# Social Pressure and Voting

Capstone Sprint 3

# Agenda

Project Overview

Initial Steps

‘Pure’ ML Modelling

ML + Causal Inference

Model Comparison

Next Steps

# #1
# Project Overview

How might we use **machine learning** to understand the role of **social pressure** in influencing **voting behavior**?

1.

Does social pressure impact voting rates?

2.

If so, by how much?

3.

For whom?

A secondary goal:

# Combine Machine Learning and Causal Inference Methods

# Data Overview

## Treatment

1 of 4 treatments, ascending in level of social pressure

## Target

Voted in 2006 Michigan Primary Election or not

## Features

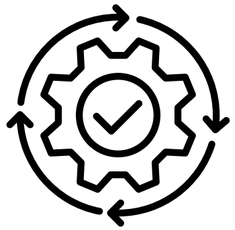
Demographic features (age, gender, ethnicity, employment)

## Levels of granularity

Individual and ZIP-code level

# #2
# Initial Steps

# Initial Steps

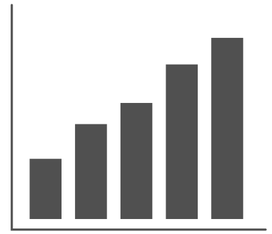


### Pre-processing

One-hot encoding treatment column

### Feature Engineering

Dropping redundant, collinear and low variance columns



### EDA

Examining the distribution of features and clustering

# #3
# ‘Pure’ ML Modelling

# Model Overview

| MODEL NAME | ACCURACY | PRECISION | RECALL | F1 |
|---|---|---|---|---|
| **LogReg: SMOTE** | 0.57 | 0.37 | 0.50 | 0.43 |
| **Tuned Decision Tree** | 0.70 | 0.57 | 0.21 | 0.31 |
| **Balanced Random Forest** | 0.60 | 0.42 | 0.68 | 0.52 |
| **Gradient Boosting Classifier** | 0.71 | 0.59 | 0.39 | 0.39 |
| **Dense Neural Network** | 0.71 | 0.60 | 0.27 | 0.36 |

# Model Overview

| MODEL NAME | ACCURACY | PRECISION | RECALL | F1 |
|---|---|---|---|---|
| **LogReg: SMOTE** | **0.57** | 0.37 | 0.50 | 0.43 |
| **Tuned Decision Tree** | **0.70** | 0.57 | 0.21 | 0.31 |
| **Balanced Random Forest** | **0.60** | 0.42 | 0.68 | 0.52 |
| **Gradient Boosting Classifier** | **0.71** | 0.59 | 0.39 | 0.39 |
| **Dense Neural Network** | **0.71** | 0.60 | 0.27 | 0.36 |

# Model Overview

| MODEL NAME | ACCURACY | PRECISION | RECALL | F1 |
|---|---|---|---|---|
| **LogReg: SMOTE** | **0.57** | 0.37 | 0.50 | **0.43** |
| **Tuned Decision Tree** | **0.70** | 0.57 | 0.21 | **0.31** |
| **Balanced Random Forest** | **0.60** | 0.42 | 0.68 | **0.52** |
| **Gradient Boosting Classifier** | **0.71** | 0.59 | 0.39 | **0.39** |
| **Dense Neural Network** | **0.71** | 0.60 | 0.27 | **0.36** |

# Feature Importance (ML)

Past Voting Behaviour

Age

Income/Employment

# Problems with ‘pure’ ML

1

Does pressure change voting rates?

**Lack of consensus between models**

2

If so, by how much?

**Magnitude of treatment uncertain**

3

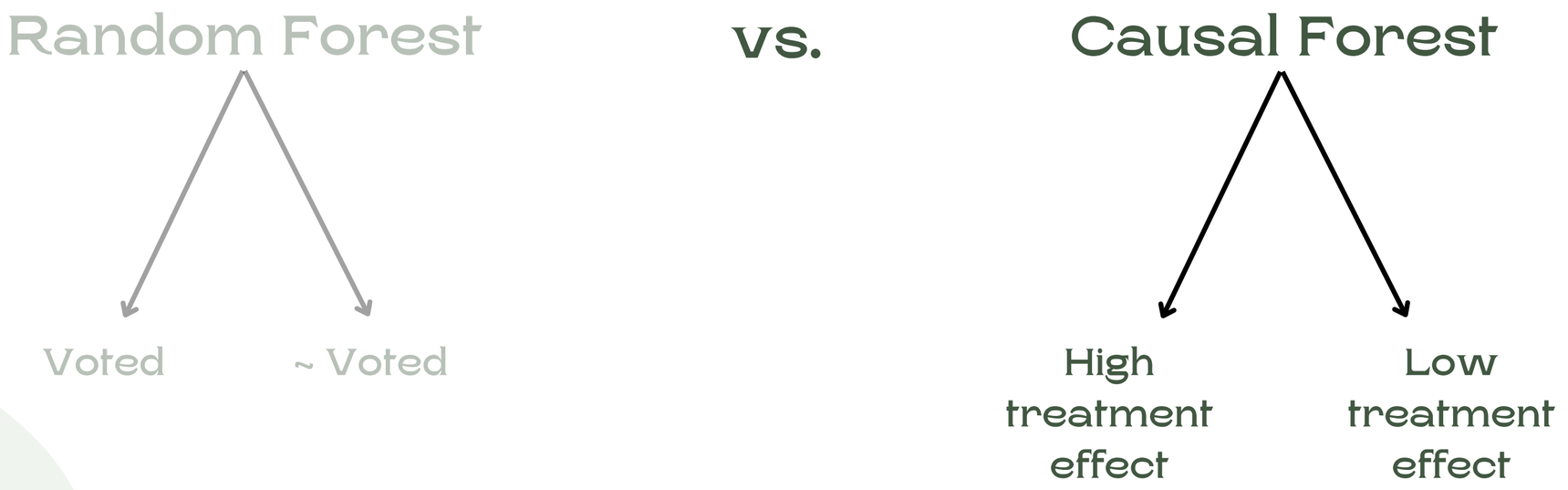
For whom, in particular?

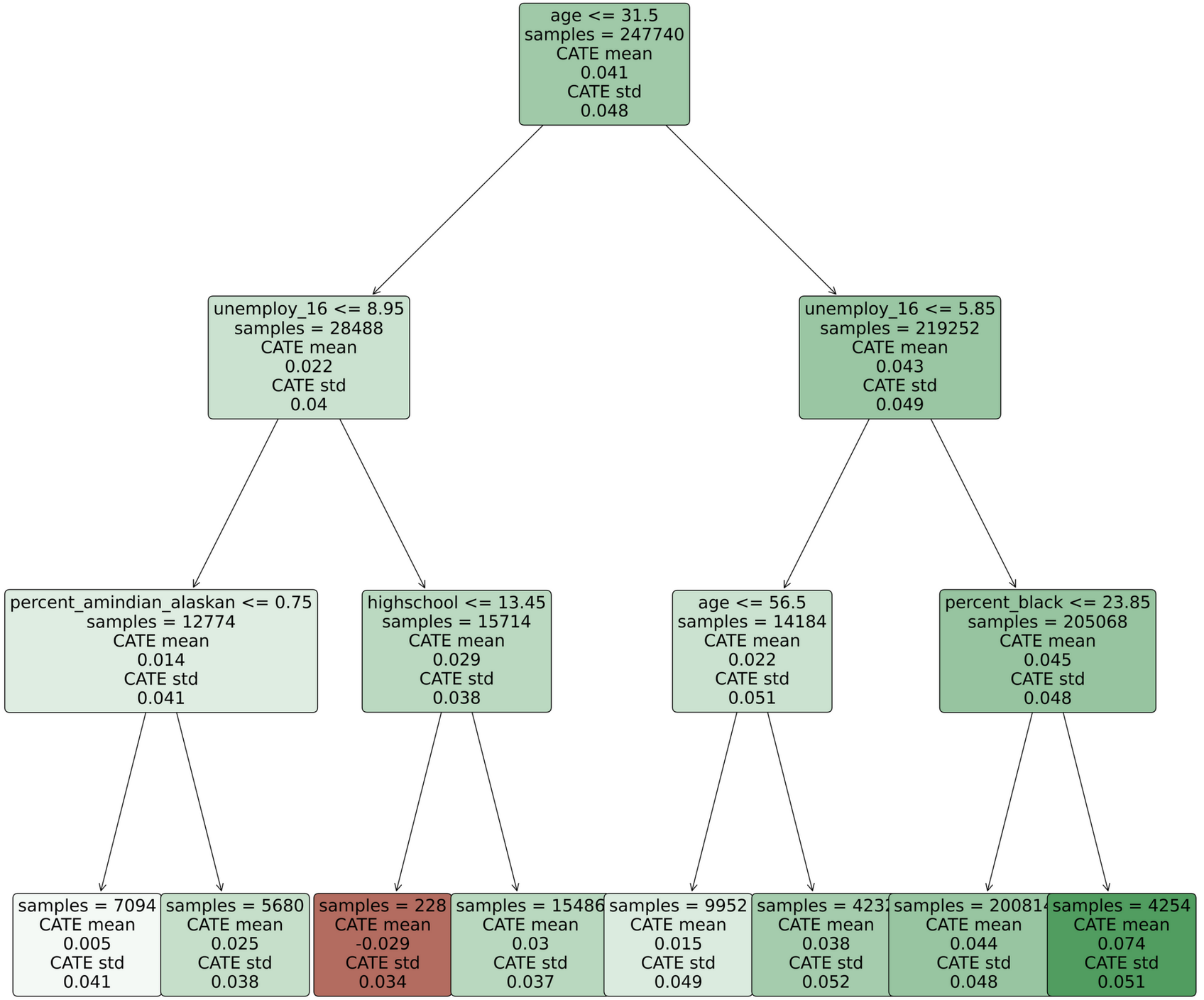
**No information on subgroups**

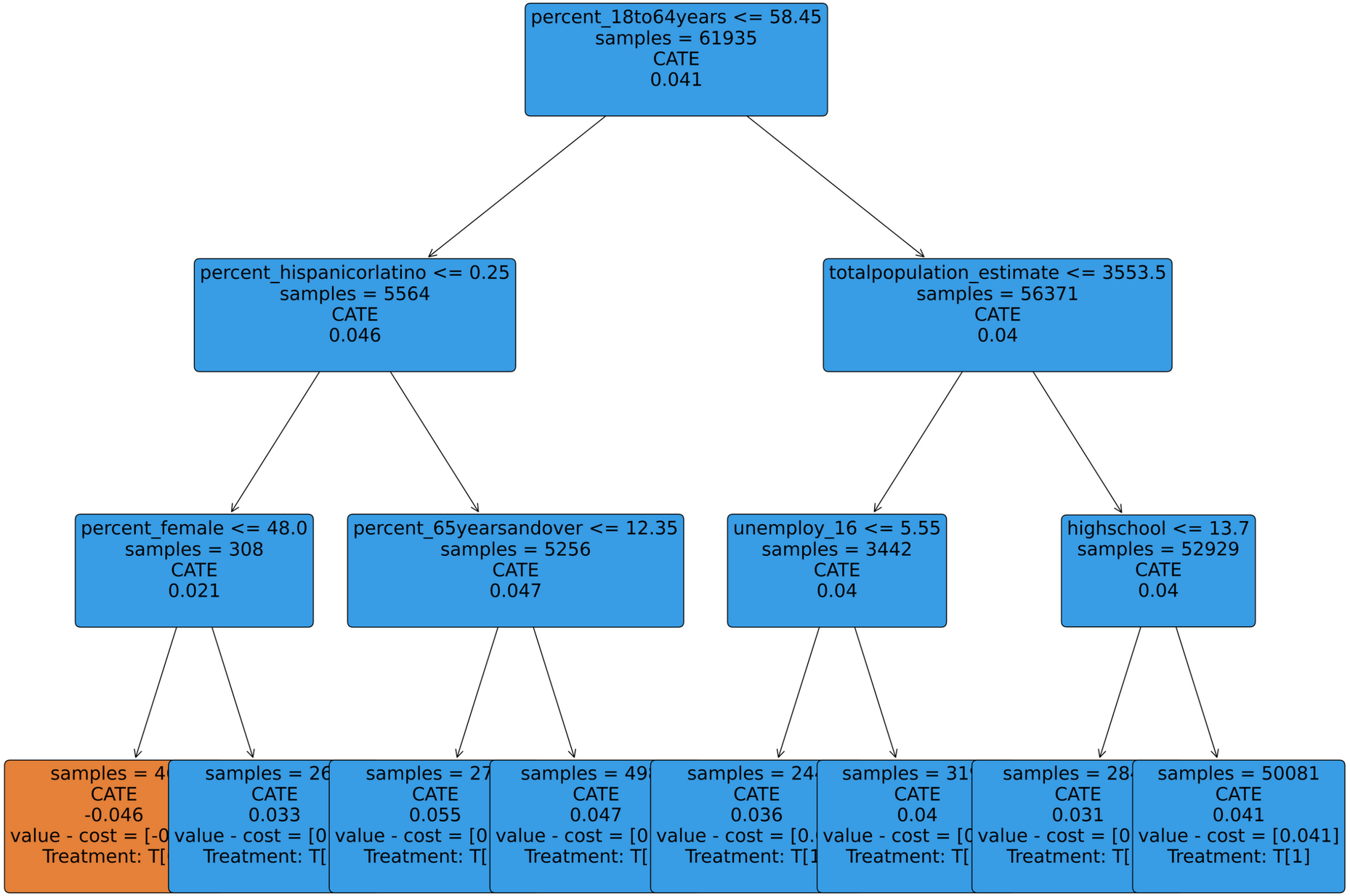
# #4
# ML + Causal Inference

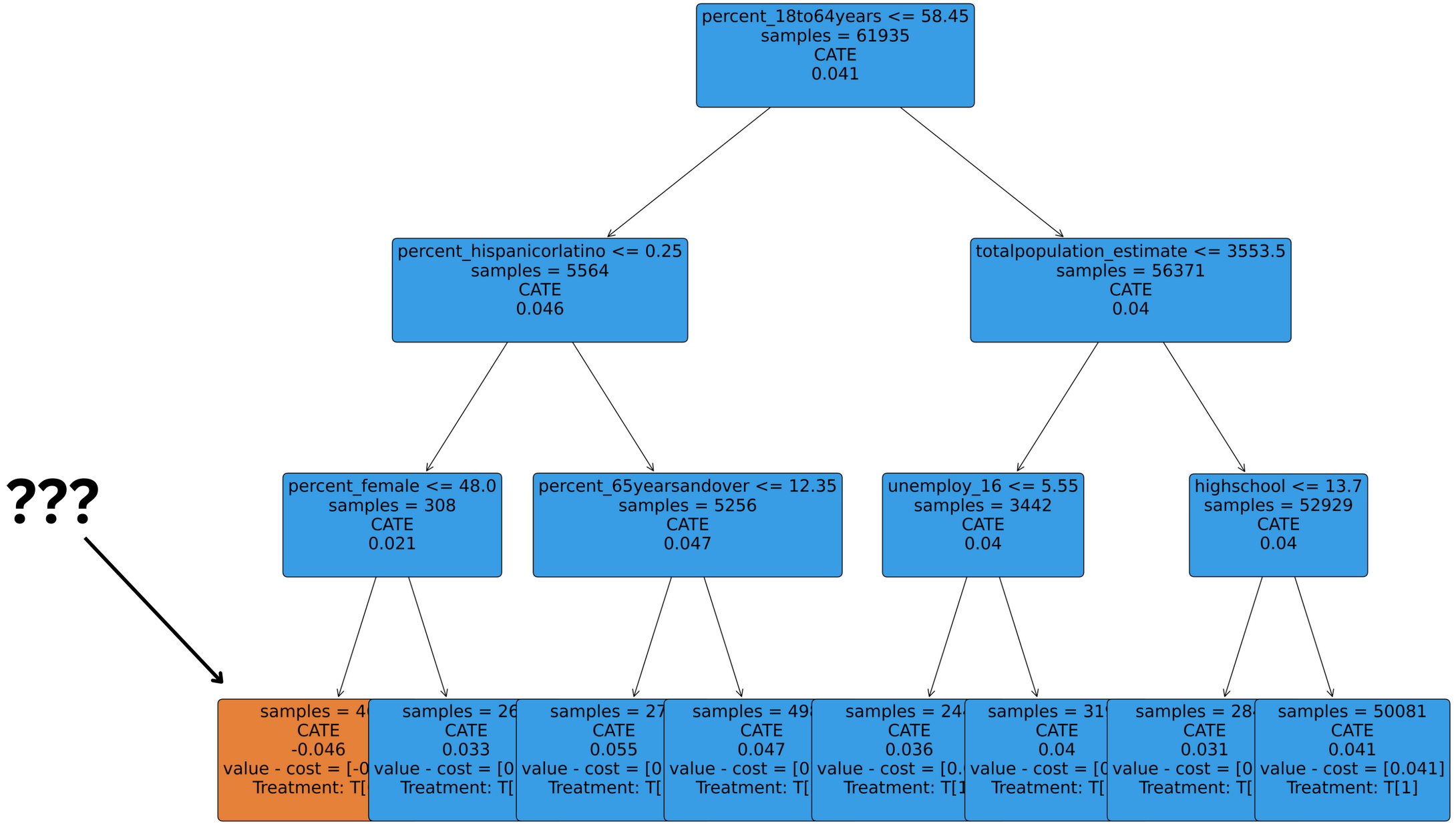
# Causal Forest Model

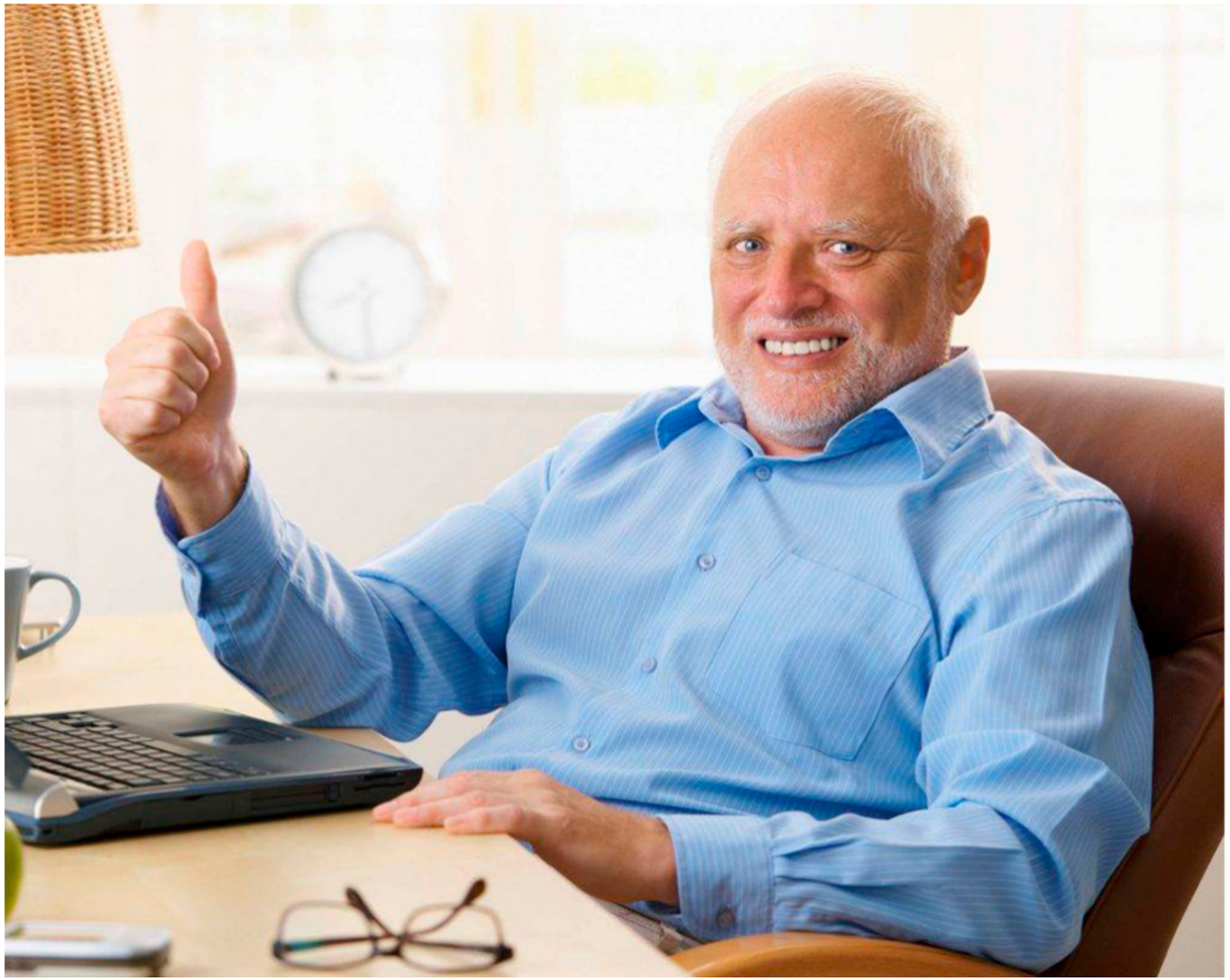
# Causal Forest Model




age <= 31.5
samples = 247740
CATE mean
0.041
CATE std
0.048
unemploy_16 <= 8.95
samples = 28488
CATE mean
0.022
CATE std
0.04
unemploy_16 <= 5.85
samples = 219252
CATE mean
0.043
CATE std
0.049
percent_amindian_alaskan <= 0.75
samples = 12774
CATE mean
0.014
CATE std
0.041
highschool <= 13.45
samples = 15714
CATE mean
0.029
CATE std
0.038
age <= 56.5
samples = 14184
CATE mean
0.022
CATE std
0.051
percent_black <= 23.85
samples = 205068
CATE mean
0.045
CATE std
0.048
samples = 7094
CATE mean
0.005
CATE std
0.041
samples = 5680
CATE mean
0.025
CATE std
0.038
samples = 228
CATE mean
-0.029
CATE std
0.034
samples = 15486
CATE mean
0.03
CATE std
0.037
samples = 9952
CATE mean
0.015
CATE std
0.049
samples = 4232
CATE mean
0.038
CATE std
0.052
samples = 200814
CATE mean
0.044
CATE std
0.048
samples = 4254
CATE mean
0.074
CATE std
0.051


percent_18to64years <= 58.45
samples = 61935
CATE
0.041
percent_hispanicorlatino <= 0.25
samples = 5564
CATE
0.046
totalpopulation_estimate <= 3553.5
samples = 56371
CATE
0.04
percent_female <= 48.0
samples = 308
CATE
0.021
percent_65yearsandover <= 12.35
samples = 5256
CATE
0.047
unemploy_16 <= 5.55
samples = 3442
CATE
0.04
highschool <= 13.7
samples = 52929
CATE
0.04
samples = 4
CATE
-0.046
value - cost = [-0
Treatment: T[
samples = 26
CATE
0.033
value - cost = [0
Treatment: T[
samples = 27
CATE
0.055
value - cost = [0
Treatment: T[
samples = 49
CATE
0.047
value - cost = [0
Treatment: T[
samples = 24
CATE
0.036
value - cost = [0.
Treatment: T[1
samples = 31
CATE
0.04
value - cost = [0
Treatment: T[
samples = 28
CATE
0.031
value - cost = [0
Treatment: T[
samples = 50081
CATE
0.041
value - cost = [0.041]
Treatment: T[1]


???
percent_18to64years <= 58.45
samples = 61935
CATE
0.041
percent_hispanicorlatino <= 0.25
samples = 5564
CATE
0.046
totalpopulation_estimate <= 3553.5
samples = 56371
CATE
0.04
percent_female <= 48.0
samples = 308
CATE
0.021
percent_65yearsandover <= 12.35
samples = 5256
CATE
0.047
unemploy_16 <= 5.55
samples = 3442
CATE
0.04
highschool <= 13.7
samples = 52929
CATE
0.04
samples = 4
CATE
-0.046
value - cost = [-0
Treatment: T[
samples = 26
CATE
0.033
value - cost = [0
Treatment: T[
samples = 27
CATE
0.055
value - cost = [0
Treatment: T[
samples = 49
CATE
0.047
value - cost = [0
Treatment: T[
samples = 24
CATE
0.036
value - cost = [0.
Treatment: T[1
samples = 31
CATE
0.04
value - cost = [0
Treatment: T[
samples = 28
CATE
0.031
value - cost = [0
Treatment: T[
samples = 50081
CATE
0.041
value - cost = [0.041]
Treatment: T[1]



# #5
# Model Comparison

# Model Comparison

| | Pure ML | ML + Causal Inference |
|---|---|---|
| Does social pressure impact voting rates? | | |
| If so, by how much? | | |
| For whom? | | |

# Model Comparison

| | Pure ML | ML + Causal Inference |
|---|---|---|
| Does social pressure impact voting rates? |  | |
| If so, by how much? |  | |
| For whom? |  | |

# Model Comparison

| | Pure ML | ML + Causal Inference |
|---|---|---|
| **Does social pressure impact voting rates?** |  | Yes |
| **If so, by how much?** |  | 4.6%, on average |
| **For whom?** |  | Well-defined subgroups |

# #6
# Next Steps

# Next Steps

S L E E P !

## Productise

Create an interactive dashboard using Streamlit for demo day.

## Clustering

Try to find more meaningful, well-defined clusters.

## Bayesian Networks

Capture more fine-grained causal relationships.

# Thank you!