

Assessment documentation

Contents

1	Data preparation and cleaning	2
2	Exploratory data analysis	3
2.1	Sample characteristics by income group	5
2.2	conditional probability estimation	5
3	Hypothesis tests (bivariate associations)	6
4	Machine learning analysis	9
4.1	Random forest	9
4.2	Logistic regression	13

Income Classification: Data and Task

- **Raw data:** Individual-level microdata from the 1994–95 US Census.
- **Target variable:** Binary indicator for whether annual income exceeds \$50,000.
- **Sample size:** $N = 299,283$ individuals; positive class: 6%.
- **Feature set:** 41 predictors.
- **Objective:** “Identifying characteristics that are associated with a person making more or less than \$50,000 per year” \rightarrow binary classification problem.

1 Data preparation and cleaning

To preprocess the data, all column names were first mapped with reference to relevant metadata, and coded categorical variables were decoded into interpretable labels. The training and testing files were merged into a single dataset to allow for unified exploratory analysis. Finally, a series of diagnostic checks confirmed that the merged data contained no missing values, no duplicated rows, and no negative values in any continuous variable.

Several additional transformations were considered but ultimately deemed unnecessary or counterproductive. Winsorising the continuous variables at the 99th percentile and scaling them to the $[0,1]$ range was explored, but this was rejected because extreme values in variables such as wage, capital gains, and dividends carry substantive information relevant to income prediction. Harmonisation of categorical variables, rule-based imputation, and standardisation of high-cardinality categories were also considered, but the dataset was already fully encoded. Finally, although the metadata mentions that there are approximately 67,652 “duplicates or conflicting instances” in the data, they cannot be identified since household IDs are not provided. No rows or columns are therefore removed at this stage.

2 Exploratory data analysis

To get a feel for the data, I first explore the distribution of features for all respondents. Some noteworthy plots that should help the reader understand the demographic composition slightly better are presented below.

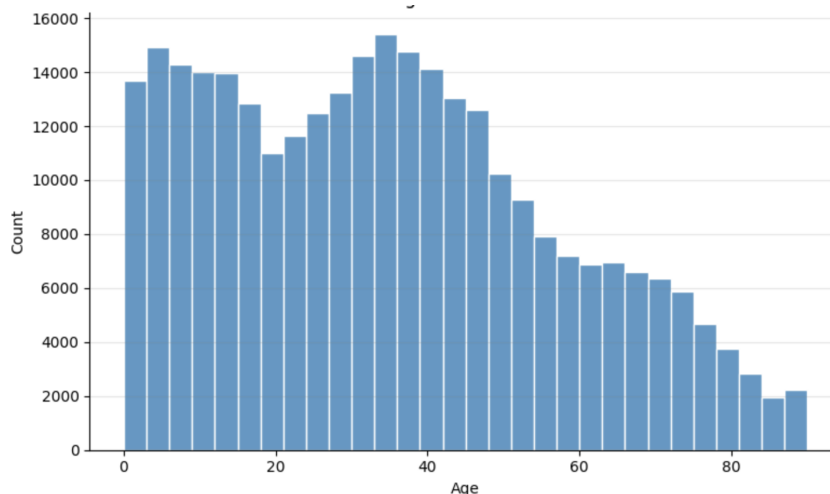


Figure 1: **Age distribution.** Histogram of the age distribution of all respondents with 30 bins, each spanning three years.

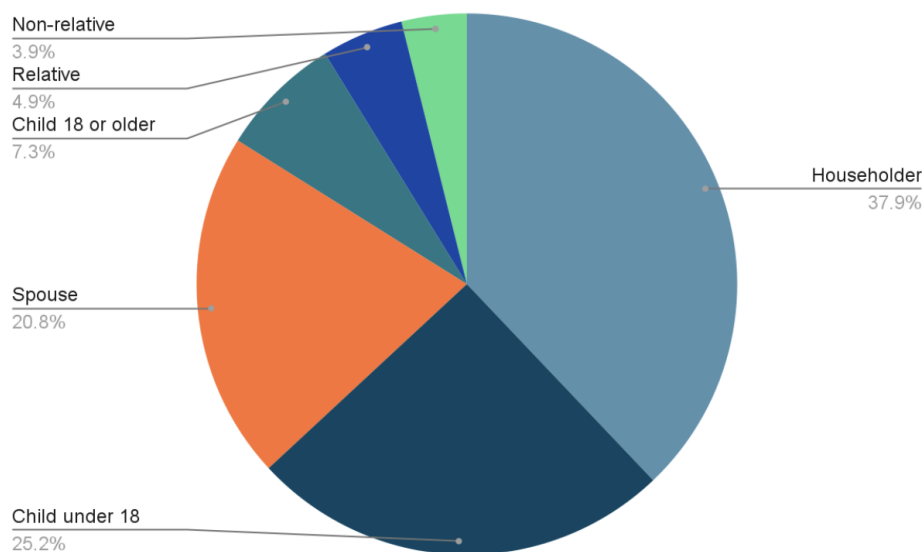


Figure 2: **Household composition.** Household composition among respondents. The categories “child under 18 never married” and “child under 18 ever married” were combined into “Child under 18”, and the categories “Nonrelative of householder” and “Group Quarters - Secondary individual” were combined into “Non-relative”.

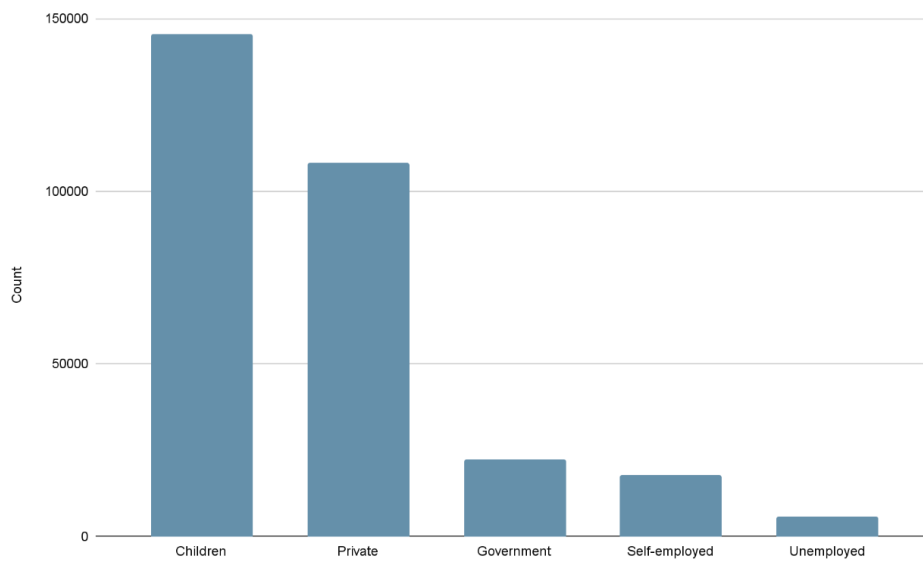


Figure 3: **Employment status distribution.** This information was primarily derived from the “class_of_worker” column with parts from the “employment_status” column. Individuals in the “Self-employed-not incorporated” and the “Self-employed-incorporated” were grouped, as well as individuals in the Federal government, State government and Local government categories. Additionally, those “Without pay” (0.08%) and those who “Never worked” (0.02%) were placed into the unemployed category.

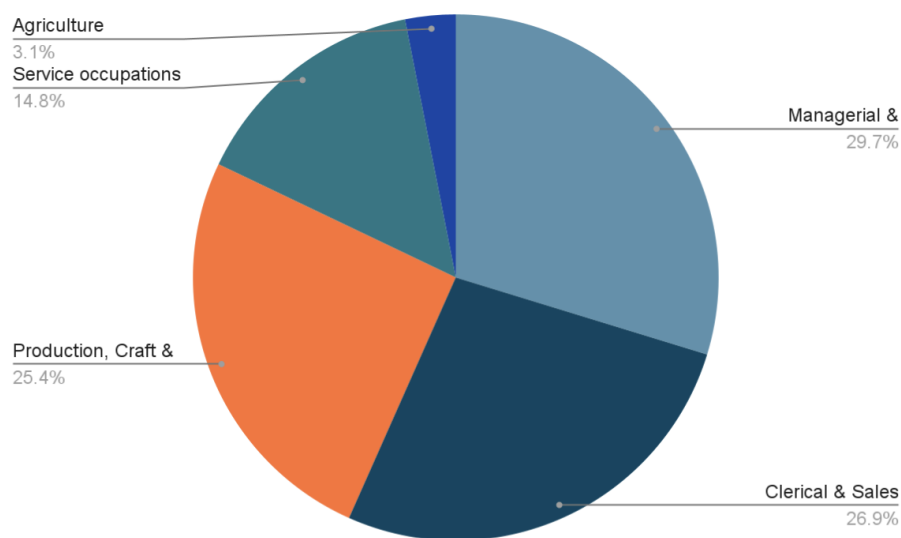


Figure 4: **Occupation distribution (grouped).** The five occupation groups shown here were created by aggregating occupations from fifteen more granular categories (see code for details). This chart therefore only includes the roughly 49% of survey respondents who were employed.

Table 1: Summary statistics

Statistic	Value
Total respondents	299,283
Average age	34.5
Male share	48%
Income above \$50,000	6%

2.1 Sample characteristics by income group

Sample characteristics at the group level provide valuable and easily interpretable insights into the structure of our data. To assess these I first convert features that I believe are most “indicative” into booleans (i.e., variables that flag whether an individual meets a particular condition. I then compare the proportions of individuals meeting each condition across income groups. These descriptive comparisons help answer:

“How do the demographic, educational, and financial attributes of high earners differ from those of lower earners?”

Table 2: Sample characteristics by annual income group

	Annual income	
	$\leq \$50,000$	$> \$50,000$
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

2.2 conditional probability estimation

Next, I examine conditional probabilities to measure how frequently a higher-income outcome occurs within a defined subgroup. This helps answer questions such as:

“What share of all individuals who have a specific trait also earn more than \$50,000 per year.”

Table 3: Conditional probability 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

3 Hypothesis tests (bivariate associations)

Next I conduct hypothesis tests. These are useful because they provide a systematic way to determine whether observed differences between income groups reflect genuine underlying patterns rather than random variation within the sample. To this end, I first ordinally encode the level of education (see code for exact mapping) under the assumption that there exists a linear relationship between this feature and income.

To assess the relationship between the numeric variables and the income predictor, I use a Mann–Whitney U test, followed by Cliff’s delta to complement the results by quantifying the magnitude of the difference between groups. A non-parametric effect size such as Cliff’s delta is preferred over parametric correlations because many of the variables (e.g., wage, capital gains) are skewed and do not satisfy normality assumptions. Values range from -1 to 1 , with 0 indicating no difference and larger absolute values signalling stronger effects. It complements the Mann–Whitney U test by focusing on effect magnitude rather than significance alone. As a robustness check I also compute point-biserial correlations; the results are nearly identical, indicating that the findings are not sensitive to the chosen method.

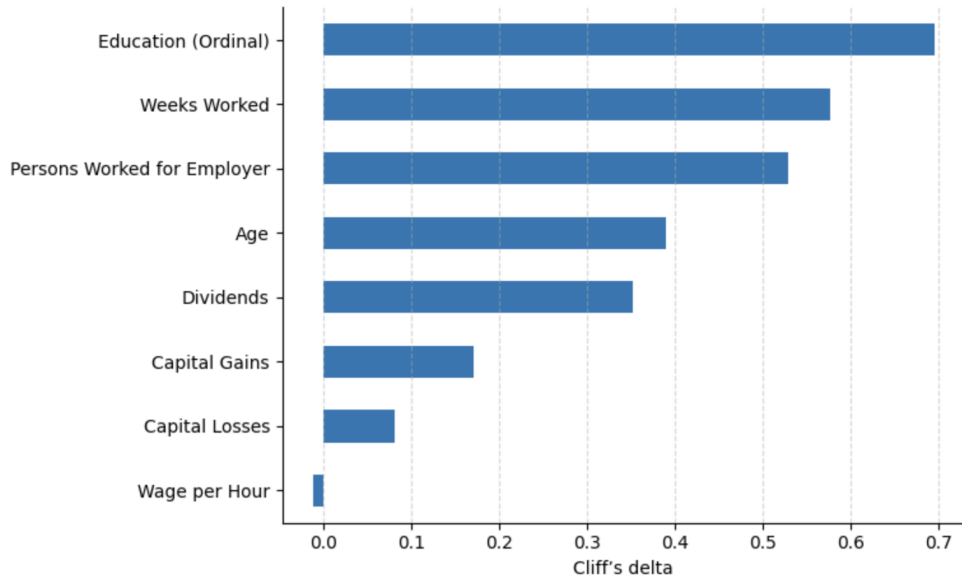


Figure 5: **Cliff’s delta values** Cliff’s delta values for each numeric predictor, showing the magnitude and direction of group differences between high- and low-income individuals. All features show statistically significant group differences with $p < 10^{-10}$ returned for the Mann–Whitney U test. See code for exact values.

To assess the statistical significance of our categorical variables as predictors of income, I conduct a Chi-square test of independence and compute Cramer’s V. Together, these indicate whether a categorical feature is associated with income status and, if so, how strong that association is. Cramer’s V complements the Chi-square test in the same way as Cliff’s delta for the Mann Whitney U test, but it ranges from 0 to 1.

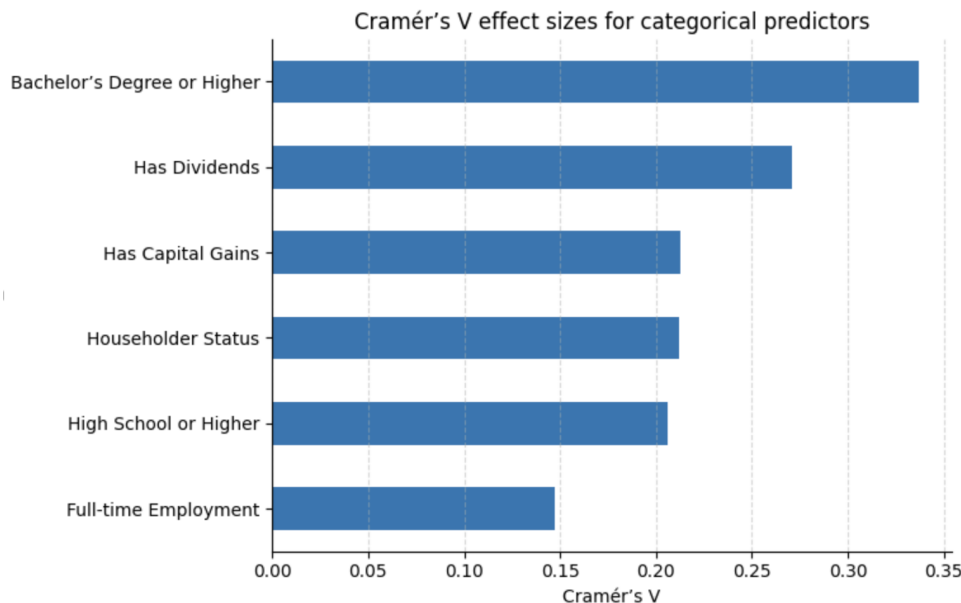


Figure 6: **Cramer’s V effect size.** Values for each numeric predictor, showing the magnitude and direction of group differences between high- and low-income individuals. All features show statistically significant group differences with $p < 10^{-10}$ returned for the Chi-squared test. See code for exact values.

Feature selection and engineering

Eight of the original 41 features were removed for several reasons. First, instance weights, which represent survey sampling weights used to recover population-level estimates, and the year of collection were dropped. Note that weights could be incorporated later by applying them during model training through weighted loss functions. Second, four variables appeared twice in the data, once in a detailed version and once in a higher-level summary. For example, a detailed industry code with 52 categories was provided alongside a less granular version with 24 categories. Since I will one-hot encode the feature, I retain the less granular summary version for all four such cases. The `state_prev_residence` variable is a particularly problematic example because it contains 52 distinct values but applies only to the small share of respondents who reported moving. Third, I remove two of three migration variables that exhibit substantial multicollinearity. Specifically, `migration_msa`, `migration_reg`, and `migration_within_reg` encode very similar moves (e.g., “MSA to MSA”, “same county”, “same county”), so only one is retained.

Finally, I considered grouping rare categories for some categorical variables to avoid excessively sparse dummy matrices. Additionally, several discrete financial variables (most notably dividends and capital gains) are highly skewed, with more than 90% of respondents reporting zeros. These could, in principle, be binarised (e.g., “any dividends” vs “none”) to reduce sparsity and improve model stability. A complete overview of all retained variables, their types, and the corresponding encoding procedures is provided in Tables 4, 5, and 6.

Table 4: 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

Table 5: 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

Table 6: 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

4 Machine learning analysis

Machine learning approaches provide a useful extension to the previous analysis, which focused on average, marginal, and mostly linear relationships. Its primary advantages for identifying characteristics associated with the target variable are:

- non-linear relationships between features and the target variable;
- interaction effects between multiple features (e.g., how receiving veterans' benefits and being privately employed jointly affect the probability of earning more than \$50,000);
- out-of-sample predictive performance to assess whether patterns generalise.

I implement two models. First, a Random Forest, which performs well on tabular survey data, provides model-level interpretability, and is computationally efficient. Second, a logistic regression, which I include as an interpretable linear baseline to benchmark the Random Forest.

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

Performance is not the primary objective here, but it is a relevant indicator of how reliable the proposed models are in their output (i.e., the features they deem most indicative of high income). Both models are trained to optimise balanced accuracy, since there is significant class imbalance in the target variable to the point where predicting "\$5,000 would yield 96% accuracy.

4.1 Random forest

The Random Forest model is an ensemble learning method that constructs many decision trees and aggregates their predictions, reducing overfitting and improving stability. It is well suited to tabular survey data because it can naturally capture complex non-linear relationships and interactions between variables without requiring strong functional form assumptions. Moreover, it handles mixed data types, is robust to outliers, and performs well even when features vary in scale. These properties make Random Forests appropriate for the objective of identifying patterns in heterogeneous household characteristics while prioritising predictive accuracy and generalisability.

For the Random Forest model, hyperparameters are tuned via grid search, including the number of trees (`n_estimators`), maximum tree depth (`max_depth`), minimum samples per leaf (`min_samples_leaf`), and the number of features considered when splitting (`max_features`). These parameters are selected using three-fold cross-validation on the training set by optimising balanced accuracy. Interpretability is ensured through feature importance plots and partial dependence plots, which illustrate the contribution and marginal effects of key predictors.

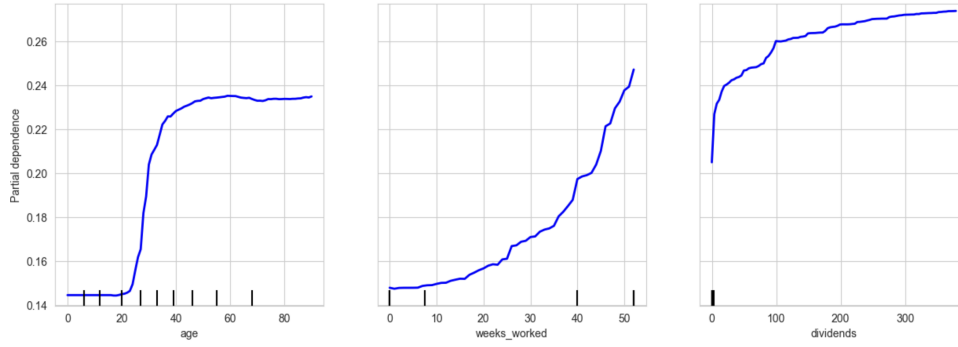


Figure 7: **Partial Dependence Plots for additional continuous features.** PDPs showing how age, weeks worked, and dividends influence the predicted probability of earning over \$50,000.

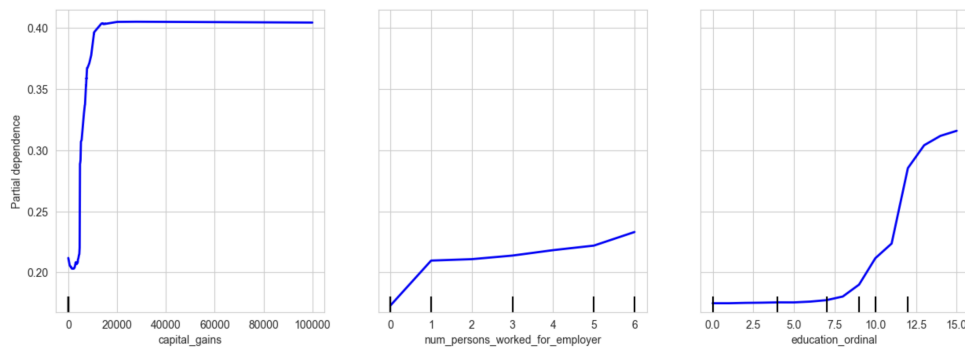


Figure 8: **Partial Dependence Plots for key continuous features.** PDPs showing how capital gains, firm size, and education ordinal influence the predicted probability of earning over \$50,000.

The numeric features show strong and often non-linear associations with the probability of earning more than \$50,000. Age, weeks worked, education, dividends, capital gains, and firm size are all influential predictors, with the partial-dependence plots above indicating that many of these effects follow logarithmic or diminishing-returns patterns. For example, the probability of high income increases steeply with full-year employment and early-career age increases, before flattening later in life. Similarly, even small amounts of dividends or capital gains sharply raise predicted income probabilities, with the marginal contribution tapering off for larger values. Education exhibits a comparable pattern: lower levels of schooling contribute little, but the predicted probability rises rapidly once post-secondary qualifications are reached. These non-linearities underscore the value of machine-learning methods, which are well suited to capturing threshold effects that linear models would understate.

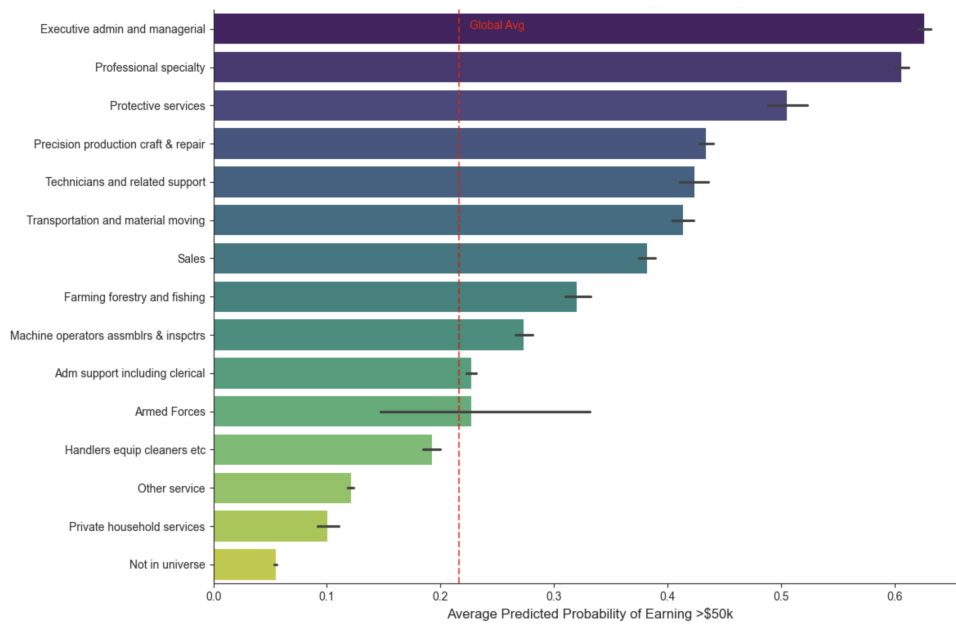


Figure 9: **Conditional Average Predictions by major occupation.** Average predicted probability of earning over \$50,000 across major occupation groups, with error bars and the global average marked.

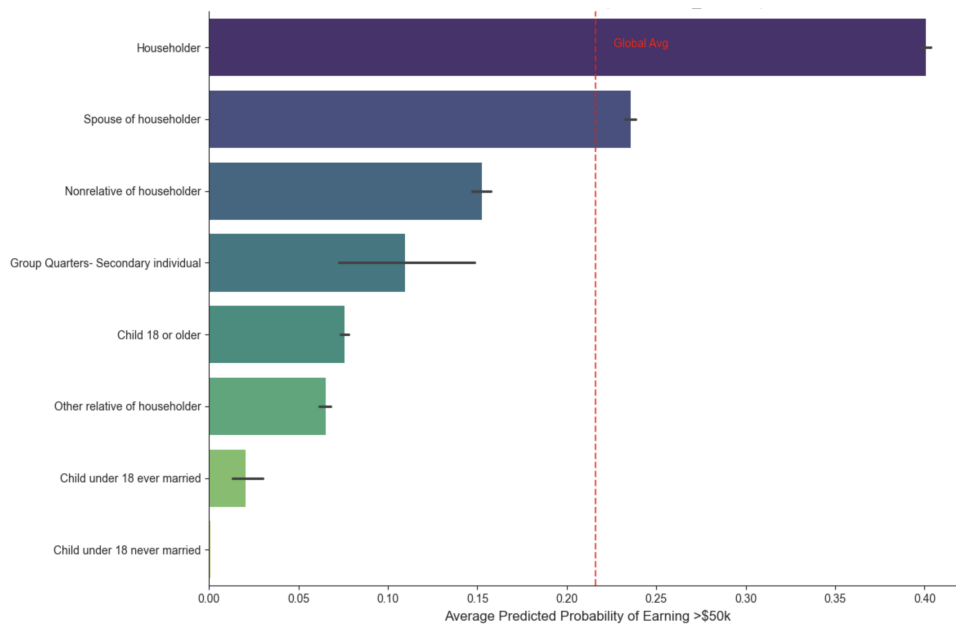


Figure 10: **Conditional Average Predictions by household role.** Average predicted probability of earning over \$50,000 across household categories, with error bars and the global average marked.

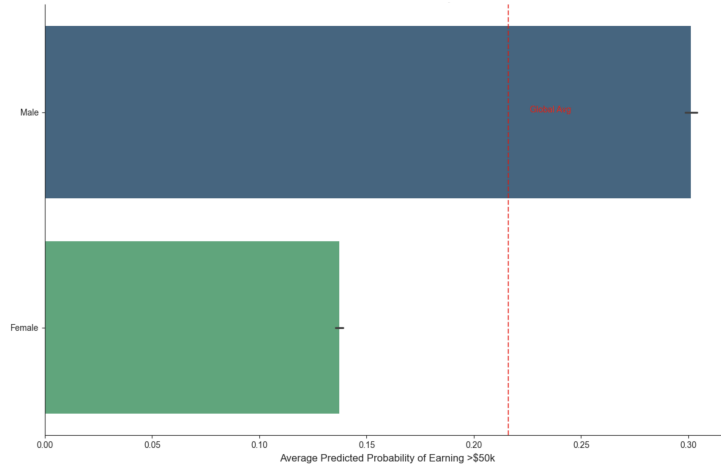


Figure 11: **Conditional Average Predictions by sex.** Average predicted probability of earning over \$50,000 for males and females, with error bars and the global average marked.

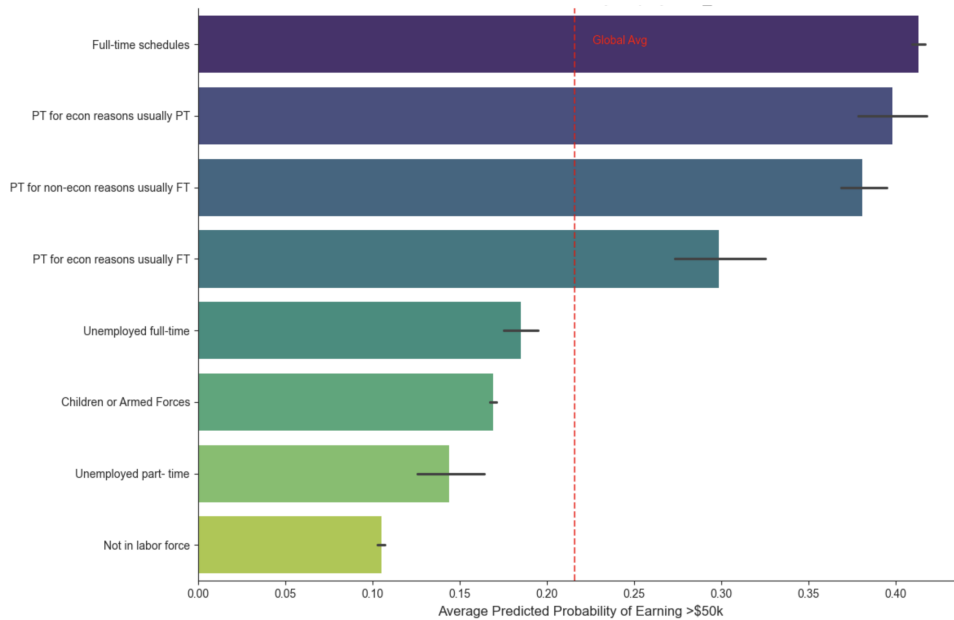


Figure 12: **Conditional Average Predictions by employment status.** Average predicted probability of earning over \$50,000 across employment status groups, with error bars and the global average marked.

The categorical variables highlight structural and demographic factors that strongly segment income outcomes. Being a householder, male, or employed in executive, administrative, or managerial occupations substantially increases the predicted probability of high income, far above the global sample average. Occupation is a particularly strong indicator, with executive and professional groups far more likely to exceed the \$50,000 threshold than service, clerical, or manual roles.

4.2 Logistic regression

Logistic regressions model the probability of a binary outcome by applying the logistic function to a linear combination of input features. The approach is appropriate for this setting because it provides an interpretable baseline model that captures linear and additive relationships between covariates and the probability of belonging to the target class. This allows for a quantification of the direction and relative magnitude of associations through the sign and size of the estimated coefficients.

In my implementation, all continuous variables were standardised using Z-score normalisation to ensure that coefficients are comparable and that the optimisation procedure behaves well. The model is trained to maximise balanced accuracy, enabling a fair comparison with the Random Forest in the presence of class imbalance. Interpretability is achieved via a coefficient plot, which displays the estimated effects of each feature on the log-odds of the outcome.

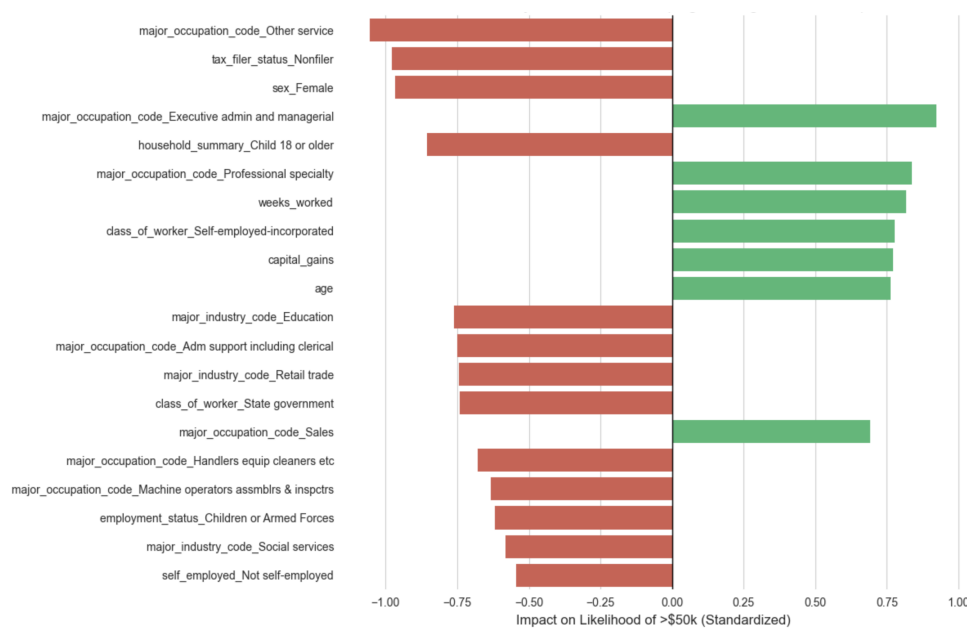


Figure 13: **Logistic regression coefficients.** These indicate the direction and strength of predictors for earning over \$50,000.

In alignment with previous findings, the plot shows that certain occupations and financial indicators strongly increase the probability of earning more than \$50,000. Specifically, executive and professional roles, being self-employed in an incorporated business, higher weeks worked, capital gains and age all have clear positive effects. By contrast, being female, filing as a non-filer and working in service, clerical, retail or manual occupations reduce the likelihood of high income.