Model Overview: Random Forest trained on the Adult (Census Income) data set

Felipe S. Dosso, Betania E. R. da Silva, Décio M. Filho

October 3, 2024

1 Model details

This Random Forest model was developed by Décio Miranda, Betania da Silva and Felipe Dosso. It was trained on the Adult dataset to classify people's income as $> 50K$ or $\leq 50K$.

2 Intended use

The primary intended use of this model is to automate the pre-selection process for individuals who may be eligible for financial assistance or social benefits. The model aims to predict whether an individual's income exceeds $50,000 annually based on various demographic and socioeconomic attributes. This prediction can be used to:

- Accelerate the distribution of financial aid to those who need it most.
- Improve the efficiency of social benefit allocation processes.
- Provide a preliminary assessment of an individual's financial status for social welfare programs.

2.1 Primary intended users

The primary intended users of this model are:

- Government agencies and social service departments responsible for administering social benefits and financial assistance programs.
- Non-governmental organizations (NGOs) involved in poverty alleviation and social welfare initiatives.
- Researchers and policymakers studying income inequality and socioeconomic factors.
- Social workers and case managers need to quickly assess an individual's potential eligibility for various assistance programs.

3 Factors

3.1 Relevant factors

The Adult Census Income dataset contains various demographic and socioeconomic attributes relevant to the model's predictions. These factors include:

- Age
- Education level
- Occupation
- Sex
- Country of origin
- Race
- Marital status
- Work hours per week
- Income (target variable: above or below $50,000 annually)

These factors are crucial in determining an individual's socioeconomic profile and potential eligibility for social benefits.

- Sex: The dataset shows a significant imbalance between men (20,380) and women (9,782). This underrepresentation of women may lead to biased predictions.
- Race: Given the historical context of racial inequalities, evaluating the model's performance across different racial groups is essential to ensure fairness.
- Country of origin: As the dataset includes information on the country of origin, it's important to assess whether the model shows any bias against immigrants or specific nationalities.
- Age: The model should be evaluated across different age groups to ensure it does not discriminate against older or younger individuals.
- Education level: The model's performance should be assessed across different education levels to avoid perpetuating educational disparities.

These evaluation factors are significant given the model's intended use in social benefit allocation, where fairness and equality are paramount. Continuous monitoring and adjustment may be necessary to mitigate biases and ensure the model contributes to social well-being rather than perpetuating existing inequalities.

4 Metrics

4.1 Model performance measures

Three performance metrics were used to evaluate the model. The first one is **roc-AUC-score** and measures the model's ability to distinguish between classes. A higher AUC indicates better performance in classifying positive and negative cases and is very efficient in obtaining unbalanced data. Another metric that was used is the **accuracy**. It is the proportion of correctly predicted instances out of the total predictions. The last metric, the **Brier Score**, evaluates the accuracy of the model's predicted probabilities. A lower Brier score indicates the predicted probabilities are closer to the actual outcomes, meaning better probability calibration.

The model performed fairly well in these metrics as seen below:

Metric	**Value**
AUC Score	0.9118
Accuracy	0.8340
Brier Score	0.1002

4.2 Fairness Performance

Two metrics were used to assess fairness in the model, **Statistical Parity Difference** and **Disparate Impact**. The first measures the difference in the probability of favorable outcomes between privileged and unprivileged groups, and the last measures the ratio of favorable outcomes between unprivileged and privileged groups. The model's results across these metrics is shown in the next table

	Without model		Random Forest	
Metric	SPD	DI	SPD	DI
Gender	0.19	0.36	0.09	0.30
Race	0.10	0.60	0.06	0.44
Native country	0.05	0.76	0.05	0.47

Table 1: Comparison between baseline model and random forest

The table 1 shows that the disparities between the sensitive features decreased significantly, but biases are still present.

5 Evaluation and Training Data

Both the Evaluation and Training data were taken from the *Adult* dataset directly from the UCI-Machine Learning Repository. The raw data was split into

3 parts, around 50% training, 33% for testing and 10% for validating the model. After that, missing values like and spaces in column names were removed. Furthermore, samples containing the "?" were excluded. Then, a one-hot encoding for the categorical features that would not be part of the sensitive group was performed (i.e., features considered sensitive, such as race, gender, and country of origin, were directly categorized in a binary form).

A lot of features are imbalanced in the data. In Income, for instance, 24.89% of people have an income greater than 50 thousand dollars, while 75.11% have an income less than or equal to 50 thousand dollars. Race remains, with 14.02% of people being non-white and 85.98% being white. For sex/gender, 32.43% are women and 67.57% are men. Additionally, for the country of origin, only 8.81% of people are non-native, while 91.19% are natives (born in the USA).

6 Quantitative analyses

The model's performance was disaggregated by sensitive factors, including race, gender, and country of origin. Fairness was evaluated using metrics such as Statistical Parity Difference (SPD) and Disparate Impact (DI) to identify potential biases. The analysis revealed variations in these fairness metrics across demographic groups, with notable biases persisting in race and gender, highlighting areas where the model may disproportionately affect certain populations.

7 Ethical considerations

The Random Forest model trained on the Adult dataset raises critical ethical concerns, particularly regarding fairness and bias. Sensitive data such as race, gender, and country of origin were used. Although fairness metrics like Statistical Parity Difference (SPD) and Disparate Impact (DI) were applied, disparities persist, especially for race and gender. This highlights the risk of perpetuating social inequalities in automated income predictions. While hyperparameter tuning and calibration improved performance, there remains a need for deeper fairness evaluations to mitigate potential harms, such as discriminatory outcomes for marginalized groups.