Income Classification Using U.S. Census Data

Jake Roll

rolljake@msu.edu

https://github.com/JakeRoll04/cmse492_project November 2, 2025

Abstract

This project aims to predict whether an individual earns more than $50,000 annually using demographic and occupational data from the U.S. Census Bureau's Adult Income dataset. By applying modern machine learning methods, the study seeks to identify key socioeconomic features influencing income levels. The analysis begins with data cleaning, exploratory visualization, and feature correlation analysis, followed by the development of baseline and advanced classification models. a baseline Logistic regression has been completed, while random forest and gradient boosting will be used in the future to compare in terms of predictive accuracy and interpretability. Model performance is evaluated using accuracy, precision, recall, and F1-score. The results will provide insight into demographic and economic disparities while serving as a case study for data preprocessing, model tuning, and evaluation in applied ML workflows.

1. Background and Motivation

Income inequality remains a central topic in social and economic research. Predicting income brackets from census data highlights structural relationships between education, occupation, and demographic variables. Prior studies on the UCI Adult dataset show that while linear models achieve moderate accuracy, ensemble methods often perform better by capturing nonlinear dependencies. This project combines statistical rigor with interpretability to showcase responsible machine learning.

2. Data Description

The dataset originates from the UCI Machine Learning Repository (Census Income). It includes 48,842 records and 14 attributes such as age, education, occupation, hours-per-week, and marital status, with a binary target variable indicating income class (`<=50K`, `>50K`). Categorical variables contain missing entries labeled "?" and are cleaned through removal and encoding. Continuous features like age and hours-per-week are standardized, and categorical fields are one-hot encoded before model training.

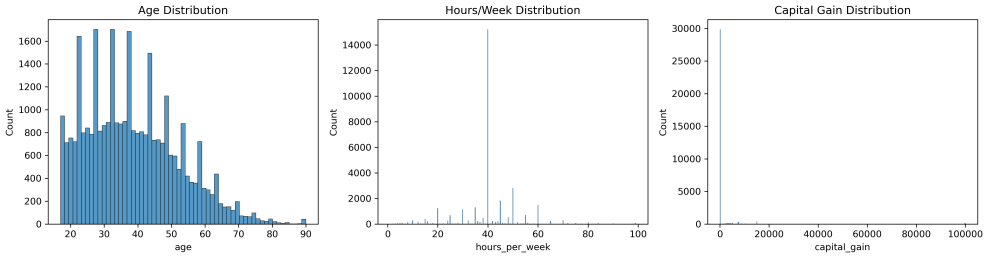

Figure 1: Distributions of key numeric features (e.g., age, hours-per-week, capital gain).

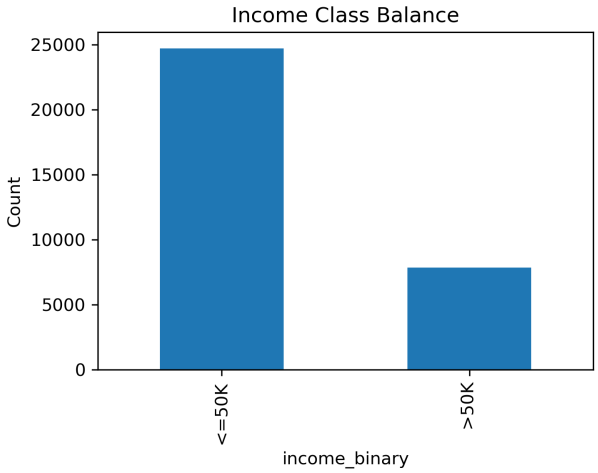

Figure 2: Class balance for the binary income target (`<=50K` vs `>50K`).

Exploratory analysis visualizes (1) feature distributions (Figure 1) and (2) class imbalance (Figure 2). These plots summarize dataset variability and highlight mild class imbalance between income groups.

3. Proposed Methodology

The study follows a supervised classification pipeline:

1. **Baseline:** Logistic Regression for interpretability.
2. **Intermediate:** Random Forest to capture nonlinear relationships.
3. **Advanced:** Gradient Boosting (XGBoost) for optimized accuracy.

Each model is trained using stratified train/validation/test splits. Hyperparameter tuning uses 5-fold cross-validation. Feature importance and confusion matrices are used to interpret performance.

4. Evaluation Framework

Performance metrics include accuracy, precision, recall, F1-score, and ROC-AUC. Baseline accuracy is defined by the majority-class predictor. Success is defined as achieving at least a 10% improvement over the baseline while maintaining interpretability. Cross-validation ensures model robustness. Results are visualized with precision–recall curves and feature importance plots.

5. Timeline and Milestones

The project timeline spans from early November 2025 through the final deadline on December 8, 2025. Major activities are structured week-by-week with explicit milestones, but some tasks run in parallel to save time and to reflect the fact that modeling, analysis, and documentation overlap.

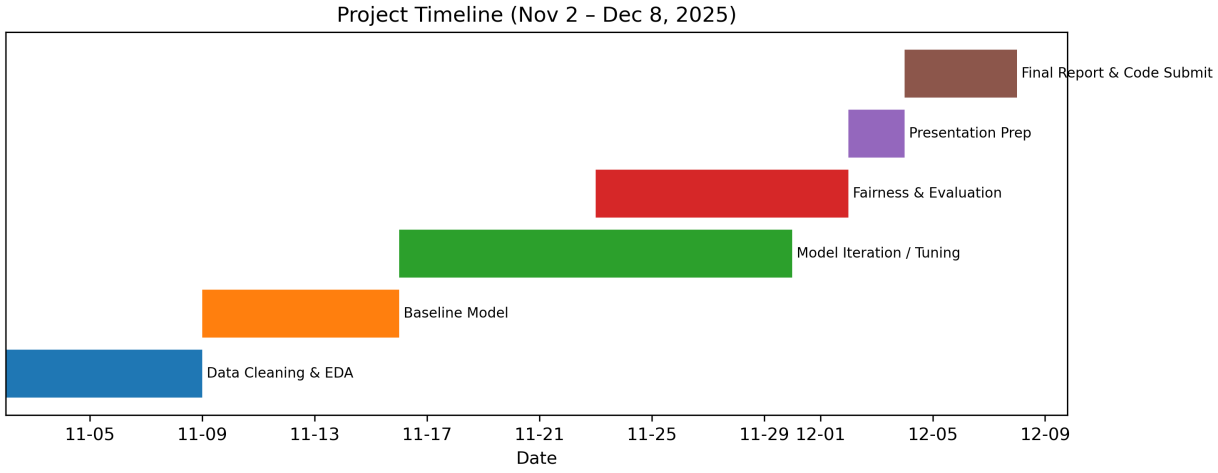

Figure 3: Week-by-week project plan showing task dependencies, overlapping work, and final deliverables. Presentation practice is scheduled for Dec 2–4 (Week 15), and the final report and code are due Dec 8.

Weeks 1–2 (Nov 2 – Nov 16): Data cleaning, exploratory data analysis (EDA), and documentation of the dataset. In parallel, I build and evaluate the baseline model (logistic regression) to

establish reference metrics. These steps produce the first deliverables: processed data, EDA figures, and baseline performance numbers.

Weeks 3–4 (Nov 16 – Nov 30): I begin model development beyond the baseline. Random Forest and Gradient Boosting models are trained, tuned, and compared. This overlaps with fairness and evaluation work: I assess how well these models perform on the minority income class (`>50K`), and I generate metrics like precision, recall, and F1. The dependency here is that I cannot tune or compare advanced models until the cleaned data and baseline are stable, but I *can* iterate on model evaluation and figure generation in parallel while tuning hyperparameters.

Week 5 / Week 15 of course (Dec 2 – Dec 4): Presentation prep and rehearsal. At this point, plots, tables, and key findings should already exist. The focus here is communicating the problem motivation, describing the modeling pipeline clearly, and explaining limitations (class imbalance, bias across demographic features). This is a hard checkpoint: I should be able to talk through method, data, and results by Dec 2–4.

Final Deliverables (Dec 4 – Dec 8): I finalize the written report and push all code, processed data samples, figures, and evaluation outputs to GitHub. The last block of time (Dec 4–Dec 8) is deliberate buffer: if a model underperforms, I can fall back to the validated logistic regression baseline and Random Forest results. If there are issues with data quality or missing values, I can regenerate the processed dataset and update the report. The final PDF report and the repository are submitted by December 8.

This schedule builds in contingency time. The critical path is: (1) produce a clean dataset and baseline results, (2) train and compare advanced models, (3) generate evaluation figures, and (4) assemble the final narrative. Tasks like figure polishing, fairness discussion, and writing can proceed in parallel with model tuning, which protects the December 8 final submission and the December 2–4 presentation window.