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

The Role of Education from 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. First Section

7. Second Section

Motivation

- Education has long been recognized as a key determinant of income, and its relationship has been widely studied in the literature.
- This relationship is shaped by structural features of the Mexican labor market, such as informality, regional disparities, and education-occupation mismatches.
- Evidence shows that changes in the earnings gap between educational groups have been the main driver of income inequality in Mexico over the 1980s and mid-1990s.
- 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 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

- [Herrera et al., 2023] use XGBoost and neural networks to show how education and digital access jointly shape income distribution in Brazil using survey data, and rely on SHAP values to interpret interactions.
- [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 predicto, 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 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.
- **Key Variables:** 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`.

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` to income to reduce skewness, stabilize model performance, and handle zero values safely.
- **Scaling:** Standardized all numeric variables to mean 0 and unit variance.
- **Feature Engineering:** Created grouped variables (`RELIGION_GROUPED`, `ETHNIC_MINORITY`) and interaction terms (e.g., `ANIO` × `ESCOLARIDAD`).
- After preprocessing, the final analytical sample includes 15,996 individuals.
- To ensure fair evaluation, the data is split into a training set (90%, 14,396) for model estimation and a testing set (10%, 1,600).

Data Exploration

	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

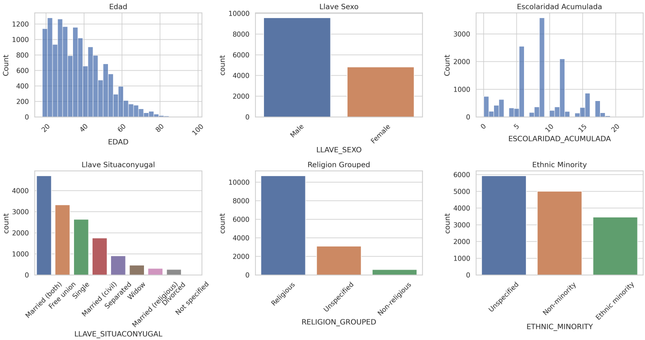

Figure: Distributions of key demographic characteristics in the training dataset.

- 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 one-third identify as part of an ethnic minority.

Data Exploration

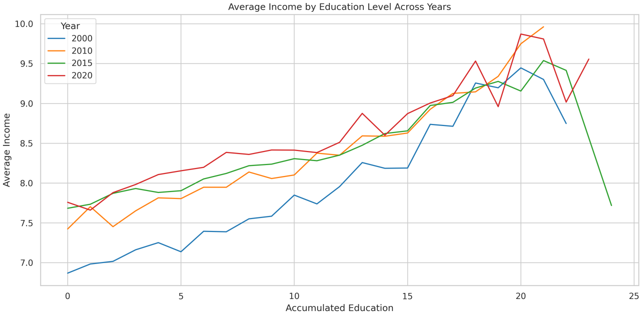

Figure: Average log-income by accumulated years of education across census years (2000–2020).

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

- **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 `ESCOLARIDAD_ACUMULADA` and `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 high-cardinality variables, local labor markets, using *Target Encoding*, which replaces each category with the mean log income (the target variable) for that group.

Modeling Approach

Five models were implemented:

- **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.

Bullet Points

- Lorem ipsum dolor sit amet, consectetur adipiscing elit
- Aliquam blandit faucibus nisi, sit amet dapibus enim tempus eu
- Nulla commodo, erat quis gravida posuere, elit lacus lobortis est, quis porttitor odio mauris at libero
- Nam cursus est eget velit posuere pellentesque
- Vestibulum faucibus velit a augue condimentum quis convallis nulla gravida

Blocks of Highlighted Text

In this slide, some important text will be highlighted because it's important. Please, don't abuse it.

Block

Sample text

Alertblock

Sample text in red box

Examples

Sample text in green box. The title of the block is "Examples".

Multiple Columns

Heading

1. Statement
2. Explanation
3. Example

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer lectus nisl, ultricies in feugiat rutrum, porttitor sit amet augue. Aliquam ut tortor mauris. Sed volutpat ante purus, quis accumsan dolor.

Table

Treatments	Response 1	Response 2
Treatment 1	0.0003262	0.562
Treatment 2	0.0015681	0.910
Treatment 3	0.0009271	0.296

Table: Table caption

Theorem

Theorem (Mass–energy equivalence)

$E = mc^2$

Figure

Uncomment the code on this slide to include your own image from the same directory as the template .TeX file.

Citation

An example of the `\cite` command to cite within the presentation:

This statement requires citation [**?**].

References

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.

Herrera, G. P., Constantino, M., Su, J.-J., and Naranpanawa, A. (2023).
The use of icts and income distribution in brazil: A machine learning explanation using shap values.
Telecommunications Policy, 47(8):102598.

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

The End