Boosted Migration: The Effect of Migration on Low-Wage Workers

Andrés Rengifo-Jaramillo *

October 25, 2025

Abstract

This paper examines the labor market effects of Venezuelan migration on low-wage Colombian workers. Using nationally representative survey data from 2012 to 2019 and a machine learning model to classify workers by their predicted probability of being in the lower tail of the wage distribution, I implement a Triple Differences-in-Differences strategy to estimate causal impacts across cities with varying levels of migrant exposure. The main finding is that, although migration had no significant effects on employment formality or self-employment, it did lead to a decline in wages among low-wage natives in high-migration cities. This effect is robust to alternative definitions of treatment and control groups. This finding suggests that immigration can exert downward pressure on earnings at the lower end of the wage distribution and underscore the importance of protecting vulnerable groups during large-scale displacement events.

Keywords: Migration; Colombia; Triple Differences-in-Differences; Machine Learning.

JEL classification: J01, J15, J61, C45

*Independent Researcher. Email: afrj1996@gmail.com. I first conceived the idea of this paper while I was working at Harvard Business School, Digital Reskilling Lab as a Research Assistant

1 Introduction

The recent mass migration from Venezuela represents one of the largest displacement crises in modern Latin American history. By May 2025, according to the inter-agency coordination platform for refugees and migrants, roughly 6.87 million Venezuelans had left their country and had settle in other countries in Latin America and the Caribbean, with Colombia hosting the largest share—more than 2.8 million people, most of whom arrived within a few years following the 2016 reopening of the border (R4V, 2025). This sudden and large inflow of migrants has raised critical questions about the capacity of host labor markets to absorb newcomers and the potential consequences for native workers, particularly those in vulnerable positions.

While the literature on the labor market impacts of immigration is extensive, most empirical evidence comes from high-income countries with relatively structured migration regimes and strong labor institutions[1]. In contrast, low- and middle-income countries face very different conditions. Labor markets in these settings are often informal, fragmented, and weakly regulated (Breza and Kaur, 2025). As a result, the mechanisms through which migration affects native workers may differ substantially from those documented in the Global North and remain understudied.

This paper examines the short-run impact of Venezuelan migration on labor market outcomes for Colombian workers, focusing on low-wage individuals who are more likely to compete with migrants for low-paying jobs: even if the skill composition of the migrant population is similar to that of the natives, they suffer from a downgrade in their returns to skills, as documented in the Venezuelan case (Santamaria, 2022), which leads to high competition for low-wage jobs[2]. Motivated by this fact, I apply several machine learning models trained on pre-crisis data to estimate natives' probability of being in the lower tail of the wage distribution. Then, I leverage the 2016 border reopening and the subsequent deterioration in the living conditions in Venezuela, which caused a sharp increase in migrant flows after this event, and historical migration patterns across cities in a triple-difference in difference (DDD) strategy that compares pre- and post-border reopening differences in outcomes of low-wage and non-low-wage individuals between cities with high and low historical Venezuelan migration.

The use of machine learning techniques to identify the potential exposed population allows for a more flexible and data-driven identification of potentially affected individuals, reducing reliance on arbitrary definitions of the treatment and control groups [3]. Furthermore, a key strength of the methodology is that, in the training stage, I only employ individual characteristics unlikely to be influenced by the shock in the short term [4]. This careful variable selection allows the prediction model to retain validity after the shock. Additionally, given that I use predictors that are defined for all individuals (not only those who are working), I can classify all individuals who are unemployed

[1] For example, from all the papers listed by Dustmann et al. (2016) only one used data from a developing country (Malaysia).

[2] See also subsection 3.2 below for an analysis of this phenomenon. Similarly, Dustmann et al. (2013) report evidence of downgrading of migrants in the UK.

[3] This approach, based on Card and Krueger (1995), was used by Cengiz et al. (2022) to analyze the impact of minimum wage increases on low-wage workers, and for Card et al. (2024) to predict stated gender preferences in job postings.

[4] For example, I use house ownership status and house amenities, such as clean water supply. A detailed description of all the variables used to train the model can be found in appendix B and table B.2.

or inactive. This allows to determine the effect of the migration shock on low-wage employees' unemployment and participation rates.

The research design relies on the assumption that in the absence of the migration shock, the difference in outcomes between classified low-wage and non-low-wage workers would evolve similarly across cities. I provide evidence of the plausibility of the parallel trends assumption before the border re-opening, using an event study specification, and show a lack of significance of the coefficients associated with the pre-treatment period for all the studied outcomes.

I first document the strong capacity of the trained models to predict low-wage workers. Out of sample performance calculations reveal that the model can achieve high levels of precision while maintaining high levels of recall. For example, for a 75% recall, the best model, XGBoost, achieves a precision of 52%[5]. Then I use the predicted probabilities as input for the DDD estimation. The main findings reveal a negative and statistically significant effect on log wages for low-wage native workers in cities with high exposure to migration. Specifically, I estimate a 7.1 percent reduction in wages for this group after the border reopening, consistent with downward pressure on wages due to increased labor supply from migration inflows. In contrast, no significant effect was found on the probability of unemployment. Additionally, I document a positive and marginally significant effect on labor market participation in the first year after the border re-opening.

Furthermore, I show that these results mask heterogeneity across different groups: The negative impact on wages increases with age as young workers (aged 15 to 28) face a wage reduction of about 10% while workers aged 41 or more face a 5% reduction. Similarly, I find a negative and significant effect on unemployment only for women. Then, I also explore heterogeneous effects on wages across industries and I find that industries with a high share of small firms or a high share of self-employed workers experience a less pronounced reduction in wages. Finally, I also explore the effect of the migration shock on other labor market variables and document null results on the probability of having a formal job or being self-employed. This indicates that the main adjustment channel is the decrease in wages for low-wage individuals.

This paper advances the broader immigration literature by extending insights from high-profile debates in developed economies, such as the Mariel Boatlift in the United States, to the context of South-to-South migration in middle-income countries like Colombia [6]. Furthermore, it introduces a robust, innovative methodology that leverages machine learning to identify workers potentially impacted by migration. I show that the machine learning models outperform other approaches conceived to identify workers with a high risk of being exposed to migration using observable characteristics such as education (Altonji and Card, 1991).

A growing body of empirical research has investigated the effects of Venezuelan migration on the Colombian labor market. For example, Bonilla-Mejía et al. (2024); Caruso et al. (2019); Delgado-Prieto (2024); Otero-Cortés et al. (2022); Pedrazzi and Peñaloza-Pacheco (2023), employ an Instrumental Variable (IV) approach leveraging spatial variation on previous migration patterns and temporal variation from different indicators of the Venezuelan economic crisis. These instruments fall into the Bartik-instrument category, which combines local pre-shock migrant shares with time-

[5] Similarly, Cengiz et al. (2022) report that for a recall of 75% their preferred model achieves a 35% precision rate. Our model also achieves a similar performance to that reported in Bazzi et al. (2022). See section B for more details about precision and recall curves.

[6] See for example Borjas (2017); Card (1990) or Monras (2021) as exponents of this debate.

varying national-level shocks (Goldsmith-Pinkham et al., 2020). While these strategies offer a compelling solution to the endogeneity of migrant location choices, they present challenges of their own. Since there is only one time varying national level shock: the deterioration of Venezuelan conditions; measured using the price index as in Bonilla-Mejía et al. (2024) and Otero-Cortés et al. (2022) or the number of migrants crossing the border, as in Delgado-Prieto (2024); the validity of the instruments hinges on the exogeneity of pre-shock migrant shares (Borusyak et al., 2021; Goldsmith-Pinkham et al., 2020). This assumption may not hold given potential unobserved confounders, such as regional policy differences or historical socioeconomic factors. For instance, border cities may have unique labor markets or social integration dynamics that could correlate with both the instruments and the outcomes, potentially violating the exclusion restriction.

The identification strategy employed in this paper, a Triple Difference in Difference, allows for differential trends for the exposed groups (low-wage workers) across high and low historical migration cities, relaxing the identification assumption of Bartik instruments. Furthermore, the methodology plausibly allows the study of other behavioral responses of potentially low-wage workers, such as labor market participation and unemployment since we can classify an individual as low-wage worker, even if that individual is not working. Also, I show that the machine learning approach performs better than simple rules that classify potentially low-wage individuals based on commonly used observable demographics such as age, gender, and education[7].

This paper is also related to the emerging empirical literature that uses machine learning techniques in applied econometrics (Kleinberg et al., 2018; Lee et al., 2010; Mullainathan and Spiess, 2017). Angrist and Frandsen (2022), in particular, explore the use of data-driven machine learning models in empirical Labor Economics. More related is the investigation of Cengiz et al. (2022) that uses machine learning techniques to identify workers more exposed to minimum wage increases. This paper extends this methodology to the study of the effects of migration and shows how this methodology can be integrated into a causal inference framework.

We start in the next section with a brief discussion of the context of the Venezuelan exodus. Section 3 describes our data and presents descriptive statistics. Section 4 presents the results of the prediction tasks and Section 5 presents our empirical strategy. Section 6 presents the main results and robustness checks. Section 7 explores heterogeneity. Finally, section 8 concludes. The Appendix contains additional tables and figures and further description of the implementation of the machine learning approach to predict low-wage workers.

2 Background

The Venezuelan crisis intensified following the death of President Hugo Chávez in 2013 and a sharp decline in global oil prices, sparking profound economic, political, and humanitarian turmoil (Chaves-González and Echeverría-Estrada, 2020). This downturn exposed the vulnerabilities of an economy overly reliant on petroleum revenues, which had funded extensive social programs under the Bolivarian Revolution initiated in 1999. Policies including expropriations of private property,

[7] Cengiz et al. (2022) point to the fact that several investigations have focused on teenagers, or low-educated individuals in search for a demographic group of individuals earning the minimum wage. I focus on similar classifications and found that the machine learning model outperforms such a demographic classification approach.

constitutional changes, and centralized control eroded institutional stability and private sector vitality, leading to endemic economic distress (Vera, 2015). By 2015, oil prices had halved, crippling government finances and triggering shortages in subsidized essentials like food and medicine (Neuman, 2015). Furthermore, the Macroeconomic repercussions were severe: hyperinflation hit approximately 130,000% in 2018, while GDP contracted by double digits annually from 2016, plummeting to -30% in 2020 (IMF, 2025).

These factors fueled a massive exodus of Venezuelans who left their country in search of better economic opportunities and to avoid political violence (European Parliament Research Service, 2018). Diplomatic strains compounded the situation; in 2015, tensions over alleged Colombian armed groups in Venezuela prompted President Maduro to close the border, halting trade and migration flows initially in Táchira state (Macias, 2015). This closure lasted until July 6, 2016, when it reopened amid protests in Venezuelan border areas (Santamaria, 2022). Figure 1 shows the impact that the reopening of the border had on the migration flows (Panel A) and the composition of Colombia's labor market (Panel B). Both series illustrate the sharp increase in Venezuelan migration starting in right after the border re-opening and the subsequent increase in the share of migrants in labor force.

Migrants primarily entered via land crossings like the Simón Bolívar International Bridge in Cúcuta, Páez Bridge in Arauca, and Paraguachón in Maicao, often undocumented and traveling onward by foot, hitchhiking, or bus. Colombian policies allowed entry with expired documents, though many used irregular paths (Cancilleria de Colombia, 2019). To facilitate integration, Colombia introduced regularization efforts, including the Special Permit of Permanence (PEP) in late 2018, enabling undocumented Venezuelans to access formal employment without sponsors or investments (Bahar et al., 2022).

3 Data and Descriptive Statistics

This section describes the data sources, the process to obtain the final sample, and a description of the process of labor market accommodation of Venezuelan migrants.

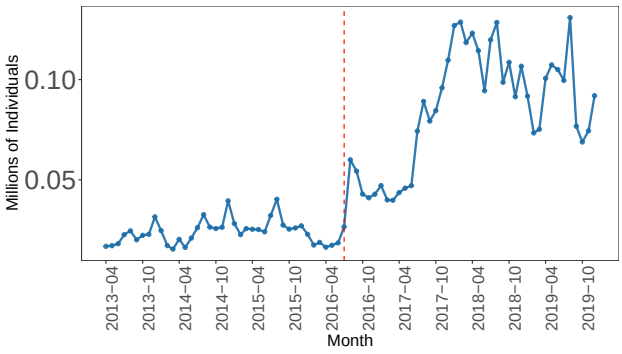
3.1 Data

The main data source is the Gran Encuesta Integrada de Hogares (GEIH, by its acronym in Spanish), Colombia's most important household survey focused on labor market outcomes. The GEIH is collected monthly and is nationally representative for the 23 largest metropolitan areas, which together represent over half of the Colombian population (Bonilla-Mejía et al., 2024). As the GEIH does not produce representative estimates for smaller cities or rural areas, our analysis is limited to urban residents in these major metropolitan areas.

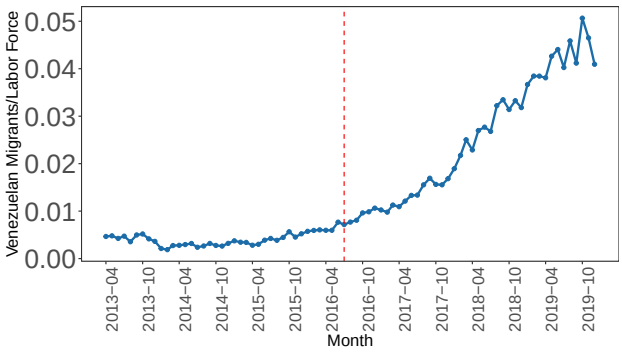
This dataset contains information about the salary, employment, and demographic information such as age, gender, family structure, education choices, and housing characteristics. I use a rich set of household characteristics and household structure variables and other demographic variables to feed the prediction model. Given our focus on labor market effects, the sample is restricted to individuals

Figure 1: Migration shock: Venezuelan Influx and Share on Labor Force

(a) Venezuelans Flux



(b) Venezuelan Share on Labor Force



Note: Figure 1 shows the evolution of the Venezuelan migration to Colombia. Panel (a) displays the monthly inflow of Venezuelan migrants (in millions). Panel (b) displays the evolution of the share of Venezuelan migrants in the labor force. Migrants are defined as individuals who were living in Venezuela 5 years prior to the survey date. *Source:* Panel A: Author's own calculations from Migración Colombia, Panel B: Author's own calculations from GEIH.

of working age (between 15 and 65 years old) in urban areas [8].

The different rounds of this survey allow the construction of a repeated cross-section for several geographic entities. I employ the data from 2012 to 2019. To construct the prediction model, I use the 2012 survey round as the primary sample. I chose this round because it predates the deepening of the Venezuelan social crisis, allowing the prediction model to be based on data generated before the large influx of Venezuelan migrants into the labor market. Section 4 describes in detail the construction of the training and test samples and the process of training the model. I use the migration module from the survey rounds of 2013 to 2019 to identify the migrants and native individuals. I define a migrant as an individual that were living in Venezuela 5 years prior to the survey date. I also use information from *Migración Colombia*, which collects data from all the foreign individuals in the country, and the extract of the 2005 Colombian census from IPUMS to calculate the share of Venezuelans in the labor force in each of the main cities in 2005 as a measure of previous settlement patterns.

[8] See table B.2 for a full description of the variables used to train the model.

3.2 Arrival of Venezuelan Migrants and Labor Market Adjustments

How does the local labor markets have absorbed the migration shock? Figure 2 reveals a high concentration of wages below the minimum wage, particularly for migrant workers, and a sharp peak for both migrants and natives. The broader distribution for native workers, with a more pronounced tail extending beyond one, indicates greater wage diversity and possibly better access to higher wage tiers.

Figure 2: Wage distribution of Natives and Migrant Workers

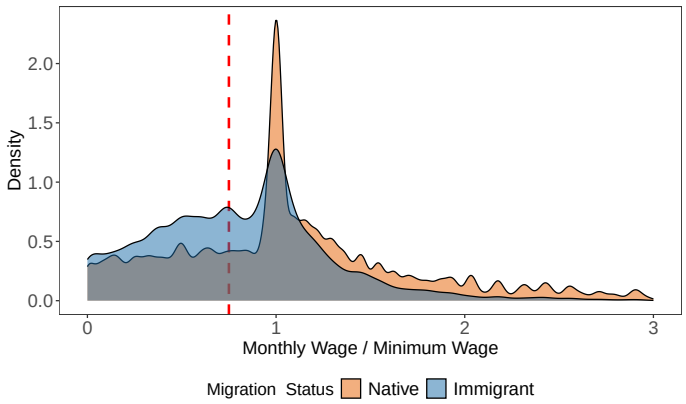


Figure 2 shows the wage distribution for native (orange) and Immigrant (blue) workers. For expositional purposes, the graph only depicts observations between 0 and 3 minimum wages, but the density is calculated for the entire distribution. The red line marks the threshold for low-wage workers, set at 75% of the monthly minimum wage. Migrants are defined as individuals who were living in Venezuela 5 years prior to the survey date. *Source:* Author's own calculations from the GEIH.

The red line at 75% of the minimum wage underscores a critical insight: a significant number of workers, especially immigrants, fall below this low-wage threshold. This finding points to economic vulnerability among immigrant populations, where approximately half of the immigrant wage distribution lies below or near this cutoff. This motivates the focus on low-wage workers as the population more likely to be affected by the migration shock since migrants are arguably workers who face a non-binding minimum wage and compete with other native workers facing the same conditions.

Is there any spatial difference in the conditions of migrants? Figure 3 suggests that there are spatial differences in the wage conditions of Venezuelan migrants depending on whether they reside in high-migration cities (measured using 2005's migrant shares in the labor force) or not, especially after 2016. Across all years from 2013 to 2019, the left tail of the distribution of wages for Venezuelan immigrants (relative to the minimum wage) tends to be thicker in high-migration cities (blue boxes) compared to other cities (orange boxes). This pattern is particularly evident in the years after 2016, where the wage gap between median wage for high and low migration cities appears to widen. This pattern is consistent with an increase in competition for low-wage jobs after the border re-opening, not only between migrants but also, potentially, between native and migrant workers. This observation motivates the empirical strategy of comparing low-wage and non–low-wage individuals across cities with different historical migration patterns.

Figure 3: Wage distribution for Venezuelan Migrants in High and Low migration Cities

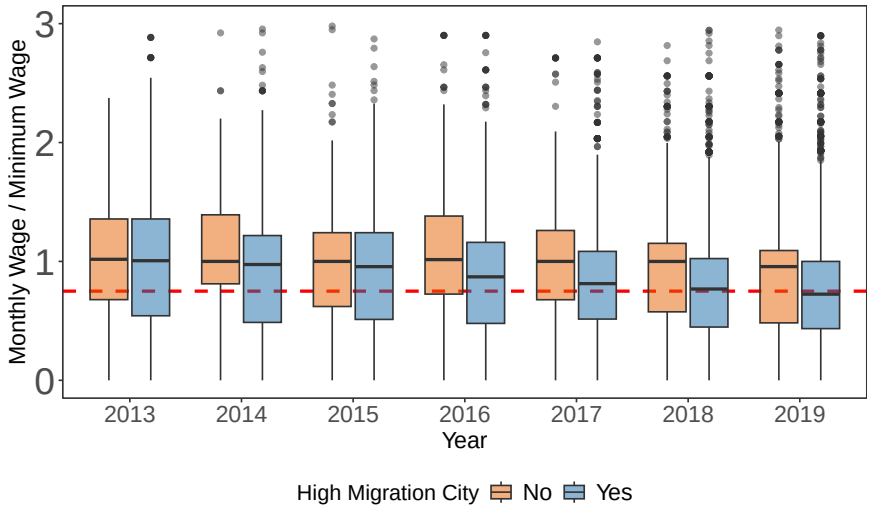


Figure 3 shows the distribution of monthly wages relative to the minimum wage for for Venezuelan migrants between 2013 and 2019. The y-axis is limited to show values between 0 and 3. Each box plot represents the interquartile range (25th to 75th percentile), with the horizontal line indicating the median. The whiskers extend to 1.5 times the interquartile range, and dots represent outliers. The red dashed line marks the 0.75 minimum wage threshold, highlighting the proportion of individuals earning close to or below this level. *Source:* Author's own calculations from the GEIH.

Table 1 presents descriptive statistics comparing native and immigrant populations, highlighting key demographic and economic differences. The data includes mean values and standard deviations (SD) for variables such as age, gender (female), inactivity, monthly wage as a proportion of the minimum wage, formality (defined by pension affiliation), probability of being a low-wage worker, and years of schooling. Notable differences include a 3.71-year age gap, with immigrants averaging 32.63 years compared to 36.35 for natives, and a higher likelihood of immigrants being low-wage workers (0.36 vs. 0.33). All differences are statistically significant (p-value ¡0.001), underscoring distinct labor market and educational profiles between the two groups.

Table 1: Descriptive Statistics

	Natives		Immigrants			
Variable	Mean	SD	Mean	SD	Difference	p-value
Age	36.35	14.08	32.63	11.92	3.71	0
Female	0.54	0.50	0.51	0.50	0.03	0
Inactive Worker	0.28	0.45	0.20	0.40	0.08	0
Monthly wage/ Minimum wage	1.47	1.98	0.92	1.36	0.55	0
Formal Employee	0.48	0.59	0.12	0.33	0.35	0
Probability of being LW worker	0.33	0.26	0.36	0.25	-0.03	0
Years of Shool	10.47	4.19	9.97	3.63	0.50	0

Note: Table 1 depicts descriptive statistics comparing natives and Immigrants, showing means, standard deviations (SD), differences, and p-values for variables including age, gender (female), inactivity, monthly wage as a proportion of the Minimum wage, formality, as defined by pension affiliation, probability of being a low-wage worker and years of school.

4 Predicting Low-Wage Employees

This section first describes the processes of predicting low-wage individuals. First I describe the process of construction of the training and test samples. Then I briefly describe the Machine Learning models trained and presents the results. See section B for more details about the training process.

4.1 Test and Training Samples

To train the models I used the 2012 GEIH sample for all cities. I kept all workers between 15 and 65 years old and used the 2012 minimum wage to define the low-wage workers. I divided the data into two parts. The first part, the training sample, with 60% of all households in the sample[9] is used to tune the model parameters using 5-fold cross-validation. The second, the test sample, which comprises 40% of all households, is used to compare the out-of-sample model performance of the competing models. [10].

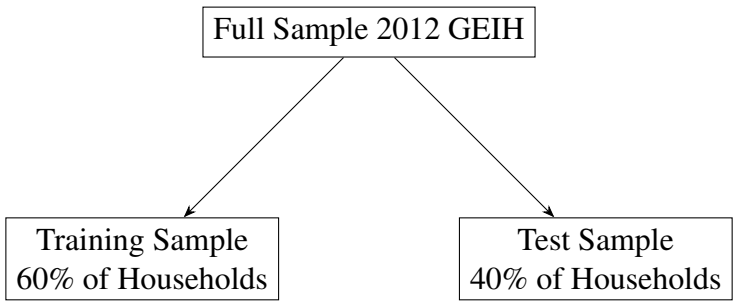
4.2 Prediction Algorithms

In this section I briefly describe the prediction algorithms used to train the model. As described in the previous section, I constructed a test sample in order to compare all these models in terms of their out-of-sample performance. This ensures that the trained model will be trained to predict unseen observations or, in other words, will not overfit.

[9]Here, a household refers to all the persons that live in a given house.

[10]To create all subsamples I used the city as strata. This ensures that all cities are proportionally represented in each subsample.

Figure 4: Process of Dividing the Sample for Model Training and Evaluation



Training Sample: Used to tune model parameters with 5-fold cross-validation.
Test Sample: Used to evaluate out-of-sample
model performance and calculate final thresholds.

Logit-Lasso Model: This algorithm extends the logistic regression estimation method by adding a penalty to the model to reduce complexity, effectively shrinking some of the coefficients to zero (Tibshirani, 1996). This results in a simpler model, often leading to less variance and to better generalization on unseen or test data.

Random Forest: This is an ensemble learning method that constructs multiple decision trees and merges their predictions to improve accuracy and control overfitting (Breiman, 2001). Each tree in the forest gives a classification, and the forest chooses the classification with the most votes (over all the trees in the forest). This approach is powerful for handling diverse feature sets and can automatically capture complex interactions between variables.

Extreme Gradient Boosting Machine (XGBoost): XGBoost is a highly efficient and scalable implementation of gradient boosting developed by Chen and Guestrin (2016). It works by building trees sequentially, where each new tree attempts to correct errors made by the previous ones. This method is known for its speed and performance, achieving high predictive accuracy by combining the outputs of multiple weak learners to form a strong predictor. However, the number of parameters to tune can be large, which makes them susceptible to overfitting and time-consuming with large datasets. In this paper, I focused on a small set of tuning parameters, and, nevertheless, this model supersedes the other models.

Neural Network: This algorithm is a one-layer Neural Network, also known as a single-layer perceptron, which consists of an input layer, a single hidden layer, and an output layer (James et al., 2021). It captures linear and some non-linear patterns in data using neurons or nodes to process inputs. The network's predictions are based on learned weights that connect these neurons. While simpler than deeper networks, a one-layer neural network can still learn the non-linearities and complexity of the DGP while maintaining a low complexity.

4.3 Results

The Table 2 presents the out-of-sample performance of five prediction models using two key evaluation metrics: the Area Under the Precision-Recall Curve (AUC-PR) and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). These metrics are particularly informative in settings with class imbalance, such as predicting rare events. Across the models, the Super Learner achieves the highest scores in both AUC-PR (0.657) and AUC-ROC (0.822), followed closely by XGBoost and the Neural Net. Notably, the Logit-Lasso model underperforms relative to the others, with the lowest AUC-PR (0.541), suggesting that its linear specification is less effective in capturing the complexities of the prediction task.

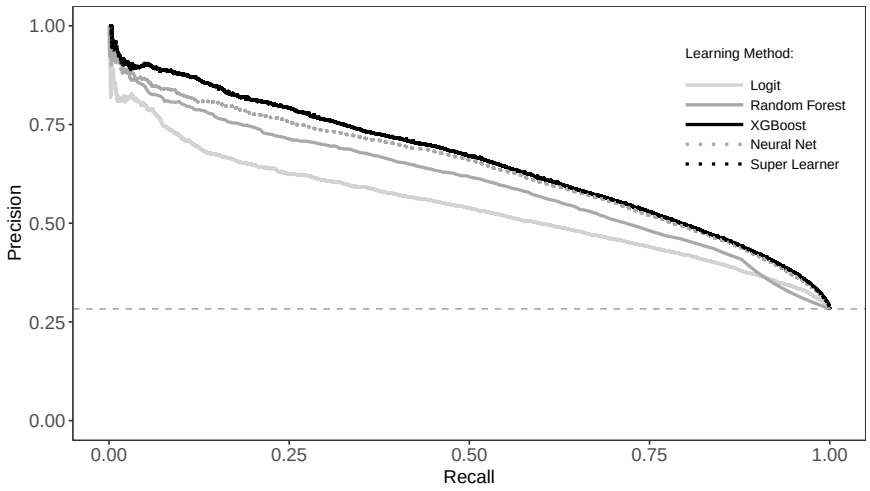
Table 2: Out-of-Sample Performance of Prediction Models,

Model	(1) AUC-PR	(2) AUC-ROC
Random Forest	0.602	0.782
Logit-Lasso	0.541	0.753
XGBoost	0.656	0.821
Neural Net	0.638	0.815
Super Learner	0.657	0.822

Note: Table 2 depicts the out-of-sample area under the Precision-Recall (AUC-PR) and Operating Characteristic Curve (AUC-ROC) for all the models trained. The estimation uses the data in the test sample.

Despite the marginally higher performance of the Super Learner, I choose XGBoost as the preferred model for subsequent analysis. This decision is motivated by two factors. First, the performance gain from using the Super Learner over XGBoost is minimal—just 0.001 in AUC-PR and AUC-ROC—implying limited value added. Second, an inspection of the ensemble's weights reveals that XGBoost contributes nearly all the predictive power to the Super Learner, indicating that the ensemble's strong performance is driven almost entirely by XGBoost. Therefore, for parsimony and interpretability, I rely on XGBoost as the main prediction model. A.5 shows how the chosen machine learning model compares with other demographic driven rules to distinguish low-wage workers. Researchers have looked at the labor market of teenagers, females and low educated individuals as potentially a low-wage labor markets since their lack of experience, discrimination and education, respectively (Cengiz et al., 2022). I compare the precision and recall scores of these simple classification rules and find that the XGBoost model outperforms this basic classifications. In particular, the model offers a Pareto improvement in all cases in the sense that it can achieve a higher precision (recall) given the same level of recall (precision) achieved by the classification rule.

Figure 5: Precision Recall Curves on the Test Set



Note: Figure 5 depicts the out-of-sample performance of each of the Machine learning methods applied to the low-wage workers classification task. The results are computed using the test sample subset. See main text and appendix B or details about each model and the training process.

4.4 Using Predicted Probability for Inference

With the predicted probability for each individual in our sample, we can define low-wage employees, allowing for different cut-offs in the estimated probability. Following Cengiz et al. (2022) and Card and Krueger (1995) I define the high-recall sample (HR) defined with a threshold (0.28) that allows a recall of 75%. This means that 75% of all minimum wage workers are included in this group. On the other hand, the precision in this case is 0.52. With this sample in hand, I define the low-wage workers as those with a predicted probability greater than 0.28. I show that the main results are robust to the choice of the threshold.

5 Methodology

This section describes the identification strategy used to estimate the effect of the migration shock on labor market outcomes, namely, a Triple Differences-in-Differences research design (DDD). I first present the methodological approach and the estimating equation, then define the causal parameter of interest and discuss the assumptions required for identification.

5.1 Set Up

The implementation of the DDD methodology leverages the fact that, based on the estimated probability, we can classify a group of low-wage employees within each city. In addition, we define high-migration cities and a post-treatment period to evaluate the effect of the border reopening in

July 2016. This approach compares the difference in outcomes between low-wage workers and other workers in highly migration-exposed cities to the same difference in less-exposed cities, before and after the policy shock. The key identifying assumption is that, in the absence of the migration shock, the evolution of this gap between low-wage and other workers would have been parallel across high- and low-migration cities.

Note that this approach allows for differential trends between workers within the same city. The DDD estimating equation is the following:

$$\begin{aligned} Y_{ict} =& \gamma_c + \delta_t + X'_{ict}\Omega + \Gamma_c\,(t \times \gamma_c) + \rho^{D}\text{LowWage}_{ict} \\ &+ \rho^{1,DD}(\text{Post}_t \times \text{HighMig}_c) + \rho^{2,DD}(\text{Post}_t \times \text{LowWage}_{ict}) + \rho^{3,DD}(\text{HighMig}_c \times \text{LowWage}_{ict}) \\ &+ \rho^{DDD}(\text{Post}_t \times \text{HighMig}_c \times \text{LowWage}_{ict}) + \varepsilon_{ict} \end{aligned} \tag{1}$$

where γ_c and δ_t denote city and month fixed Effects. LowWage_{ict} is an indicator of the worker being a low-wage worker; that is, their predicted probability is higher than 0.28. Post_t is an indicator variable of the re-opening of the border and HighMig_c is an indicator of the city having a Venezuelan migrant labor share in 2005 superior to the median. This last indicator is intended to allow us to distinguish between cities with high exposure to migration, since migrants tend to settle in cities with previous migration flows. Finally, X_{ict} and $(t \times \gamma_c)$ denote individual controls and city trends. The interest lies on estimating ρ^{DDD}. The set of demographic controls includes age and its square, a dummy variable for female individuals, interaction terms between age (and age squared) and this female dummy, literacy status, years of schooling, a variable that records the highest grade completed, marital status, and the individual's relationship to the household head.

This is a triple Differences-in-Differences model with non-staggered adoption since we define the treatment as the reopening of the border. Hence, the problems arising with "forbidden" comparisons between already treated groups will not be an issue here (Arkhangelsky and Imbens, 2024). We can see that this estimator uses two Differences-in-Differences (DiDs) comparisons to construct the triple difference estimator. First, it compares low-wage individuals and high-wage individuals within high-migration cities before and after the reopening of the frontier. This is the first DiDs. The second DiDs compares the mean outcome for low-wage individuals and high-wage individuals in low-migration cities before and after the event.

To explore the dynamics of the impact of migration, I also estimate a dynamic version of the equation 1 that replaces $(\text{Post}_t \times \text{HighMig}_c \times \text{LowWage}_{ict})$ with:

$$\sum_{s \in G} \rho_s^{DDD}(\mathbf{1}[t \in s] \times \text{HighMig}_c \times \text{LowWage}_{ict})$$

where G is a set of disjoint subsets of $[0, T]$. $s \in G$ is a subset of the time frame. Identification does not require parallel trends in each DiD comparison to hold. There can be differences in potential

outcomes between different groups of workers within a city, but these differences in counterfactual trends should be equal across different cities. I provide evidence that this is the case when estimating the dynamic version of equation 1.

5.2 Fourth Differences

With the purpose of studying the heterogeneous effects of the immigration shock, I extend equation 1 to include an interaction with a subpopulation indicator to explore potential heterogeneous effects. The Fourth difference in difference specification is:

$$\begin{aligned} Y_{ict} =& \gamma_c + \delta_t + X'_{ict}\Omega + \Gamma_c\,(t \times \gamma_c) + \sum_{m=1}^{4} \rho^{m,D}(\text{4 Linear Terms}) \\ &+ \sum_{m=1}^{6} \rho^{m,DD}(\text{6 Double Interactions}) + \sum_{m=1}^{4} \rho^{m,DDD}(\text{4 Triple Interactions}) \\ &+ \rho^{DDDD}(\text{Post}_t \times \text{HighMig}_c \times \text{LowWage}_{ict} \times \text{Indicator}_i) + \varepsilon_{ict} \end{aligned} \tag{2}$$

Where Indicator_i is a dummy variable for the subpopulation of interest and the rest of the terms have the same interpretation as these of equation 1. In this specification, we are interested in the forth interaction coefficient, ρ^{DDDD} that can be interpreted as the differential impact on the outcome on individuals with $\text{Indicator}_i = 1$, and in the third interaction coefficient, $\rho^{1,DDD}$, of $\text{Post}_t \times \text{HighMig}_c \times \text{LowWage}_{ict}$.

6 Results

6.1 Effect of Migration on Wages, Unemployment and Participation

Table 3 reports the results from the DDD specification described above, where we examine the effects of the reopening of the border in July 2016 on two main outcomes: log wages, the and indicator or unemployment, and for participation in the labor market (being employed or unemployed *and* looking or a job). The coefficient of interest, ρ^{DDD}, corresponds to the interaction term "Low Wage × High Mig. City × Post," which captures the differential change in outcomes for low-wage workers in high-migration cities after the policy shock, relative to other groups.

The analysis focuses on low-wage individuals, defined as those with an estimated probability of being greater than 0.28. As mentioned before, this threshold corresponds with recall of 75% of all low-wage employees in the test sample. The treatment group consists of low-wage workers in cities with high exposure to Venezuelan immigration—defined as cities where the share of Venezuelan migrants in the labor force exceeds the national median in 2005—after the border opening.

Table 3: Triple Difference In Differences Estimates of the Effect of Immigration on Wages, Employment and Participation

	(1)	(2)	(3)
Dependent Variable	Log Wage	Unemployment	Participation
Low Wage × High Mig. City × Post	-0.071	-0.004	0.018
	(0.019)***	(0.004)	(0.009)*
	[0.014] **	[0.317]	[0.184]
Low Wage × Post	0.098	0.007	-0.030
	(0.011)***	(0.002)***	(0.006)***
	[0.001] ***	[0.300]	[0.001]***
High Mig. City × Post	0.026	-0.002	-0.002
	(0.011)	(0.005)	(0.007)
	[0.294]	[0.771]	[0.845]
Low Wage × High Mig. City	0.069	0.003	-0.017
	(0.021)***	(0.010)	(0.012)
	[0.002] ***	[0.833]	[0.308]
Low Wage	-0.424	0.052	-0.069
	(0.017)***	(0.005)***	(0.012)***
	[0.000] ***	[0.000]***	[0.015]**
Observations	1,472,614	1,905,710	2,676,042
Adjusted R-squared	0.358	0.056	0.283
Dep. Mean	13.426	0.122	0.712
Dep. Sd	0.912	0.328	0.453
City FE	Yes	Yes	Yes
Time FE	Yes	Yes	Yes
Demographic Controls	Yes	Yes	Yes
City Trends	Yes	Yes	Yes
Prediction Model	XGB	XGB	XGB

Note: Table 3 shows the results of estimating 1 for Log Wages (colum 1), an indicator for unemployment (Column 2) and labor market participation (Column 3). Low-wage workers are those with an estimated probability greater than 0.28. High migration cities are cities with a share of Venezuelan immigrants in the Labor force superior to the median across cities in 2005. Standard errors robust to intra-city correlation in parentheses. P-values calculated using Wild Bootstrap are presented in brackets. Significance levels are: * p<0.01, ** p<0.05, * p<0.1

Standard errors (in parenthesis) are robust to intra-city correlation. This is, I allow for the residual to be correlated across individuals inside the same city. To avoid the problems in inference that may arise from a small number of clusters (Cameron et al., 2008) I also report the wild bootstrap P-values[11]. Since the p-values and confidence intervals associated constructed using the wild bootstrap procedure are, in general, more conservative, I will report only these p-values in the following analysis. Column (1) shows that the estimated effect on log wages of -0.0712 that is statistically significant (p-value < 0.05). This suggests that, following the reopening of the border,

[11] See Roodman et al. (2019) for a guide to implement this procedure.

low-wage workers in cities with high migration exposure experienced an additional 7.1 percent reduction in wages compared to similar workers in low-migration cities and relative to higher-wage workers. This result is consistent with downward pressure on wages due to increased labor supply from migration inflows. Prior to the shock, the average monthly salary of workers earning 75% or less of the minimum wage was approximately 82 USD. A 7% decline thus corresponds to a reduction of about 5.7 USD per month, or roughly 69 USD per year, representing a substantial loss for low-wage workers

In contrast, column (2) indicates no significant effect on the probability of unemployment. This implies that while low-wage workers faced wage adjustments, there was no corresponding increase in unemployment risk. Finally, column 3 reports the effects on participation. I found a positive and non-significant increase in the probability of participating in the labor market. To explore possible dynamic impacts, table A.1 divides the post-opening period into two periods before and after 2018. This table reports evidence about a stable negative effect over the course of the post period, but a short run impact on participation in the first part of the post period: right after the border opening, participation increased by a little less than 3%, which represents 7% of the standard deviation in participation across individuals.

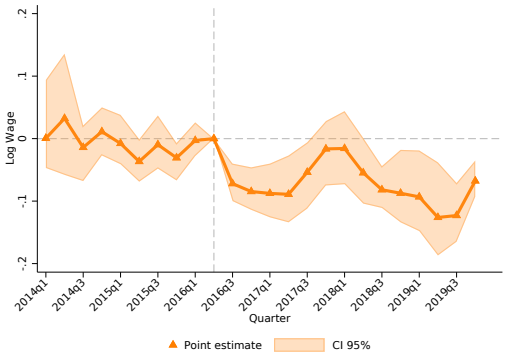
It is important to note that the validity of the DDD estimator relies on the parallel trends assumption: absent the migration shock, the gap in outcomes between low-wage and high-wage workers would have evolved similarly across high- and low-migration cities. While the DDD design partially relaxes the need for strict parallel trends between cities, it still requires that differences in potential trends between low- and high-wage workers are the same across cities. Therefore, testing for pre-trends using the dynamic specification is essential to assess the credibility of this identifying assumption. We can shed light on these dynamics and provide graphical evidence supporting the plausibility of parallel trends before the policy change by estimating the dynamic version of equation 1. Figure 6 shows the results of this exercise for (log) wages and unemployment.

Figure 6 presents the dynamic estimates of the effect of migration on low-wage natives, using the event-study specification described earlier. Panel A displays the evolution of log wages for low-wage workers in high-migration cities relative to other groups, while Panel B shows the corresponding estimates for the probability of unemployment, and panel C presents the dynamic effects over participation.

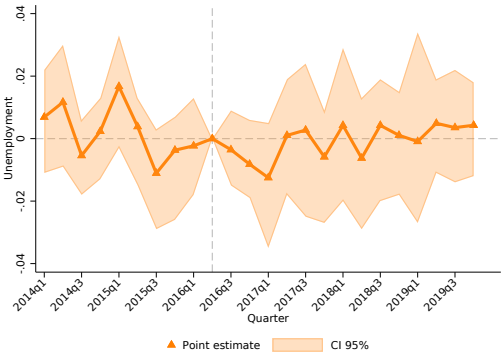
The estimates in Panel A suggest that before the reopening of the border in mid-2016 (marked by the vertical dashed line), there is no evidence of diverging trends between treatment and control groups. The coefficients fluctuate around zero, and their confidence intervals overlap with zero throughout the pre-treatment period, supporting the validity of the parallel trends assumption. After the policy shock, however, there is a noticeable and persistent decline in wages for low-wage workers in high-migration cities, consistent with the significant negative estimate found in Table 3.

In contrast, Panel B shows no clear evidence of an impact on unemployment rates for this group. Both before and after the policy change, the estimates hover around zero with wide confidence intervals, suggesting no significant displacement effect on employment. Together, these dynamics reinforce the interpretation that the migration shock primarily manifested through wage adjustments rather than through job losses among low-wage natives.Finally, and as discussed earlier, panel C reports an increase in participation in the short run.

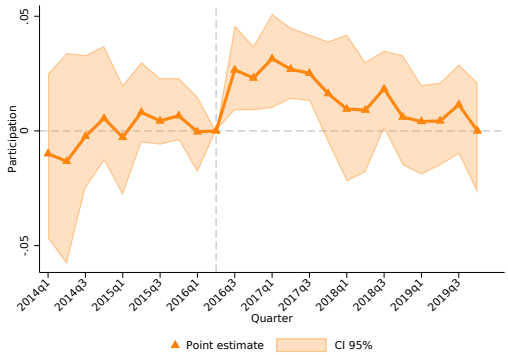
Figure 6: Dynamic Estimates of the Effect of Migration on Low-Wage Natives



A) Log Wages



B) Unemployment



C) Participation

Note: Figure 6 shows the results from estimating the dynamic version of equation 1 with s being groups of months that belong to the same quarter. Treatment is defined as the quarter of the reopening of the border. Panel A, depicts the results using Log Wages as the LHS variable. In panel B, the LHS variable is the probability of being unemployed. Finally, in Panel C, the LHS variable serves as an indicator of labor market participation. All specifications include demographic controls, city, and time (monthly) fixed effects and city trends. Low-wage workers are those with an estimated probability greater than 0.28. High migration cities are cities with a share of Venezuelan immigrants in the Labor force superior to the median across cities in 2005.Confidence Intervals at the 95% are calculated using Wild Bootstrap with 999 repetitions.

6.2 Robustness

The definition of the treated group based on the probability threshold that achieves a high recall of 75% ensures that this group captures roughly 75% of all low-wage workers and will be correctly classified. Nevertheless, as shown in Figures A.1,A.2, and, A.3, which plots the point estimates and 95% confidence intervals from estimating equation 1 under alternative definitions of both treatment and control groups, the results are robust to several definitions of the probability threshold (0.48, 0.38, 0.28, or 0.18) and also to defining the control group (non-low-wage workers) as those that have a low probability (less than 0.08 or 0.18) of being a low-wage worker.

Each dot in the figure corresponds to a different specification. The red dot represents the baseline

specification, and Blue dots denote alternative specifications. Across all variations, the estimated coefficients remain negative and statistically similar in magnitude, ranging from approximately -0.05 to -0.10, with overlapping confidence intervals.

Another potential concern is that the results are driven by any single city in the sample. Since certain cities may have experienced particularly large inflows of migrants or idiosyncratic labor market shocks, their inclusion could disproportionately influence the overall results.To address this concern I sequentially exclude one city at a time and re-estimate the dynamic specification in Figure A.4. Consistent estimates across these iterations indicate that the results are not dominated by outliers or city-specific factors, strengthening confidence in the validity and generalizability of the findings.

6.3 Other Outcomes

Table 4 presents the results of estimating DDD model to assess the effect of immigration on various labor market outcomes: the probability of working in a small firm (Column 1), the probability of having an open-ended contract (Column 2), a formality index (Column 3), and the probability of being self-employed (Column 4). All specifications include city and time fixed effects, a rich set of demographic controls, and use predictions from an XGBoost model to classify individuals by wage level. Standard errors are clustered at the city level.

Across all columns, the estimates of this interaction term are small and statistically insignificant, suggesting limited average effects of Venezuelan immigration on the outcomes studied. In contrast, the interaction term Low Wage × High Mig. City is statistically significant in Columns 2 and 4, indicating that even before the treatment period, low-wage individuals in high-migration cities had lower probabilities of holding open-ended contracts and higher probabilities of being self-employed. Additionally, consistent with prior expectations, the coefficient on Low Wage is significant across all specifications and indicates that low-wage individuals are more likely to work in small firms, less likely to hold formal jobs or open-ended contracts, and more likely to be self-employed.

Overall, the results show no strong evidence that immigration negatively affected other worker conditions of the native employees, namely firms size, formality and the probability of being self-employed.

6.4 Discussion

The results presented so far indicate that the Colombian labor market absorbed the Venezuelan inflow mainly through wage compression at the lower end of the distribution, with limited effects on employment, formality, or self-employment. This suggests that in highly flexible and segmented labor markets, migration shocks may translate more into wage adjustments than into job losses, as firms and workers adapt through informal mechanisms rather than formal restructuring. The wage effects concentrated among low-wage natives are consistent with increased competition in the informal or semi-formal sectors, where barriers to entry are low and minimum wage regulations are often non-binding.

The evidence also has clear policy implications. While the Colombian labor market demonstrated resilience in absorbing a large number of Venezuelan migrants, the burden of adjustment fell

disproportionately on vulnerable native workers. Targeted interventions aimed at protecting the earnings of low-wage workers—such as wage subsidies, training programs, and incentives for formal hiring—could help mitigate these pressures. In parallel, policies that facilitate the labor market integration of migrants, such as recognition of qualifications and simplified formalization procedures, can reduce informality and promote productivity gains. Ultimately, the challenge for policymakers is to balance protection and inclusion: designing frameworks that safeguard vulnerable natives without restricting migrants' access to decent work, thereby ensuring that migration contributes to shared economic gains.

Table 4: Triple Difference In Differences Estimates of the Effect of Immigration on Native Labor Market Conditions

	(1)	(2)	(3)
Dependent Variable	Self Employee	Formality	Works Small Firm
Low Wage × High Mig. City × Post	0.001	-0.013	0.001
	[0.947]	[0.165]	[0.855]
Low Wage × Post	-0.007	0.009	-0.009
	[0.457]	[0.209]	[0.021]**
High Mig. City × Post	0.003	0.007	-0.001
	[0.765]	[0.126]	[0.825]
Low Wage × High Mig. City	0.030	0.019	0.003
	[0.023] **	[0.298]	[0.698]
Low Wage	0.046	-0.144	0.097
	[0.003] ***	[0.000]***	[0.000]***
Observations	1,472,614	1,472,614	1,472,614
Adjusted R-squared	0.113	0.220	0.188
Dep. Mean	0.434	0.438	0.557
Dep. Sd	0.496	0.496	0.497
City FE	Yes	Yes	Yes
Time FE	Yes	Yes	Yes
Demographic Controls	Yes	Yes	Yes
City Trends	Yes	Yes	Yes
Prediction Model	XGB	XGB	XGB

Note: Table 4 shows the results of estimating 1. Low-wage workers are those with an estimated probability greater than 0.28. High migration cities are cities with a share of Venezuelan immigrants in the Labor force superior to the median across cities in 2005. Wild Bootstrap P-Values on Brackets. * p<0.01, ** p<0.05, * p<0.1

7 Heterogeneous Effects

Table 5 presents the results from the Fourth Difference-in-Differences specification in equation (2). First, to investigate the heterogeneous effects of migration on log wages across different industries,

Table 5: Fourth Difference In Differences Estimates of the Heterogeneous Effect of Migration Shock on Wages

Industry with:	(1) High Kaitz Index	(2) High share of Pension Affiliates	(3) High Share of Small Firms	(4) High Share of Self Emplyment
Low Wage × High Mig. City × Post × Indicator	0.008	-0.003	0.060	0.040
	[0.527]	[0.739]	[0.042]**	[0.123]
Low Wage × High Mig. City × Post	-0.081	-0.071	-0.123	-0.103
	[0.009] ***	[0.011]**	[0.007]***	[0.023]**
Linear Combination	-0.073	-0.074	-0.063	-0.063
	[0.009] ***	[0.004]***	[0.018]**	[0.004]***
Observations	1,472,614	1,467,904	1,467,904	1,467,904
Adjusted R-squared	0.368	0.368	0.367	0.367
Dep. Mean	13.426	13.426	13.426	13.426
Dep. Sd	0.912	0.911	0.911	0.911
City FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes
Demographic Controls	Yes	Yes	Yes	Yes
City Trends	Yes	Yes	Yes	Yes
Prediction Model	XGB	XGB	XGB	XGB

Note: Table 5 shows the results of estimating a 4 interaction model for Log Wage. Low-wage workers are those with an estimated probability greater than 0.28. High migration cities are cities with a share of Venezuelan immigrants in the Labor force superior to the median across cities in 2005. Wild Bootstraped P-values robust to intra-city correlation in brackets. * p<0.01, ** p<0.05, * p<0.1

I interact the treatment variable with industry-level indicators, defined based on pre-border-opening conditions. Specifically, column (1) classifies industries according to their Kaitz index at the 2-digit level, designating those above the median as "High Kaitz" industries. Columns (2), (3) and (4) follow a similar approach, using measures of informality: the proportion of workers affiliated with the pension system, the proportion employed in small firms (fewer than five employees), and the proportion of self-employed workers, respectively.

The coefficient on the triple interaction term—Low Wage × High Migration City × Post—is negative and statistically significant across all four columns, indicating that low-wage workers in high-migration cities experienced a consistent wage penalty following the migration shock. In contrast, the fourth interaction term, which interacts the treatment with the industry-level indicator, is small in magnitude and statistically insignificant in all but column 4, where the triple difference coefficient shows that the reduction in wages for low-wage workers in sectors with a low share of small firms is twice as large as the impact for workers in sectors with a high share of small firms. This suggests that the wage effect is relatively homogeneous across industries, regardless of differences in informality levels or the bindingness of the minimum wage. Therefore, the main conclusion is that the migration shock disproportionately affected low-wage workers across the board, irrespective of industry characteristics such as informality.

Table 6, on the other hand, reports the estimates from the Fourth Differences-in-Differences specification, which further interacts the main triple difference term with indicators for different subpopulations. This allows us to explore heterogeneity in the effects of migration on low-wage

Table 6: Fourth Difference In Differences Estimates of the Heterogeneous Effect of Migration Shock

Sub Population	(1) Age from 15 to 28	(2) Age from 29 to 40	(3) Age 41+	(4) Females	(5) Low Education
Panel A. Log Wage					
Low Wage × High Mig. City × Post × Sub Population	-0.042	0.005	0.034	-0.001	-0.003
	[0.072] *	[0.266]	[0.101]	[0.971]	[0.838]
Low Wage × High Mig. City × Post	-0.059	-0.072	-0.086	-0.075	-0.064
	[0.014] **	[0.003]***	[0.008]***	[0.053]*	[0.066]*
Linear Combination	-0.101	-0.067	-0.052	-0.076	-0.067
	[0.011] **	[0.011]**	[0.015]**	[0.027]**	[0.018]**
Observations	1,472,614	1,472,614	1,472,614	1,472,614	1,472,614
Adjusted R-squared	0.358	0.358	0.358	0.359	0.358
Dep. Mean for Sub Pop	13.295	13.571	13.401	13.228	13.454
Dep. Sd for Sub Pop	0.844	0.854	0.981	1.015	0.961
Panel B. Unemployment					
Low Wage × High Mig. City × Post × Sub Population	-0.006	-0.008	0.013	-0.020	0.017
	[0.754]	[0.147]	[0.301]	[0.007]***	[0.165]
Low Wage × High Mig. City × Post	-0.001	-0.001	-0.012	0.006	-0.016
	[0.781]	[0.719]	[0.250]	[0.266]	[0.064]*
Linear Combination	-0.007	-0.009	0.001	-0.014	0.001
	[0.616]	[0.133]	[0.845]	[0.052]*	[0.863]
Observations	1,905,710	1,905,710	1,905,710	1,905,710	1,905,710
Adjusted R-squared	0.057	0.056	0.058	0.056	0.056
Dep. Mean for Sub Pop	0.204	0.104	0.076	0.148	0.123
Dep. Sd for Sub Pop	0.403	0.306	0.265	0.355	0.328
Panel C. Participation					
Low Wage × High Mig. City × Post × Sub Population	-0.003	0.004	-0.001	0.001	0.002
	[0.604]	[0.734]	[0.899]	[0.960]	[0.703]
Low Wage × High Mig. City × Post	0.019	0.016	0.020	0.014	0.017
	[0.136]	[0.200]	[0.297]	[0.332]	[0.329]
Linear Combination	0.016	0.020	0.019	0.015	0.019
	[0.274]	[0.255]	[0.229]	[0.271]	[0.161]
Observations	2,676,042	2,676,042	2,676,042	2,676,042	2,676,042
Adjusted R-squared	0.294	0.284	0.289	0.284	0.284
Dep. Mean for Sub Pop	0.567	0.873	0.747	0.632	0.683
Dep. Sd for Sub Pop	0.496	0.332	0.435	0.482	0.465
City FE	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes
Demographic Controls	Yes	Yes	Yes	Yes	Yes
City Trends	Yes	Yes	Yes	Yes	Yes
Prediction Model	XGB	XGB	XGB	XGB	XGB

Note: Table 6 shows the results of estimating a 4 interaction model for Log Wage. Low-wage workers are those with an estimated probability greater than 0.28. High migration cities are cities with a share of Venezuelan immigrants in the Labor force superior to the median across cities in 2005. Wild Bootstraped P-values robust to intra-city correlation in brackets. * p<0.01, ** p<0.05, * p<0.1

workers across age groups, gender, and educational attainment for all the main variables, log wages, employment and participation. The coefficient of interest in each column, "Low Wage × High Mig. City × Post × Sub Population," measures how the impact on low-wage workers in high-migration cities after the border reopening varies across these subgroups.

Columns (1)-(3) show results by age group. Young workers aged 15-28 (column 1) experience a large and statistically significant decline in log wages (-0.101) which is twice as large as the one for the rest of workers (-0.059), suggesting they are particularly vulnerable to wage competition from migrant inflows. In contrast, workers aged 29-40 (column 2) and older workers aged 41 and above (column 3) experience an effect similar to the average effect.

Columns (4) and (5) present results by gender and education. There are no statistically significant differential effects for males or females, or for low educated workers (secondary incomplete or less), indicating that the wage impact of migration on low-wage workers does not differ systematically by gender or education in this context.

In panels B and C I explore differential effects on unemployment and participation, respectively, and find evidence of homogeneous effects across the different sub-populations. Except for the unemployment rate of female individuals, which decreases after the migration shock by 1.4 percent, a 10% decrease compared with the female unemployment rate of 14%. Together, these results highlight important heterogeneity in the wage response to migration shocks, especially when related to age profiles (for wages) and gender (for unemployment). Finally, in Table A.3, I present the results from estimating equation 2 for only the first year after the border re-opening. This is because, for participation, the main impact occurs in the first year of the post-period. I find that the effect on participation is homogeneous across different sub-populations

8 Conclusions

This paper investigates the short-run labor market effects of the Venezuelan migration crisis on Colombian workers, focusing on those at the lower end of the wage distribution. Exploiting the 2016 border reopening as a natural experiment, I combine a machine learning–based classification of low-wage individuals with a Triple Differences-in-Differences identification strategy. This approach allows for a flexible, data-driven definition of exposure to migration and enables causal inference on how migration inflows affect vulnerable native workers.

The results reveal that Venezuelan migration led to a significant decline in wages among low-wage natives in high-migration cities, while employment formality, self-employment, and unemployment rates remained largely unaffected. These findings suggest that labor market adjustments to migration occurred primarily through wage compression rather than job displacement or changes in employment type. Heterogeneity analyses indicate that the wage impact was stronger among younger workers.

Methodologically, this paper contributes to the growing intersection of machine learning and causal inference in labor economics by demonstrating how predictive models can enhance the identification of exposed populations in quasi-experimental settings. Substantively, it extends the evidence on migration shocks beyond high-income countries, showing that even in highly informal labor markets

such as Colombia's, large inflows of migrants can generate measurable but contained wage effects concentrated at the bottom of the income distribution. Overall, the evidence points to the resilience of urban labor markets in absorbing large migrant populations, albeit with distributional consequences for the most vulnerable workers. Future research should examine longer-term dynamics, including potential effects on occupational mobility, firm productivity, and local public services, to better understand how migration reshapes economic opportunity in middle-income host countries.

Data Availability Statement: The main data used in this paper can be download from DANE data portal: https://microdatos.dane.gov.co/index.php/catalog/MERCLAB-Microdatos. A replication package for this article is available at https://zenodo.org/records/17444504.

References

Altonji, J. G. and Card, D. (1991). The effects of immigration on the labor market outcomes of less-skilled natives.

Angrist, J. D. and Frandsen, B. (2022). Machine labor. *Journal of Labor Economics*, 40(S1):97–140.

Arkhangelsky, D. and Imbens, G. (2024). Causal models for longitudinal and panel data: a survey. *The Econometrics Journal*, 27(3):C1–C61.

Bahar, D., Dooley, M., and Huang, C. (2022). Integración de los venezolanos en el mercado laboral colombiano. *Brookings Institution*. Avaliable at: https://www.brookings.edu/es/articles/integracion-de-los-venezolanos-en-el-mercado-laboral-colombiano/.

Bazzi, S., Blair, R. A., Blattman, C., Dube, O., Gudgeon, M., and Peck, R. (2022). The promise and pitfalls of conflict prediction: Evidence from colombia and indonesia. *The Review of Economics and Statistics*, 104(4):764–779.

Bonilla-Mejía, L., Morales, L. F., Hermida, D., and Flórez, L. A. (2024). The labor market effect of south-to-south migration: Evidence from the venezuelan crisis. *International Migration Review*, 58(2):764–799.

Borjas, G. J. (2017). The wage impact of the marielitos: A reappraisal. Technical Report 5.

Borusyak, K., Hull, P., and Jaravel, X. (2021). Quasi-experimental shift-share research designs. *The Review of Economic Studies*, 89(1):181–213.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breza, E. and Kaur, S. (2025). Labor markets in developing countries. *Annual Review of Economics*, 17(Volume 17, 2025):747–776.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427.

Cancilleria de Colombia (2019). Resolución 0872 de 2019. Accessed: 2025-09-22.

Card, D. (1990). The impact of the mariel boatlift on the miami labor market. *Industrial and Labor Relations Review*, 43(2):245–257.

Card, D., Colella, F., and Lalive, R. (2024). Gender preferences in job vacancies and workplace gender diversity. *The Review of Economic Studies*, 92(4):2437–2471.

Card, D. and Krueger, A. B. (1995). *Myth and measurement: The new economics of the minimum wage*. Princeton University Press, Princeton, NJ.

Caruso, G., Canon, C. G., and Mueller, V. (2019). Spillover effects of the venezuelan crisis: migration impacts in colombia. *Oxford Economic Papers*, 73(2):771–795.

Cengiz, D., Dube, A., Lindner, A., and Zentler-Munro, D. (2022). Seeing beyond the trees: Using machine learning to estimate the impact of minimum wages on labor market outcomes. *Journal of Labor Economics*, 40(S1):S203–S247.

Chaves-González, D. and Echeverría-Estrada, C. (2020). Venezuelan migrants and refugees in latin america and the caribbean: A regional profile. Accessed: 2025-09-22.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Delgado-Prieto, L. (2024). Immigration, wages, and employment under informal labor markets. *Journal of Population Economics*, 37(2):55.

Dustmann, C., Frattini, T., and Preston, I. P. (2013). The effect of immigration along the distribution of wages. *The Review of Economic Studies*, 80(1):145–173.

Dustmann, C., Schönberg, U., and Stuhler, J. (2016). The impact of immigration: Why do studies reach such different results? *The Journal of Economic Perspectives*, 30(4):31–56.

European Parliament Research Service (2018). Venezuela: The political crisis and its impact on migration. Technical report, European Parliament. Accessed: 2025-09-23.

Goldsmith-Pinkham, P., Sorkin, I., and Swift, H. (2020). Bartik instruments: What, when, why, and how. *American Economic Review*, 110(8):2586–2624.

IMF (2025). World Economic Outlook: Venezuela. https://www.imf.org/external/datamapper/profile/VEN/WEO. Accessed: 2025-09-22.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R*. Springer, New York, NY, USA, 2 edition.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293.

Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3):337–346.

Macias, A. (2015). Venezuela closed 2 of its borders with colombia after a violent shoot-out. Accessed: 2025-09-22.

Monras, J. (2021). Local adjustment to immigrant-driven labor supply shocks. *Journal of Human Capital*, 15(1):204–245.

Mullainathan, S. and Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Neuman, W. (2015). Strict rationing in venezuela as plunging oil prices hurt economy. *The New York Times*. Accessed: 2025-09-22.

Otero-Cortés, A., Tribín-Uribe, A. M., and Mojica-Urueña, T. (2022). Heterogeneous labor market effects of the venezuelan exodus on female workers: Evidence from colombia. Technical Report Borradores de Economía No. 1172, Banco de la República de Colombia.

Pedrazzi, J. and Peñaloza-Pacheco, L. (2023). Heterogeneous effects of forced migration on the female labor market: The venezuelan exodus in colombia. *The Journal of Development Studies*, 59(3):324–341.

Polley, E. C. and van der Laan, M. J. (2010). Super learner in prediction. Working Paper 266, U.C. Berkeley Division of Biostatistics.

R4V (2025). Refugees and migrants from venezuela: Population update - june 2025. https://www.r4v.info/en/population-update-june2025. Accessed: 27 July 2025.

Roodman, D., Ørregaard Nielsen, M., MacKinnon, J. G., and Webb, M. D. (2019). Fast and wild: Bootstrap inference in stata using boottest. *The Stata Journal*, 19(1):4–60.

Santamaria, J. (2022). When a stranger shall sojourn with thee': The impact of the venezuelan exodus on colombian labor markets. Documentos de trabajo - Alianza EFI 20046, Alianza EFI.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Vera, L. (2015). Venezuela 1999-2014: Macro-policy, oil governance and economic performance. *Comparative Economic Studies*, 57(3):539–568.

Online Appendix to "Boosted Migration: The Effect of Migration on Low-Wage Workers"

Andres Rengifo-Jaramillo

October 25, 2025

A Additional Tables and Figures

Table A.1: Triple Difference In Differences Estimates For Main LM outcomes

Dependent Variable	(1) Log Wage	(2) Unemployment	(3) Participation
Low Wage × High Mig. City × Post[t <2018]	-0.064	-0.008	0.028
	(0.020)***	(0.004)**	(0.009)***
	[0.005] ***	[0.097]*	[0.024]**
Low Wage × High Mig. City × Post[t ≥ 2018]	-0.077	-0.001	0.010
	(0.020)***	(0.004)	(0.010)
	[0.010] ***	[0.858]	[0.429]
Low Wage × Post	0.098	0.007	-0.030
	(0.011)***	(0.002)***	(0.006)***
	[0.002] ***	[0.267]	[0.000]***
High Mig. City × Post	0.023	0.000	-0.008
	(0.020)	(0.005)	(0.007)
	[0.320]	[0.989]	[0.327]
Low Wage × High Mig. City	0.069	0.003	-0.017
	(0.021)***	(0.010)	(0.012)
	[0.002] ***	[0.839]	[0.294]
Low Wage	-0.424	0.052	-0.070
	(0.017)***	(0.005)***	(0.012)***
	[0.000] ***	[0.000]***	[0.014]**
Observations	1,472,614	1,905,710	2,676,042
Adjusted R-squared	0.358	0.056	0.283
Dep. Mean	13.426	0.122	0.712
Dep. Sd	0.912	0.328	0.453
City FE	Yes	Yes	Yes
Time FE	Yes	Yes	Yes
Demographic Controls	Yes	Yes	Yes
City Trends	Yes	Yes	Yes
Prediction Model	XGB	XGB	XGB

Note: Table A.1 shows the results of estimating 1 for Log Wages (colum 1), an indicator for unemployment (Column 2) and labor market participation (Column 3). Low-wage workers are those with an estimated probability greater than 0.28. High migration cities are cities with a share of Venezuelan immigrants in the Labor force superior to the median across cities in 2005. Standard errors robust to intra-city correlation in parentheses, Wild Bootstrap P-values in brackets. * p<0.01, ** p<0.05, * p<0.1

Table A.2: Triple Difference In Differences Estimates

Dependent Variable	(1) Self Employee	(2) Formality	(3) Works Small Firm
Low Wage × High Mig. City × Post[t <2018]	0.003	-0.012	0.001
	(0.012)	(0.008)	(0.004)
	[0.810]	[0.174]	[0.884]
Low Wage × High Mig. City × Post[t ≥ 2018]	-0.001	-0.014	0.000
	(0.011)	(0.008)	(0.004)
	[0.971]	[0.176]	[0.928]
Low Wage × Post	-0.007	0.009	-0.009
	(0.007)	(0.006)	(0.001)***
	[0.421]	[0.240]	[0.023]**
High Mig. City × Post	0.002	0.007	-0.001
	(0.011)	(0.004)	(0.004)
	[0.834]	[0.152]	[0.822]
Low Wage × High Mig. City	0.030	0.019	0.003
	(0.010)***	(0.014)	(0.006)
	[0.025] **	[0.297]	[0.651]
Low Wage	0.046	-0.144	0.097
	(0.005)***	(0.009)***	(0.002)***
	[0.003] ***	[0.000]***	[0.000]***
Observations	1,472,614	1,472,614	1,472,614
Adjusted R-squared	0.113	0.220	0.188
Dep. Mean	0.434	0.438	0.557
Dep. Sd	0.496	0.496	0.497
City FE	Yes	Yes	Yes
Time FE	Yes	Yes	Yes
Demographic Controls	Yes	Yes	Yes
City Trends	Yes	Yes	Yes
Prediction Model	XGB	XGB	XGB

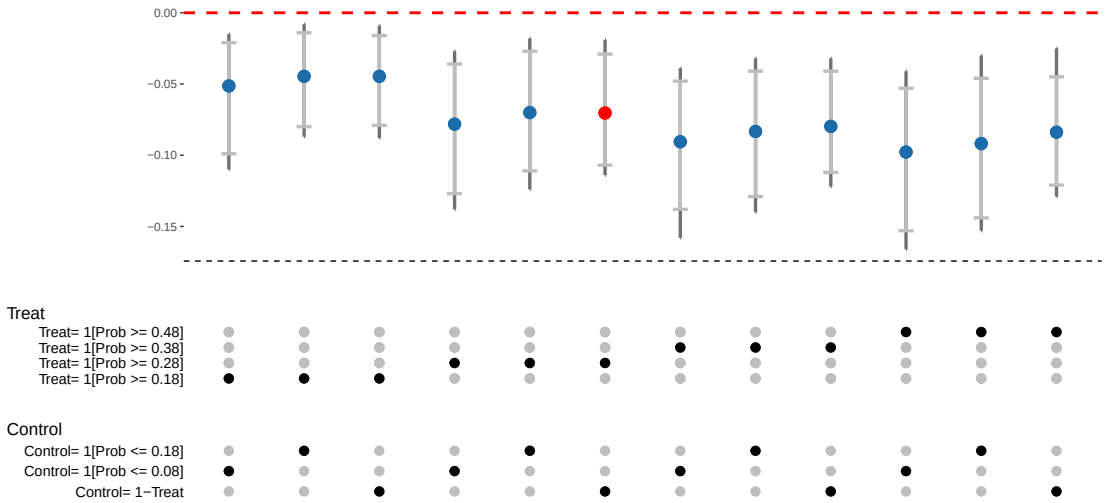
Note: Table A.2 shows the results of estimating 1 for Log Wages (colum 1), an indicator for unemployment (Column 2) and labor market participation (Column 3). Low-wage workers are those with an estimated probability greater than 0.28. High migration cities are cities with a share of Venezuelan immigrants in the Labor force superior to the median across cities in 2005. Standard errors robust to intra-city correlation in parentheses, Wild Bootstrap P-values in brackets. * p<0.01, ** p<0.05, * p<0.1

Table A.3: Fourth Differences-in-Differences Estimates of the Heterogeneous Effect of Migration Shock

Sub Population	(1) Age from 15 to 28	(2) Age from 29 to 40	(3) Age 41+	(4) Females	(5) Low Education
Panel A. Log Wage					
Low Wage × High Mig. City × Post[t <2018] × Sub Population	-0.028	-0.007	0.030	-0.020	-0.010
	[0.212]	[0.668]	[0.154]	[0.558]	[0.534]
Low Wage × High Mig. City × Post[t <2018]	-0.057	-0.063	-0.077	-0.053	-0.052
	[0.013] **	[0.011]**	[0.011]**	[0.135]	[0.067]*
Linear Combination	-0.085	-0.070	-0.047	-0.073	-0.062
	[0.016] **	[0.014]**	[0.018]**	[0.028]**	[0.009]***
Observations	1,472,614	1,472,614	1,472,614	1,472,614	1,472,614
Adjusted R-squared	0.358	0.358	0.358	0.359	0.358
Dep. Mean for Sub Pop	13.295	13.571	13.401	13.228	13.454
Dep. Sd for Sub Pop	0.844	0.854	0.981	1.015	0.961
Panel B. Unemployment					
Low Wage × High Mig. City × Post[t <2018] × Sub Population	-0.009	-0.012	0.018	-0.014	0.018
	[0.550]	[0.083]*	[0.121]	[0.021]**	[0.122]
Low Wage × High Mig. City × Post[t <2018]	-0.004	-0.004	-0.019	-0.003	-0.021
	[0.148]	[0.276]	[0.091]*	[0.538]	[0.037]**
Linear Combination	-0.013	-0.016	-0.001	-0.017	-0.003
	[0.334]	[0.039]**	[0.708]	[0.018]**	[0.628]
Observations	1,905,710	1,905,710	1,905,710	1,905,710	1,905,710
Adjusted R-squared	0.057	0.056	0.058	0.056	0.056
Dep. Mean for Sub Pop	0.204	0.104	0.076	0.148	0.123
Dep. Sd for Sub Pop	0.403	0.306	0.265	0.355	0.328
Panel C. Participation					
Low Wage × High Mig. City × Post[t <2018] × Sub Population	0.007	0.003	-0.010	-0.002	-0.002
	[0.304]	[0.717]	[0.444]	[0.857]	[0.830]
Low Wage × High Mig. City × Post[t <2018]	0.025	0.026	0.034	0.027	0.030
	[0.013] **	[0.018]**	[0.045]**	[0.075]*	[0.151]
Linear Combination	0.032	0.029	0.024	0.025	0.028
	[0.014] **	[0.016]**	[0.035]**	[0.039]**	[0.002]***
Observations	2,676,042	2,676,042	2,676,042	2,676,042	2,676,042
Adjusted R-squared	0.294	0.284	0.289	0.284	0.284
Dep. Mean for Sub Pop	0.567	0.873	0.747	0.632	0.683
Dep. Sd for Sub Pop	0.496	0.332	0.435	0.482	0.465
City FE	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes
Demographic Controls	Yes	Yes	Yes	Yes	Yes
City Trends	Yes	Yes	Yes	Yes	Yes
Prediction Model	XGB	XGB	XGB	XGB	XGB

Note: Table A.3 shows the results of estimating equation 2 for Log Wages (Panel A), Unemployment indicator (Panel B) and Participation (Panel C). Low-wage workers are those with an estimated probability greater than 0.28. High migration cities are cities with a share of Venezuelan immigrants in the Labor force superior to the median across cities in 2005. Wild Bootstraped P-values robust to intra-city correlation in brackets. * p<0.01, ** p<0.05, * p<0.1

Figure A.1: Robustness of the Effect of Migration on Wages: Different Thresholds



Note: Figure A.1 shows the estimated coefficient ρ^{DDD} and the associated confidence interval from estimating equation 1 for Log Wages using different definitions of low-wage individuals and control individuals. Red denotes the baseline specification, and blue denotes an alternative specification. Confidence Intervals are calculated using the wild Bootstrap procedure.

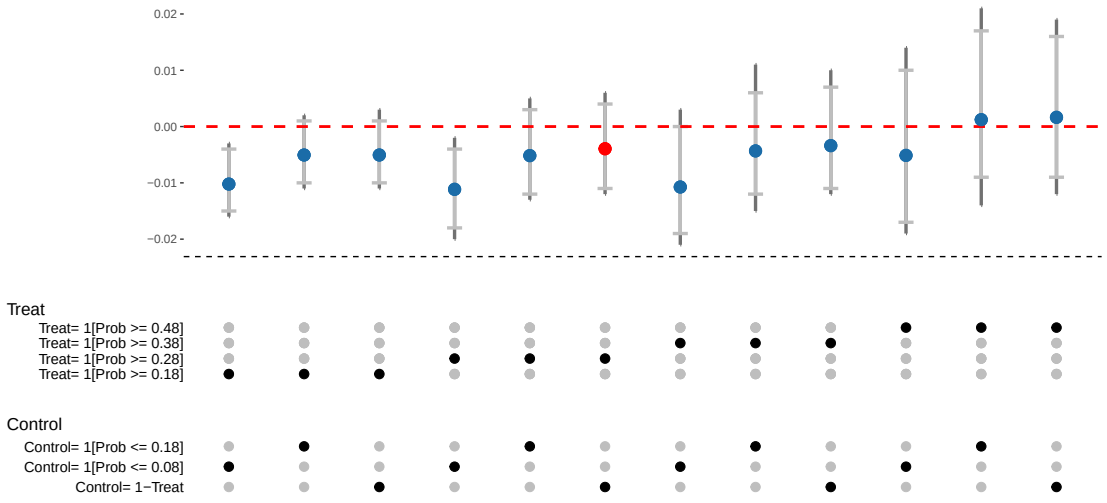


Figure A.2: Robustness of the Effect of Migration on Unemployment: Different Thresholds

Note: Figure A.2 shows the estimated coefficient ρ^{DDD} and the associated confidence interval from estimating equation 1 for Unemployment using different definitions of low-wage individuals and control individuals. Red denotes the baseline specification, and blue denotes an alternative specification. Confidence Intervals are calculated using the wild Bootstrap procedure.

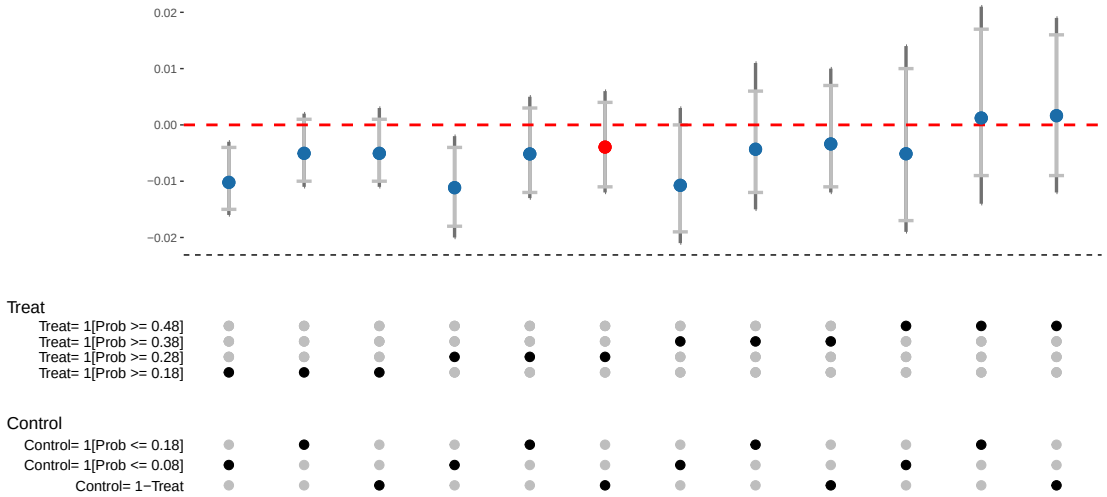
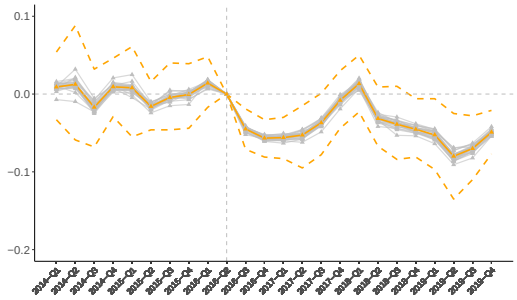


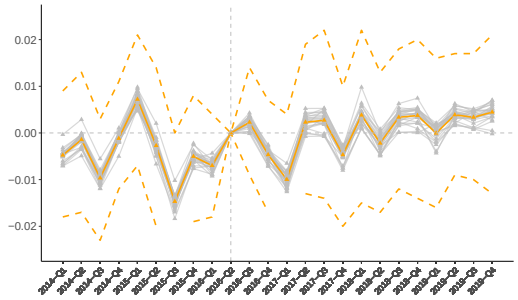
Figure A.3: Robustness of the Effect of Migration on Participation: Different Thresholds

Note: Figure A.3 shows the estimated coefficient ρ^{DDD} and the associated confidence interval from estimating equation 1 for Participation using different definitions of low-wage individuals and control individuals. Red denotes the baseline specification, and blue denotes an alternative specification. Confidence Intervals are calculated using the wild Bootstrap procedure.

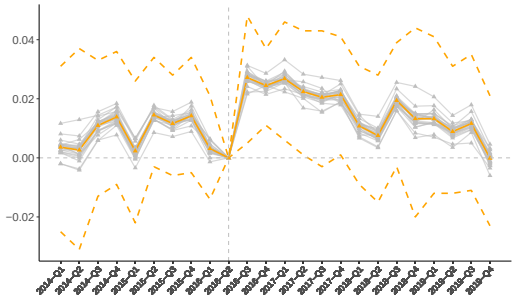
Figure A.4: Dynamic Estimates of the Effect of Migration on Low-Wage Natives



A) Log Wages



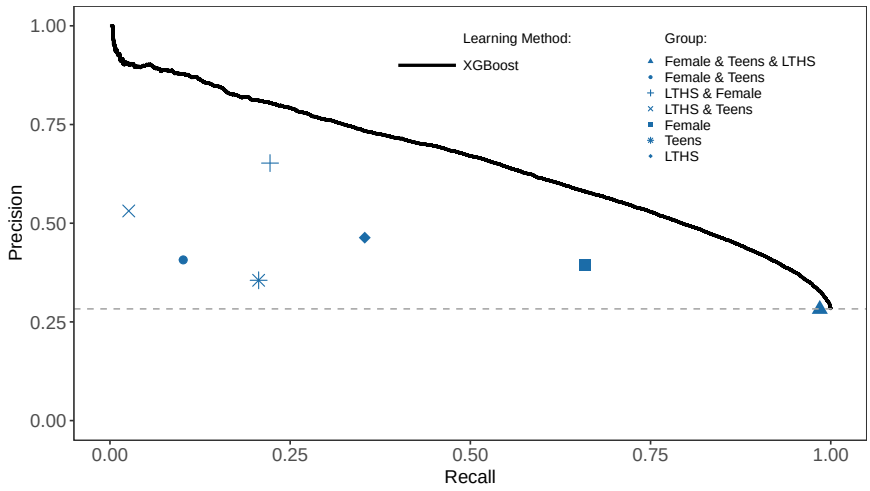
B) Unemployment



C) Participation

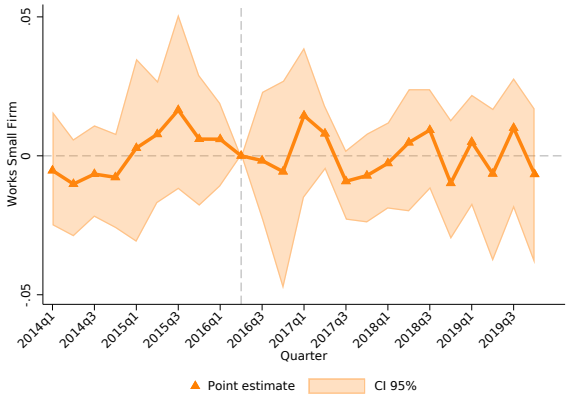
Note: Figure A.4 shows the results from estimating the dynamic version of equation 1 with s being groups of months that belong to the same quarter, eliminating a city at a time. Grey lines and triangles denote the results from each one of this exercises. The orange solid line denotes the baseline coefficients, dashed lines denote the baseline confidence intervals calculated using Wild Bootstrap with 999 repetitions. Panel A, depicts the results using Log Wages as the LHS variable. In panel B, the LHS variable is the probability of being unemployed. Finally, in Panel C, the LHS variable serves as an indicator of labor market participation. All specifications include demographic controls, city, and time (monthly) fixed effects and city trends. Low-wage workers are those with an estimated probability greater than 0.28. High migration cities are cities with a share of Venezuelan immigrants in the Labor force superior to the median across cities in 2005.

Figure A.5: Precision Recall Curves on the Test Set

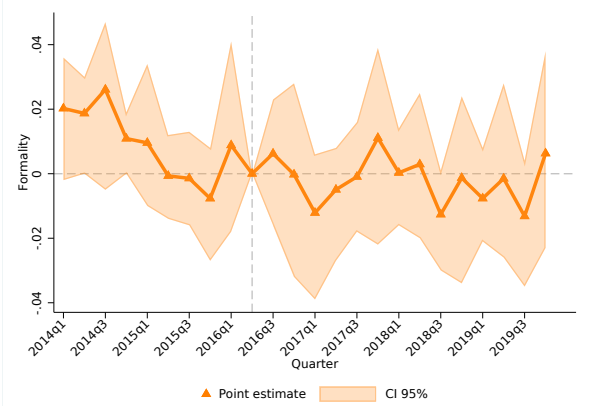


Note: Figure A.5 depicts the out-of-sample performance of each of the XGBoost model and the precision and recall of several demographic classification of low-wage workers using demographic observables. The data used is the test data from the 2012 GEIH sample. LTHS refers tp workers with Less Than High School schooling. Teens are workers from 15 to 25 years old. See the main text for details about the test sample construction and estimation of the XGBoost model.

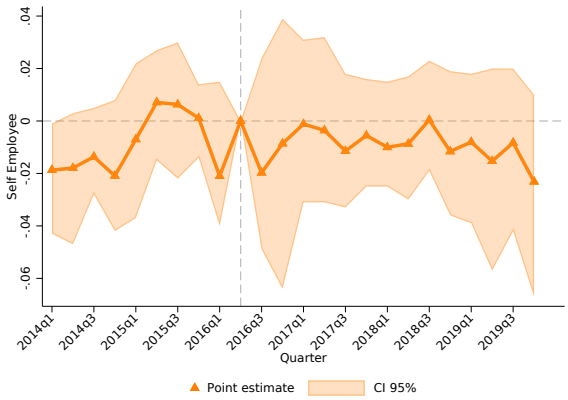
Figure A.6: Dynamic Estimates of the Effect of Migration on Low-Wage Natives



A) Formality: Small Firm



B) Formality: Pension Affiliate



C) Self Employed

Note: Figure A.6 shows the results from estimating the dynamic version of equation 1 with s being groups of months that belong to the same quarter. Treatment is defined as the quarter of the reopening of the border. Panel A, depicts the results using Log Wages as the LHS variable. In panel B, the LHS variable is the probability of being unemployed.

B Machine Learning Models

B.1 Evaluation Metrics: Precision-Recall Curve and AUC-PR

From any binary classification exercise we can obtain the following possible outcomes:

- **True Positives (TP):** Instances correctly predicted as positive.
- **False Negatives (FN):** Instances that are actually positive but incorrectly predicted as negative.
- **False Positives (FP):** Instances that are actually negative but incorrectly predicted as positive.
- **True Negatives (TN):** Instances correctly predicted as negative.

This can be summarized with the *confusion matrix*, which records all possible outcomes of the prediction compared with the true class:

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Table B.1: Confusion matrix of classification outcomes.

Obviously, the objective is to minimize the number of incorrect instances, i.e., FP and FN. But in many application the loss incurred by a FP or a FN may be different. Precision is a measure of how well the model is at reducing False Positives:

$$\text{Precision} = \frac{TP}{TP+FP}$$

In the extreme a Precision of 1 implies not FP. Note that this can be achieve if the model classifies all instances as negative cases. On the other hand Recall is a measure of how well the model is at reducing False Negatives:

$$\text{Recall} = \frac{TP}{TP+FN}$$

Similarly, any model that predicts all instances as positive will yield a 100% recall rate.

This extremes are not optimal in any classification problem. So the objective is to find a balance between Precision and Recall. The precision-recall curve plots Precision against Recall for different threshold values. The AUC-PR is defined as the area under this curve:

$$\text{AUC-PR} = \int_0^1 P(R)\,dR$$

where $P(R)$ is the precision at recall R.

B.2 Lasso Logistic Regression

Lasso logistic regression is a type of regression analysis used for binary classification tasks, which integrates both variable selection and regularization to enhance the prediction accuracy and interpretability of the statistical model it produces. The logistic regression model predicts the probability of the binary response variable Y given the predictor variables $\mathbf{X}$. The functional form of the logistic model is given by:

$$P(Y = 1 \mid \mathbf{X}) = \frac{e^{\mathbf{X}\beta}}{1 + e^{\mathbf{X}\beta}}$$

where $\mathbf{X}$ represents the matrix of input features and β the vector of coefficients to be estimated.

In a lasso logistic regression, an ℓ_1 penalty term is added to the loss function to enforce sparsity in the solution. This penalty term helps in shrinking some of the coefficients exactly to zero, automatically performing variable selection. The regularized loss function for lasso logistic regression is:

$$L(\beta) = -\left(\sum_{i=1}^{n} [y_i \log(P_i) + (1 - y_i) \log(1 - P_i)] \right) + \lambda \sum_{j=1}^{p} |\beta_j|$$

where n is the number of observations, y_i is the observed outcome for observation i, P_i is the predicted probability for i, p is the number of predictors, and λ is the regularization parameter that controls the strength of the penalty.

The lasso logistic regression model is estimated using the `glmnet` package in R, which efficiently implements coordinate descent algorithms to optimize the loss function. The `glmnet` function automatically searches over a grid of λ values to determine the optimal regularization strength, allowing the selection of a model with a balance between complexity and interpretability. Typically, cross-validation is employed to select the best λ parameter that minimizes the error on unseen data.

Figure B.1: Cross-Validation Results for Logistic Regression

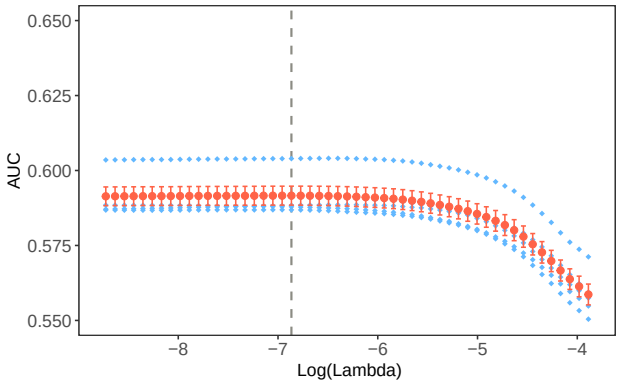


Figure B.1 shows the results of the 5-fold cross-validation exercise for the Lasso-Logit model. The dashed gray line shows the best parameter.

B.3 Random Forest

Random Forest is an ensemble learning method widely used for classification and regression tasks, which constructs multiple decision trees during training and outputs either the mode of the classes (classification) or the mean prediction (regression) of the individual trees. This method enhances predictive accuracy and controls overfitting by leveraging the diversity among the trees.

The basic building block of a Random Forest is a decision tree. For a classification task, each decision tree provides a class prediction, and the Random Forest takes a majority vote to decide the final class. The Random Forest algorithm introduces two main sources of randomness to ensure tree diversity: Bootstrap Sampling: It creates different training subsets by randomly sampling with replacement from the original data features. Feature Selection: At each node of the tree, it selects the best split among a random subset of predictors rather than considering all predictors.

The idea behind integrating multiple decision trees in this way is that individual trees might have high variance, but when aggregated as an ensemble, they provide robust and reliable predictions.

Although the Random Forest model does not require an explicit functional form like parametric models (e.g., logistic regression), it operates based on simple if-then rules structured in decision trees. For classification, the prediction $\hat{y}$ for a given input $\mathbf{X}$ is given by:

$$\hat{y} = \text{mode}(\{T_1(\mathbf{X}), T_2(\mathbf{X}), \ldots, T_m(\mathbf{X})\})$$

where T_k represents the k-th decision tree in the forest, and m is the total number of trees.

The Random Forest model is estimated using packages such as `randomForest` in R. The model is built by growing multiple decision trees as follows:

- Tree Construction: Each tree is grown using a bootstrap sample from the training dataset. A random subset of features is selected at each split for the node decision. - Out-of-Bag (OOB) Error Estimation: Trees are trained using bootstrap samples, leaving one-third of the samples as out-of-bag data to estimate error rates and variable importance internally. - Aggregation: The final prediction for classification tasks is made by majority voting (mode) across all the trees.

Random Forest is advantageous due to its ability to handle large datasets with higher dimensionality and multicollinearity and provides a measure of feature importance, making it both a robust and interpretable model.

Figure B.2: Cross-Validation Results for Random Forest

(a) Mtry (b) Max Nodes

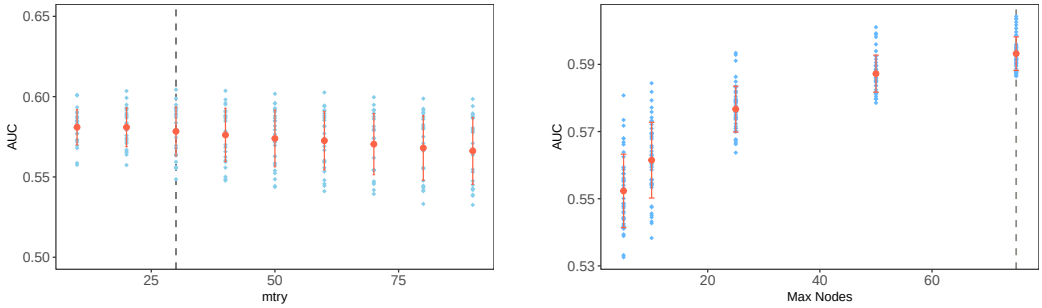


Figure B.2 shows the results of the 5-fold Cross-validation exercise for the Random Forest model. Panel (a) shows the results for the Number of Randomly selected variables in each node to be considered (Mtry) and panel (b) shows the results for the Maximum number of nodes of each tree. The dashed gray line shows the best parameter.

B.4 Boosting

XGBoost (Extreme Gradient Boosting) is a powerful and efficient implementation of gradient-boosted decision trees. It is well-suited for regression and classification tasks due to its performance and scalability. The primary strength of XGBoost lies in its use of second-order gradient information, flexibility for user-defined objectives and evaluation criteria, and a robust handling of missing data.

Functional Form

The model prediction $\hat{y}_i$ for an input $\mathbf{X}_i$ is given as the sum of the predictions from K individual trees:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\mathbf{X}_i), \quad f_k \in \mathscr{F}$$

where $\mathscr{F}$ denotes the space of regression trees (CART). The objective function to be minimized consists of a loss function to measure the goodness of fit and a regularization term to control the complexity of the model:

$$\mathscr{L}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Here, l is a differentiable convex loss function, and $\Omega(f_k)$ is the regularization term defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|\mathbf{w}\|^2$$

where T is the number of leaves in the tree f, $\mathbf{w}$ represents the vector of scores on leaves, γ is the penalty for each leaf, and λ is the L_2 regularization term on leaf weights.

Parameter Tuning

Effective tuning of XGBoost parameters is critical for achieving optimal model performance. Key parameters include:

- **Learning Rate (η):** Controls the step size during the updating of trees. Smaller values require more boosting rounds. Typical values range from 0.01 to 0.3.

- **Max Depth (max_depth):** Determines the maximum depth of a tree. Controls model complexity; deeper trees can capture more complex patterns but are prone to overfitting. Commonly set between 3 to 10.

- **Subsample:** Proportion of samples used for training each tree. Helps prevent overfitting by introducing randomness. Values usually range from 0.5 to 1.

- **Colsample_bytree:** Fraction of features to be randomly sampled for each tree. Helps in controlling overfitting. Usual values range from 0.3 to 0.8.

- **Gamma (γ):** Minimum loss reduction required to make a further partition in a leaf node. Higher values lead to more conservative models.

- **Lambda (λ):** L_2 regularization term on weights, analogous to Ridge regression, providing a way to penalize large coefficients to prevent overfitting.

- **Alpha (α):** L_1 regularization term on weights, analogous to Lasso regression, helpful for inducing sparsity in feature weights.

The optimization of these parameters is generally done through techniques like grid search or automatic hyperparameter tuning libraries, possibly leveraging cross-validation to evaluate performance across different parameter settings. These tuning practices ensure that the model generalizes well on unseen data.

Figure B.3: Cross-Validation Results for XGboost Model

(a) Log Learning Rate (b) Maximum Depth

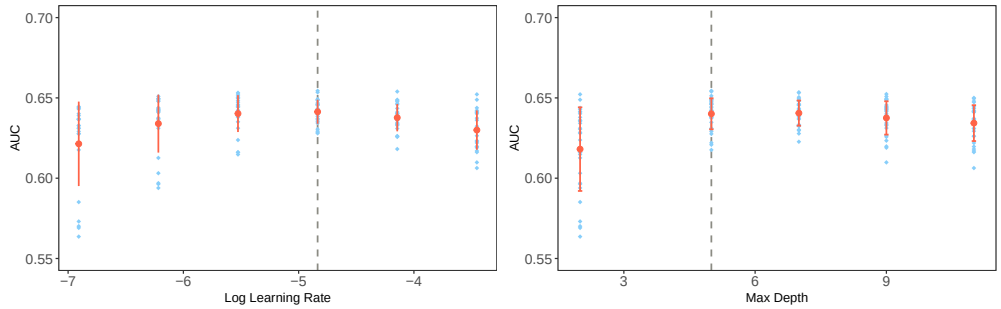


Figure B.3 shows the results of the 5-fold Cross-validation exercise for the XGBoost model. Panel (a) shows the results for the Number of Randomly selected variables in each node to be considered (Mtry) and panel (b) shows the results for the Maximum number of nodes of each tree. The dashed gray line shows the best parameter.

B.5 Neural Network Model

A neural network is a computational model inspired by the way biological neural networks in the human brain process information. It consists of interconnected processing nodes (neurons) which work collectively to solve specific problems. Neural networks are particularly effective for tasks like classification, regression, and pattern recognition. The basic structure of a simple feedforward neural network includes:

- Input Layer: Consists of input neurons, each representing a feature of the data.
- Hidden Layer(s): Consists of one or more hidden layers of neurons which transform the input into meaningful representations. The activation of a neuron in the hidden layer is calculated as:

  $$h_j = \sigma\left(b_j + \sum_{i=1}^{I} x_i w_{ij}\right)$$

  where σ is an activation function (often a sigmoid in `nnet`), b_j is the bias, x_i are input features, and w_{ij} are the weights connecting the input unit i to the hidden unit j.
- Output Layer: Produces the final prediction of the network. For classification tasks, the output might be transformed using a softmax function to produce probabilities.

  Training a neural network involves finding the optimal set of weights W and biases b that minimize a loss function. Commonly, the loss function for a classification task is the cross-entropy loss:

  $$L(y, \hat{y}) = -\sum_{c} y_c \log(\hat{y}_c)$$

  where y is the true distribution (often one-hot encoded) and $\hat{y}$ is the predicted distribution.

  The `nnet` package uses the following process for training:

- Backpropagation The algorithm computes gradients of the loss function with respect to the weights using the backpropagation algorithm and updates them in the direction that reduces the error. The update rule for stochastic gradient descent (SGD), often used in `nnet`, is:

  $$w_{ij}^{(t+1)} = w_{ij}^{(t)} - \eta \frac{\partial L}{\partial w_{ij}}$$

  where η is the learning rate, controlling the step size of each update.
- Convergence Training continues iteratively for a predefined number of epochs or until the change in loss is smaller than a tolerance level, indicating convergence.

The key parameters in the implementation of the model are:

- Size: Number of units in the hidden layer. This affects the model's capacity to learn complex patterns.
- Decay: A regularization parameter to prevent over-fitting by penalizing large weights through an L_2 penalty term.
- Epochs:Number of iterations for which the optimization algorithm will run.

The `nnet` package thus provides a straightforward approach to building and training simple neural network models in R. This fundamental model serves as a building block for more complex architectures, accommodating various types of data.

Figure B.4: Cross-Validation Results for Neural Net Model

(a) Lambda (b) Epchs (c) Number of Neurons

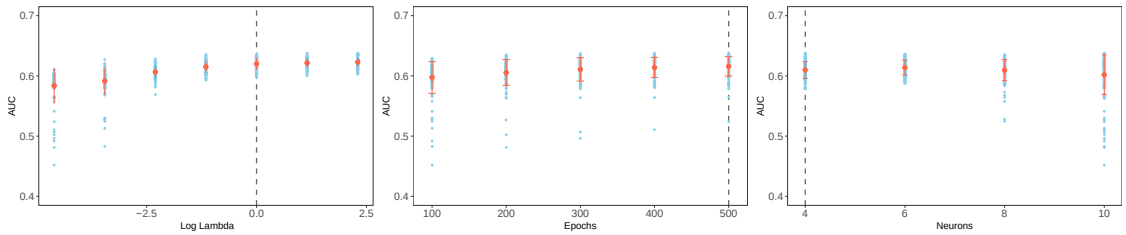


Figure B.4 shows the results of the 5-fold Cross-validation exercise for the Neural Net model. Panel (a) shows the results for the (b) shows the results for the, panel (c)... . The dashed gray line shows the best parameter.

B.6 Super Learner

I implemented the Super Learner algorithm described in Polley and van der Laan (2010). The Super Learner is an ensemble machine learning method that creates an optimal weighted combination of predictions from a library of candidate algorithms. It aims to improve predictive performance by leveraging the strengths of multiple models.

The Super Learner model comprises several steps. First, define a set of candidate learner models, $M_1, M_2, \ldots, M_K$, which include all algorithms previously trained. Second, split the training data into V folds. In the $v = 1 \ldots V$ step, each candidate model is trained on all folds but v , generating $\hat{Y}^{(k)}$ predictions for learner M_k on the validation fold, v.The result is a $N \times K$ matrix with out of sample- cross-validation probabilities.

These probabilities serve as inputs for the meta-learning phase. The task of the meta-learner is to find weights α_k for each model such that:

$$\text{Final Prediction} = \sum_{k=1}^{K} \alpha_k \hat{Y}^{(k)}$$

where $\sum_{k=1}^{K} \alpha_k = 1$, ensuring the predictions are also probabilities.

I select the combination of weights that yield the highest AUC-PR on validation data. Cross-validation ensures that these results generalize to unseen data and is crucial in determining the

robustness of the combined model. This methodology offers a flexible yet powerful approach to building ensembles, providing a path to robust and interpretable models even in challenging scenarios with skewed class distributions by focusing on precision and recall.

B.7 Further Results

Figure B.5, show the histogram of the estimated probability for each year and shows that the distribution of probabilities is stable across time.

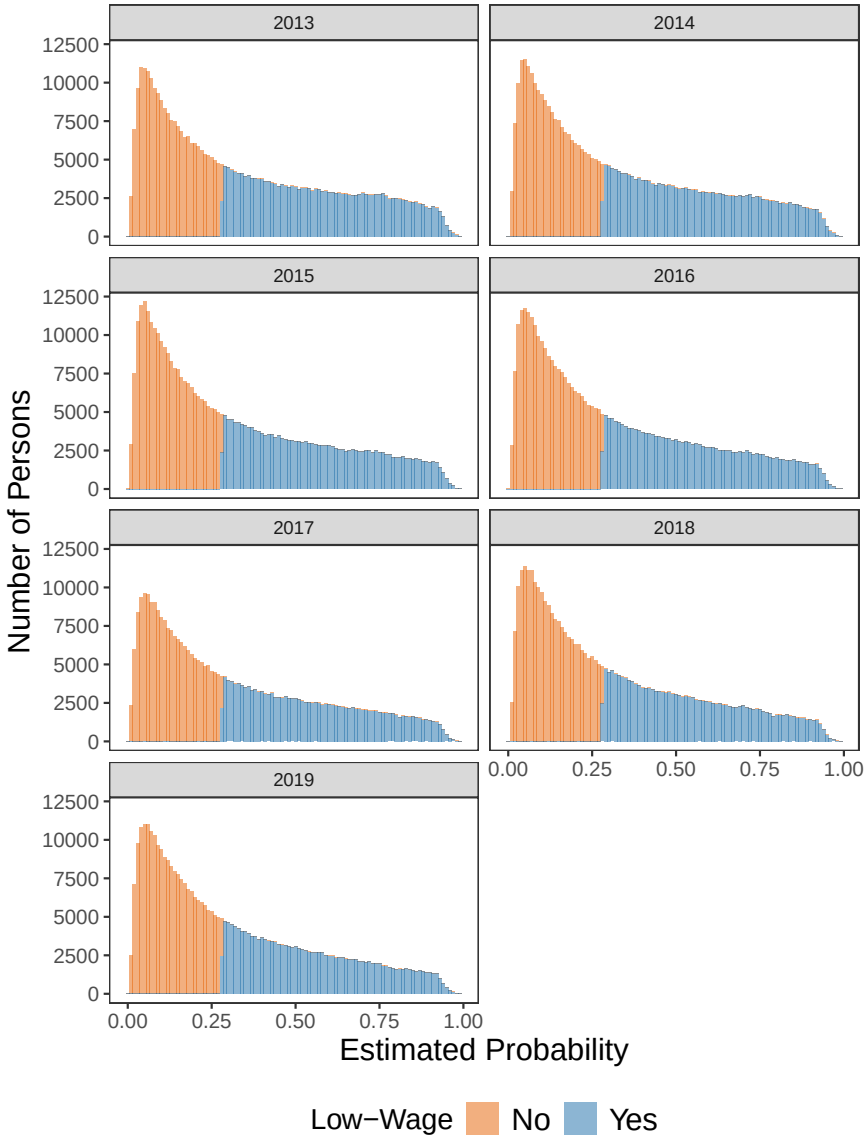


Figure B.5: Distribution of Estimated Probability Across Years

Table B.2: Description of Variables Used in Machine Learning Model

Variable	Description
sex	Gender of the individual (e.g., male or female).
age	Age of the individual in years.
relation_with_head	Relationship of the individual to the household head (e.g., spouse, child, parent).
marital_status	Marital status of the individual (e.g., single, married, divorced).
literacy	Literacy status of the individual (e.g., ability to read and write).
month	Month of the survey, event, or data collection.
city	City or location of residence.
h_usedrooms	Number of rooms used in the household.
h_bathroom_type	Type of bathroom facility in the household (e.g., flush toilet, pit latrine).
h_bathroom_shared	Indicator whether the bathroom is shared with other households.
h_garbage_disposal	Method of garbage disposal in the household (e.g., collected, burned).
h_water_supp	Indicator of household having clean water supply.
h_household_ownership	Ownership status of the household (e.g., owned, rented).
h_amenities01	Household has home phone service.
h_amenities02	Household has water heater or electric shower.
h_amenities03	Household has color television.
h_amenities04	Household has DVD player.
h_amenities05	Household has sound system.
h_amenities06	Household has computer for household use.
h_amenities07	Household has vacuum cleaner or polisher.
h_amenities08	Household has air conditioning.
h_amenities09	Household has fan.
h_amenities10	Household has bicycle.
h_amenities11	Household has motorcycle.
h_amenities12	Household has private car.
h_amenities13	Household has vacation home or apartment.
h_amenities14	Household has washing machine.
h_amenities15	Household has refrigerator.
h_amenities16	Household has blender.
h_amenities17	Household has electric or gas stove.
h_amenities18	Household has electric or gas oven.
h_amenities19	Household has microwave oven.
h_size	Size of the household (number of members).
h_type	Type of housing (e.g., apartment, single-family house).
h_walls	Material of the household walls (e.g., brick, wood).
h_floors	Material of the household floors (e.g., tile, concrete).
h_gas_supp	Source of gas supply for the household (e.g., natural gas, propane).
h_sewerage_supp	Type of sewerage system in the household (e.g., public sewer, septic).
h_stratum	Socioeconomic stratum or level of the household.
h_const_water_supp	Indicator of constant water supply availability.
h_amenities	Overall count or index of household amenities.
h_imputed_rent	Imputed rental value of the household dwelling.
nminors	Number of minors (children under 18) in the household.
offspring	Number of offspring of the individual.
sons	Number of sons of the individual.
brothers	Number of brothers of the individual.
older_brother	Indicator or count of older brothers.