US Census Assessment

Gian Jaeger

Income classification: data and task

The raw data:

- Individual-level microdata from the US Census
- Target variable: binary indicator for whether annual income exceeds $50,000
- Total sample size (N): 299,283 individuals; positive class: 6%
- Feature set: 41 variables

Objective:

“identify characteristics that are associated with a person making more or less than $50,000 per year”

→ binary classification problem

Outline of proposed approach

- Data preparation and cleaning
    - Mappings, NaNs, duplicates, conflicting values and more
- Exploratory data analysis (all data)
    - Capture the underlying distributions of our primary indicators
    - Group level descriptive statistics
    - Conditional probabilities
    - Hypothesis testing
- Feature selection and engineering
    - Drop unnecessary and highly collinear features
    - Ordinally encode or one hot encode categorical features
- Machine learning analysis (train/test split)
    - Random forest to identify non-linear relationships and assess variable importance
    - Logistic regression to estimate interpretable effects and quantify marginal contributions of each feature

Data preparation and cleaning

Operations conducted:

- Mapped column names from metadata
- Decoded numeric categorical variables
- Merged train and test data
- Checked for NaN values
- Checked for duplicates
- Checked for negative values across continuous variables

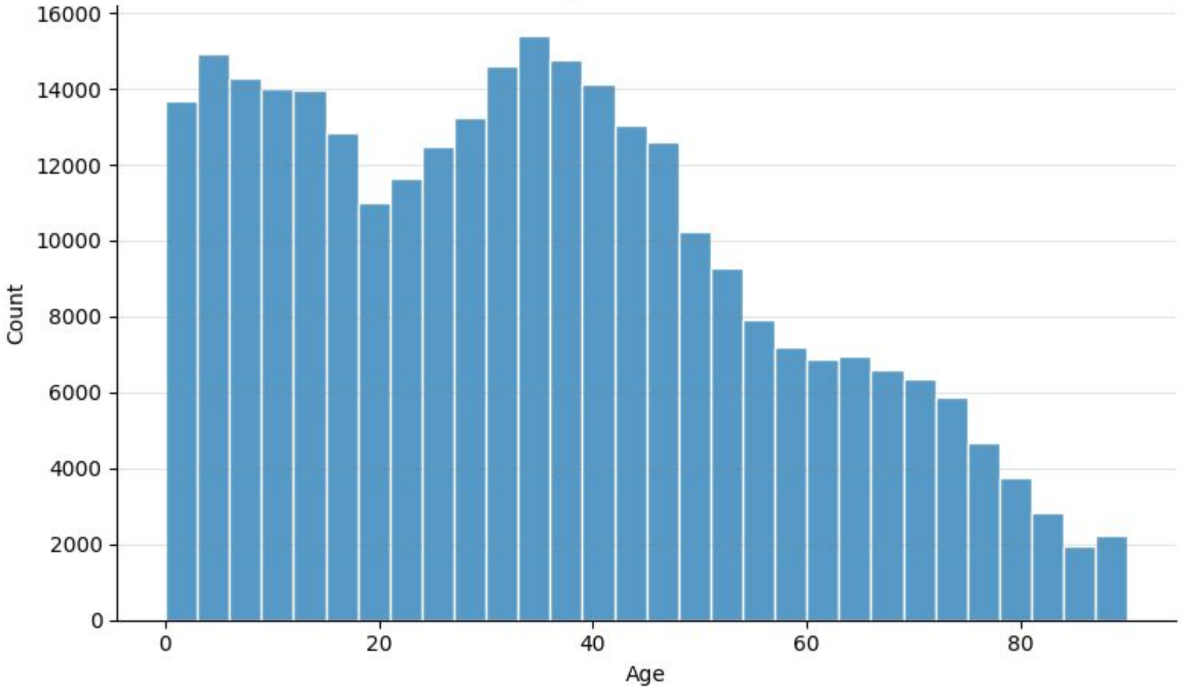
Additional considerations (decided against):

- Winsorize continuous variables at a 99% limit and subsequently standardized to a [0, 1] range
- Harmonization of categorical variables → not necessary
- Rule-based imputation → not necessary
- Standardising categorical variables → see machine learning section
- Metadata mentions 67,652 “duplicates or conflicting instances”

Exploratory analysis

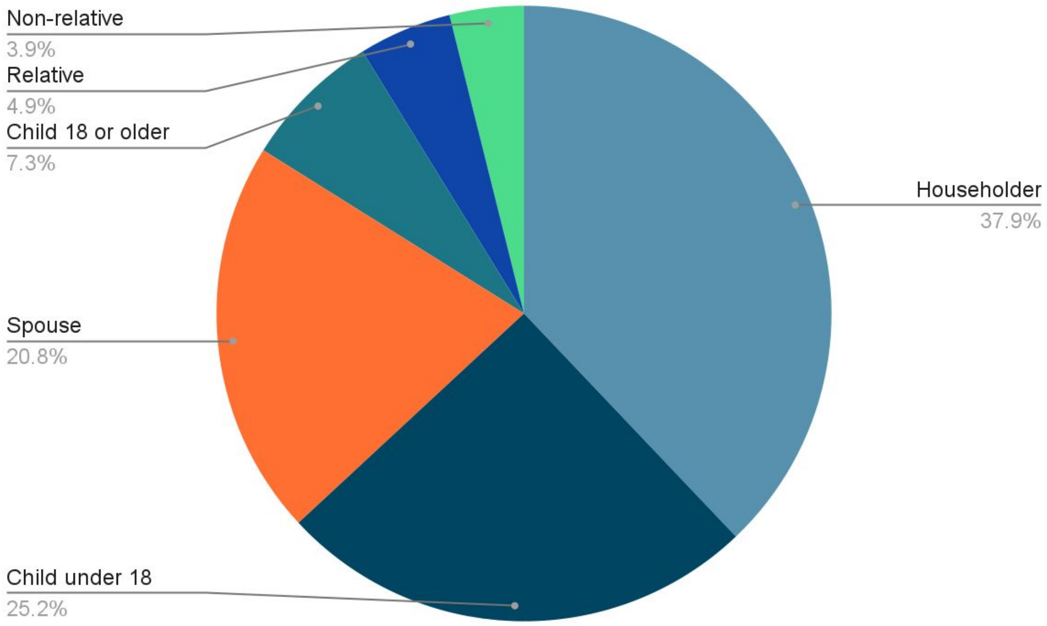
- Understanding basic distributions within the data (age, gender, household role, employment, occupation etc.)
- Boolean encode important variables to get more interpretable insights (e.g., how does having a bachelor’s degree or higher impact the likelihood of earning more than $50,000?)
- Group level descriptive statistics
- Conditional probabilities for relevant features
- Hypothesis testing to evaluate one-to-one correlations between the target variable and relevant features
    - Numeric features → Mann–Whitney U test and Cliff’s delta
    - Categorical features → Chi-square test of independence and Cramer’s V

Age



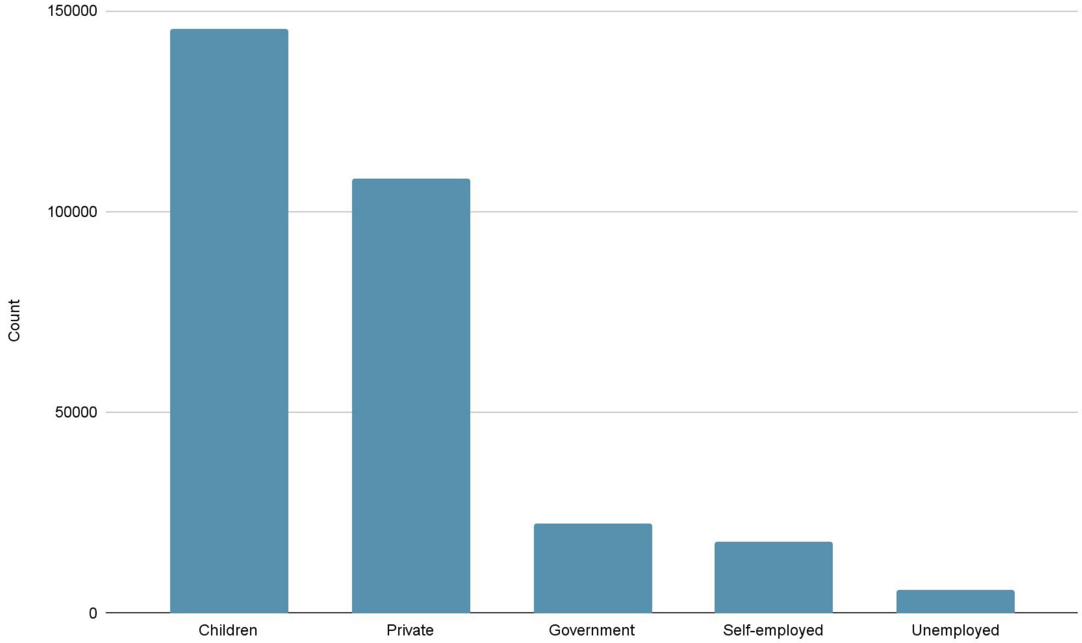
***Gender split**: 48% male, 52% female*

Household composition



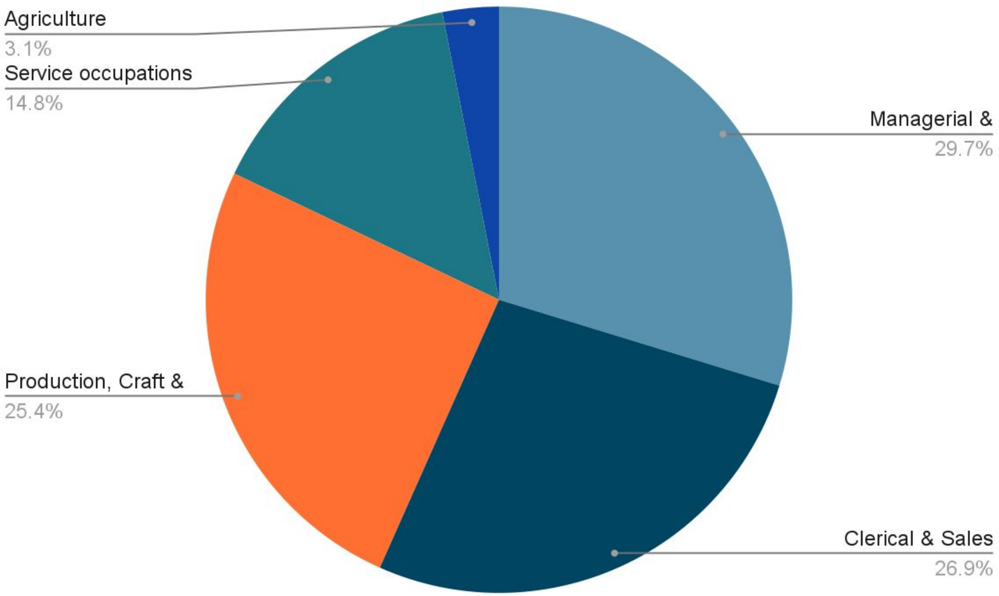
The categories “child under 18 never married” and the category “child under 18 ever married” were combined into “child under 18”. Additionally, the categories “Nonrelative of householder" and "Group Quarters- Secondary individual" were combined into "Non-relative".

Employment status



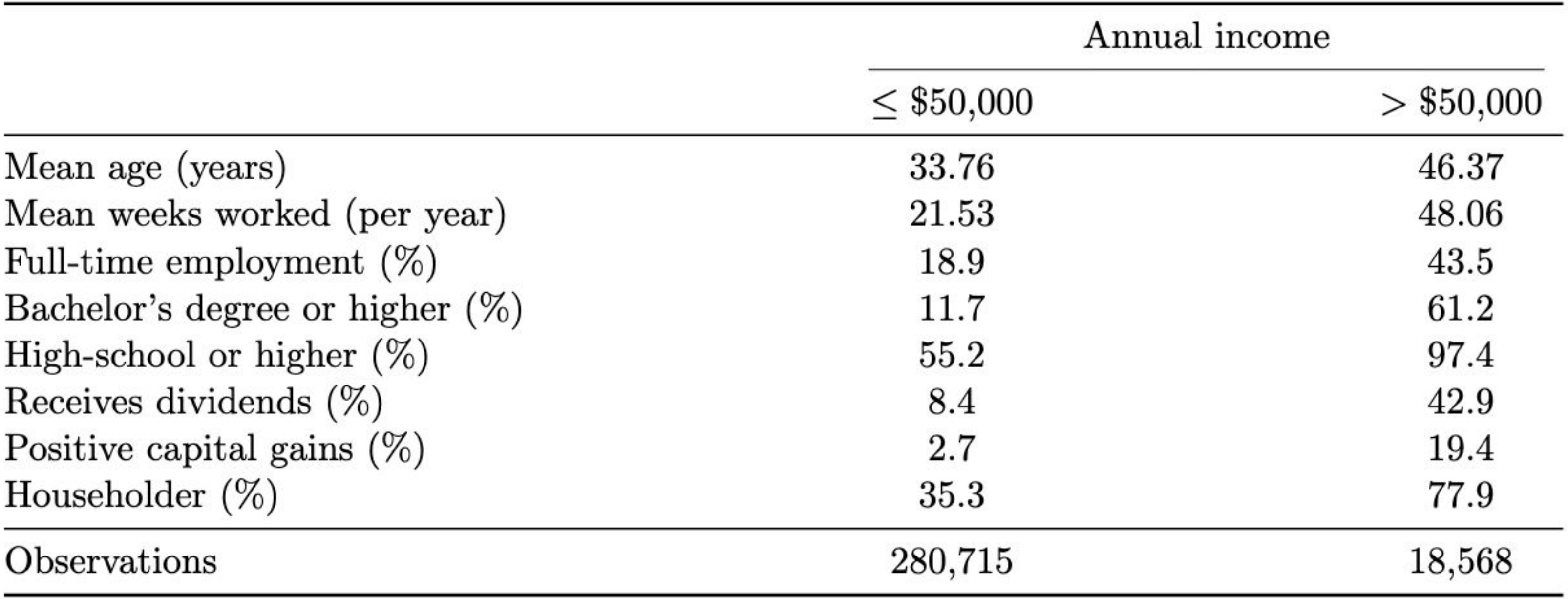
This information was primarily derived from the "class_of_worker" column with parts from the "employment_status" column. Individuals in the "Self-employed-not incorporated" "Self-employed-incorporated" were grouped, as well as individuals in the Federal government, State government and Local government categories. Additionally, those "Without pay" (0.08%) and those who "Never worked" (0.2%) were placed into the unemployed category.

Occupation distribution (split within employed category)



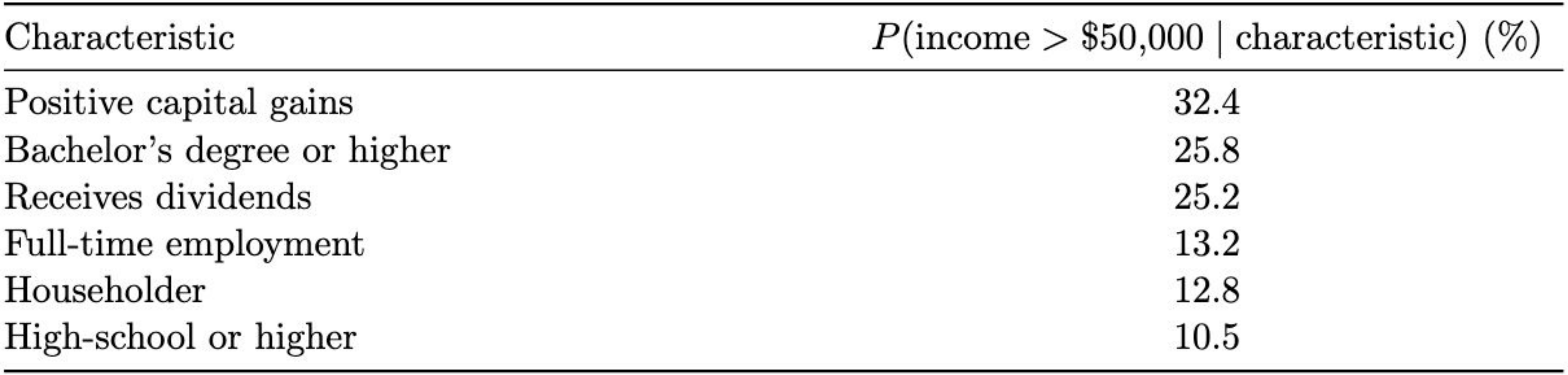
These groups were constructed by combining occupations from 15 more granular categories. See code for a detailed breakdown. They contain occupations for the roughly 49% of people in the survey who are employed.

Sample characteristics by income group



	Annual income	
	≤ $50,000	> $50,000
Mean age (years)	33.76	46.37
Mean weeks worked (per year)	21.53	48.06
Full-time employment (%)	18.9	43.5
Bachelor’s degree or higher (%)	11.7	61.2
High-school or higher (%)	55.2	97.4
Receives dividends (%)	8.4	42.9
Positive capital gains (%)	2.7	19.4
Householder (%)	35.3	77.9
Observations	280,715	18,568

Conditional probabilities of earning more than $50,000



Characteristic	$P(\text{income} > \$50{,}000 \mid \text{characteristic})$ (%)
Positive capital gains	32.4
Bachelor's degree or higher	25.8
Receives dividends	25.2
Full-time employment	13.2
Householder	12.8
High-school or higher	10.5

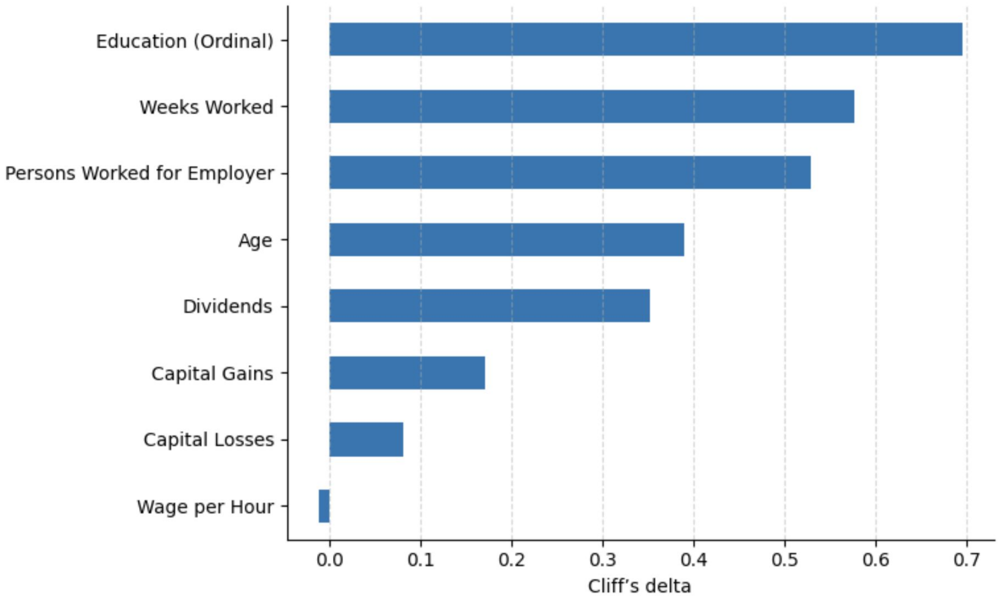
Bivariate associations for numeric features

- **Mann-Whitney U test**: A nonparametric significance test that evaluates whether the distribution of a numeric variable differs between the low-income and high-income groups

→ result: $p < 10^{-10}$ for all features

(see code for exact values)

- **Cliff's delta**: correspondingly captures how strongly values in the high-income group tend to exceed (or fall below) those in the low-income group (–1 to 1 scale)



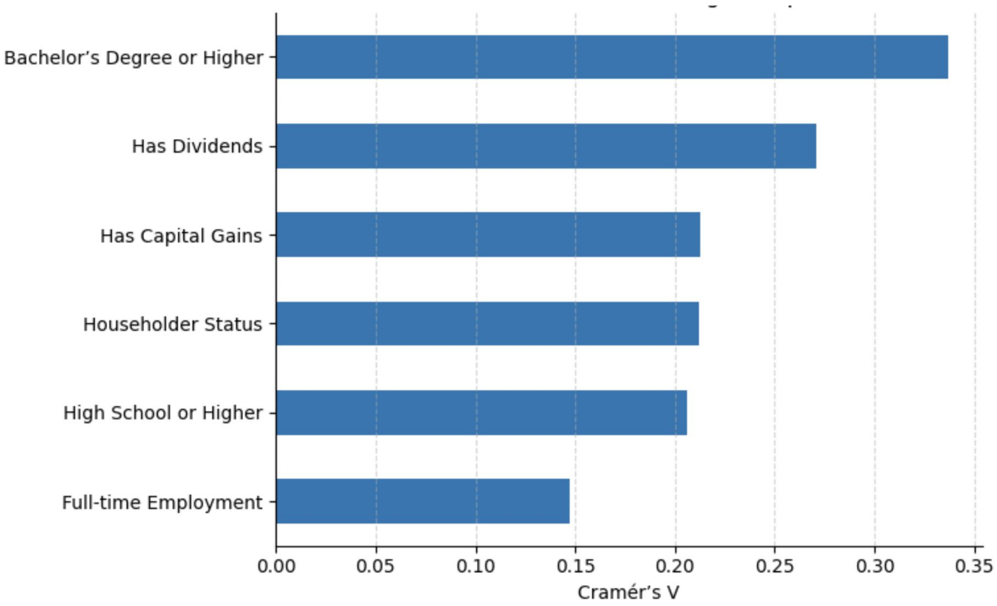
Bivariate associations for categorical features

- **Chi-square test of independence**: assesses whether two categorical variables are statistically associated by comparing observed and expected frequencies

→ result: $p < 10^{-10}$ for all features

(see code for exact values)

- **Cramer's V**: standardised effect-size measure from the chi-square statistic that quantifies the strength of association between categorical variables (0–1 scale)



Feature selection and engineering

- Applied binary encoding to the target variable (1 = 50,000+, else 0)
- Removed the following features
    - 'instance_weight'
    - 'year'

    not relevant

    - 'state_prev_residence'
    - 'household_family_stat'
    - 'detailed_occupation_recode'
    - 'Detailed_industry_recode'

    have a summary column

    - 'migration_code_change_in_reg'
    - 'Migration_code_change_in_msa'

    significant multicollinearity
- Encoding for categorical variables:
    - Education → ordinal encoding
    - 33 other variables → one hot encoding
- Additionally considered grouping rare categorical answers to improve model stability

Machine learning analysis

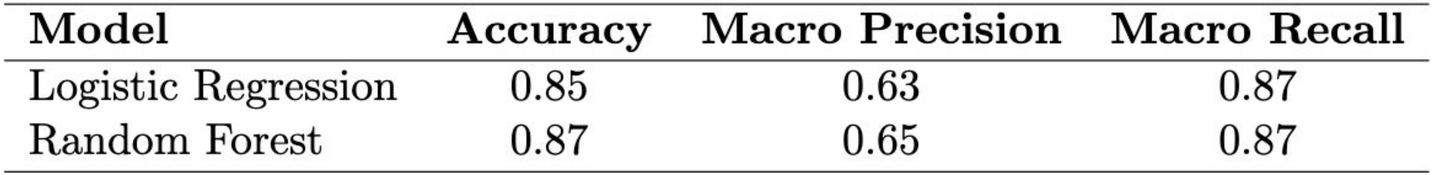
- Why machine learning?
  - Our previous analysis has focused on average, marginal, and mostly linear relationships
- Machine learning models can extend this by capturing:
  a. Non-linear relationships between features and the target variables
  b. Interaction effects between 2+ features and the target variable (e.g., how does receiving veterans benefits and being privately employed affect the chance of earning >$50,000)
  c. Out-of-sample predictive performance to evaluate whether patterns generalise
- Models:
  - **Random forest**: Performs well on tabular survey data, offers model-level interpretability and is efficient
  - **Logistic regression**: Captures linear relationships; it is included as an interpretable baseline to validate the Random Forest

Machine learning models

- Random forest
    - Hyperparameters tuned via grid search:
        - Number of trees (n_estimators)
        - Maximum tree depth (max_depth)
        - Minimum samples per leaf (min_samples_leaf)
        - Number of features considered when “splitting” (max_features)
    - These parameters are chosen via three-fold cross-validation on the training set by optimising for balanced accuracy
    - Interpretability achieved via feature importance plots and partial dependence plots
- Logistic regression
    - Implemented standard scaling (Z-score normalization) for continuous variables
    - Trained to optimise balanced accuracy to allow for a meaningful comparison
    - Interpretability achieved via coefficient plots

Model performance

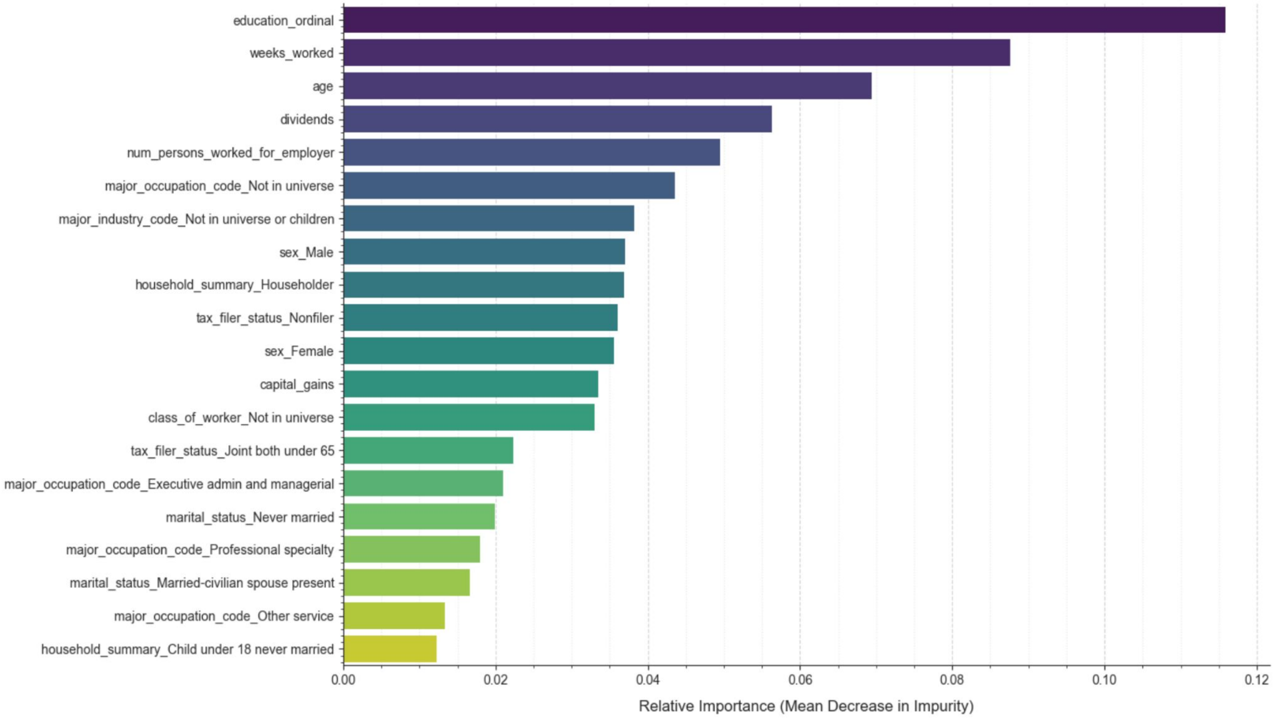
Key finding: Model performance is not our primary focus, but the convergence of the two distinct modeling techniques provides support for the notion that the identified characteristics are robust, reproducible, and statistically significant



Model	Accuracy	Macro Precision	Macro Recall
Logistic Regression	0.85	0.63	0.87
Random Forest	0.87	0.65	0.87

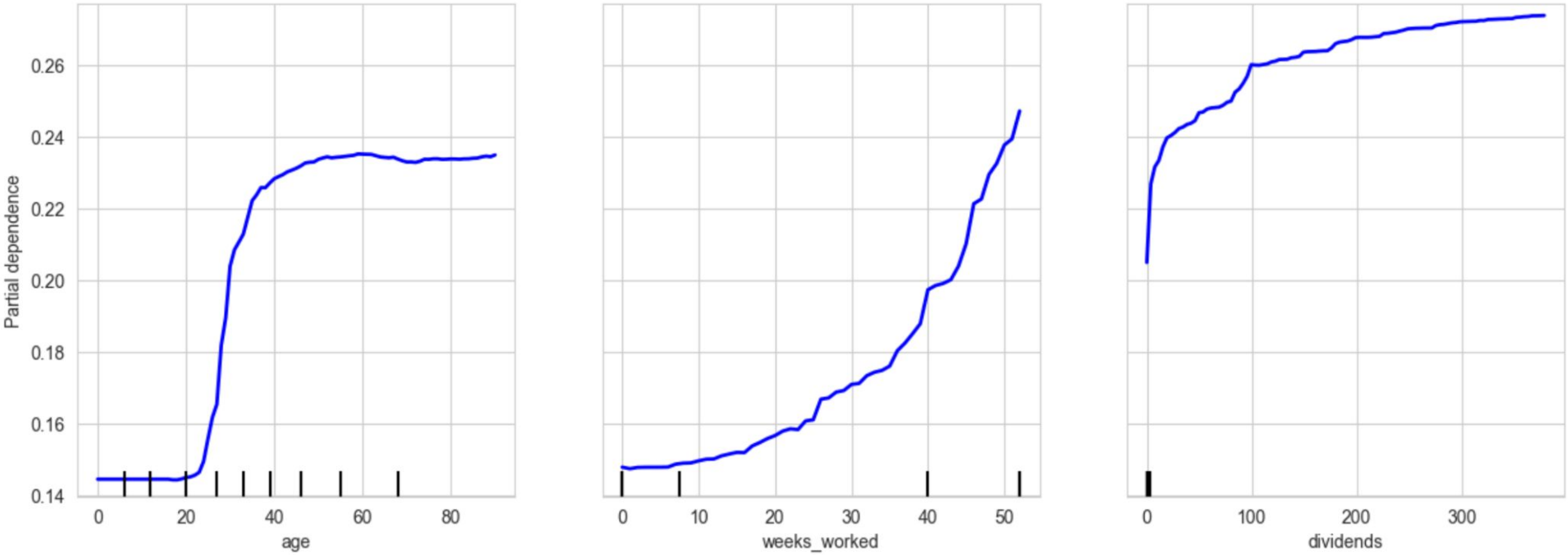
See appendix for logistic regression findings; the remainder of the ML analysis will focus on Random Forest.

Feature importance plot

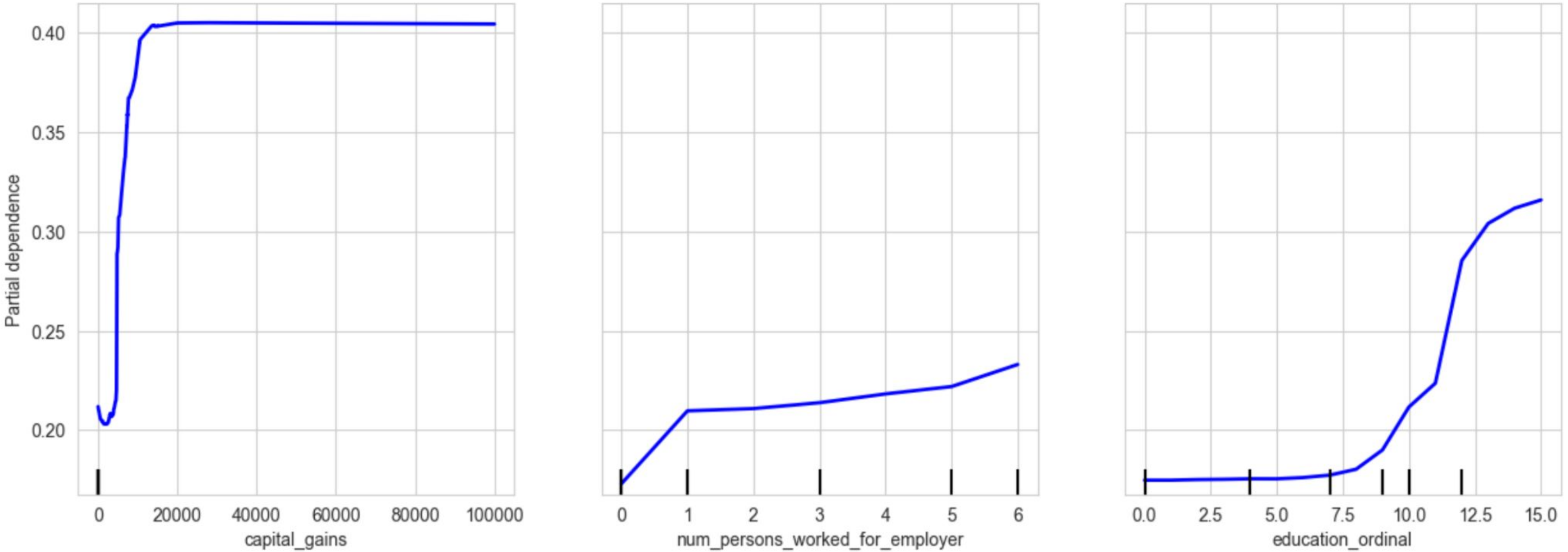


***Interpretation**: The length of each bar in the visualization represents the relative predictive power of that characteristic. The values are normalized to sum to 1.0.*

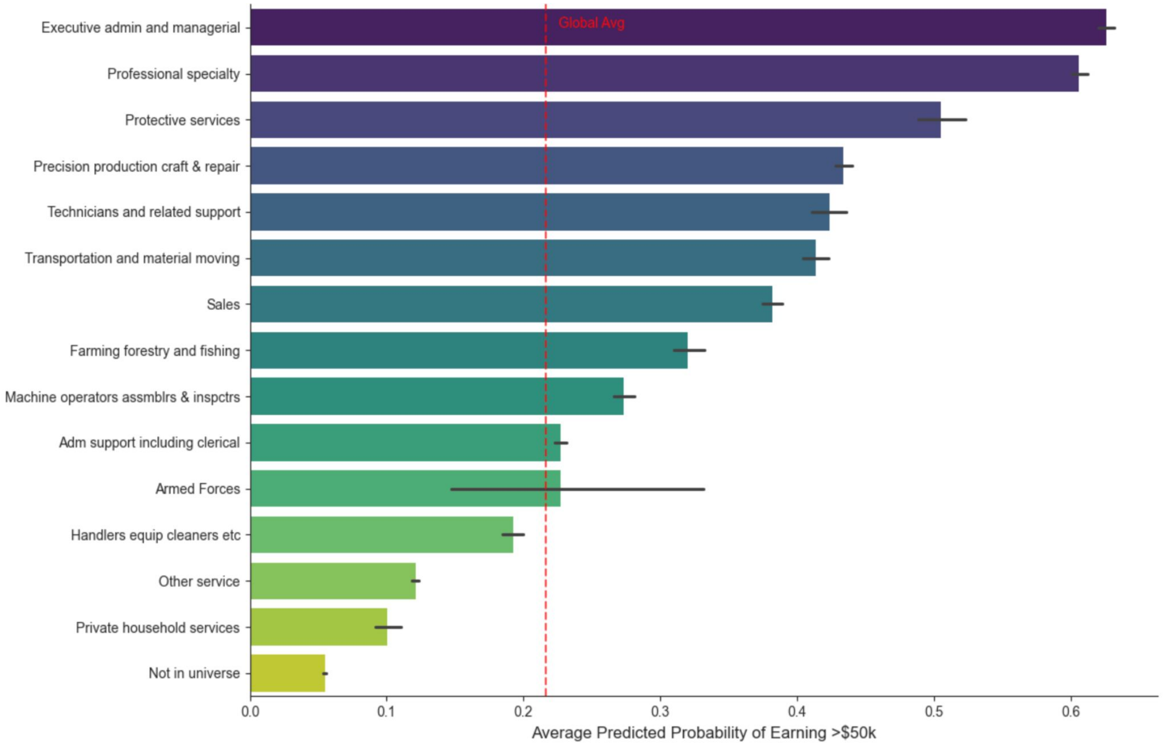
Partial Dependence Plots - numeric features I



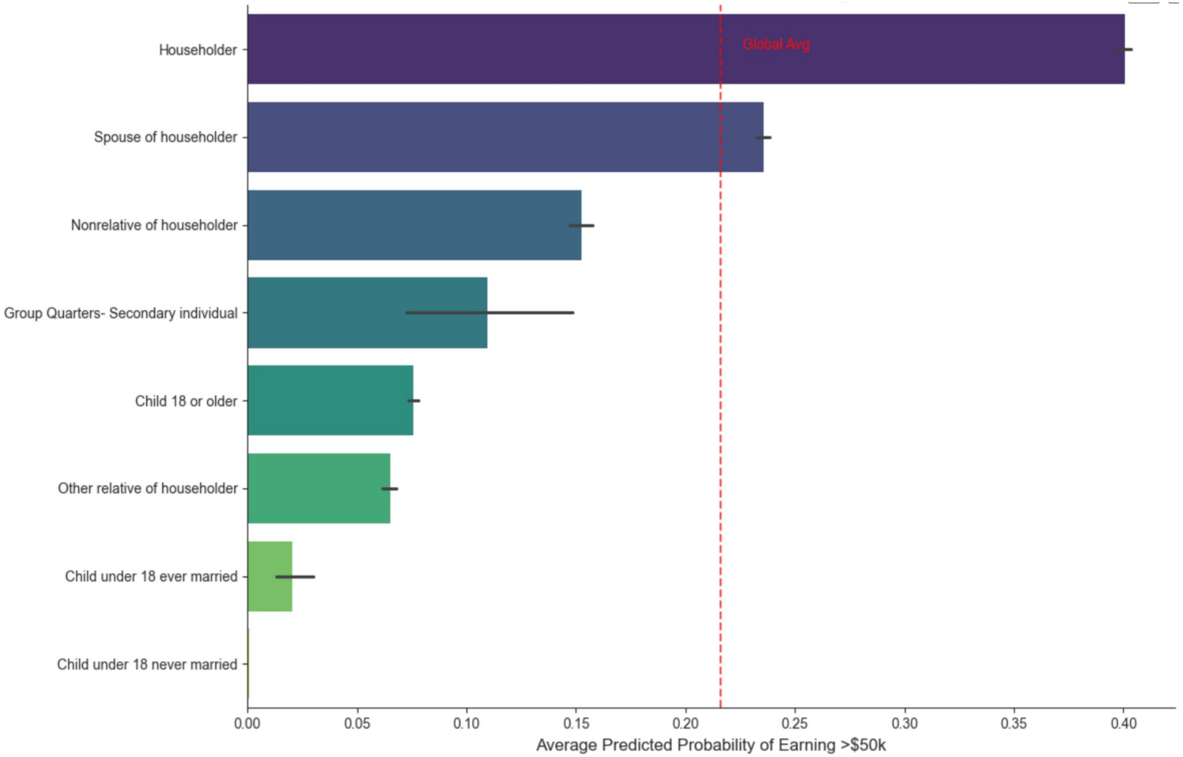
Partial Dependence Plots - numeric features II



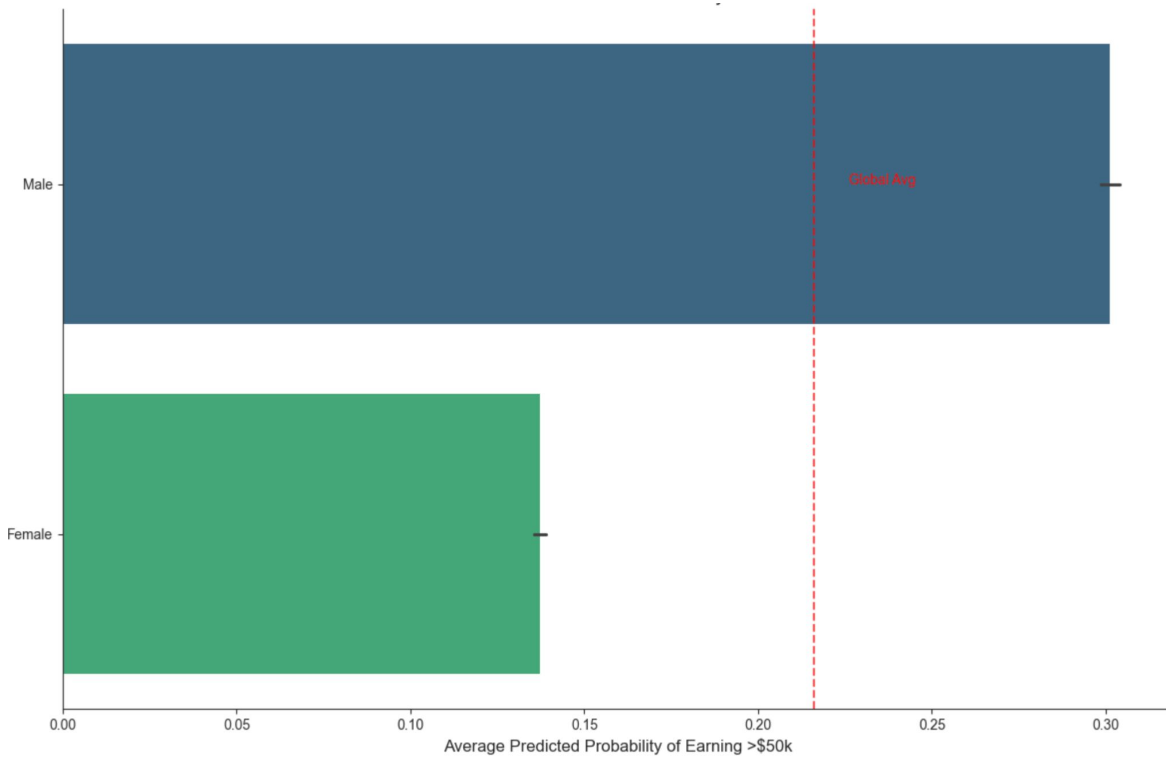
Mean Predicted Probability by occupation



Mean Predicted Probability by household role



Mean Predicted Probability by sex

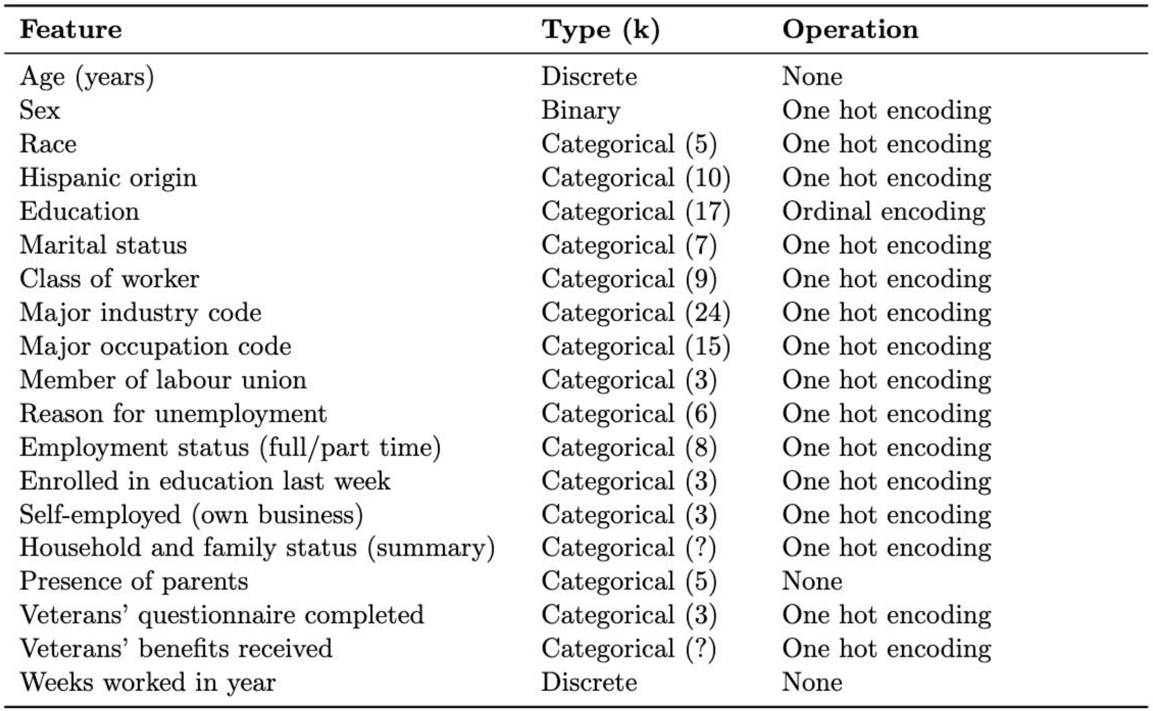


Main findings

- Statistical analysis:
    - Conditional probabilities reveal that individuals reporting capital gains and individuals with a bachelor’s degree or higher have a 25%+ chance of earning >$50,000
    - Hypothesis tests reveal statistically significant distributional differences between income and all features considered
    - Key indicators (in order): education, weeks worked, corresponding firm size, age and dividend yield
    - Surprisingly, hourly wage had a negative correlation with the probability of earning >$50,000 → most likely because high earners do not receive/report an "hourly wage"
- Machine learning analysis:
    - Key indicators we are all numeric; specifically they were: education, weeks worked, age, dividend yield and corresponding firm size
    - Age, dividend yields and capital gains have a logarithmic relationship with the probability of earning >$50,000
    - Being a householder, a male or active in executive/admin/managerial occupations significantly increases the probability of earning >$50,000

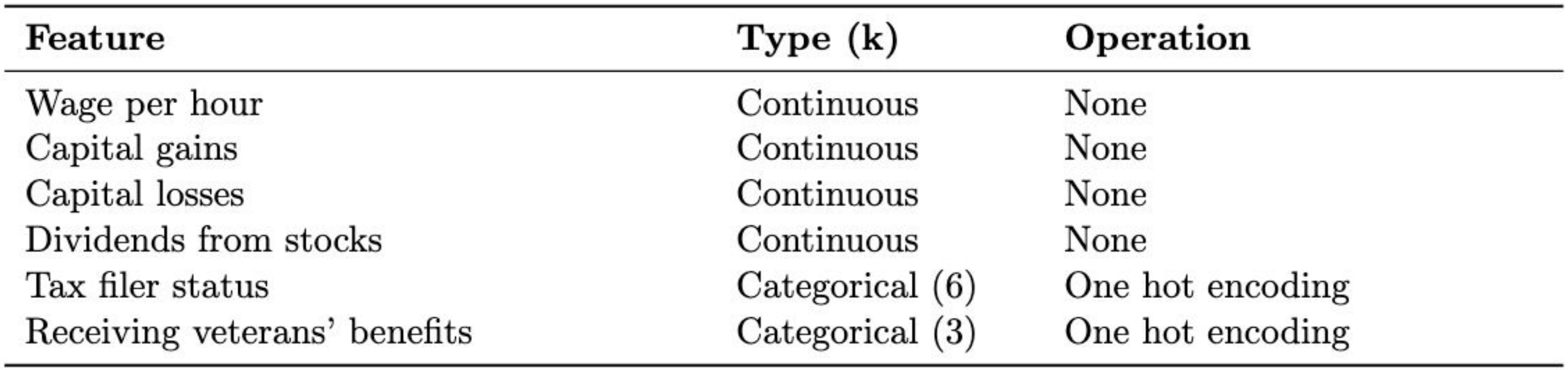
Appendix

Demographic, labour market and household features



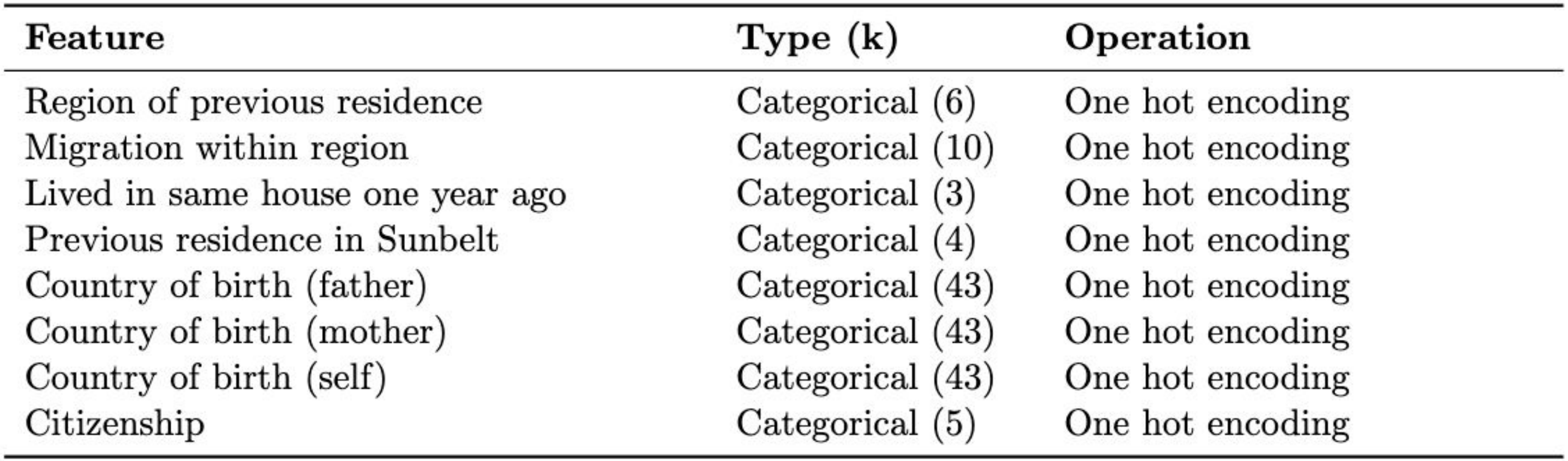
Feature	Type (k)	Operation
Age (years)	Discrete	None
Sex	Binary	One hot encoding
Race	Categorical (5)	One hot encoding
Hispanic origin	Categorical (10)	One hot encoding
Education	Categorical (17)	Ordinal encoding
Marital status	Categorical (7)	One hot encoding
Class of worker	Categorical (9)	One hot encoding
Major industry code	Categorical (24)	One hot encoding
Major occupation code	Categorical (15)	One hot encoding
Member of labour union	Categorical (3)	One hot encoding
Reason for unemployment	Categorical (6)	One hot encoding
Employment status (full/part time)	Categorical (8)	One hot encoding
Enrolled in education last week	Categorical (3)	One hot encoding
Self-employed (own business)	Categorical (3)	One hot encoding
Household and family status (summary)	Categorical (?)	One hot encoding
Presence of parents	Categorical (5)	None
Veterans' questionnaire completed	Categorical (3)	One hot encoding
Veterans' benefits received	Categorical (?)	One hot encoding
Weeks worked in year	Discrete	None

Financial features



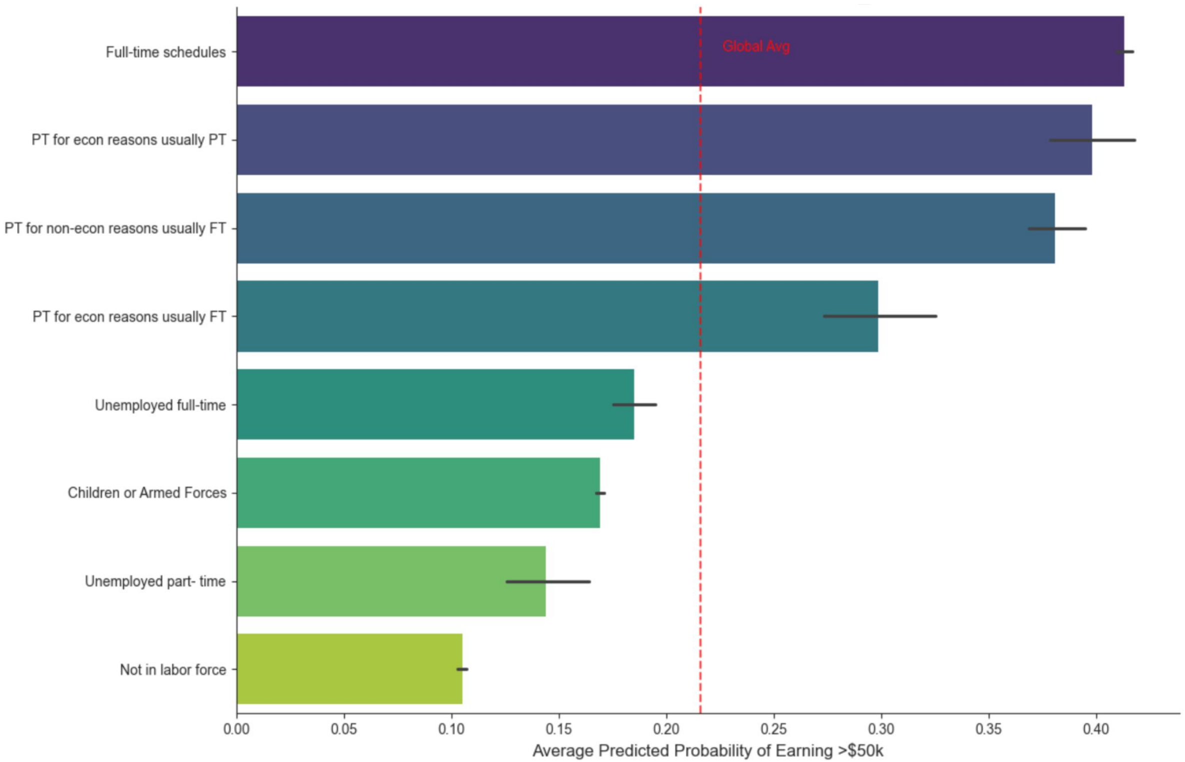
Feature	Type (k)	Operation
Wage per hour	Continuous	None
Capital gains	Continuous	None
Capital losses	Continuous	None
Dividends from stocks	Continuous	None
Tax filer status	Categorical (6)	One hot encoding
Receiving veterans' benefits	Categorical (3)	One hot encoding

Geographic and migration Features

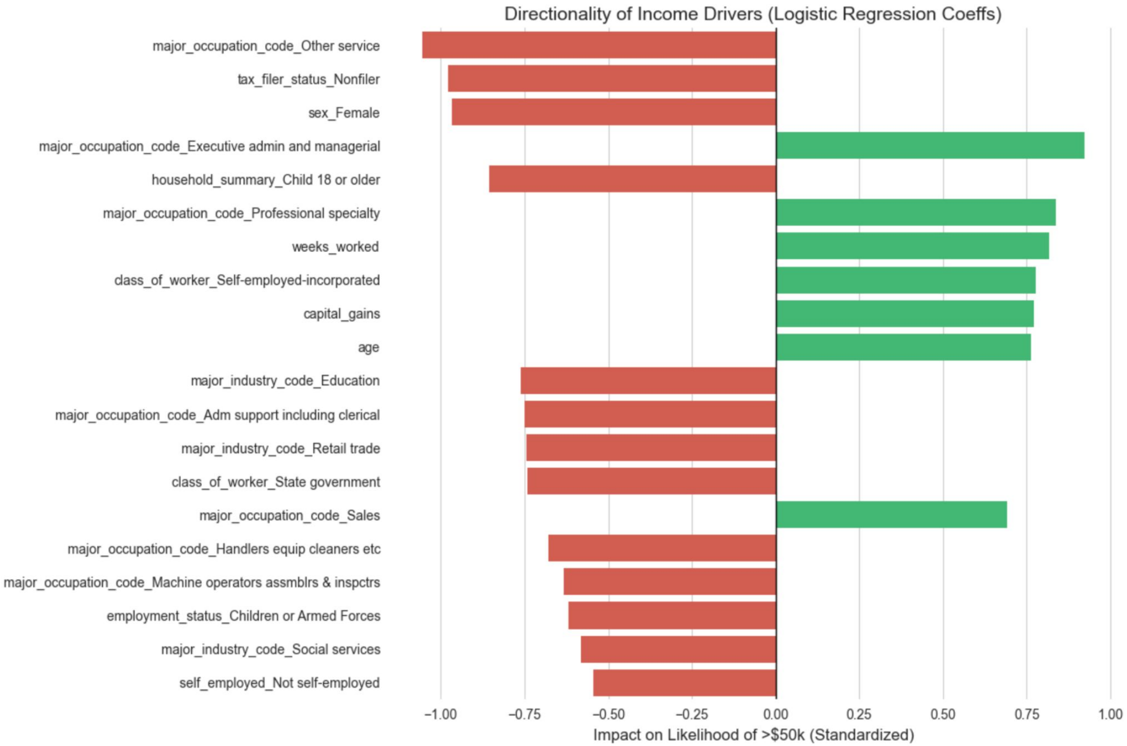


Feature	**Type (k)**	**Operation**
Region of previous residence	Categorical (6)	One hot encoding
Migration within region	Categorical (10)	One hot encoding
Lived in same house one year ago	Categorical (3)	One hot encoding
Previous residence in Sunbelt	Categorical (4)	One hot encoding
Country of birth (father)	Categorical (43)	One hot encoding
Country of birth (mother)	Categorical (43)	One hot encoding
Country of birth (self)	Categorical (43)	One hot encoding
Citizenship	Categorical (5)	One hot encoding

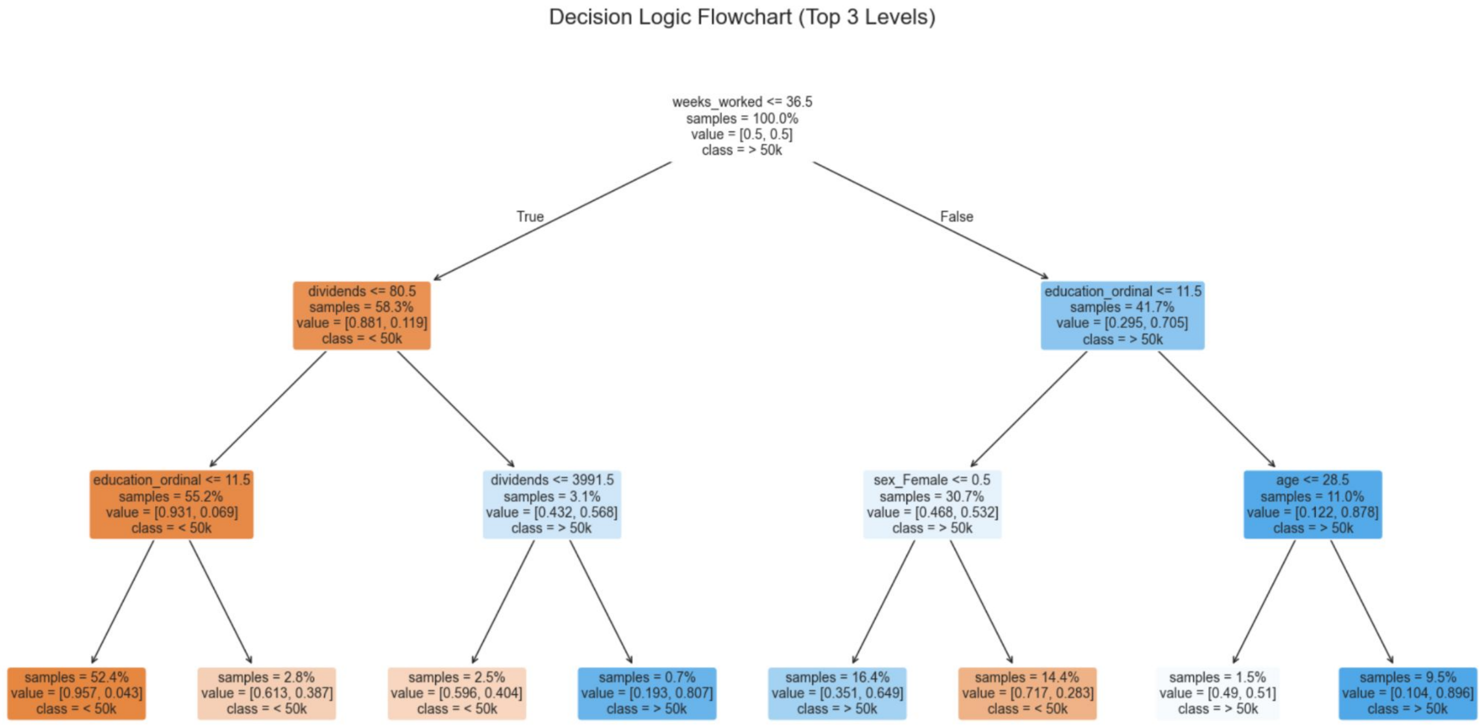
Mean Predicted Probability by employment status



Logistic regression - key features



Understanding random forest (single tree)



Decision Logic Flowchart (Top 3 Levels)

weeks_worked <= 36.5
samples = 100.0%
value = [0.5, 0.5]
class = > 50k

True

False

dividends <= 80.5
samples = 58.3%
value = [0.881, 0.119]
class = < 50k

education_ordinal <= 11.5
samples = 41.7%
value = [0.295, 0.705]
class = > 50k

education_ordinal <= 11.5
samples = 55.2%
value = [0.931, 0.069]
class = < 50k

dividends <= 3991.5
samples = 3.1%
value = [0.432, 0.568]
class = > 50k

sex_Female <= 0.5
samples = 30.7%
value = [0.468, 0.532]
class = > 50k

age <= 28.5
samples = 11.0%
value = [0.122, 0.878]
class = > 50k

samples = 52.4%
value = [0.957, 0.043]
class = < 50k

samples = 2.8%
value = [0.613, 0.387]
class = < 50k

samples = 2.5%
value = [0.596, 0.404]
class = < 50k

samples = 0.7%
value = [0.193, 0.807]
class = > 50k

samples = 16.4%
value = [0.351, 0.649]
class = > 50k

samples = 14.4%
value = [0.717, 0.283]
class = < 50k

samples = 1.5%
value = [0.49, 0.51]
class = > 50k

samples = 9.5%
value = [0.104, 0.896]
class = > 50k