Exploring Socio-Demographic Patterns in Income Classification Using Machine Learning

Sara Borello[1] **and Keita Jacopo Viganò**[1]

[1]MSc Students in Data Science, University of Milano-Bicocca (Unimib)

This study aims to build a predictive model for classifying individuals earning over $50k annually using a structured and systematic approach. The process began with a comprehensive preprocessing phase, including feature engineering, to enhance the dataset's quality and relevance. Multiple machine learning models were then trained, with tailored preprocessing steps applied to individual models when necessary to optimize their performance. To ensure a robust and threshold-independent evaluation, model performance was assessed using ROC curves, allowing for unbiased comparisons across models. The best-performing model was selected, and an optimal threshold was determined to maximize classification accuracy.

Contents

1 Introduction 1

2 Data & Data Exploration 1

3 Pre-Processing 2
- A Missing Values . . . 2
- B Feature Engineering . . . 2
  - B.1 Grouping . . . 2
- C Zero Variance and Separation . . . 3
- D Model Selection . . . 3
- E Under Sampling . . . 3

4 Model Training 4

5 Classification Performance Evaluation 5

6 Conclusion 6

7 Bibliography 7

1 Introduction

The idea is to focus on analyzing data from 1994 to build predictive models that identify the socio-economic factors influencing whether an individual earned more than $50K at the time. Studies have shown that historical socio-economic data provide valuable insights into patterns of inequality and mobility (1). The study could then compare these findings with equivalent data from 2024, offering a lens to understand how these determinants have evolved over three decades. This approach aligns with research that emphasizes the importance of longitudinal data in tracking economic disparities and their drivers (2).

For example, in 2024, an algorithm that predicts whether a person earned more than $50K in 1994 could help uncover the key drivers of income during a pivotal historical period. A guiding research question might be: *"What socio-demographic and occupational factors influenced income levels in the United States in 1994, and how have these dynamics evolved by 2024?"*

Research on the evolution of labor markets and technological disruptions suggests that such comparative analyses can reveal how economic transformations shape opportunities and outcomes (3). This analysis could inform how economic, technological, and social changes over the last thirty years have reshaped labor markets and income distribution. By leveraging historical insights, it could support the development of more equitable economic policies and advanced predictive models for today's labor markets, with applications in professional training, workforce development, and efforts to reduce income inequality (4).

2 Data & Data Exploration

The dataset, derived from the 1994 Census database, consists of 15 variables, including 14 features and 1 target variable. These variables capture socio-demographic and economic attributes of individuals living in the United States, providing a combination of categorical and numerical data to predict whether an indi-

vidual earns over $50K annually:

1. **Age**: Numerical, representing an individual's age in years.
2. **Workclass**: Categorical, describing the type of employment (e.g., "Private," "Self-emp-not-inc").
3. **fnlwgt (Final Weight)**: Represents a population weighting factor used by the Census Bureau to adjust for sampling biases
4. **Education**: Categorical, indicating the highest education level (e.g., "HS-grad," "Some-college").
5. **Education.num**: Numerical, representing education levels (1 for lowest, 16 for highest).
6. **Marital Status**: Categorical, describing marital status (e.g., "Married-civ-spouse," "Never-married").
7. **Occupation**: Categorical, describing the type of job (e.g., "Prof-specialty," "Craft-repair," "Sales").
8. **Relationship**: Categorical, indicating family roles (e.g., "Husband," "Not-in-family").
9. **Race**: Categorical, representing racial identity (e.g., "White," "Black").
10. **Sex**: Categorical, indicating gender ("Male," "Female").
11. **Capital.gain**: Numerical, representing capital gains from investments.
12. **Capital.loss**: Numerical, representing capital losses from investments.
13. **Hours.per.week**: Numerical, indicating weekly work hours.
14. **Native.country**: Categorical, indicating country of origin.
15. **Income**: Binary target variable, indicating "<=50K" or ">50K."

The dataset contains 32,561 rows, featuring both categorical variables, such as "workclass" and "education," and numerical variables, such as "age" and "hours.per.week." The target variable is binary, with a distribution of "<=50K" at 76% and ">50K" at 24%.

3 Pre-Processing

A. Missing Values. : The dataset has missing values in **workclass** (5.64%), **occupation** (5.66%), and **native.country** (1.79%), while the remaining variables are complete. Assuming the missing data are **Missing Completely at Random (MCAR)**, their occurrence is independent of both observed and unobserved variables, meaning the missing cases can be considered a random sample. To maintain data integrity and simplify analysis, these missing values were removed, preventing potential biases or inaccuracies that could arise from imputation.

B. Feature Engineering. : The variable **fnlwgt (Final Weight)** represents a population weighting factor used by the Census Bureau to adjust for sampling biases and ensure that the dataset reflects the broader population demographics. However, it is not directly informative for predictive modeling, as it does not provide individual-level information relevant to the target variable, such as personal attributes or socio-economic characteristics. Instead, it introduces redundancy and potential noise into the model, as its primary purpose is to adjust for survey design rather than predict individual outcomes. Therefore, to improve model efficiency and focus on variables directly associated with income prediction, **fnlwgt** was excluded from the analysis.

B.1. Grouping. : Variables with many levels were grouped to reduce dimensionality, prevent sparsity, and improve model performance while retaining key distinctions.

Native Country: This variable was grouped due to its many levels, some with very low frequencies. This approach provided a more balanced distribution across the levels, following the rules below:

- **Europe**: Germany, Italy, England, Poland, Portugal, Greece, France, Ireland, Scotland, Hungary,

Yugoslavia, Holand-Netherlands

- **Asia**: Philippines, India, China, Vietnam, Japan, Iran, Taiwan, Cambodia, Thailand, Laos, Hong
- **South America**: Cuba, Columbia, Haiti, Dominican-Republic, Ecuador, Guatemala, Peru, El-Salvador, Nicaragua, Trinadad&Tobago, Jamaica, Puerto-Rico, Honduras
- **Africa**: South Africa
- **Canada**: Canada
- **Mexico**: Mexico
- **Oceania**: Outlying-US(Guam-USVI-etc)

Mexico and Canada were retained as a separate category due to their significant representation in the dataset.

Education: This variable was grouped into broader categories to simplify analysis and enhance interpretability. Lower levels like "Preschool" and "1st-4th" were grouped as **Primary**, intermediate levels as **Middle**, high school levels as **High School**, post-secondary as **Higher Education**, and advanced degrees as **Postgraduate**. This reduced granularity provides a clearer overview and improves model efficiency by minimizing sparsity. Retaining both variables, Education and Education.num, is appropriate because, during feature engineering, grouping the levels of Education is necessary due to the large number of categories. Keeping Education.num alongside this grouped variable ensures that the model retains precise and detailed information about education, complementing the grouped representation. This decision is validated by XGBoost's feature importance, where Education.num emerges as a key predictor.

Education Values	**Mapped Category**
Preschool, 1st-4th, 5th-6th	Primary
7th-8th, 9th, 10th, 11th	Middle
12th, HS-grad	High School
Some-college, Assoc-voc, Assoc-acdm, Bachelors, Prof-school	Higher Education
Masters, Doctorate	Postgraduate

Table 1. Mapping of Education Levels to Categories

C. Zero Variance and Separation. : All variables in the dataset have non-zero variance, meaning there are no constant features. This ensures that all features contribute meaningful information to the analysis, as none are invariant across the dataset. Additionally, there are no issues of separation, meaning that no single feature or combination of features perfectly predicts the target variable.

D. Model Selection. : Model selection was performed using the **Boruta** feature selection algorithm, chosen for its ability to identify all relevant features by comparing the importance of actual features to randomized shadow features. This ensures that only truly informative variables are retained while discarding irrelevant ones. Following the application of this method, all variables in the dataset were found to be relevant and were retained for the analysis.

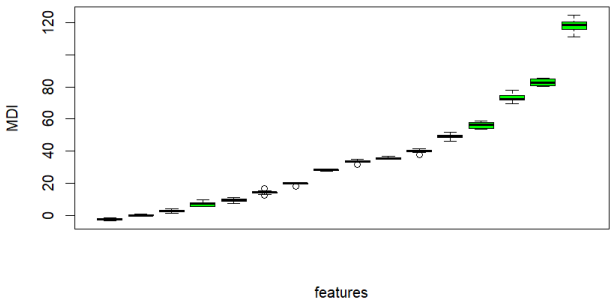

Fig. 1. Model Selection by Boruta

E. Under Sampling. : Given the target variable's distribution, with "<=50K" at 76% and ">50K" at 24%, and the large size of the dataset, undersampling was applied to balance the two classes. This approach ensures that the model is not biased toward the majority class, improving its ability to generalize and perform well on both classes. After the undersampling process, the dataset was split into training and testing sets, with 80% of the data used for training and 20% for testing.

Specific preprocessing steps were applied directly to the selected models to optimize their performance.

4 Model Training

In models where it is permitted, hyperparameters will be tuned using 10-fold cross-validation to maximize performance. The metric chosen for hyperparameter tuning is **sensitivity**, which focuses on minimizing false negatives. Focusing on sensitivityis crucial in this context because the goal is to correctly identify individuals earning more than $50K, as missing these cases would distort the analysis of socio-economic factors that contribute to higher income levels. Sensitivity ensures that the model captures these positive instances accurately, enabling a more reliable understanding of income dynamics. This is particularly important when comparing 1994 to 2024, as it allows for a consistent and unbiased analysis of how the key drivers of income have evolved, ensuring that critical patterns and trends are not overlooked.

Logistic Regression. The *logistic regression* model required specific preprocessing steps to ensure proper functionality. Missing values, already handled during the initial preprocessing phase, were addressed to ensure a clean dataset. A check for *collinearity* was conducted, as high collinearity among independent variables can lead to issues in parameter estimation; in this case, no significant correlation was found between the variables. Additionally, data *normalization* was applied to mitigate the impact of outliers and maintain stable model performance. After completing these preprocessing steps, the model was trained and achieved a final accuracy of 0.815 and a sensitivity of 0.834 on the test set.

Naive Bayes. The *Naive Bayes* model required preprocessing similar to logistic regression. *Collinearity* was checked to ensure the *independence assumption* underlying this model, and no significant collinearity was found among the variables. The data were then *normalized* to standardize feature scales and improve model performance. After these preprocessing steps, the model was trained and achieved an accuracy of 0.74 and a sensitivity of 0.569 on the test set.

Random Forest. For the *Random Forest* model, hyperparameters were tuned over 10 iterations of cross-validation to maximize *sensitivity*. The tuning process adjusted critical parameters: the fraction of features considered at each split (*col_frac*), set to 0.564 to enhance tree diversity, and the minimum number of samples in a leaf node (*Min_child_size*), set to 19 to reduce overfitting. The maximum tree depth (*Max_depth*) was limited to 19 to balance complexity and performance, while the number of trees in the forest (*Nmodel*) was set to 293 to improve generalization through ensemble predictions. After training, the model achieved an accuracy of 0.829 and a sensitivity of 0.857.

Gradient Boosting. In the *Gradient Boosting* model, hyperparameters were tuned over 10 iterations of cross-validation to maximize *sensitivity*. The tuning process optimized key parameters, including the data fraction (*Data_fraction*), set to 0.8336 to determine the proportion of data used at each iteration, and the minimum number of samples in a leaf node (*Min_child_size*), adjusted to 174. The maximum tree depth (*Max_depth*) was limited to 4, while the learning rate (*Learning_rate*) was set to 0.156 to control the contribution of each tree to the overall model. Additionally, the number of trees in the ensemble (*Nmodel*) was set to 140. After training, the model achieved an accuracy of 0.831 and a sensitivity of 0.856, reflecting its effectiveness in identifying positive cases while maintaining robust accuracy.

XGBoost. For the *XGBoost* model, hyperparameters were tuned over 10 iterations of cross-validation to maximize *sensitivity*. Parameters such as maximum tree depth (*Max_depth = 4*), minimum child size (*Min_child_size = 2*), and learning rate (*Learning_rate = 0.07*) were tuned to control complexity and optimize learning.

Unique to XGBoost, additional parameters were adjusted to leverage its differences from Gradient Boosting. The *colsample_bytree* parameter, set to 0.734, determines the fraction of features sampled for each tree, promoting tree diversity. The *Gamma* parameter, tuned to 1,011, sets a minimum loss reduction required for further tree splitting, helping to regularize the model

by limiting unnecessary splits. The *Subsample* parameter, set to 0.927, specifies the fraction of data used for training each tree, reducing overfitting by introducing stochasticity.
After training, the *XGBoost* model achieved an accuracy of 0.839 and a sensitivity of 0.866, demonstrating its strength in effectively identifying positive cases while maintaining robust overall performance.

Bagging. For the *Bagging* model, hyperparameters were tuned to maximize *sensitivity*. The key parameter *BagSizePercent*, set to 54, defines the percentage of the training data used in each bootstrap sample, ensuring diversity among the base models while maintaining sufficient data for robust learning. After training, the *Bagging* model achieved an accuracy of 0.807 and a sensitivity of 0.832, showcasing its effectiveness in balancing generalization and accurate identification of positive cases.

Stacking. For the *Stacking* model, the three best-performing models—*XGBoost*, *Gradient Boosting*, and *Random Forest*—were used as base learners. The predicted probabilities from these models were utilized as independent variables for the meta-model, which in this case was a *logistic regression* model, to generate the final prediction. The optimized parameters for the base models were those previously identified through cross-validation, ensuring that each model contributed its best performance to the ensemble.
After training, the *Stacking* model achieved an accuracy of 0.840 and a sensitivity of 0.860, showcasing its effectiveness in balancing generalization and accurate identification of positive cases.

5 Classification Performance Evaluation

The goal of this phase is to evaluate the classification performance of the previously trained models. Evaluation metrics are introduced to enable a fair comparison of classifiers by addressing the threshold dependency issue. Metrics like **sensitivity** or **accuracy** cannot be directly compared across models, as they are calculated for a specific threshold value. To ensure robustness, evaluation measures must be independent of the threshold.

The primary method used is the **ROC curve analysis**. ROC curves illustrate how the **True Positive Rate (sensitivity)** changes relative to the **False Positive Rate (1-specificity)** across varying threshold values (ROC points). The steepness of the curve indicates the model's ability to correctly classify events while minimizing errors on non-events (1-specificity). In the initial part of the ROC curve, high thresholds correspond to the posterior probabilities, while lower thresholds are represented toward the latter part of the curve.

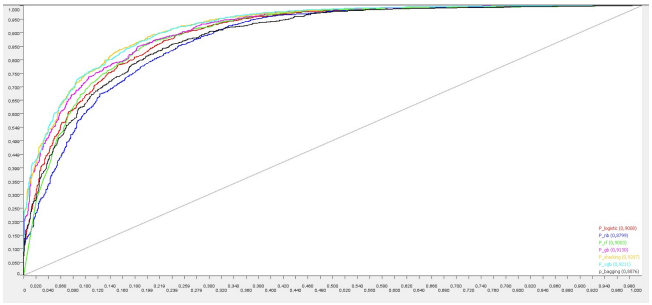

Fig. 2. Roc Curves

Model	Performance AUC
P_logistic	0.9008
P_nb	0.8799
P_rf	0.9003
P_gb	0.9130
P_stacking	0.9207
P_xgb	0.9211
P_bagging	0.6376

Table 2. Performance of different models.

The results, shown in both the table 2 and figure 2, demonstrate that ensemble methods, particularly **XGBoost** and **Stacking**, provided the best performance, outperforming standalone models such as **Logistic Regression**, **Random Forest**, and **Naive Bayes**. This underscores the power of combining models or using advanced boosting techniques for robust classification tasks.

Threshold Optimization for XGBoost. In this step, the classification performance of the best-performing model, **XGBoost**, was analyzed with a focus on determining an optimal threshold for application to new data. The threshold was specifically chosen to maximize the model's **accuracy**, while ensuring a balanced

trade-off with sensitivity, to correctly identify as many positive cases (`>50K` income) as possible, without excessively compromising specificity. This approach balances the trade-off between correctly classifying positive cases and minimizing false positives, aligning the model's performance with the primary objective of reducing critical misclassifications.

The Figure 3 illustrates the distribution of positive class probabilities, with the selected threshold marked by an orange line. This visualization illustrates how probabilities are distributed for the two classes (`>50K` and `<=50K`) and how the threshold separates them. By setting the threshold at 0.4256, the model balances sensitivity and specificity, aligning with the objective of minimizing false negatives while maintaining reasonable precision and overall accuracy.

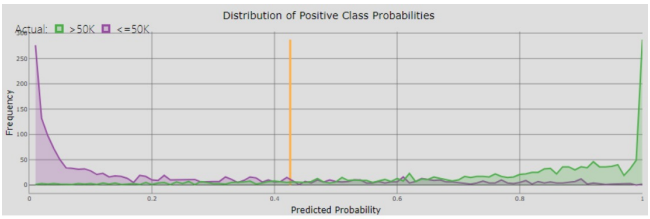

Fig. 3. Selection of the threshold: Determining the income level that separates individuals into meaningful categories for analysis

The analysis of the **XGBoost** model's classification performance at the chosen threshold of **0.4256** is presented in Table 3.

Metric	Value
Sensitivity	0.8981
Specificity	0.7816
Precision	0.8044
Accurancy	0.8395

Table 3. Performance Metrics of the Model

With the chosen threshold, the model achieved a high sensitivity of 0.898, an improvement over the baseline Gradient Boosting model without threshold optimization, which had a sensitivity of 0.86. This indicates the model's enhanced ability to correctly identify individuals earning more than $50K, a critical metric for the study's objective. While improving sensitivity, the specificity remains balanced at 0.782, ensuring a reasonable trade-off by minimizing false positives without significantly compromising the accuracy of predictions for individuals earning $50K or less.

6 Conclusion

The next step in this analysis would be to construct a predictive model using recent data from 2024, following the same methodology applied to the 1994 dataset. By replicating the process, it will be possible to ensure consistency and comparability between the two time periods. Leveraging the variable importance derived from these models will allow for identifying which factors have the most significant global impact on income disparities in 2024, and contrasting them with the key drivers identified for 1994. This comparison will offer valuable information on how socio-demographic and occupational income determinants have evolved over three decades, reflecting larger economic, technological, and social changes (5, 6).

To ground this analysis, the global feature importance from the 1994 dataset in Figure 4 provides a snapshot of the socio-economic dynamics of that time, highlighting the most influential factors in predicting whether an individual earned more than $50k annually. The feature importance was computed using the total gain from the XGBoost model, which sums up the gain across all splits where the feature is used, providing a comprehensive measure of each feature's contribution to the model's predictive performance.

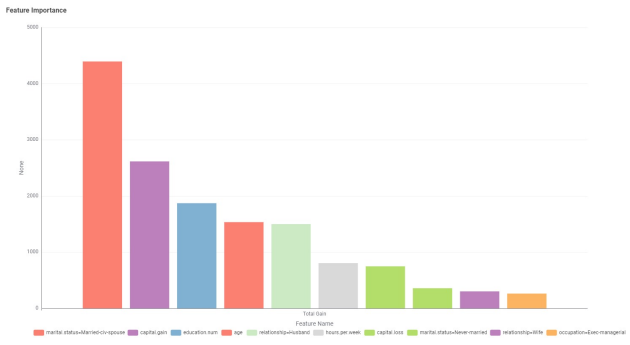

Fig. 4. Global Feature Importance for Predicting Income Levels

The results reveal the following key points

- **Marital status (Married-civ-spouse):** During the 1990s, being in a stable marital relationship often correlated with greater economic stability. This may be attributed to dual-income households

or traditional societal norms where married individuals were more likely to hold steady, higher-paying jobs.

- **Capital gain:** Investment income was a significant factor in distinguishing high earners in the 1990s. Those with substantial capital gains were typically part of the higher-income strata.
- **Education level (education.num):** Educational attainment played a crucial role, as more years of formal education often provided access to professional or managerial roles, which offered higher salaries.

These factors collectively reflect the economic and social priorities of the era, emphasizing the importance of education, stable relationships, and financial investments as drivers of financial success.

An additional intriguing avenue for exploration is the use of SHAP, a method that enables a local understanding of model predictions. SHAP provides insights into how individual variables contribute to a specific prediction by quantifying their influence on the posterior probability for a given instance (5). Applying SHAP to both the 1994 and 2024 models would allow for the analysis of income determinants at an individual level (Fig. 5)

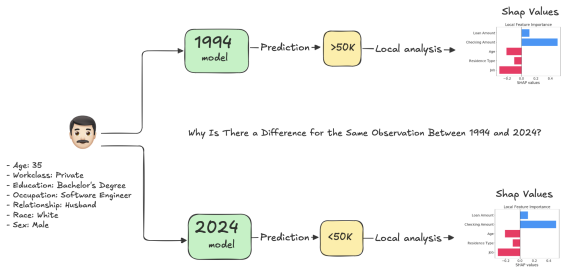

Fig. 5. Local analysis comparing SHAP values for the same individual using the 1994 and 2024 models. The differences highlight evolving feature importance and socio-economic dynamics over three decades.

This could uncover how the importance of specific variables varies across different time periods, thereby revealing nuanced differences in income dynamics. For instance, it could show how factors like education or occupational sector contribute differently to income levels in 1994 compared to 2024. This approach provides a comprehensive framework not only for understanding global trends but also for examining individual-level representations of income determinants, fostering a more granular perspective on the evolving nature of economic inequalities.

7 Bibliography

1. Thomas Piketty. Capital in the twenty-first century. Harvard University Press. 2014.
2. Raj Chetty, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. Where is the land of opportunity? the geography of intergenerational mobility in the united states. Quarterly Journal of Economics. 2014.
3. David H. Autor. Why are there still so many jobs? the history and future of workplace automation. Journal of Economic Perspectives. 2015.
4. James J. Heckman and Stefanie Mosso. The economics of human development and social mobility. Annual Review of Economics. 2014.
5. Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems. 2017.
6. Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics. 2007.