Predicting Income Levels Using U.S. Census Data

Agenda

- Project Overview
- Understanding the Data
- Data Preparation & Quality
- Modeling & Results
- Insights & Client Impact
- Recommendations & Next Steps

Project Overview

- The project centers on predicting income levels using U.S. Census data.
- A classification approach is applied to distinguish individuals earning above and below $50K annually.
- The analysis is structured in **Dataiku**, integrating preparation, modeling, and validation within a single workflow.
- The focus is on interpretability and reliability—building results that can be trusted and scaled.

Understanding the Data

Overview

- ~300K records from the U.S. Census Bureau (1994–1995).
- Data includes **demographic**, **socio-economic**, and **employment** information per individual.
- Target variable: **Income level** (`>50K` vs `≤50K` annual income).

Key Attributes Used

- **Demographic:** Age, Sex, Marital Status, Education
- **Employment:** Occupation, Industry, Class of Worker, Work Hours,
- **Financial:** Capital Gains, Losses, Dividends, Federal Tax Liability, Weeks Worked

Highlights

- Data shows strong **class imbalance** (≈6% earn >50K).
- Income-related attributes align strongly with **class of worker and weeks worked in year**.

Data Preparation & Quality

- Conducted a full quality audit across all variables — dataset was complete but contained **redundant and low-value columns**.
- **Removed non-informative features** (≥ 99 % identical values, duplicate encodings, or administrative fields) to enhance model clarity.
- **Engineered new, interpretable features** to capture economic behavior and life stage:
  - **Has Assets:** presence of capital gains, losses, or dividends
  - **Age Group:** categorized by life stage (Child → Retired)
  - **Work Status:** whole-year, partial-year, or non-worker

→ **Result:** a clean, de-duplicated, and insight-rich dataset, ready for reliable modeling and business interpretation.

Modeling & Results

- Conducted **multiple modeling experiments** using different data designs and preparation levels — from raw, unprocessed data to the final polished, model-ready dataset.
- Explored several **algorithmic approaches** to assess stability and predictive power:
  - **Logistic Regression** – as a baseline, interpretable benchmark
  - **Decision Tree** – for rule-based explainability
  - **Random Forest** – for ensemble learning and feature robustness
  - **XGBoost** – for optimized gradient boosting and handling data imbalance
- Tested **different feature configurations**, iteratively refining the dataset to evaluate the impact of cleaning, encoding, and engineered variables.
- Applied **class weighting** to address target imbalance and ensure fair model comparison.

→ **Result:** Progressive data refinement and model experimentation led to a more stable, interpretable, and higher-performing model suitable for actionable insights.