Predicting Mexico-to-US Migration with Machine Learning for Counterfactual Analysis

Parth Chawla and J. Edward Taylor
Department of Agricultural and Resource Economics, University of California, Davis

Selected Poster prepared for presentation at the 2025 AAEA & WAEA Joint Annual Meeting in Denver, CO: July 27-29, 2025

Email for correspondence: chawla@ucdavis.edu

Predicting Mexico-to-US Migration with Machine Learning for Counterfactual Analysis

Parth Chawla [1] J. Edward Taylor [1]

[1]Department of Agricultural and Resource Economics, UC Davis

UC DAVIS

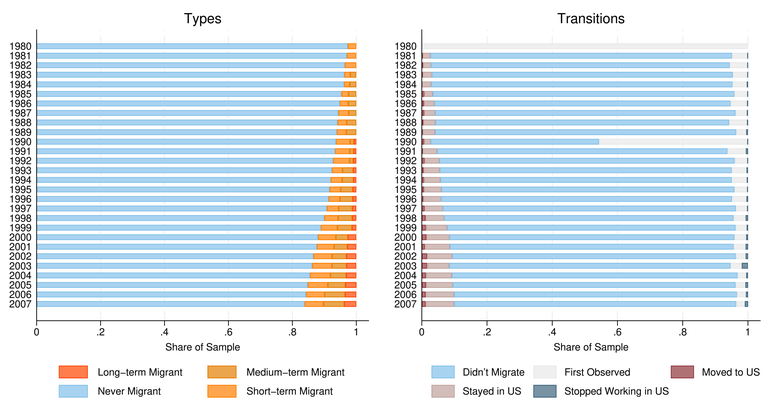
Introduction

- Risks from climate change, conflict, demographic shifts, and economic and political instability are increasing the need for reliable tools to predict migration patterns.
- Mexican migrants play a critical role in the US economy, comprising 63% of the farm workforce and large shares in construction, hospitality, and other labor-intensive sectors.
- Machine learning (ML) methods, particularly tree-based algorithms, can uncover complex, nonlinear relationships that conventional linear models often miss.
- ML models with high out-of-sample predictive accuracy can be used to simulate migrant responses to economic and policy shocks. These simulations can help identify the factors that shape how migration patterns may evolve in response to future shocks.
- However, collecting detailed migrant data through surveys is often expensive and time-consuming. There is a need for models that maintain strong predictive performance using data that is either publicly available or relatively inexpensive to collect.

Data and Descriptive Statistics

- We use data from the Mexico National Rural Household Survey, covering 1,514 rural households across 80 communities spanning all five census regions in Mexico.
- Our *ideal dataset* is a panel tracking the employment locations of **10,739 individuals** annually from 1980 to 2007.
- **11.8%** of individuals in the sample migrated to work in the US at least once between 1980 and 2007, with the annual share rising steadily over time.
- On average, **7.1%** of person-year observations correspond to individuals migrating in a given year. In 2007 (the year used as our out-of-sample test year) the migrant share is **10%**, corresponding to **735 migrants**.

Share of Migrant Types and Transitions Over Time



Long–term Migrant: More than 10 years observed in the US; Medium–term Migrant: 4 to 10 years observed in the US; Short–term Migrant: 1 to 3 years observed in the US.

Model Training and Testing

- We trained our benchmark model using LightGBM, a tree-based gradient-boosting framework, on the *ideal dataset*, the full 1980-2006 panel with basic household and individual characteristics, and individual migration histories.
- The outcome variable is a binary indicator for **whether an individual migrated to work in the US in a given year**.
- The model was tested out-of-sample on 2007. We used a sliding 5-year cohort strategy: training on the first four years of each cohort and validating on the fifth to tune hyperparameters, before retraining on all pre-2007 data.
- Using the same data, we also trained a simpler logistic regression model for comparison. Since LightGBM builds many small decision trees that capture non-linear patterns and interactions, it performs substantially better.

Model	Data Used for Model Training				Precision	Recall	F1
	Years	Migration History	Basic HH & Ind. Chars.	Weather			
LightGBM (Benchmark)	**1980-2006**	**Yes**	**Yes**	**No**	**0.81**	**0.89**	**0.85**
Logistic Regression	1980-2006	Yes	Yes	No	0.51	0.92	0.66
LightGBM	2003-2006	No	Yes	No	0.79	0.25	0.38
LightGBM	**2003-2006**	**No**	**Yes**	**Yes**	**0.81**	**0.69**	**0.75**
Logistic Regression	2003-2006	No	Yes	No	0.18	0.81	0.29
Logistic Regression	2003-2006	No	Yes	Yes	0.17	0.81	0.28

- Next, we tested both models using a *restricted dataset*: training only on the last four years (2003-2006) and excluding migration history variables.
- Performance declines for both models under this *restricted* setting.
- However, adding readily available weather variables to the restricted dataset significantly improves performance for LightGBM, which comes within 0.1 points of the benchmark.
- Therefore, even with *less-than-ideal data*, decision tree-based models can approach benchmark performance, and outperform simpler parametric models like logistic regression, by incorporating simple, publicly available inputs like weather.

Definitions of Model Evaluation Metrics

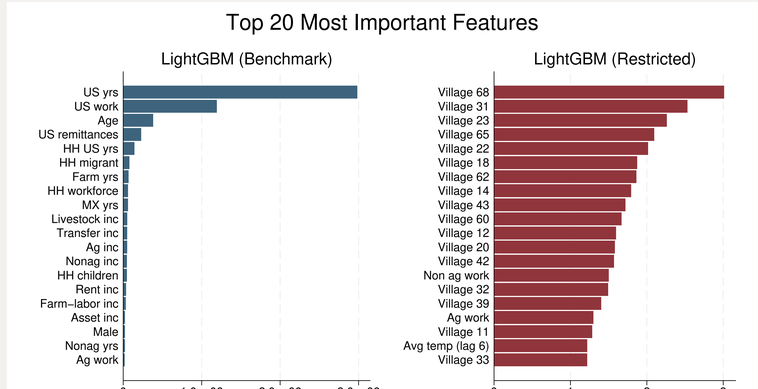
- **Precision:** Of all the individuals *predicted* to migrate in 2007, the fraction who actually did.
- **Recall:** Of all the individuals who *did* migrate in 2007, the fraction the model correctly identified.
- **F1:** Harmonic mean of Precision and Recall, balancing "how correct" vs. "how complete".

Model Features or Variables

- **Basic Household & Individual Characteristics** (lagged): Gender, age, no. of children, household size, employment sector (ag, nonag), and income from various sources.
- **Migration History** (lagged): Indicators for worked in Mexico/US, whether another household member migrated, and cumulative years in the US (individual and household).
- **Weather Variables** (lagged, seasonal): Average temperature, total precipitation, growing-degree days, and heating-degree days.

Feature or Variable Importance

- The benchmark LightGBM model (trained on 1980-2006 with individual migration history) draws most of its predictive power from past US migration experience, remittances, age, and household migrant indicators.
- The restricted LightGBM model (trained on 2003-2006 without migration history or remittances but including weather data) achieves a close F1 score (within 0.1), relying more evenly on location, sectoral employment, and weather conditions.



All predictors are one–year lagged. Values show relative contribution to reduction in binary log–loss.
Benchmark trained on 19807?2006 with migration history. Restricted trained on 20037?2006 without migration history but with weather data.

Counterfactual Simulations of Shocks

- We use the restricted model to run counterfactual simulations by shocking policy-relevant variables in the test year. Since predictors are lagged, outcomes in 2007 are affected by shocks applied in 2006.
- A 10% increase in avg. temperature in 2006 reduces predicted migration by 13% in 2007. Simulating demographic or economic shocks, such as increasing age by 10% or reducing income by 10%, lowers migration by 17% and 18%, respectively.