Assessing important features in predicting U.S. income level

Yang Chen | August 2025

Agenda

01

Key results

02

Overview

03

Data insights and methodology

04

Results

05

Recommendations

06

Future improvements

01

Key results

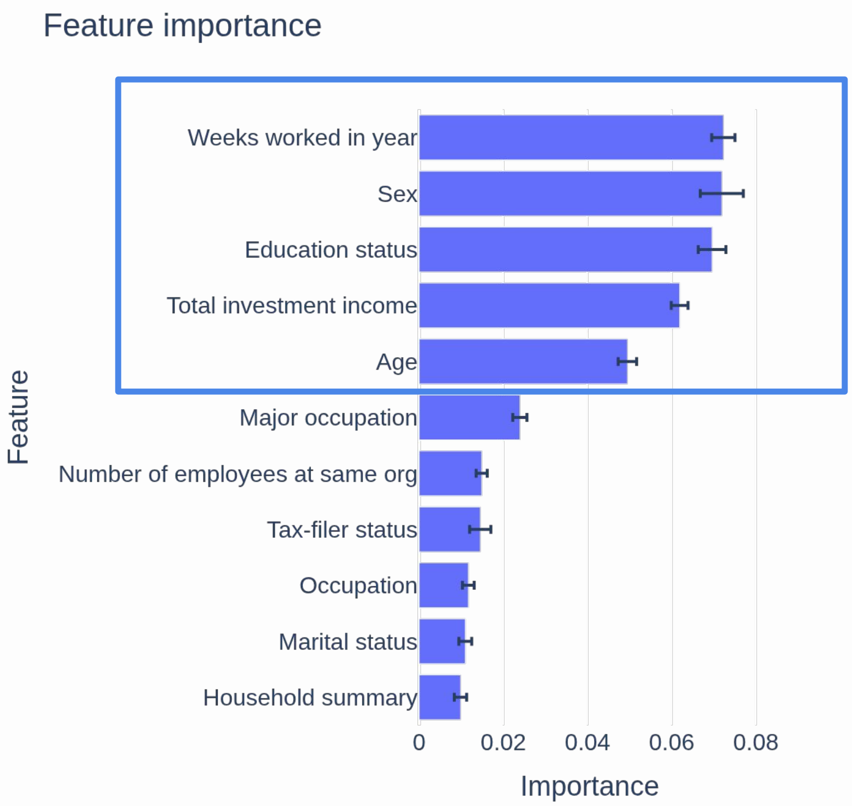
Important features and trends on income level

Top 5 important features

1. Weeks worked in year
2. Sex
3. Education status
4. Total investment income
5. Age

Positive effects predicting income level

Weeks worked in a year and education status have increasing, positive effects on predicting above 50K earners



02

Overview

Analysis goals and data overview

Overview

Goal

Identify important characteristics associated with a person making **more or less than $50,000 per year.**

01

Observations

Sampled 1994-1995 U.S. Census archive data with **~300K individuals**

02

Features

40 features (range of economic, demographic, employment, geographic)

03

Income Level

93.8% in Below $50K
6.2% in Above $50K

Full data

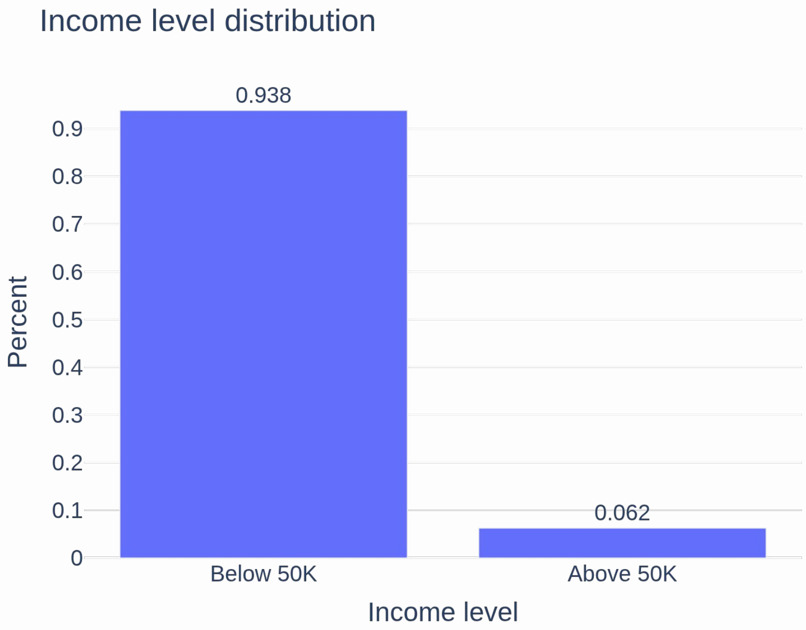
03

Data insights and methodology

Analytical trends, data preparation, and modeling strategy

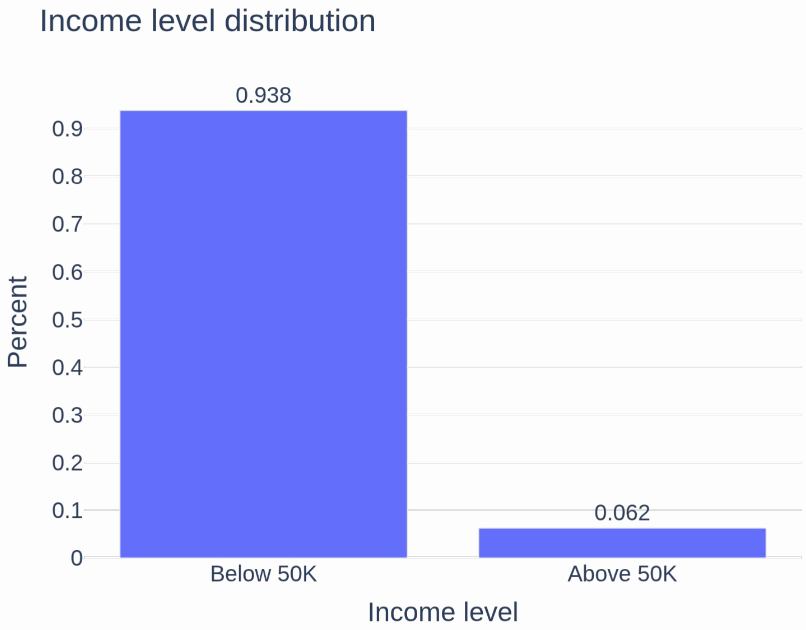
Insight - Income level

- Income split:
  - Below 50K earners represent large majority of the dataset, only 6% are Above 50K



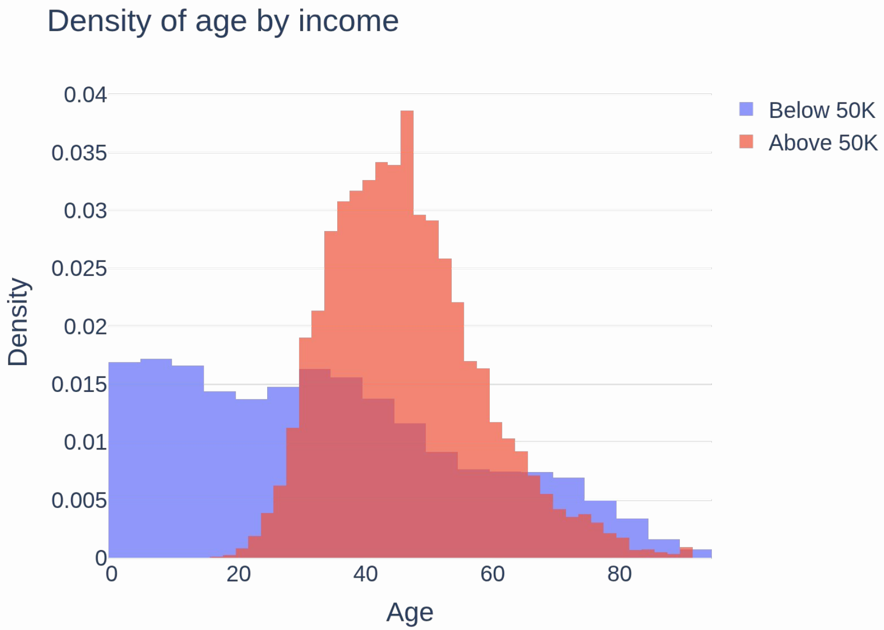
Insight - Income level

- Income split:
  - Below 50K earners represent large majority of the dataset, only 6% are Above 50K
- Technical consideration:
  - Balance training data to improve model predictions on both groups



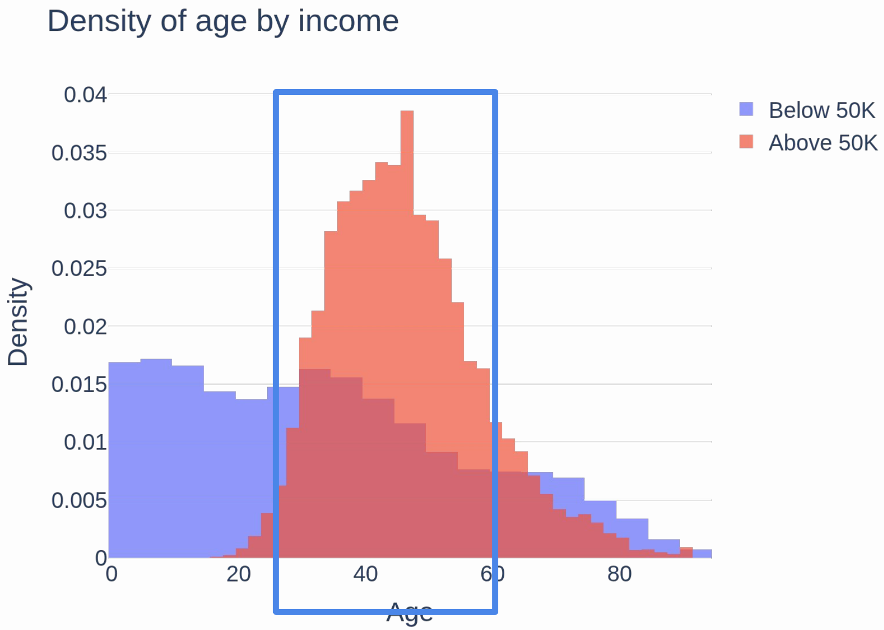
Insight - Age and income level

- Above 50K:
  - Predominantly middle-age (35-50)



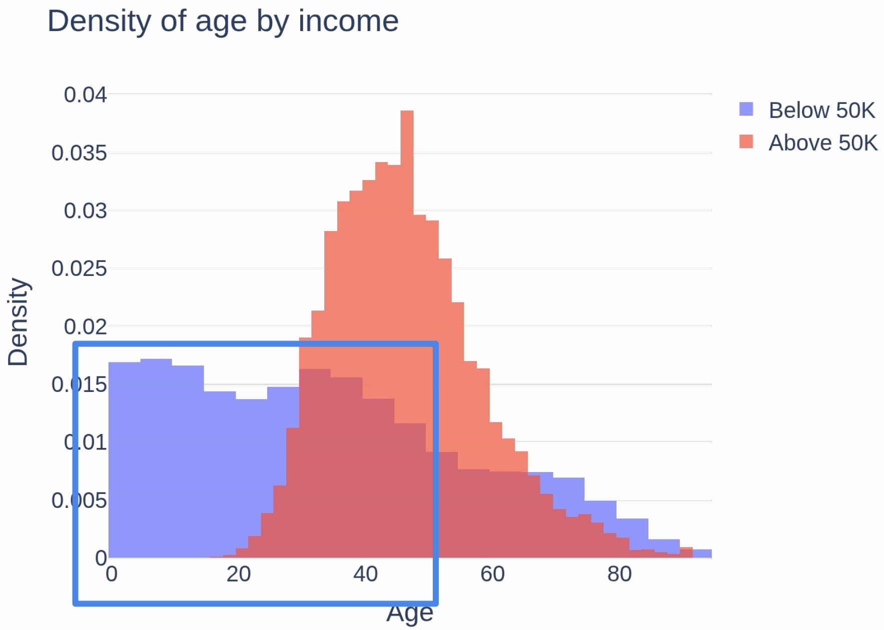
Insight - Age and income level

- Above 50K:
  - Predominantly middle-age (35-50)
  - Significant overlap, but higher earners more common



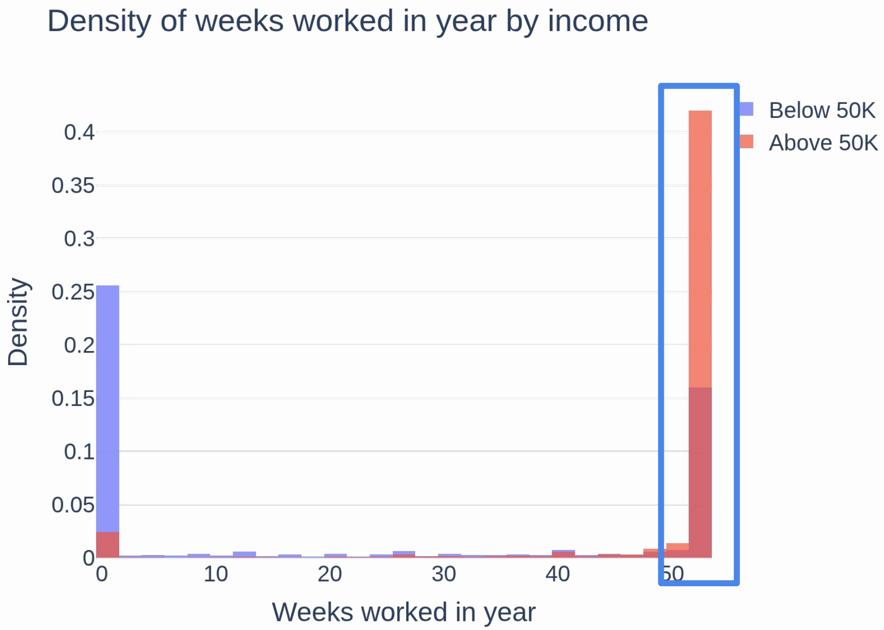
Insight - Age and income level

- Above 50K:
  - Predominantly middle-age (35-50)
  - Significant overlap, but higher earners more common
- Below 50K:
  - Predominantly <40
  - Flatter distribution across younger ages



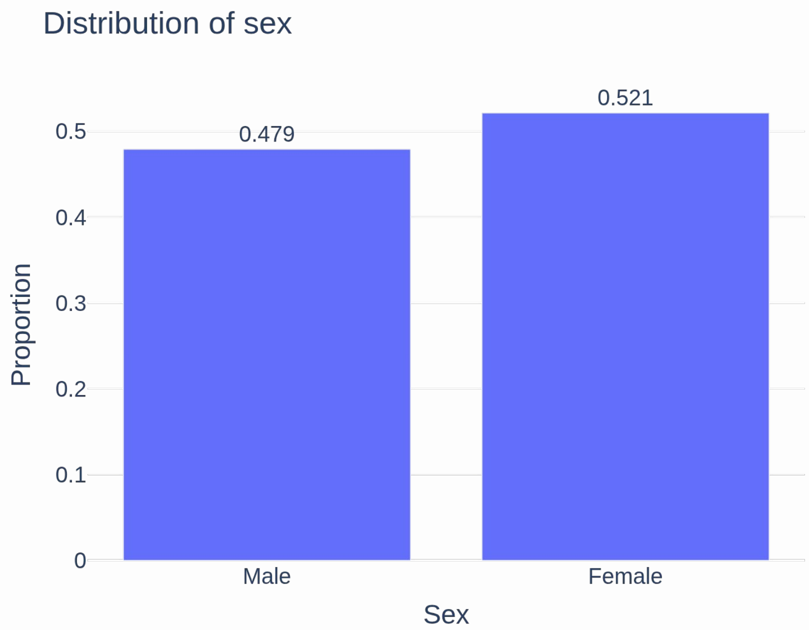
Insight - Weeks worked in year and income level

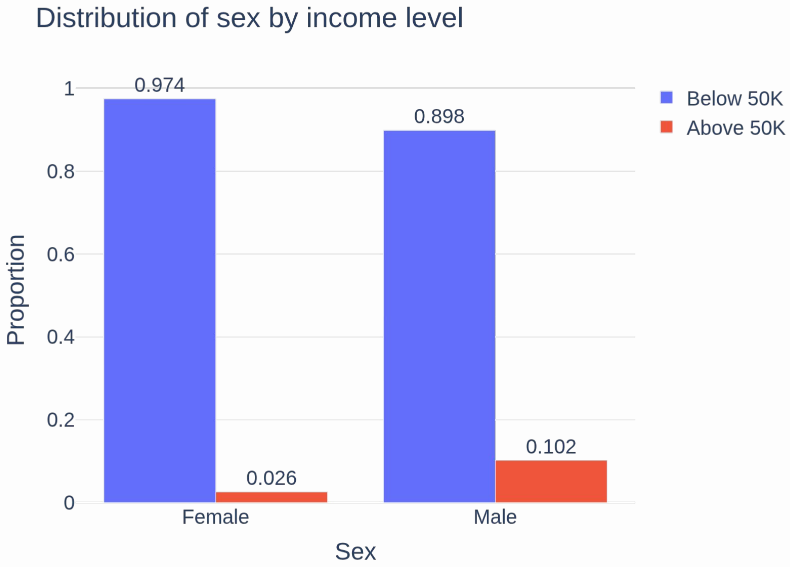
- Above 50K:
  - Concentrated in **full-time work** 52 weeks worked in a year



Insight - Gender and income level

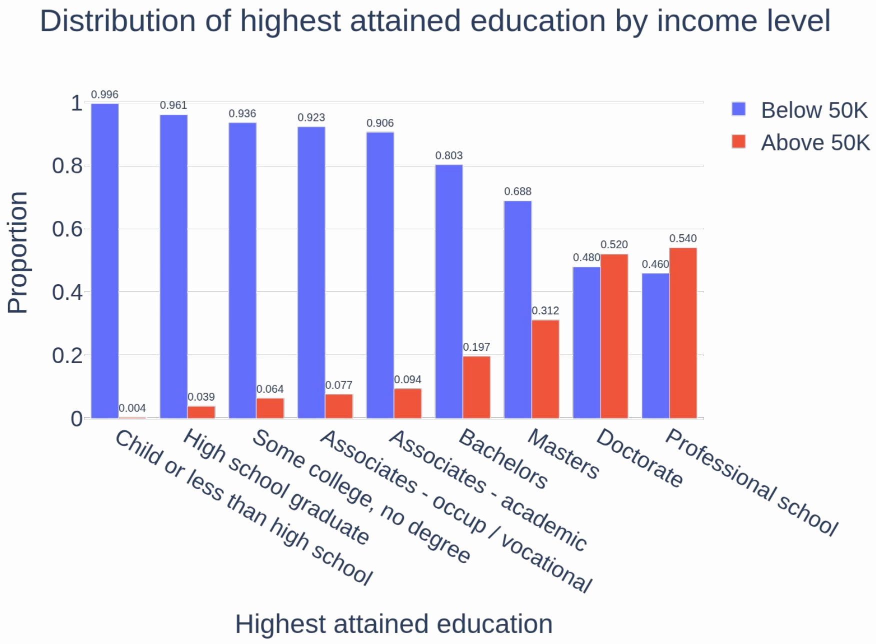
Despite similar overall male-female proportions in the dataset, males earning above $50K are significantly more represented at nearly **5 times** the rate of females in the same income level





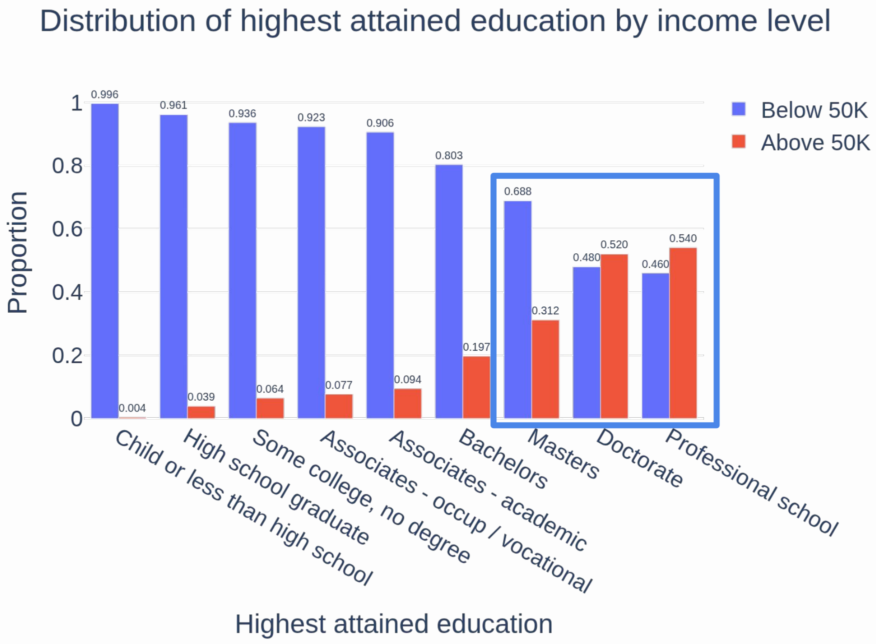
Insight - Education and income level

- Higher education levels show increased representation in Above 50K



Insight - Education and income level

- Higher education levels show increased representation in Above 50K
- Advanced degrees (Masters, Doctorate, Professional) have significantly* higher proportions than prior levels



*: Evaluated using pairwise chi-square tests at 0.01 significance level with multiple test correction

Methodology

Preprocessing	**Model selection**	**Model evaluation**	**Feature importance**

- Migration feature missing values
- Univariate feature selection (correlations, chi-square)
- Separation concerns for baseline modeling
- Feature engineering

Missing values in migration features

- **Data quality issue:** Migration features had missing values for observations where:
  - “Lived in this house 1 year ago” = “Not in universe under 1 year old”
  - Age > 0 for these observations (qualifies for migration questions)

Missing values in migration features

- **Data quality issue:** Migration features had missing values for observations where:
    - "Lived in this house 1 year ago" = "Not in universe under 1 year old"
    - Age > 0 for these observations (qualifies for migration questions)
- **Bias validation:**
    - Compared income level patterns and feature distributions between:
        - Observations with complete migration data
        - Observations with missing migration data

Missing values in migration features

- **Data quality issue:** Migration features had missing values for observations where:
  - "Lived in this house 1 year ago" = "Not in universe under 1 year old"
  - Age > 0 for these observations (qualifies for migration questions)
- **Bias validation:**
  - Compared income level patterns and feature distributions between:
    - Observations with complete migration data
    - Observations with missing migration data
- **Result:**
  - Excluding migration features won't bias our predictions and make feature importance unreliable for recommendations

Missing values in migration features

49.3%

Observations impacted

7

Migration features with all missing values or NIU

Features

40

33

Feature engineering and selection

1

Feature created: total investment income
Collapsed categories in employment stat, country to reduce dimensions

4

Features removed: capital gains, capital losses, stock dividends, household stat

Features

33

30

Methodology

Preprocessing	Model selection	Model evaluation	Feature importance

- Models tested:
    - Logistic regression with regularization (mitigate multicollinearity)
    - Random forest (handle non-linear patterns)

Methodology

Preprocessing	**Model selection**	Model evaluation	Feature importance

- Models tested:
  - Logistic regression with regularization (mitigate multicollinearity)
  - Random forest (handle non-linear patterns)
- Cross validation (CV) assessment
  - Hyperparameter tuning and selecting final model
  - Oversampling during CV to address income level imbalance

Methodology

Preprocessing	Model selection	**Model evaluation**	Feature importance

- Evaluation metrics:
    - **Precision**:
        - Of all the times the model predicted Above 50K, how often was it correct?

Methodology

Preprocessing	Model selection	**Model evaluation**	Feature importance

- Evaluation metrics:
  - **Precision**:
    - Of all the times the model predicted Above 50K, how often was it correct?
  - **Recall**:
    - Of all the actual Above 50K observations, what percentage did the model catch?

Methodology

Preprocessing	Model selection	**Model evaluation**	Feature importance

- Evaluation metrics:
  - **Precision**:
    - Of all the times the model predicted Above 50K, how often was it correct?
  - **Recall**:
    - Of all the actual Above 50K observations, what percentage did the model catch?
  - **F1**:
    - How good is the model at catching Above 50K and be accurate when making predictions?

Methodology

Preprocessing	Model selection	Model evaluation	**Feature importance**

- Feature importance ranking:
  - **Permutation feature importance**
    - How important is a feature based on the model error when we break its relationship?

Methodology

Preprocessing	Model selection	Model evaluation	**Feature importance**

- Feature importance ranking:
  - **Permutation feature importance**
    - How important is a feature based on the model error when we break its relationship?
  - **Random variable ranking**
    - How important is a feature compared to randomly guessing?

Methodology

Preprocessing	Model selection	Model evaluation	**Feature importance**

- Feature importance ranking:
  - **Permutation feature importance**
    - How important is a feature based on the model error when we break its relationship?
  - **Random variable ranking**
    - How important is a feature compared to randomly guessing?
- Important feature influence:
  - **Partial dependence plots**
    - What is the isolated effect of this feature on our income level predictions?

04

Results

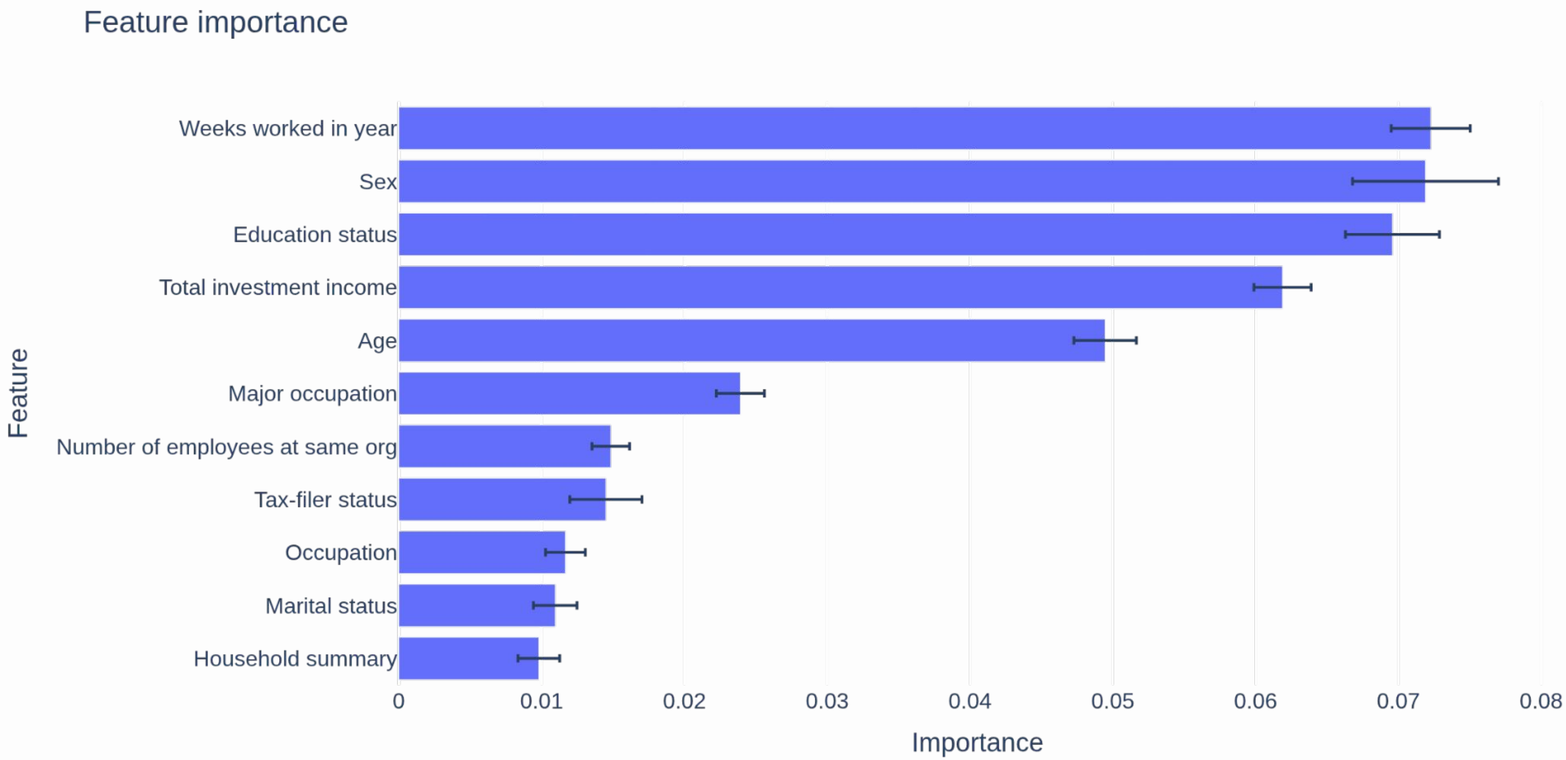
Model results and important features

Results - Performance

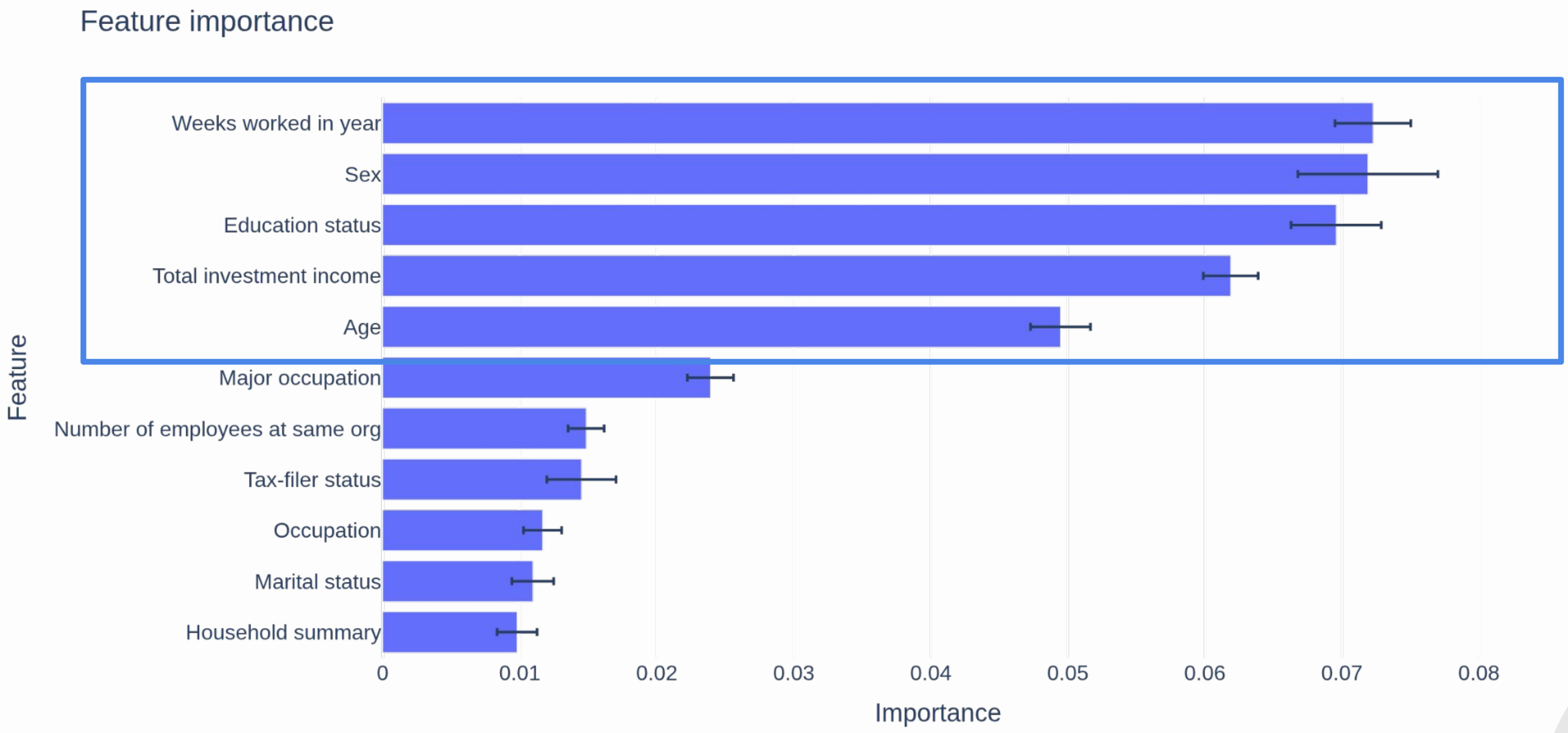
Random forest was selected as the final model and evaluated on the test set.

	Precision	Recall	F1
Below 50K	0.97	0.96	**0.97**
Above 50K	0.53	0.62	**0.57**
Macro Avg.	0.75	0.79	**0.77**

Results - 11 important features



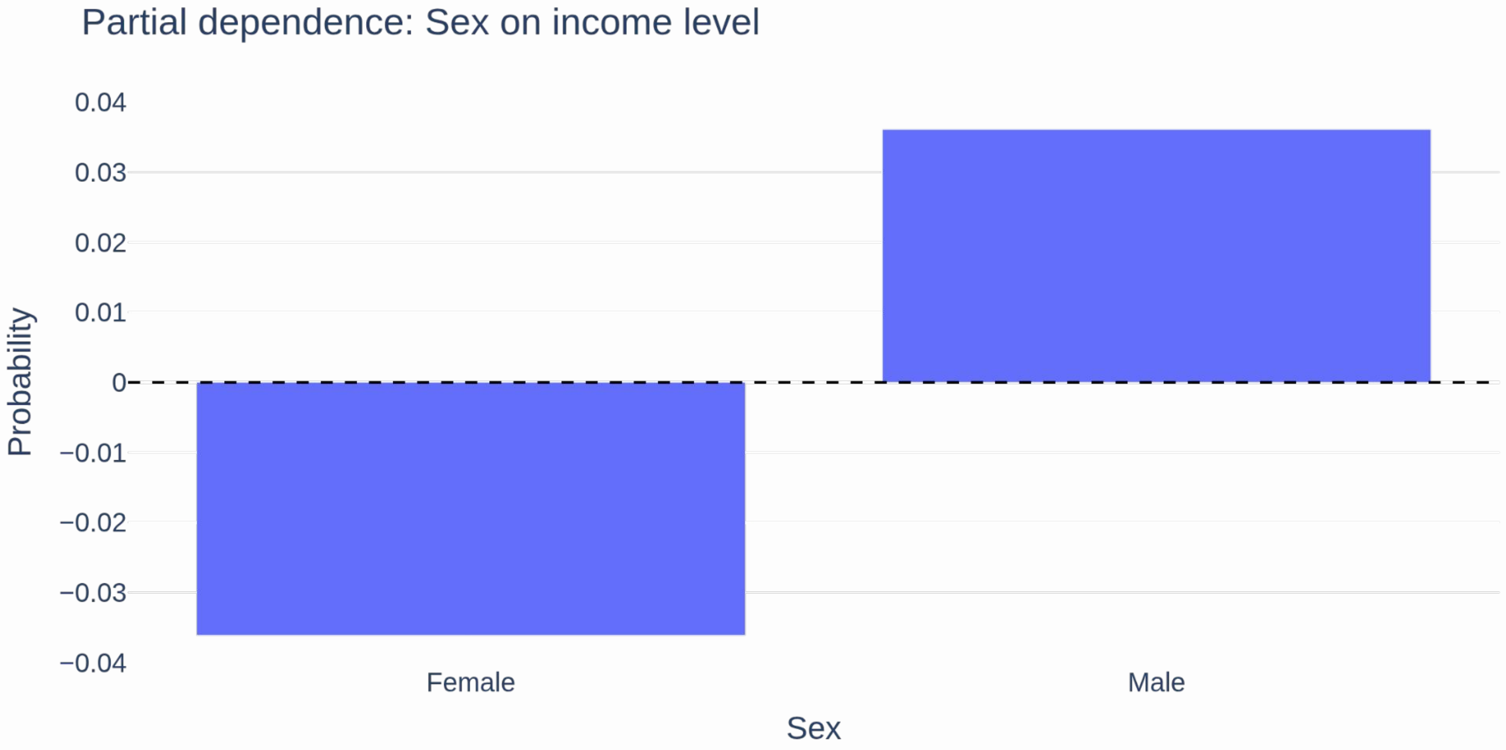
Results - Top 5 important features



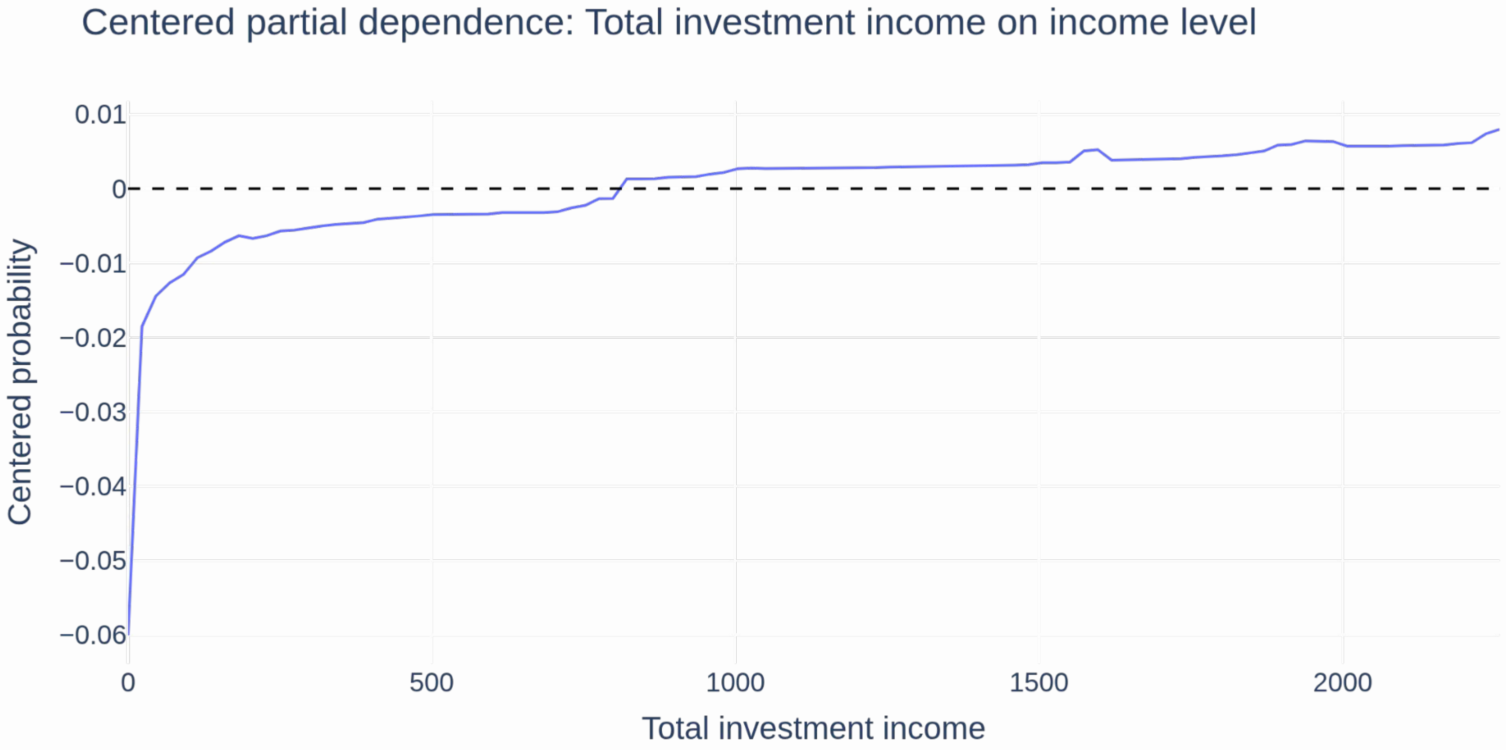
Results - Weeks worked in year effect

Results - Weeks worked in year effect

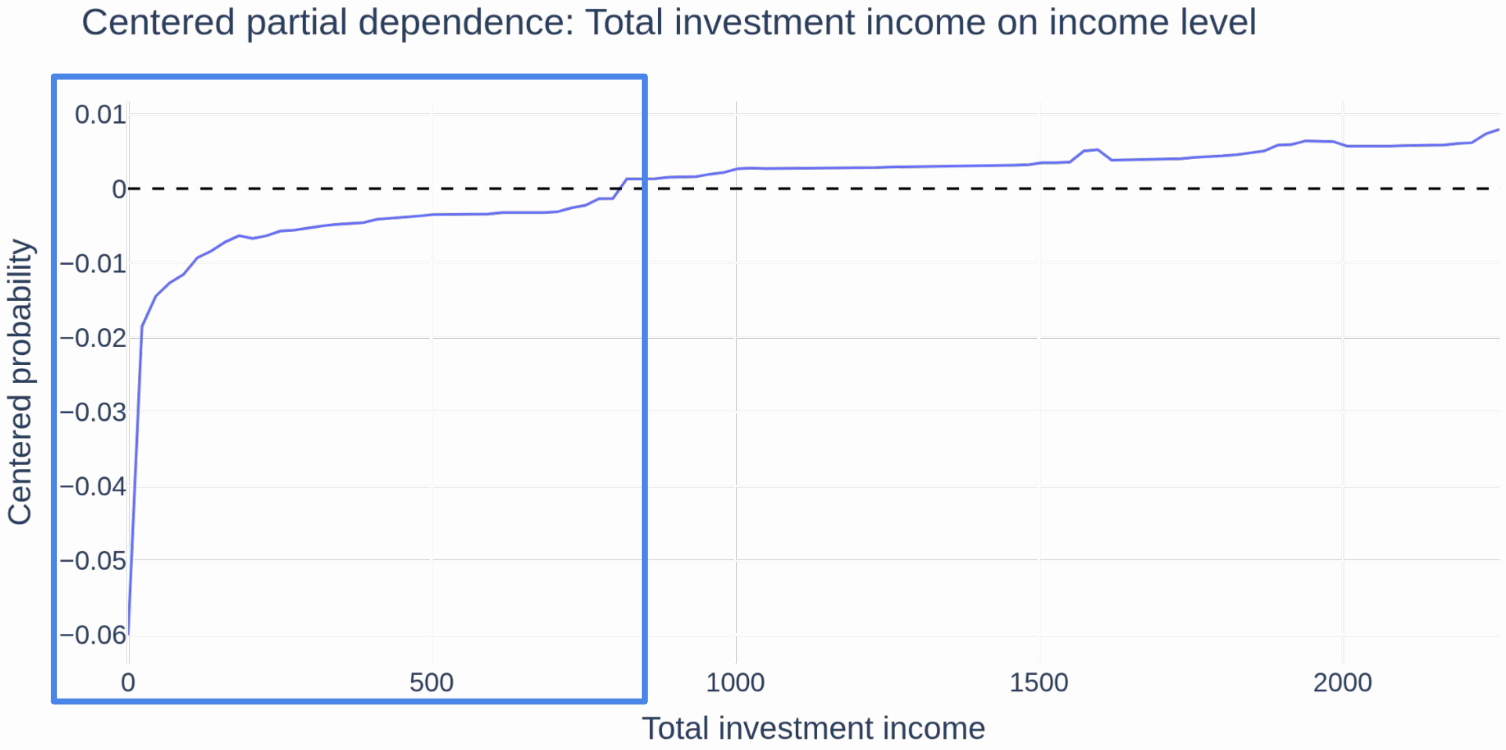
Results - Sex effect



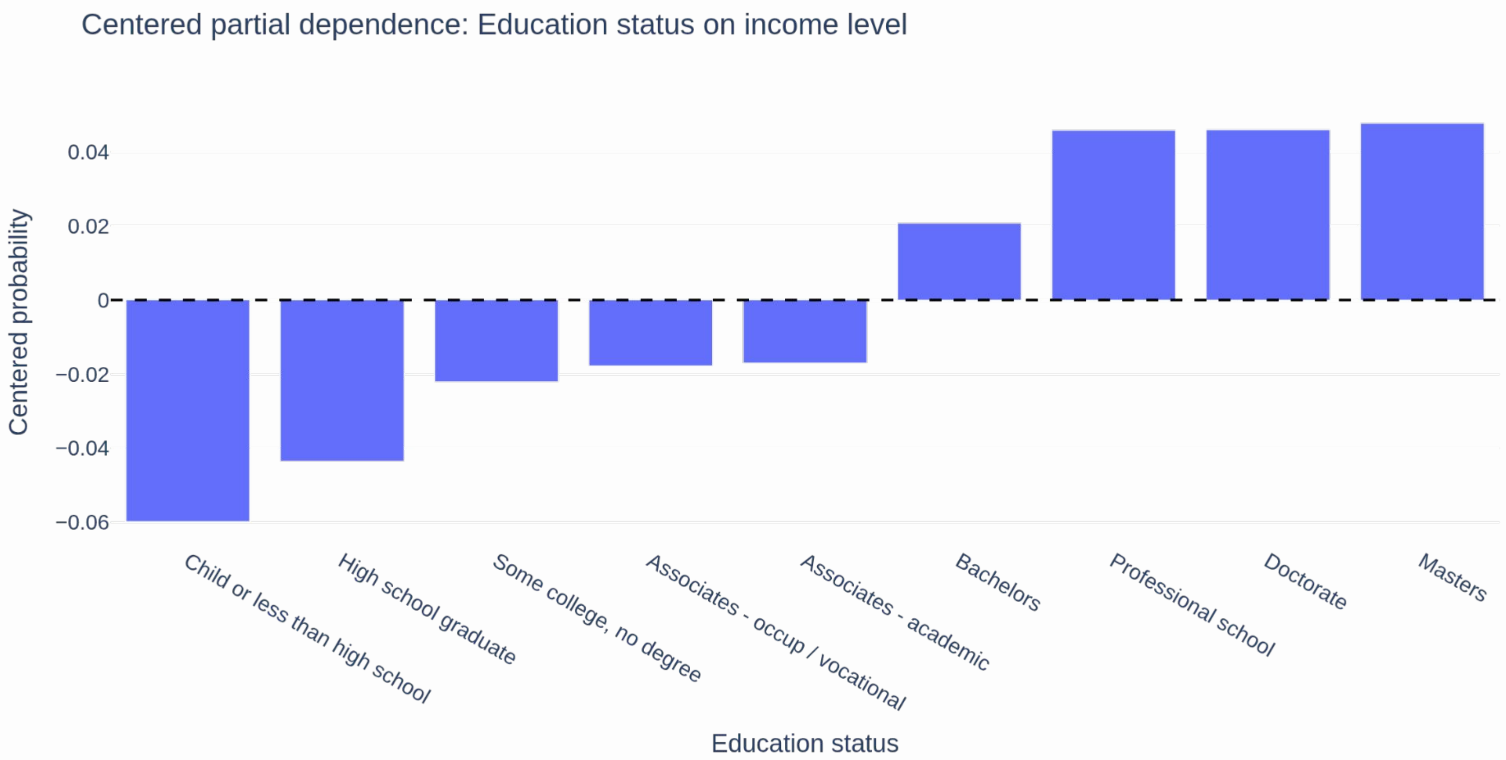
Results - Investment income effect



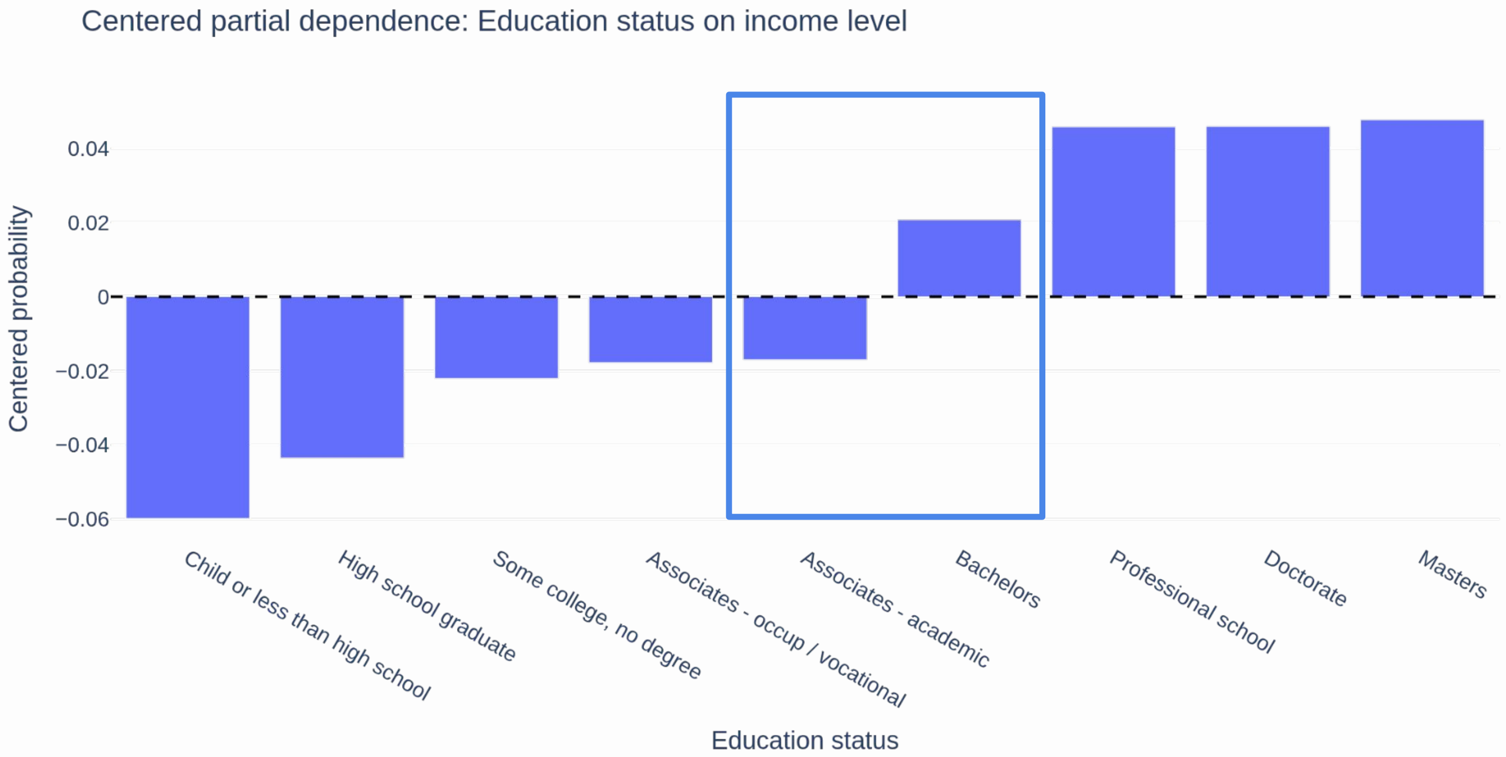
Results - Investment income effect



Results - Education effect



Results - Education effect



05

Recommendations

Recommendations

- Compare with more recent time frame
- Focus investigative efforts in important features
- Investigate potential gender bias
- Leverage education to income connection

06

Future Improvements

Future improvements

- Determine missing value pattern for migration features using other demographic information
- Develop more interactions between features
- Create additional features on business context
- Implement additional feature selection methods
- Evaluate model ensembles

Thank you!
Questions?