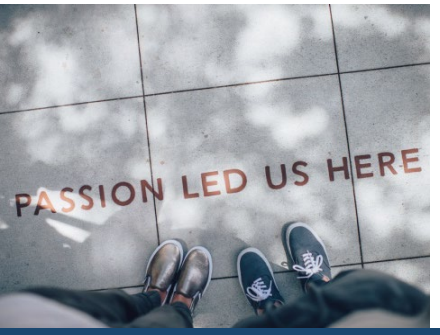


Lab 6: Predicting Social Mobility Using Decision Trees and Regression

Gregory Bruich, Ph.D.

HARVARD UNIVERSITY

OPPORTUNITY INSIGHTS

Lab 6: Predicting Social Mobility Using Decision Trees and Regression

- In the next several labs, we will discuss policies related to **prediction**
  - Criminal justice application: who will fail to appear in court?
- In contrast, previous labs were about **causal inference**
- Causal inference is predictive too, but predictive under *manipulation*
- In today's lab, we will develop two methods for prediction:
  1. Multivariate regression
  2. Decision trees
- **Application:** Predicting Social Mobility using Community Characteristics

Key Lessons from Lab 6

- Substantive question: how best to predict upward mobility using community characteristics?
- Key machine learning idea is to split data into “training” and “test samples”
- Cherry picking or “over fitting” in the training data will lead to bad performance when trying to make out of sample prediction
- Key methodological tools:
    1. Multivariable regressions and decision trees to make prediction
    2. Controlling complexity of a decision tree by changing tuning parameters
    3. Using root mean squared prediction error (RMSPE) to assess accuracy

Predicting Social Mobility Using a Linear Regression

Milwaukee, WI

- In Milwaukee, WI, there are 5.72 bowling alleys per 100,000 residents and 22.6% single parent families

$$\hat{Y}_i = 56.9 + 0.41 \times Bowling\ Alleys_i - 70.1 \times Single\ Parents_i$$

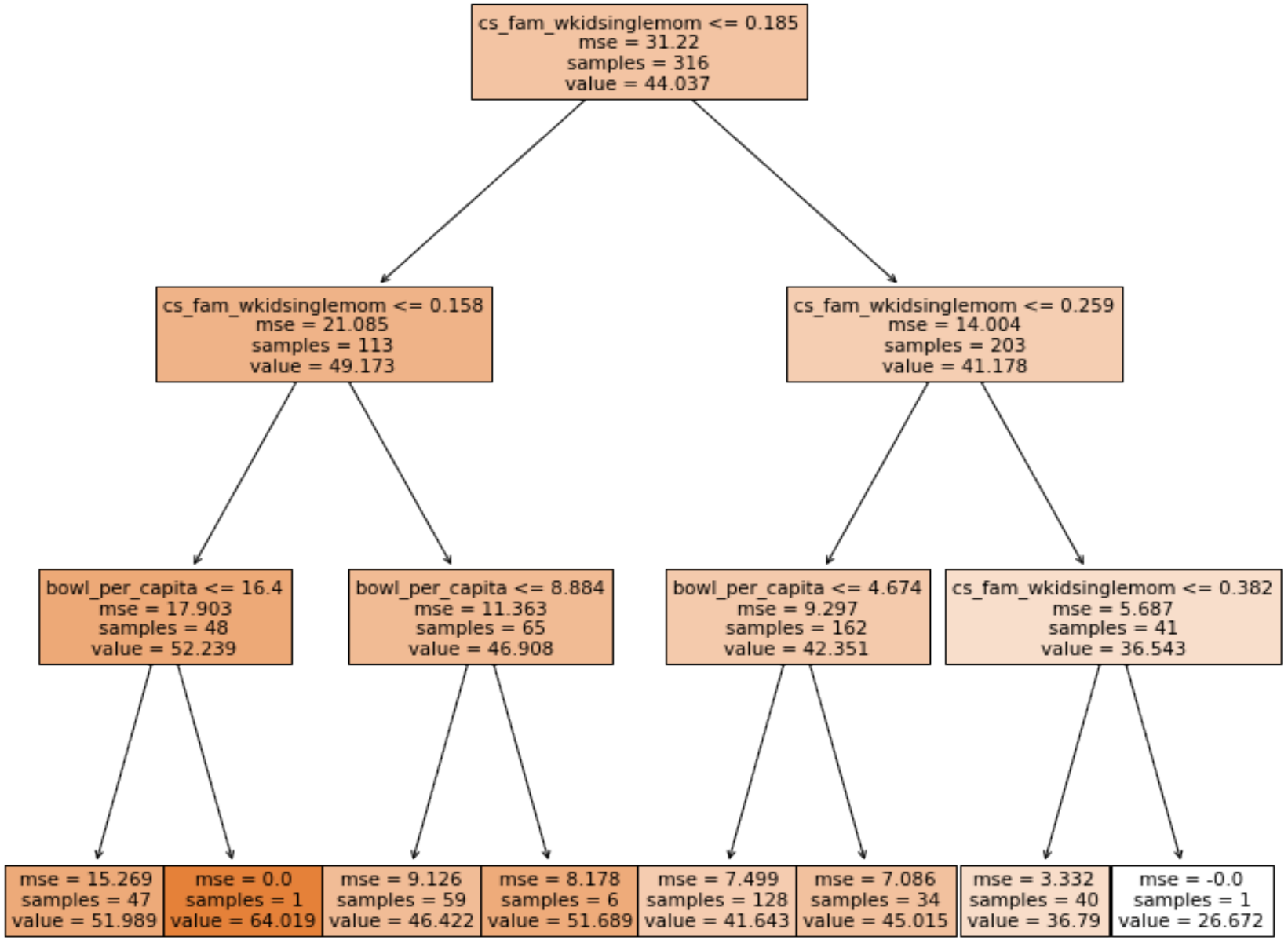
$$= 56.9 + 0.41 \times 5.72 - 70.1 \times 0.226$$

$$= 43.4 \text{ percentile, or about } \$34{,}000$$

- The prediction error is: $38.9 - 43.4 = -4.5$ percentiles

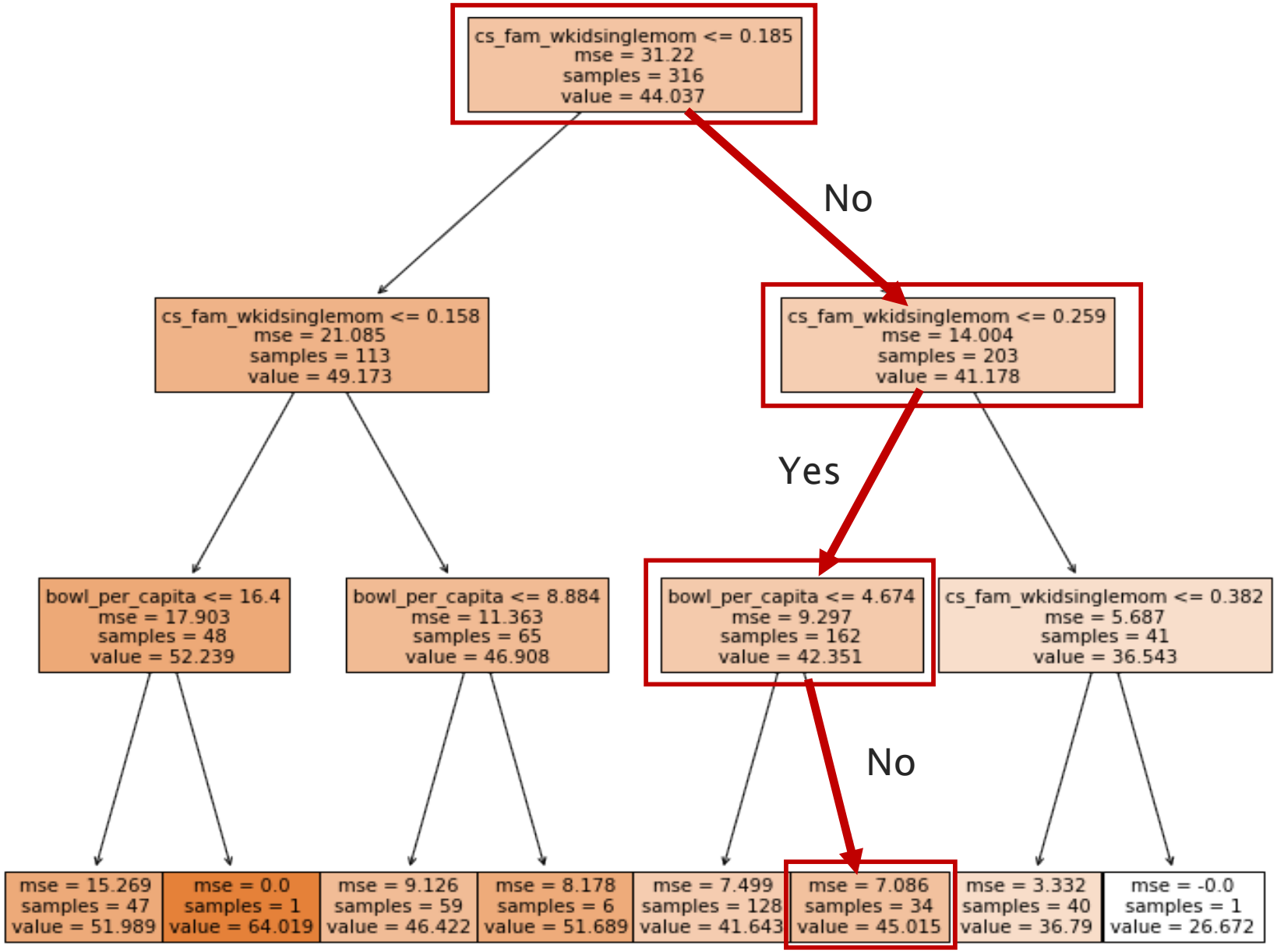
Predicting Social Mobility Using a Decision Tree

Milwaukee, WI



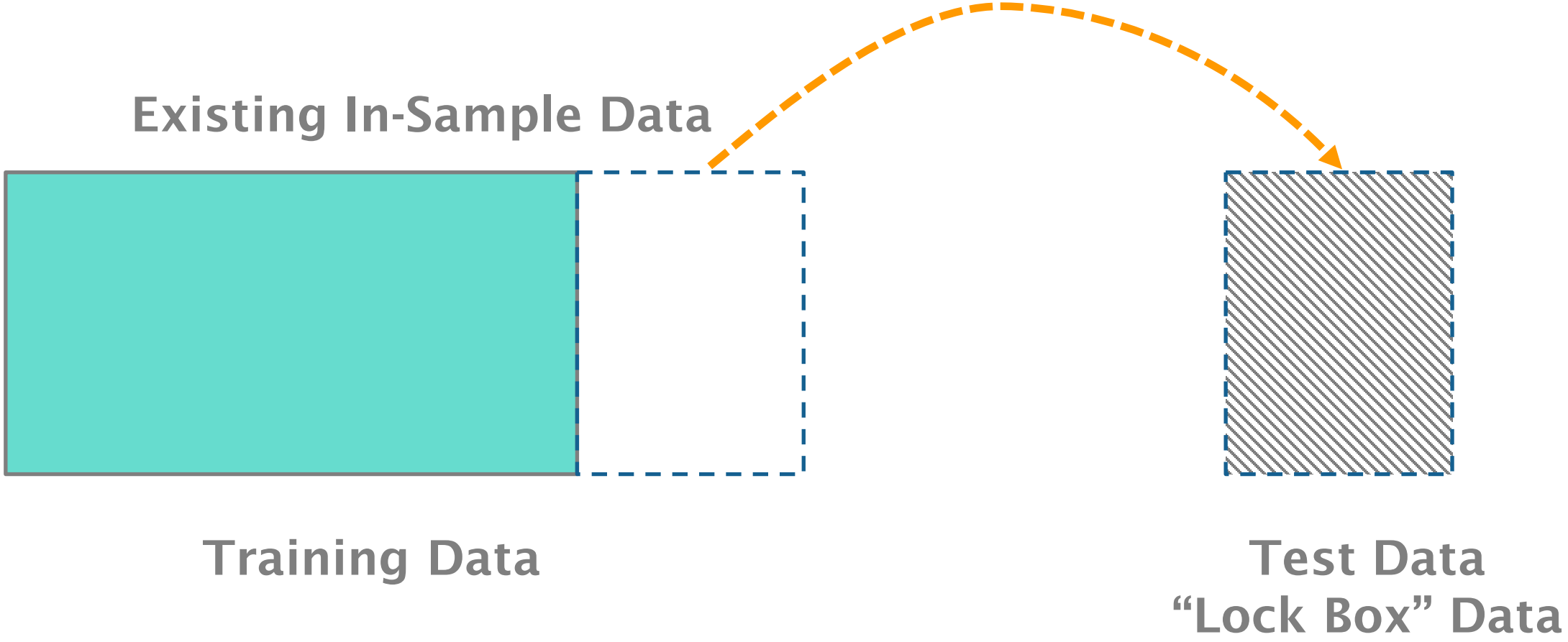
Predicting Social Mobility Using a Decision Tree

Milwaukee, WI



Pseudo Out-of-Sample Validation of Prediction Models

Splitting Existing Data into Training and Test Data



Prediction Errors vs. Size of Decision Tree