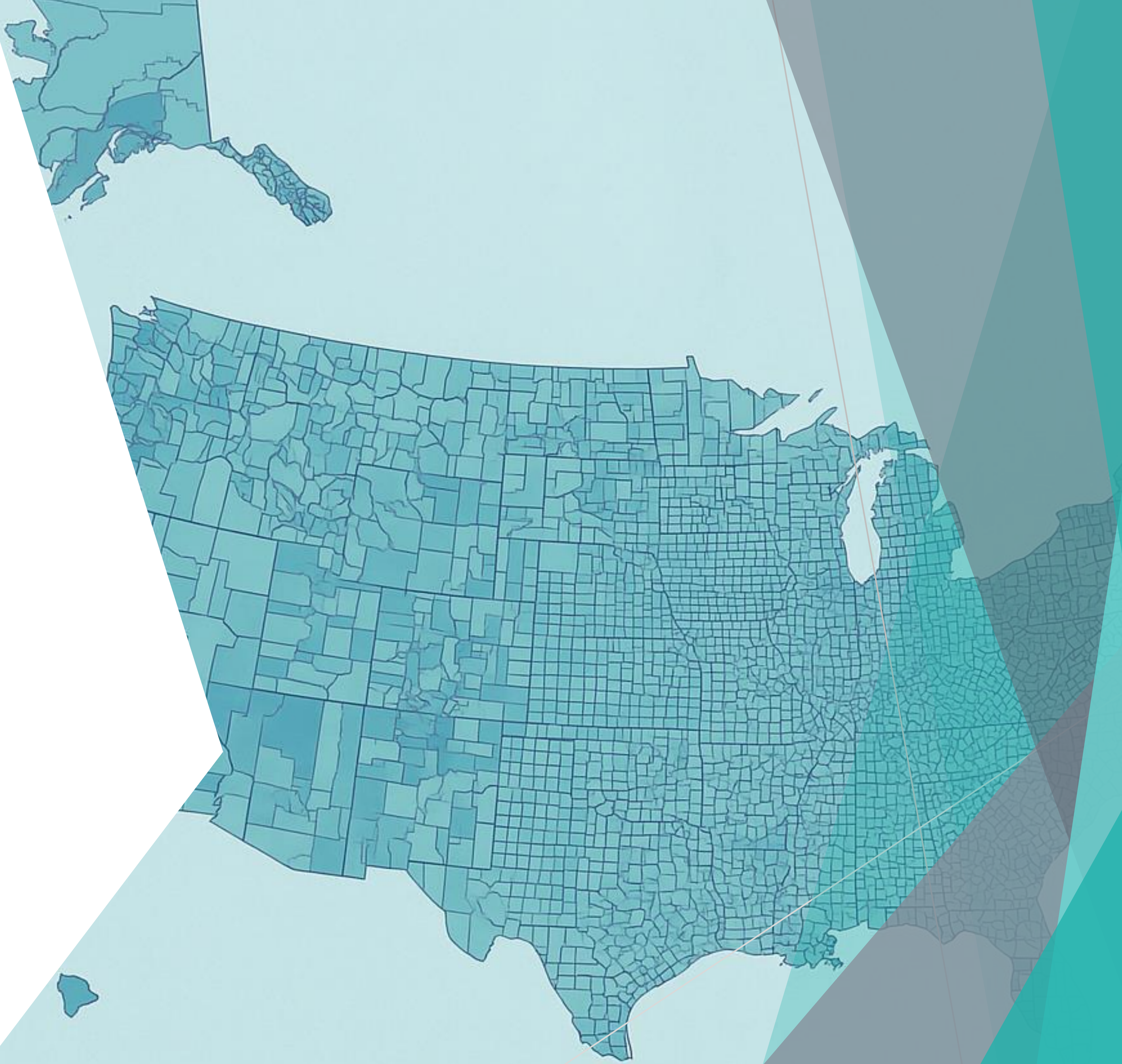
Understanding Income Drivers in the US Population

Uncovering actionable insights

07-04-2025

Gabriele Coli



dataiku

The Task

Every ten years, the census is conducted to collect and organize information regarding the US population with the intention of effectively allocating billions of dollars of funding to various endeavours.

Additionally, the collection of census information helps to examine the demographic characteristics of subpopulations across the country.

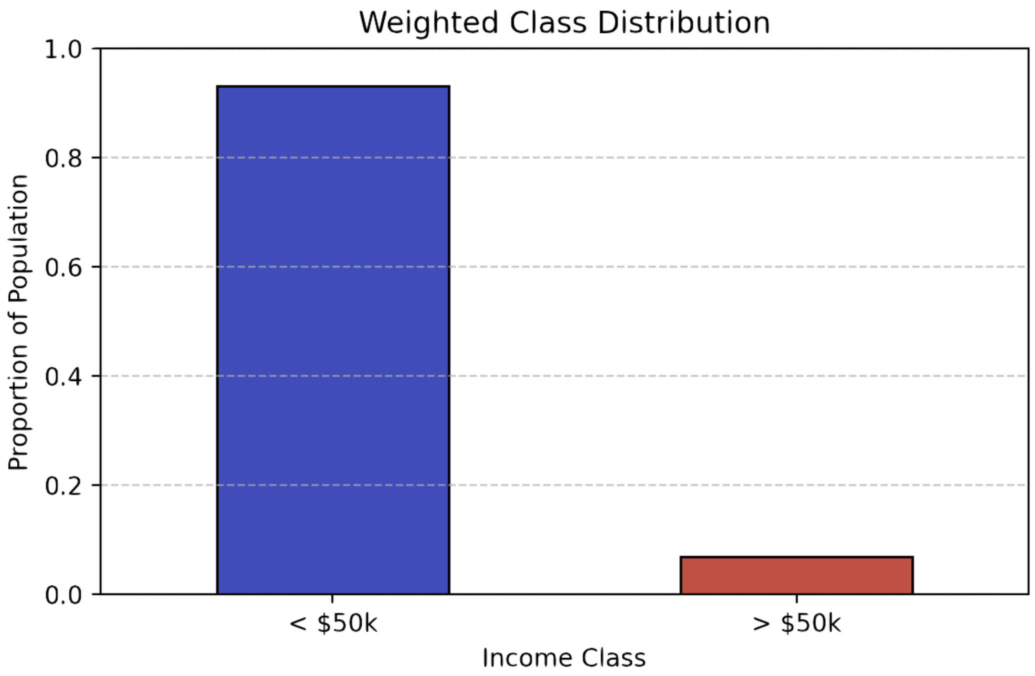
Help policymakers identify groups that are more or less likely to earn >$50K, to guide economic support strategies.

The Dataset

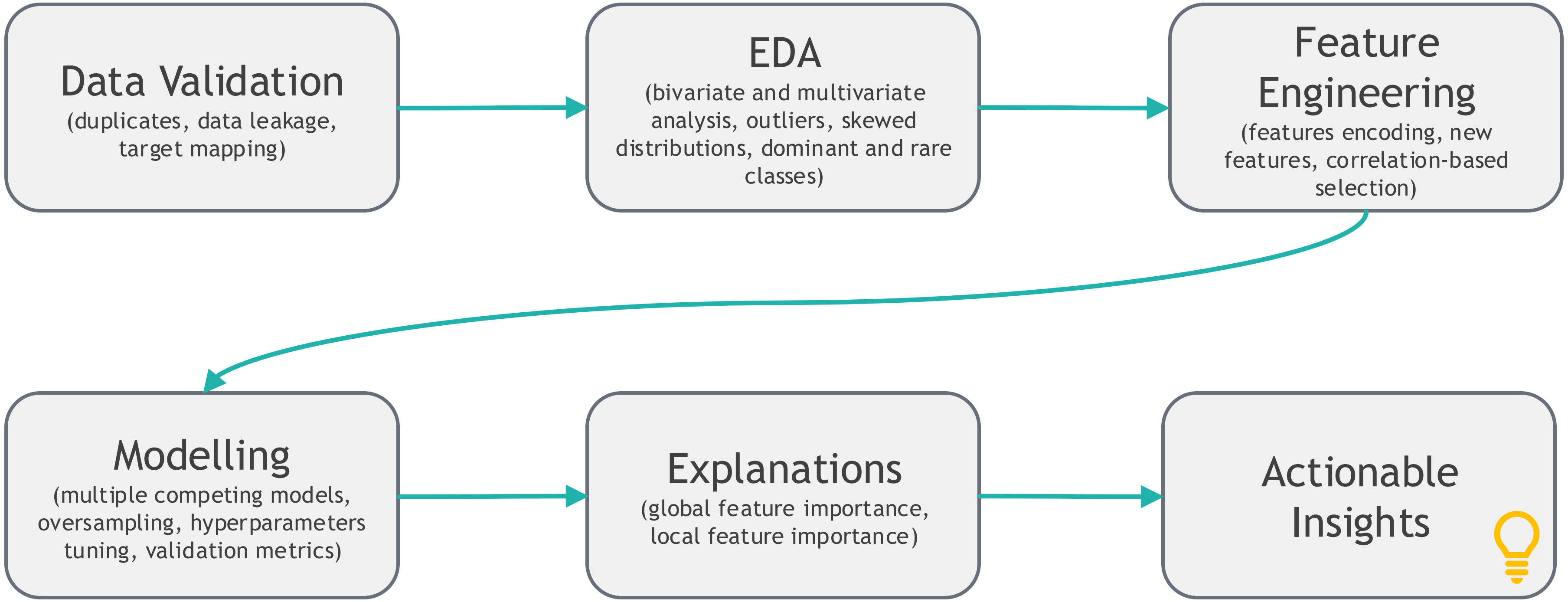
~300k anonymised individuals from US census, already split between training and test sets.

Features around age, education, occupation, citizenship, migration condition, sex, ethnicity, and more.

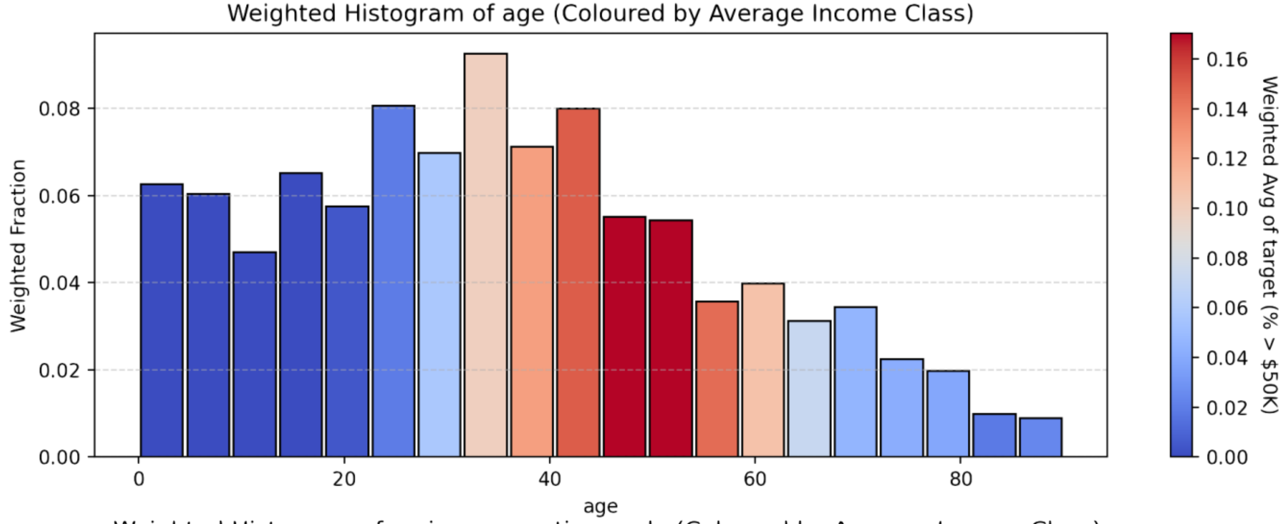
Collected with stratified sampling: each record has an associated weight.

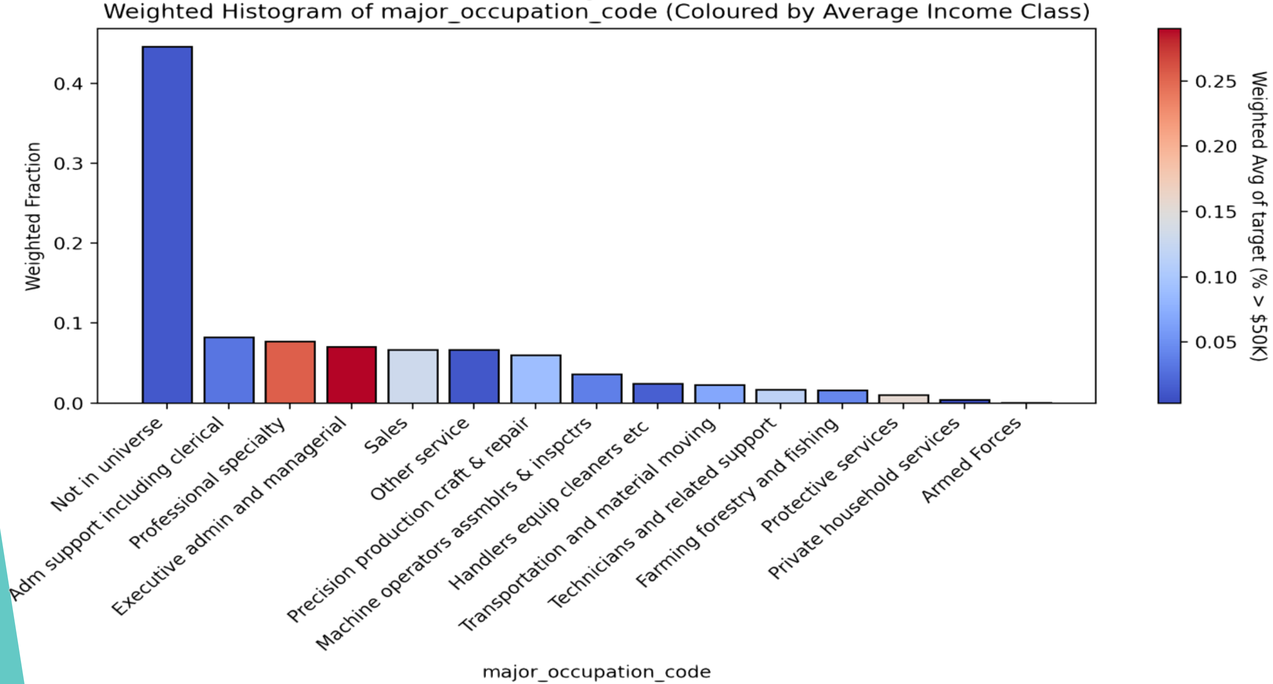


Methodology Overview



Key EDA Insights

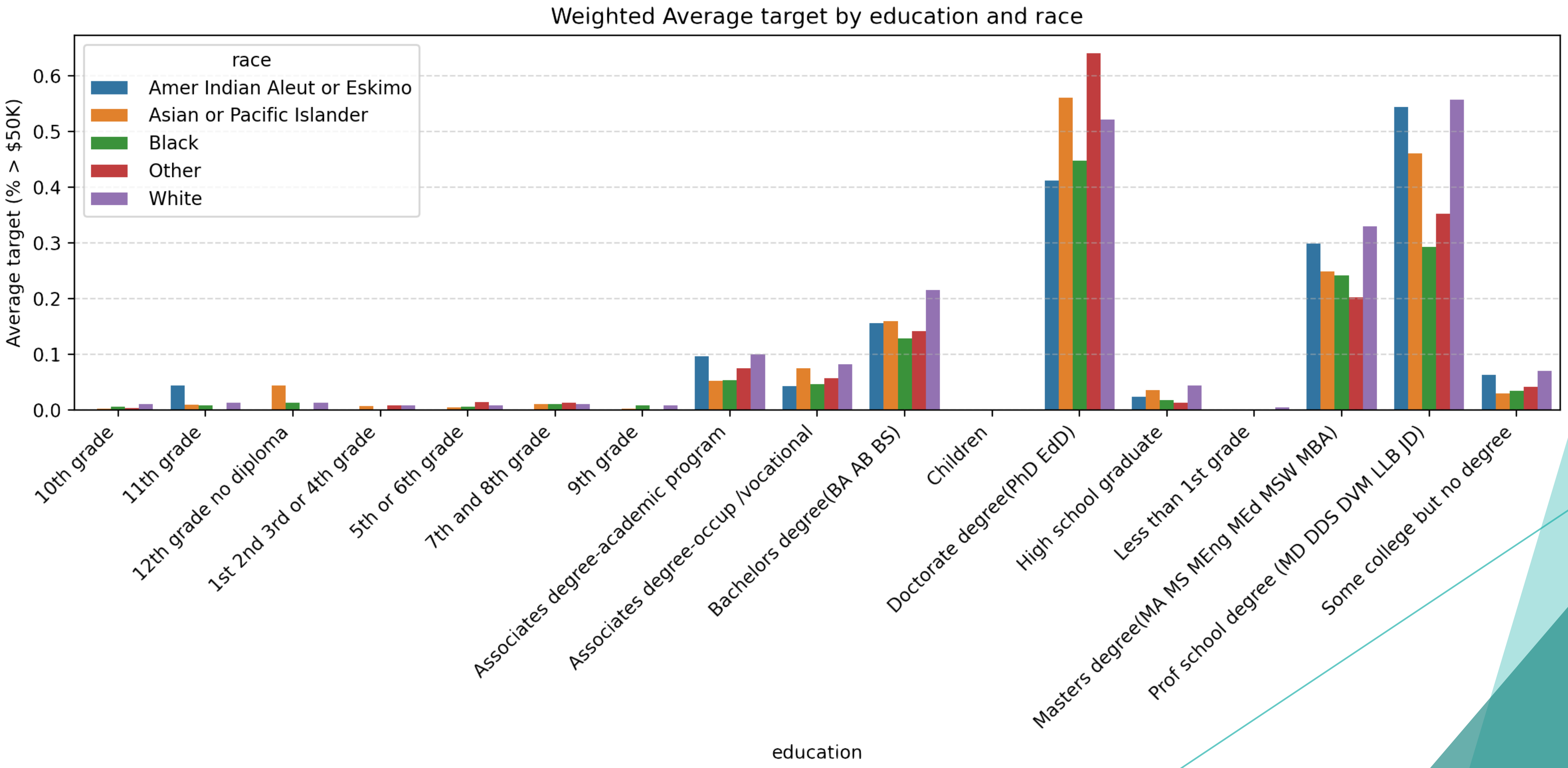




- Average target greatly varies with age;
- Occupation strongly correlates with income levels;
- Investment activities does too, but shows a very skewed distribution;
- Whether the individual is a male or a female also correlates with income levels.

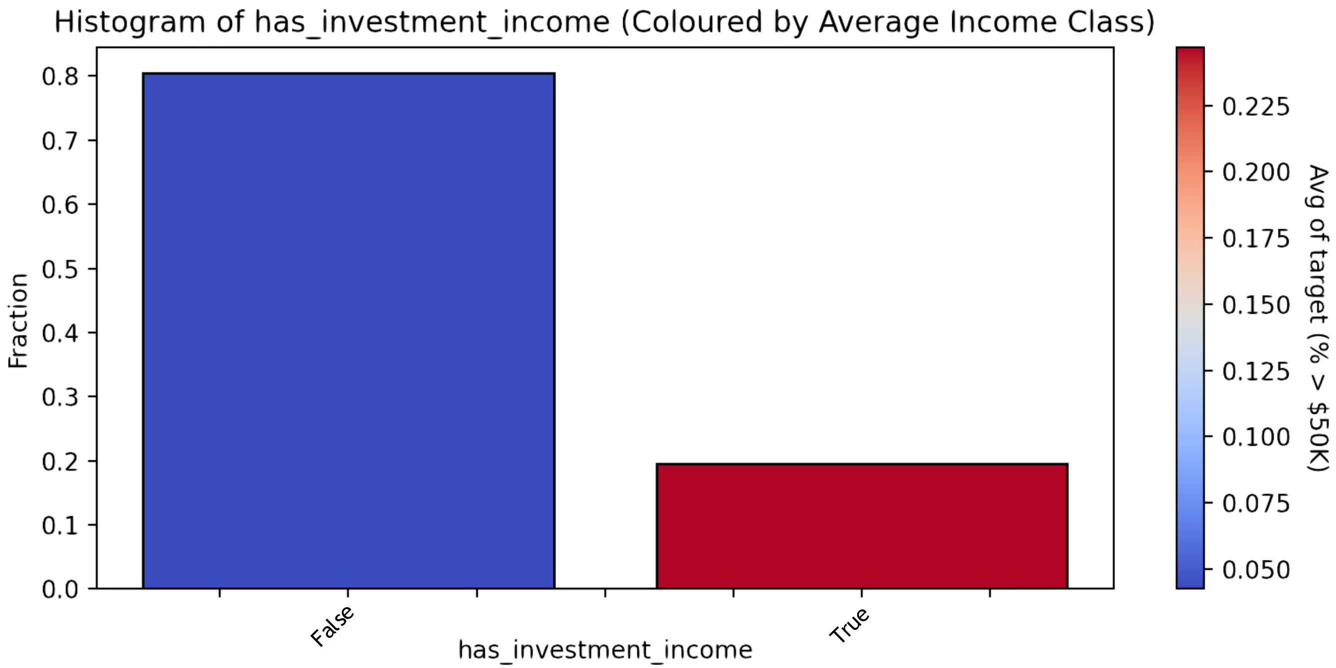
Key EDA Insights

Possibly the strongest predictor seems to be the education level of the individual. The multivariate analysis below simultaneously shows an income level disparity based on individual's ethnicity.



Data Preparation (numerical)

Feature	Distribution	Predictive Power	Action
Age	Balanced	High	Keep + Group ✓
Capital Gains	Skewed	High	Bool + Group ✓
Veterans Benefit	Unbalanced	Low	Drop ✗



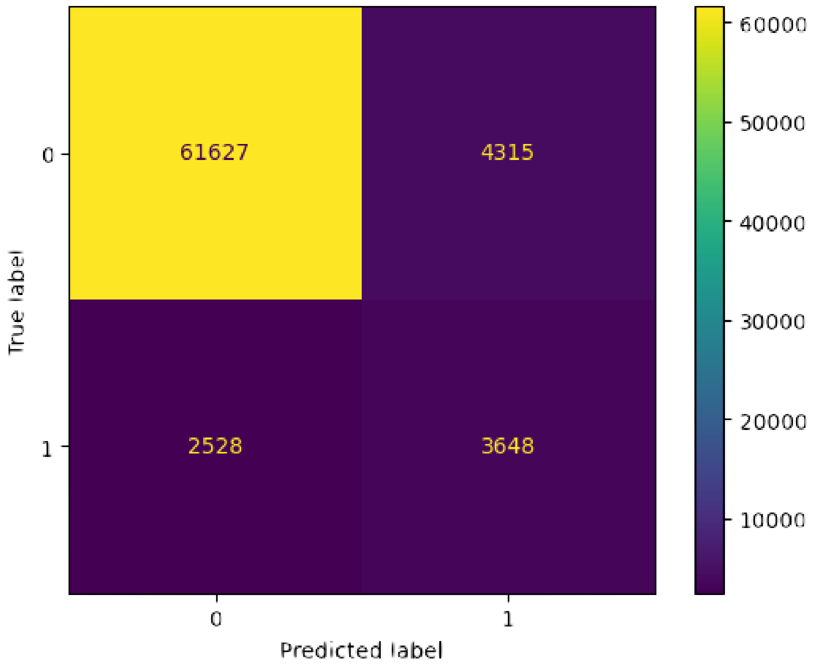
Data Preparation (nominal)

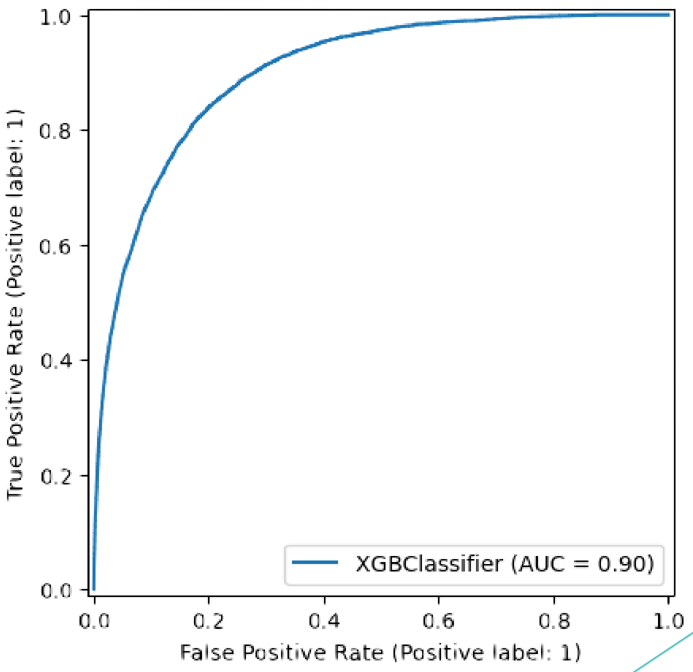
Feature	Dominance	Cardinality	Predictive Power	Comments	Actions
Education	26%	17	Strong	There is an existing order of values	Ordinal Encoding ✓
Occupation Code	44%	15	Strong	Frequency correlates with target	Frequency Encoding ✓
Hispanic Origins	88%	10	Medium	Natural way of grouping some of the values	To boolean ✓
Tax Filer Status	35%	6	Medium		One-hot encoding ✓
Previous Residence in Sunbelt	49%	4	Weak	Many NaNs	Drop ✗
Enrolled in education programme	93%	3	Medium	Redundant	Drop ✗

Modelling Approach & Results

Given the resulting feature matrix, we tried **several models and parameters**, and ultimately we made the following modelling choices:

- **Oversampling** of the minority class: given the strong class imbalance, we need to force the model to weight more the high income records in order to learn meaningful patterns.
- **XGBoost Classifier**: notoriously high-performance model, particularly suited for a mix of numerical and categorical features.



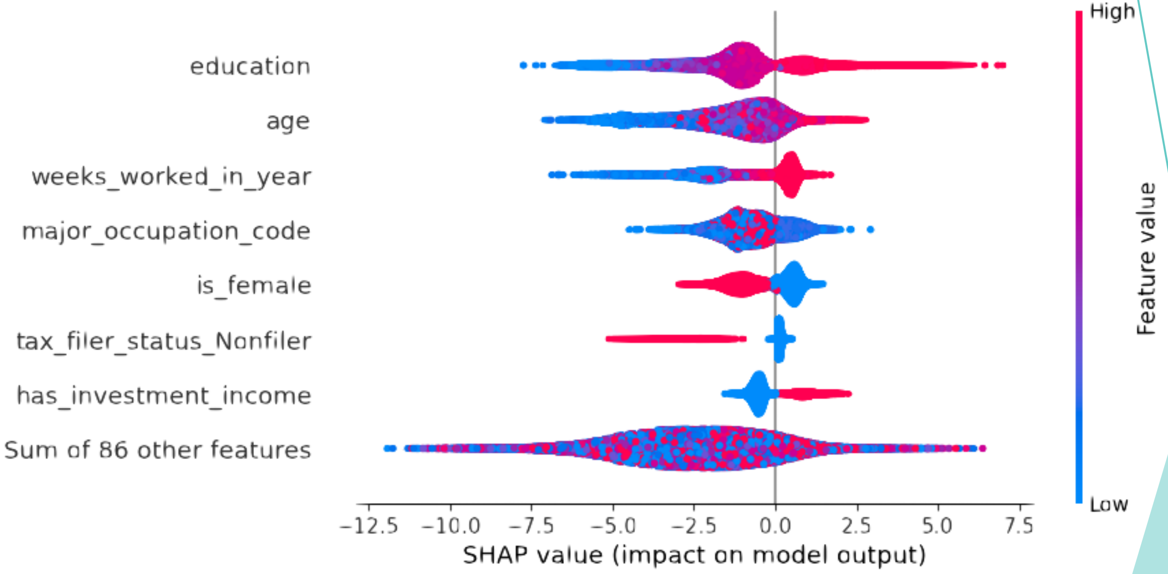


Feature Importance: Key Drivers

The key driver for income level is **education** above all. The specific **occupation** of an individual, together with her **investment activities** and her **sex** also strongly contribute.

When it comes to characterising the profile of an individual with high income, **age** and **tax filer status** are also crucial predictors, but it’s important to be mindful of the **causality relationships** here.

Other less crucial drivers are the **ethnicity** of the individual, her potential **Hispanic origins**, and the **birth country of her parents**.



Feature Importance: From Modelling to Policy

The key driver for income level is **education** above all. The specific **occupation** of an individual, together with her **investment activities** and her **sex** also strongly contribute.

When it comes to characterising the profile of an individual with high income, **age** and **tax filer status** are also crucial predictors, but it's important to be mindful of the **causality relationships** here.

Other less crucial drivers are the **ethnicity** of the individual, her potential **Hispanic origins**, and the **birth country of her parents**.

a. **Education accessibility** could radically improve economic outcomes for certain subpopulations.
b. Increase funding or access to **adult education** and degree-completion programs, especially in underprivileged areas.
c. Invest in **mid-career upskilling programs** to maintain productivity and income levels as the population ages.
d. Design **financial literacy** and investment education programs, especially for underserved communities.
e. **Align workforce development** initiatives with high-paying industries.
f. Promote **inclusive hiring and pay transparency** initiatives.
g. Promote **initiatives to reduce the gender income gap**, such as improving access to affordable childcare.

Limitations & Future Work

- Limitations:
  - No residency geographical data: it would unlock tailored interventions in specific areas/states.
  - No information on family social class: it would enable a deeper understanding of social lift and its effectiveness.
  - Limited temporal data: not possible to assess temporal trends and policy outcomes.

- Future Work:
  - Deep-dive in specific population segments and their correlation with income level.
  - Rigorous feature selection mechanism to enhance model predictions and explainability.
  - Automated HTML report generation from pipeline.
  - Pipeline robustness enhancement.
  - Causality study.

Questions?