Using Machine Learning to Estimate the Effect of Undocumented Status on Education-Occupation Mismatch for College Graduates

Dr. Veronica Sovero[1]
Mario Arce Acosta [2]

[1]University of California, Riverside
[2]University of California, Davis

Contents

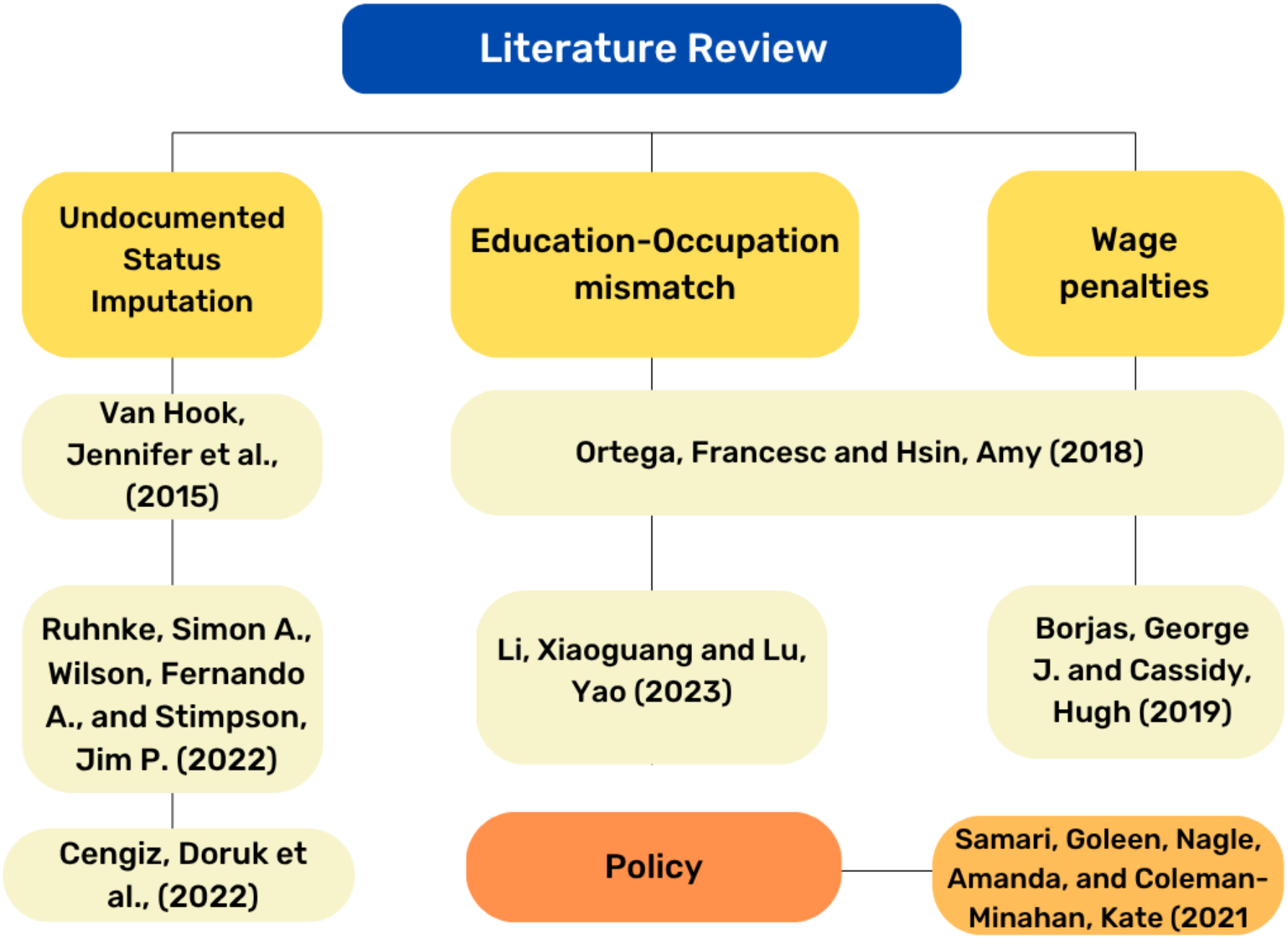
Introduction

Motivation

- Many states offer a combination of in-state tution and financial aid to undocumented immigrants. [1]
- There are roughly 1.71 million undocumented immigrants with a college education residing in the United States. [2]
- For college graduates, how does undocumented status affect:
  - Occupation-education mismatch?
  - Wages?

[1] https://www.higheredimmigrationportal.org/states/

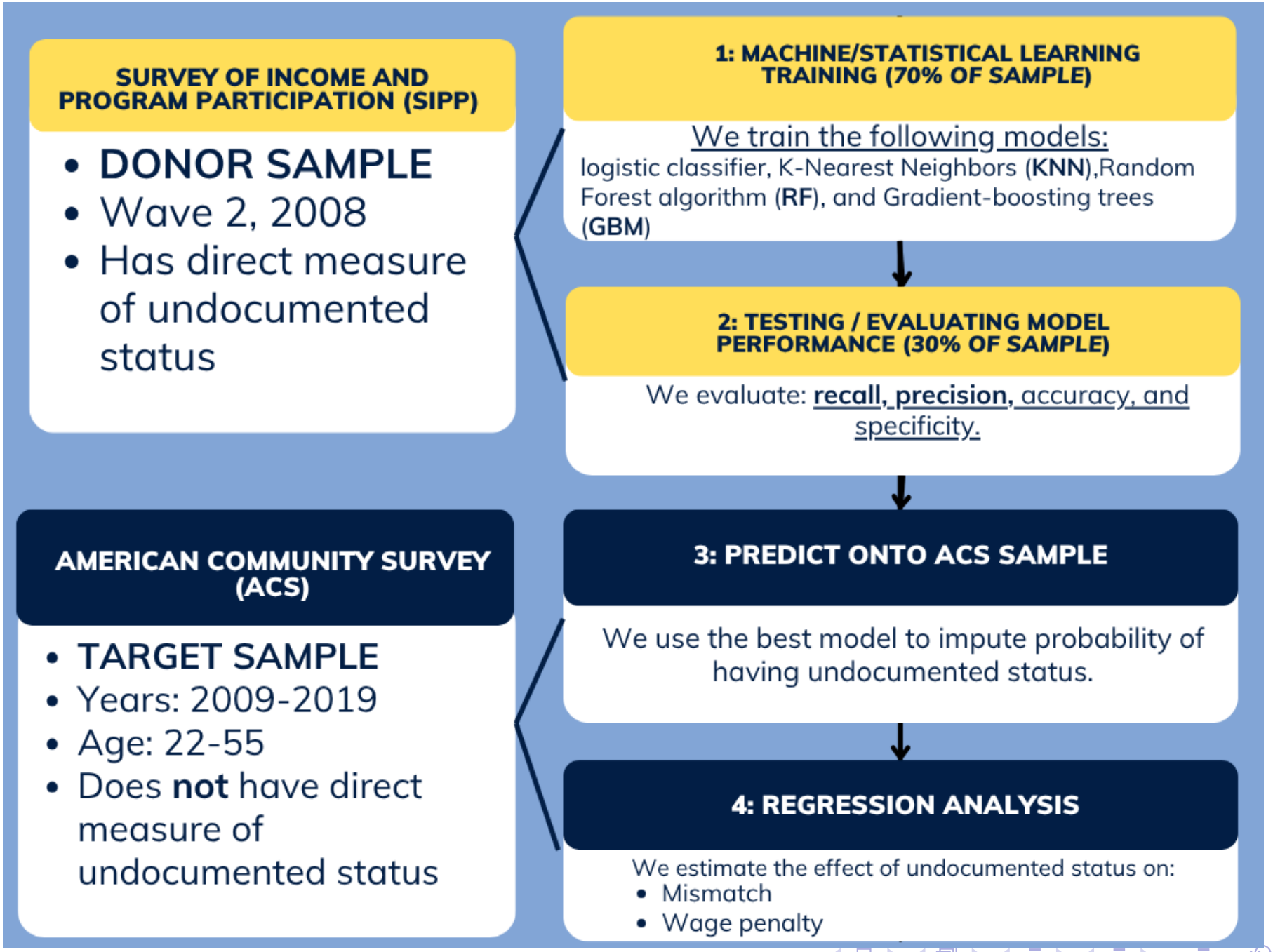
[2] https://cmsny.org/educated-immigrants-millet-080122/


Literature Review
Undocumented Status Imputation
Education-Occupation mismatch
Wage penalties
Van Hook, Jennifer et al., (2015)
Ortega, Francesc and Hsin, Amy (2018)
Ruhnke, Simon A., Wilson, Fernando A., and Stimpson, Jim P. (2022)
Li, Xiaoguang and Lu, Yao (2023)
Borjas, George J. and Cassidy, Hugh (2019)
Cengiz, Doruk et al., (2022)
Policy
Samari, Goleen, Nagle, Amanda, and Coleman-Minahan, Kate (2021

Our contribution

- Utilize machine learning methods for undocumented status imputation in the American Community Survey (ACS)
- Estimate the impact of undocumented status on education-occupation mismatch and wage penalties for college graduates (smaller, hidden population)
- Examine the role of federal and state-level policy on labor market outcomes for undocumented immigrants

Data and Methods


SURVEY OF INCOME AND PROGRAM PARTICIPATION (SIPP)
• DONOR SAMPLE
• Wave 2, 2008
• Has direct measure of undocumented status
1: MACHINE/STATISTICAL LEARNING TRAINING (70% OF SAMPLE)
We train the following models:
logistic classifier, K-Nearest Neighbors (KNN),Random Forest algorithm (RF), and Gradient-boosting trees (GBM)
2: TESTING / EVALUATING MODEL PERFORMANCE (30% OF SAMPLE)
We evaluate: recall, precision, accuracy, and specificity.
AMERICAN COMMUNITY SURVEY (ACS)
• TARGET SAMPLE
• Years: 2009-2019
• Age: 22-55
• Does not have direct measure of undocumented status
3: PREDICT ONTO ACS SAMPLE
We use the best model to impute probability of having undocumented status.
4: REGRESSION ANALYSIS
We estimate the effect of undocumented status on:
• Mismatch
• Wage penalty

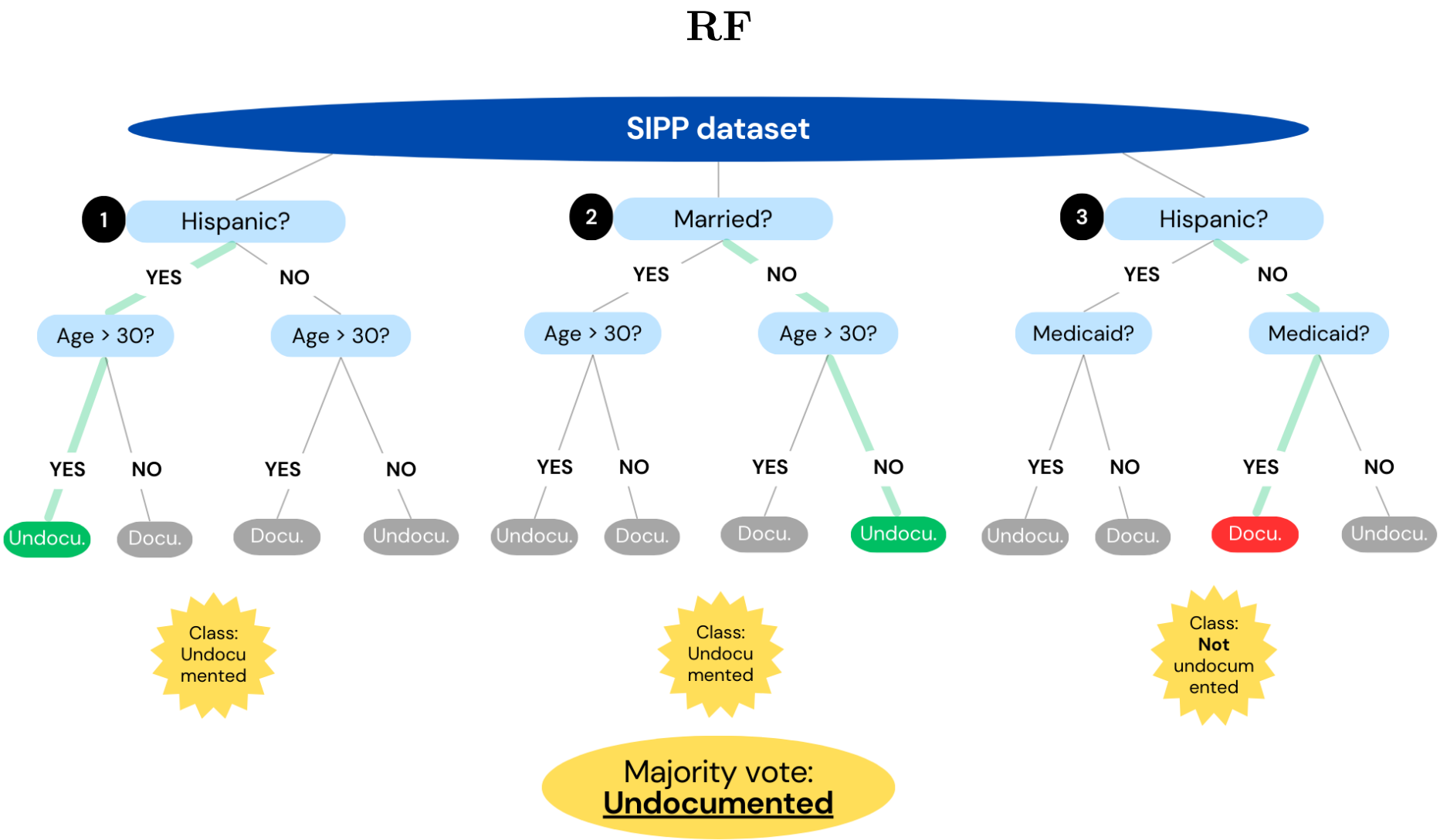
Training Sample

We apply ***logical imputation*** to generate the SIPP training sample, classifying a noncitizen as possibly undocumented if they **do not** meet any of the following conditions:

- Veteran status
- Medicare receipt
- Social Security income receipt
- Arrived before January 1st, 1982

Undocumented status is observed in this training sample.

Random Forest algorithm



Gradient-boosting trees algorithm

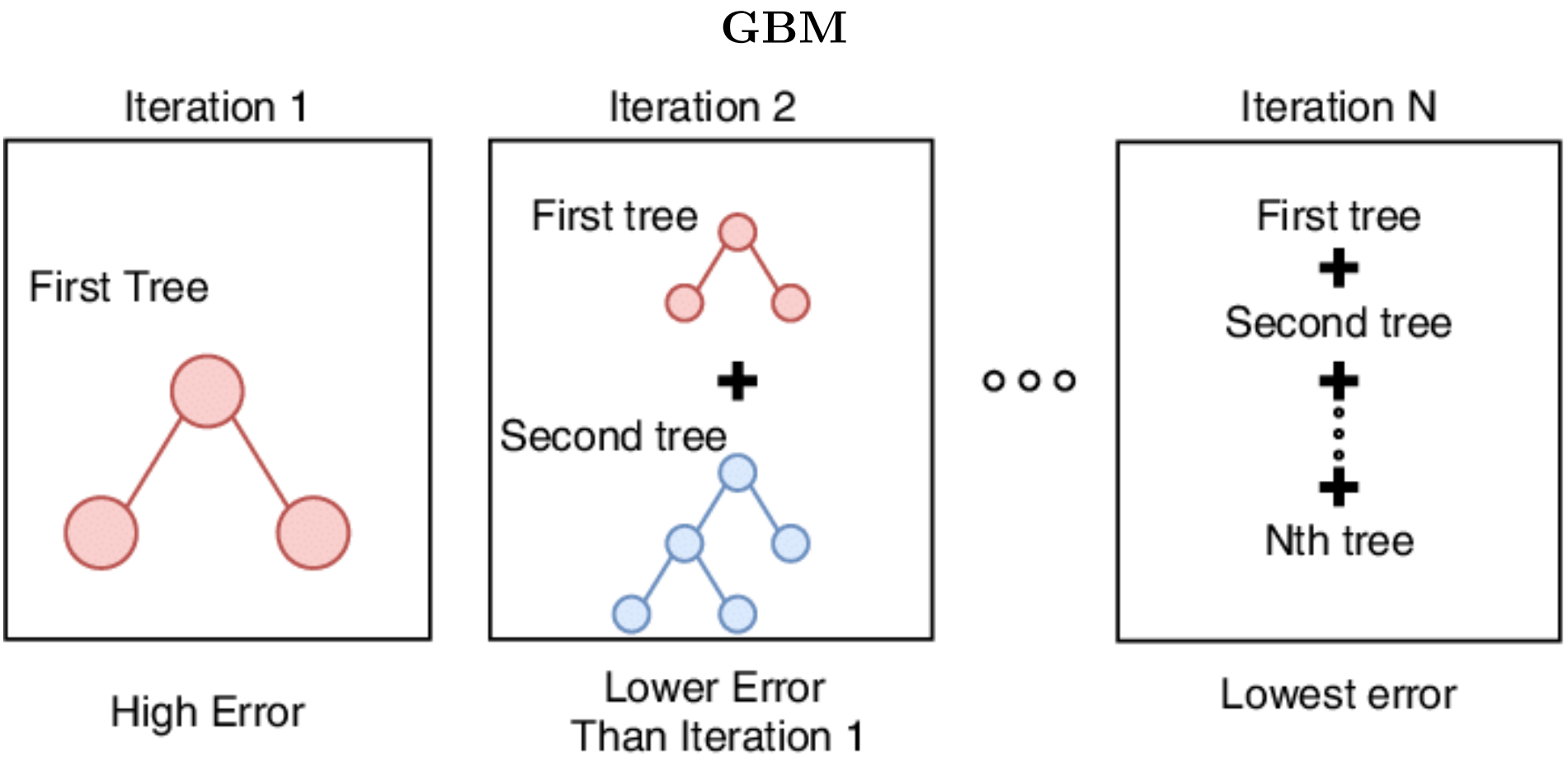


Figure: (Source) https://www.researchgate.net/figure/Schematic-representation-of-the-Gradient-boosted-tree-model_fig4_351856003

Model performance metrics

Precision:

$$\frac{True\ Positives}{True\ Positives+False\ Positives} = \frac{\#\ Correctly\ classified\ as\ undocumented}{\#\ Classified\ as\ undocumented}$$

Recall:

$$\frac{True\ Positives}{True\ Positives+False\ Negatives} = \frac{\#\ Correctly\ classified\ as\ undocumented}{\#\ Truly\ undocumented}$$

Specificity: $\frac{True\ Negatives}{True\ Negatives+False\ Positives}$

Accuracy: $\frac{True\ Positives+True\ Negatives}{Total}$

Defining Vertical and Horizontal mismatch

People are considered mismatched within their occupation using the following definitions:[3]

Vertically mismatched:
$Educational\ attainment \neq Most\ common\ educational\ attainment$

Horizontally mismatched:
$Field\ of\ study \neq Two\ most\ common\ fields\ of\ study$

Horizontally undermatched:

$Median\ wage\ of\ occupation < Median\ wage\ of\ field\ of\ study\ (of\ those\ horizontally\ matched)$

Horizontally overmatched:

$Median\ wage\ of\ occupation > Median\ wage\ of\ field\ of\ study\ (of\ those\ horizontally\ matched)$

[3] U.S. born workers are the reference group for modal occupations and fields of study.

Econometric model

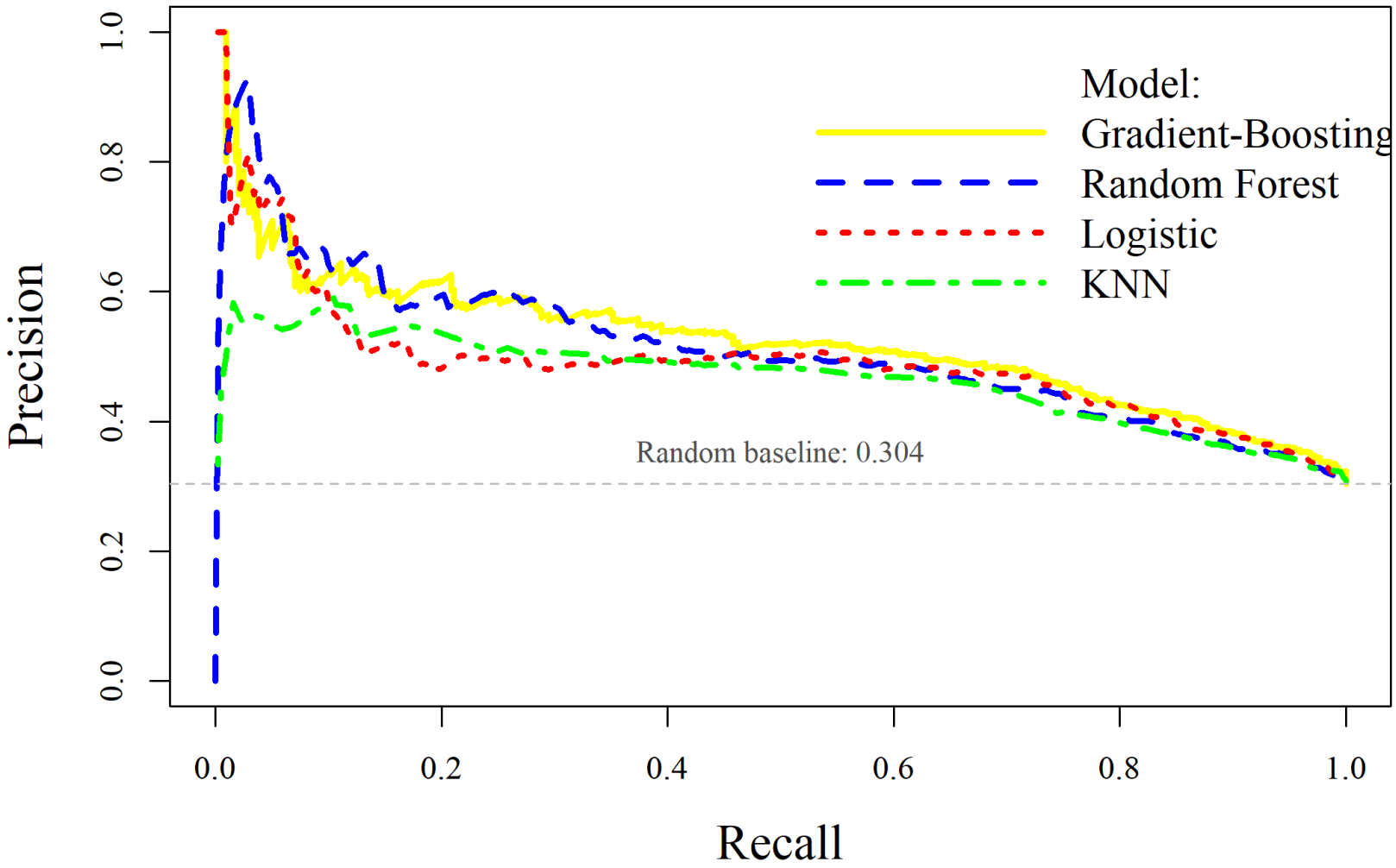
(1) $Vertical\ Mismatch_i = \beta X_i + \beta_U Hundermatch_i + \beta_O Hovermatch_i + \beta_u Undocu_i + \varepsilon_i$

(2) $Horizontal\ Mismatch_i = \beta X_i + \beta_V Vmismatch_i + \beta_u Undocu_i + \varepsilon_i$

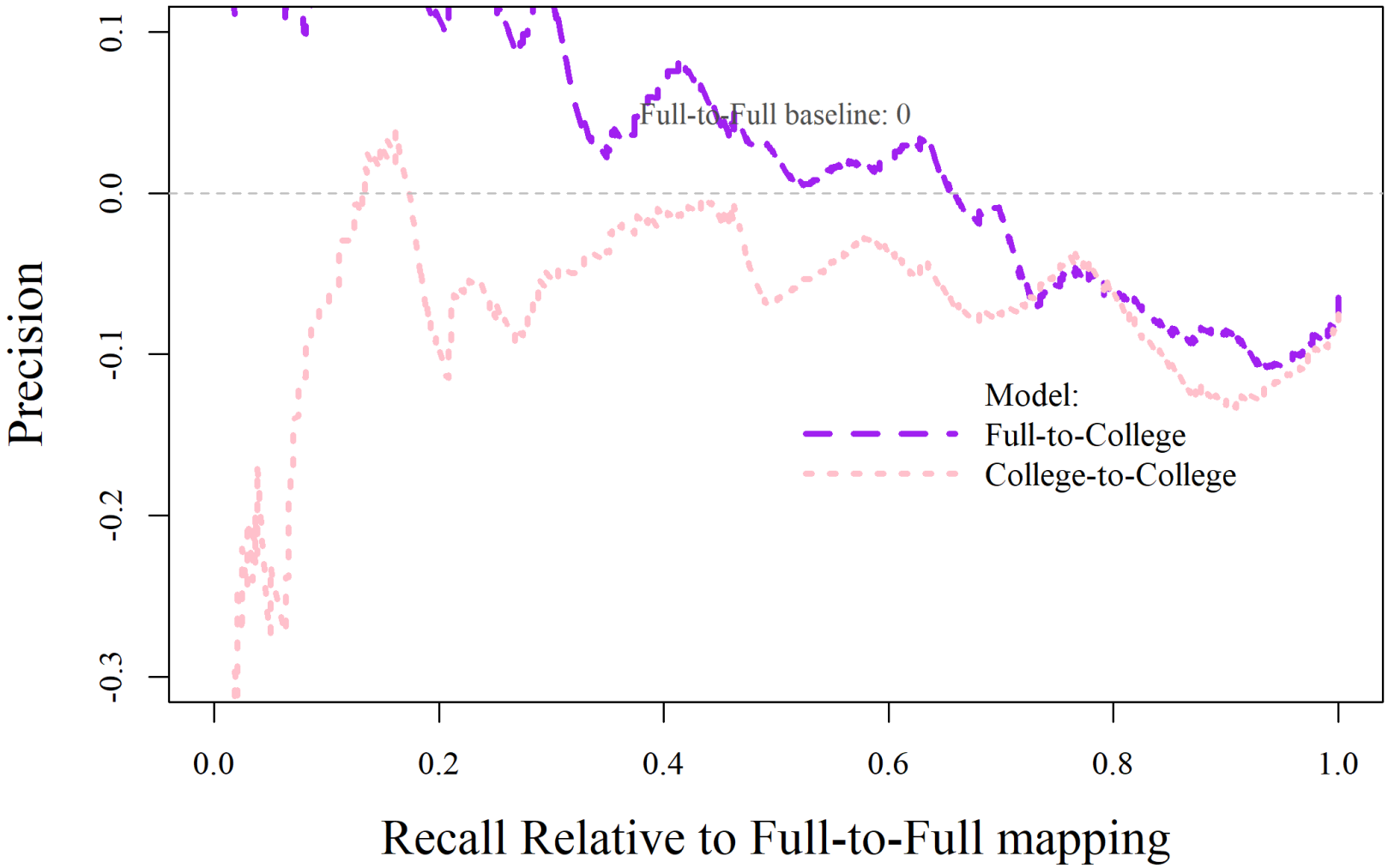
(3) $\log wage_i = \beta X_i + \beta_V Vmismatch_i + \beta_U Hundermatch_i + \beta_O Hovermatch_i + \beta_u Undocu_i + \varepsilon_i$

Results

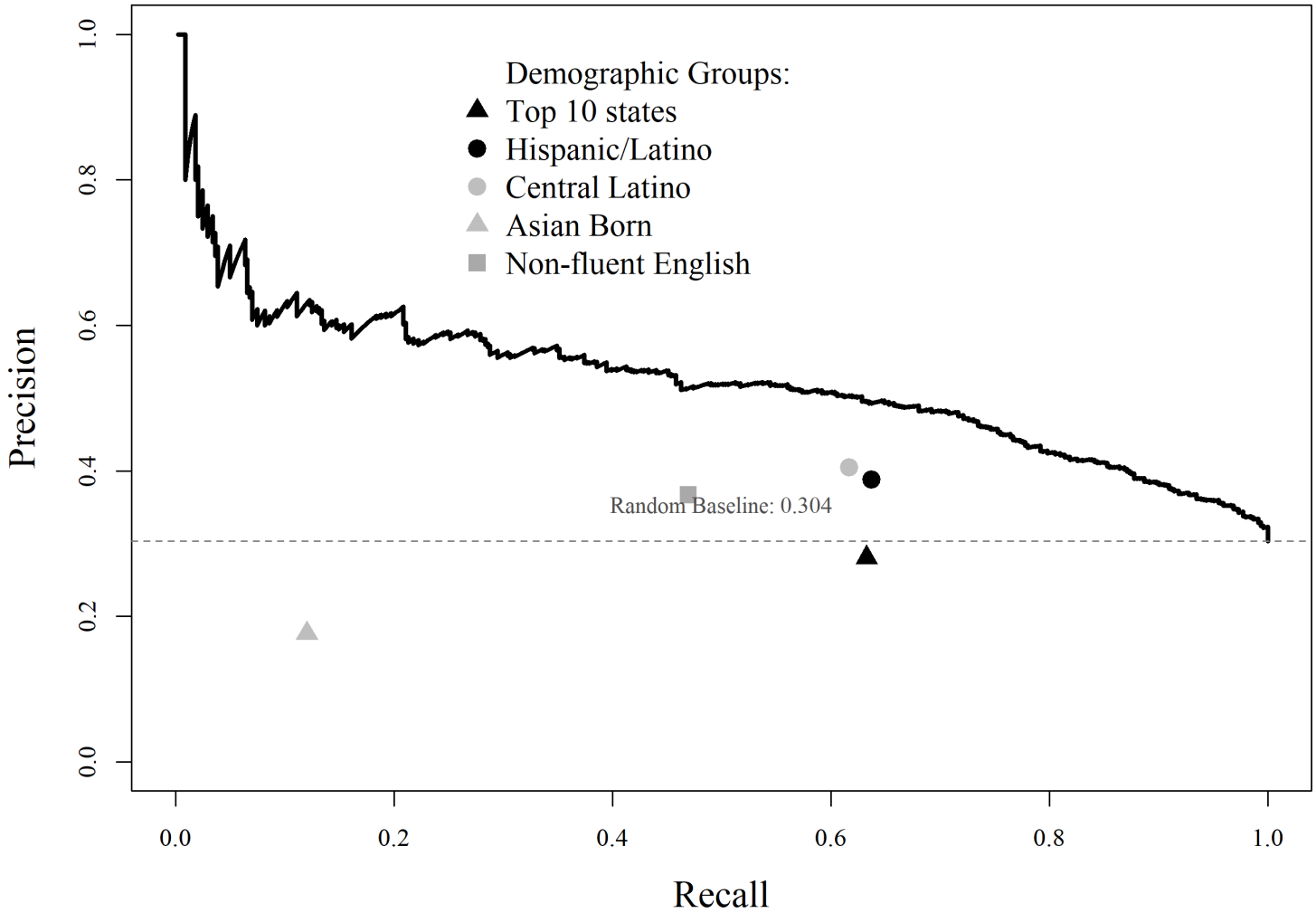
Algorithm Performance



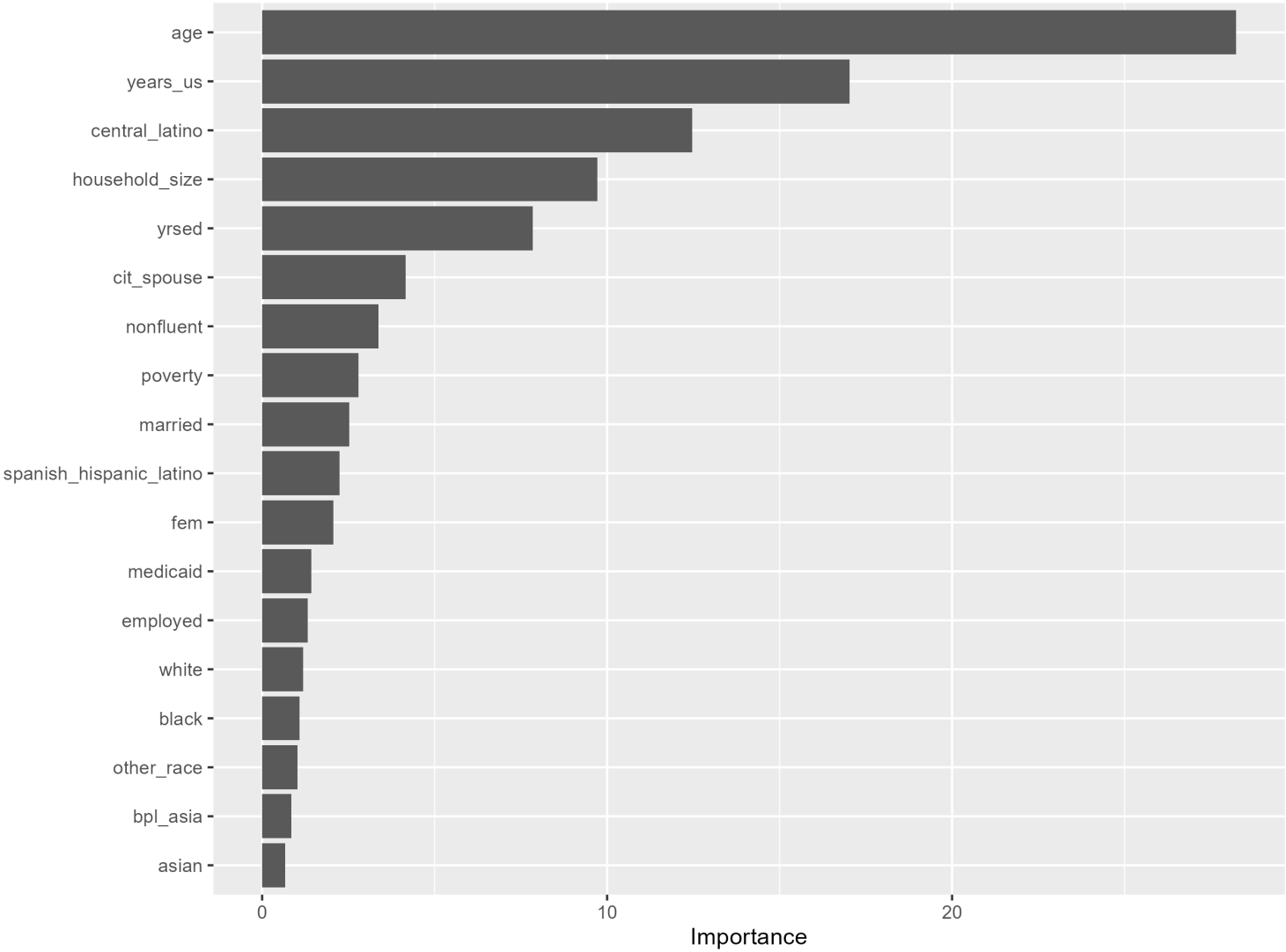
Varying training and testing samples



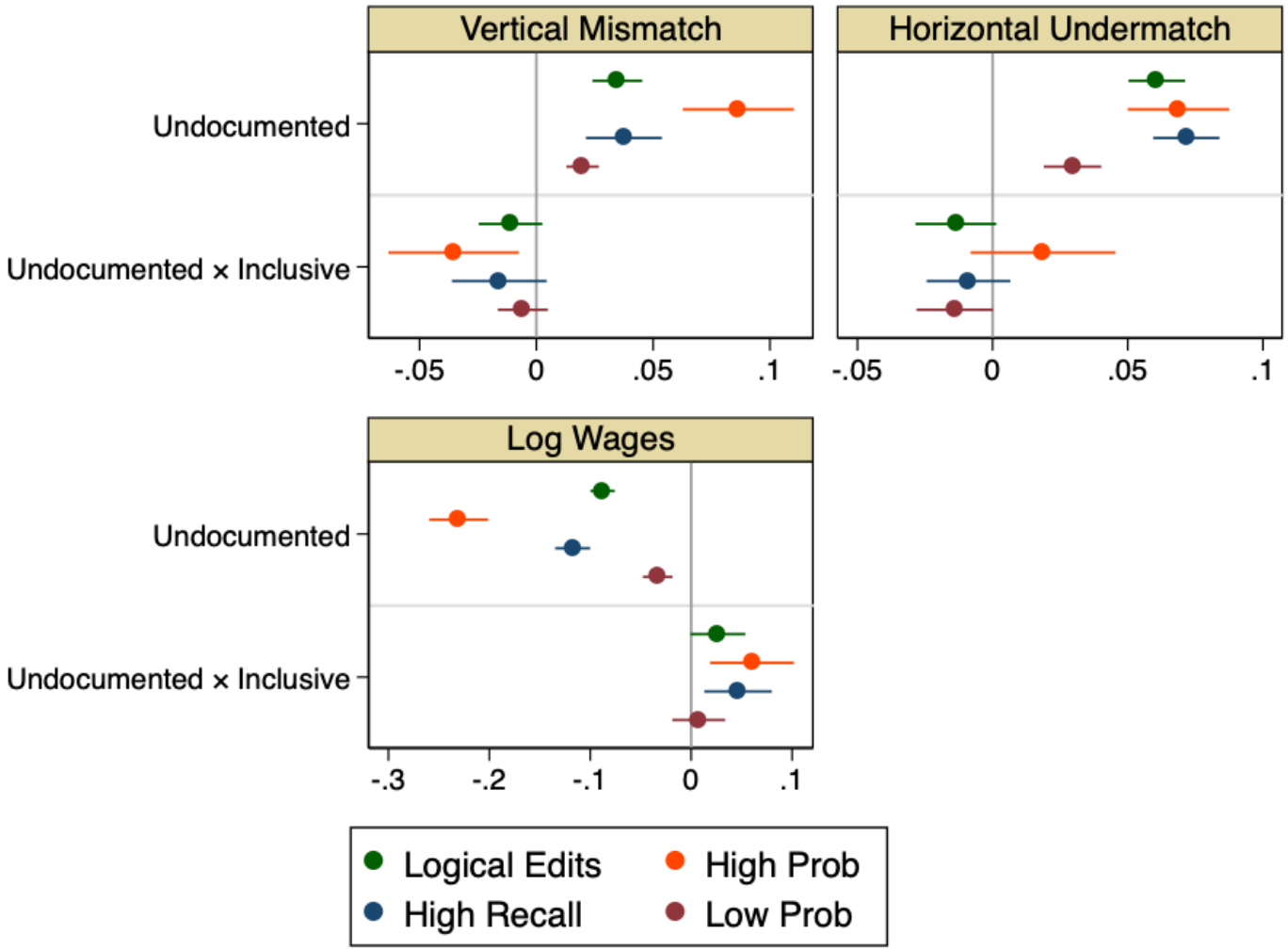
Logical edits with demographic subgroups



Feature importance (GBM)



Immigration Policy Climate (GBM; Samari Goleen, 2021)



Summary

- The Gradient-boosting trees algorithm has the best precision compared to previous imputation methods
- Undocumented college graduates experience education-occupation mismatch and wage penalties
- Suggestive evidence that inclusive policy climates help reduce mismatch rates and wage penalties

Conclusion

Contact information

Dr. Veronica Sovero: vsovero@ucr.edu
Mario Arce Acosta: maarceacosta@ucdavis.edu

HTTPS://GITHUB.COM/NEWTRINO0/UNDOCU_MISMATCH_WAGE_RESEARCH_2024

Acknowledgments

Thank you to the University of California, Riverside and the Academic Preparation, Recruitment & Outreach department for hosting this research in its infancy.