Census Income Features Analysis

Agenda

Introduction/Background
Problem Statement & Objective
Executive Summary
Approach Summary
Data Exploration, Cleansing, Preparation, and Features Engineering
Data Modeling
Model Evaluation
Model Interpretation & Findings
Potential Next Steps and Enhancement

Introduction/Background

The U.S. Census Bureau collects economic and demographic data to support federal policy and funding decisions

This analysis explores how demographic and employment-related factors influence income levels

Problem

Identify characteristics associated with a person making more or less than $50,000/year

Objective

Build an explainable model pipeline using EDA, data processing, modeling, and result interpretation to understand what feature affect income

Executive Summary

- Cleaned and processed ~300,000 census records, this include handling missing values and encoding values.
- Investigated the correlation between all features and income class
- Built Logistic Regression, Random Forest, and XGBoost models.
- Used class weights to address imbalance,
- Random forest selected (F1 = 88%, AUC = 0.93).
- Key features reveled: Education, Capital Gains, Marital Status, Major, Industry, nd Weeks Worked.

Approach

Data Exploration, Cleansing, Preparation, and Features Engineering

- What is the data?
- What are the type of features (numerical or categorical)?
- Are their missing values?
- How are features correlated?

Data Modeling

What algorithms to use and fit our data? And why?

- How to tune the model parameter and control the training?
- How to address imbalance issue

Models Evaluation

- How does the different models behave?
- How to evaluate the performance of the models?
- What is the best model?

Model Interpretation & Findings

- What are the features that affects our models the most?
- What are the characteristics of the 50k> income ?

Data Exploration, Cleansing, Preparation, and Features Engineering

Data Exploration
- List all features and their unique values
- Identify Classes distribution
- Check and Adjust Features Type

Data Cleansing
- Check & remove duplicate rows
- Identify missing values
- Handle missing values

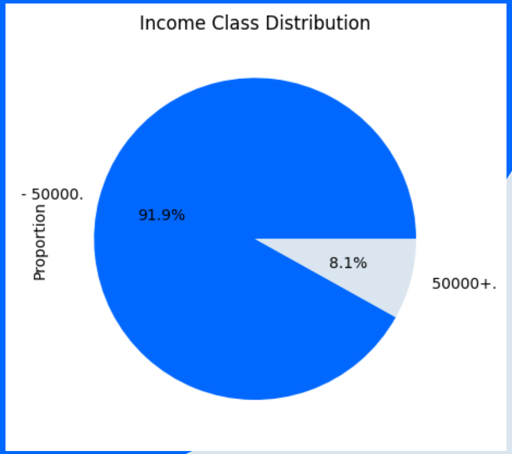
Features Engineering
- Encode categorical data
- Add new features

Data Visualization
- Visualize numerical features distribution
- Explore features correlation to income class

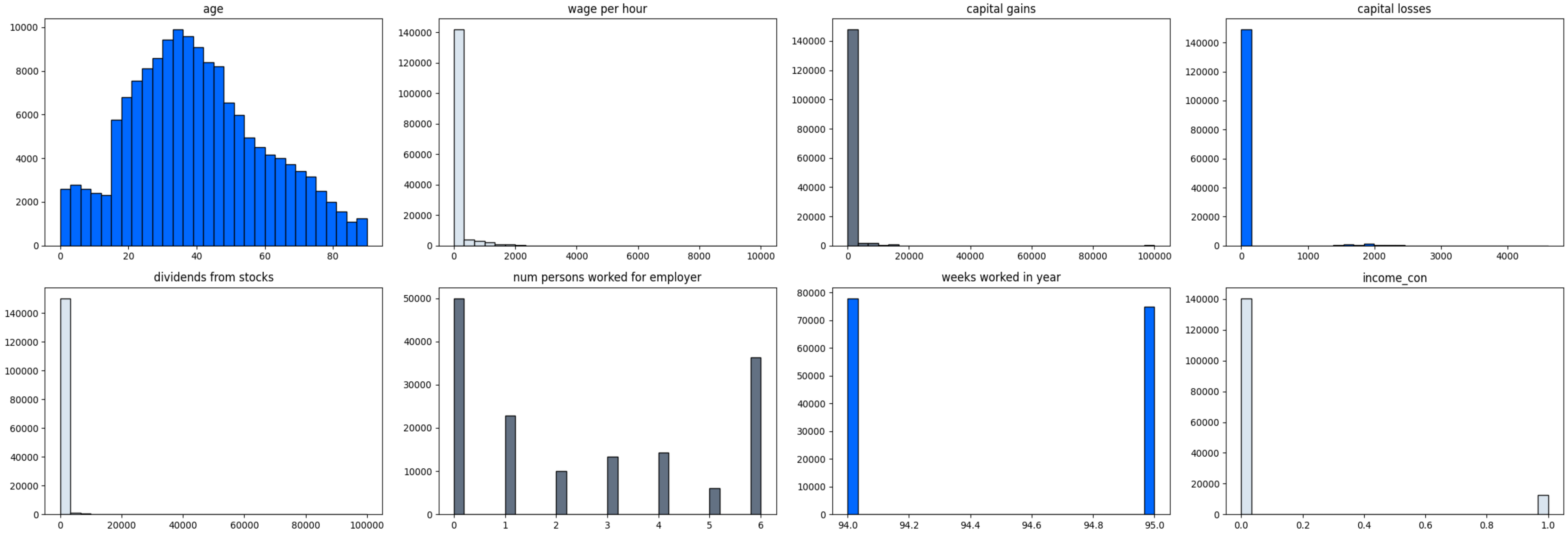
Data Processing & Preparation
- Nominal Features encoding
- Numerical Features Scaling

Findings & Actions:

- Detected Data Imbalance
- Dropped features with more than 20% data missing
- Handled missing values from the other features by imputation (mean and mode)
- Numerical Features correlation using Pearson reveled low to mid positive correlation of features
- Categorical using Chi square test an Carmer also reveled mid to low correlation

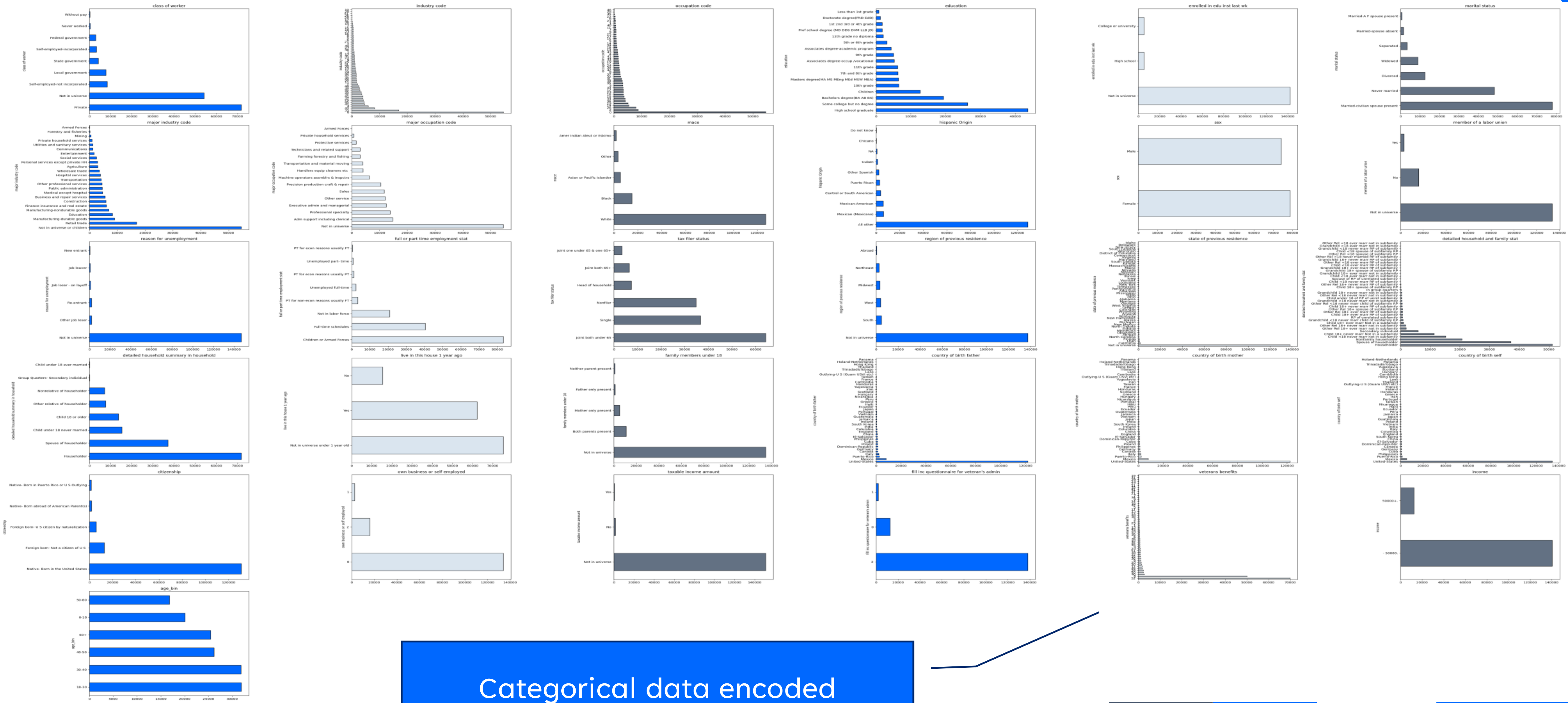


Numerical Features Distribution



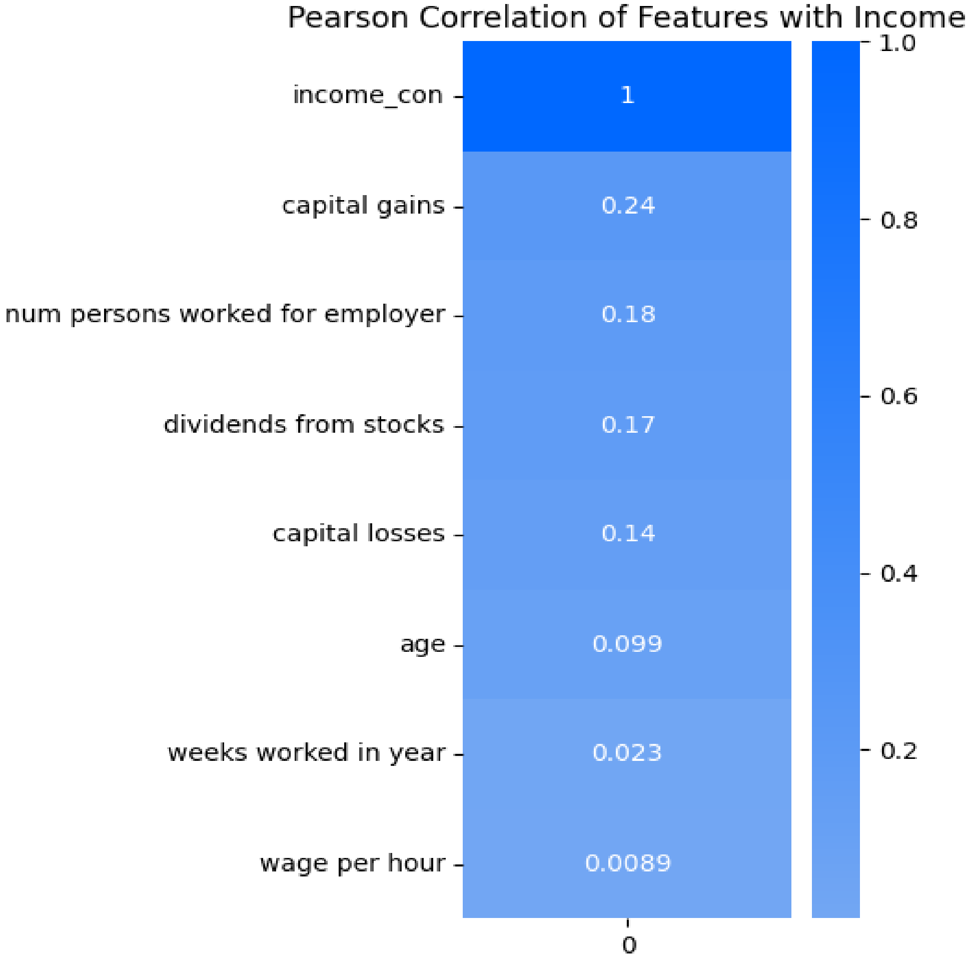
Features will be scaled using MinMax

Categorical Features Distribution



Categorical data encoded

Correlations of Numerical Features with Income



- Pearson method is used to show correlations between numerical variable and class
- +/- determine the type of relationship, while the number determine strength
- Based on Pearson method, all features has impact
- Capital gains, losses, dividends, and people working for employer has the strongest correlation

Correlations of Categorical Features with Income Chi Square

Feature	Chi-Square Statistic	p-value
income	152883	0.0
occupation code	26980	0.0
education	21129	0.0
major occupation code	17982	0.0
industry code	10401	0.0
major industry code	9053	0.0
veterans benefits	8669	0.0
class of worker	7628	0.0
detailed household and family stat	6870	0.0
age_bin	6717	0.0
tax filer status	5992	0.0
detailed household summary in household	5974	0.0
sex	4839	0.0
marital status	4342	0.0
full or part time employment stat	2748	0.0
family members under 18	1783	0.0
fill inc questionnaire for veteran's admin	1242	0.0
hispanic Origin	1229	0.0
country of birth father	1117	0.0
country of birth mother	1091	0.0
enrolled in edu inst last wk	939	0.0
own business or self employed	778	0.0
country of birth self	764	0.0
mace	727	0.0
citizenship	445	0.0
state of previous residence	431	0.0
member of a labor union	404	0.0
region of previous residence	395	0.0
live in this house 1 year ago	387	0.0
reason for unemployment	256	0.0
taxable income amount	74	0.0

- Chi Square test is a statistical method used to check if two categorical variables are related
- All Features seems to have a relationship with the class income based on p value.
- Strongest Features based on Chi test are occupation, education, industry, and major.

Correlations of Categorical Features with Income Cramér's V

- Cramér's V measures the strength of association between categorical variables. It ranges from 0 to 1, where 0 indicates no association and 1 indicates a perfect association.
- Based on the test, strongest features are occupation, education, major, and class of worker.

Feature	Cramér's V
income	1.0
occupation code	0.42
education	0.37
major occupation code	0.34
industry code	0.26
veterans benefits	0.24
major industry code	0.24
class of worker	0.22
detailed household and family stat	0.21
age_bin	0.21
detailed household summary in household	0.2
tax filer status	0.2
sex	0.18
marital status	0.17
full or part time employment stat	0.13
family members under 18	0.11
hispanic Origin	0.09
country of birth father	0.09
fill inc questionnaire for veteran's admin	0.09
country of birth mother	0.08
enrolled in edu inst last wk	0.08
mace	0.07
country of birth self	0.07
own business or self employed	0.07
state of previous residence	0.05
live in this house 1 year ago	0.05
member of a labor union	0.05
citizenship	0.05
region of previous residence	0.05
reason for unemployment	0.04
taxable income amount	0.02

Data Modeling

- Our problem is a **Binary Classification** with **unbalanced classes**, and need of model Interpretation and explainability
- Classes imbalance can **bias** the model toward the majority class
- Class imbalance is handled using **class weights** when training, it give higher weight to minority class
- GridSearchCV: Exhaustively tests combinations of **hyperparameters** Uses 3-fold cross-validation to avoid overfitting
- Tuned parameters: C for Logistic Regressionn_estimators, max_depth for RF/XGBoostlearning_rate for XGBoost

- Models Used:

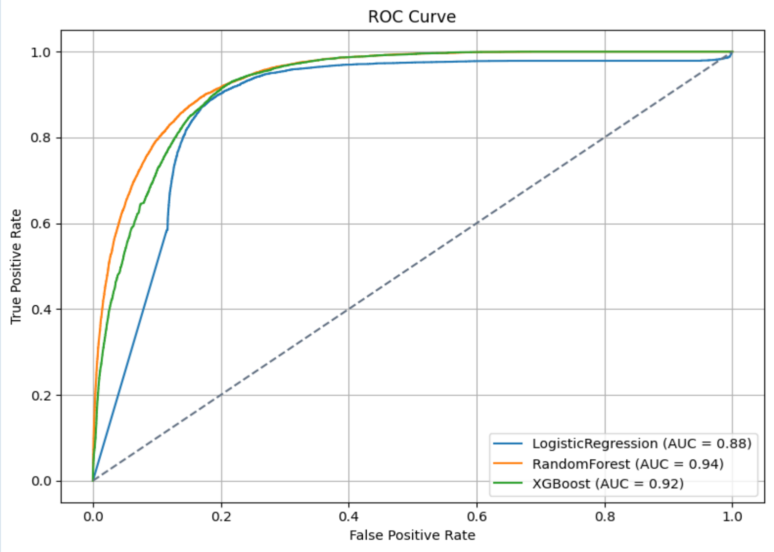
Logistic Regression:	Random Forest:	XGBoost:
• Simple, interpretable baseline, Estimates the probability of class using a weighted linear combination of features	• Handles non-linear relationships and noisy data and imbalance, it works by Building multiple decision trees and averages their predictions	• Gradient Boosted Trees—High-performing and optimized for structured data, Builds trees sequentially, each one correcting previous errors

Model Evaluation

Metrics Used:

- **ROC AUC**: Measures model's ability to rank predictions
- Precision: Focuses on correct positive predictions.
- Recall: Important due to minority class.
- **F1-Score:** Harmonic mean of Precision & Recall.

Model	F1	ROC AUC
Logistic Regression	81%	0.87
Random Forest	88%	**0.93**
XGBoost	82%	0.92



Best Model: Random forest with F1-Score: 88% and ROC AUC: 0.93
Excellent balance of recall and precision and handle data imbalance

Model Interpretation & Findings

Using SHAP to explain models

Attribute	How It Affects Income	Explanation
Age	Positive	Older individuals are more likely to earn higher income due to experience
Capital Gains	Strong Positive	Investment income is highly correlated with total income
Full-Time Employment	Positive	Full-time workers earn significantly more than part-time or unemployed
Education Level	Positive	Higher education (e.g., Bachelor's or above) leads to better-paying jobs
Occupation	Varies	Roles in management, technical, and professional fields increase income
Marital Status	Positive	Married individuals tend to report higher incomes (often dual-income households)
Citizenship	Slight Negative	Non-citizens were slightly less likely to earn >$50K on average
Gender	Male-dominated at higher incomes	The model reflects existing socioeconomic disparities in income

Results Summary

- Top features for >$50K:

Education: Advanced degrees strongly correlated

Capital Gains: Non-zero values key to high income.

Marital Status: Married individuals more likely >$50K.

Weeks Worked: Higher count associated with higher income

Age: 30–50 most frequent

Occupation: Managerial and technical jobs dominate

Potential Next Steps and Enhancement

- Perform deeper feature selection to reduce complexity.
- Investigate balance of the categorical features and any possible biases
- More Feature Engineering
- Explore SHAP for different age or race groups.
- Add explainability dashboards (e.g., Streamlit).

Questions?

Thank you

Nouf Alkedewi