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

Dr. Veronica Sovero
Mario Arce Acosta

University of California, Riverside

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)

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

AMERICAN COMMUNITY SURVEY (ACS)

- **TARGET SAMPLE**
- Years: 2013-2019
- Age: 22-55
- Does **not** have direct measure of undocumented status

1: MACHINE/STATISTICAL LEARNING TRAINING (70% OF SAMPLE)

We train the following models:

- K-Nearest Neighbors (**KNN**)
- Random Forest algorithm (**RF**)

2: TESTING / EVALUATING MODEL PERFORMANCE (30% OF SAMPLE)

We evaluate: accuracy, sensitivity, specificity, and positive predictive value.

3: PREDICT ONTO ACS SAMPLE

We use the best models to impute 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.

K-Nearest Neighbors classifier

KNN

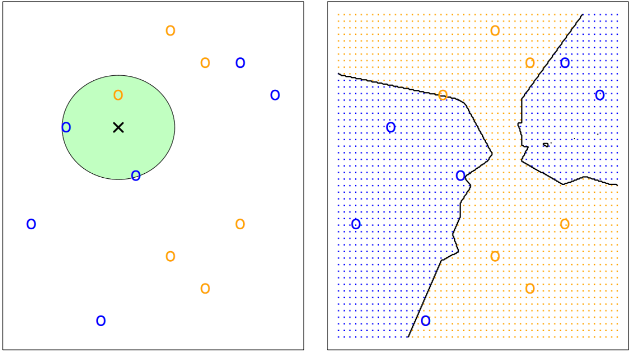
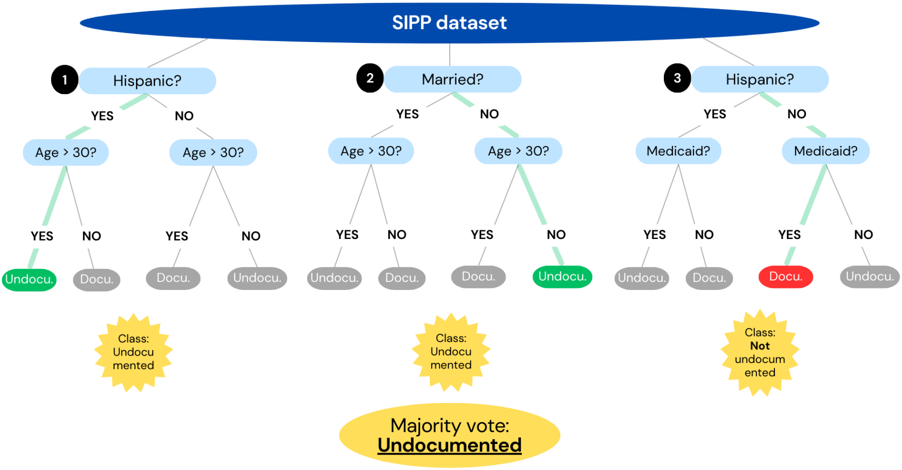


Figure: Source: *Intro to Statistical Learning with R* at https://www.statlearning.com/

Random Forest algorithm

RF



Model performance metrics

	Logical edits	Logistic	KNN	RF
Sensitivity	1.0000	0.6401	0.6424	0.6629
Specificity	0.9547	0.6614	0.6028	0.7190
Positive predictive value	0.2857	0.4518	0.4135	0.5070
Accuracy	0.9555	0.6549	0.6148	0.7019

Table: Imputation Methods Model Metrics

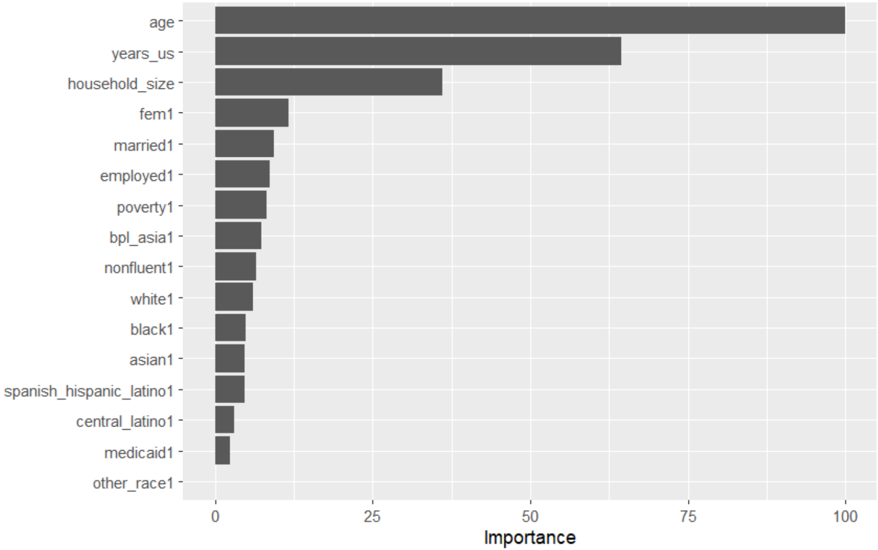
Sensitivity: $\frac{True\ Positives}{True\ Positives+False\ Negatives}$

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

Positive-predictive value: $\frac{True\ Positives}{True\ Positives+FalsePositives}$

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

Feature importance (RF model)



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

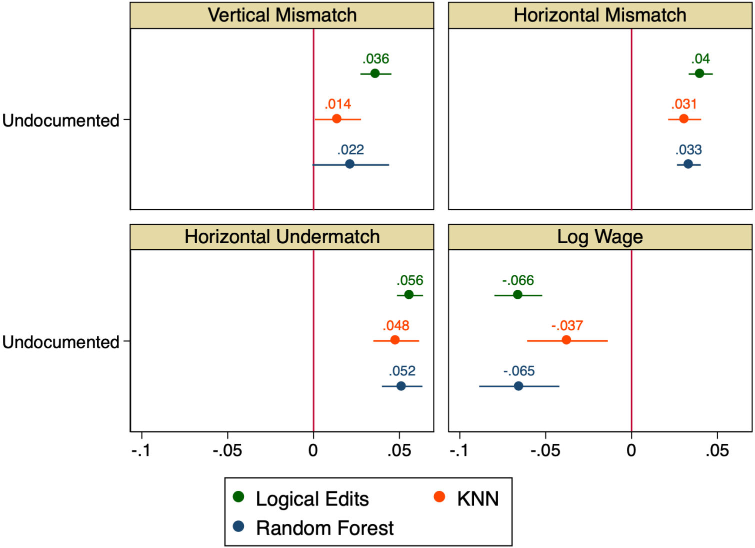
$$(1)\ Vertical\ Mismatch_i = \beta X_i + \beta_U Hundermatch_i + \beta_O Hovermatch_i + \beta_u Undocu_i + \varepsilon_i$$

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

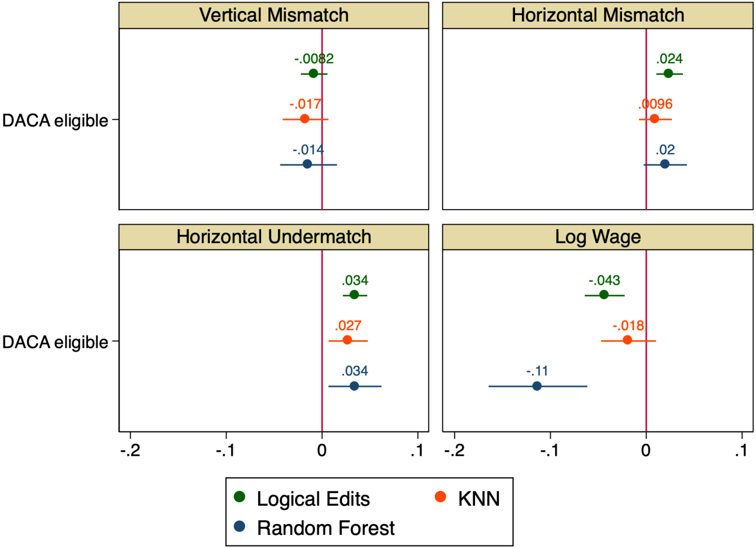
$$(3)\ \log wage_i = \beta X_i + \beta_V Vmismatch_i + \beta_U Hundermatch_i + \beta_O Hovermatch_i + \beta_u Undocu_i + \varepsilon_i$$

Results

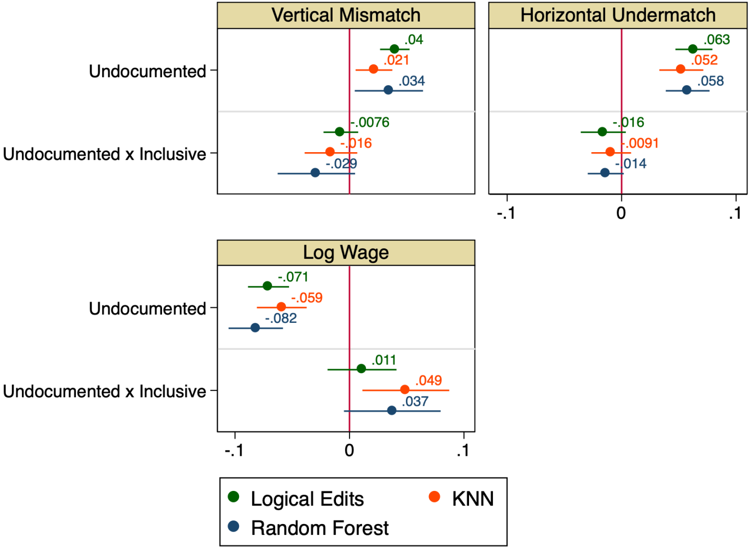
Main Results



DACA


Vertical Mismatch
Horizontal Mismatch
Horizontal Undermatch
Log Wage
DACA eligible
DACA eligible
-.0082
-.017
-.014
.024
.0096
.02
.034
.027
.034
-.043
-.018
-.11
-.2
-.1
0
.1
-.2
-.1
0
.1
Logical Edits
KNN
Random Forest

Immigration Policy Climate (Samari Goleen, 2021)



Summary

- RF algorithm has the best positive predictive value 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: mario.arce1615@gmail.com

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 making this trip possible!