


Lab 7: Predicting Social Mobility using Cross Validation and Random Forests

Gregory Bruich, Ph.D

Lab 7: Predicting Social Mobility using Cross Validation and Random Forests

- Reminder: Project Part 2 is due Fri. April 14, 2023 → budget your time wisely

- Last week's coding exercise highlighted the issue of **overfit**: the model that does well in the training data does not do well for a *new* observation

- In this lab, we will implement two methods for prediction that address overfit:

  1. **Cross validation** to choose complexity of a decision tree

  2. **Random forests** as a specific improvement upon decision trees

- We will also review the dynamic model of intergenerational mobility that Professor Chetty introduced in Lecture to study long-run racial disparities

- These three ideas are connected: they all have a recursive structure

Key Lessons from Lab 7

- Substantive question: what do differences in intergenerational mobility imply about the long-run evolution of racial disparities in economic outcomes?

- Then we will return to predicting social mobility using community characteristics

- Key methodological tools:

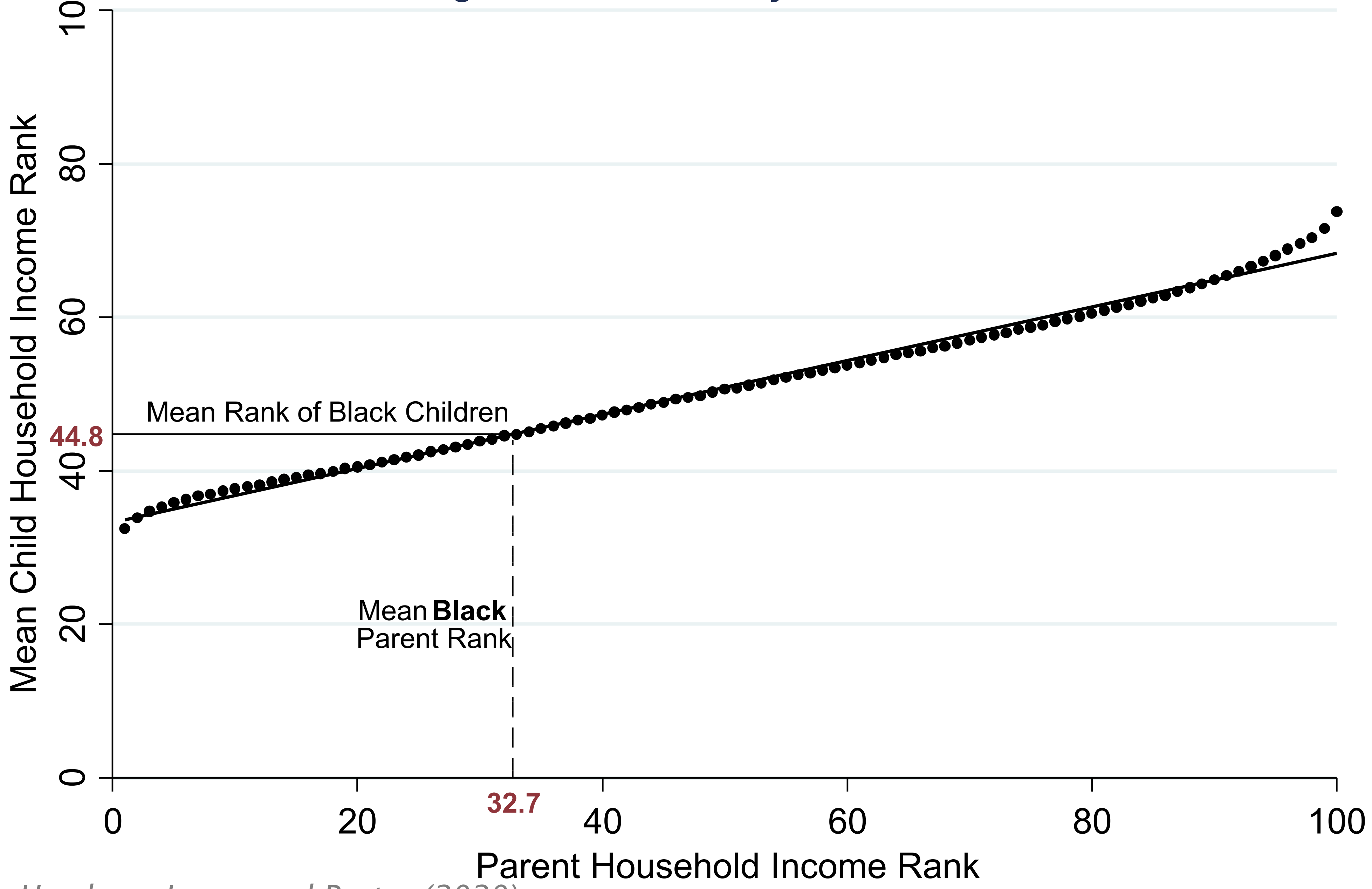
  1. Writing loops to perform iterative/recursive computations

  2. Cross-validation to solve the overfit problem by using the data you have to choose a low dimensional measure of model complexity

  3. Random forests to address overfit using “bagging” and “input randomization”

  4. Interpreting variable importance summary plots for random forests models

Becker and Tomes (1979) Model Predicts Convergence in Incomes Across Race and Ethnicity if Intergenerational Mobility is Race-Invariant



Source: Chetty, Hendren, Jones, and Porter (2020)

Dynamics of Intergenerational Mobility: Becker and Tomes (1979)

- Average of $Rank_p$ for Black children in U.S. is $Rank_p = 32.7$
- Chetty et al. (2020) report the following rank-rank regression pooling all races and genders:

$$Rank_k = 33.31 + 0.351 \times Rank_p$$

Dynamics of Intergenerational Mobility: Becker and Tomes (1979)

- Average of $Rank_p$ for Black children in U.S. is $Rank_p = 32.7$

- Chetty et al. (2020) report the following rank-rank regression pooling all races and genders:

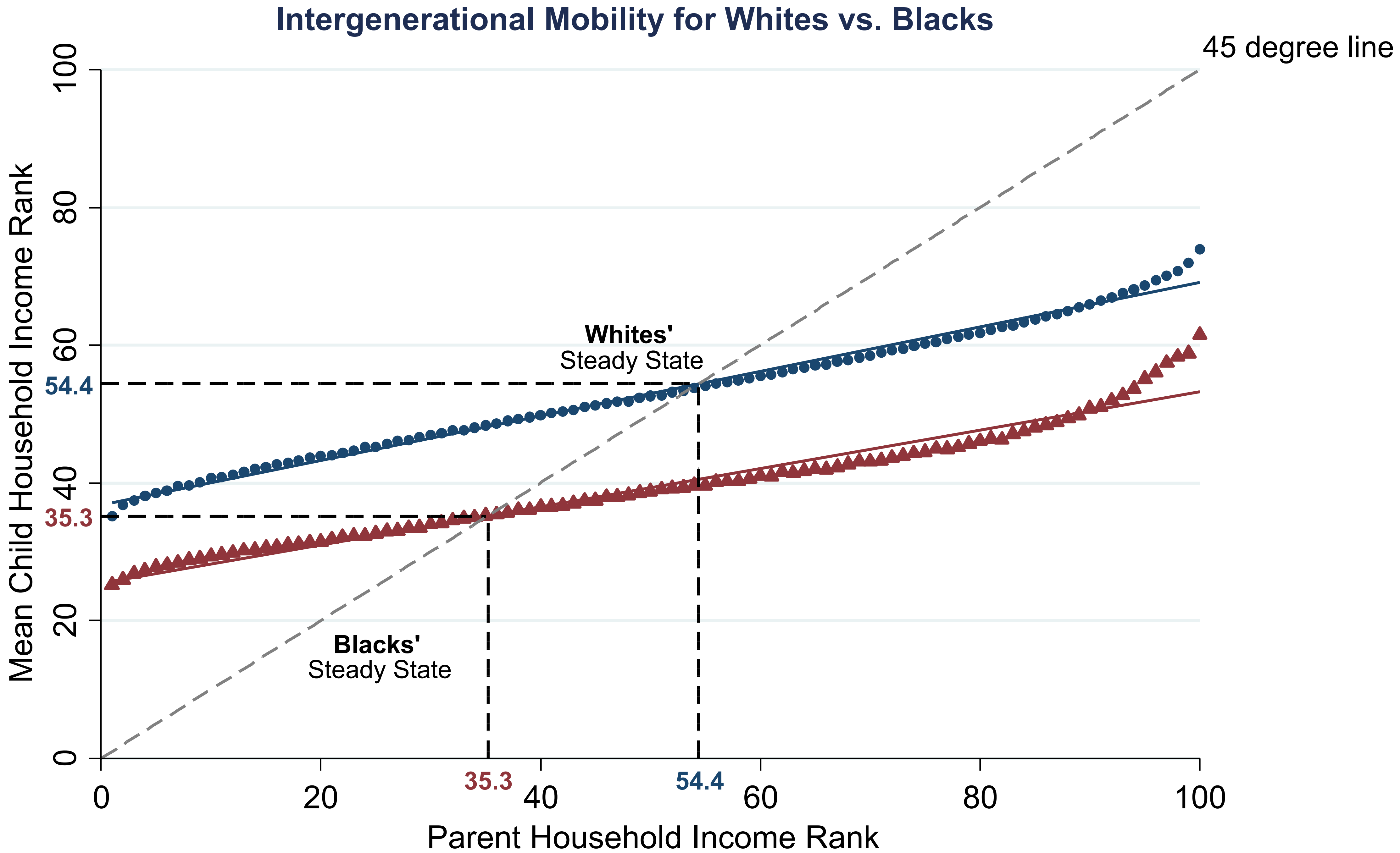
$$Rank_k = 33.31 + 0.351 \times Rank_p$$

$$= 33.31 + 0.351 \times 32.7$$

$$= 44.8^{\text{th}} \text{ percentile}$$

- Using the sample code, show that this model makes the unrealistic prediction of convergence in incomes across racial groups

- Key result from Chetty et al. (2020): children of different races experience very different rates of upward mobility across generations → inequality will persist



Source: Chetty, Hendren, Jones, and Porter (2020)

Dynamics of Intergenerational Mobility with Race-specific Rank-Rank Regressions

- Average of $Rank_p = 32.7$ for Black children in U.S.

- Rank-rank regression predicts mean rank as an adult:

$$\begin{aligned} Rank_k &= 25.4 + 0.28 \times Rank_p \\ &= 25.4 + 0.28 \times 32.7 \\ &= {\color{red}34.6} \end{aligned}$$

Generation 1

Dynamics of Intergenerational Mobility with Race-specific Rank-Rank Regressions

- Average of $Rank_p = 32.7$ for Black children in U.S.

- Rank-rank regression predicts mean rank as an adult:

$$\begin{aligned} Rank_k &= 25.4 + 0.28 \times Rank_p \\ &= 25.4 + 0.28 \times 32.7 \\ &= {\color{red}34.6} \end{aligned}$$

Generation 1

- When those children go on to have children of their own, we would predict their children will be at:

$$\begin{aligned} Rank_k &= 25.4 + 0.28 \times Rank_p \\ &= 25.4 + 0.28 \times {\color{red}34.6} \\ &= 35.1 \end{aligned}$$

Generation 2

Dynamics of Intergenerational Mobility: Steady state (or fixed point)

- We can keep iterating on this equation (computers are good at this)

- For the next generation, we would predict:

$$Rank_k = 25.4 + 0.28 \times 35.1 = 35.2$$ Generation 3

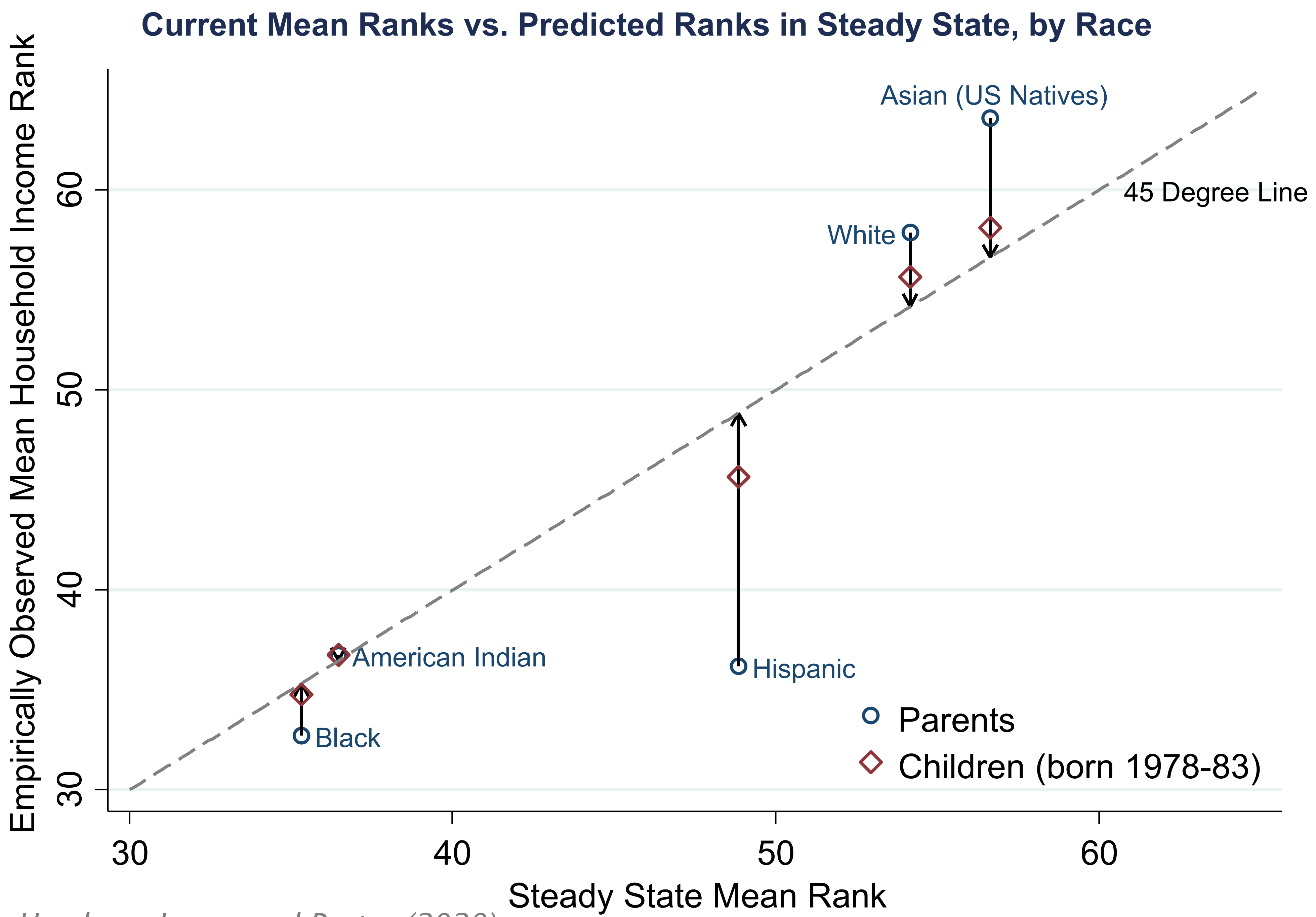
- For the generation after that, we would predict:

$$Rank_k = 25.4 + 0.28 \times 35.2 = 35.3$$ Generation 4

- And for the generation after that, we would predict

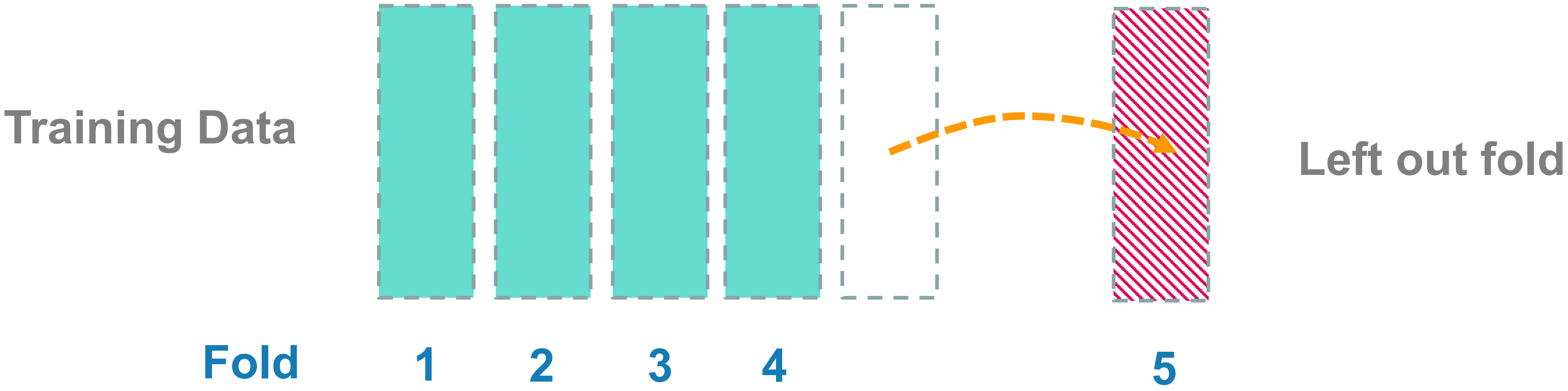
$$Rank_k = 25.4 + 0.28 \times 35.3 = 35.3$$ Generation 5

- This is a fixed point or steady state: $Rank_k = Rank_p$, which means that no further improvement in income is predicted by the model

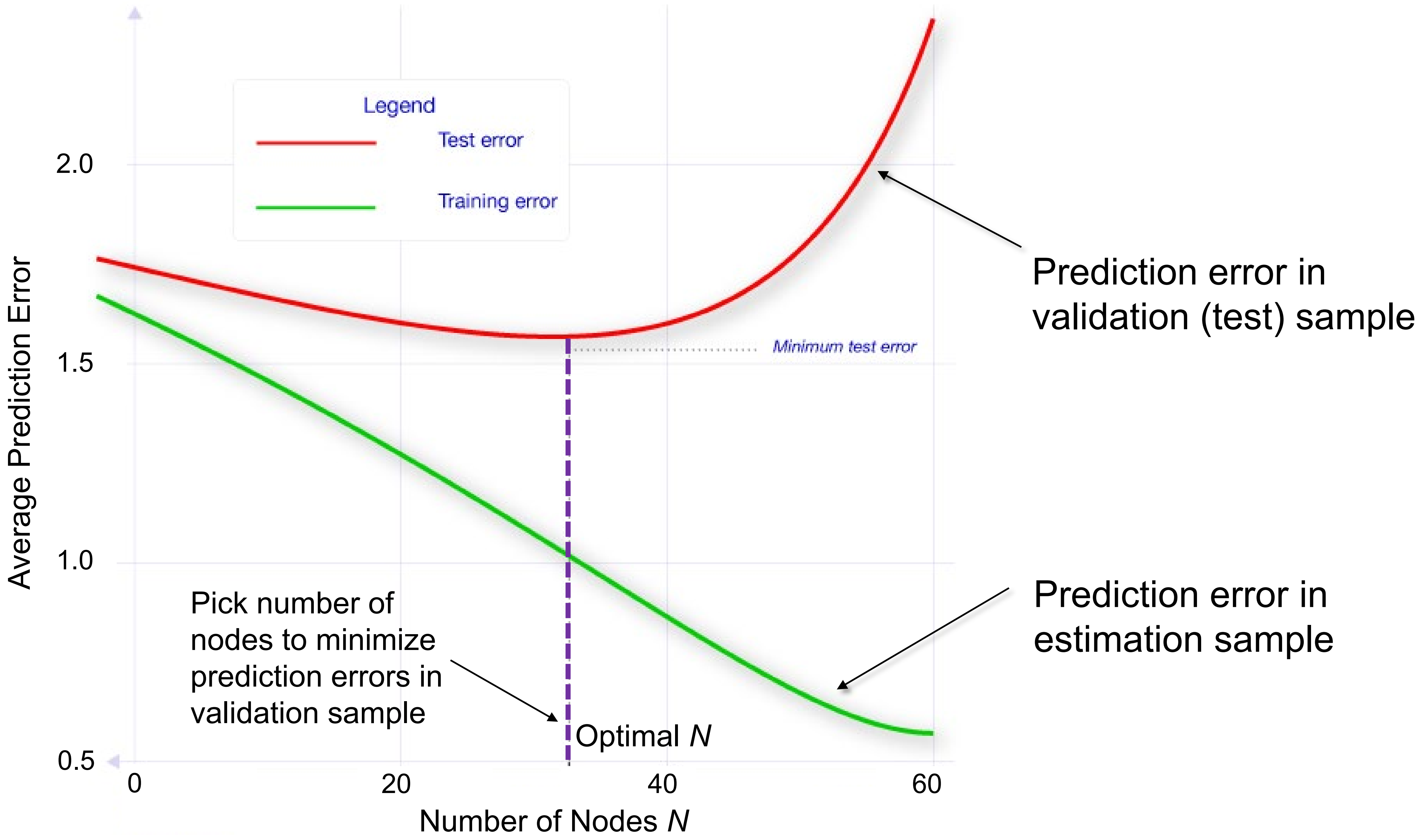


Source: Chetty, Hendren, Jones, and Porter (2020)

Primer on Cross Validation



Prediction Errors vs. Size of Decision Tree



Primer on Random Forests: “Bagging” and “Input Randomization”