Predicting Income Levels Using US Census Data

A Data Science Approach to Understanding $50K+ Income Characteristics

Omar Kadim

Problem Statement & Data Overview

- Research Question: What characteristics are associated with earning more/less than $50,000 per year?
- Dataset: ~300,000 individuals from US Census archive
- Key Challenge: Highly imbalanced data (only 6.2% earn >$50K)
- Approach: End-to-end ML pipeline with interpretability focus

Dataset	Income Class	Count	Percentage
Train	≤ $50K	187,141	93.79%
	> $50K	12,382	6.21%
Test	≤ $50K	93,576	93.80%
	> $50K	6,186	6.20%

Data Preprocessing Highlights

Initial challenges:

- Missing values up to 98% in some columns
- String formatting issues (leading spaces, trailing periods)
- Duplicate records (~3,229 duplicates removed)

Solutions:

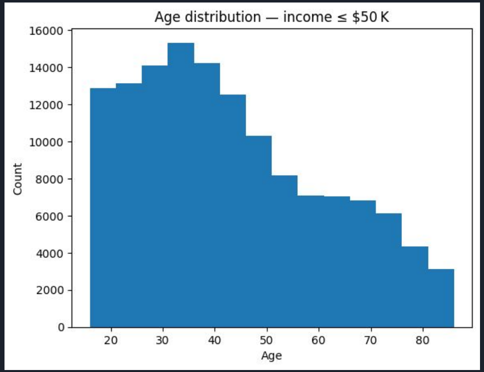
- Dropped 7 columns with >80% missing data
- Standardized categorical values
- Final training dataset: 151,196 rows × 33 features

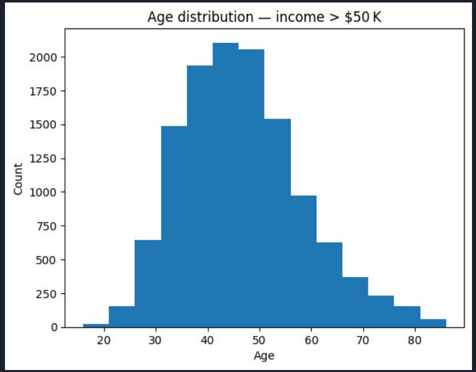
Key Finding #1 - Age Distribution

Age is the strongest predictor (18.6% feature importance)

Peak earning years: 35-50 years old
Young adults (<25) rarely high earners

Insight: Non-linear relationship - probability rises steeply from 30-50, then declines



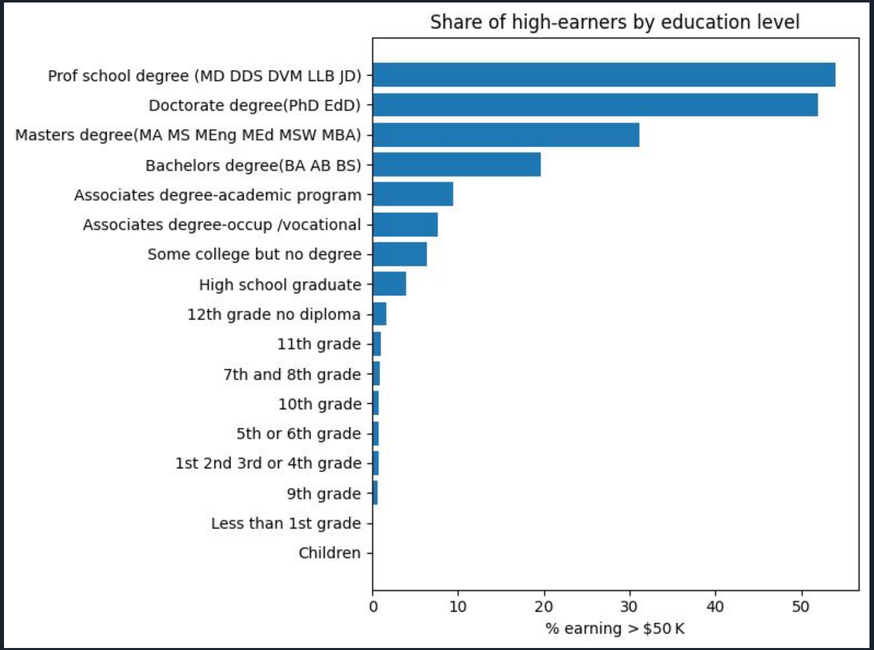


Key Finding #2 - Education Impact

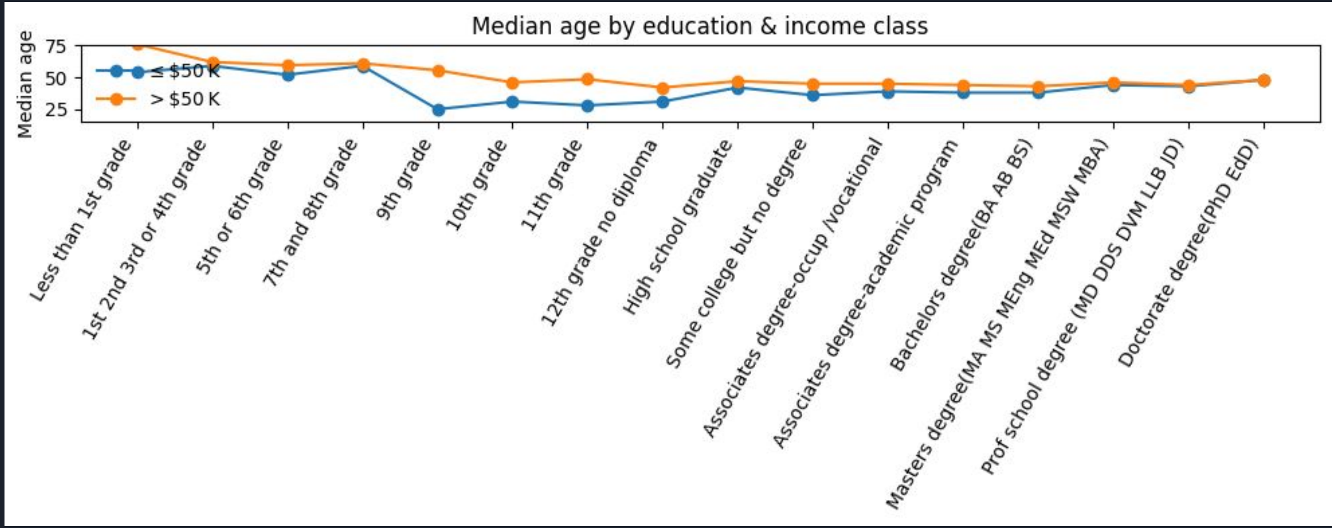
Education is second most important (8.2% feature importance)

Step-function relationship:

- High school & below: <5% earn >$50K
- Bachelor's degree: ~20% earn >$50K
- Master's degree: ~30% earn >$50K
- Professional/Doctoral: >50% earn >$50K



Key Finding #2 - Education Impact



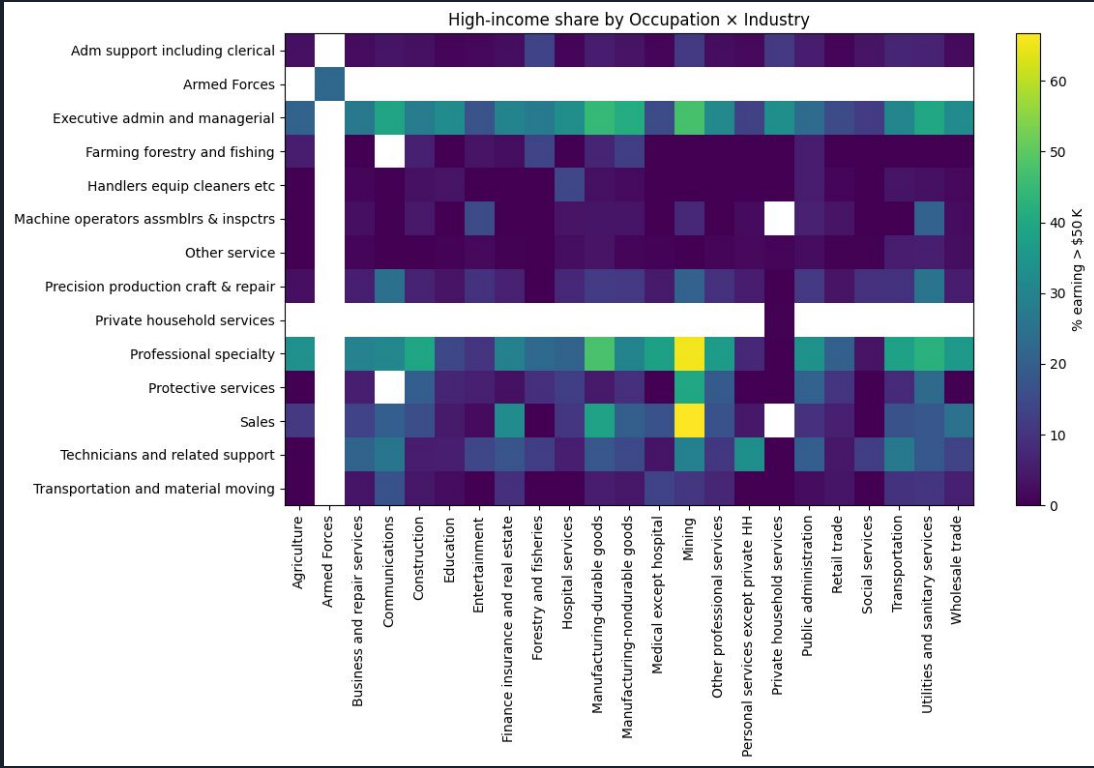
Experience premium shrinks as education rises.

Key Finding #3 - Occupation & Gender Gaps

Occupation matters:

- Executive/Professional roles dominate high earners

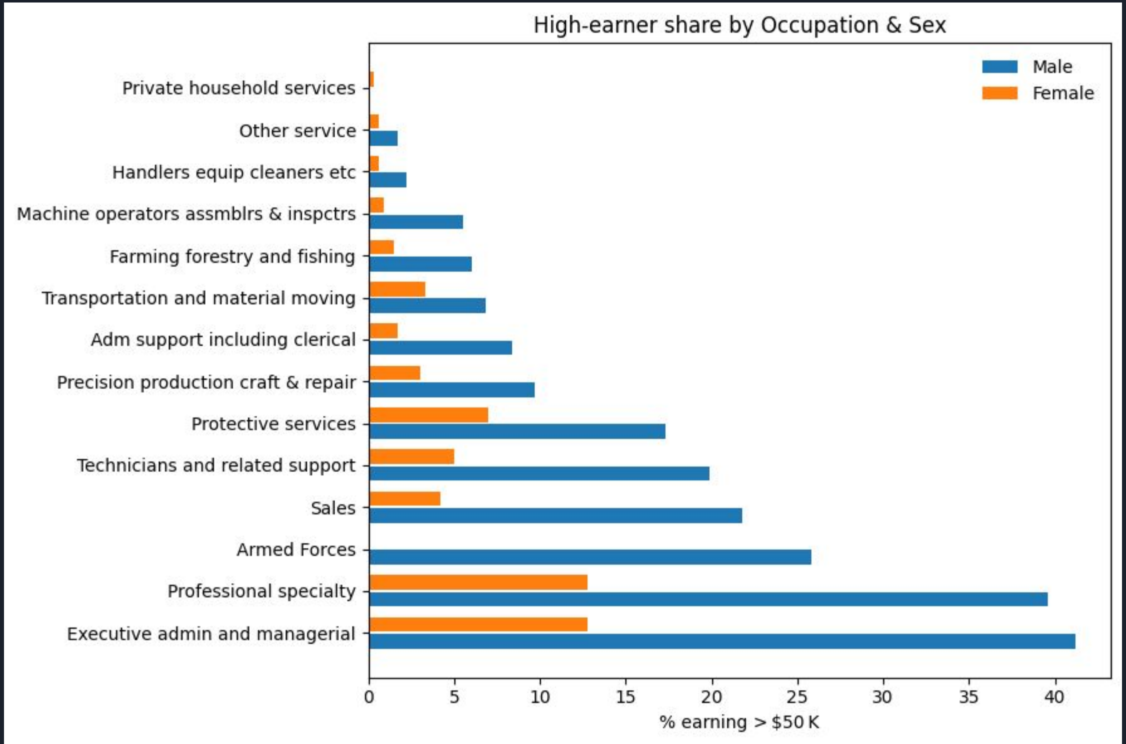
Industry + Occupation interaction crucial



Key Finding #3 - Occupation & Gender Gaps

Gender disparities evident:

- Executive roles: Male 41.2% vs Female 12.8% earn >$50K
- Professional roles: Male 39.6% vs Female 12.8% earn >$50K



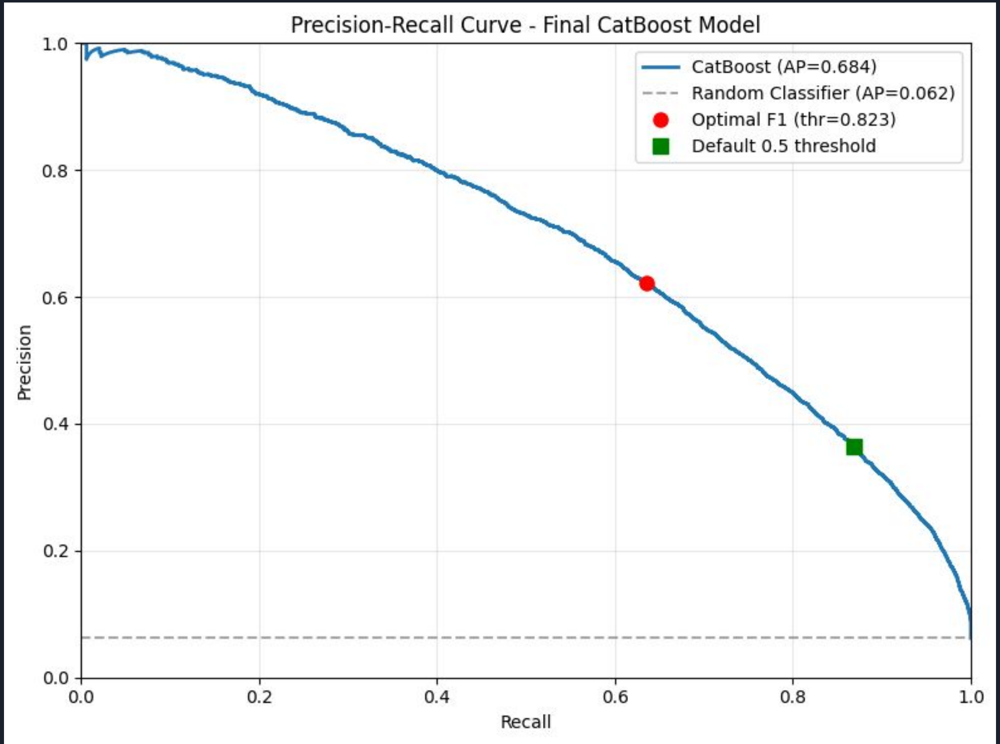
Machine Learning Model Results

Model Choice: CatBoost (gradient boosting)

Performance:

- AUC: 0.956 on test set (excellent discrimination)
- Precision: 64% at optimal threshold
- Recall: 62% at optimal threshold

Key advantage: Handles categorical features natively, captures non-linear patterns



Machine Learning Model Results

Model Choice: CatBoost (gradient boosting)

Performance:

- AUC: 0.956 on test set (excellent discrimination)
- Precision: 64% at optimal threshold
- Recall: 62% at optimal threshold

Key advantage: Handles categorical features natively, captures non-linear patterns

Dataset	AUC	Threshold	Precision	Recall	F1
Train (CV)	0.939	0.50	0.38	0.89	0.53
		0.83	0.65	0.64	0.64
Held-out Test	0.956	0.50	0.37	0.87	0.51
		0.83	0.64	0.62	0.63

Model Interpretability - Top Features

Top 7 Features driving predictions:

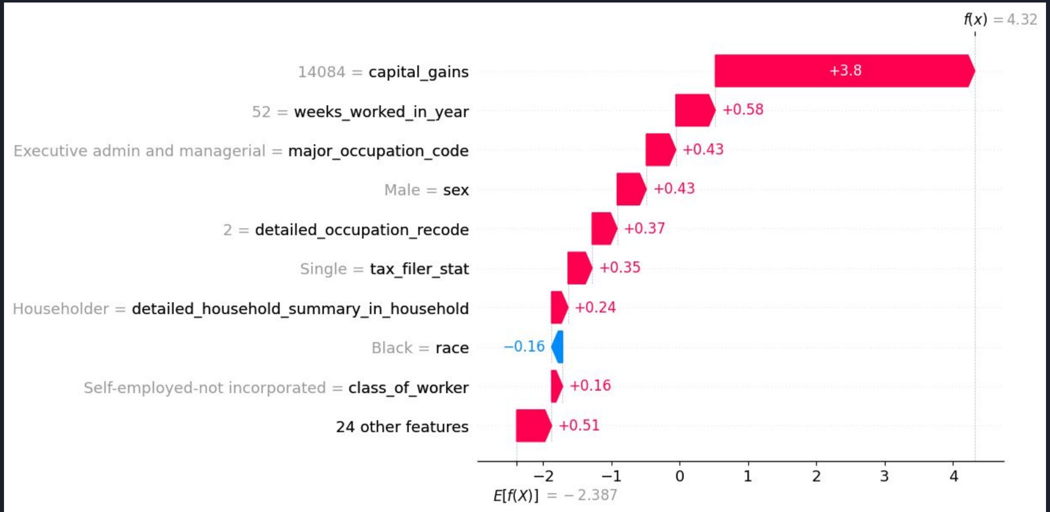
Rank	Feature	Importance (%)
1	Age	18.6%
2	Education	8.2%
3	Weeks worked per year	7.8%
4	Major occupation	6.7%
5	Sex	5.5%
6	Detailed occupation	5.4%
7	Capital gains	4.9%

Model Explainability with SHAP

SHAP Analysis reveals:

- Age and weeks worked are strongest individual drivers
- Higher education consistently increases income prediction
- Gender bias evident in model predictions
- Capital gains indicate investment wealth
- Ethnic race had a notable negative impact

Transparency: Can explain individual predictions



Business Impact & Recommendations

Threshold Selection Based on Use Case:

- High Recall (87%): Use 0.5 threshold for broad screening
- High Precision (64%): Use 0.83 threshold for targeted campaigns

Key Drivers for Policy:

- Education programs have measurable ROI
- Address gender pay gaps in professional roles
- Age-based career development programs

Limitations & Future Work

Current Limitations:

- Data from 1994-1995 (may not reflect current economy)
- Class imbalance challenges (6.2% positive class)
- Potential bias in protected characteristics

Next Steps:

- Model calibration for better probability estimates
- Ensemble methods for improved performance
- Fairness constraints to address bias

Q&A