Uncovering Determinants of Income in the United States

Building a classification model using US Census archive data

06/09/2023

Context

- Every ten years, a census is conducted to collect and organize information regarding the US population to effectively allocate billions of dollars of funding to various endeavors.
- We were tasked with utilizing data analysis and machine learning to highlight the factors that influence income levels. Our objective was to build a machine learning model capable of classifying individuals into 'low income' (less than 5000) and 'high income' (greater than 5000) categories.
- This mission spanned over 7 days and included the following deliverables:
  - A notebook and the associated environment, available on GitHub
  - This presentation deck

Data

- We have 3 datasets :
  - **Training Data**: `census_income_learn.csv` - This dataset was used to train our machine learning model.
  - **Test Data**: `census_income_test.csv` - Was used to evaluate the model's performance.
  - **Metadata**: `census_income_metadata.csv` - This file contains basic information about variables, we used it to retrieve column details.

- Training data
  - 199 522 rows and 42 columns
  - 34 categorical & 8 numerical
  - High-income (>$5000/year): 6.2% of dataset

Types of variables present in the dataset

- Demographic Information
- Income and Financial Information
- Geographic and Migration Data
- Veteran and Military Status
- Household Information
- Race and Ethnicity

Methodology and Selected Variables

Methodology

Data Importation

Handling Missing Values

Correlation Analysis

Variable Selection Based on Correlations

Variable Recategorization

Variable Selection Using L1 Regression

ElasticNet Logistic Regression

- F1 Score
- Class Weight 'Balanced'
- L1 and L2 Penalization

GridSearchCV for Parameter Optimization

Optimizing F1 with Best Threshold

Influential Factors Analysis

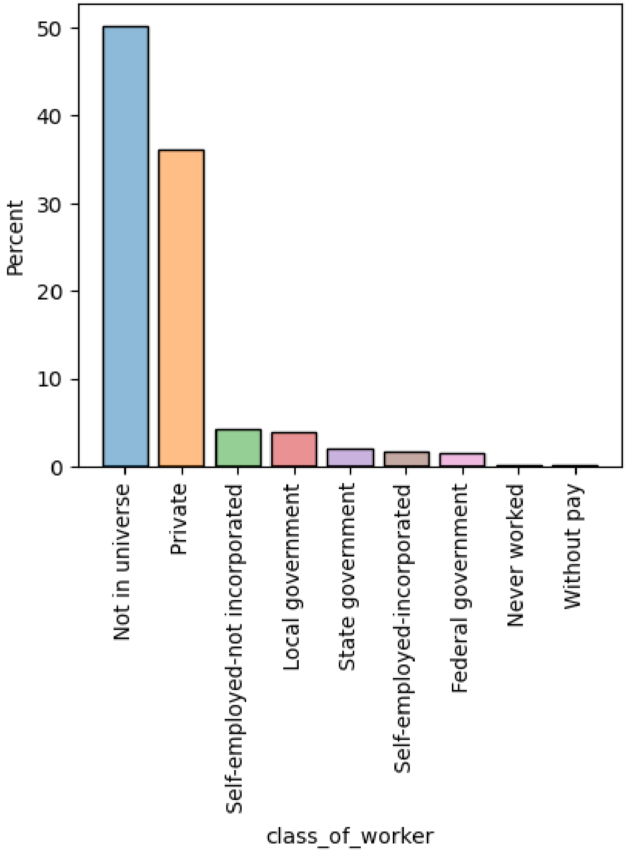
Selected variables

	Variable name	**Variable type**
1	class_of_worker	Categorical
2	education	Categorical
3	marital_stat	Categorical
4	major_industry_code	Categorical
5	major_occupation_code	Categorical
6	race	Categorical
7	sex	Categorical
8	member_of_a_labor_union	Categorical
9	full_or_part_time_employment_stat	Categorical
10	tax_filer_stat	Categorical
11	detailed_household_summary_in_household	Categorical
12	num_persons_worked_for_employer	Categorical
13	family_members_under_18	Categorical
14	own_business_or_self_employed	Categorical
15	veterans_benefits	Categorical
16	age	Categorical
17	weeks_worked_in_year	Categorical
18	investment_feature	Categorical

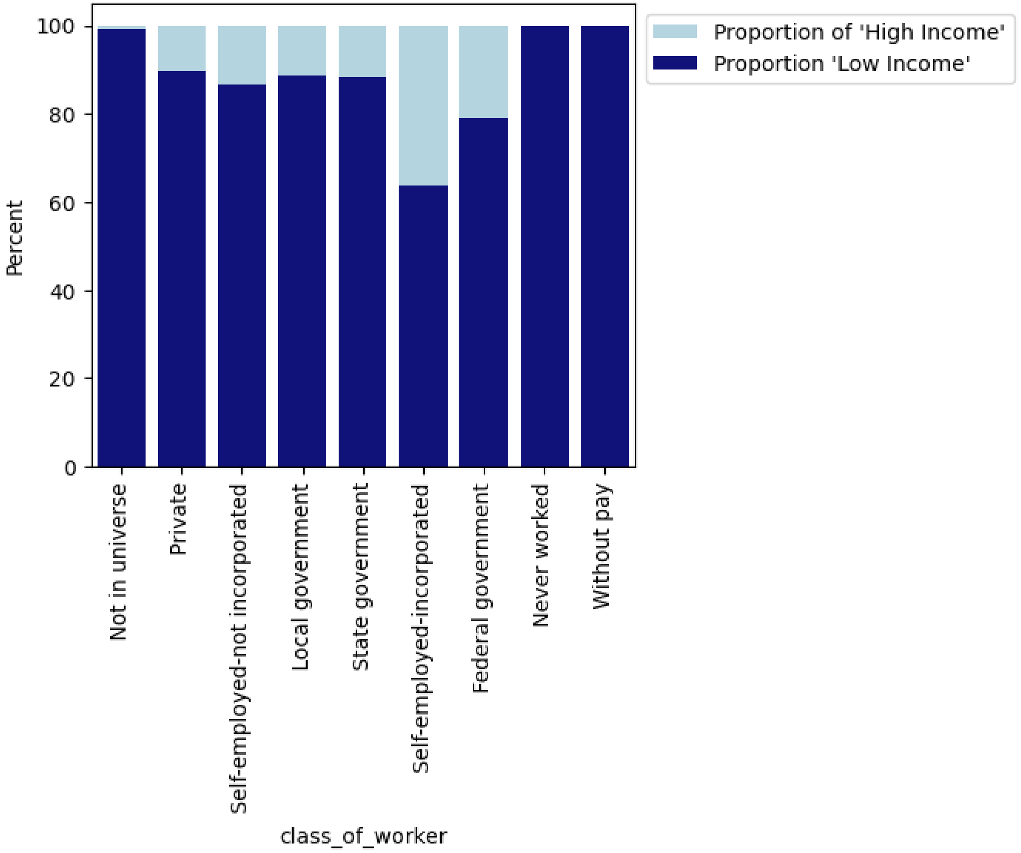
Graphical analysis

Class of worker

Distribution of class_of_worker Categories



Income Distribution by Category for class_of_worker

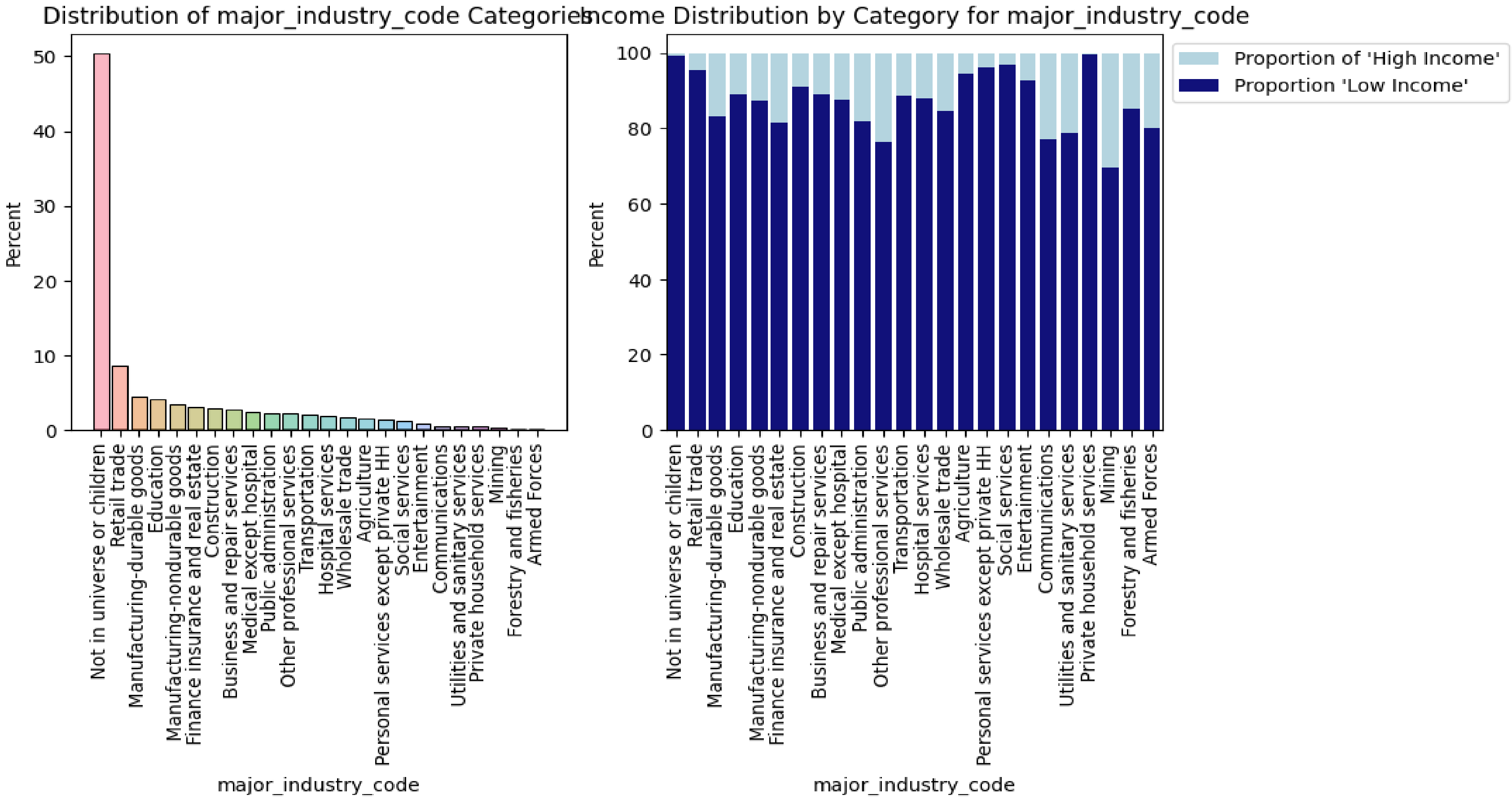


Key insights

- State government, local government, and private employment sectors have a relatively reasonable distribution of high and low-income individuals.
- Categories like "not in universe," "never worked," and "without pay" predominantly consist of individuals with low income, which is logically expected.
- The federal government employment sector shows slightly higher income levels compared to the local and state government sectors.
- Self-employed individuals with incorporated businesses tend to have higher incomes.

Graphical analysis

Industry

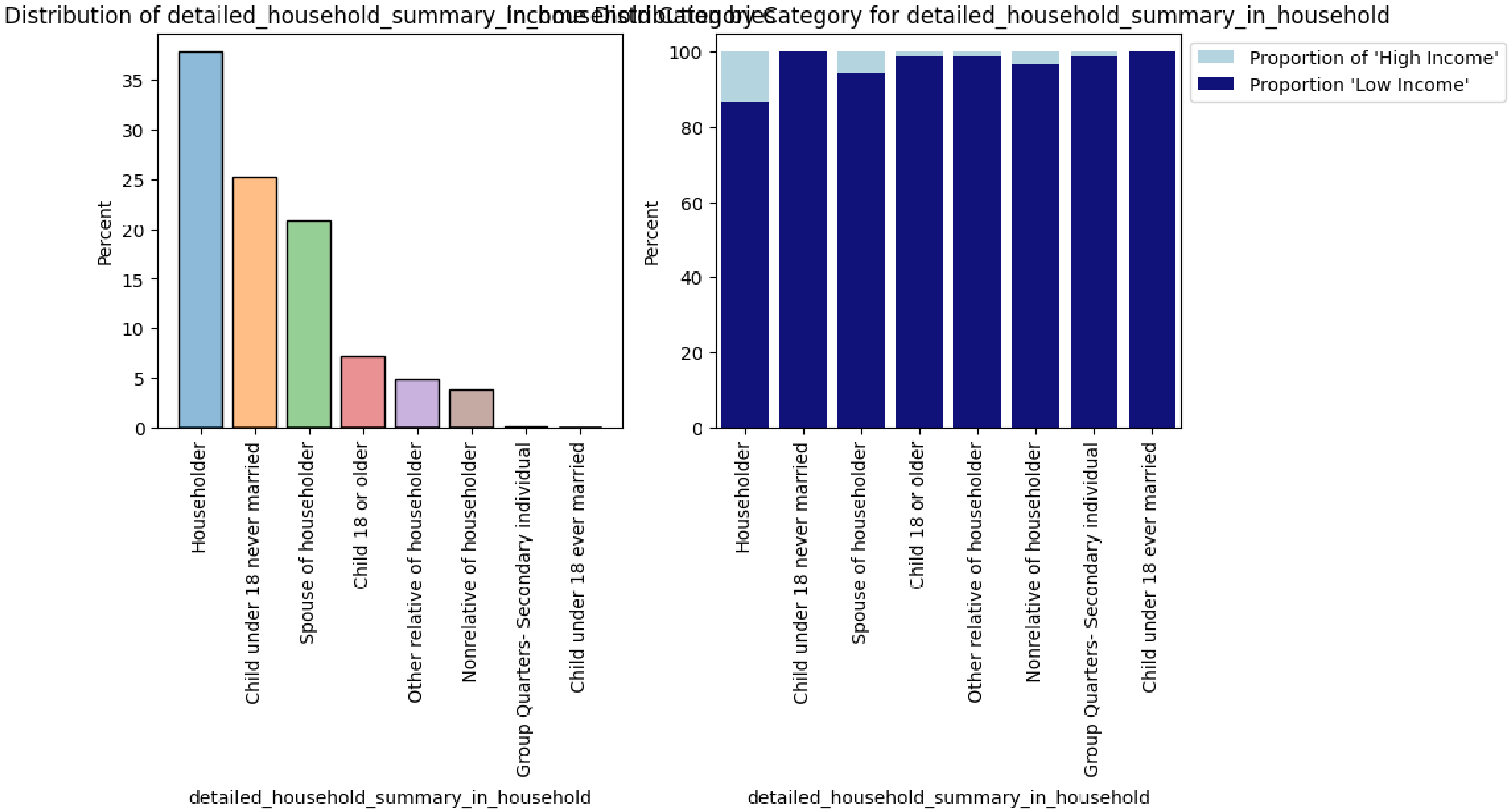


Key insights

- The variable contains various categories, with "Not in universe or children" being the most common category, accounting for approximately 99.06% of the samples.
- The sectors with the highest representation include Retail Trade, Manufacturing of durable goods, and Education.
- Categories such as "Mining" (69.73%), "Communications" (77.09%), and "Other professional services" (76.41%) have relatively lower percentages of low-income individuals.
- Categories like "Private household services" (99.45%), "Personal services except private HH" (96.29%), and "Social services" (96.69%) have higher percentages of low-income individuals.
- "Private household services" has the highest percentage of low-income individuals (99.45%), indicating that a significant proportion of individuals in this category have low income.

Graphical analysis

Household

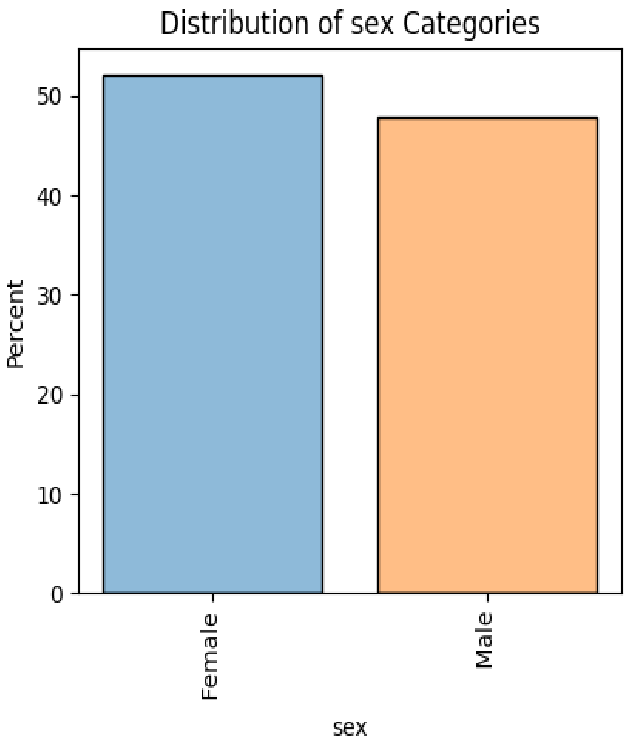


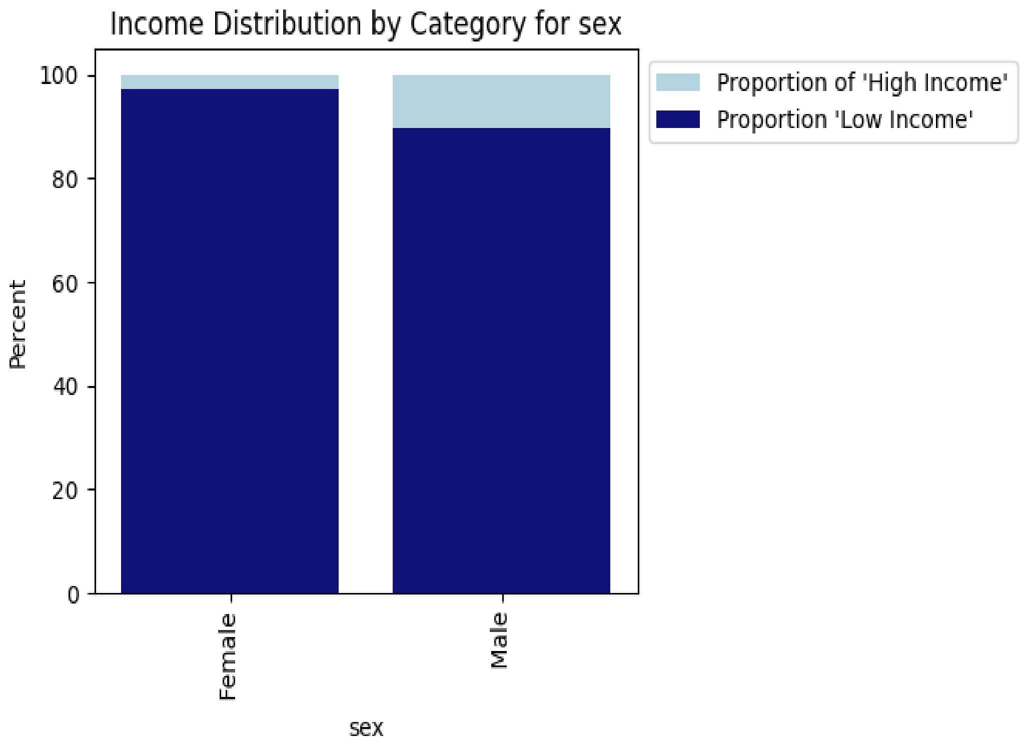
Key insights

- The most common household relationship category is "Householder," wich represent more than 35 % of the individuals.
- The second most prevalent category is "Child under 18 never married," with around 25% of the individuals.
- Householders have the highest proportion of high-income individuals, followed by spouses of householders.
- Non-relatives have slightly lower incomes, while children and other relatives of householders have the lowest incomes.

Graphical analysis

Sex



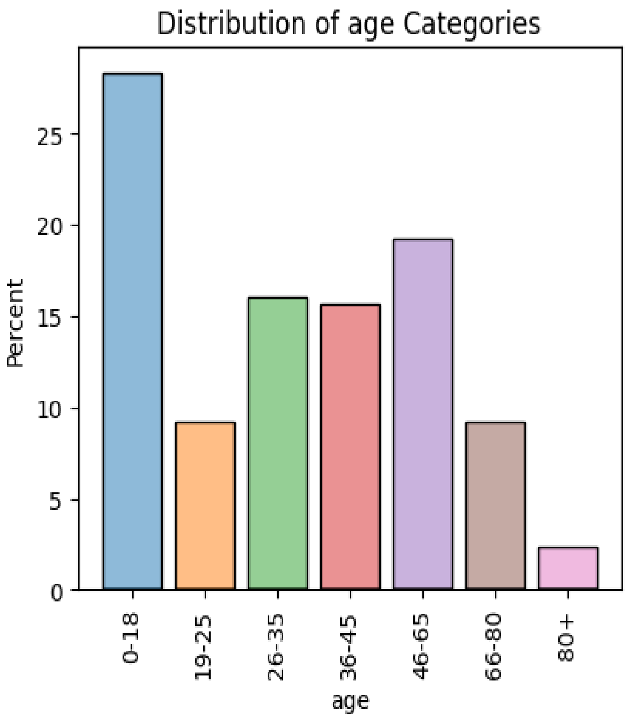


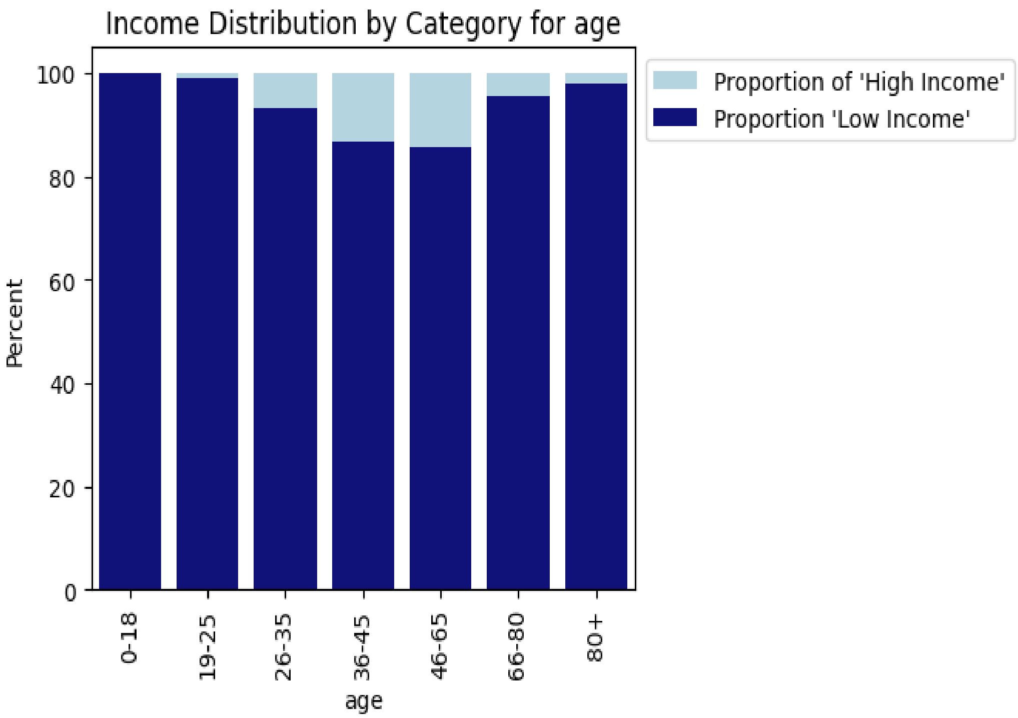
Key insights

- The analysis reveals notable income disparities between the two gender categories.
- Among women, a significant percentage (approximately 97.33%) fall into the "Low Income" category, indicating that a higher proportion of women in the dataset have lower incomes.
- In contrast, among men, a lower percentage (approximately 89.68%) are classified as "Low Income," suggesting that a relatively smaller proportion of men have lower incomes.

Graphical analysis

Age





Key insights

- Individuals within the 0-25 age range tend to have lower incomes
- While those in the 26-65 age group generally enjoy higher incomes
- Beyond the age of 65, income levels tend to decline

Results (models)

Classification report

	Precision	Recall	F1-Score	Support
< $50000/year	0.98	0.96	0.97	93575
>$50000/year	0.53	0.63	0.57	6186
Accuracy			0.94	99761
Macro Avg	0.75	0.80	0.77	99761
Weighted Avg	0.95	0.94	0.94	99761

Confusion matrix

		ACTUAL VALUES	
		< $50000/year	>$50000/year
PREDICTED VALUES	>$50000/year	90112	3463
	>$50000/year	2296	3890

Based on these results, the following observations can be made:

- The model performs well in terms of precision and recall for class 0 (<= $ 50,000 income), with high values for both metrics.

- However, the model's performance is less balanced for class 1 (> $ 50,000 income). While precision is moderate, recall is relatively low, indicating that the model tends to miss a significant portion of individuals with high income.

- The F1-score for class 1 reflects this trade-off between precision and recall, resulting in a lower F1-score.

- The model's overall accuracy is high, but it's important to consider the class imbalance. The high accuracy is primarily driven by the correct predictions for class 0, which is the majority class.

Results & Recommandations

Several variables appear to have a significant impact on income levels. Below are the most influential variables found by our model and associated recommendations:

1. **Women**:
    - **Recommendation**: Promote gender equality in the workforce and ensure fair wages for women.
    - **Public Endeavors**: Implement policies that close the gender pay gap, support work-life balance, and encourage women's participation in higher-paying industries and leadership roles.

2. **Age Category 19-25**:
    - **Recommendation**: Invest in education and job opportunities for young adults.
    - **Public Endeavors**: Create apprenticeship programs, provide affordable access to higher education, and offer job placement services for this age group.

3. **Social Services field**:
    - **Recommendation**: Enhance career prospects and wages within the social services sector.
    - **Public Endeavors**: Increase funding for social services, provide professional development opportunities, and advocate for fair wages and benefits for workers in this field.

Results & Recommandations

Several variables appear to have a significant impact on income levels. Below are the most influential variables found by our model and associated recommendations:

4. **Unemployed or Part-Time Workers**:
   - **Recommendation**: Promote full-time employment and job stability.
   - **Public Endeavors**: Implement policies that incentivize businesses to offer full-time positions with benefits, provide unemployment assistance, and create job training programs for unemployed or part-time workers.

5. **Education industry**:
   - **Recommendation**: Strengthen the education sector and enhance job opportunities within it.
   - **Public Endeavors**: Invest in education infrastructure, offer competitive salaries and benefits for educators, and provide resources for students in underserved communities.

6. **Class of Worker - State Government**:
   - **Recommendation**: Improve compensation and career advancement opportunities for state government employees.
   - **Public Endeavors**: Review and adjust salary structures, offer professional development programs, and ensure that state government positions provide competitive compensation packages.

Next steps

Business check

- Review the results from your perspective
- Engage your business experts and data team.
- Confirm and validate the analysis of the project.

Modifications

- Plan a meeting for reviewing the code and the results together
- On our side, apply modifications to fix potential issues

Automate your analyses

- Our data expert can assist you in conducting additional analyses, automating them and build dashboard for you needs.