under both treated and untreated conditions, enabling us to estimate the Average Treatment Effect (ATE) and Individual Treatment Effect (ITE) for any given input.

Key Features & Insights:

✓ *Data Handling & Preprocessing:* We implemented functionality to load real datasets and generate synthetic data when needed, ensuring flexibility in testing.

✓ *Causal Inference Modeling:* Two separate Random Forest models were trained for treated and control groups to predict factual and counterfactual outcomes.

✓ *Model Evaluation:* We assessed model performance using R^2 score, Mean Squared Error (MSE), and feature importance analysis to ensure reliability.

✓ *Covariate Balance Checking:* The system checks whether confounders are balanced between treatment and control groups to validate causal assumptions.

✓ *User Interaction & Visualization:* The web interface allows users to adjust input values, analyze individual treatment effects, and visualize the results through bar plots and distributions.

THANK YOU!