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 email (vsovero@ucr.edu)

[2]University of California, Riverside (mario.arce1615@gmail.com)

February 7, 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 machine learning models on the Survey of Income and Program Participation (SIPP) and use the predictions to impute status in the ACS. This approach enables new analyses of labor market outcomes for undocumented college graduates in nationally representative surveys. Results indicate that undocumented status is linked to higher rates of both vertical and horizontal mismatch, as well as wage penalties of roughly 4–7 percent. These penalties are smaller among the DACA-eligible and in states with more inclusive immigrant policy climates.

Keywords: Undocumented; Education-occupation mismatch; Legal status; Labor; Wage; DACA; 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 2013–2019 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 positive predictive value (PPV) 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 American Community Survey (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 regression, k-nearest neighbors, and random forest models, 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, including sensitivity, specificity, and positive predictive value (PPV). In validation tests, the random forest achieves a PPV of about 0.50 for college graduates, compared to only 0.28 for logical imputation. This improvement enables us to apply the trained models to the 2013–2019 ACS sample of college graduates, providing the scale necessary to study education–occupation mismatch and wage outcomes among undocumented workers.

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 four to seven percentage points, depending on the undocumented imputation method.

We also examine how policy context shapes these disadvantages. At the federal level, the Deferred Action for Childhood Arrivals (DACA) program, which provides work authorization and protection from deportation, is associated with lower rates of mismatch among eligible college graduates, although the evidence on wage penalties is more mixed. 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 both federal and 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, comparing federal eligibility under DACA with variation across state immigrant policy climates (Amuedo-Dorantes and Antman, 2017; Hsin and Ortega, 2018; Kuka *et al.*, 2020; 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 (**?**). 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. At the federal level, 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 positive predictive value, leading to frequent misclassification of legally present immigrants as undocumented. Statistical imputation provides a different strategy, relying on regression-based or cross-survey methods (**?**). 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

We use the 2013 to 2019 ACS surveys for our main estimation sample. Although the college degree field variable is available as far back as 2009, we are restricting our sample period to the post-DACA policy environment[1]. Relatedly, we do not include ACS data after the onset of the Covid-19 pandemic. The estimation sample consists of respondents of prime working age adults (age 20 to 42 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. 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 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]

[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.

[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.

3.2 IPC Index

To capture the underlying police climate towards immigrants, we use the Immigrant Policy Climate Index (IPC) , which 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 (Samari *et al.*, 2021). The fourteen policies are coded as either 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 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 utilize several machine learning algorithms. The general approach is as follows: we construct a donor sample that contains information on a respondent's legal status to 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 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

[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.

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

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 DACA eligibility imputation method.[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.[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.[9] 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.

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. 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 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).

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 dataset, we use the k-fold cross validation procedure, which splits the training data into folds, 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 to address class imbalance (only around 25 percent of the possibly undocumented

[7] Borjas and Cassidy use a method that marks immigrants that i) are within the occupations making up 80% of H-1B beneficiaries, ii) have resided in the U.S. for 6 years or less, iii) and are college graduates.

[8] We inlcuded 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.

[9] The different sets of codes spanned 2002-2009, 2010-2017, 2018-2022

sample is actually undocumented).

We use the following as predictors of legal status: age, years in the US, 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 DACA Sample

To identify possible DACA recipients, we apply additional filters. An immigrant without lawful status is eligible for DACA if i) they arrived in the U.S. before the age of 16, ii) they have continuously resided in the U.S. states since June 15, 2007, iii) are under 31 years old as of June 15 2012, iv) are currently in school or have completed a high school equivalent education or are an honorary veteran v) committed no felonies or significant misdemeanors and less than 3 other misdemeanors vi) and are 15 years or older at the time of application.[10] We use conditions 1-4 as additional filters (the rest are not observed in the ACS).

3.5 Econometric Model

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

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

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. $Mismatch_i$ refers to our three mismatch models: vertical mismatch, horizontal mismatch, and horizontal undermatch. 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. The coefficient of interest is for $Undocu_i$, an indicator of whether a worker is undocumented.

[10]For more information: `https://www.uscis.gov/DACA`

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 (sensitivity), the rate at which truly documented people were classified as documented (specificity), the rate at which those classified as undocumented were truly undocumented (positive predictive value, or precision), and the rate at which truly undocumented and documented people were correctly classified (accuracy)[11]. Results are reported in Table 1 for the SIPP test sample (not used in the model training).

For the logical edits imputation, the sensitivity is 1 because by definition, everyone in the test data sample is possibly undocumented. We can also see that the positive predictive value is quite low: only 28 percent of the possibly undocumented are actually undocumented. For all three algorithms, the positive predictive value is consistently higher at around 40 to 50 percent. The Random Forest algorithm has much higher positive predictive value compared to logisitic and KNN, which made this the preferred model. Nonetheless, we will present regression results with all of the imputation methods. Importantly, these models have higher positive predictive values than the commonly used sample restrictions used to isolate the truly undocumented (Appendix A.1). 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 Random Forest model is reported in 1, where

[11] Sensitivity: $\frac{TP}{TP+FN}$
Specificity: $\frac{TN}{TN+FP}$
Precision / Positive-predictive value: $\frac{TP}{TP+FP}$
Accuracy: $\frac{TP+TN}{total}$

a larger number signifies greater importance. Age and years in the US are the most highly ranked predictors, followed by years of education and Hispanic ethnicity/birthplace.

4.2 SIPP Descriptive Statistics

The SIPP data allows for us to investigate the demographic characteristics of the undocumented college graduate population, which can then be compared to the summary statistics in the target sample (ACS dataset). Table **??** presents descriptive statistics for the college graduate population in the SIPP data. Column 1 represents the pool of possibly undocumented (logical edits), whereas Column 2 restricts to the truly undocumented with the caveat that this includes those with legal temporary resident status. Interestingly, a large proportion of undocumented college graduates were born in Asia (51 percent) and a relatively smaller proportion we born in Latin America (12.9 percent). This is consistent with the Center for Migration Studies report on undocumented college graduates, which finds that India is the most common country of origin, representing 28 percent of the undocumented college graduate population.

The descriptive statistics for individuals predicted to be undocumented using KNN and the Random Forest imputations are largely similar to the truly undocumented (Column 2), with a few key differences. The KNN predictions generate a higher proportion of Asian immigrants and a lower proportion of Hispanic immigrants compared to the demographics of the actual undocumented population. For the random forest predictions, the opposite is true (higher proportion of Hispanic, lower proportion of Asian).

4.3 ACS Descriptive Statistics

In Table **??**, we compare the descriptive statistics of the ACS sample across the various imputation methods. Column 1 restricts to the foreign born population of college graduates, Column 2 applies the logical edits to identify the possibly undocumented. Our descriptive statistics show that the rates of vertical and horizontal mismatch are higher for undocumented college graduates compared to the overall foreign born population across all of the undocumented imputation methods. Interestingly, undocumented immigrants are slightly more likely to major in STEM compared to the overall foreign born population. In the regression analysis we will further examine the interactions between degree and education-occupation mismatch.

Column 3 further restricts to DACA-eligible. The DACA-eligible group has lower average earnings compared to the larger population of undocumented immigrants, possibly due to the fact that the group is much younger, on average, than the overall undocumented population. The DACA-eligible are also much more likely to be Hispanic and less likely to be Asian. In terms of degree attainment, they are also less likely to have obtained a STEM degree.

Columns 3 and 4 use the KNN and Random Forest imputation methods respectively to identify

undocumented immigrants. The demographic profiles are less imbalanced across the various imputation methods compared to the SIPP sample. For example, the proportion of Hispanic individuals is similar across the KNN and RF samples, with a slightly higher proportion of Asian individuals in the RF sample.

4.4 Mismatch by occupation

In order to investigate occupational sorting trends, we created a descriptive table listing the 10 occupations containing the most undocumented college graduates using the Random Forest imputation (Table 4). Most of the top occupation are in the service industry, which includes cooks, waiters and waitressses and childcare workers. Eight of the top 10 occupations are vertically mismatched, meaning that most workers in these occupations have less than a college degree. Horizontal undermatch rates are also high, which means that workers are employed in a occupation that pays less than what is typical for the degree.

4.5 Mismatch by degree field

We also list the degree fields that most undocumented college graduates choose in order to investigate differences in education-occupation mismatch by college major (Table 5). Business is the most common degree pursued by undocumented college graduates, followed by engineering and education. The rates of mismatch vary a great deal across these degrees. For example, rates of vertical mismatch are highest for business and education majors, whereas horizontal undermatch rates are higher for engineering majors. Whether the mismatch is due to occupational barriers will be explored further in the regression analysis.

4.6 Mismatch Regression Results

After we have imputed undocumented status in the ACS sample using predictions from the RF model trained on SIPP data, we proceed with the primary objective of this study and estimate regressions of vertical and horizontal mismatch on undocumented status. We use the three different methods for imputing undocumented status: logical edits, KNN, and Random Forest.

We see from Table **??** that undocumented status increases the likelihood of being vertically mismatched anywhere from 1.5 to 3.6 percentage points, with the largest effect coming from the logical edits imputation. We find slightly larger effects on horizontal mismatch, which is when an individual is employed in an occupation that does not match their degree. Specifically, the likelihood of being horizontally mismatched is around 3 to 4 percentage points higher for undocumented immigrants (Table **??**). 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 **??**). Importantly, these coefficients are statistically significant across all of the undocumented imputation methods, which provides robust evidence that undocumented status increases education-occupation mismatch.

4.7 Wage Regression Results

Consistent with previous studies, we observe a statistically significant negative correlation between undocumented status and log wages. From Table **??** we see that undocumented status is associated with a wage penalty ranging from 4 percent to almost 7 percent, holding vertical and horizontal mismatch constant. This suggests that there are additional labor market barriers associated with undocumented status , aside from occupational sorting, that contribute to lower wages.

4.8 DACA College Graduates

With our improved method for imputing undocumented status, we revisit the impact of DACA eligibility on mismatch and wages. We find that the relationship between DACA eligibility and vertical mismatch is not statistically significant (Table **??**), which suggests that DACA was effective in allowing college graduates to pursue an occupation that required a college degree.

On the other hand, we find evidence that DACA-eligibility status still increases the likelihood of being horizontal mismatched (and horizontal undermatched). Tables 11 and **??** show that DACA eligibility is correlated with an increase in likelihood of being horizontally mismatched and undermatched by around 2 percentage points and 3 percentage points, respectively. It is worth noting that the coefficients for DACA-eligibility and measures of mismatch are generally smaller compared to the effect size for the undocumented population as a whole.

We next investigate the wage penalty for the DACA-eligible college graduates (Table **??**). The results appear to be more sensitive to the imputation method, which means these results should be interpreted with more caution. For example, there is no statistically significant wage penalty using the KNN imputation, but the random forest imputation shows a large 10 percent wage penalty. Given the small proportion of DACA-eligible in our sample, it is likely that our estimates are too noisy to draw any conclusions about the size of the wage penalty.

4.9 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 Arts and Humanities (reference category in the regression tables). The coefficients are presented graphically in Figure 2 (regression tables presented in the appendix). All of the results are robust across the various imputation methods for undocumented status.

For STEM fields, undocumented college graduates are less likely to be vertically mismatched and have similar rates of horizontal undermatch compared to graduates with legal status. After controlling for mismatch, undocumented STEM majors appear to enjoy an earning premium. For STEM related fields, there is no increase in vertical mismatch from being undocumented. However, undocumented college graduates in STEM related fields are more likely to be horizontally undermatched (employed in an occupation that pays less than what is typical for a horizontally matched worker in the same field). This additional mismatch could be due to occupational barriers to STEM-related fields, which include math teacher education, nursing, and medical and health services. After controlling for vertical and horizontal mismatch, undocumented college graduates face additional wage penalties when majoring in STEM related fields.

For business and education majors, undocumented college graduates are more likely to be vertically mismatched compared to those who are not undocumented. Controlling for vertical mismatch, undocumented business majors are actually less likely to be horizontally undermatched. In contrast, education majors are more likely to be horizontally undermatched. This is again likely related to the fact that there are licensing barriers associated with teaching. However, after controlling for both vertical and horizontal mismatch, both business and education majors still face a wage penalty.

4.10 Policy Climate Interactions

The degree interactions provide suggestive evidence that fields that are more tightly regulated generate additional mismatch/wage penalties for undocumented college graduates. To investigate further, we interact undocumented status with an indicator for states that have net inclusive immigration policies (IPC index). Results are presented in Figure 3 and the corresponding regression tables are presented in the Appendix.

For the vertical mismatch regressions, there is evidence of a decrease in the likelihood of being vertically mismatched in states that are net inclusive. The strongest correlation is observed for the random forest imputation, which estimates a 4 percentage point decrease in the likelihood of being vertically mismatched when an undocumented immigrant is living in a net inclusive policy climate. Horizontal undermatch also appears to be less likely for undocumented immigrants living in net inclusive states, but the decrease is smaller compared to the vertical mismatch effects, with the effect size ranging from 1 to 1.6 percentage points. The RF estimates are statistically significant at the 5 percent significance level.

Holding mismatch constant, the KNN and random forest imputations suggest that wages are higher for undocumented college graduates when the state policy is net inclusive. Specifically, wages are around 5 percentage points higher using the machine learning imputations. These results suggests that an inclusive immigrant policy climate may reduce employer exploitation, separate

from it's effect on removing occupational barriers.

The interaction coefficient in the wage estimate is much smaller and statistically significant for the logical imputation. The fact that the correlations are strongest for the machine learned algorithms provide additional evidence that these imputation methods have a larger proportion of truly undocumented compared to the logical edits imputation. The idea behind this is the fact that inclusive state policies should primarily affect the actually undocumented and should have little to no effect on those with legal status. The logical imputation likely has a larger proportion of false positives, which leads to a weaker correlation with the immigrant policy climate.

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, and whether the state cooperates with federal immigration enforcement. Results are presented graphically in Figure 4 and the corresponding regression tables are presented in the Appendix.

For vertical mismatch, immigration enforcement appears to generate the largest reduction in vertical mismatch versus policies more related to employment and identification. Specifically, vertical mismatch is approximately 2 to 3 percentage points lower in states that do not cooperate with federal immigration enforcement (machine learning imputations). The effect size is again smaller for the the logical imputation estimates. On the other hand, prohibiting E-Verify plays a more important role in reducing horizontal undermatch for undocumented immigrants. Estimates range from 1 to three percentage point reductions in horizontal undermatch. After controlling for mismatch, prohibiting federal immigration cooperation has a large positive effect on wages for undocumented immigrants (estimates ranging from 3 to 7 percentage points).

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 acknowleged 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. We do not explicitly meet the same universe condition on the account

that our SIPP sample was for the year 2008-2009, and that our ACS sample spans the years 2013-2019. 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 2013-2019.

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 tree 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 Tables

	Logical edits	Logistic	KNN	RF
sensitivity	1.0000	0.6401	0.6424	0.6629
specificity	0.9547	0.6614	0.6028	0.7190
ppv	0.2857	0.4518	0.4135	0.5070
accuracy	0.9555	0.6549	0.6148	0.7019

Table 1: Imputation Methods Model Metrics

	Logical edits	Logistic	KNN	RF
sensitivity	1.0000	0.7312	0.6720	0.6811
specificity	0.9547	0.6445	0.5829	0.7061
ppv	0.2857	0.4728	0.4126	0.5025
accuracy	0.9555	0.6708	0.6100	0.6985

Table 2: SIPP Summary Statistics of Undocumented Imputation Methods

	(1) Undocumented (Logical edits)	(2) Undocumented (Actual)	(3) Undocumented (KNN)	(4) Undocumented (RF)
age	40.30	34.37	35.42	31.43
	(12.19)	(9.593)	(10.33)	(7.402)
fem	0.505	0.457	0.436	0.461
	(0.500)	(0.499)	(0.497)	(0.500)
married	0.730	0.643	0.646	0.544
	(0.444)	(0.480)	(0.479)	(0.499)
nonfluent	0.109	0.0860	0.0987	0.0984
	(0.312)	(0.281)	(0.299)	(0.299)
household_size	3.067	2.919	2.863	2.637
	(1.523)	(1.616)	(1.505)	(1.582)
poverty	0.142	0.181	0.178	0.197
	(0.349)	(0.386)	(0.383)	(0.399)
asian	0.445	0.466	0.475	0.513
	(0.497)	(0.500)	(0.500)	(0.501)
black	0.0742	0.0679	0.0510	0.0674
	(0.262)	(0.252)	(0.220)	(0.251)
white	0.463	0.448	0.455	0.394
	(0.499)	(0.498)	(0.499)	(0.490)
other_race	0.0180	0.0181	0.0191	0.0259
	(0.133)	(0.134)	(0.137)	(0.159)
bpl_asia	0.489	0.502	0.513	0.539
	(0.500)	(0.501)	(0.501)	(0.500)
Latino, born in Central America	0.0837	0.136	0.156	0.192
	(0.277)	(0.343)	(0.363)	(0.395)
spanish_hispanic_latino	0.115	0.176	0.182	0.223
	(0.320)	(0.382)	(0.386)	(0.417)
employed	0.747	0.738	0.752	0.756
	(0.435)	(0.441)	(0.433)	(0.430)
years_us	9.497	6.905	6.551	4.503
	(8.596)	(7.013)	(5.321)	(3.996)
yrsed	16.83	16.98	16.97	16.86
	(1.555)	(1.649)	(1.736)	(1.400)
Undocumented (Actual)	0.234	1	0.449	0.736
	(0.424)	(0)	(0.498)	(0.442)
Observations	944	221	314	193

mean coefficients; sd in parentheses

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

Table 3: ACS U.S. born workers and Undocumented immigrants Summary Statistics

	(1) Foreign-born	(2) Undocumented (Logical edits)	(3) DACA-eligible	(4) Undocumented (KNN)	(5) Undocumented (RF)
Surveyed Age	40.39	38.48	28.50	36.08	32.88
	(8.557)	(8.039)	(3.717)	(7.808)	(6.181)
Female	0.483	0.443	0.523	0.517	0.449
	(0.500)	(0.497)	(0.500)	(0.500)	(0.497)
Vertically Mismatched	0.234	0.272	0.294	0.312	0.339
	(0.423)	(0.445)	(0.456)	(0.463)	(0.473)
Horizontally Mismatched	0.613	0.648	0.609	0.645	0.653
	(0.487)	(0.478)	(0.488)	(0.478)	(0.476)
Horizontally Undermatched	0.444	0.491	0.478	0.509	0.531
	(0.497)	(0.500)	(0.500)	(0.500)	(0.499)
Horizontally Overmatched	0.169	0.158	0.130	0.136	0.122
	(0.375)	(0.364)	(0.336)	(0.343)	(0.327)
Poor English	0.0398	0.0776	0.0420	0.111	0.127
	(0.196)	(0.268)	(0.201)	(0.315)	(0.334)
STEM Degree	0.496	0.512	0.376	0.498	0.521
	(0.500)	(0.500)	(0.484)	(0.500)	(0.500)
Inflation-adjusted Hourly wage	49.07	46.27	32.33	40.64	37.04
	(34.09)	(34.42)	(22.19)	(32.08)	(29.30)
ln_adj	3.651	3.557	3.267	3.411	3.319
	(0.741)	(0.791)	(0.675)	(0.804)	(0.803)
White	0.279	0.265	0.207	0.231	0.133
	(0.448)	(0.441)	(0.405)	(0.422)	(0.339)
Black	0.0648	0.0534	0.0707	0.0493	0.0295
	(0.246)	(0.225)	(0.256)	(0.216)	(0.169)
Asian	0.465	0.478	0.313	0.388	0.410
	(0.499)	(0.500)	(0.464)	(0.487)	(0.492)
Hispanic	0.163	0.182	0.383	0.315	0.411
	(0.370)	(0.385)	(0.486)	(0.465)	(0.492)
Observations	404226	123867	6241	53868	26166

Note: Log wage is adjusted for inflation with CPI values starting January 2009, every year in January until January 2019.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4: Top 10 occupations populated by Undocumented (RF) workers

Occupation	Horizontal Mismatch Rate	Horizontal Under-match Rate	Horizontal Over-match Rate	Average of years of Vertical Mismatch	Vertically mis-matched occupa-tion?	Number of Undocu-likely workers by occupation
Waiters and Waitresses	0.80	0.80	0.00	3.75	1	56
Construction Laborers	0.78	0.78	0.00	4.27	1	36
Childcare Workers	0.78	0.78	0.00	3.35	1	65
Janitors and Building Cleaners	0.78	0.78	0.00	4.34	1	34
Cooks	0.77	0.77	0.00	4.28	1	77
Cashiers	0.76	0.76	0.00	4.32	1	51
Managers, all other	0.73	0.43	0.30	0.87	0	39
Clergy	0.73	0.72	0.01	1.55	0	32
Maids and Housekeeping Cleaners	0.72	0.72	0.00	4.36	1	74
Miscellaneous Agricultural Workers	0.68	0.68	0.00	4.29	1	71

Note: We restricted to the top 10 occupations populated by likely-Undocumented workers, and then sorted by horizontal mismatch. Horizontal mismatch is the mismatch of a worker's field of study and the two modal degree fields for their current occupation. Undermatch and overmatch are versions of horizontal mismatch corresponding to whether a mismatched worker's occupation holds a median wage different than that of matched workers within the degree field of the mismatched worker.

Table 5: Top 10 fields of study populated by Undocumented (RF) workers

Field of study	Horizontal Mismatch Rate	Horizontal Under-match Rate	Horizontal Over-match Rate	Average of years of Vertical Mismatch	Average rate of Vertical Mismatch	Number of Undocu-likely workers by field of study
social sciences	0.87	0.81	0.06	1.41	0.64	102
psychology	0.86	0.50	0.36	1.37	0.63	75
physical sciences	0.82	0.71	0.11	1.75	0.69	61
fine arts	0.82	0.42	0.40	1.34	0.50	78
engineering	0.64	0.49	0.15	1.46	0.72	240
biology and life sciences	0.61	0.57	0.04	1.44	0.35	81
computer and information sciences	0.55	0.46	0.08	1.16	0.67	61
medical and health sciences and services	0.33	0.28	0.04	0.79	0.51	101
education administration and teaching	0.29	0.17	0.12	0.19	0.78	181
business	0.18	0.12	0.05	1.38	0.73	397

Notes: We restricted to the top 10 fields of study populated by likely-Undocumented workers, and then sorted by horizontal mismatch. Horizontal mismatch is the mismatch of a worker's field of study and the two modal degree fields for their current occupation. Undermatch and overmatch are versions of horizontal mismatch corresponding to whether a mismatched worker's occupation holds a median wage different than that of matched workers within the degree field of the mismatched worker.

Table 6: Regressions of Undocumented Status on Vmismatch

	(1) Logical edits	(2) KNN	(3) RF
Horizontally Undermatched	0.2458***	0.2459***	0.2460***
	[0.0044]	[0.0044]	[0.0044]
Horizontally Overmatched	-0.0920***	-0.0920***	-0.0920***
	[0.0022]	[0.0022]	[0.0022]
Undocumented	0.0356***		
	[0.0047]		
Undocumented (KNN)		0.0516***	
		[0.0067]	
Undocumented (RF)			0.0573***
			[0.0079]
Mean of Dep. Var.	0.23	0.23	0.23
R-squared	0.18	0.18	0.18
N	2,406,486	2,406,486	2,406,486

Additional controls include:
dummy age indicators, gender, Medicaid reception, race/ethnicity, metropolitan residence, government occupation, English-speaking fluency, foreign born, immigration by age 10, Broad degree category indicators, years of schooling, state and year interaction fixed effects.
Robust standard errors are all clustered by state.
* $p < .1$, ** $p < 0.05$, *** $p < 0.01$

Table 7: Regressions of Undocumented Status on Hmismatch

	(1) Logical edits	(2) KNN	(3) RF
Vertically Mismatched	0.1498***	0.1499***	0.1500***
	[0.0031]	[0.0031]	[0.0031]
Undocumented	0.0402***		
	[0.0033]		
Undocumented (KNN)		0.0336***	
		[0.0043]	
Undocumented (RF)			0.0446***
			[0.0053]
Mean of Dep. Var.	0.58	0.58	0.58
R-squared	0.34	0.34	0.34
N	2,406,819	2,406,819	2,406,819

Additional controls include:
dummy age indicators, gender, Medicaid reception, race/ethnicity, metropolitan residence, government occupation, English-speaking fluency, foreign born, immigration by age 10, Broad degree category indicators, years of schooling, state and year interaction fixed effects.
Robust standard errors are all clustered by state.
* $p < .1$, ** $p < 0.05$, *** $p < 0.01$

Table 8: Regressions of Undocumented Status on H. undermatch

	(1) Logical edits	(2) KNN	(3) RF
Vertically Mismatched	0.3387***	0.3389***	0.3390***
	[0.0035]	[0.0035]	[0.0034]
Undocumented	0.0557***		
	[0.0034]		
Undocumented (KNN)		0.0525***	
		[0.0043]	
Undocumented (RF)			0.0632***
			[0.0075]
Mean of Dep. Var.	0.41	0.41	0.41
R-squared	0.27	0.27	0.27
N	2,406,486	2,406,486	2,406,486

Additional controls include:
dummy age indicators, gender, Medicaid reception, race/ethnicity, metropolitan residence, government occupation, English-speaking fluency, foreign born, immigration by age 10, Broad degree category indicators, years of schooling, state and year interaction fixed effects.
Robust standard errors are all clustered by state.
* $p < .1$, ** $p < 0.05$, *** $p < 0.01$

Table 9: Regressions of Undocumented Status on Log-Wage

	(1) Logical edits	(2) KNN	(3) RF
Vertically Mismatched	-0.3250***	-0.3249***	-0.3251***
	[0.0048]	[0.0048]	[0.0048]
Horizontally Undermatched	-0.1602***	-0.1603***	-0.1605***
	[0.0042]	[0.0042]	[0.0042]
Horizontally Overmatched	0.0486***	0.0485***	0.0485***
	[0.0034]	[0.0034]	[0.0034]
Undocumented	-0.0663***		
	[0.0067]		
Undocumented (KNN)		-0.1010***	
		[0.0084]	
Undocumented (RF)			-0.1059***
			[0.0101]
Mean of Dep. Var.	3.56	3.56	3.56
R-squared	0.29	0.29	0.29
N	2,406,486	2,406,486	2,406,486

Additional controls include:
dummy age indicators, gender, Medicaid reception, race/ethnicity, metropolitan residence, government occupation, English-speaking fluency, foreign born, immigration by age 10, Broad degree category indicators, years of schooling, state and year interaction fixed effects.
Robust standard errors are all clustered by state.
* $p < .1$, ** $p < 0.05$, *** $p < 0.01$

Table 10: Regressions of DACA-eligible Status on Vmismatch

	(1) Logical edits	(2) KNN	(3) RF
Horizontally Undermatched	0.2464***	0.2464***	0.2463***
	[0.0044]	[0.0044]	[0.0044]
Horizontally Overmatched	-0.0920***	-0.0920***	-0.0920***
	[0.0022]	[0.0022]	[0.0022]
DACA-eligible	-0.0094		
	[0.0071]		
DACA-eligible (KNN)		0.0131	
		[0.0081]	
DACA-eligible (RF)			0.0584***
			[0.0136]
Mean of Dep. Var.	0.23	0.23	0.23
R-squared	0.18	0.18	0.18
N	2,406,486	2,406,486	2,406,486

Additional controls include:
dummy age indicators, gender, Medicaid reception, race/ethnicity, metropolitan residence, government occupation, English-speaking fluency, foreign born, immigration by age 10, Broad degree category indicators, years of schooling, state and year interaction fixed effects.
Robust standard errors are all clustered by state.
* $p < .1$, ** $p < 0.05$, *** $p < 0.01$

Table 11: Regressions of DACA-eligible Status on Hmismatch

	(1) Logical edits	(2) KNN	(3) RF
Vertically Mismatched	0.1517***	0.1517***	0.1517***
	[0.0031]	[0.0031]	[0.0031]
DACA-eligible	0.0241***		
	[0.0066]		
DACA-eligible (KNN)		0.0098	
		[0.0085]	
DACA-eligible (RF)			0.0200*
			[0.0115]
Mean of Dep. Var.	0.58	0.58	0.58
R-squared	0.33	0.33	0.33
N	2,291,489	2,291,489	2,291,489

Additional controls include:
dummy age indicators, gender, Medicaid reception, race/ethnicity, metropolitan residence, government occupation, English-speaking fluency, foreign born, immigration by age 10, Broad degree category indicators, years of schooling, state and year interaction fixed effects.
Robust standard errors are all clustered by state.
* $p < .1$, ** $p < 0.05$, *** $p < 0.01$

Table 12: Regressions of DACA-eligible Status on H. undermatch

	(1) Logical edits	(2) KNN	(3) RF
Vertically Mismatched	0.3394***	0.3394***	0.3394***
	[0.0034]	[0.0034]	[0.0034]
DACA-eligible	0.0292***		
	[0.0063]		
DACA-eligible (KNN)		0.0427***	
		[0.0112]	
DACA-eligible (RF)			0.0253*
			[0.0140]
Mean of Dep. Var.	0.41	0.41	0.41
R-squared	0.27	0.27	0.27
N	2,406,486	2,406,486	2,406,486

Additional controls include:
dummy age indicators, gender, Medicaid reception, race/ethnicity, metropolitan residence, government occupation, English-speaking fluency, foreign born, immigration by age 10, Broad degree category indicators, years of schooling, state and year interaction fixed effects.
Robust standard errors are all clustered by state.
* $p < .1$, ** $p < 0.05$, *** $p < 0.01$

Table 13: Regressions of DACA-eligible Status on Log-Wage

	(1) Logical edits	(2) KNN	(3) RF
Vertically Mismatched	-0.3256***	-0.3256***	-0.3256***
	[0.0048]	[0.0048]	[0.0048]
Horizontally Undermatched	-0.1610***	-0.1610***	-0.1610***
	[0.0043]	[0.0043]	[0.0043]
Horizontally Overmatched	0.0485***	0.0485***	0.0485***
	[0.0034]	[0.0034]	[0.0034]
DACA-eligible	-0.0350***		
	[0.0093]		
DACA-eligible (KNN)		-0.0495***	
		[0.0096]	
DACA-eligible (RF)			-0.0839***
			[0.0121]
Mean of Dep. Var.	3.56	3.56	3.56
R-squared	0.29	0.29	0.29
N	2,406,486	2,406,486	2,406,486

Additional controls include:
dummy age indicators, gender, Medicaid reception, race/ethnicity, metropolitan residence, government occupation, English-speaking fluency, foreign born, immigration by age 10, Broad degree category indicators, years of schooling, state and year interaction fixed effects.
Robust standard errors are all clustered by state.
* $p < .1$, ** $p < 0.05$, *** $p < 0.01$

7 Figures

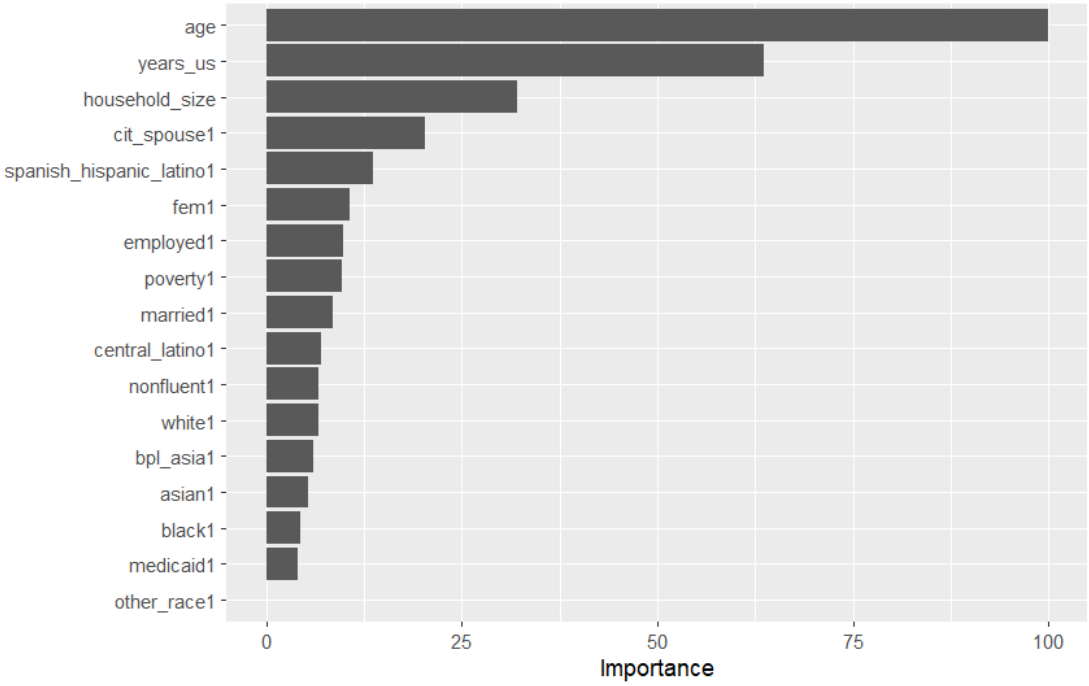

Figure 1: Feature importance plot of RF model

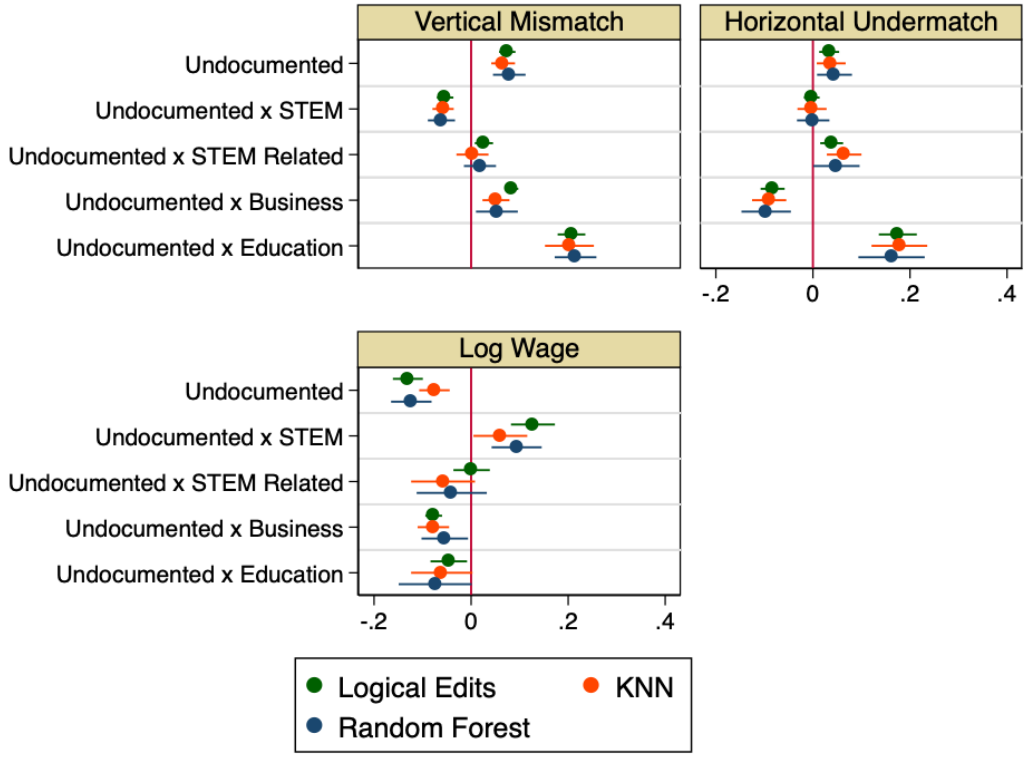

Figure 2: Coefficient plot of Undocumented status and Field of Study interactions

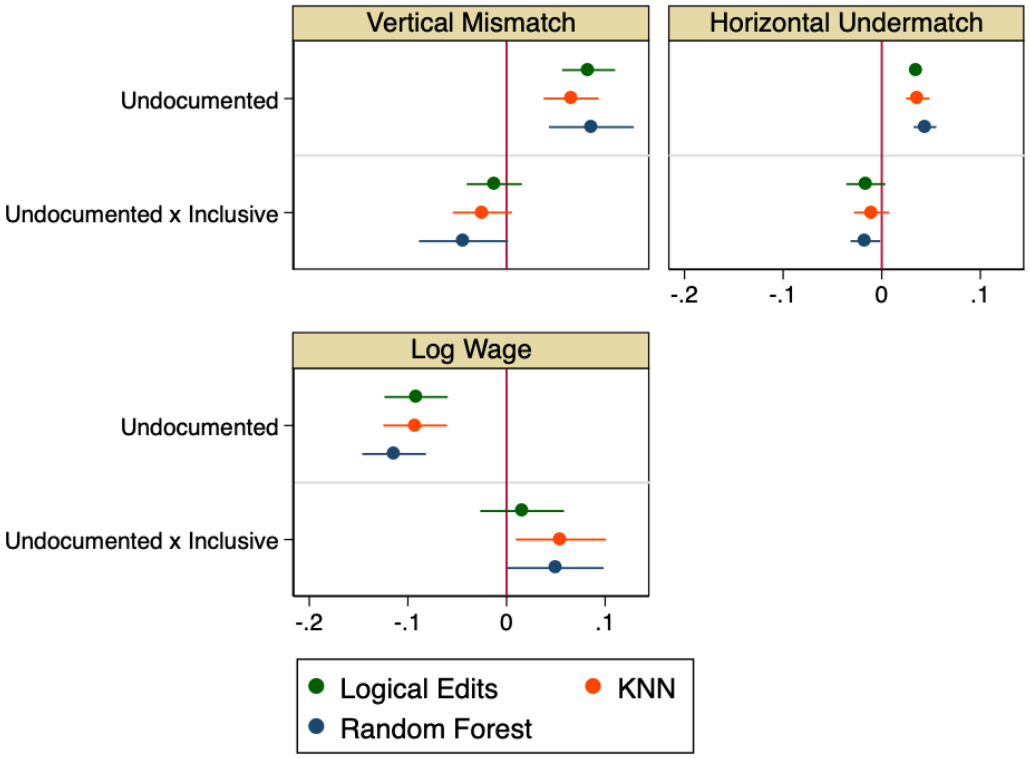

Figure 3: Coefficient plot of Undocumented Status and Inclusive State Policy Interactions

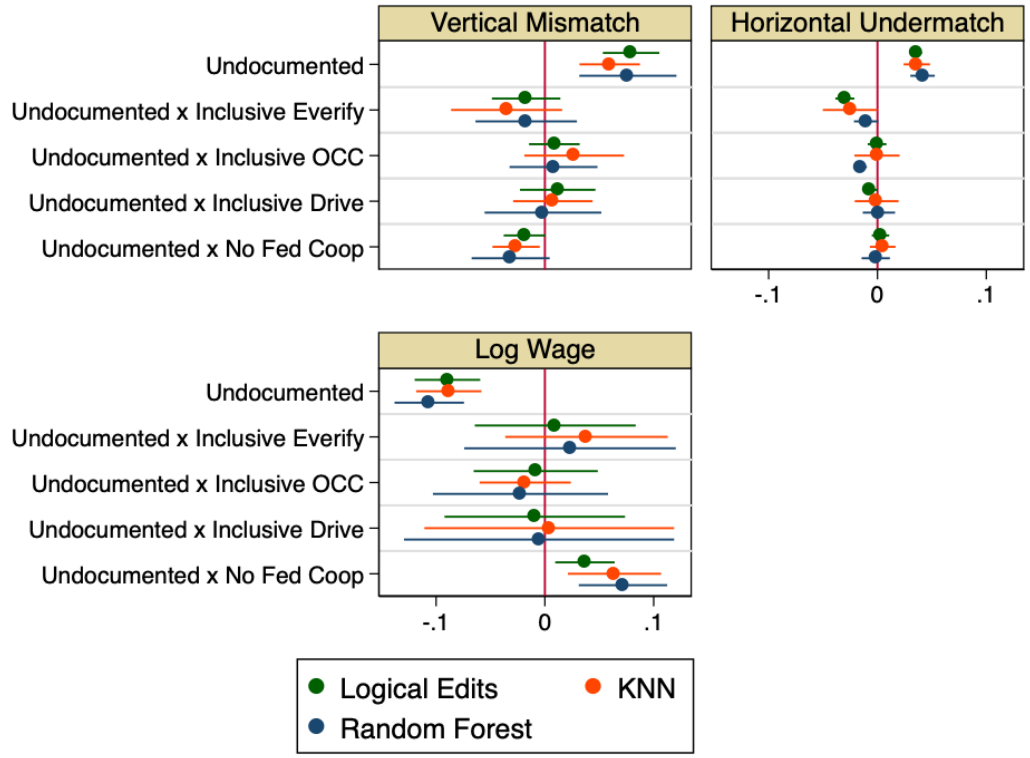

Figure 4: Coefficient plot of Undocumented status and State policy interactions

8 Appendix

	Initial SIPP sample	Noncitizens	Top 10 states	Hispanic / Latino / Spanish	Central America and Latino
logical_sensitivity	1.0000	1.0000	1.0000	1.0000	1.0000
logical_specificity	0.9547	0.0013	0.9411	0.8788	0.6304
logical_ppv	0.2857	0.2634	0.2316	0.3361	0.3780
logical_accuracy	0.9555	0.2642	0.9421	0.8858	0.6982

Table A.1: Commonly Used Sample Restrictions and their Positive Predictive Value

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

	(1) Logical edits	(2) KNN	(3) RF
Horizontally Undermatched	0.2673***	0.2695***	0.2696***
	[0.0049]	[0.0049]	[0.0049]
Horizontally Overmatched	-0.0901***	-0.0888***	-0.0887***
	[0.0021]	[0.0021]	[0.0021]
Undocumented	0.0745***	0.0660***	0.0786***
	[0.0084]	[0.0122]	[0.0167]
STEM	-0.1300***	-0.1306***	-0.1308***
	[0.0036]	[0.0037]	[0.0037]
STEM Related	-0.0893***	-0.0854***	-0.0855***
	[0.0056]	[0.0061]	[0.0062]
Business	0.0526***	0.0588***	0.0592***
	[0.0049]	[0.0053]	[0.0053]
Education	-0.0842***	-0.0778***	-0.0772***
	[0.0049]	[0.0058]	[0.0059]
Undocumented x STEM	-0.0535***	-0.0578***	-0.0612***
	[0.0084]	[0.0109]	[0.0139]
Undocumented x STEM Related	0.0264***	0.0028	0.0182
	[0.0093]	[0.0165]	[0.0165]
Undocumented x Business	0.0843***	0.0513***	0.0531**
	[0.0065]	[0.0138]	[0.0215]
Undocumented x Education	0.2069***	0.2026***	0.2152***
	[0.0142]	[0.0251]	[0.0214]
Mean of Dep. Var.	0.23	0.23	0.23
R-squared	0.14	0.14	0.14
N	2,287,291	2,287,291	2,287,291

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, state and year interaction fixed effects.
Robust standard errors are all clustered by state.
* $p < .1$, ** $p < 0.05$, *** $p < 0.01$

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

	(1) Logical edits	(2) KNN	(3) RF
Vertically Mismatched	0.3516***	0.3526***	0.3526***
	[0.0038]	[0.0037]	[0.0037]
Undocumented	0.0335***	0.0377**	0.0445**
	[0.0102]	[0.0148]	[0.0179]
STEM	0.1047***	0.1056***	0.1057***
	[0.0080]	[0.0082]	[0.0082]
STEM Related	-0.1775***	-0.1755***	-0.1749***
	[0.0052]	[0.0051]	[0.0051]
Business	-0.3959***	-0.3988***	-0.3991***
	[0.0055]	[0.0055]	[0.0055]
Education	-0.2998***	-0.2954***	-0.2948***
	[0.0049]	[0.0046]	[0.0046]
Undocumented x STEM	-0.0024	-0.0019	0.0005
	[0.0080]	[0.0150]	[0.0166]
Undocumented x STEM Related	0.0386***	0.0643***	0.0486**
	[0.0117]	[0.0178]	[0.0237]
Undocumented x Business	-0.0830***	-0.0902***	-0.0964***
	[0.0123]	[0.0176]	[0.0253]
Undocumented x Education	0.1749***	0.1782***	0.1620***
	[0.0195]	[0.0286]	[0.0340]
Mean of Dep. Var.	0.42	0.42	0.42
R-squared	0.26	0.26	0.26
N	2,287,291	2,287,291	2,287,291

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, state and year interaction fixed effects.
Robust standard errors are all clustered by state.
* $p < .1$, ** $p < 0.05$, *** $p < 0.01$

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

	(1) Logical edits	(2) KNN	(3) RF
Vertically Mismatched	-0.3670***	-0.3705***	-0.3705***
	[0.0066]	[0.0069]	[0.0069]
Horizontally Undermatched	-0.1810***	-0.1818***	-0.1818***
	[0.0054]	[0.0053]	[0.0053]
Horizontally Overmatched	0.0734***	0.0721***	0.0720***
	[0.0032]	[0.0033]	[0.0033]
Undocumented	-0.1303***	-0.0756***	-0.1233***
	[0.0154]	[0.0156]	[0.0208]
STEM	0.2290***	0.2328***	0.2328***
	[0.0076]	[0.0080]	[0.0082]
STEM Related	0.1588***	0.1561***	0.1556***
	[0.0072]	[0.0076]	[0.0078]
Business	0.1144***	0.1085***	0.1079***
	[0.0084]	[0.0091]	[0.0091]
Education	-0.1895***	-0.1915***	-0.1917***
	[0.0096]	[0.0095]	[0.0094]
Undocumented x STEM	0.1274***	0.0603**	0.0939***
	[0.0225]	[0.0276]	[0.0257]
Undocumented x STEM Related	0.0011	-0.0578*	-0.0400
	[0.0187]	[0.0328]	[0.0360]
Undocumented x Business	-0.0771***	-0.0778***	-0.0543**
	[0.0086]	[0.0161]	[0.0239]
Undocumented x Education	-0.0461**	-0.0604*	-0.0741*
	[0.0187]	[0.0315]	[0.0375]
Mean of Dep. Var.	3.56	3.56	3.56
R-squared	0.26	0.26	0.26
N	2,287,291	2,287,291	2,287,291

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, state and year interaction fixed effects.
Robust standard errors are all clustered by state.
* $p < .1$, ** $p < 0.05$, *** $p < 0.01$

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

	(1) Logical edits	(2) KNN	(3) RF
Horizontally Undermatched	0.2685***	0.2697***	0.2697***
	[0.0049]	[0.0049]	[0.0049]
Horizontally Overmatched	-0.0890***	-0.0885***	-0.0885***
	[0.0021]	[0.0021]	[0.0021]
Undocumented	0.0832***	0.0655***	0.0860***
	[0.0134]	[0.0139]	[0.0214]
Undocumented x Inclusive	-0.0124	-0.0245	-0.0436*
	[0.0139]	[0.0150]	[0.0225]
Constant	0.1334***	0.1365***	0.1367***
	[0.0020]	[0.0020]	[0.0020]
Mean of Dep. Var.	0.23	0.23	0.23
R-squared	0.14	0.14	0.14
N	2,287,291	2,287,291	2,287,291

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, state and year interaction fixed effects.
Robust standard errors are all clustered by state.
* $p < .1$, ** $p < 0.05$, *** $p < 0.01$

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

	(1) Logical edits	(2) KNN	(3) RF
Vertically Mismatched	0.3520***	0.3527***	0.3527***
	[0.0038]	[0.0037]	[0.0037]
Undocumented	0.0348***	0.0365***	0.0438***
	[0.0027]	[0.0059]	[0.0057]
Undocumented x Inclusive	-0.0162	-0.0103	-0.0166**
	[0.0099]	[0.0089]	[0.0075]
Constant	0.3315***	0.3325***	0.3326***
	[0.0009]	[0.0009]	[0.0009]
Mean of Dep. Var.	0.42	0.42	0.42
R-squared	0.26	0.26	0.26
N	2,287,291	2,287,291	2,287,291

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, state and year interaction fixed effects.
Robust standard errors are all clustered by state.
* $p < .1$, ** $p < 0.05$, *** $p < 0.01$

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

	(1) Logical edits	(2) KNN	(3) RF
Vertically Mismatched	-0.3688***	-0.3707***	-0.3707***
	[0.0068]	[0.0069]	[0.0069]
Horizontally Undermatched	-0.1810***	-0.1818***	-0.1818***
	[0.0053]	[0.0053]	[0.0053]
Horizontally Overmatched	0.0725***	0.0718***	0.0718***
	[0.0033]	[0.0033]	[0.0033]
Undocumented	-0.0917***	-0.0927***	-0.1140***
	[0.0160]	[0.0161]	[0.0161]
Undocumented x Inclusive	0.0158	0.0551**	0.0495**
	[0.0211]	[0.0228]	[0.0244]
Constant	3.7197***	3.7166***	3.7165***
	[0.0031]	[0.0031]	[0.0031]
Mean of Dep. Var.	3.56	3.56	3.56
R-squared	0.26	0.26	0.26
N	2,287,291	2,287,291	2,287,291

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, state and year interaction fixed effects.
Robust standard errors are all clustered by state.
* $p < .1$, ** $p < 0.05$, *** $p < 0.01$

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

	(1) Logical edits	(2) KNN	(3) RF
Horizontally Undermatched	0.2685***	0.2697***	0.2697***
	[0.0049]	[0.0049]	[0.0049]
Horizontally Overmatched	-0.0890***	-0.0885***	-0.0885***
	[0.0021]	[0.0021]	[0.0021]
Undocumented	0.0792***	0.0596***	0.0763***
	[0.0129]	[0.0138]	[0.0222]
Undocumented x Inclusive Everify	-0.0173	-0.0352	-0.0172
	[0.0157]	[0.0254]	[0.0232]
Undocumented x Inclusive OCC	0.0087	0.0269	0.0079
	[0.0116]	[0.0228]	[0.0201]
Undocumented x Inclusive Drive	0.0118	0.0073	-0.0018
	[0.0173]	[0.0182]	[0.0267]
Undocumented x No Fed Coop	-0.0186*	-0.0265**	-0.0315*
	[0.0096]	[0.0108]	[0.0178]
Mean of Dep. Var.	0.23	0.23	0.23
R-squared	0.14	0.14	0.14
N	2,287,291	2,287,291	2,287,291

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, state and year interaction fixed effects.
Robust standard errors are all clustered by state.
* $p < .1$, ** $p < 0.05$, *** $p < 0.01$

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

	(1) Logical edits	(2) KNN	(3) RF
Vertically Mismatched	0.3520***	0.3527***	0.3527***
	[0.0038]	[0.0037]	[0.0037]
Undocumented	0.0358***	0.0361***	0.0414***
	[0.0026]	[0.0060]	[0.0056]
Undocumented x Inclusive Everify	-0.0300***	-0.0251**	-0.0104*
	[0.0043]	[0.0125]	[0.0056]
Undocumented x Inclusive OCC	-0.0004	-0.0004	-0.0159***
	[0.0043]	[0.0103]	[0.0031]
Undocumented x Inclusive Drive	-0.0070*	-0.0007	0.0013
	[0.0035]	[0.0100]	[0.0074]
Undocumented x No Fed Coop	0.0028	0.0048	-0.0016
	[0.0040]	[0.0059]	[0.0065]
Mean of Dep. Var.	0.42	0.42	0.42
R-squared	0.26	0.26	0.26
N	2,287,291	2,287,291	2,287,291

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, state and year interaction fixed effects.
Robust standard errors are all clustered by state.
* $p < .1$, ** $p < 0.05$, *** $p < 0.01$

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

	(1) Logical edits	(2) KNN	(3) RF
Vertically Mismatched	-0.3688***	-0.3707***	-0.3707***
	[0.0067]	[0.0069]	[0.0069]
Horizontally Undermatched	-0.1810***	-0.1818***	-0.1818***
	[0.0053]	[0.0053]	[0.0053]
Horizontally Overmatched	0.0725***	0.0718***	0.0718***
	[0.0033]	[0.0033]	[0.0033]
Undocumented	-0.0895***	-0.0883***	-0.1063***
	[0.0150]	[0.0149]	[0.0159]
Undocumented x Inclusive Everify	0.0095	0.0382	0.0231
	[0.0368]	[0.0372]	[0.0484]
Undocumented x Inclusive OCC	-0.0084	-0.0181	-0.0225
	[0.0284]	[0.0208]	[0.0401]
Undocumented x Inclusive Drive	-0.0093	0.0040	-0.0054
	[0.0413]	[0.0571]	[0.0618]
Undocumented x No Fed Coop	0.0368***	0.0640***	0.0718***
	[0.0136]	[0.0213]	[0.0202]
Mean of Dep. Var.	3.56	3.56	3.56
R-squared	0.26	0.26	0.26
N	2,287,291	2,287,291	2,287,291

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, state and year interaction fixed effects.
Robust standard errors are all clustered by state.
* $p < .1$, ** $p < 0.05$, *** $p < 0.01$