A Machine Learning Approach to Analyze Income Determinants in the Mexican Labor Market

The Evolving Role of Education (2000–2020)

Anahí Reyes Miguel

Introduction to Machine Learning
École Polytechnique, ENSAE, Télécom Paris

April 15, 2025

Overview

1. Motivation
2. Research Question
3. Literature & Prior Work
4. Data
5. Methodology
6. Results
7. Conclusion

Motivation

- Changes in the earnings gap between educational groups were the main driver of inequality in Mexico between the 1980s and mid-1990s [Lopez-Acevedo, 2006].
- This relationship is shaped by structural features of the Mexican labor market, such as informality, regional disparities [Hausmann et al., 2021], and education-occupation mismatches [Quinn and Rubb, 2006].
- Education has long been recognized as a key determinant of income, and its relationship has been widely studied in the literature using parametric and non-parametric methods [Becker, 1964, Mincer, 1974].
- This project contributes by updating the analysis of education's role in the determination of income by using recent data and a predictive Machine Learning (ML) approach.

Research Question

Has the importance of education in predicting income increased over time (2000–2020), relative to other factors such as age, gender, occupation, or local labor market conditions in Mexico?

Literature & Prior Work

- ML is particularly promising for economics when empirical analysis is framed as an algorithmic process that estimates and compares multiple models, by integrating tuning—or data-driven model selection [Athey, 2019].
- [Matkowski, 2021] compares traditional and ML-based income prediction models using U.S. Census data. His work demonstrates that Gradient Boosting and Random Forest outperform OLS, with education consistently emerging as a key predictor, even in the presence of high-dimensional features.
- In the Mexican context, [Gomez-Cravioto et al., 2022] show that, overall, the Gradient Boosting model performed best in both regression and classification tasks, with SHAP values used to interpret feature contributions. However, they also found that linear and logistic regression models were more adequate for describing the relationships between variables and the first income after graduation.

Data Description

- **Source:** Banco de México harmonized census microdata (2000, 2010, 2015, 2020)[1].
- **Sample size:** A random sample of 72,000 individuals was drawn from a total of approximately 8 million census records.
- **Variables:** The dataset contains 66 variables, including income, education, age, gender, occupation, ethnicity, religion, economic activity, commute time, and local labor market.
- Each row represents an individual surveyed in the census.

[1] See the SIDIE Datasets documentation and methodological notes from Banco de México, available at: `https://www.banxico.org.mx/DataSetsWeb/dataset?ruta=LLM&idioma=en` collected by Instituto Nacional de Estadística y Geografía (INEGI).

Data Processing

- **Missing Data:** Variables with >10% missing values were removed.
- **Sample Filtering:** Restricted to individuals aged 18+ with strictly positive labor income, focusing on the economically active population.
- **Log Transformation:** Applied `log1p` ($\log(1 + x)$) to income to reduce skewness, stabilize model performance, and handle zero values safely.
- **Feature Engineering:** Created a grouped variable `ETHNIC_MINORITY` that captures self-reported Indigenous or Afrodescendant identity.
- After preprocessing, the final analytical sample includes 15,996 individuals.
- At this step, the data was split into a training set (90%, 14,396) for model estimation and a testing set (10%, 1,600).

Data Exploration

	Accumulated Schooling Years	Age	Income
Count	14,396	14,396	14,396
Mean	8.83	37.19	32,929.40
Std	4.53	13.49	1,666,680.71
Min	0.00	18.00	2.00
25%	6.00	26.00	2,143.00
Median	9.00	35.00	3,440.00
75%	12.00	46.00	6,000.00
Max	23.00	98.00	99,999,999.00

Table: Descriptive Statistics of Key Numerical Variables

- On average, individuals have 8.8 years of schooling and are 37 years old.
- Reported income is highly dispersed, with a median of 3,440 pesos and a maximum close to 100,000,000 pesos.

Data Exploration

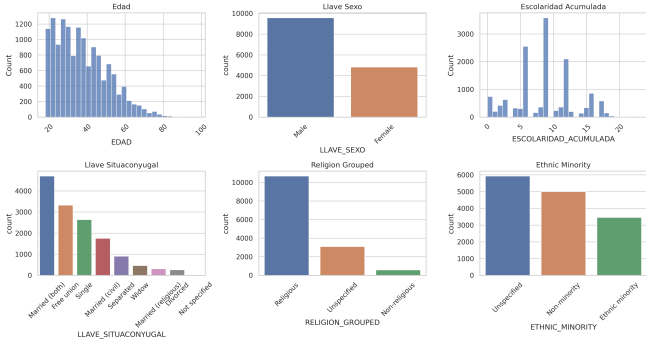

Figure: Distributions of key demographic characteristics in the training dataset.

- Most individuals are aged 20–50, two-thirds are male, and schooling typically ranges from 6 to 12 years.
- The majority are married and religious, and about 30% identify as part of an ethnic minority.

Data Exploration

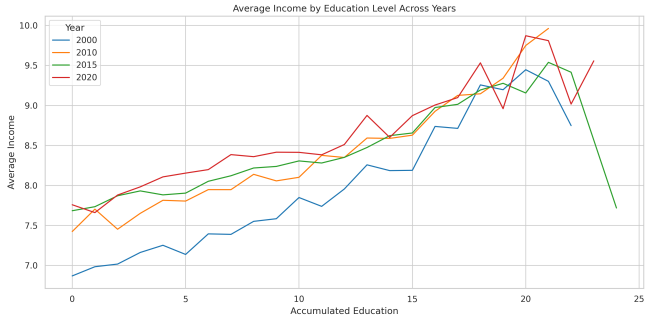

Figure: Average log-income by accumulated years of education across census years (2000–2020).

- Income increases with education across all years, although the slope flattens or is even negative at higher education levels.
- Differences in the slopes between years suggest temporal shifts in returns to education.

Data Preparation and Transformation

I implemented a series of transformations to ensure compatibility with scikit-learn pipelines, enhancing model performance and interpretability [Pedregosa et al., 2011].

- **Missing values:** Missing values were handled using `SimpleImputer`. The median was used for numerical variables and the most frequent category (mode) for categorical variables.
- **Standardization of numerical variables:** All numerical variables were standardized using `StandardScaler`.

Data Preparation and Transformation

- **Interaction Variable:** Interaction terms were calculated by applying a polynomial transformation (degree 2) to Accumulated Schooling Years (`ESCOLARIDAD_ACUMULADA`) and Year (`ANIO`).
- **Encoding Categorical Variables:**
  - One-Hot Encoding for low-cardinality variables (fewer than or equal to 10 categories), such as gender and marital status.
  - Target Encoding for the local labor market with 713 categories, using *Target Encoding*, which replaces each category with the mean log income (the target variable) for that group.

Modeling Approach

- Five models were implemented to balance interpretability and predictive power:
  - **Linear Regression**
  - **Lasso** and **Ridge Regression** (regularized linear models)
  - **Random Forest** and **Gradient Boosting** (tree-based ensemble models)
- All models evaluated using **5-fold cross-validation**.
- Lasso and Ridge tuned using `LassoCV` and `RidgeCV`.
- Tree-based models optimized via **grid search** on a data subset to reduce runtime.

Results

Model	MSE	MAE	R^2
Gradient Boosting	0.4804	0.4980	0.4018
Linear Regression	0.4823	0.4984	0.3995
Ridge Regression	0.4823	0.4984	0.3994
Lasso Regression	0.4898	0.5021	0.3901
Random Forest	0.5526	0.5401	0.3119
Baseline (Mean Predictor)	0.8031	0.6726	0.0000

Table: Performance of each model.

- Gradient Boosting achieved the best overall results with the lowest MSE and MAE, and the highest R^2, meaning it explains about 40% of the variance in log-transformed income.
- Linear and Ridge regression performed similarly ($R^2 \approx 0.3995$), while Lasso was slightly less accurate.
- Random Forest showed lower predictive accuracy ($R^2 = 0.3119$).

Results

Model	R^2	Top 2 Features
Gradient Boosting	0.4018	1. MERCADO_TRABAJO_LOCAL 2. ANIO $\times$ ESCOLARIDAD_ACUMULADA
Linear Regression	0.3995	1. ANIO $\times$ ESCOLARIDAD_ACUMULADA 2. ESCOLARIDAD_ACUMULADA
Ridge Regression	0.3994	1. ANIO $\times$ ESCOLARIDAD_ACUMULADA 2. ESCOLARIDAD_ACUMULADA
Lasso Regression	0.3901	1. EDAD^2 2. ANIO
Random Forest	0.3119	1. ESCOLARIDAD_ACUMULADA $\times$ ANIO 2. MERCADO_TRABAJO_LOCAL

Table: Comparative of the two top predictive variables across models.

Feature Importance and Interpretation

- In terms of absolute predictive error, an MAE of 0.50 in log-income terms implies that predictions deviate from the true values by a factor of about 1.65 on average, that is, predicted income is typically 65% higher or lower than the actual value.
- For example, a predicted income of 5,000 pesos might correspond to a true value between roughly 3,030 and 8,250 pesos (among 0.82 and 2.2 minimum salaries at 2020).

Feature Importance and Interpretation

- I chose the two best peforming models to analyze further how education predicts income over time: Gradient Boosting Regressor and Linear Regressor.
- I use SHAP (SHapley Additive exPlanations) values for interpreting feature contributions across both linear and non-linear models by quantifying the marginal impact of each input feature on a given prediction.

Gradient Boosting Regressor

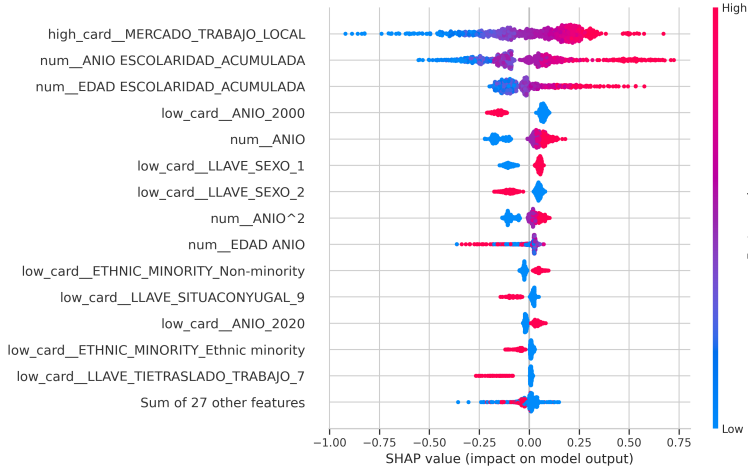

Figure: SHAP Beeswarm Plot for Gradient Boosting Regressor

Linear Regression

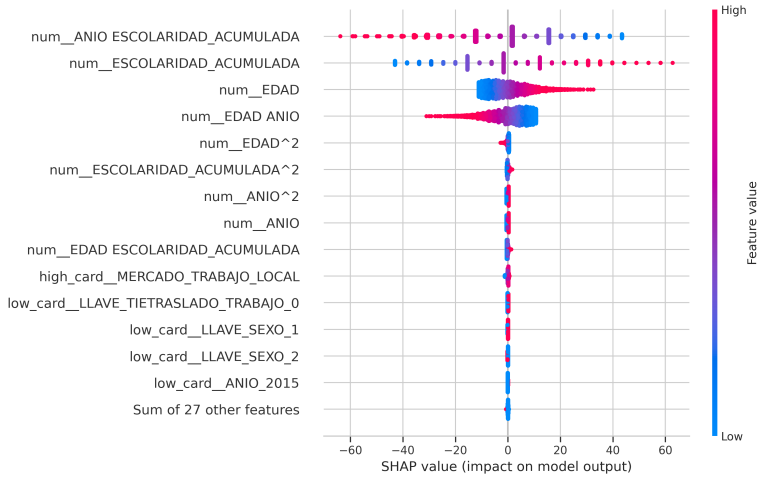

Figure: SHAP Beeswarm Plot for Linear Regression

Interpretation of Mixed Results

- My findings suggest that the local labor market context in which individuals work may influence their income, suggesting that earnings are not solely determined by personal characteristics but are also shaped by structural and geographic factors.
- This raises an important question: Is income more strongly influenced by local labor market conditions or education?
- The evolving role of education cannot be fully captured by a single modeling strategy and it could be useful to adopt other modeling approaches.

Conclusion

- This study aimed to explore whether the importance of education in predicting income in Mexico has increased over time, using interpretable machine learning methods.
- Understanding income inequality requires not just better data—but better tools to interpret it. This study shows how ML can guide that path.
- The results were mixed, making it unclear whether education is the most important predictor.
- My findings suggest that both regional context and changing returns to education play a role in shaping income inequality in Mexico.
- Future research could benefit from incorporating household-level variables, more granular labor market indicators, and richer data to better understand the drivers of income.

Questions?

References I


Athey, S. (2019).
The impact of machine learning on economics.
In Agrawal, A., Gans, J., and Goldfarb, A., editors, *The Economics of Artificial Intelligence: An Agenda*, chapter 14, pages 507–547. University of Chicago Press, Chicago.


Becker, G. (1964).
Human Capital.
National Bureau of Economic Research, New York.


Gomez-Cravioto, D. A., Diaz-Ramos, R. E., Hernandez-Gress, N., and Ceballos, H. G. (2022).
Supervised machine learning predictive analytics for alumni income.
Journal of Big Data, 9(1):11.


Hausmann, R., Pietrobelli, C., and Santos, M. A. (2021).
Place-specific determinants of income gaps: New sub-national evidence from mexico.
World Development, 146:105566.

References II


Lopez-Acevedo, G. C. (2006).
Mexico: Two decades of the evolution of education and inequality.
Research Working Paper WPS3919, The World Bank, Washington, DC.
Disclosed in July 2010.


Matkowski, M. (2021).
Prediction of individual income: A machine learning approach.
Bachelor's thesis, Bryant University.
CC-BY-NC-ND licensed.


Mincer, J. (1974).
Schooling, Experience, and Wages.
National Bureau of Economic Research, New York.

References III

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011).

Scikit-learn: Machine learning in Python.

Journal of Machine Learning Research, 12:2825–2830.

Quinn, M. A. and Rubb, S. (2006).

Mexico's labor market: The importance of education-occupation matching on wages and productivity in developing countries.

Economics of Education Review, 25(2):147–156.