Income Prediction Model - Census Data

Identify key economic and demographic drivers that predicts earning potential (>$50K)

Dataiku Technical Assessment
Nitant Pandey

August 2025

Executive Summary

The model accurately identifies high-income profiles and highlights education and employment patterns as key drivers

- **High-income drivers:**
  - Education, weeks worked, business ownership, and executive/professional roles
- **Lower-income drivers:**
  - Younger age (<25), social services, and retail industries
- **Model Selection:**
  - Logistic Regression provides **clear, interpretable odds-based predictions**

Data Exploration

Data // ~200K Records Analyzed

Numeric Variables

- Age
- Wage Per Hour
- Capital Gains, Losses, & Dividends
- Employer Size
- Weeks Worked Per Year

Income Markers

- Education
- Marital Status
- Worker Class
- Sex, Race
- Full Time or Part Time Employment
- Occupation, Industry
- Labor Union Membership
- Tax Filer Status
- Owning Business or Self Employment

Household

- Household & Family Status
- Household Summary
- Family Members under 18

Veteran Status

- Veteran Benefits
- Questionnaire for Veteran Admin

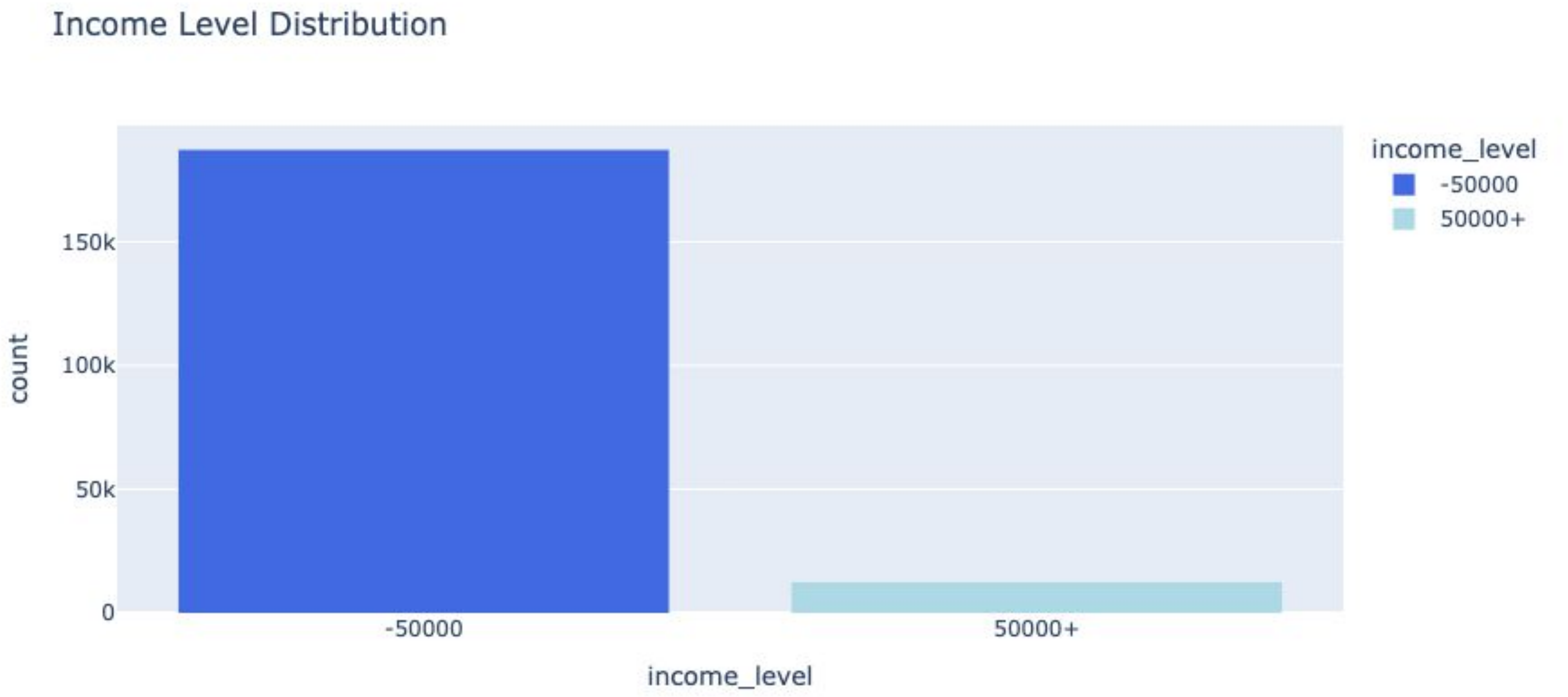
Migration

- Change in MSA
- Change in Region
- Within Region
- Lived in the same house 1 year ago
- Sunbelt
- Previous Region
- Previous State

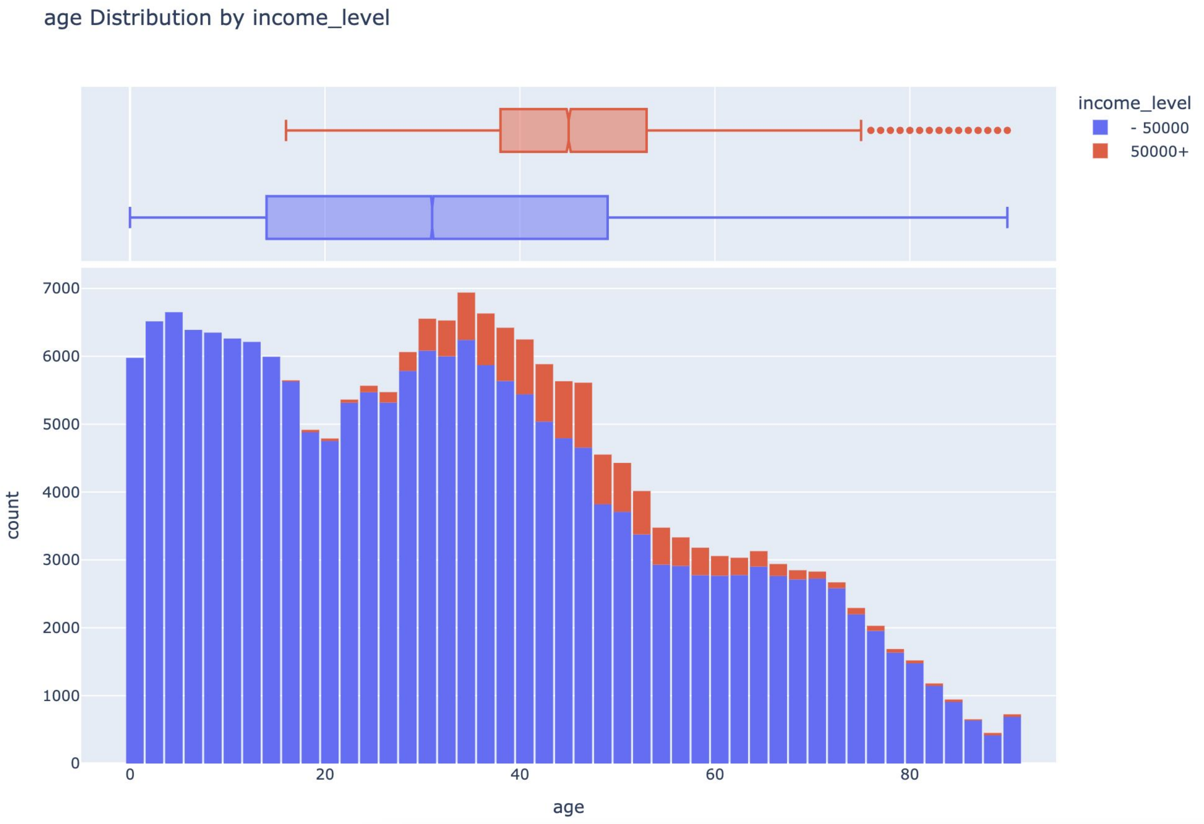
Demographic

- Hispanic Origin
- Citizenship
- Country of Birth
- Country of Parents Birth

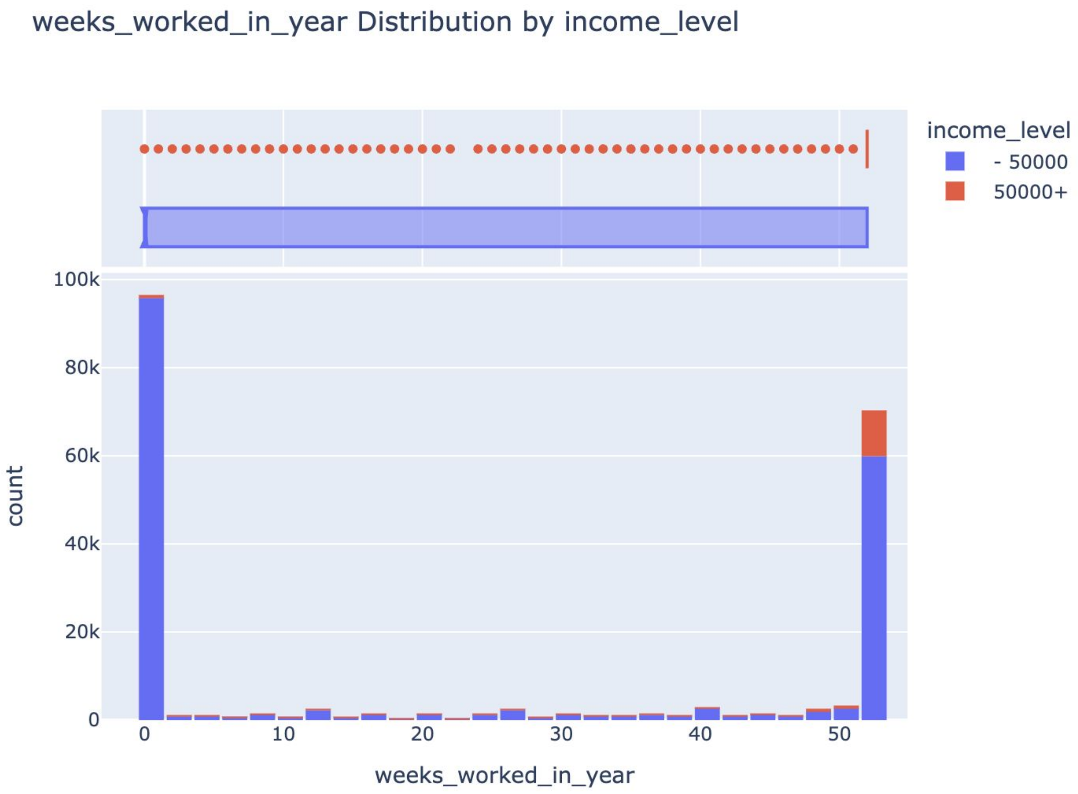
Distribution of Income Level



Most High Earners are of the age 30 or more

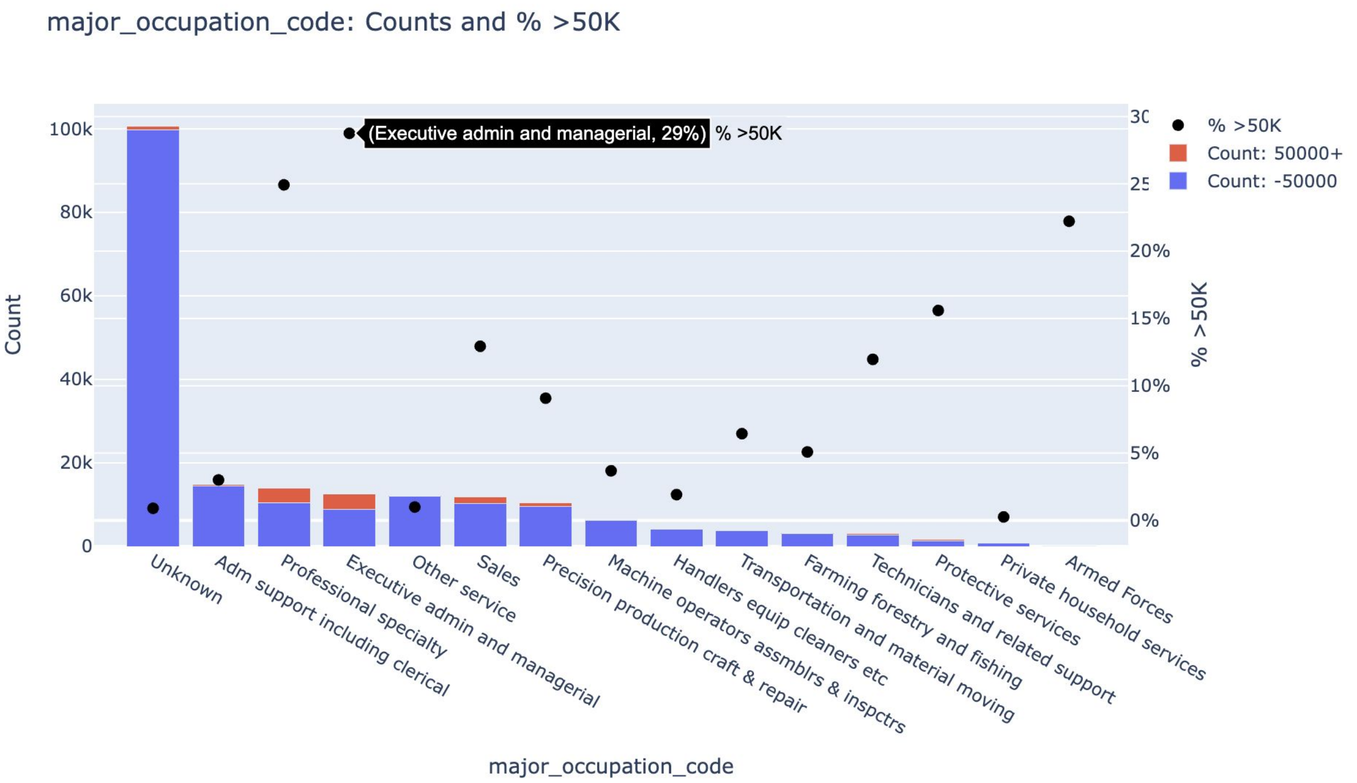


Of High Earners, Most Individuals Worked 50+ Weeks Per Year

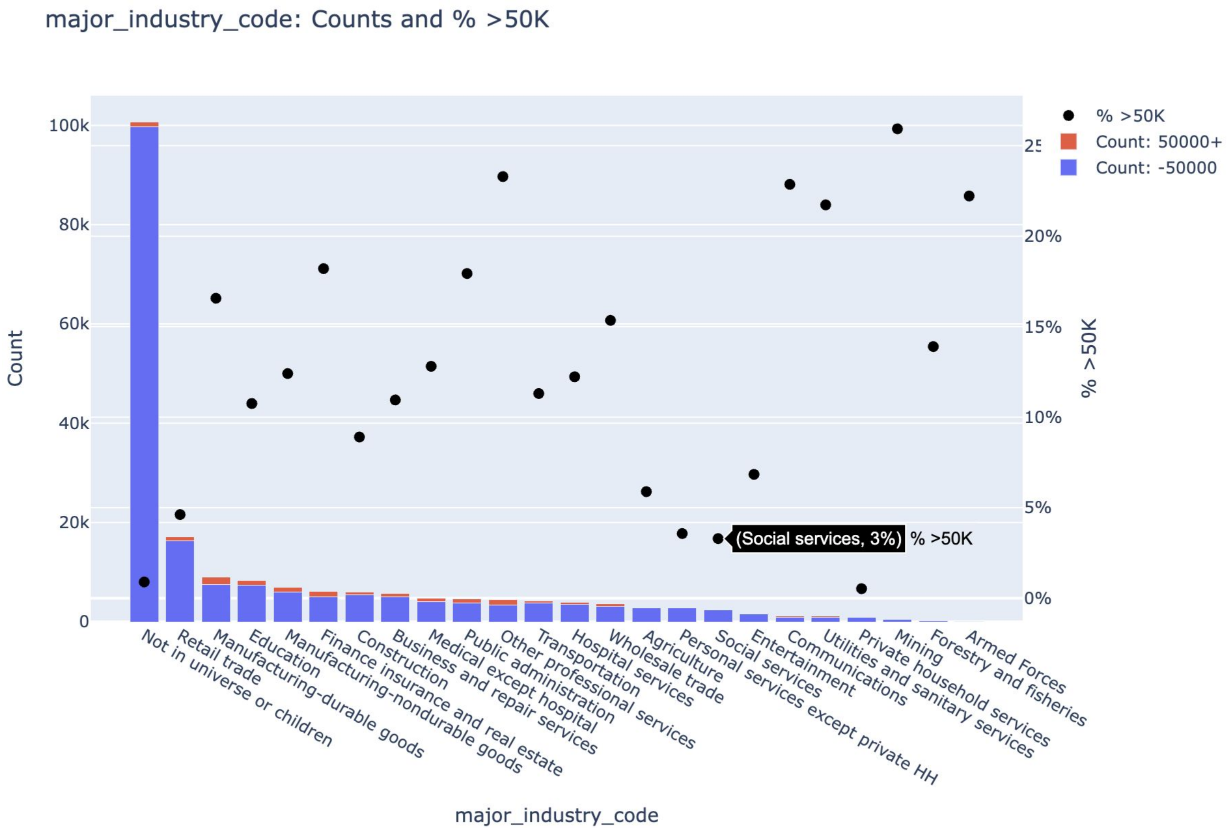


- **Lower-income individuals (≤ $50K)**: Most either worked **0 weeks or the full year**
- **Higher-income individuals ($50K+)**: Appear mostly as **outliers above the box**, meaning they are fewer in number and have varied weeks worked, mostly clustered near 52 weeks.
- The **concentration of red dots** around 52 weeks suggests that **high earners typically work the entire year**.

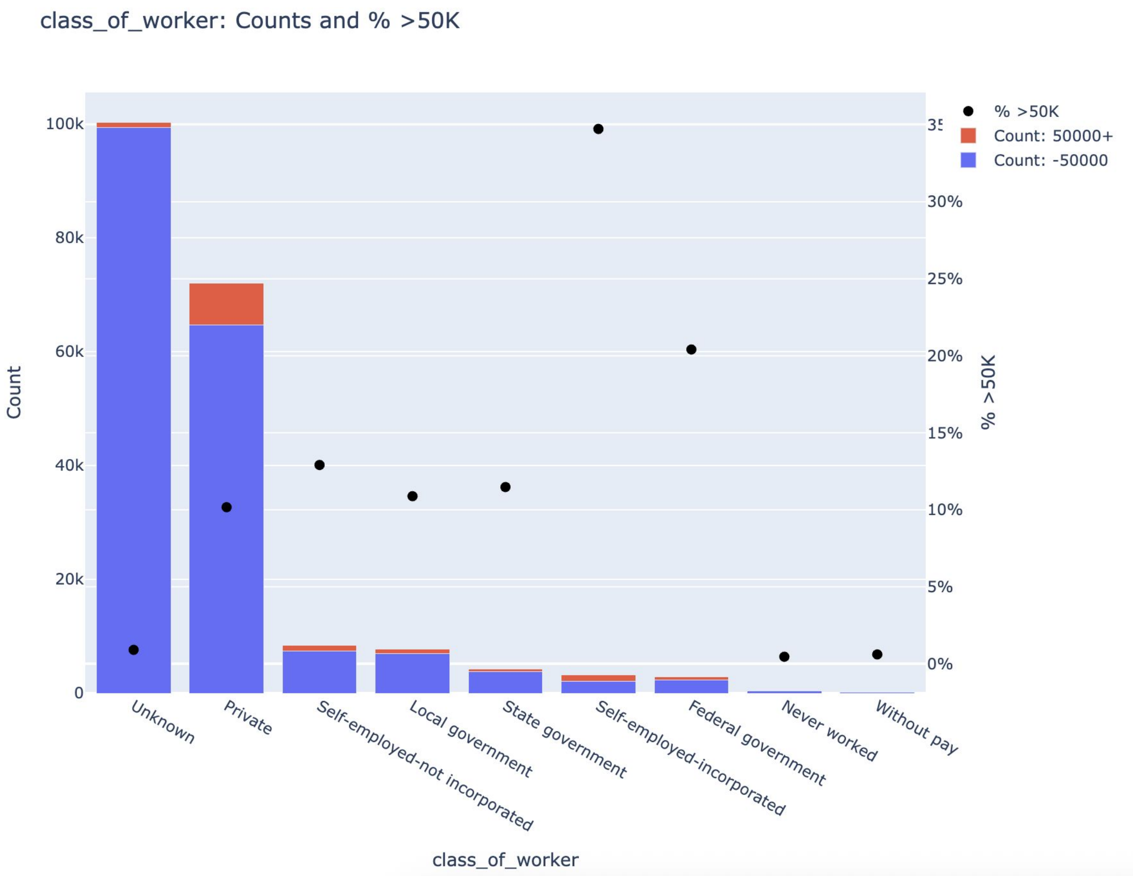
Executive & Professional Roles Dominate >$50K



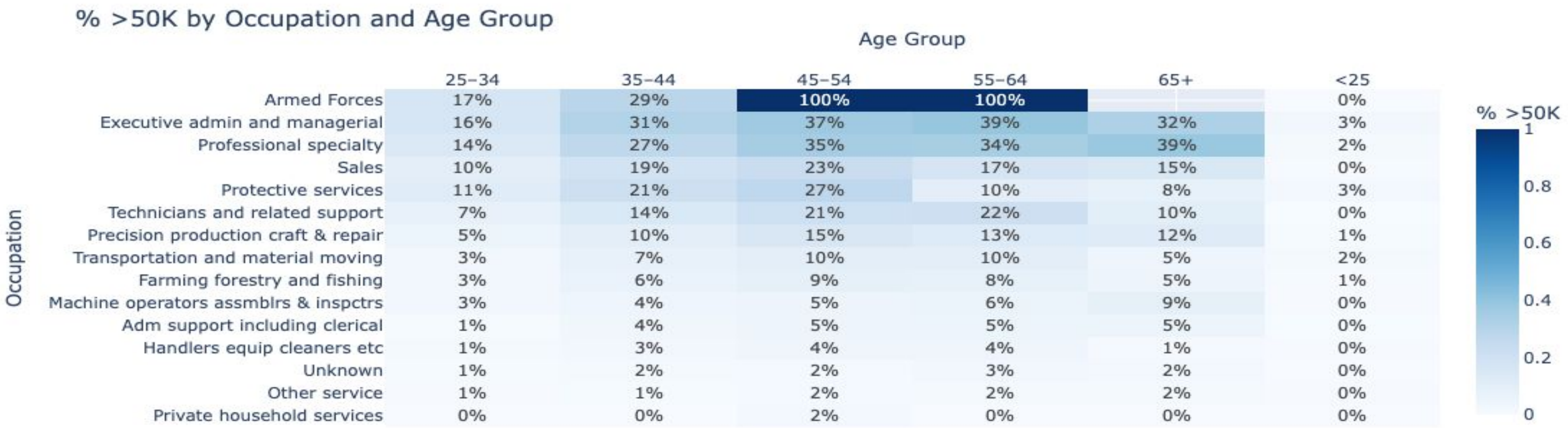
Social Services & Education tend to be <$50K



Private & Self Employed Individuals have higher earning Potential



Income Potential Is Occupation Dependent



- *Age Amplifies Earning Potential*
- *Certain Occupations Rarely cross 50K*
- *Transitioning Occupations Drive Income*
- *Armed Forces & Executive roles dominate 50K+ Earners*

Model, Performance, & Trade Offs

Note: more than $50K earning is interchangeably referred to as “high earners”, “high income potential”, or “class 1”

Model Choice: Explainability First

- Logistic Regression selected for **transparent, odds-based predictions**
- Random Forest used as **validation** to confirm top drivers
- Key drivers: education, weeks worked, business ownership, age
- *Logistic Regression balances recall and interpretability, ideal if missing a high earner is costlier than a false alarm. Random Forest is conservative and better when false positives are more expensive.*

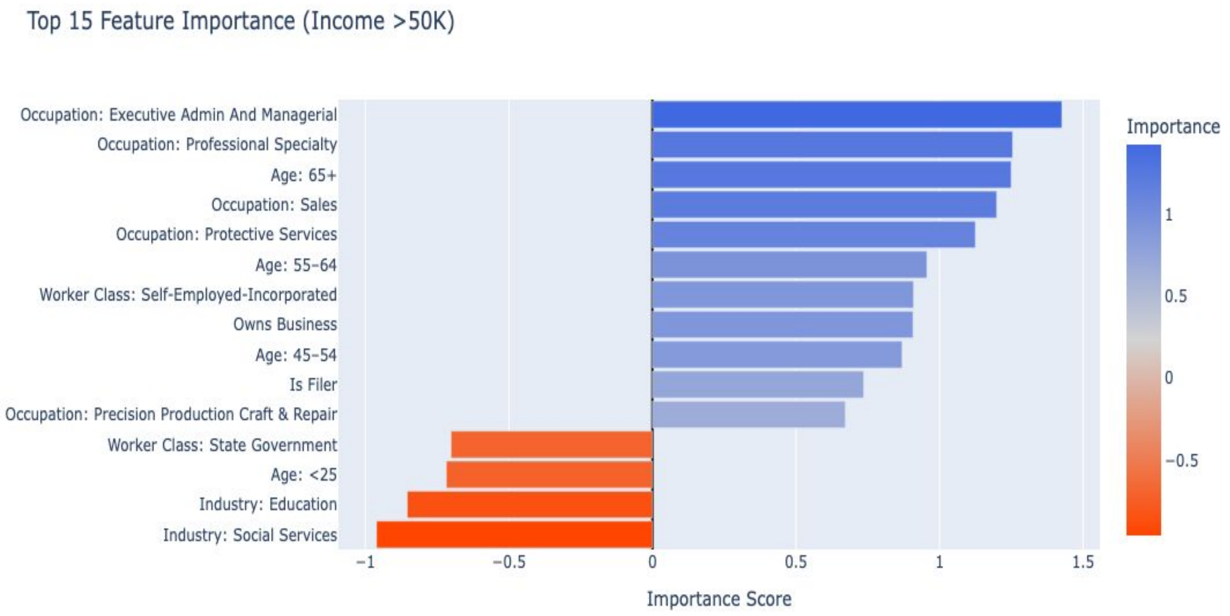
Model Performance Overview

Model greatly amplifies the targeting efficiency over no model

	Actual < $50K	Actual > $50K
Predicted < $50K	**TN = 89.5K** *correctly identified low-income individuals*	**FP = 3.9K**
Predicted > $50K	**FN = 2.3K**	**TP = 3.8K** *correctly identified high earners*

- Model already filters the population so that 1 in 2 predicted high earners is correct, vs 1 in 16 without the model.”
- Model correctly identifies 62% of the high earners with 50% precision.
- Model misses about 38% high earners, meaning **opportunity loss** if used for targeting
- Business rules or secondary checks can further refine this to improve ROI.

Top 15 Drivers of Income >$50K From the Model



Positive drivers (Blue bars):

- Executive, professional, and sales occupations have the **highest positive impact** on earning >$50K.
- Older age groups (**45–65+**) significantly increase income likelihood.
- Self-employment and business ownership are strong positive drivers.

Negative drivers (Red bars):

- Younger workers (**<25**) and public sector/state government roles are **less likely** to earn >$50K.
- Certain industries like **Education** and **Social Services** are associated with lower income.

Owning A Business Is A Strong Income Booster

1. **Owning a business significantly increases the likelihood of earning >$50K** for people in lower-paying industries like **Education** or **Social Services**.

2. The **boost is similar** (~21%) across both industries.

3. **Why this matters:**

    - Even roles that typically pay less can cross into the higher income bracket **when combined with business ownership**.

    - Business ownership acts as a **multiplier** for income potential.

For someone in Education, the chance of earning >$50K rises from 30% to 51%.
For someone in Social Services, it rises from 28% to 49%.

This shows that even in lower-income potential industries, owning a business can nearly double the likelihood of crossing the $50K mark.

Strategy to Improve ROI from Model Predictions

Business Rules

- Adjust model probability threshold
- Apply **hybrid rule**: Model prediction × Business logic
    - Model probability > 0.7
      AND Age 35–65
      AND Executive/Admin
    - Model predicts >$50K AND
      Owns Business
      OR Files Taxes
    - Predicted >$50K
      AND Urban + Married + Works Full-Time

Advanced Models

- Gradient Boosting (XGBoost, LightGBM) for higher precision
- AdaBoost or Random Forest to reduce false positives
- Stacking models (LogReg + RF + Boosting) for best performance
- Calibrate probabilities + tune thresholds for ROI optimization

By combining business rules with advanced models, we can:

- ➢ Increase precision and reduce wasted outreach
- ➢ Capture more true high earners while maintaining interpretability
- ➢ Continuously improve ROI through threshold and model tuning

Hybrid Rule vs Original Model: Precision-Recall Tradeoff

Precision = accuracy of identifying high earners
Recall = ability to find all high earners

Metric	Original Model	Hybrid Rule v1	Hybrid Rule v2
Precision (Class 1)	0.50	0.28	0.47
Recall (Class 1)	0.62	0.73	0.22
Balance or F1 Score (Class 1)	0.55	0.41	0.30
False Positive Rate	4.1%	12.1%	1.6%

- ➔ Original Model: Balanced precision and recall.
- ➔ Hybrid Rule v1: Maximizes recall but sacrifices precision (false positives).
- ➔ Hybrid Rule v2: Very precise, minimal false positives, but misses true positives.

Business Trade Off: *Choose v1 for coverage-focused campaigns, v2 for ROI-focused targeting.*

Model Blind Spots And Uncertainties

⚠️ High-income individuals with non-traditional signals may be missed (False Negatives)	✅ Monitor False Negative Rate monthly; add hybrid rules for edge cases
⚠️ Over-reliance on model predictions could lead to missed opportunities	✅ Combine model with business logic; review borderline cases with human-in-loop
⚠️ Certain occupations or industries (e.g., Social Services) have lower model accuracy	✅ Track precision/recall by occupation; collect more data or upsample
⚠️ Bias risk: Age or occupation may indirectly impact fairness if not monitored	✅ Perform regular fairness audits; reweigh or remove sensitive variables if needed
⚠️ Data drift risk: Income patterns may shift over time, requiring retraining	✅ Monitor data distributions; retrain model quarterly; use drift dashboards

Recommendation

- Deploy **Logistic Regression** as the **primary model** for explainable and fair predictions
    - Combine model probability with **simple business rules** (e.g., owns business, full-time)
    - Aligns with business needs for **transparent decision-making**
- Random Forest serves as **secondary validation**, confirming top drivers
- Hybrid Rule Layer
- Monitoring and Maintenance Plan
- Future Enhancements

Why Model Over Raw Correlation?

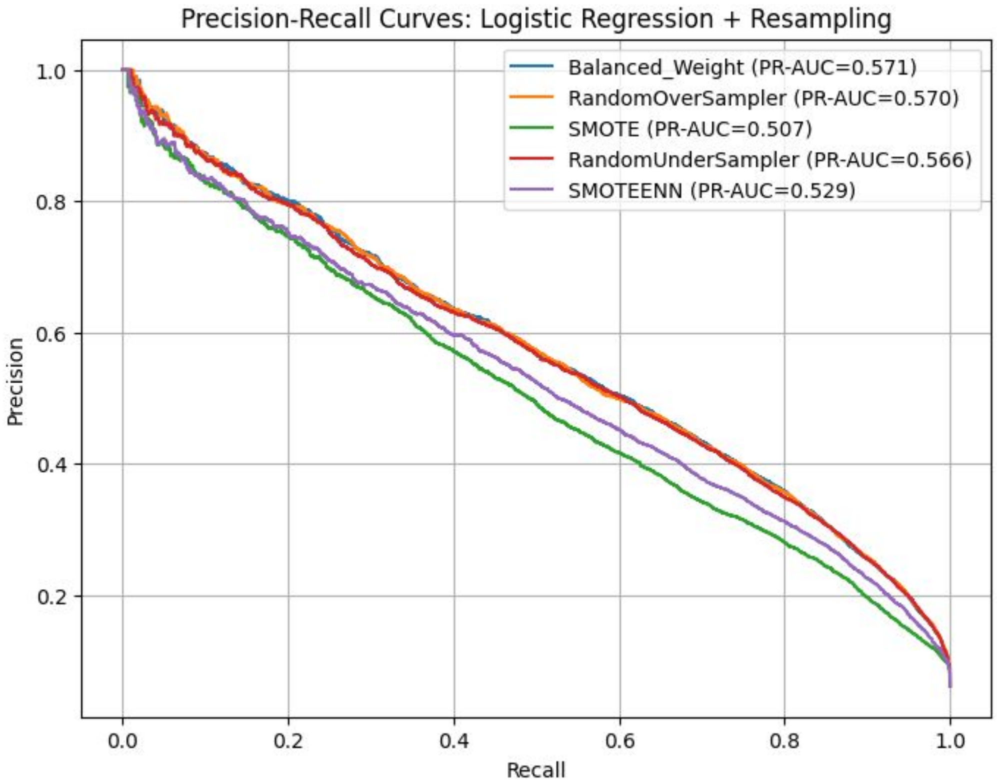
- Classification doesn't discover *causal* markers but does provide the **most predictive and interpretable signals** for practical decision-making
- **Trained specifically to distinguish high-income vs low-income individuals**, so its values directly highlight **drivers of income level**
- More interpretable than raw correlations because the model **controls for other variables simultaneously**
- Classification ensures markers are **predictive**, not just **coincidental**
- Model can **prioritize follow-ups or interventions** (e.g., marketing, program targeting) while also explaining **why** a feature matters

Appendix

Turning Raw Data into Signals // Feature Engineering

- **Data Cleaning & Standardization** (“?”, leading/trailing spaces, “Not in Universe”)
- **Categorical Encoding**
  - **One-hot encoding** for key categorical variables:
    - major_industry_code, major_occupation_code, class_of_worker, marital_status, citizenship, etc.
  - Created **age group buckets**:
    - <25, 35–44, 45–54, 55–64, 65+
- **Derived Binary Flags (business, tax filer, US born)**
- **Log Transformations for Skewed Features** (dividends and gains/losses)
- **Household & Demographic Indicators**
  - Collapsed **detailed household status** into simplified flags:
    - marital_status_Married (spouse present
    - Child or grandchild in household
  - family_members_under_18 simplified into **presence/absence flags**

Engineering / Tuning the Base Model



Insights:

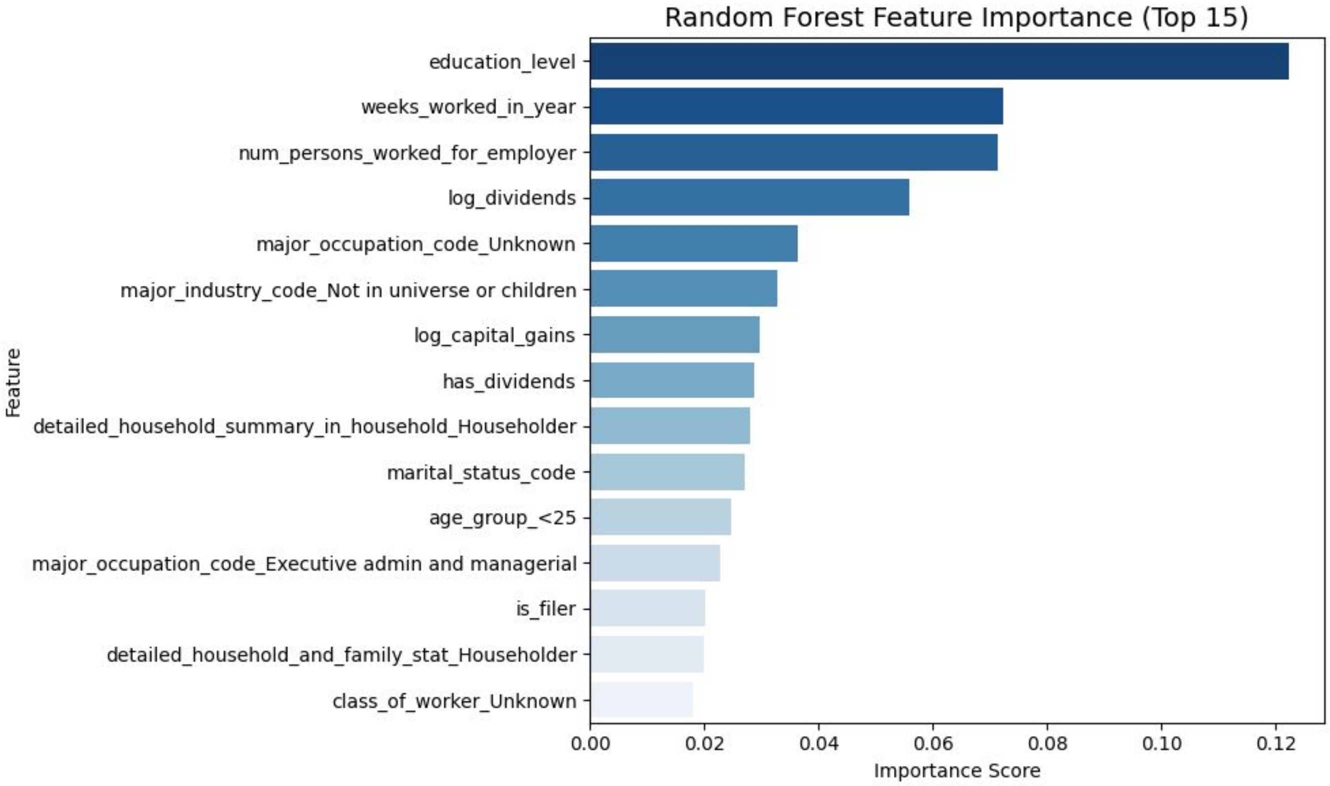
- Class balancing and random over-sampling performed **similarly and best**.
- Complex resampling (SMOTE/SMOTEENN) **did not improve results** and introduced **noise**.
- **Business takeaway:** Simple balancing works best for this dataset; avoid overengineering.

Random Forest Validation

Model / Scenario	Accuracy	Precision (Class 1)	Recall (Class 1)	F1 (Class 1)	Notes
Logistic Regression	0.940	0.490	0.620	0.550	Balanced recall; moderate precision
Hybrid Rule v1	0.869	0.284	0.731	0.409	Aggressive recall, many false positives
Hybrid Rule v2 (Stricter)	0.936	0.468	0.220	0.299	Reduced FPs, but low recall
Random Forest (Calibrated)	0.950	0.742	0.300	0.427	Very precise but misses many positives

- **Random Forest =** 🎯 – Shoots fewer targets but usually hits the right ones.
- **Logistic/Hybrid =** 🕸️ – Catches more, but with some False Positives.

Random Forest Confirms The Same Major Drivers



Correlation Matrix

	age	wage_per_hour	capital_gains	capital_losses	dividends_from_stocks	instance_weight	num_persons_worked_for_employer	weeks_worked_in_year
age	1.000000	0.036938	0.053590	0.063351	0.104976	-0.001611	0.140887	0.206181
wage_per_hour	0.036938	1.000000	-0.001082	0.010993	-0.005731	0.012353	0.191543	0.195687
capital_gains	0.053590	-0.001082	1.000000	-0.012700	0.131476	0.002549	0.058015	0.083549
capital_losses	0.063351	0.010993	-0.012700	1.000000	0.042427	0.008052	0.084255	0.100762
dividends_from_stocks	0.104976	-0.005731	0.131476	0.042427	1.000000	-0.000009	0.007206	0.013823
instance_weight	-0.001611	0.012353	0.002549	0.008052	-0.000009	1.000000	0.042778	0.029240
num_persons_worked_for_employer	0.140887	0.191543	0.058015	0.084255	0.007206	0.042778	1.000000	0.747302
weeks_worked_in_year	0.206181	0.195687	0.083549	0.100762	0.013823	0.029240	0.747302	1.000000