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

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

[1]University of California, Riverside, Economics (vsovero@ucr.edu)

[2]University of California, Davis (maarceacosta@ucdavis.edu)

March 20, 2026

Abstract

This study estimates the extent of education-occupation mismatch and the associated wage penalties for undocumented college graduates. Using data from the American Community Survey (ACS), we classify workers as vertically mismatched (higher educational attainment than is typical for the occupation) or horizontally mismatched (field of degree is not typical for the occupation). Because the ACS does not identify undocumented status, we train a gradient boosting machine (GBM) model on the Survey of Income and Program Participation (SIPP) and use predicted probabilities to impute status in the ACS. This approach enables new analyses of labor market outcomes for undocumented college graduates in nationally representative surveys. Undocumented college graduates have higher rates of both vertical and horizontal mismatch and face a wage penalty of about 8 percent overall, rising to over 20 percent among those most likely undocumented. In STEM fields, mismatch differences are small, but a sizable wage penalty remains in the high-probability sample, indicating that pay disparities arise largely within occupations rather than from differences in job placement. Wage and mismatch penalties are smaller in states with inclusive immigrant policy climates, underscoring the role of institutional context.

Keywords: Undocumented; Education-occupation mismatch; Legal status; Labor; Wage; Policy; Income inequality

1 Introduction

A substantial number of undocumented immigrants in the United States have completed higher education - recent estimates suggest that roughly 1.7 million undocumented adults hold a college degree. However, relatively little is known about their labor market experiences. In this paper, we investigate whether college-educated undocumented workers are more vulnerable to education-occupation mismatch and wage penalties. Vertical mismatch occurs when workers hold more education than their occupation requires, while horizontal mismatch occurs when workers are employed outside the field of their degree. Both forms of mismatch are associated with lower wages and reduced occupational mobility (Li and Lu, 2023), making them important outcomes for assessing whether undocumented college graduates are able to translate their education into commensurate labor market returns.

We use data from the American Community Survey (ACS) for the years 2009-2019 to examine mismatch rates and associated wage penalties among undocumented college graduates. A central challenge is that large, nationally representative datasets such as the ACS do not identify undocumented status. A common approach in the literature has been logical imputation, which identifies likely undocumented individuals by eliminating cases where legal status can be confirmed (Warren, 2014; Bachmeier *et al.*, 2014). Although widely used, this method has low predictive accuracy for college graduates: our estimates indicate a precision of only about 28 percent in the second wave of the 2008 Survey of Income and Program Participation (SIPP). Another common strategy in the literature is to restrict analyses to subsamples defined by national origin - for example, Mexican- or Central American-born (e.g., Amuedo-Dorantes and Sparber, 2014). Yet undocumented college graduates differ demographically from the broader undocumented population. For example, nearly half are Asian (*Estimates of the Unauthorized Immigrant Population Residing in the United States* 2024), suggesting that these strategies may be less suitable for this subgroup.

To address these limitations, we develop a machine learning approach to classify undocumented status. We draw on the second wave of the 2008 Survey of Income and Program Participation (SIPP), which directly measures legal immigration status, to construct a training set (donor sample). We then apply the model predictions to the ACS, the target sample, where legal status is not observed. Following recent work that applies machine learning to impute legal status in survey data (Ruhnke *et al.*, 2022), we train supervised classifiers, including logistic classifiers, k-nearest neighbors, random forest, and gradient-boosting trees models (GBM) using demographic and socioeconomic predictors such as age, years in the United States, education, race/ethnicity, English proficiency, and employment status. Model performance is evaluated using standard metrics such as recall (sensitivity) and precision (positive predictive value). In validation tests, the gradient-boosting trees model has higher precision values for any given level of recall. This improvement enables us to apply the trained models to the 2009-2019 ACS sample of college graduates, pro-

viding the scale necessary to study education-occupation mismatch and wage outcomes among undocumented workers.

We evaluate classification performance using precision-recall curves, which summarize the trade-off between correctly identifying undocumented individuals (recall) and minimizing false positives (precision). Across models trained on the SIPP, the gradient boosting machine (GBM) consistently dominates the precision-recall frontier, indicating superior performance in distinguishing undocumented from documented respondents. The GBM achieves both higher precision at comparable levels of recall and higher recall at comparable levels of precision than the logistic classifier, random forest, and k-nearest neighbor models. This means that for any chosen threshold, the GBM model identifies a larger share of truly undocumented individuals while generating fewer false positives. Given this balanced improvement in both predictive accuracy and classification stability, we use GBM-based probabilities to impute undocumented status in the ACS and construct the high-probability, high-recall, and low-probability estimation samples.

Our regression results show that undocumented immigrants are more likely to be vertically mismatched and horizontally undermatched. Holding mismatch constant, we also estimate a wage penalty ranging from approximately three to twenty-one percentage points, depending on the undocumented imputation method.

We also examine how policy context shapes these disadvantages. At the state level, policies related to work authorization verification, occupational licensing, driver's licenses, immigration enforcement, and identification also moderate outcomes (Samari *et al.*, 2021). Our results show that undocumented college graduates in states with more inclusive immigrant policy climates face lower risks of mismatch and smaller wage penalties than those in more restrictive environments. Taken together, our results show that state policies can moderate the labor market disadvantages associated with undocumented status.

This study makes three contributions. First, it provides national estimates of vertical and horizontal mismatch and associated wage penalties for undocumented college graduates, a group that has received limited attention in prior work. Second, it advances measurement by training supervised classifiers on observed legal status in the SIPP to impute status in the ACS, improving on rule-based approaches and building on recent applications of machine learning in this area (Warren, 2014; Van Hook *et al.*, 2015; Ruhnke *et al.*, 2022; Cengiz *et al.*, 2022). Third, it situates these outcomes in policy context, implementing indicators in state-level immigration policy climates (Samari *et al.*, 2021).

In the sections that follow, we provide a brief overview of the related literature, then outline the methods for identifying the undocumented immigrant population and our measures of education-occupation mismatch. We present the results of our estimated regression models and conclude with a discussion of the limitations and avenues for future research.

2 Literature Review

Education-occupation mismatch is a well-documented feature of U.S. labor markets. Highly educated workers frequently experience both vertical and horizontal mismatch, with important implications for career trajectories (Li and Lu, 2023). Immigrants are more likely than natives to work in positions below their level of schooling, reflecting challenges in transferring foreign credentials and restrictions on occupational access (Ortega and Hsin, 2018). For undocumented immigrants, lack of work authorization further limits opportunities for appropriate job placement. Policies that expand legal access, such as DACA, have been shown to improve schooling and labor market outcomes, in part by reducing mismatch (Amuedo-Dorantes and Antman, 2017; Hsin and Ortega, 2018). More recent work highlights ongoing enrollment and labor market challenges for undocumented students even as policy contexts evolve (Kidder and Johnson, 2025). Yet the extent of mismatch among undocumented college graduates remains largely unexamined.

Beyond mismatch, undocumented immigrants face significant wage penalties. Research shows that lacking legal status lowers earnings even after accounting for education, experience, and other characteristics (Borjas and Cassidy, 2019). These penalties arise in part because undocumented workers are excluded from higher-paying occupations and remain concentrated in lower-wage sectors. Such findings underscore that wages reflect not only human capital but also the constraints imposed by legal status. However, little is known about how these penalties manifest for undocumented college graduates, whose formal qualifications suggest access to higher returns but whose lack of authorization may limit them to lower-quality matches.

The labor market opportunities of undocumented immigrants are also structured by policy context. In the United States, DACA provides temporary protection and work authorization, which research shows improves schooling and employment outcomes for eligible young adults (Amuedo-Dorantes and Antman, 2017; Hsin and Ortega, 2018; Kuka *et al.*, 2020). At the state level, policies governing occupational licensing, driver's licenses, and access to higher education create additional opportunities or restrictions (Chung, 2023; Cho, 2022; Amuedo-Dorantes and Arenas-Arroyo, 2020; Amuedo-Dorantes and Sparber, 2014). More restrictive enforcement regimes have been linked to negative consequences for immigrant families and communities (Amuedo-Dorantes and Arenas-Arroyo, 2019). Together, this work shows that policy variation conditions the extent to which undocumented immigrants are able to translate their education into occupational and wage gains.

Studying these outcomes requires reliable ways to identify undocumented immigrants in survey data. One widely used approach combines logical edits with demographic reweighting (Warren, 2014), and has been adopted in applied research estimating the size and characteristics of the undocumented population (e.g., Warren, 2014; Borjas and Cassidy, 2019). This method, however, tends to have low precision, leading to frequent misclassification of legally present immigrants as undocumented. Statistical imputation provides a different strategy, relying on regression-based or

cross-survey methods (Van Hook *et al.*, 2015). These methods can improve classification, but only under strict conditions that are difficult to satisfy in practice. Their limitations have motivated recent interest in machine learning approaches. In migration research, decision tree-based models have been used to impute undocumented status with improved accuracy (Ruhnke *et al.*, 2022), and in labor economics, machine learning has been applied to estimate heterogeneous effects of minimum wage increases, underscoring its broader potential in applied research (Cengiz *et al.*, 2022).

Building on this literature, our analysis estimates the extent of mismatch and wage penalties for undocumented college graduates. We also assess how these outcomes vary across federal and state policy contexts. Methodologically, we advance prior approaches by applying machine learning to improve the imputation of legal status in survey data.

3 Data and Methods

Our sample is derived from the 2009 to 2019 ACS surveys. We restrict the analysis to undocumented status in general, acknowledging that there are already many studies that examine the impact of DACA on schooling and labor market outcomes[1]. We exclude ACS data after the onset of the Covid-19 pandemic to ensure such a unique event does not affect our analysis. The estimation sample consists of respondents of prime working age adults (age 22 to 55 years old). We filtered our sample to only include observations with nonmissing and nonzero values for wage, occupation, degree field for college, and employment.

3.1 Vertical and horizontal mismatch

We measure vertical and horizontal mismatch using the methods described by Li and Lu (Li and Lu, 2023). We used US-born citizens to identify the typical wages and degree fields for each occupation[3]. Vertical mismatch occurs when a worker, a college graduate in our case, is employed in an occupation where the modal level of educational attainment does not match their educational attainment (Li and Lu, 2023).

Horizontal mismatch occurs when the worker's degree field during college does not match the

[1]The impact of DACA on schooling outcomes is still being studied today, with different conclusions being made about its effect on aspects such as school completion, enrollment, and attendance. One paper finds that DACA led to a decrease in the probability of enrolling in school for young, noncitizen adults[2] with at least a high school diploma or equivalent (Amuedo-Dorantes and Antman, 2017). It has also lead to an increase in the likelihood of employment for eligible individuals. Existing literature observing the impact of DACA finds dropout rates to have increased between 7.3 to 14.6 percentage points for DACA recipients at 4-year universities and full-time enrollment by DACA recipients to have increased between 5.5 to 11.0 percentage points at community colleges (Hsin and Ortega, 2018). As Amuedo-Dorantes and Antman and Kuka *et al.*, imply, DACA may increase the opportunity cost of attending college which DACA recipients may respond to by choosing to prioritize work over college.

[3]We also make edits to the occupation codes, resolving discrepancies in the different updates made to the code in the years 2010 and 2018. This leads to three different sets of occupational codes, which we converted to the coding starting 2010, before it was updated again in 2018. The different sets of codes spanned 2002-2009, 2010-2017, 2018-2022. We made this decision because the modification in 2010 created too many new occupation codes and categories, and this conversion required the least amount of assumptions on how a worker's occupational code is classified after the update.

two most common degree fields of workers within the occupation they are working in. Additionally, horizontal mismatch is then categorized into two types by creating two variables: horizontal undermatch and horizontal overmatch. A worker is considered horizontally undermatched if the median wage of workers with the same degree field is less than the median wage of horizontally matched workers in their occupation. Respectively, a worker is considered horizontally overmatched if the median wage of workers with the same degree field is more than the median wage of horizontally matched workers in their occupation.[4]

3.2 IPC Index

To capture the underlying police climate towards immigrants, we use the Immigrant Policy Climate Index (Samari *et al.*, 2021) which we dub IPC, and tracks state-level policies over time across five domains: access to public health benefits, higher education, labor and employment, driver's licenses and identification, and immigration enforcement. The fourteen policies are coded either as inclusive towards the immigrant population, neutral, or exclusive. We create an indicator variable for whether a state is net inclusive across the fourteen policies. Because the IPC index includes policy domains that may not be directly relevant for occupational matching (for example access to health insurance and higher education), we also create indicator variables for a few select policies: whether the professional licensure is available for undocumented individuals, whether E-Verify is prohibited, whether the state allows undocumented immigrants to obtain driver's licenses, and whether the state cooperates with federal immigration enforcement.

3.3 Undocumented Status Imputation

To impute undocumented status, we begin by training and evaluating a handful of machine learning algorithms. The general approach is as follows: we construct a donor sample that contains information on a respondent's legal status to train the model, then apply the model predictions to the target sample (the ACS data). Our donor sample is Wave 2 of the Survey of Income and Program Participation (SIPP). SIPP is one of the only nationally representative surveys to directly measure immigrants' legal status. Respondents were asked their immigration status upon entering the United States and whether they have adjusted their immigration status since their initial entry. Immigrant respondents who listed their immigration status as "Other" instead of "Permanent resident" and who had responded "No" to whether they adjusted their immigration status were taken to be truly undocumented immigrants.[5] This will serve as the classification variable for model

[4]There will be a small group of worker that are horizontally mismatched but are neither undermatched nor overmatched due to the median wage for their degree field matching the median wage of horizontally matched workers for their occupation. The fact that wages within the ACS data extract are reported in fixed amounts after rounding also contributes to this.

[5]This assignment of status holds limitations of self-reporting and lack granularity. Undocumented status assigned in this manner overlooks those that have work authorization through other means: namely visas and DACA recipient status.

training[6]. Prior work has shown that self-reported legal status in the SIPP is generally reliable for analytic use (Bachmeier *et al.*, 2014).

We create a training dataset of the possibly undocumented using logical imputation methods. We apply the following criteria:

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

Immigrant respondents that met any of the above conditions were assigned documented status, and the remainder are classified as possibly undocumented. We also apply the same age and education restrictions as the target ACS sample (college educated, ages 22-55).

Additionally, we utilized the method for filtering out H-1B immigrants (Borjas and Cassidy, 2019) to improve the accuracy of the logical edits imputation method in ACS. [7] We deviate from Borjas and Cassidy by broadening our occupations associated with H-1B beneficiaries, accordingly to the 2021 DHS and USCIS report (*Characteristics of H 1B Specialty Occupation Workers* 2024) by including a much longer list of occupations based on the shorter list found in the USCIS report. We included the top 7 detailed occupations excluding Other Occupations: Occupations in Systems Analysis and Programming, Computer-Related Occupations, Electrical/Electronics Engineering Occupations, Occupations in College and University Education, Occupations in Architecture, Engineering, and Surveying, Accountants, Auditors, and Related Occupations, Occupations in Administrative Specializations. These occupations made up to 80.7% of approved H-1B petitions.[8]

Following in the steps of Ruhnke *et al.*, we estimate logistic classifier, k-nearest neighbors (KNN) and random forest (RF) machine learning models on the possibly undocumented sample predominantly using the caret package in R. Inspired by Cengiz *et al.*, an additional model - gradient-boosting decision trees - is incorporated and displays more favorable precision-recall tradeoffs.

KNN identifies K observations with respect to their distance from a test observation and, using the conditional probability of points belonging to a class within the proximity of the test observation, classifies the test observation as the class with the largest probability within the proximity. Tree-based machine learning models are based on recursive binary splitting to grow

[6] In principle, SIPP contains the data needed to estimate education occupation mismatch, but the main limitation is sample size.

[7] Due to the discrepancies between occupational coding in the ACS and the SIPP, we refrain from applying this same logical edits filter of H-1B status in the SIPP.

[8] We also make edits to the occupation codes, resolving discrepancies in the different updates made to the code in the years 2010 and 2018. This leads to three different sets of occupational codes, which we converted to the coding starting 2010, before it was updated again in 2018. We made this decision because the modification in 2010 created too many new occupation codes and categories, and this conversion required the least amount of assumptions on how a worker's occupational code is classified after the update.

trees, where an observation is classified based on the most commonly occurring class. The RF algorithm grows n decision trees and, when a tree is split or grown, m predictors are randomly taken from our entire set of predictors of a class. This provides the RF algorithm an advantage in classifying observations over other techniques because of the large number of trees which reduces the variance of predictions, and the subset of predictors which decreases correlation between the trees (James *et al.*, 2023). Gradient-boosting trees is a machine learning algorithm, that consists of many "weak" learners to construct a "strong" learner. These "weak" learners are individual trees, much like in the case of the random forest algorithm, and each subsequent "weak" learner is used to construct a "strong" model that is ultimately the final accumulation of all "weak" learners, or trees. This is a recursive process, where for each step involving a "weak" learner, there is a ""strong" contemporary model that is an ensemble of the previous "weak" learners.

For an in-depth explanation of the hyperparameter values (such as shrinkage and tree depth) used for tuning the gradient-boosting trees model (GBM), we defer to Cengiz *et al.*, We currently set hyperparameter values that are identical to Cengiz *et al.*, when training the GBM, though future work would involve testing different hyperparameter values.

We split the SIPP into a training and test sample (70/30 split), where the test sample is used to evaluate model performance based on commonly used machine learning metrics. Within the training datasets for the logistic, KNN, and random forest models, we use the k-fold cross validation procedure, which splits the training data into folds and then trains the model on the k-1 folds. This is repeated k times, and the results are averaged. We follow standard practices and select 10 folds. Additionally, we up-sample the truly undocumented group for the random forest model to address class imbalance (only around 25 percent of the possibly undocumented sample is actually undocumented). Since the gradient-boosting trees model is our newest imputation method and displays great precision rates, we have currently shelved the idea of up-sampling for the gradient-boosting trees training sample.

We use the following as predictors of legal status: age, years in the United States, gender, race/ethnicity, region of birth, English fluency, marital status, years of education, poverty status, Medicaid receipt status, and employment status. After training our machine learning models, we apply the model predictions to the target sample of the possibly undocumented in the ACS data.

3.4 Econometric Model

We build regression models to predict vertical mismatch, horizontal undermatch, and log wage. Our first is the mismatch regression model:

$$Vmismatch_i = \alpha X_i + \alpha_U Hundermatch_i + \alpha_O Hovermatch_i + \alpha_u Undocu_i + \varepsilon_i$$

$$Hundermatch_i = \alpha X_i + \alpha_V Vmismatch_i + \alpha_u Undocu_i + \varepsilon_i$$

where X_i represents a vector of socioeconomic factors believed to be correlated with a worker's earnings such as: age, age squared, gender, race and ethnicity, nativity to the United States, immigrated before 10 years of age, metropolitan residence, degree field (five categories), years of education, and class of worker (government or not). $Vmismatch_i$ equals 1 if a worker is vertically mismatched; $Hundermatch_i$ equals 1 if a worker is horizontally undermatched; $Hovermatch_i$ equals 1 if a worker is horizontally overmatched.

Our current econometric model specification consists of linear probability models for predicting different dimensions of mismatch. Following "Education–Occupation Mismatch and Nativity Inequality Among Highly Educated U.S. Workers", the vertical mismatch indicator is included as a control for Horizontal undermatch and overmatch; similarly, Horizontal undermatch and overmatch are included as controls in the vertical mismatch model. Because each model controls for the other dimension of mismatch, the coefficients represent the association of undocumented status with one form of misallocation conditional on the other. The coefficient of interest is for $Undocu_i$, an indicator of whether a worker is undocumented.

Our log hourly wage model is specified as:

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

where w_i represents the hourly wage of each individual worker i. β_V, β_U, and β_O correspond to the wage penalties of a worker that is either vertically mismatched, horizontally undermatched, or horizontally overmatched, respectively. $Undocu_i$ is an indicator of whether a worker is undocumented, and β_u is its associated wage penalty.

Both models include U.S. state-by-year fixed effects. This is to control for policy reforms that vary by state and over time. Other factors such as minimum wage, earnings, and opportunities for human capital investment also vary by state.

4 Results

4.1 Machine Learning Imputation Results

When evaluating the performance of a machine learning model, we look at: the rate at which truly undocumented people were correctly classified as undocumented (recall, or sensitivity), the rate at which truly documented people were classified as documented (specificity), the rate at which those classified as undocumented were truly undocumented (precision, positive predictive value), and the

rate at which truly undocumented and documented people were correctly classified (accuracy)[9]. Precision-recall curves are displayed in Figure 1, where each machine learning algorithm is evaluated and compared against each for its performance on the SIPP test sample (not used in the model training). Given that our positive cases are those that are truly undocumented, we aim for an imputation method with the best precision-recall tradeoff outside extreme recall ranges. That is, we value the metric of recall over precision. This ensures we prioritize correctly classifying the few true positive cases (undocumented college graduates) out of all true positive cases, instead of prioritizing the rate of cases that were correctly classified as positive out of all cases classified as positive. An emphasis on recall means we correctly classify as many true positives as a ratio to all true positive cases, rather than as a ratio to all positive classifications themselves.

Following this emphasis on recall, we selected the gradient-boosting trees model (GBM) to compare performance across different training and test samples. We mapped training samples consisting of our full sample and college graduate sample, onto full (referred to as Full-to-Full) and college graduate (referred to as Full-to-College and College-to-College) test samples as seen in Figure 3. Additionally, we used our best variation of these GBMs to compare performance across demographic groups in Figure 5.

For the logical edits imputation, the recall is 1 because by definition, everyone in the test data sample is possibly undocumented. We can also see that the precision is quite low: only 28 percent of the possibly undocumented are actually undocumented. The precision-recall curves also includes a random baseline, shown as a horizontal reference line equal to the share of undocumented individuals in the test data. This baseline represents the precision that would be achieved by randomly guessing - classifying individuals as undocumented without using any predictors. The GBM curve remains well above the random baseline across the full range of recall values, confirming that the model provides substantial improvement over random classification. After exploring additional machine learning techniques, we learned that the gradient-boosting trees model (GBM) has higher precision-recall values than the commonly used sample restrictions - restricting to Hispanic/Latino workers for example - used to isolate the truly undocumented (Figure 5). This includes restricting the sample to the top 10 states where undocumented immigrants reside, restricting to individuals of Hispanic ethnicity, and restricting to individuals born in Mexico/Central America.

The feature importance of each predictor in the gradient-boosting trees models are reported in Figure7. A larger number signifies a greater contribution in reducing the model's prediction error. Age and years in the United States are the most highly ranked predictors, followed by years in the United States. This means that for every split in the trees for this method, the predictors of age

[9]
Specificity: $\frac{TN}{TN+FP}$
Recall (Sensitivity): $\frac{TP}{TP+FN}$
Precision (Positive predictive value): $\frac{TP}{TP+FP}$
Accuracy: $\frac{TP+TN}{total}$

and years in the United States were the leading variables that reduced prediction error the most and improved the model the most.

4.2 ACS Descriptive Statistics by Probability Quartile

In Table 2, we compare the descriptive statistics of the ACS sample across probability quartiles created from the gradient-boosting trees model predictions. These quartiles are based on the predicted probabilities of being undocumented for the GBM model trained on the full sample and also evaluated with the test sample. Quartile 1 represents the lowest predicted probability of undocumented status and Quartile 4 the highest, with thresholds defined using the SIPP calibration sample.

As predicted likelihood increases, the share Hispanic rises from 12 percent in the lowest quartile to 42 percent in the highest, while the share Asian declines from 48 to 39 percent. High-probability individuals are younger on average (about 30 years compared with 42 years in the lowest quartile), less likely to be married, and report shorter durations in the United States - approximately 4 years compared to 14 years for the low-probability group. They also have lower rates of English proficiency. Differences by field of study are also evident. The high-probability group is somewhat more likely to hold STEM degrees and slightly less likely to major in business or education.

Table 1 compares the estimation samples used in the regression analysis: the logical-edits baseline and the high-recall, high-probability, and low-probability groups derived from the GBM predictions. Our high-recall group is current defined by the probability threshold that classifies at least 75 percent of the truly undocumented college graduates in SIPP as undocumented. Each sample in Appendix Table 1 represents a different position along the precision-recall frontier reported in Figure 1. The logical-edits sample achieves full recall by construction but has low precision (0.29 in the SIPP validation sample). The GBM model improves classification substantially: the high-recall sample, defined by the threshold that captures roughly 75 percent of truly undocumented respondents, attains precision around 0.39; the high-probability sample, restricted to the top quartile of predicted probabilities, reaches precision of about 0.51. The low-probability group, drawn from the bottom quartile, has precision near 0.10 and serves as a placebo sample composed primarily of legally present immigrants.

Across these estimation samples, demographic characteristics and mismatch rates vary systematically with predicted undocumented probability. The high-probability group is younger, has fewer years in the United States, and experiences the highest rates of both vertical and horizontal mismatch (39 percent and 65 percent, respectively), while the low-probability group closely resembles the broader foreign-born college-educated population. The logical-edits and high-recall samples fall between these extremes, offering alternative estimation bounds that balance coverage against precision.

4.3 Mismatch Regression Results

The descriptive patterns presented above provide suggestive evidence of differences in labor-market outcomes by predicted legal status. We now turn to the regression analysis, which examines these relationships more formally. We estimate models of vertical and horizontal mismatch by undocumented status for the high-probability, high-recall, and low-probability groups.

Table 4 presents results for vertical mismatch. Relative to the logical-edits baseline (Column 1), the GBM model yields larger coefficients when undocumented status is restricted to the high-probability group, indicating a stronger association between undocumented status and employment in occupations requiring less education. Undocumented workers in the high-probability group are about 7.5 percentage points more likely to be vertically mismatched, compared with 3.0 percentage points using logical edits and 3.1 percentage points in the high-recall group. The low-probability group also shows a smaller but statistically significant coefficient (1.8 percentage points). The clearer separation across groups suggests that the GBM model improves classification, though some overlap remains. Additionally, the estimates may also reflect unobserved characteristics of those classified as possibly undocumented, which may contribute to higher baseline undermatch rates even apart from immigration status.

Similarly, the likelihood of being horizontally undermatched is greater for undocumented immigrants, where horizontal undermatch occurs when an individual is employed in an occupation that is lower paying compared to what is typical for a horizontally matched worker with the same degree (Table 5). Consistent with the vertical mismatch results, the coefficients are largest for the high-probability group, suggesting that undocumented status is associated with a greater likelihood of working in lower-paying occupations relative to one's field of study. Undocumented workers in the high-probability group are about 7.6 percentage points more likely to be horizontally undermatched, compared with 5.6 percentage points using logical edits and 6.8 percentage points in the high-recall group. The low-probability group also shows a smaller but statistically significant coefficient (2.5 percentage points).

The larger overall magnitudes relative to the vertical mismatch results reflect that horizontal undermatch captures a broader set of labor-market frictions. Vertical mismatch captures a narrower form of misallocation, namely whether a worker's formal education exceeds what is typical for their occupation. Horizontal undermatch, by contrast, captures a wider range of wage-related frictions within fields of study including sorting into lower-paying industries or employer segments. Even when undocumented workers obtain positions aligned with their educational level, these constraints can limit earnings and advancement. Taken together, the two dimensions indicate that undocumented college graduates face both limited access to higher-skill occupations and weaker returns within their areas of study.

4.4 Wage Regression Results

The mismatch regression results suggest that undocumented college graduates face both limited access to higher-skill occupations and weaker returns within their areas of study. These patterns carry through to wage outcomes. As shown in Table 6, the wage regressions estimate the wage penalty associated with undocumented status after accounting for both vertical and horizontal mismatch. Using logical edits, undocumented status is associated with a 7.7 percent reduction in hourly wages, compared with 10 percent for the high-recall group and 21 percent for the high-probability group. The larger penalty for the high-probability group suggests that the GBM model isolates workers facing the strongest legal and institutional constraints in the labor market. The smaller penalty for the low-probability group (around 3 percent) reinforces this interpretation, indicating that most of the wage disadvantage is concentrated among those most likely to lack legal status. By controlling for vertical and horizontal mismatch, the results suggest that additional labor market barriers beyond occupational sorting contribute to lower wages for undocumented immigrants.

4.5 Degree Interactions

We next examine whether undocumented college graduates face greater education-occupation mismatch and wage penalties in certain fields of study. To do do, we interact undocumented status with the five degree field categories: STEM, STEM- related, Business, Education, and other majors (reference category in the regression tables). The coefficients are presented graphically in Figure 6 (regression tables presented in the appendix).

Among STEM majors, undocumented workers show limited evidence of occupational mismatch but a substantial wage penalty concentrated in the high-probability sample. In the logical-edits and high-recall groups, differences in vertical and horizontal mismatch are small (2 to 3 percentage points) and wage gaps are modest and statistically insignificant. In contrast, in the high-probability sample, the wage penalty rises sharply to about 20 percent, even though mismatch rates remain similar to those of documented STEM workers. This pattern suggests that for the STEM graduates most likely to be undocumented, wage differences are not driven by education-occupation mismatch but instead occur within occupations.

Among business majors, undocumented college graduates exhibit higher rates of vertical mismatch. Depending on the imputation method, rates of vertical mismatch range from 8 to 12 percentage points relative to the other major category. Interestingly, there are lower rates of horizontal undermatch for business majors relative to other fields. Holding the educational requirements (vertical mismatch) of their occupation constant, undocumented business graduates are 6 to 20 percentage points less likely to be in occupations that typically pay less for business majors. The wage regressions, which control for both mismatch dimensions, show a remaining

earnings penalty of about 13 percentage points in the high-probability sample. Taken together, these results indicate that undocumented business graduates are more likely to be employed below their educational level and, conditional on occupation, earn less than comparable documented workers.

4.6 Policy Climate Interactions

The results by degree field highlight variation in how undocumented status relates to job placement and earnings across educational specializations. Some of this heterogeneity likely reflects field-specific exposure to occupational licensing and credentialing barriers, which are shaped in part by state-level policy regimes. To examine how these institutional contexts condition the relationship between undocumented status and labor-market outcomes, we interact undocumented status with an indicator for states that have net inclusive immigration policies (IPC index). The indicator for *net inclusive* immigrant policy climate equals one for states whose overall policy index is net positive and zero otherwise. Results are presented in Figure 6 and the corresponding regression tables are presented in the Appendix.

Across outcomes, the IPC interactions consistently attenuate penalties associated with undocumented status, though the magnitudes differ. For vertical mismatch, residing in an inclusive state is associated with a roughly 2-3 percentage point smaller undocumented-documented gap in the high-probability sample. The corresponding interaction for horizontal undermatch is smaller (1-2 percentage points) and less precisely estimated. The relatively modest IPC effects for mismatch outcomes likely reflect the persistence of educational and credentialing barriers that state policy can only partially offset. The moderating effect of an inclusive state policy climate is substantially larger for wages. In the high-probability sample, undocumented workers in inclusive states experience a 5-6 percent smaller wage penalty compared to those in restrictive states. The weaker moderation for mismatch and stronger moderation for wages imply that state policy may be less effective at altering where undocumented college graduates work but more effective at shaping pay and job quality once they are employed.

As with earlier results, the IPC effects are strongest in the high-probability sample, smaller in the high-recall sample, and negligible in the low-probability (placebo) sample. This gradient reinforces that the observed moderation is concentrated among workers most likely to be undocumented, providing validation of the classification model.

Because the IPC includes policy domains that may not be directly relevant for occupational matching (for example access to health insurance and higher education), we also interact undocumented status with a few select policies: whether the professional licensure is available for undocumented individuals, whether E-verify is prohibited, whether the state allows undocumented immigrants to obtain driver's licenses. Results are presented graphically in Figure 6 and the

corresponding regression tables are presented in the Appendix.

The estimates are noisy and often move in unexpected directions across samples. In several cases, coefficients are similar in magnitude for the high- and low-probability groups or even larger for the latter, suggesting limited precision in identifying heterogeneous effects by policy type. While some interactions point toward smaller penalties in inclusive policy settings, the lack of consistent sign and significance indicates that no single policy dimension drives the moderation observed in the composite IPC measure. The overall state policy climate, rather than individual policy components, appears to more reliably capture differences in labor-market outcomes for undocumented college graduates.

5 Discussion/Conclusion and Future Work

This paper synthesizes methods and existing literature to provide insight on the educational mismatch and wage penalties of undocumented college graduates. We examine labor market penalties through the lens of education, policies, and immigration status using improved statistical imputation methods of undocumented status on large public surveys.

Although we find the machine learning algorithms have greater positive predictive value than logical imputation, there are a few limitations that need to be acknowledged and directions for future work. There are two conditions discussed by (Van Hook *et al.*, 2015) that the imputation method's bias will depend on: joint observation and same universe. Future iterations of the imputation methods will include mismatch and wages in the training data in order to satisfy the joint observation condition. While the SIPP and ACS are national surveys within the United States, we exercise caution and recognize that the characteristics of immigrant since 2008 may have changed enough to affect our imputation method on data taken from a survey during the years 2009-2019.

Given that the demographics of the undocumented samples are slightly different based on the specific imputation method, it is likely that more model tuning is needed to better match the demographics of the undocumented college graduate population. Additionally, if there is any systematic sorting of undocumented immigrants across degree fields, the model should be also be trained on degree field. We also wish to explore other algorithms such as the gradient-boosting trees model, which has been shown to perform well in classifying minimum wage workers ((Cengiz *et al.*, 2022)).

References

Amuedo-Dorantes, Catalina and Antman, Francisca (2017). "Schooling and labor market effects of temporary authorization: evidence from DACA", *Journal of Population Economics*, pp. 339–373. ISSN: 0933-1433, 1432-1475.

Amuedo-Dorantes, Catalina and Arenas-Arroyo, Esther (2020). "Labor market impacts of states issuing of driver's licenses to undocumented immigrants", *Labour Economics*, ISSN: 09275371.

Amuedo-Dorantes, Catalina and Arenas-Arroyo, Esther (2019). "Immigration Enforcement and Children's Living Arrangements", *Journal of Policy Analysis and Management*, pp. 11–40. ISSN: 0276-8739, 1520-6688.

Amuedo-Dorantes, Catalina and Sparber, Chad (2014). "In-state tuition for undocumented immigrants and its impact on college enrollment, tuition costs, student financial aid, and indebtedness", *Regional Science and Urban Economics*, pp. 11–24. ISSN: 01660462.

Bachmeier, James D., Van Hook, Jennifer, and Bean, Frank D. (2014). "Can We Measure Immigrants' Legal Status? Lessons from Two U.S. Surveys", *International Migration Review*, Vol. 48 No. 2. Compares imputation strategies including Warren's logical edits with reweighting., pp. 538–566. DOI: 10.1111/imre.12059.

Borjas, George J. and Cassidy, Hugh (2019). "The wage penalty to undocumented immigration", *Labour Economics*, p. 101757. ISSN: 09275371.

Cengiz, Doruk *et al.*, (2022). "Seeing beyond the Trees: Using Machine Learning to Estimate the Impact of Minimum Wages on Labor Market Outcomes", *Journal of Labor Economics*, Vol. 40 No. S1, S203–S247. DOI: 10.1086/718497. eprint: https://doi.org/10.1086/718497. **available at**: https://doi.org/10.1086/718497.

"Characteristics of H 1B Specialty Occupation Workers", (2024). **available at**: https://www.uscis.gov/sites/default/files/document/data/H1B_Characteristics_Congressional_Report_FY2021-3.2.22.pdf (accessed 25 Sept. 2024).

Cho, Heepyung (2022). "Driver's license reforms and job accessibility among undocumented immigrants", *Labour Economics*, ISSN: 09275371.

Chung, Bobby W (2023). "Effects of Occupational License Access on Undocumented Immigrants Evidence from the California Reform",

"Estimates of the Unauthorized Immigrant Population Residing in the United States", (2024). **available at**: https://ohss.dhs.gov/sites/default/files/2024-06/2024_0418_ohss_estimates-of-the-unauthorized-immigrant-population-residing-in-the-united-states-january-2018%25E2%2580%2593january-2022.pdf (accessed 23 Sept. 2024).

Hsin, Amy and Ortega, Francesc (2018). "The Effects of Deferred Action for Childhood Arrivals on the Educational Outcomes of Undocumented Students", *Demography*, pp. 1487–1506. ISSN: 0070-3370, 1533-7790.

James, Gareth *et al.*, (2023). *Intro to Stat. Learning with R*,

Kidder, William C. and Johnson, Kevin R. (2025). "California Dreamin': DACA's Decline and Undocumented College Student Enrollment in the Golden State", *Journal of College University Law*,

Kuka, Elira, Shenhav, Na'ama, and Shih, Kevin (2020). "Do Human Capital Decisions Respond to the Returns to Education? Evidence from DACA", *American Economic Journal: Economic Policy*, pp. 293–324. ISSN: 1945-7731, 1945-774X.

Li, Xiaoguang and Lu, Yao (2023). "Education–Occupation Mismatch and Nativity Inequality Among Highly Educated U.S. Workers", *Demography*, pp. 201–226. ISSN: 0070-3370, 1533-7790.

Ortega, Francesc and Hsin, Amy (2018). "Occupational Barriers and the Labor Market Penalty from Lack of Legal Status", *SSRN Electronic Journal*, ISSN: 1556-5068.

Ruhnke, Simon A., Wilson, Fernando A., and Stimpson, Jim P. (2022). "Using machine learning to impute legal status of immigrants in the National Health Interview Survey", *MethodsX*, ISSN: 22150161.

Samari, Goleen, Nagle, Amanda, and Coleman-Minahan, Kate (2021). "Measuring structural xenophobia: US State immigration policy climates over ten years", *SSM - Population Health*, ISSN: 23528273.

Van Hook, Jennifer *et al.*, (2015). "Can We Spin Straw Into Gold? An Evaluation of Immigrant Legal Status Imputation Approaches", *Demography*, pp. 329–354. ISSN: 0070-3370, 1533-7790.

Warren, Robert (2014). "Democratizing Data about Unauthorized Residents in the United States: Estimates and Public-Use Data, 2010 to 2013", *Journal on Migration and Human Security*,

6 Figures

Figure 1: Precision Recall curves for Full Training and Test sample, different models

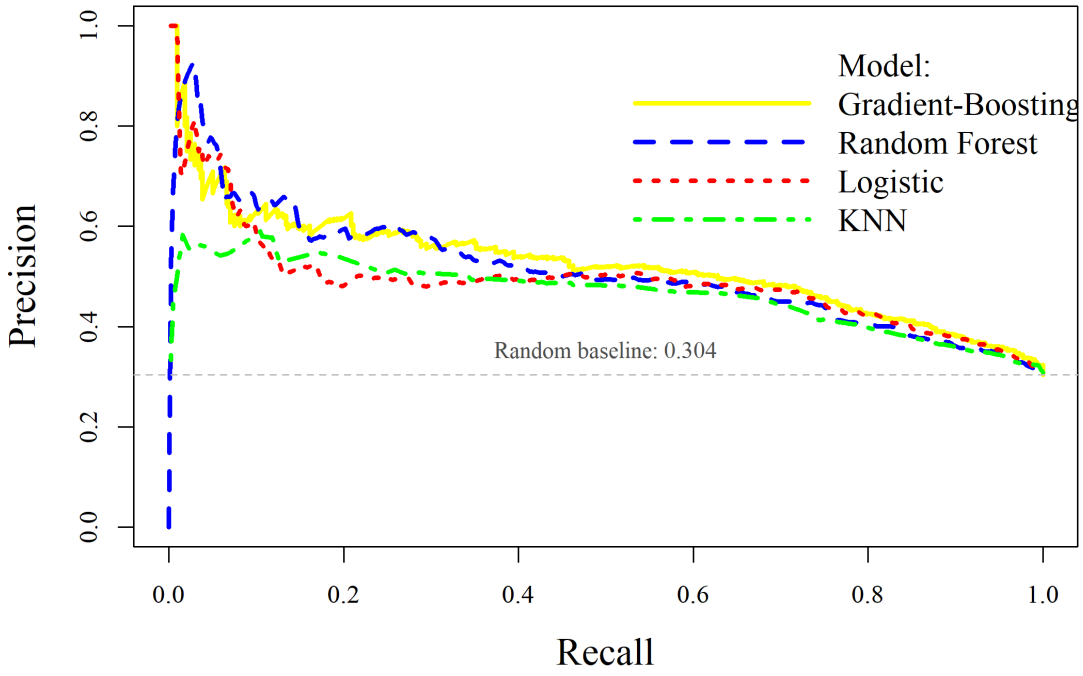


Figure 2: Precision Recall curves relative to GBM curve, different models

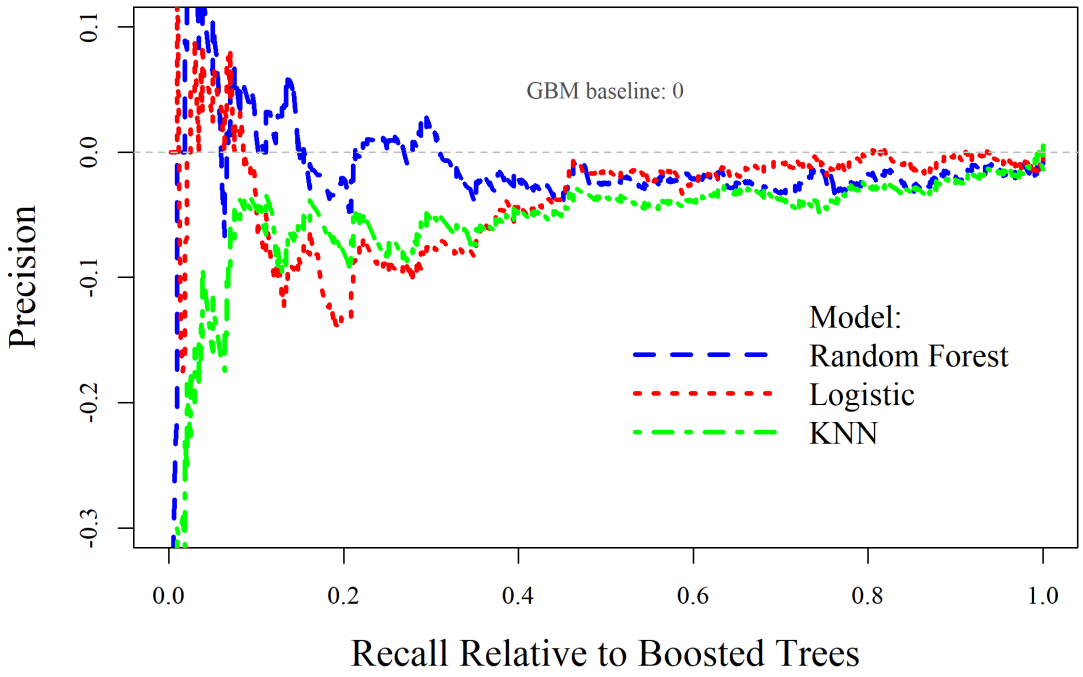
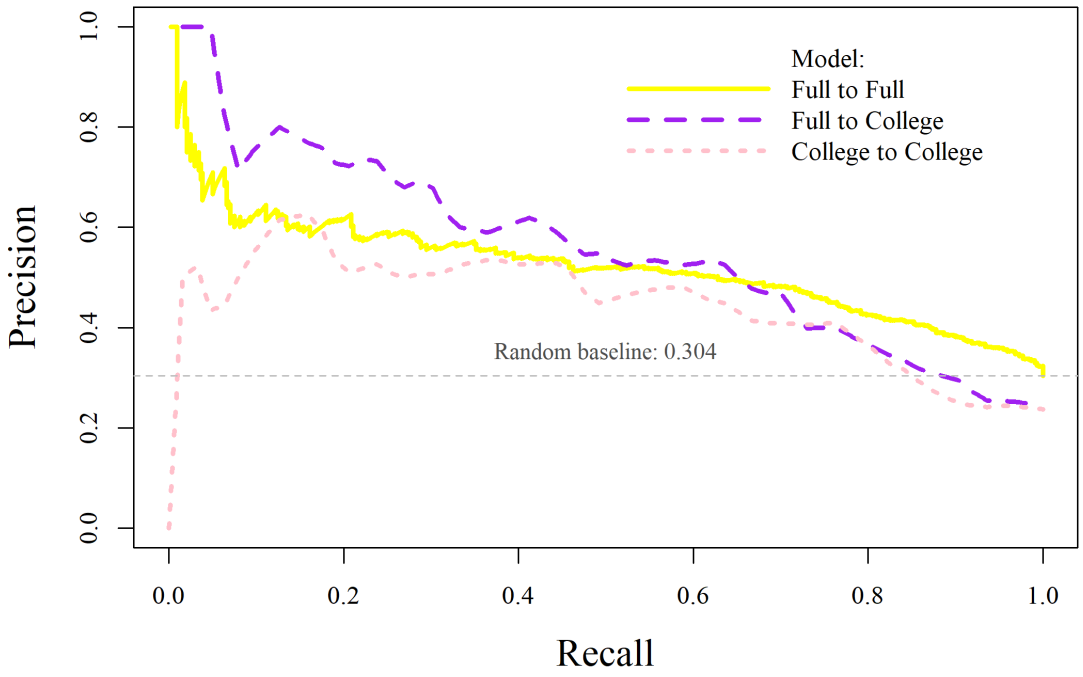
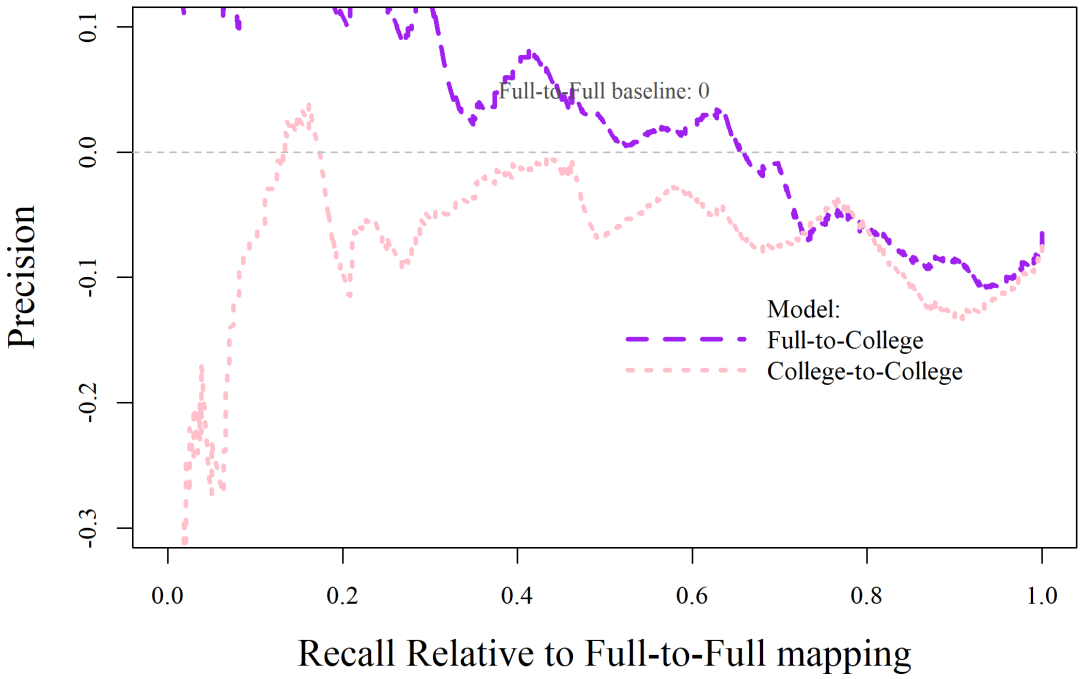


Figure 3: Precision-Recall Curves for Full and College Graduate Training Samples of the gradient boosting trees model mapped onto different test samples



Note: Full-to-Full refers to the performance of the GBM trained on the full training sample and evaluated with the full test sample. Full-to-college referes to the GBM trained on the full training sample but evaluated with the college graduate test sample. College-to-College refers to the GBM trained on the college graduate training sample and evaluated with the college graduate test sample.

Figure 4: Precision-Recall Curves for Full and College Graduate Training Samples of the gradient boosting trees model mapped onto different test samples



Note: Full-to-Full refers to the performance of the GBM trained on the full training sample and evaluated with the full test sample. Full-to-college referes to the GBM trained on the full training sample but evaluated with the college graduate test sample. College-to-College refers to the GBM trained on the college graduate training sample and evaluated with the college graduate test sample.

Figure 5: Precision-Recall values for demographic groups for Gradient-Boosting Trees (Full-to-Full)

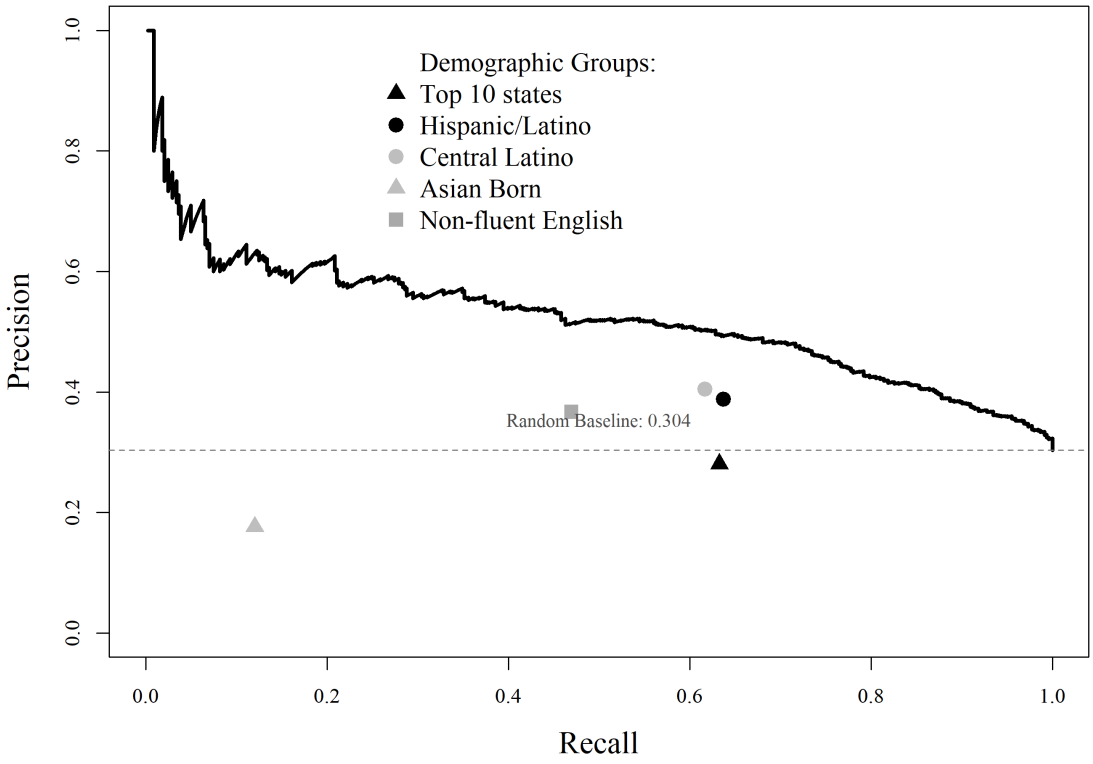


Figure 6: Coefficient plot of Undocumented Status and Inclusive State Policy Interactions

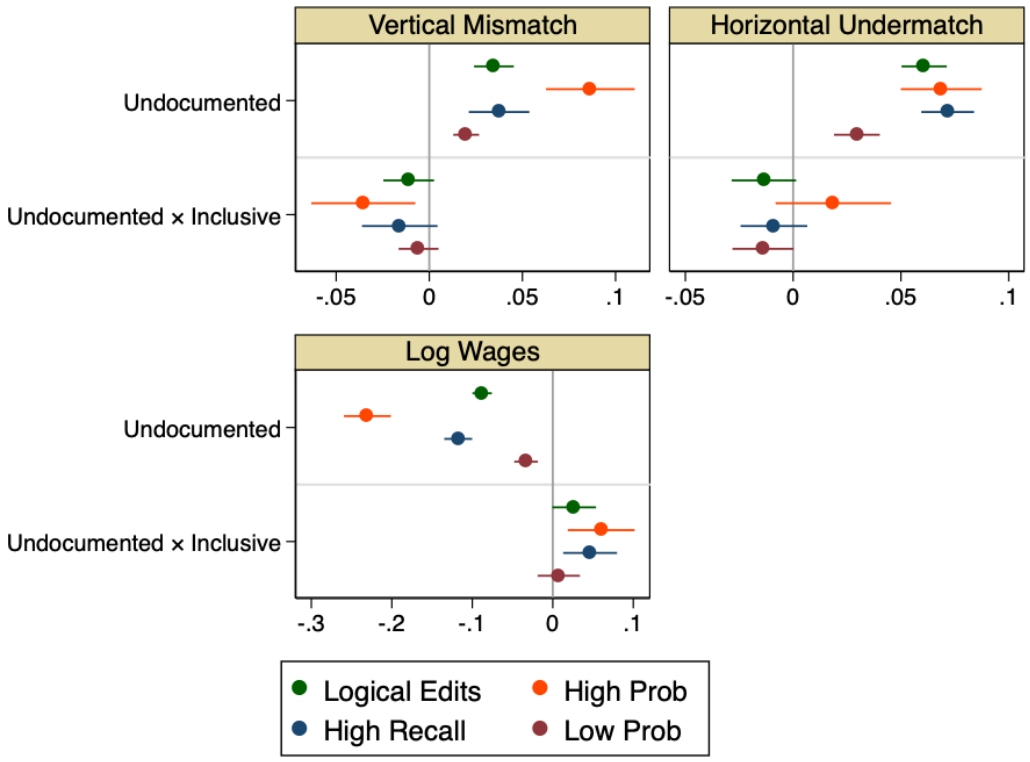
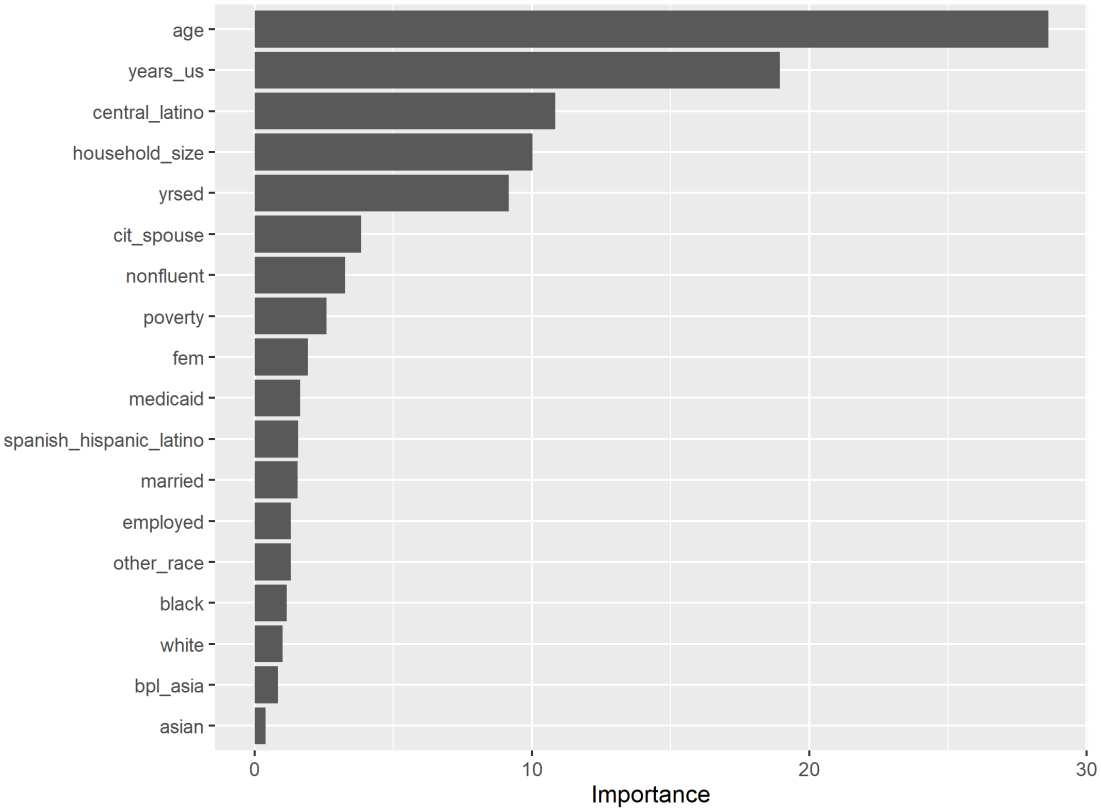


Figure 7: Gradient boosting trees Feature Importance (Full Training Sample)



7 Tables

Table 1: Probability Quartiles and Proportions of truly undocumented college graduates

Probability quartiles	Number of truly undocumented graduates	Number of observations in quartile	Proportion of correctly classified (Precision)
Full-to-Full			
Q1	36	363	0.0992
Q2	73	363	0.2011
Q3	138	363	0.3802
Q4	194	363	0.5344
Full-to-College			
Q1	5	67	0.0746
Q2	8	67	0.1194
Q3	14	67	0.2090
Q4	36	67	0.5373

Note: These quartiles were created from the SIPP by dividing it into four same-sized quartiles.

Table 2: ACS Summary Statistics by Probability Quartiles

	(1) Q1 (Low)	(2) Q2	(3) Q3	(4) Q4 (High)
Hispanic	0.115	0.231	0.346	0.420
	(0.319)	(0.421)	(0.476)	(0.494)
White	0.318	0.229	0.181	0.139
	(0.466)	(0.420)	(0.385)	(0.346)
Black	0.0668	0.0424	0.0309	0.0221
	(0.250)	(0.201)	(0.173)	(0.147)
Asian	0.479	0.477	0.421	0.395
	(0.500)	(0.499)	(0.494)	(0.489)
Female	0.471	0.417	0.375	0.323
	(0.499)	(0.493)	(0.484)	(0.468)
Surveyed Age	41.81	34.98	32.60	30.26
	(7.419)	(6.200)	(5.554)	(5.081)
Married	0.777	0.705	0.616	0.472
	(0.416)	(0.456)	(0.486)	(0.499)
Poor English	0.0604	0.0768	0.108	0.218
	(0.238)	(0.266)	(0.311)	(0.413)
years in the united states	13.98	7.688	6.156	3.717
	(8.578)	(5.826)	(4.797)	(3.738)
STEM Major	0.475	0.550	0.571	0.566
	(0.499)	(0.498)	(0.495)	(0.496)
STEM Related Major	0.107	0.0980	0.0834	0.0860
	(0.309)	(0.297)	(0.277)	(0.280)
Business Major	0.205	0.183	0.176	0.177
	(0.404)	(0.386)	(0.381)	(0.381)
Education Major	0.0509	0.0395	0.0444	0.0466
	(0.220)	(0.195)	(0.206)	(0.211)
Other Major	0.162	0.130	0.124	0.125
	(0.369)	(0.336)	(0.330)	(0.330)
Observations	111269	46806	15743	12158

Note: Probability Quartile thresholds were defined based on SIPP quartile thresholds.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3: ACS Summary Statistics

	(1) All	(2) Logical Edits	(3) High Recall	(4) High Prob	(5) Low Prob
Hispanic	0.0740	0.184	0.286	0.420	0.115
	(0.262)	(0.387)	(0.452)	(0.494)	(0.319)
White	0.745	0.272	0.204	0.139	0.318
	(0.436)	(0.445)	(0.403)	(0.346)	(0.466)
Asian	0.0982	0.468	0.452	0.395	0.479
	(0.298)	(0.499)	(0.498)	(0.489)	(0.500)
Black	0.0613	0.0547	0.0365	0.0221	0.0668
	(0.240)	(0.227)	(0.188)	(0.147)	(0.250)
Female	0.522	0.440	0.393	0.323	0.471
	(0.500)	(0.496)	(0.488)	(0.468)	(0.499)
Surveyed Age	39.53	38.56	33.70	30.26	41.81
	(9.257)	(7.997)	(6.160)	(5.081)	(7.419)
Poor English	0.00817	0.0789	0.107	0.218	0.0604
	(0.0900)	(0.270)	(0.309)	(0.413)	(0.238)
STEM Major	0.365	0.508	0.557	0.566	0.475
	(0.481)	(0.500)	(0.497)	(0.496)	(0.499)
STEM Related Major	0.0931	0.101	0.0930	0.0860	0.107
	(0.291)	(0.302)	(0.291)	(0.280)	(0.309)
Business Major	0.217	0.195	0.180	0.177	0.205
	(0.412)	(0.396)	(0.384)	(0.381)	(0.404)
Education Major	0.0937	0.0472	0.0418	0.0466	0.0509
	(0.291)	(0.212)	(0.200)	(0.211)	(0.220)
Other Major	0.232	0.148	0.128	0.125	0.162
	(0.422)	(0.356)	(0.334)	(0.330)	(0.369)
Vertically Mismatched	0.234	0.286	0.288	0.388	0.285
	(0.424)	(0.452)	(0.453)	(0.487)	(0.451)
Horizontally Mismatched	0.576	0.640	0.639	0.652	0.640
	(0.494)	(0.480)	(0.480)	(0.476)	(0.480)
Horizontally Undermatched	0.406	0.490	0.512	0.562	0.475
	(0.491)	(0.500)	(0.500)	(0.496)	(0.499)
Horizontally Overmatched	0.170	0.150	0.127	0.0899	0.165
	(0.375)	(0.357)	(0.333)	(0.286)	(0.371)
Inflation-adjusted Hourly wage	44.81	45.72	41.42	30.76	48.58
	(31.46)	(34.01)	(31.44)	(25.72)	(35.33)
Observations	3599938	185979	74373	12158	111269

Note: Probability Quartiles and 75 percent of positive cases thresholds were defined based on SIPP thresholds. The high recall group was defined by taking the predicted probability threshold that captured 75 percent of truly undocumented workers.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4: Regressions of Undocumented Status on Vertical Mismatch

	(1) Logical Edits	(2) High Prob	(3) High Recall	(4) Low Prob
Horizontal Undermatch	0.2554***	0.2557***	0.2556***	0.2558***
	[0.0043]	[0.0043]	[0.0043]	[0.0042]
Horizontal Overmatch	-0.0935***	-0.0934***	-0.0934***	-0.0935***
	[0.0022]	[0.0022]	[0.0022]	[0.0022]
Undocumented	0.0308***	0.0750***	0.0321***	0.0177***
	[0.0042]	[0.0094]	[0.0065]	[0.0022]
Mean of Dep. Var.	0.25	0.25	0.25	0.25
R-squared	0.18	0.18	0.18	0.18
N	3,598,848	3,598,848	3,598,848	3,598,848

Additional controls include:
dummy age indicators, gender, Medicaid receipt, race/ethnicity, metropolitan residence, government occupation, English-speaking fluency, foreign born, immigration by age 10, broad degree category indicators, years of schooling, and state×year interaction fixed effects.
Robust standard errors are clustered by state.
* $p < .1$, ** $p < .05$, *** $p < .01$

Table 5: Regressions of Undocumented Status on Horizontal Undermatch

	(1) Logical Edits	(2) High Prob	(3) High Recall	(4) Low Prob
Vertically Mismatched	0.3389***	0.3393***	0.3390***	0.3394***
	[0.0033]	[0.0033]	[0.0033]	[0.0033]
Undocumented	0.0560***	0.0748***	0.0687***	0.0246***
	[0.0031]	[0.0105]	[0.0049]	[0.0020]
Mean of Dep. Var.	0.41	0.41	0.41	0.41
R-squared	0.27	0.27	0.27	0.27
N	3,598,848	3,598,848	3,598,848	3,598,848

Additional controls include:
dummy age indicators, gender, Medicaid receipt, race/ethnicity, metropolitan residence, government occupation, English-speaking fluency, foreign born, immigration by age 10, broad degree category indicators, years of schooling, and state×year interaction fixed effects.
Robust standard errors are clustered by state.
* $p < .1$, ** $p < .05$, *** $p < .01$

Table 6: Regressions of Undocumented Status on Log Wages

	(1) Logical Edits	(2) High Prob	(3) High Recall	(4) Low Prob
Vertically Mismatched	-0.3168***	-0.3170***	-0.3170***	-0.3173***
	[0.0042]	[0.0043]	[0.0043]	[0.0043]
Horizontal Undermatch	-0.1476***	-0.1483***	-0.1479***	-0.1484***
	[0.0037]	[0.0037]	[0.0037]	[0.0037]
Horizontal Overmatch	0.0373***	0.0371***	0.0370***	0.0372***
	[0.0032]	[0.0032]	[0.0032]	[0.0032]
Undocumented	-0.0784***	-0.2108***	-0.1017***	-0.0306***
	[0.0054]	[0.0144]	[0.0105]	[0.0062]
Mean of Dep. Var.	3.57	3.57	3.57	3.57
R-squared	0.29	0.29	0.29	0.29
N	3,598,848	3,598,848	3,598,848	3,598,848

Additional controls include:
dummy age indicators, gender, Medicaid receipt, race/ethnicity, metropolitan residence, government occupation, English-speaking fluency, foreign born, immigration by age 10, broad degree category indicators, years of schooling, and state×year interaction fixed effects.
Robust standard errors are clustered by state.
* $p < .1$, ** $p < .05$, *** $p < .01$

8 Appendix

Figure A.1: Gradient-Boosted Trees Feature Importance (College Training Sample)

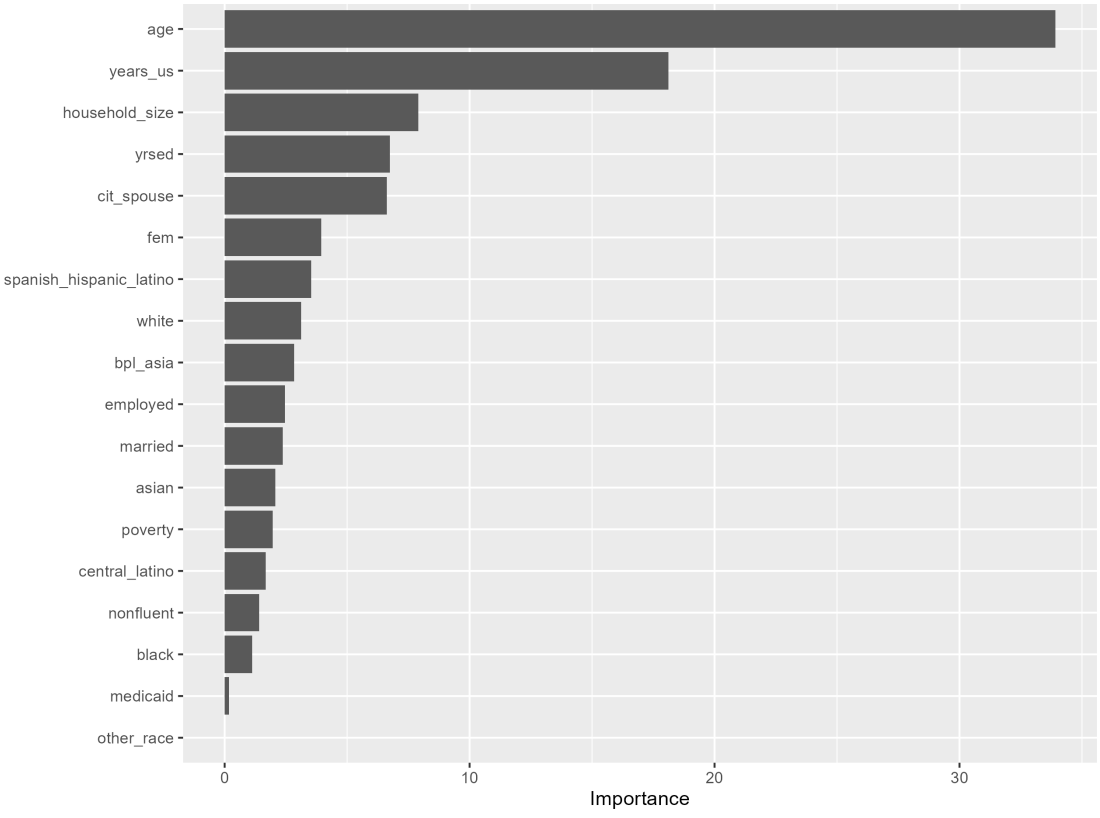


Figure A.2: Coefficient plot of Undocumented status and Field of Study interactions

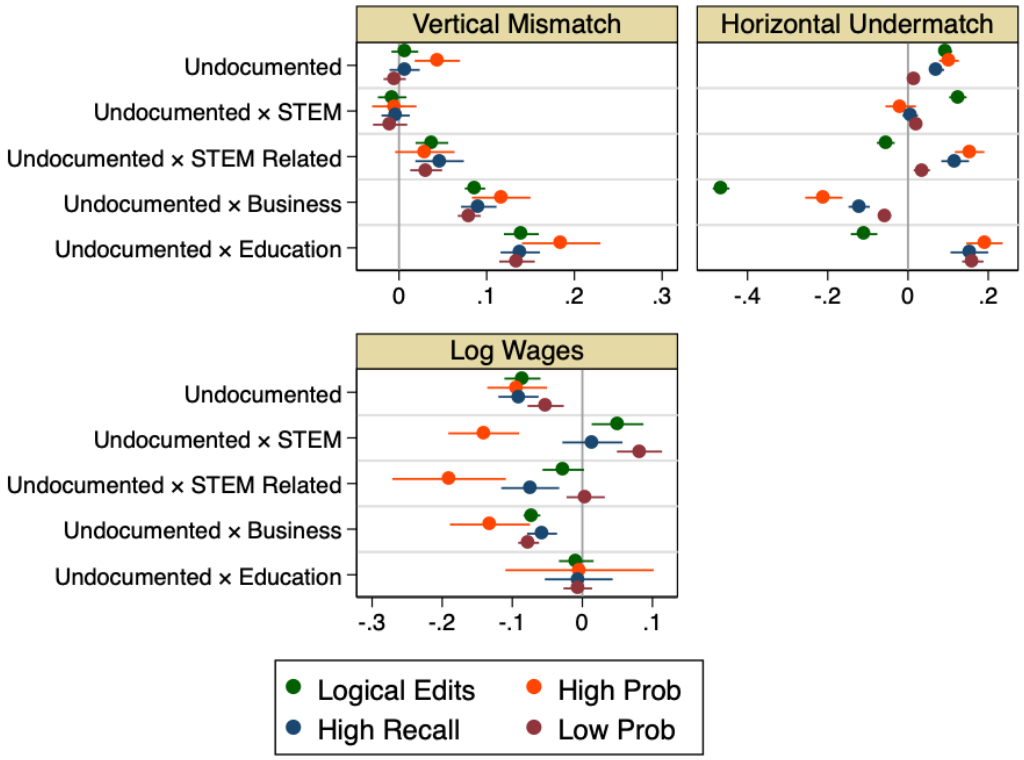


Figure A.3: Coefficient plot of Undocumented status and State policy interactions

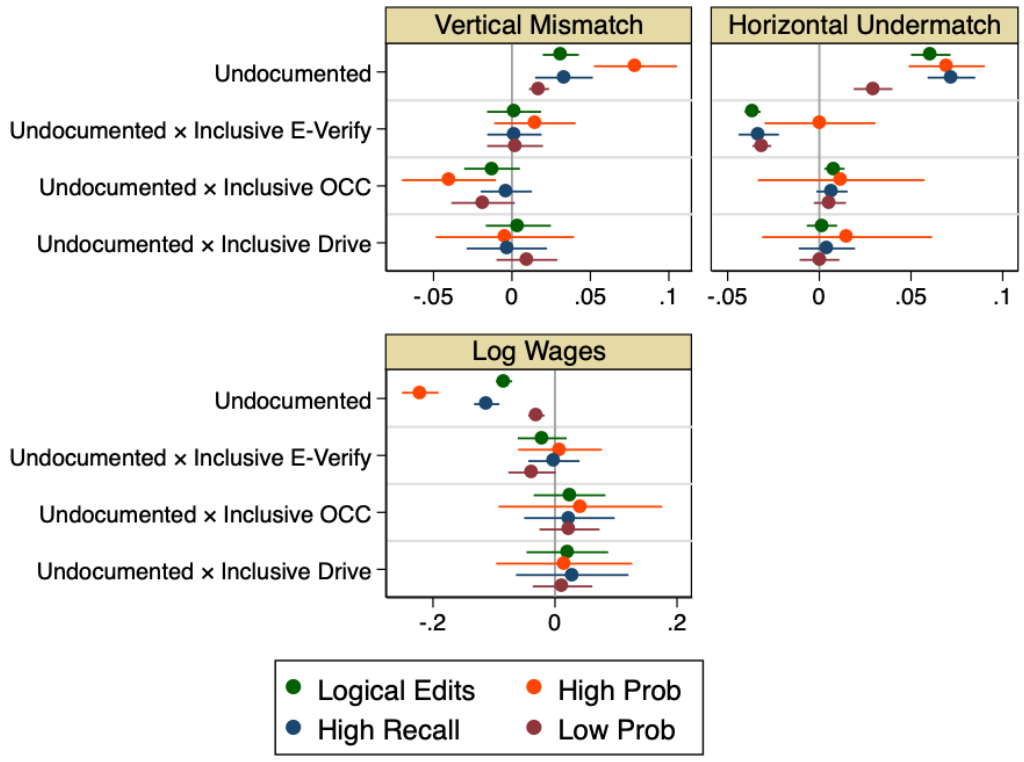


Table A.1: Commonly Used Sample Restrictions and their performance metrics, based in SIPP

	Base logical edits sample	Noncitizens	Top 10 states	Hispanic / Latino / Spanish	Central America and Latino
Recall	1.0000	1.0000	1.0000	1.0000	1.0000
Specificity	0.9547	0.0013	0.9411	0.8788	0.6304
Precision	0.2857	0.2634	0.2316	0.3361	0.3780
Accuracy	0.9555	0.2642	0.9421	0.8858	0.6982

Table A.2: Regressions of Undocumented Status on Vmismatch (Degree Interaction Terms)

	(1) Logical Edits	(2) High Prob	(3) High Recall	(4) Low Prob
Undocumented	0.0066	0.0437***	0.0064	-0.0049
	[0.0076]	[0.0128]	[0.0086]	[0.0063]
Undocumented × STEM	-0.0078	-0.0054	-0.0038	-0.0102
	[0.0081]	[0.0126]	[0.0080]	[0.0097]
Undocumented × STEM Related	0.0375***	0.0295*	0.0463***	0.0310***
	[0.0093]	[0.0169]	[0.0138]	[0.0090]
Undocumented × Business	0.0867***	0.1165***	0.0910***	0.0801***
	[0.0058]	[0.0166]	[0.0101]	[0.0065]
Undocumented × Education	0.1395***	0.1850***	0.1383***	0.1346***
	[0.0099]	[0.0223]	[0.0112]	[0.0101]
Horizontally Undermatched	0.2551***	0.2557***	0.2554***	0.2556***
	[0.0043]	[0.0043]	[0.0043]	[0.0043]
Horizontally Overmatched	-0.0942***	-0.0935***	-0.0938***	-0.0938***
	[0.0021]	[0.0022]	[0.0022]	[0.0022]
Mean of Dep. Var.	0.25	0.25	0.25	0.25
R-squared	0.18	0.18	0.18	0.18
N	3,598,848	3,598,848	3,598,848	3,598,848

Standard errors in brackets

* $p < .1$, ** $p < .05$, *** $p < .01$

Table A.3: Regressions of Undocumented Status on Horizontal Undermatch (Degree Interaction Terms)

	(1) Logical Edits	(2) High Prob	(3) High Recall	(4) Low Prob
Undocumented	0.0494***	0.1022***	0.0712***	0.0155***
	[0.0067]	[0.0122]	[0.0094]	[0.0050]
Undocumented × STEM	0.0183**	-0.0182	0.0055	0.0206***
	[0.0070]	[0.0189]	[0.0095]	[0.0070]
Undocumented × STEM Related	0.0697***	0.1536***	0.1175***	0.0345***
	[0.0119]	[0.0182]	[0.0172]	[0.0101]
Undocumented × Business	-0.0850***	-0.2097***	-0.1218***	-0.0584***
	[0.0102]	[0.0231]	[0.0132]	[0.0082]
Undocumented × Education	0.1626***	0.1908***	0.1530***	0.1616***
	[0.0166]	[0.0226]	[0.0234]	[0.0132]
Vertically Mismatched	0.3387***	0.3392***	0.3389***	0.3393***
	[0.0033]	[0.0033]	[0.0033]	[0.0033]
Mean of Dep. Var.	0.41	0.41	0.41	0.41
R-squared	0.27	0.27	0.27	0.27
N	3,598,848	3,598,848	3,598,848	3,598,848

Standard errors in brackets

* $p < .1$, ** $p < .05$, *** $p < .01$

Table A.4: Regressions of Undocumented Status on Log Wages (Degree Interaction Terms)

	(1) Logical Edits	(2) High Prob	(3) High Recall	(4) Low Prob
Undocumented	-0.0855***	-0.0930***	-0.0912***	-0.0524***
	[0.0128]	[0.0213]	[0.0142]	[0.0129]
Undocumented × STEM	0.0502***	-0.1406***	0.0143	0.0815***
	[0.0184]	[0.0252]	[0.0213]	[0.0160]
Undocumented × STEM Related	-0.0272*	-0.1902***	-0.0743***	0.0047
	[0.0147]	[0.0403]	[0.0206]	[0.0136]
Undocumented × Business	-0.0720***	-0.1320***	-0.0575***	-0.0766***
	[0.0060]	[0.0284]	[0.0107]	[0.0074]
Undocumented × Education	-0.0087	-0.0040	-0.0052	-0.0065
	[0.0123]	[0.0527]	[0.0241]	[0.0102]
Vertically Mismatched	-0.3162***	-0.3170***	-0.3168***	-0.3168***
	[0.0042]	[0.0043]	[0.0043]	[0.0042]
Horizontally Undermatched	-0.1479***	-0.1483***	-0.1478***	-0.1487***
	[0.0038]	[0.0037]	[0.0038]	[0.0037]
Horizontally Overmatched	0.0377***	0.0371***	0.0373***	0.0373***
	[0.0032]	[0.0032]	[0.0032]	[0.0032]
Mean of Dep. Var.	3.57	3.57	3.57	3.57
R-squared	0.29	0.29	0.29	0.29
N	3,598,848	3,598,848	3,598,848	3,598,848

Standard errors in brackets

* $p < .1$, ** $p < .05$, *** $p < .01$

Table A.5: Regressions of Undocumented Status on Vmismatch (IPC Interaction Terms)

	(1) Logical Edits	(2) High Prob	(3) High Recall	(4) Low Prob
Horizontally Undermatched	0.2554***	0.2557***	0.2556***	0.2558***
	[0.0043]	[0.0043]	[0.0043]	[0.0042]
Horizontally Overmatched	-0.0935***	-0.0934***	-0.0934***	-0.0935***
	[0.0022]	[0.0022]	[0.0022]	[0.0022]
Undocumented	0.0347***	0.0865***	0.0375***	0.0198***
	[0.0053]	[0.0119]	[0.0081]	[0.0034]
Undocumented × Inclusive	-0.0111	-0.0354**	-0.0159	-0.0058
	[0.0068]	[0.0139]	[0.0101]	[0.0053]
Mean of Dep. Var.	0.25	0.25	0.25	0.25
R-squared	0.18	0.18	0.18	0.18
N	3,598,848	3,598,848	3,598,848	3,598,848

Additional controls include:
dummy age indicators, gender, race/ethnicity, metropolitan residence, statefipyear age, government occupation, English-speaking fluency, foreign born, immigration by age 10, STEM degree indicators, years of schooling, and state×year interaction fixed effects.
Robust standard errors are all clustered by state.
* $p < .1$, ** $p < .05$, *** $p < .01$

Table A.6: Regressions of Undocumented Status on Horizontal Undermatch (IPC Interaction Terms)

	(1) Logical Edits	(2) High Prob	(3) High Recall	(4) Low Prob
Vertically Mismatched	0.3388***	0.3393***	0.3390***	0.3394***
	[0.0033]	[0.0033]	[0.0033]	[0.0033]
Undocumented	0.0608***	0.0687***	0.0717***	0.0296***
	[0.0052]	[0.0093]	[0.0061]	[0.0053]
Undocumented × Inclusive	-0.0136*	0.0186	-0.0089	-0.0140*
	[0.0074]	[0.0133]	[0.0077]	[0.0071]
Mean of Dep. Var.	0.41	0.41	0.41	0.41
R-squared	0.27	0.27	0.27	0.27
N	3,598,848	3,598,848	3,598,848	3,598,848

Additional controls include:
dummy age indicators, gender, race/ethnicity, metropolitan residence, statefipyear age, government occupation, English-speaking fluency, foreign born, immigration by age 10, STEM degree indicators, years of schooling, and state×year interaction fixed effects.
Robust standard errors are all clustered by state.
* $p < .1$, ** $p < .05$, *** $p < .01$

Table A.7: Regressions of Undocumented Status on Log Wages (IPC Interaction Terms)

	(1) Logical Edits	(2) High Prob	(3) High Recall	(4) Low Prob
Vertically Mismatched	-0.3168***	-0.3170***	-0.3170***	-0.3173***
	[0.0042]	[0.0043]	[0.0043]	[0.0043]
Horizontally Undermatched	-0.1476***	-0.1483***	-0.1478***	-0.1484***
	[0.0037]	[0.0037]	[0.0037]	[0.0037]
Horizontally Overmatched	0.0373***	0.0371***	0.0370***	0.0372***
	[0.0032]	[0.0032]	[0.0032]	[0.0032]
Undocumented	-0.0878***	-0.2303***	-0.1175***	-0.0332***
	[0.0060]	[0.0145]	[0.0086]	[0.0073]
Undocumented × Inclusive	0.0266*	0.0602***	0.0464***	0.0075
	[0.0135]	[0.0207]	[0.0166]	[0.0131]
Mean of Dep. Var.	3.57	3.57	3.57	3.57
R-squared	0.29	0.29	0.29	0.29
N	3,598,848	3,598,848	3,598,848	3,598,848

Additional controls include:
dummy age indicators, gender, race/ethnicity, metropolitan residence, statefipyear age, government occupation, English-speaking fluency, foreign born, immigration by age 10, STEM degree indicators, years of schooling, and state×year interaction fixed effects.
Robust standard errors are all clustered by state.
* $p < .1$, ** $p < .05$, *** $p < .01$

Table A.8: Regressions of Undocumented Status on Vmismatch (Policy Interaction Terms)

	(1) Logical Edits	(2) High Prob	(3) High Recall	(4) Low Prob
Horizontally Undermatched	0.2554***	0.2557***	0.2556***	0.2558***
	[0.0043]	[0.0043]	[0.0043]	[0.0043]
Horizontally Overmatched	-0.0935***	-0.0934***	-0.0934***	-0.0935***
	[0.0022]	[0.0022]	[0.0022]	[0.0022]
Undocumented	0.0311***	0.0788***	0.0332***	0.0172***
	[0.0057]	[0.0132]	[0.0092]	[0.0032]
Undocumented × Inclusive Everify	0.0014	0.0146	0.0015	0.0020
	[0.0086]	[0.0129]	[0.0086]	[0.0089]
Undocumented × Inclusive OCC	-0.0127	-0.0402***	-0.0036	-0.0184*
	[0.0089]	[0.0149]	[0.0081]	[0.0101]
Undocumented × Inclusive Drive	0.0041	-0.0046	-0.0033	0.0095
	[0.0103]	[0.0220]	[0.0128]	[0.0097]
Mean of Dep. Var.	0.25	0.25	0.25	0.25
R-squared	0.18	0.18	0.18	0.18
N	3,598,848	3,598,848	3,598,848	3,598,848

Additional controls include:
dummy age indicators, gender, race/ethnicity, metropolitan residence, statefipyear age, government occupation, English-speaking fluency, foreign born, immigration by age 10, STEM degree indicators, years of schooling, and state×year interaction fixed effects.
Robust standard errors are all clustered by state.
* $p < .1$, ** $p < .05$, *** $p < .01$

Table A.9: Regressions of Undocumented Status on Horizontal Undermatch (Policy Interaction Terms)

	(1) Logical Edits	(2) High Prob	(3) High Recall	(4) Low Prob
Vertically Mismatched	0.3388***	0.3393***	0.3390***	0.3394***
	[0.0033]	[0.0033]	[0.0033]	[0.0033]
Undocumented	0.0608***	0.0695***	0.0720***	0.0293***
	[0.0054]	[0.0104]	[0.0065]	[0.0052]
Undocumented × Inclusive Everify	-0.0364***	0.0004	-0.0330***	-0.0314***
	[0.0022]	[0.0151]	[0.0055]	[0.0025]
Undocumented × Inclusive OCC	0.0082***	0.0119	0.0069	0.0058
	[0.0028]	[0.0226]	[0.0042]	[0.0044]
Undocumented × Inclusive Drive	0.0014	0.0151	0.0041	0.0001
	[0.0041]	[0.0231]	[0.0076]	[0.0054]
Mean of Dep. Var.	0.41	0.41	0.41	0.41
R-squared	0.27	0.27	0.27	0.27
N	3,598,848	3,598,848	3,598,848	3,598,848

Additional controls include:
dummy age indicators, gender, race/ethnicity, metropolitan residence, statefipyear age, government occupation, English-speaking fluency, foreign born, immigration by age 10, STEM degree indicators, years of schooling, and state×year interaction fixed effects.
Robust standard errors are all clustered by state.
* $p < .1$, ** $p < .05$, *** $p < .01$

Table A.10: Regressions of Undocumented Status on Log Wages (Policy Interaction Terms)

	(1) Logical Edits	(2) High Prob	(3) High Recall	(4) Low Prob
Vertically Mismatched	-0.3168***	-0.3170***	-0.3170***	-0.3173***
	[0.0042]	[0.0043]	[0.0043]	[0.0042]
Horizontally Undermatched	-0.1476***	-0.1483***	-0.1478***	-0.1485***
	[0.0037]	[0.0037]	[0.0037]	[0.0037]
Horizontally Overmatched	0.0373***	0.0371***	0.0370***	0.0372***
	[0.0032]	[0.0032]	[0.0032]	[0.0032]
Undocumented	-0.0835***	-0.2208***	-0.1119***	-0.0305***
	[0.0066]	[0.0150]	[0.0103]	[0.0066]
Undocumented × Inclusive Everify	-0.0210	0.0081	-0.0017	-0.0377*
	[0.0200]	[0.0341]	[0.0209]	[0.0193]
Undocumented × Inclusive OCC	0.0239	0.0415	0.0237	0.0238
	[0.0292]	[0.0668]	[0.0369]	[0.0245]
Undocumented × Inclusive Drive	0.0202	0.0151	0.0281	0.0123
	[0.0333]	[0.0555]	[0.0459]	[0.0243]
Mean of Dep. Var.	3.57	3.57	3.57	3.57
R-squared	0.29	0.29	0.29	0.29
N	3,598,848	3,598,848	3,598,848	3,598,848

Additional controls include:
dummy age indicators, gender, race/ethnicity, metropolitan residence, statefipyear age, government occupation, English-speaking fluency, foreign born, immigration by age 10, STEM degree indicators, years of schooling, and state×year interaction fixed effects.
Robust standard errors are all clustered by state.
* $p < .1$, ** $p < .05$, *** $p < .01$