Income Prediction using US Census Data

Exploring Characteristics Associated with Income Levels

Manish Patel

24/11/24

Problem Statement

Exploring Key Drivers of Income Levels in the U.S

Background

- The U.S. Census Bureau collects data to support strategic decisions on resource allocation and policy making.
- Income prediction is vital for understanding economic disparities and tailoring interventions.

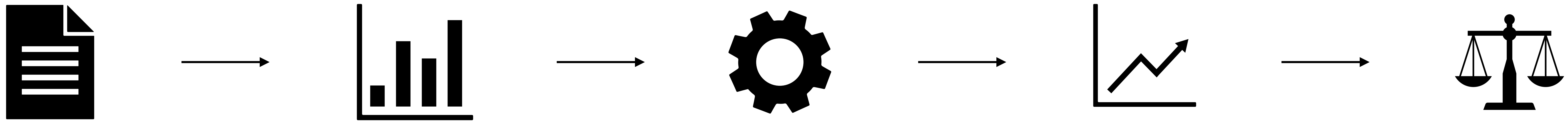
Objective

- Develop a model to predict whether an individual earns >$50K or <=$50K annually based on demographic and economic features.

Key Questions

1. What are the most influential factors in determining income level?
2. How accurately can machine learning models classify income groups?

Scope of Work



Input: Census Data (Demographics, Education, etc.)

EDA: Exploratory Data Analysis

Feature Engineering: Prepare Data for Modelling

Modelling: Build and Evaluate ML Models

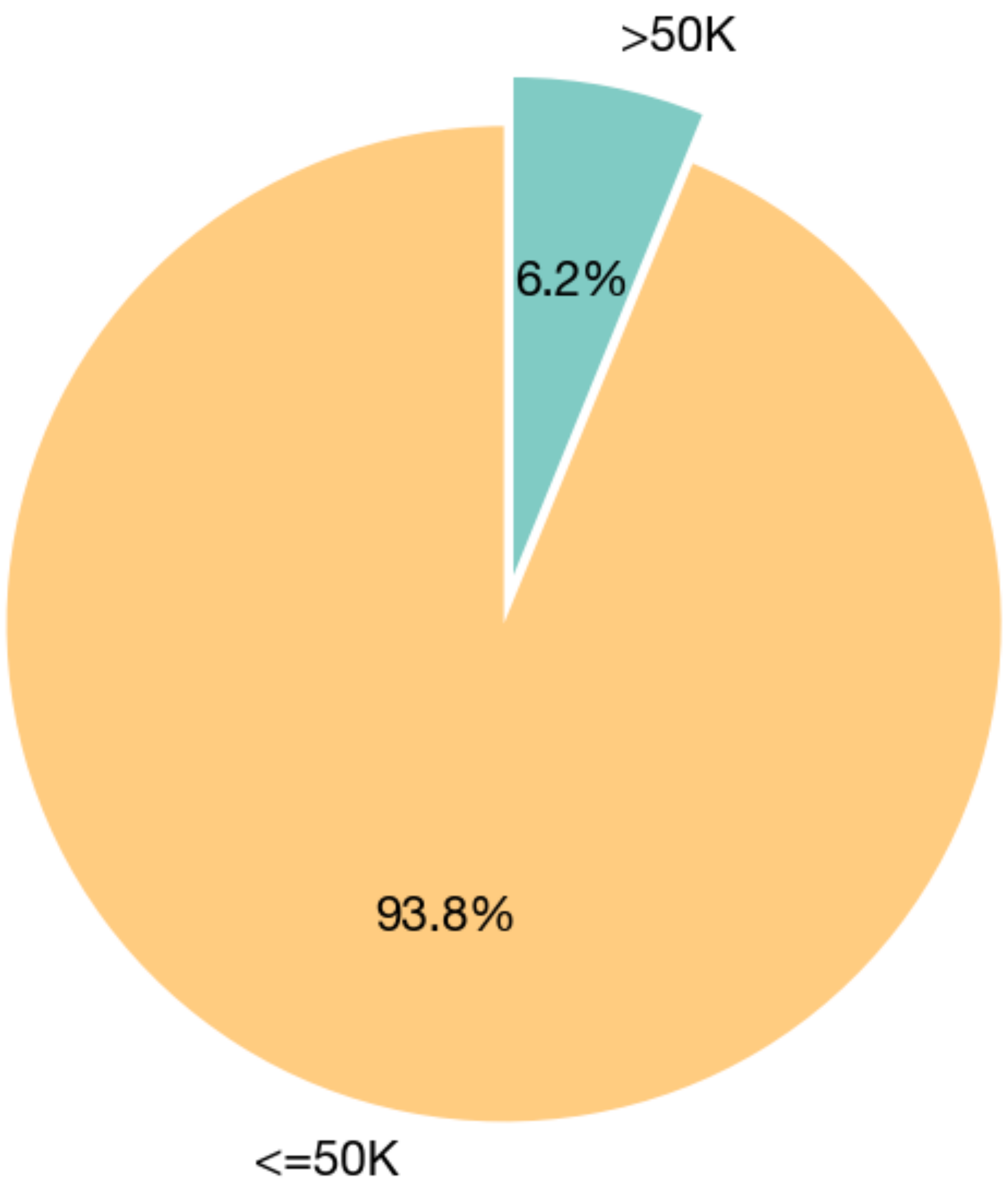
Output: Income Classification of >$50K vs. <= $50K

EDA (Exploratory Data Analysis)

Dataset Overview

Exploring Key Features and Challenges of the Dataset

- **Training**: 199,523 rows
- **Testing**: 99,762 rows
- **40 Features**: 7 numerical, 33 categorical
- **Target**: >50K or <=50K
- **Source**: U.S. Census Bureau
- **Challenge**: Class imbalance (93.8% <=50K, 6.2% >50K)



Income Class Distribution

EDA

Educational Attainment

What is this?

The chart shows the proportion of individuals earning more or less than 50K, based on **educational attainment** (i.e., the highest level of education completed by an individual).

Insights

Advanced Degrees Dominate:

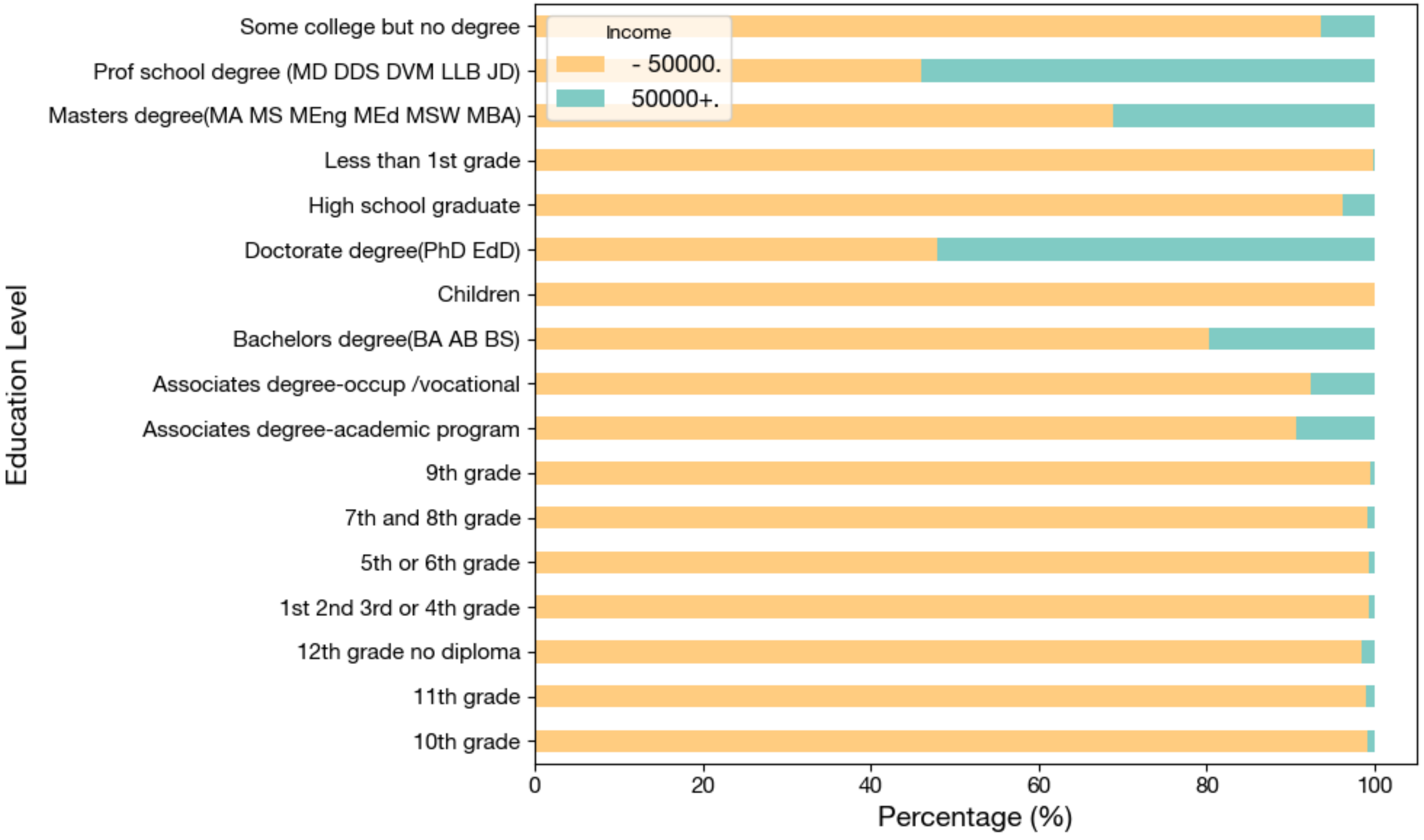
- Individuals with advanced degrees (e.g., Master's, Professional School, PhD) are far more likely to earn >50K
- For example, more than 50% of PhDs earn >50K.

Minimal Education → Low Income

- Groups with lower education levels (e.g., high school or less) overwhelmingly earn <=50K.

Prediction Relevance:

- Advanced degrees provide a clear signal for income above $50K, making educational attainment a strong predictive feature
- A Bachelor's degree offers some predictive power, but its role is more nuanced compared to higher education levels



EDA

Class of Worker

What is this?

This chart shows the proportion of individuals earning <=50K and >50K across selected worker categories, highlighting meaningful differences within government and self-employment groups.

Insights

Self employed (incorporated):

- These individuals have the highest proportion of earnings >50K, likely reflecting the benefits of structured entrepreneurship.

Federal government workers

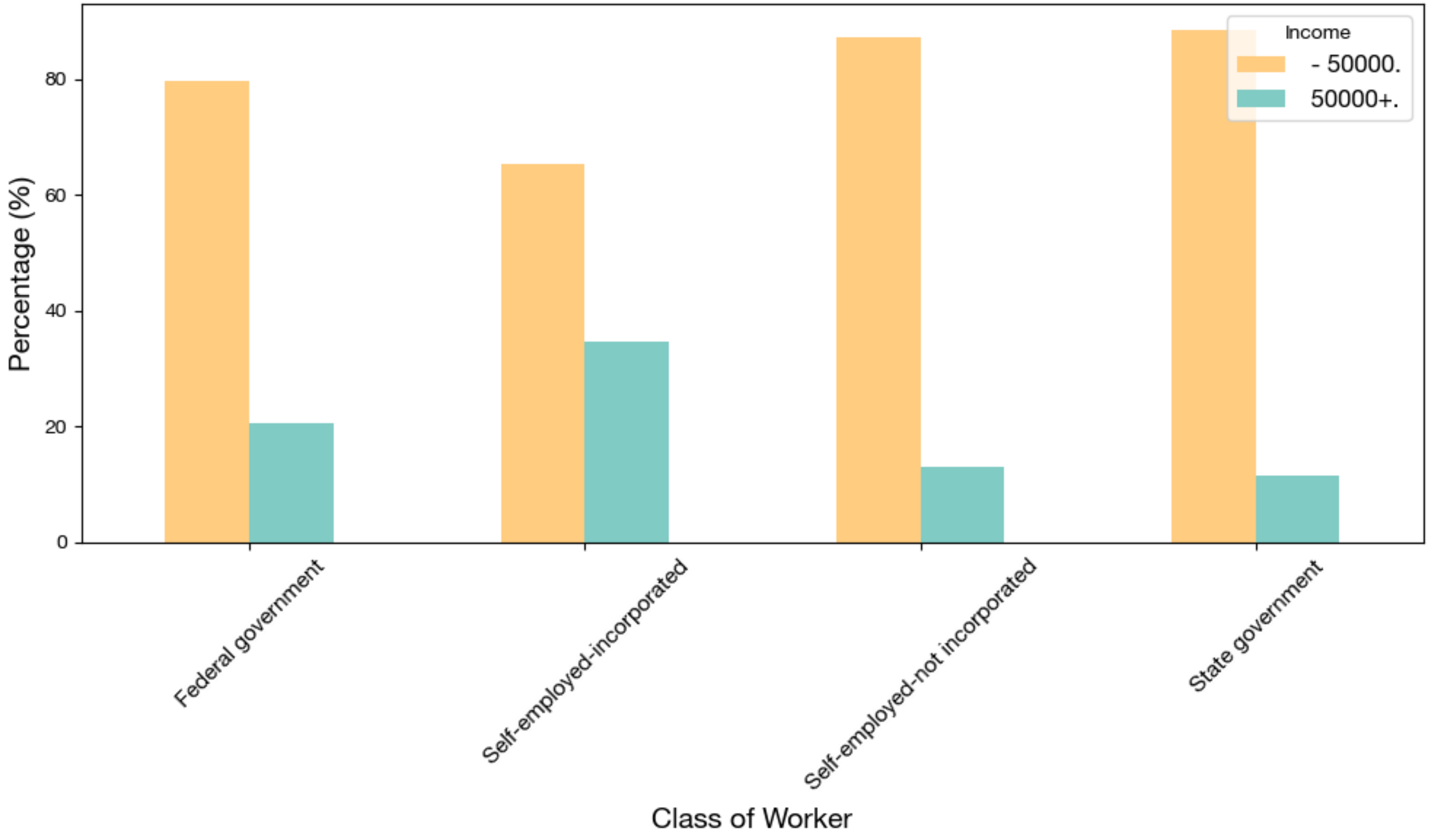
- More likely to earn >50K compared to **state government** workers, possibly due to better pay scales and benefits at the federal level.

Self-employed (not incorporated)

- Self-employed (not-incorporated) individuals and **state government** workers have similar patterns, with a high majority earning <=50K.

Prediction Relevance:

- Class of worker, when narrowed to key categories, can act as a **useful predictor** for income.
- Features like government level (federal vs. state) and incorporation status for self-employed individuals could provide distinct signals for classification models.



Income Proportion for Key Worker Types

EDA

Major Occupation Code

What is this?

This heat map shows the proportion of individuals earning <=50K and >50K for each major occupation category.

Insights

High-Proportion >50K **Occupations**:

- **Professional specialties** (e.g., doctors, engineers) have ~25% earning >50K
- **Executive/admin and managerial roles** follow closely with ~29%.

High-Proportion <=50K Occupations:

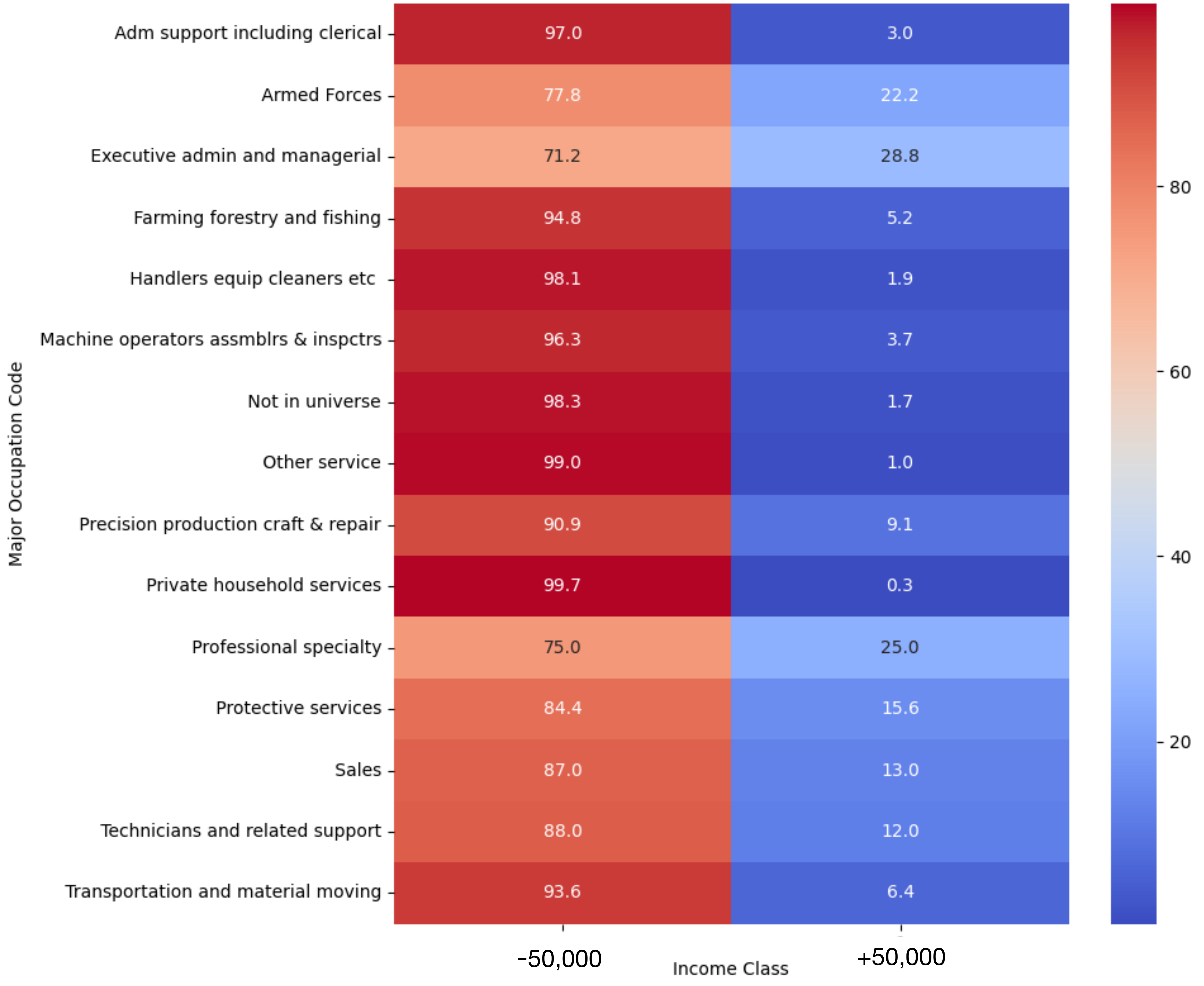
- Manual labor jobs like **Handlers**, **Machine Operators**, and **Private Household Services** have the lowest likelihood of earning >50K (<5%)

General Trends:

- Occupations requiring higher skill levels or specialised knowledge correlate strongly with higher income

Prediction Relevance:

- Occupation is a strong predictor of income due to the clear disparity in income proportions across different categories.
- Occupation is a strong predictor of income due to the clear disparity in income proportions across different categories.

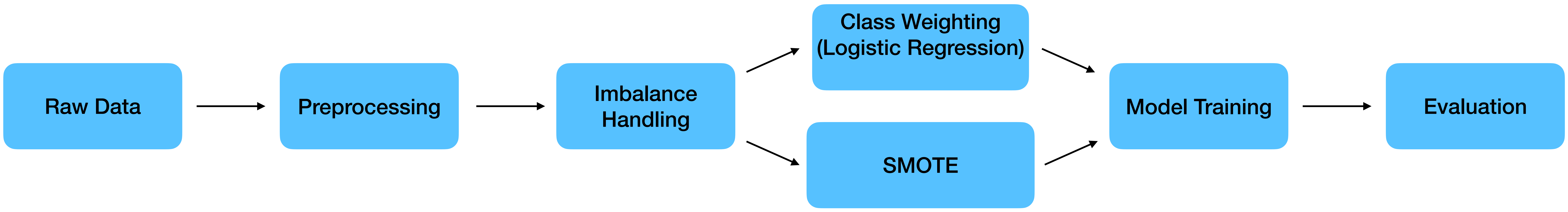


Income Proportion by Major Occupation Code

Modelling

Modelling Approach

This pipeline illustrates the process from raw data to model evaluation



Data Preparation

Dropped columns with a majority of '?' values.

Replaced sparse **'?' values** with ‘unknown’.

Dropped ‘Instance Weight’ for simplicity.

Removed duplicate rows for clean input.

Preprocessing

Scaled numerical features with ‘StandardScaler’

Encoded categorical features with ‘OneHotEncoder’.

Imbalance Handling

Class Weighting: Adjusted the Logistic Regression loss function.

SMOTE: Oversampled the minority class for a balanced training dataset.

Evaluation

Trained **Logistic Regression** and **Random Forest.**

Evaluated models on an imbalanced validation set.

Model Evaluation

Logistic Regression - under-sampling majority class (<= 50k)

Validation Set

	precision	recall	f1-score	support
0	0.86	0.84	0.85	2474
1	0.85	0.86	0.85	2473
Accuracy			0.85	4947
Macro avg	0.85	0.85	0.85	4947
Weighted avg	0.85	0.85	0.85	4947
	Validation AUC-ROC Score: 0.93			

Test Set

	precision	recall	f1-score	support
0	0.99	0.87	0.93	93,576
1	0.31	0.87	0.46	6186
Accuracy			0.87	99762
Macro avg	0.65	0.87	0.69	99762
Weighted avg	0.95	0.87	0.90	99762
	Validation AUC-ROC Score: 0.95			

1. **Validation vs. Test Imbalance**: The validation set was balanced via undersampling, which does not reflect the natural class imbalance in the test set, leading to a drop in Precision for >50K on the test set (31%).
2. **Loss of Majority Class Information**: Under-sampling reduced the diversity of the majority class (<=50K) in the training data, limiting the model's ability to generalise to the imbalanced test set.

Model Evaluation

Logistic Regression

Validation Set (No balancing with class weights)

	precision	recall	f1-score	support
0	0.99	0.84	0.91	28,001
1	0.32	0.86	0.47	2,473
Accuracy			0.84	30,474
Macro avg	0.65	0.85	0.69	30,474
Weighted avg	0.93	0.84	0.87	30,474
	Validation AUC-ROC Score: 0.95			

Validation Set (SMOTE)

	precision	recall	f1-score	support
0	0.98	0.87	0.92	28,001
1	0.36	0.81	0.50	2,473
Accuracy			0.87	30,474
Macro avg	0.67	0.84	0.71	30,474
Weighted avg	0.93	0.87	0.89	30,474
	Validation AUC-ROC Score: 0.92			

1. **Class Weights Performance:** Logistic Regression with class weights achieved **high Recall for** >50K **(86%)** but suffered from low Precision (32%), leading to a weak F1-score (0.47) despite a strong overall AUC-ROC of 0.93.
2. **SMOTE Trade-Offs**: SMOTE improved Precision for >50K to **36%** and **boosted overall accuracy (87%)** but failed to substantially improve Recall (81%) or F1-score (0.50), achieving a slightly lower AUC-ROC (0.92).

Model Evaluation

Decision Trees

Validation Set (Random Forest)

	precision	recall	f1-score	support
0	0.95	0.99	0.97	28,001
1	0.73	0.37	0.49	2,473
Accuracy			0.94	30,474
Macro avg	0.84	0.68	0.73	30,474
Weighted avg	0.93	0.94	0.93	30,474
	Validation AUC-ROC Score: 0.92			

1. **Majority Class Bias**: Random Forest performed strongly on the majority class (<=50K) with high Precision (95%) and Recall (99%), but struggled with the minority class (>50K), achieving only **37% Recall** and a low F1-score of **0.49**
2. **Trade-Off Between Classes**: Despite an overall AUC-ROC of **0.92**, the model heavily favored the majority class, indicating insufficient sensitivity to the minority class due to the class imbalance.

Future Work

Future Work and Next Steps

These steps provide a roadmap for refining the model pipeline and improving predictions.

1. Enhanced Exploratory Data Analysis (EDA)

- Perform deeper analysis of **numerical features** like capital gains, dividends from stocks, and age, as they likely play a significant role in income prediction.
- Investigate feature relationships (e.g., correlations, interactions) to guide **feature engineering** for creating more informative variables.

2. Advanced Feature Engineering

- Incorporate the **instance weight** column to better reflect the population distribution.
- Explore new features (e.g., feature interactions or transformations) derived from existing data.

3. Robust Model Training

- Use **hyperparameter optimisation** (e.g., grid search or random search) to fine-tune model parameters.
- Implement **k-fold cross-validation** to improve robustness and ensure generalisation.

4. More Complex Models

- Experiment with **XGBoost** to leverage feature diversity and handle class imbalance effectively.
- Investigate **neural networks** to determine if they better capture non-linear relationships in the data.

5. Addressing Class Imbalance

- Explore **hybrid sampling techniques** (e.g., SMOTE combined with undersampling).
- Consider **cost-sensitive learning** approaches to improve precision-recall trade-offs for the minority class (>50K).