# U.S. Census Data
# Income Prediction & Socioeconomic Insights Analysis
*04/01/2025*

# Problem Statement

| | | |
|---|---|---|
| 01 | Objective | • Understand trends in demographic & economic features<br>• Classify whether an individual earns more or less than $50,000 per annum |
| 02 | Data Domain | • Income distribution of U.S. residents between 1994-95<br>• Roughly 300,000 raw records of individuals<br>• 40 descriptors of data (job codes, education, age, hours worked etc.) |
| 03 | Key Challenges | • Handling large-scale, anonymized data<br>• Managing duplicates, missing values & conflicting records<br>• Building interpretable & accurate models |

# Data Wrangling

## Removed Features

| Feature | Missing |
|---|---|
| migration code-change in msa | 49% |
| migration code-change in reg | 49% |
| migration code-move within reg | 49% |
| migration prev res in sunbelt | 49% |

## Imputed Features (Using KNN)

| Feature | Missing |
|---|---|
| country of birth father | 4.1% |
| country of birth mother | 3.7% |
| country of birth self | 2.1% |
| hispanic origin | 0.5% |
| state of previous residence | 0.4% |

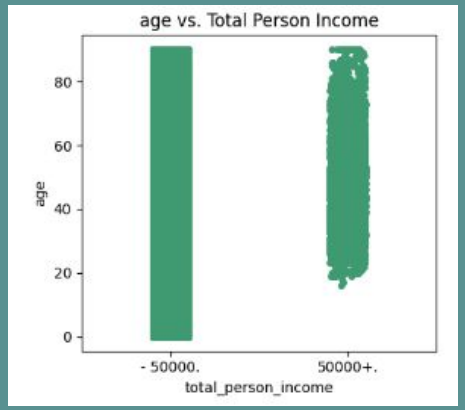
## Deduplication

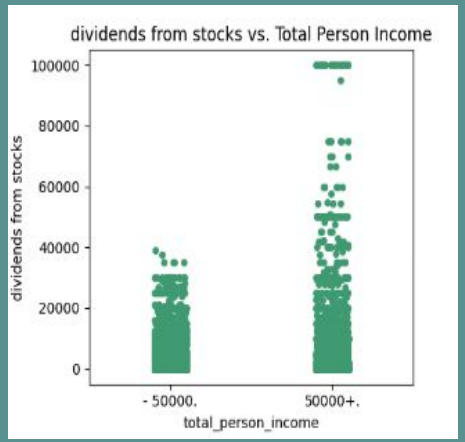
| Issue | Learn % | Test % |
|---|---|---|
| Duplicates (Dropped) | 23.4% | 20.9% |
| Target/Label Conflicts (Resolved via Weighted Majority Vote) | 0.0004% | 0.0003% |

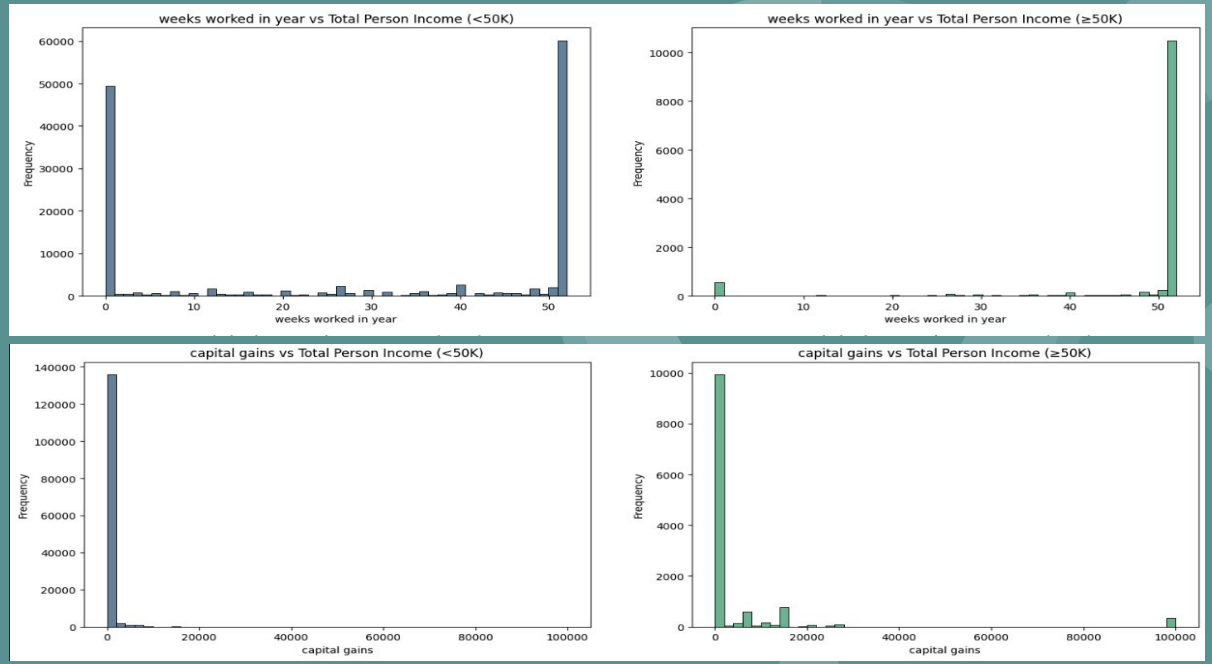
23% Total Reduction in Learn (199.5k to 152.8k)

21% Total Reduction in Test (99.7k to 78.8k)

# Exploratory Data Analysis: Continuous Features



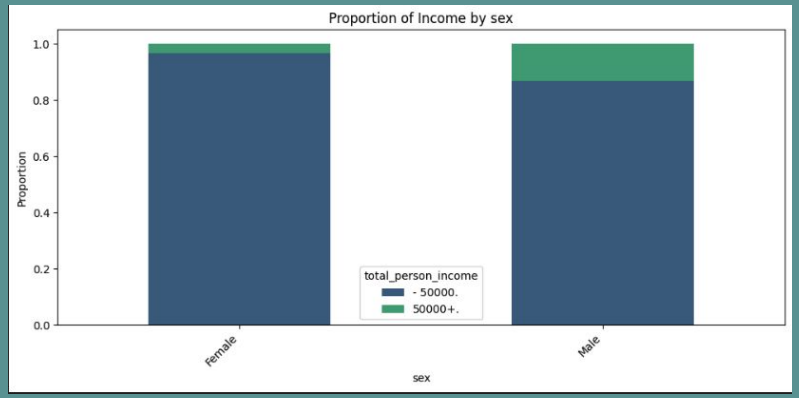



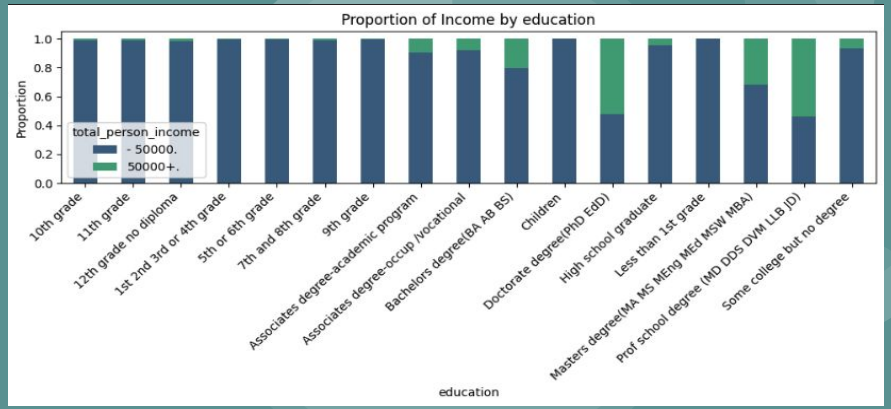



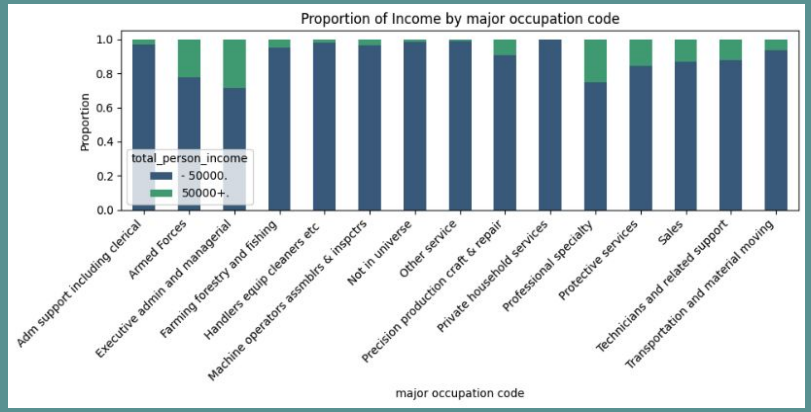

## Key Insights:

- Most people don't have access to capital gains whilst those who do are strongly associated with higher income possibly due to investment activities
- People under the age bracket of 18-20 are solely in lower income class implying no or low income (through part time work) as they are kids
- Most people who earn more than $50,000 tend to work for 52 weeks of the year
- Dividends from stocks over $40,000 are mainly received by people earning over $50,000

# Exploratory Data Analysis: Nominal Features



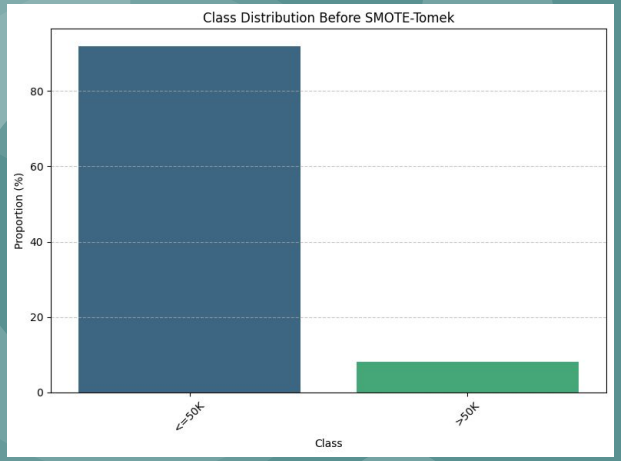



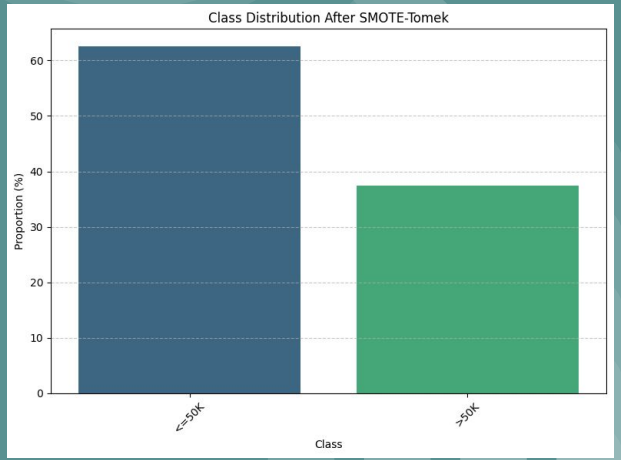

## Key Insights:

- Higher education levels like Bachelors, Doctorate, Masters and Prof school degrees show higher proportion of records with income above $50,000
- Clerical, managerial and professional specialties have higher proportion of incomes above 50,000, while lower income proportions are observed for farming, cleaning and labor-intensive roles
- Despite females outnumbering the males in the census study, males have a higher proportion of income above $50,000 highlighting gender income disparity
- Householders are dominant in the income class earning more than $50,000 whilst other relatives are not in the class
- Self employed people with their own businesses and people working for the government dominate the income class earning more than $50,000 as well

# Preprocessing

- New features:
  - is_full_time: If weeks worked in a year > 40
  - has_capital_gains: If capital_gains > 0
  - income potential: age x weeks worked x wage per hour
  - wage_per_age: Wage per hour / age
- Nominal features with high cardinalities grouped & target encoded
- Nominal features with low cardinalities label encoded
- Highly skewed continuous features like capital gains, losses, dividends from stocks log transformed
- All continuous features standard scaled
- Shortlisted features based on Cramer's V for nominal & Mann-Whitney for continuous features
- Class imbalance addressed via hybrid approach combining:
  - Synthetic Minority Oversampling Technique: Upto 50% of majority class to maintain generalization capability of models
  - Tomek Links (undersamples majority class)



# Modelling

## Logistic Regression

| Class | Precision | Recall | F1 Score |
|---|---|---|---|
| <50k | 98% | 92% | 95% |
| >50k | 44% | 75% | 55% |
| **AVG** | **71%** | **83%** | **75%** |

## XGBoost

| Class | Precision | Recall | F1 Score |
|---|---|---|---|
| <50k | 96% | 98% | 97% |
| >50k | 66% | 56% | 61% |
| **AVG** | **81%** | **77%** | **79%** |

## LightGBM

| Class | Precision | Recall | F1 Score |
|---|---|---|---|
| <50k | 96% | 98% | 97% |
| >50k | 73% | 51% | 60% |
| **AVG** | **84%** | **75%** | **79%** |

## Ranking

1. **LightGBM**
   - Avg F1-Score: 79%
   - ROC-AUC: 94.3%
2. **XGBoost**
   - Avg F1-Score: 79%
   - ROC-AUC: 94.03%
3. **Logistic Regression**
   - Avg F1-Score: 75%
   - ROC-AUC: 93.07%

## Top Features -LightGBM

- Education
- Detailed Occupation Recode
- Detailed Industry Recode
- Age
- Major Occupation Code
- Age x Persons Worked for Employer
- Age x Weeks Worked in Year
- Major Industry Code

# Potential Improvements

- **Enhanced Hyperparameter Tuning** – Explore a broader range of parameters to optimize model performance.
- **Feature Selection with Mutual Information Scores** – Identify and eliminate noisy variables.
- **Interaction Features for Categorical Variables** – Create new features to capture relationships between categories.
- **Advanced Binning Techniques** – Apply more granular binning for continuous features to improve data representation.
- **Exploring Neural Networks** – Test models like **Multi-Layer Perceptron (MLP)** for capturing complex patterns.
- **Outlier Detection and Removal** – Refine preprocessing by identifying and removing outliers to improve model robustness.