

DML Analysis of Occupational Outcomes Among U.S. Immigrants

Nahom Tsegaye, Isaac Mitchell, Lingfeng Zhu

College of Engineering, Northeastern University
INFO 6105 INFO Data Science Engineering Methods and Tools
December 8, 2025

DML Analysis of Occupational Outcomes Among U.S. Immigrants.....	1
1. Introduction.....	3
2. Data and Sample Construction.....	4
2.1 Data Source.....	4
2.2 Variable Definitions.....	4
2.3 Feature Engineering.....	4
2.4 Sample Construction.....	5
3. Methodology.....	6
3.1 The Identification Challenge.....	6
3.2 Double Machine Learning Framework.....	6
3.3 XGBoost Implementation.....	7
3.4 Heterogeneity Analysis.....	8
3.5 Limitations.....	8
4.1 Model Diagnostics.....	9
4.2 Average Treatment Effect.....	10
4.3 Heterogeneity by Years in the United States.....	11
4.4 Heterogeneity by STEM Status.....	11
4.5 Heterogeneity by Origin Region.....	12
4.6 High-Error Cluster Analysis.....	12
5. Interpretation.....	14
5.1 Summary of Findings.....	14
5.2 Alignment with Prior Literature.....	14
5.3 Policy Implications.....	14
References.....	15
DataSet.....	15
Source Code.....	15

1. Introduction

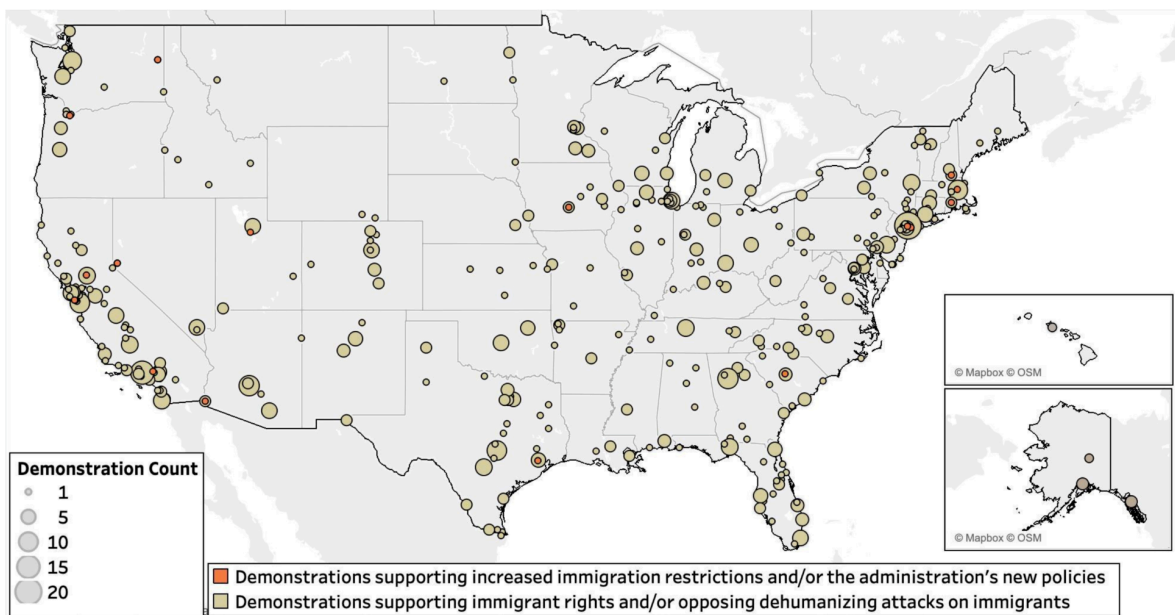
Immigration has emerged as one of the most contentious political issues across Western democracies. Public discourse frequently frames immigrants as economic threats, competitors to native-born workers, and a strain on public resources. The prevailing political narrative assumes a zero-sum framing linked to anti-immigrant policy preferences (Chinoy et al., 2023)—one in which immigrants successfully convert their human capital into occupational attainment at natives' expense. Yet public discourse and policy debates remain largely uninformed by causal evidence addressing this issue (Card & Peri, 2016).

This evidentiary gap has real consequences. As documented by the Bridging Divides Initiative (2025), immigration-related demonstrations have surged in early 2025, reflecting deepening polarization around immigrant economic integration. Understanding whether immigrants face systematic advantages or disadvantages in the labor market is therefore not merely an academic question but one with immediate policy relevance.

This study is going to analyze the effect of immigration status on occupational success using machine learning methods designed for causal inference to empirically show how the causal effect of immigrant status on occupational attainment, providing an unbiased estimate that runs counter to the rhetoric of unjust advantage.

Immigration-Related Demonstrations

January-March 2025



Data Source: ACLED with additional BDI coding

Figure 1.1 Bridging Divides Initiative. (2025). *Issue brief: Mapping the rise in immigration-related demonstrations in early 2025* [Map]. Princeton University. bridgingdivides.princeton.edu

2. Data and Sample Construction

2.1 Data Source

This study uses microdata from IPUMS USA (Integrated Public Use Microdata Series), a harmonized database of U.S. Census and American Community Survey samples (Ruggles et al., 2025). The analysis covers survey years 2000–2023, providing over two decades of labor market observations spanning multiple economic cycles, policy changes, and demographic shifts. The initial extract contained 35,805,107 person-year observations. The unit of observation is the weighted employed adult, with survey weights (PERWT) applied to generate population-representative estimates.

2.2 Variable Definitions

Outcome Variable. Our primary outcome is OCCSCORE, a constructed variable that assigns occupational income scores to each occupation based on median earnings. This measure provides a continuous indicator of occupational attainment that is comparable across time periods and geographic regions.

Treatment Variable. Immigration status (is_immigrant) is defined as a binary indicator derived from birthplace (BPL). Individuals born outside the United States and its territories are coded as immigrants.

Control Variables. The analysis controls a comprehensive set of demographic, human capital, and geographic variables including age, sex, race/ethnicity, educational attainment (EDUC), English proficiency, family structure, state of residence, and year fixed effects. Immigration-specific controls include years in the United States, age at arrival, citizenship status, and origin region.

2.3 Feature Engineering

Raw IPUMS variables were transformed into research-relevant features. The pipeline addressed structural zeros (e.g., natives have no immigration year) separately from true missing values through median imputation with missingness flags.

Key engineered features include detailed human-capital, migration, and family-structure measures. STEM degree status is derived from the respondent’s reported field of study (DEGFIELD), coded as 1 for engineering, science, mathematics, or related technical majors and 0 otherwise (5.4% of the sample). English proficiency is constructed from SPEAKENG and equals 1 for individuals who report speaking English “well” or “very well,” capturing functional language ability (96.7% proficient). Origin region is based on birthplace (BPL), originally categorized into 18 regions but collapsed into six aggregated development-level groups to reflect broad economic context and migration selectivity.

Family-structure variables include family burden, defined as the total number of children and siblings (NCHILD + NSIBS), along with binary indicators for marital status (is_married = 1 if currently married) and parenthood (has_children = 1 if NCHILD > 0).

Migration-trajectory variables are also engineered. Years in the U.S. is computed as (YEAR – YRIMMIG) for immigrants, while for natives it is set equal to their age to maintain a consistent exposure metric. This harmonized measure was specifically engineered to support the stratified DML framework used later in the analysis, to capture the assimilation process.

2.4 Sample Construction

The analytic sample was constructed through the following exclusion criteria. First, we restricted the sample to adults aged 18–65 to focus on the prime working-age population. Second, we limited the analysis to employed individuals to examine occupational outcomes conditional on labor force participation. Third, given the computational demands of causal inference methods requiring cross-fitting, we drew a 10% random sample from the full dataset. The final analytic sample contains approximately 1.79 million observations. Immigrants constitute 14.6% of the sample.

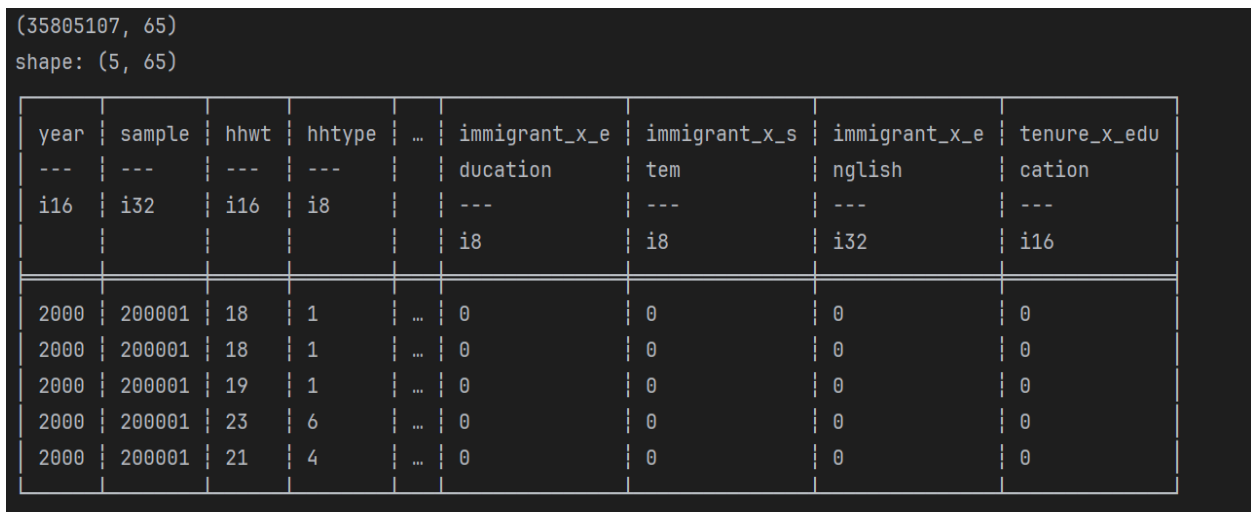


Figure 1.2 Descriptive statistics

3. Methodology

3.1 The Identification Challenge

Answering questions about immigrant labor market outcomes requires more than comparing average outcomes between immigrants and natives. Such comparisons suffer from fundamental selection bias: immigrants may possess unobservable characteristics that influence occupational success independent of immigration status itself. A naive comparison of immigrant and native OCCSCORE conflates the true causal effect of immigration status with these pre-existing differences.

3.2 Double Machine Learning Framework

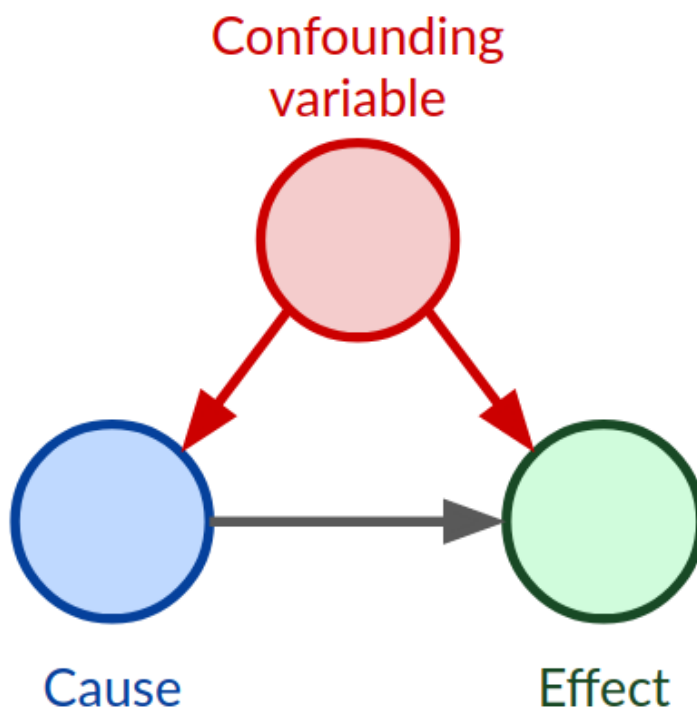


Figure 1.3 Directed acyclic graph (DAG) for the assumed causal structure.

To address this challenge, we employ Double Machine Learning (DML), a method specifically designed to estimate causal effects in high-dimensional settings with complex, nonlinear confounding (Chernozhukov et al., 2018). This approach generalizes the Frisch-Waugh-Lovell theorem (Frisch & Waugh, 1933), which demonstrates that partialling out controls yields identical coefficients to full regression. DML extends this logic to settings where the functional form of confounding is unknown. This "residual-on-residual" regression recovers quasi-exogenous variation.

$$Y = \theta T + g(X) + \varepsilon \text{ (outcome equation)}$$

$$T = m(X) + \eta \text{ (treatment equation)}$$

Where:

- Y = OCCSCORE (occupational income score)
- T = Immigration status (0/1)
- X = Vector of confounders
- θ = Causal effect of interest
- $g(X), m(X)$ = Unknown nuisance functions

Then the residuals are calculated, which are the parts of Y and T not explained by X :

- $\tilde{Y} = Y - m(X)$
- $\tilde{T} = T - g(X)$

Finally, we run a simple linear regression of \tilde{Y} on \tilde{T} :

- $\theta = \operatorname{argmin}_{\theta} \sum (\tilde{Y} - \theta \tilde{T})^2$

3.3 XGBoost Implementation

We employ XGBoost (Extreme Gradient Boosting; Chen & Guestrin, 2016) as the core estimator for both nuisance components. XGBoost is well suited for population-scale census microdata due to its strong predictive performance, robustness to outliers, and computational efficiency.

A key advantage is its capacity to approximate highly nonlinear structural relationships. Census variables such as age, education, migration history, and occupational characteristics interact in complex ways that traditional linear models cannot capture (Mullainathan & Spiess, 2017). XGBoost's additive tree structure allows flexible modeling of heterogeneous returns to education, interaction effects between immigrant tenure and occupation, and non-linear regional labor-market gradients.

For the outcome model, we fit a GPU-accelerated XGBoost regressor targeting the multimodal OCCSCORE variable, which exhibits 45 distinct occupational values with varying variance across groups. For the treatment model, we employ a binary XGBoost classifier predicting immigration status. All models incorporate IPUMS sampling weights (PERWT) to preserve representativeness across cross-validation folds.

To evaluate robustness, we conduct systematic hyperparameter sensitivity analysis across tree depth (4–8), learning rate (0.05–0.2), and regularization ($\lambda = 1-5$). Across all configurations, estimated treatment effects remain within a narrow range, suggesting that the inferred immigrant penalty is not driven by tuning choices.

3.4 Heterogeneity Analysis

While average treatment effects are informative, they may obscure substantial heterogeneity in the immigrant occupational penalty. Prior work indicates that immigrant labor-market assimilation is highly uneven across tenure, education, and country-of-origin groups (Borjas, 1999; Abramitzky et al., 2014). We therefore estimate stratified DML models along several theoretically and policy-relevant dimensions.

Years in the United States. We distinguish recent arrivals (0–5 years), mid-tenure immigrants (5–10 years), established immigrants (10–20 years), and long-term immigrants (20+ years). This stratification separates early assimilation dynamics from longer-run economic integration.

STEM Status. STEM and non-STEM degree holders often experience divergent occupational pathways due to differences in skill transferability, credential recognition, and employer preferences. Stratifying by STEM background helps isolate whether the immigrant penalty is concentrated in occupations with lower skill portability.

Origin-Country Development Level. Migrants from high-income countries differ substantially from those from developing regions in pre-migration human capital and skill transferability. This dimension helps disentangle structural skill differences from host-country discrimination effects.

Additionally, we conduct model-based clustering on high-error cases to identify data-driven subgroups whose outcomes deviate from learned patterns, potentially revealing unobserved barriers not captured by standard covariates.

3.5 Limitations

Several limitations warrant careful interpretation. OCCSCORE is constructed using 1950 income rankings with subsequent adjustments and may not fully reflect contemporary occupational wage structures (Autor & Dorn, 2013). Future work using direct wage data from the American Community Survey could address this.

Despite flexible XGBoost controls, DML inference relies on the conditional independence assumption; unobserved factors such as employer discrimination or informal networks may confound treatment effects (Fuhr et al., 2024). Audit studies pairing experimental and observational methods could help disentangle skill mismatch from demand-side barriers.

Due to computational constraints, we analyze a subsample rather than the full 35 million observations. The cross-sectional nature of our data prevents cleanly separating assimilation from cohort effects; longitudinal analysis tracking individuals across census waves would strengthen causal claims about within-person integration trajectories.

Our clustering analysis on high-error cases, though revealing heterogeneous subgroups, relies on features selected by variance rather than theoretical relevance, and the modest silhouette scores (0.15) suggest fuzzy cluster boundaries.

4. Results

4.1 Model Diagnostics

Outcome Model Performance. The XGBoost regressor achieves $R^2 = 0.277$ on held-out test data, with RMSE of 9.21 points. While modest, this performance is acceptable for DML applications where the goal is capturing confounding relationships rather than maximizing predictive accuracy. The minimal train-test gap (R^2 of 0.299 vs. 0.277) indicates appropriate regularization without overfitting.

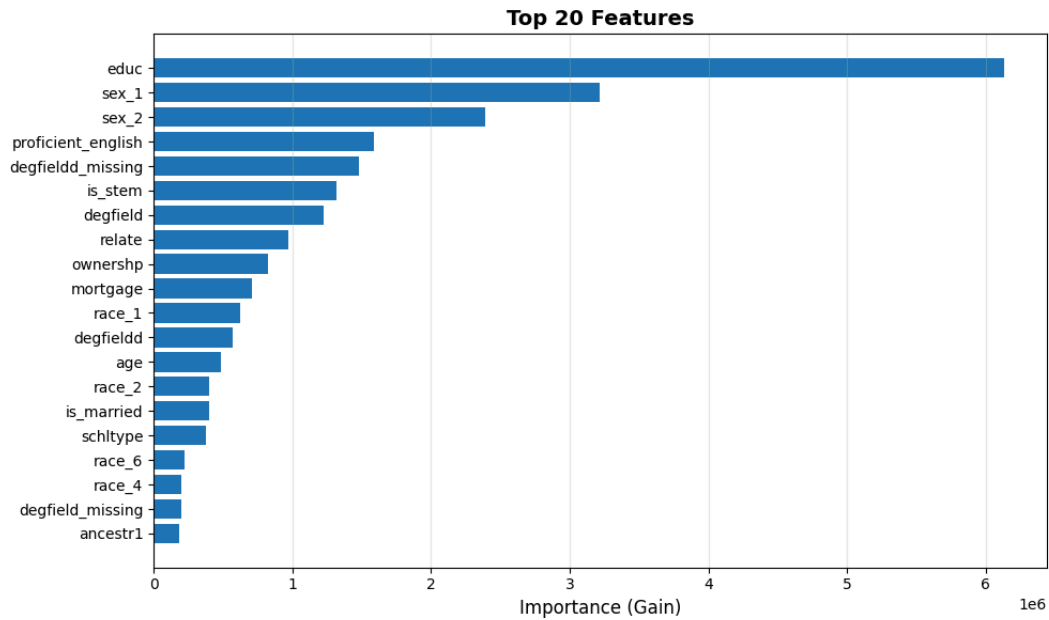


Figure 1.4 Feature importance for Outcome Model

Treatment Model Performance. The XGBoost classifier achieves test accuracy of 92.9% and AUC-ROC of 0.944, indicating excellent discrimination between immigrants and natives based on observable characteristics. This strong performance confirms that confounders substantially predict immigration status—a prerequisite for effective bias removal. Class-specific metrics show precision of 0.95 and recall of 0.98 for natives, and precision of 0.84 and recall of 0.68 for immigrants.

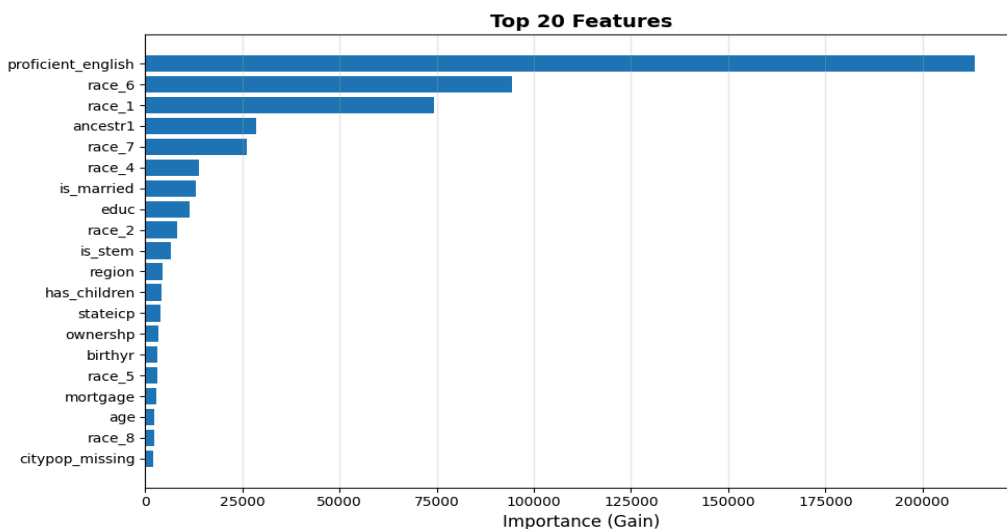


Figure 1.5 Feature importance for Treatment Model

Regression Diagnostics: the OLS model estimated on the residualized outcome indicates that the specification is statistically well-behaved. The coefficient estimates are stable, and the HC3 robust standard errors closely match the conventional OLS standard errors, suggesting that heteroskedasticity is not a meaningful concern in this setting.

Model fit statistics show an R-squared near zero, which is expected in a residual-on-residual regression after partialling out high-dimensional controls. The F-statistic (812.2, $p < 0.001$) confirms that the residualized predictor retains significant explanatory power even after orthogonalization. Distributional diagnostics—including the Omnibus and Jarque–Bera tests—indicate non-normality in the residuals, which is typical with large samples of over 1.7 million observations and does not threaten consistency. The Durbin–Watson statistic (≈ 2.00) shows no evidence of serial correlation, and the very low condition number (4.29) signals an absence of multicollinearity problems.

4.2 Average Treatment Effect

The estimated causal effect of immigrant status on occupational attainment is -0.84 OCCSCORE points (95% CI: $-0.90, -0.78$; $p < 0.001$). This effect is statistically significant at all conventional levels and robust to heteroskedasticity correction.

The effect size is modest in absolute terms—approximately 0.84 points on a scale with standard deviation of 9.2—but represents a 2.8–3% relative penalty when evaluated against mean OCCSCORE of 28–30. The narrow confidence interval indicates high estimation precision.

This finding contradicts the zero-sum narrative suggesting immigrants gain occupational advantage at natives' expense. Instead, immigrants face systematic disadvantage in converting human capital to occupational attainment.

4.3 Heterogeneity by Years in the United States

Table 1: Immigrant Penalty by Tenure

Years in the U.S.	N	Effect	SE	95% CI	p-value
0–5 years	3,127,190	–1.80	0.069	[–1.93, –1.66]	<0.0001
5–10 years	3,128,207	–1.94	0.064	[–2.07, –1.81]	<0.0001
10–20 years	3,204,923	–1.34	0.047	[–1.43, –1.25]	<0.0001
20+ years	3,298,237	–0.40	0.038	[–0.47, –0.32]	<0.0001

Taken together, the interaction between time in the United States and occupational attainment exhibits a clear pattern of gradual but incomplete economic assimilation. Immigrants who have been in the country less than five years face a large occupational penalty of –1.80 points, reflecting in part the formidable barriers at the point of entry due to language, the lack of local work experience, and limited recognition of foreign credentials. The penalty is largest among immigrants who have been in the United States for 5–10 years (–1.94), indicating that early integration does not immediately translate to improved occupational outcomes. Among those who have been in the United States for 10–20 years, the penalty begins to dissipate somewhat (–1.34), which suggests meaningful-though uneven-progress toward socioeconomic integration. For long-term immigrants who have resided more than twenty years in the country, the penalty substantially narrows to –0.40, demonstrating considerable occupational convergence with U.S.-born workers. Overall, these findings comport with a refined assimilation framework wherein immigrant occupational mobility improves steadily with time but remains persistently shaped by systemic challenges that impede complete convergence.

4.4 Heterogeneity by STEM Status

Table 2: Immigrant Penalty by Field

Field	N	Effect	SE	95% CI	p-value
STEM	3,110,144	–1.35	0.100	[–1.55, –1.16]	<0.0001
Non-STEM	3,529,715	–0.71	0.031	[–0.77, –0.65]	<0.0001

Contrary to the expectation that technical skills transfer easily across borders, STEM degree holders face a penalty nearly twice that of non-STEM workers. This finding suggests that highly educated immigrants systematically work below their qualification levels. This also indicates credential recognition barriers in regulated STEM occupations (engineering licensure).

4.5 Heterogeneity by Origin Region

Table 3: Immigrant Penalty by Origin

Origin Region	N	Effect	SE	95% CI	p-value
High Income Western	3,097,410	+0.60	0.072	[0.46, 0.74]	<0.0001
Upper Middle Europe	3,116,326	-0.57	0.062	[-0.69, -0.45]	<0.0001
MENA/Africa	3,089,852	-0.59	0.099	[-0.79, -0.40]	<0.0001
High Income Asia	3,110,833	-0.87	0.106	[-1.08, -0.66]	<0.0001
Latin America	3,302,345	-1.23	0.044	[-1.32, -1.15]	<0.0001
Developing Asia	3,158,053	-1.27	0.093	[-1.45, -1.09]	<0.0001

The gradient follows economic development of origin countries. Only immigrants from high-income Western countries experience an occupational premium (+0.60 points), suggesting their credentials and skills are readily recognized in U.S. labor markets. Immigrants from developing Asia and Latin America face the largest penalties, consistent with greater barriers to credential recognition and potentially discrimination. High Income Asia penalty (-0.87) is surprising and may reflect language barriers and occupational licensing restrictions.

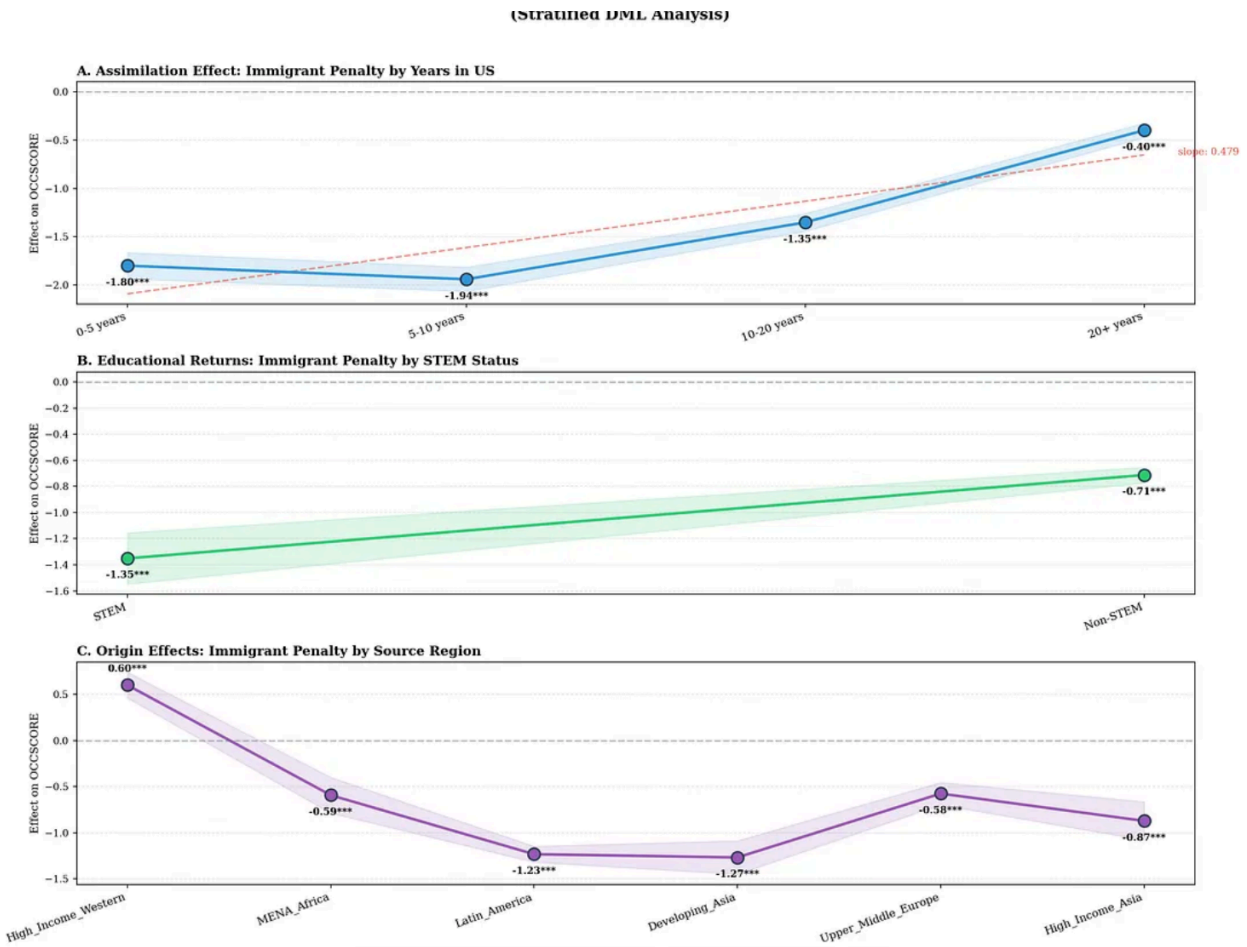


Figure 1.6 Heterogeneity plot graphs

4.6 High-Error Cluster Analysis

To identify sources of heterogeneity not captured by pre-specified stratification, we conducted K-means clustering on the top 5% of cases by absolute prediction error ($|\text{residual}| > 17.43$; $N = 89,513$). The optimal number of clusters was $K = 3$ based on silhouette scores.

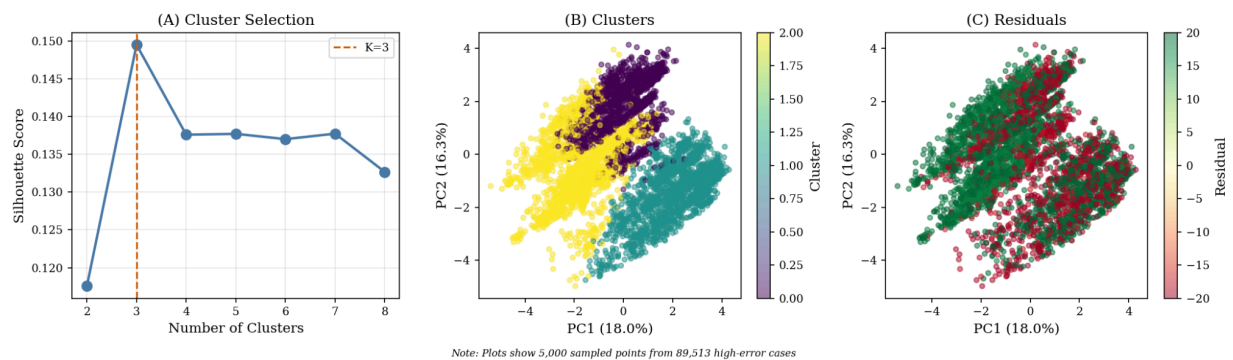


Figure 1.7 Cluster selection and separation

Cluster 0: Low Education, Largest Penalty. This cluster exhibits the strongest immigrant penalty (−3.23 points, $p < 0.001$), the lowest educational attainment (low degree-field codes), and mean OCCSCORE of 39.1. These individuals likely face structural barriers including credential scarcity and vulnerability to occupational segmentation.

Cluster 1: Highly Educated, Moderate Penalty. Despite having the strongest human capital profiles (highest degree-field codes, mean OCCSCORE of 46.1), immigrants in this cluster still experience significant penalties (−2.04 points, $p = 0.008$). This pattern reflects brain waste—highly educated immigrants working below qualification levels due to licensing barriers and credential non-recognition.

Cluster 2: Mixed Background, No Significant Penalty. This cluster shows no statistically significant penalty (−0.82 points, $p = 0.095$) and the highest average residual (+13.23), indicating the model consistently underpredicts their success. Distinct ancestral composition suggests these may be communities with strong socioeconomic networks or advantageous settlement patterns that facilitate integration.

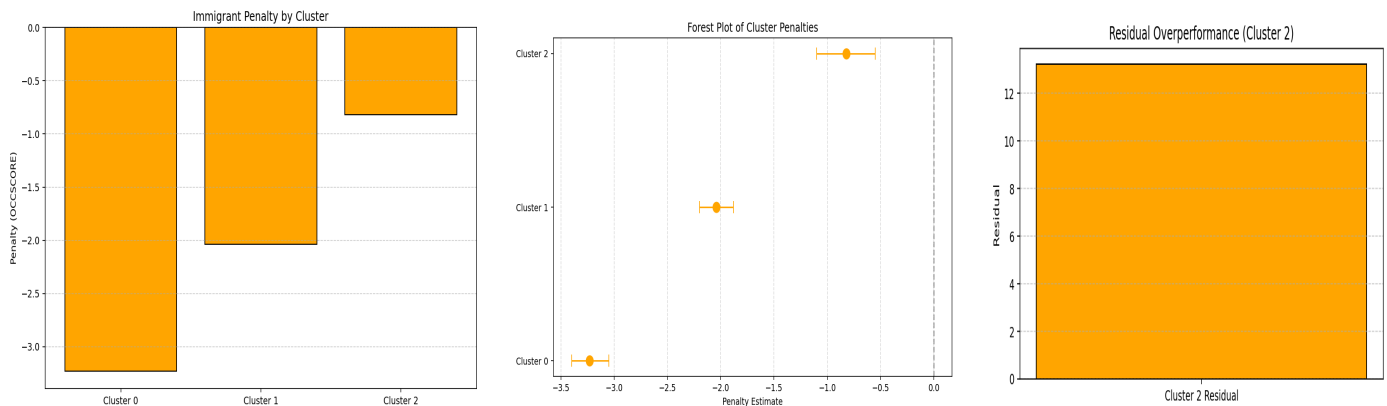


Figure 1.8 Cluster diagnostics

5. Interpretation

5.1 Summary of Findings

This study provides causal evidence on the relationship between immigrant status and occupational attainment using Double Machine Learning methods applied to over two decades of U.S. Census microdata. Our findings challenge prevailing narratives that frame immigrants as gaining unfair economic advantage.

The average immigrant faces a modest but statistically significant occupational penalty of -0.84 points, representing approximately 3% lower occupational attainment than comparable native-born workers. However this effect is averaged across all time periods, regions, and origin groups. It cannot discriminate between important cofounder effects like credential non-recognition and English proficiency. This was the main driving force behind using stratified DML analysis

5.2 Alignment with Prior Literature

Our findings align with and extend several established results in immigration economics. The overall penalty magnitude is consistent with the 4–6% wage penalty documented by Borjas and Cassidy (2019). The assimilation pattern—with penalties declining over time but never fully converging matches Dustmann and Preston's (2008) findings on occupational downgrading dynamics. The STEM paradox supports Batalova and Fix's (2021) brain waste hypothesis regarding credential non-recognition for highly skilled immigrants. The origin-region heterogeneity reflects global inequality in credential recognition systems documented by Kreimer (2024).

Our methodological contribution lies in applying DML with flexible machine learning estimators, providing more robust causal estimates than the OLS regressions common in earlier work (cf. Chiswick, 1978). This approach flexibly controls for complex confounding while maintaining valid statistical inference.

5.3 Policy Implications

These findings have several policy implications. The persistence of occupational penalties even among long-settled immigrants suggests structural barriers beyond individual assimilation efforts. Policies targeting credential recognition and occupational licensing reform may be more effective than those focused solely on language training or job search assistance. The large penalty for STEM workers highlights inefficiencies in how the U.S. labor market integrates highly skilled immigrants. Given policy emphasis on attracting STEM talent, the brain waste documented here represents both individual hardship and aggregate economic loss. Integration support should be tailored to specific immigrant high volume populations rather than applied uniformly.

For data practitioners, the DML + XGBoost pipeline demonstrated here is readily adaptable to other policy evaluation contexts — the modular codebase can be retrained on state-level data to identify regional variation in immigrant penalties, helping local agencies prioritize interventions. Workforce development agencies could integrate similar models into their intake systems, flagging immigrants with high predicted skill-occupation mismatch for targeted credential evaluation programs. For policymakers, the stratified analysis offers a template: rather than treating immigrants as a monolithic group, agencies should segment by tenure, origin region, and field of study when designing integration programs. Concretely, the finding that STEM penalties exceed non-STEM penalties suggests that licensing boards in engineering and healthcare should be priority targets for credential recognition reform. The origin-region gradient further implies that bilateral credential agreements with high-volume source countries — particularly in Latin America and Asia — would yield the largest returns. Finally, the 5–10 year penalty peak indicates that integration support should extend well beyond the initial settlement period, with employment services proactively reaching out to mid-tenure immigrants rather than focusing exclusively on new arrivals.

References

- Abramitzky, R., Boustan, L., & Eriksson, K. (2014). *A nation of immigrants: Assimilation and economic outcomes in the early 20th century*. *Journal of Political Economy*, 122(3), 467–506.
- Autor, D., & Dorn, D. (2013). *The growth of low-skill service jobs and the polarization of the US labor market*. *American Economic Review*, 103(5), 1553–1597.
- Batalova, J., & Fix, M. (2021). *Leaving money on the table: The persistence of brain waste among college-educated immigrants*. Migration Policy Institute.
- Borjas, G. J. (1999). *The economic analysis of immigration*. In O. Ashenfelter & D. Card (Eds.), *Handbook of Labor Economics* (Vol. 3, pp. 1697–1760). Elsevier.
- Borjas, G. J., & Cassidy, H. (2019). *The wage penalty to undocumented immigration*. *Labour Economics*, 61, 101757.
- Bridging Divides Initiative. (2025). *Issue brief: Mapping the rise in immigration-related demonstrations in early 2025*. Princeton University.
- Card, D., & Peri, G. (2016). *Immigration economics by George J. Borjas: A review essay*. *Journal of Economic Literature*, 54(4), 1333–1349.
- Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). *Double/debiased machine learning for treatment and structural parameters*. *The Econometrics Journal*, 21(1), C1–C68.
- Chinoy, S., Nunn, N., Sequeira, S., & Stantcheva, S. (2023). *Zero-sum thinking and the roots of US political divides* (NBER Working Paper No. 31688). National Bureau of Economic Research.
- Chiswick, B. R. (1978). *The effect of Americanization on the earnings of foreign-born men*. *Journal of Political Economy*, 86(5), 897–921.
- Dustmann, C., Frattini, T., & Preston, I. (2013). *The effect of immigration along the distribution of wages*. *Review of Economic Studies*, 80(1), 145–173.
- Frisch, R., & Waugh, F. V. (1933). *Partial time regressions as compared with individual trends*. *Econometrica*, 1(4), 387–401.
- Fuhr, J., Berens, P., & Papies, D. (2024). *Estimating causal effects with double machine learning—A method evaluation* (arXiv:2403.14385).
- Kreimer, L. (2024, November 14). *Highly skilled immigrants face a changing landscape for credential recognition*. Migration Policy Institute.
- Mullainathan, S., & Spiess, J. (2017). *Machine learning: An applied econometric approach*. *Journal of Economic Perspectives*, 31(2), 87–106.

DataSet

IPUMS Dataset: Ruggles, S., Flood, S., Sobek, M., Backman, D., Cooper, G., Richards, S., Rodgers, R., Schroeder, J., & Williams, K. C. W. (2025). *IPUMS USA: Version 16.0 [Dataset]*. Minneapolis, MN: IPUMS.

Source Code

Source code available at <https://github.com/nahomaraya/migration-occupation-model>

