
# Understanding Income Drivers

## A Census-Based Data Science Analysis

Prepared by: Kazmeen Chaudhary

Assessment: Dataiku Data Scientist – Technical Evaluation

Objective: Demonstrate a clear, explainable, and production-ready data science approach while communicating effectively with a mixed technical audience.

# Problem Context & Data Overview

## Business Question

The US Census Bureau collects demographic and economic data to understand population characteristics and inform policy decisions.

Our Task: Predict whether individuals earn ⩽$50K or >$50K annually using census demographic features

## Data Sources Provided

We worked with four files:

- Training dataset – historical census records used to build models
- Test dataset – unseen data used to validate generalization
- Metadata file – definitions, encodings, and domain context
- Supplemental documentation – survey structure and caveats

## Dataset at a Glance

| ~300K | 40+ | 2:1 |
|---|---|---|
| Records | Features | Train/Test Split |
| Individual-level census data | Demographics, education, employment, financial | 2/3 training, 1/3 held-out test |

Binary classification with class imbalance: ~6% earn >$50K

## Data Structure

Hierarchical microdata structure (Person → Family → Household). Analysis focuses on person-level predictions while accounting for nested dependencies.

## Why This Matters

### Policy Impact

- Target social programs
- Understand mobility barriers

### Business Value

- Identify income drivers
- Inform resource allocation

# Analytical Approach

## End-to-End Methodology

To answer the question responsibly, I followed a standard but disciplined pipeline:

01

Exploratory Data Analysis

02

Data cleaning & preprocessing

03

Feature engineering & grouping

04

Model training (multiple approaches)

05

Model evaluation & comparison

06

Explainability & interpretation

07

Limitations & next steps

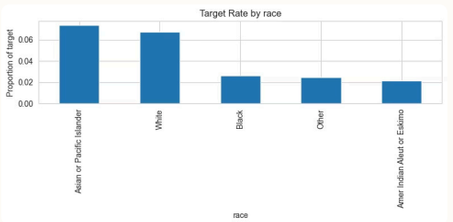
# Exploratory Data Analysis (EDA)

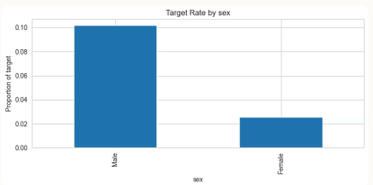
## What We Explored First

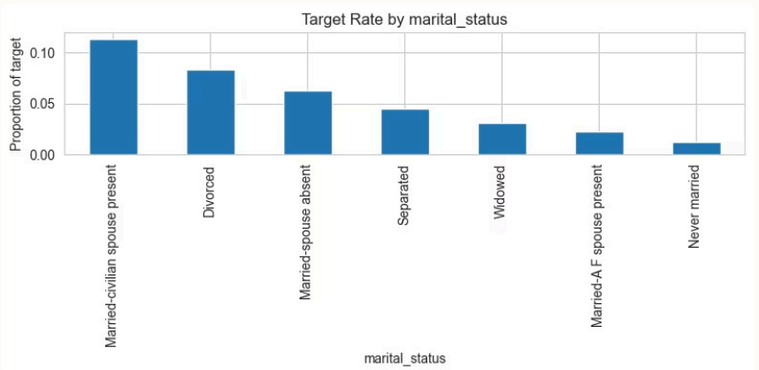
Before modeling, we explored:

- Income imbalance (only ~6% earn > $50K)
- Distribution of age, education, work duration
- Income rates across:
  - Education levels
  - Marital status
  - Sex and race groupings
  - Employment type

> **Key takeaway:** Income is highly skewed, and relationships are non-linear. This immediately told us that accuracy alone would be misleading.



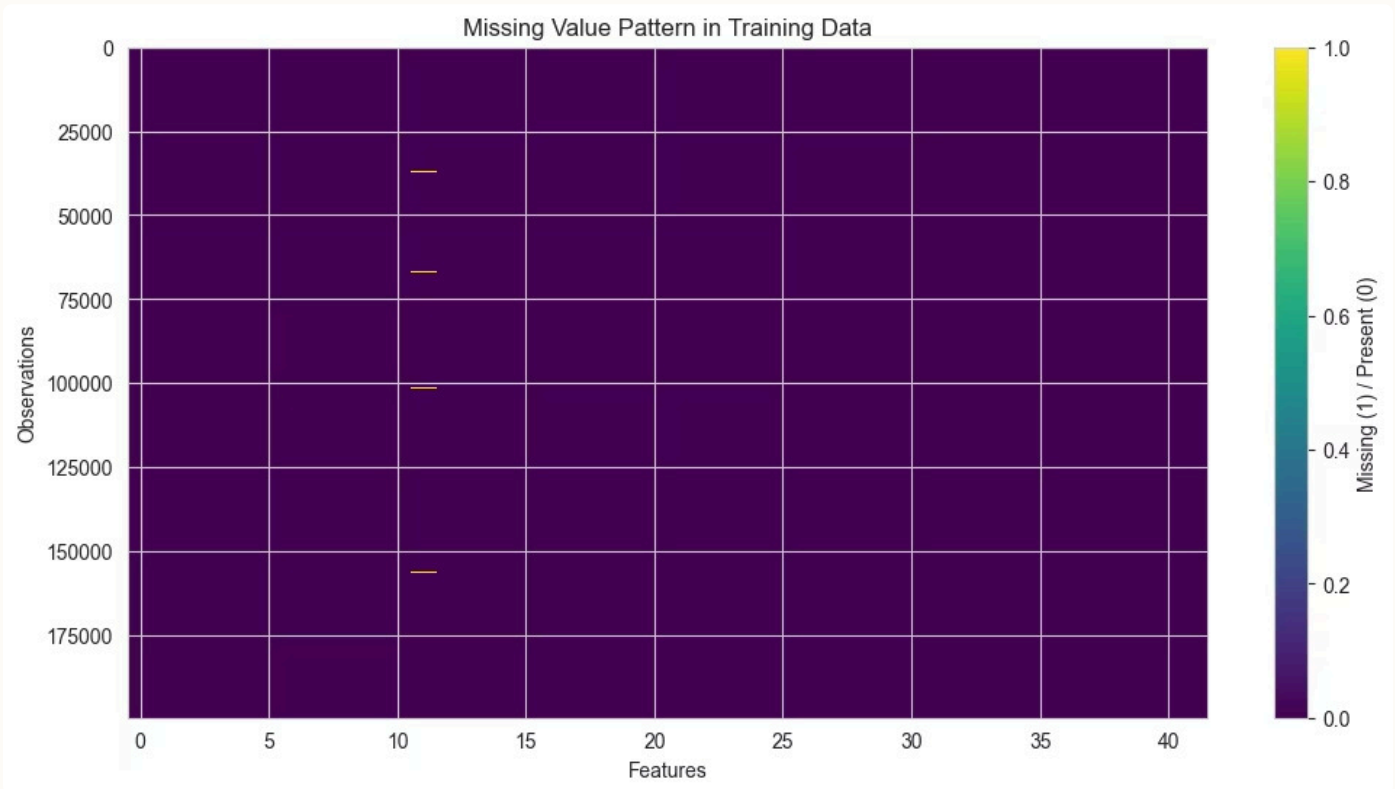



## Data Cleaning & Preprocessing Decisions

Key steps included:

- Handling "Not in universe" values explicitly (not blindly dropped)
- Preserving meaningful missingness
- Log-transforming highly skewed numeric variables:
  - Wages
  - Capital gains/losses
  - Dividends
- Grouping rare categorical levels

### Missing Value Pattern

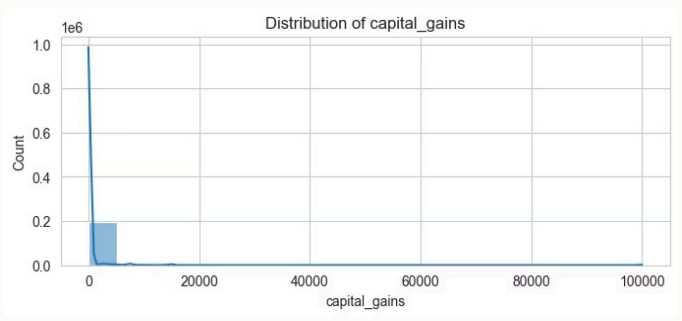



Heatmap showing missing value patterns across features. Dark purple indicates complete data, lighter colors show missing values. Most features have complete data with minimal missingness.

### Outlier Analysis

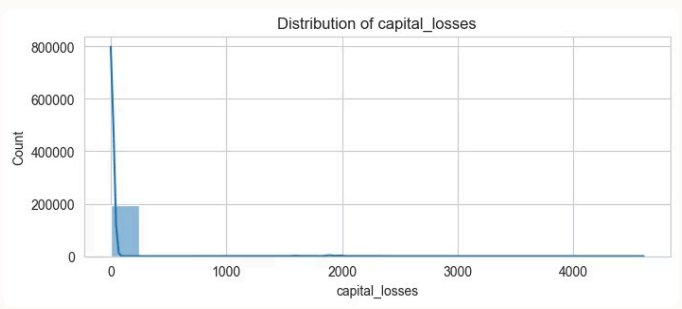
Financial variables showed significant outliers representing legitimate high earners. These were preserved and log-transformed rather than removed to maintain predictive signal from high-income individuals.

### Distribution of Financial Variables

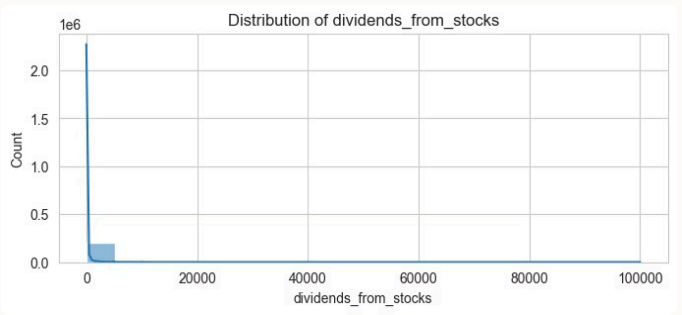



Distribution of capital_gains


Distribution of wage_per_hour




Distribution of capital_losses




Distribution of dividends_from_stocks

All four financial variables show extreme right-skew with most values concentrated near zero and long tails extending to high values. This confirms the need for log transformation to normalize distributions for modeling.

# Feature Engineering Strategy

## Thematic Feature Grouping

To improve interpretability, features were grouped by real-world meaning:

### Labor Market & Employment

Worker type, weeks worked, unemployment reason

*Captures job stability & income risk*

### Education & Occupation

Education level, occupation group, industry

*Represents human capital*

### Household & Demographics

Marital status, household role, race grouping

### Migration & Citizenship

Domestic vs international movement, generation

### Tax & Financial Signals

Filing status, senior indicator

*Strong proxies for income level*

# Feature Transformation Flow

## From Raw Data to Model-Ready Features

### 40+ raw survey features

Initial dataset from census surveys

### 73 engineered features

After grouping & transformation

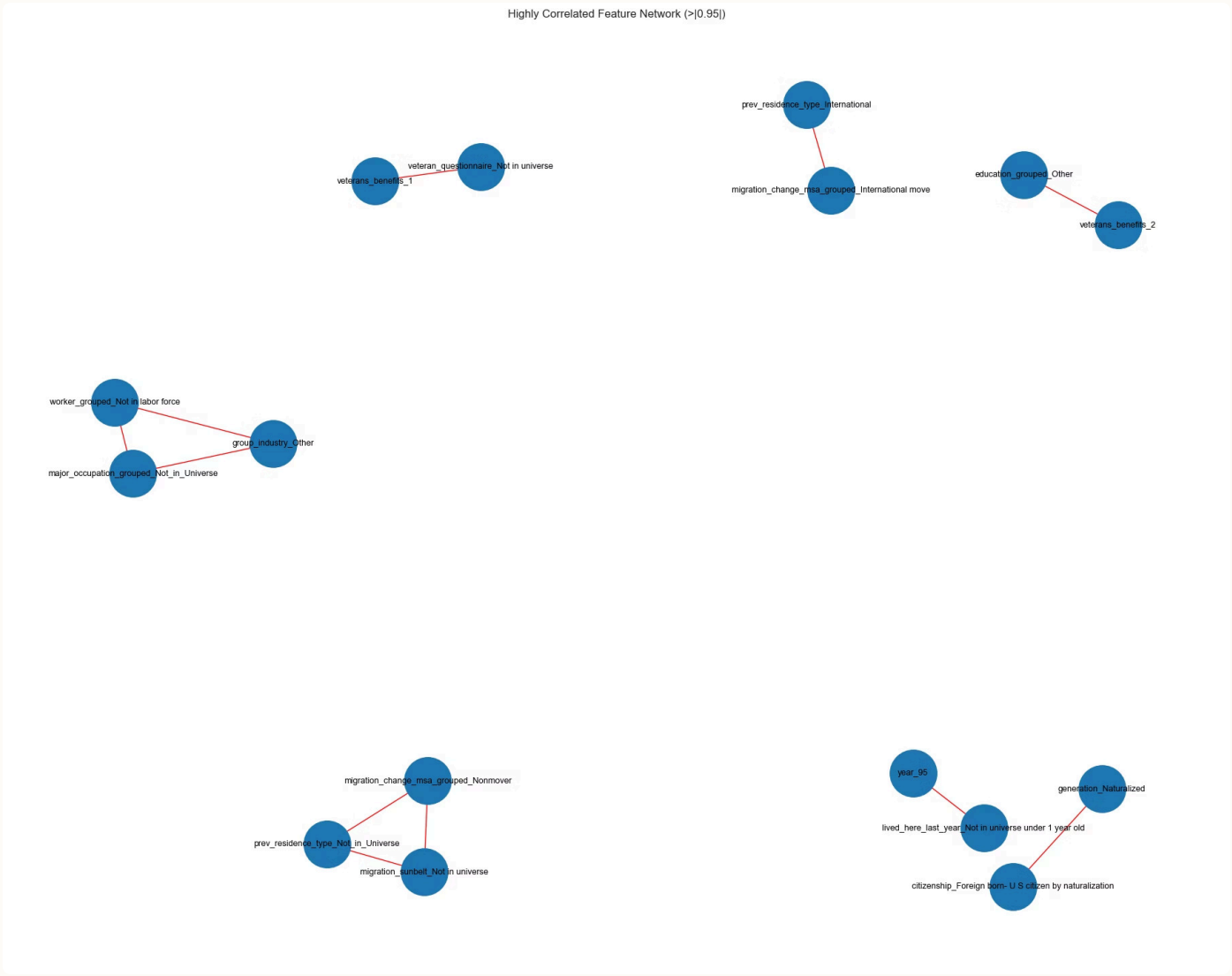
### 61 final features

After correlation pruning

# Handling Multicollinearity

## Why We Dropped Highly Correlated Features

Identified feature pairs with correlation > 0.95




Visualization of the 12 highly correlated feature pairs

**Correlated Feature Pairs Identified:**

### "Not in Universe" Overlaps:

- Occupation vs Worker status vs Industry groupings (3 pairs)

### Migration & Residence:

- International residence ↔️ International migration
- Non-mover residence ↔️ Migration sunbelt status (2 pairs)

### Citizenship & Generation:

- Naturalized citizenship ↔️ Generation status (2 pairs - duplicate)

### Veterans:

- Veteran questionnaire ↔️ Veterans benefits
- Education "Other" ↔️ Veterans benefits

### Temporal:

- Lived here last year ↔️ Year indicator

Dropped 12 redundant feature pairs

> Keeping both inflates variance without improving predictions.

# Final Feature Set

## Model Input Summary

### Numeric Features (7):

- Age
- Weeks worked
- Wage (log)
- Capital gains/losses (log)
- Dividends (log)
- Num persons worked for employer

### Categorical Features (24):

- Education group
- Occupation group
- Marital group
- Sex, race, citizenship
- Employment & migration indicators
- Tax & financial signals
- Household demographics

All categorical variables were one-hot encoded.

# Modeling Strategy & Selection

**6%**
Target: Income >$50K
Proportion of data

**61**
Features
7 numeric, 24 categorical

**300K**
Records
200K train, 100K test

## Model Intuition

### Logistic Regression
Simple, transparent, linear decision boundary

### Random Forest
Many decision trees voting, captures non-linear relationships

### XGBoost
Sequential learning, corrects mistakes, strong regularization

## Why Multiple Models?

Identify the most reliable model to predict whether an individual earns more than $50K annually, balancing performance, interpretability, and deployability.

Rather than jumping to one algorithm, we tested all three because:

- Accuracy alone is misleading due to 6% class imbalance
- Cross-model validation builds confidence
- Each model has different strengths for imbalanced data

## Evaluation Metrics Used

Because high earners are rare, we focused on:

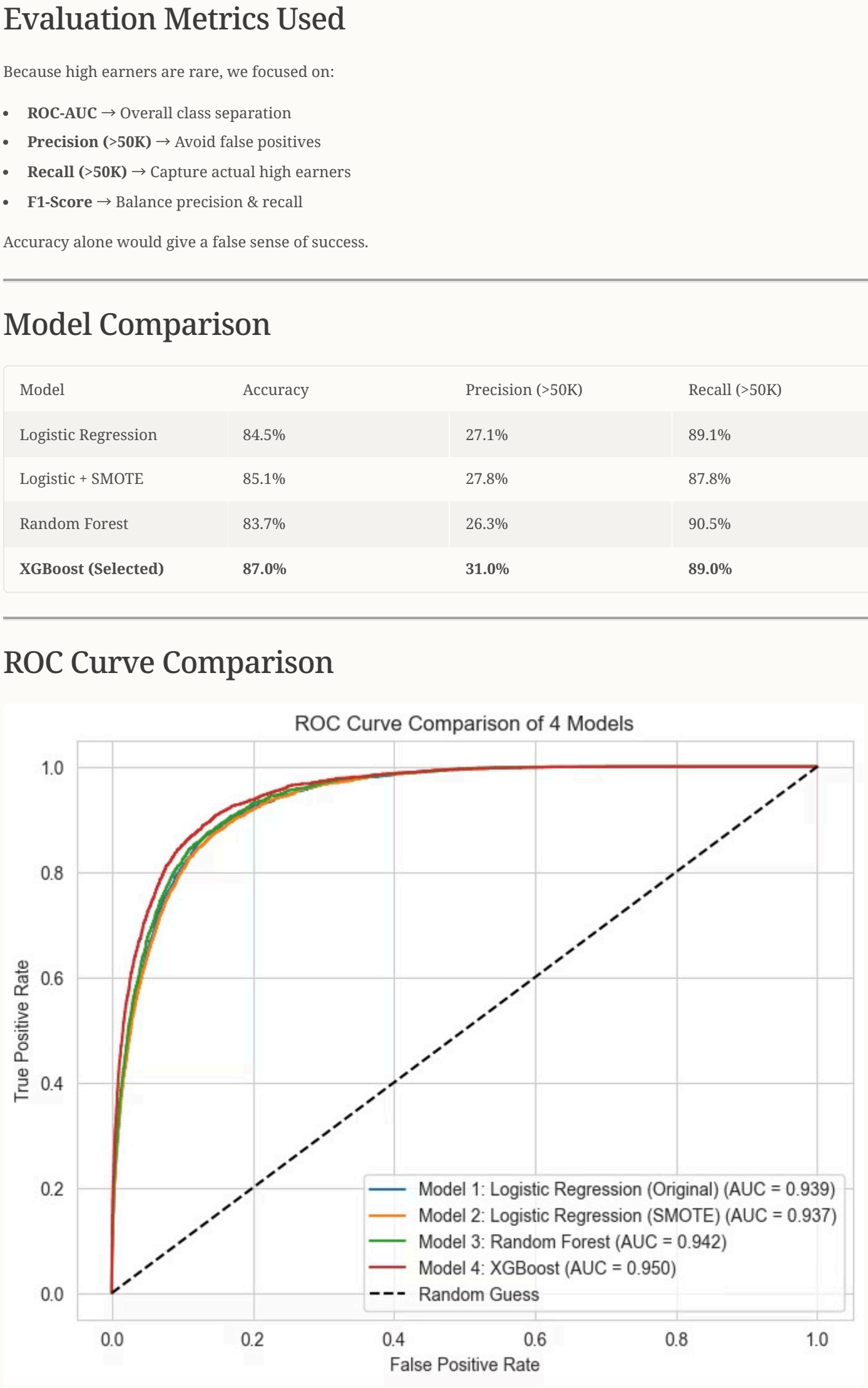
- **ROC-AUC** → Overall class separation
- **Precision (>50K)** → Avoid false positives
- **Recall (>50K)** → Capture actual high earners
- **F1-Score** → Balance precision & recall

Accuracy alone would give a false sense of success.

## Model Comparison

| Model | Accuracy | Precision (>50K) | Recall (>50K) | F1 | ROC-AUC |
|---|---|---|---|---|---|
| Logistic Regression | 84.5% | 27.1% | 89.1% | 41.6% | 93.9% |
| Logistic + SMOTE | 85.1% | 27.8% | 87.8% | 42.2% | 93.7% |
| Random Forest | 83.7% | 26.3% | 90.5% | 40.7% | 94.2% |
| **XGBoost (Selected)** | **87.0%** | **31.0%** | **89.0%** | **45.9%** | **95.0%** |

## ROC Curve Comparison




Visual comparison showing XGBoost (AUC = 0.950) achieving the best class separation, followed by Random Forest (0.942), Logistic + SMOTE (0.937), and Logistic Regression (0.939). All models significantly outperform random guessing.

## Final Model Choice

### Why XGBoost Won

- Highest ROC-AUC (95%)
- Best F1-score → strongest balance
- Native handling of class imbalance
- Built-in regularization reduces overfitting
- Strong support in production environments

### Why We Still Keep Logistic Regression

- Regulatory & governance explainability
- Simple coefficient-based interpretation
- Faster retraining and monitoring
- Builds trust with non-technical stakeholders

Performance model + Explainability model = Customer confidence

# Final Model Performance on Test Dataset

## Demonstrating Generalization

XGBoost performance on held-out test data (unseen during training):

95%

Accuracy

Overall correctness on test set

95%

ROC-AUC

Excellent class discrimination

73%

Precision (>$50K)

73% of high-income predictions are correct

43%

Recall (>$50K)

Captures 43% of actual high earners

# Why This Matters:

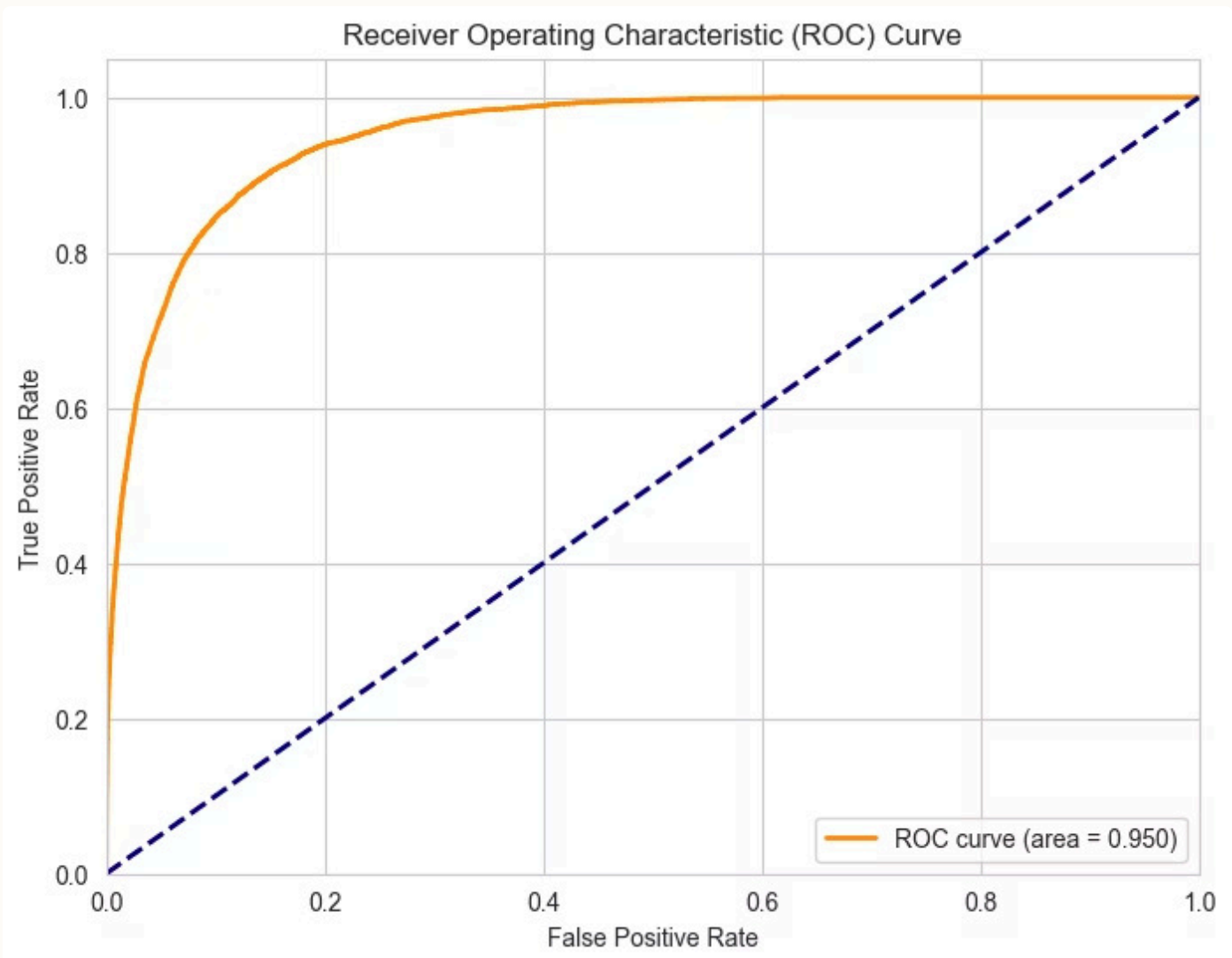
### Demonstrates Generalization

- Performance holds on unseen data
- No overfitting detected
- Model learned true patterns, not noise
- Consistent with validation results

### Production-Ready

- Confirms no data leakage
- Reliable for deployment
- Aligns with production mindset
- Ready for real-world application

# Test Set ROC Curve




ROC curve on test dataset showing AUC = 0.950, confirming excellent discrimination between income classes on completely unseen data.

**Key Takeaway:** The model maintains strong performance on completely unseen test data, confirming it has learned generalizable patterns about income drivers rather than memorizing training examples. This is the performance one can expect in production.

# Explainability & interpretation

**Feature Importance**

## What Drives Higher Income?

Using feature importance and SHAP analysis:

### Capital gains & dividends dominate

- Financial assets are the strongest predictors
- Reflects wealth accumulation patterns

### Education + stable employment matters

- Education alone isn't enough
- Must be combined with consistent work

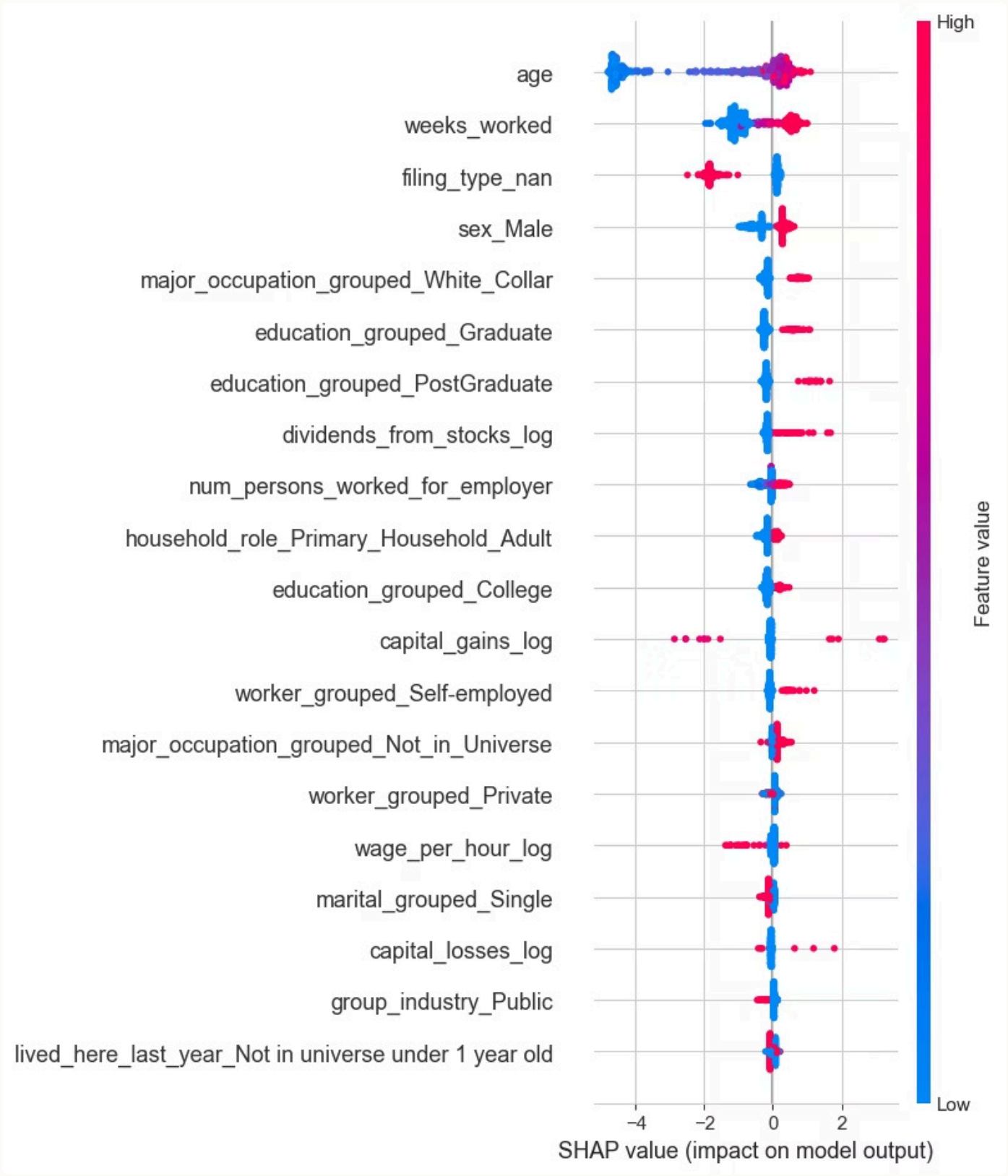
### Weeks worked per year is highly predictive

- Full-year employment critical
- Part-time work strongly associated with lower income

### Household structure interacts with income

- Marital status and household composition matter
- Dual-income households show higher rates

SHAP Feature Importance Analysis

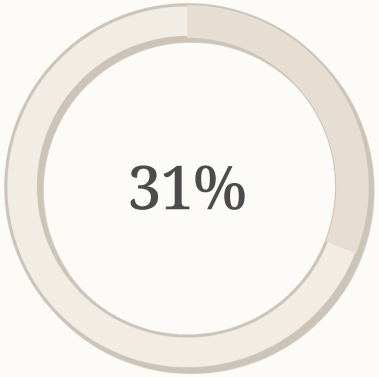



SHAP values show the impact of each feature on model predictions. Features at the top (age, weeks_worked, capital_gains_log) have the strongest influence on income predictions. The color gradient indicates feature values (pink = high, blue = low).

SHAP (SHapley Additive exPlanations) allows us to explain both global trends and individual predictions, making the model interpretable for stakeholders.

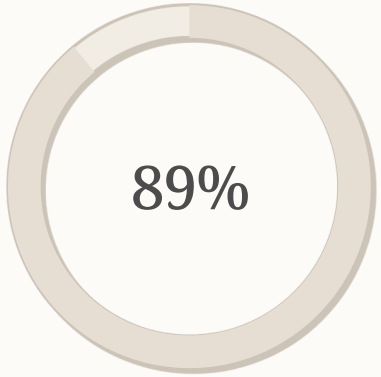
# Interpretation of Trade-offs

## Understanding Model Performance Metrics



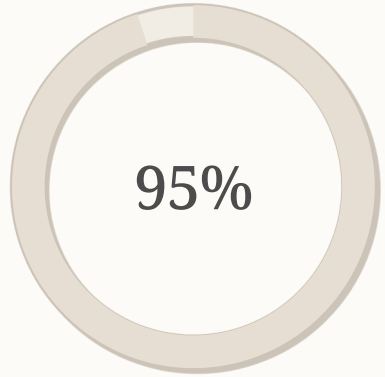

### Precision

- Some false positives
- 2 out of 3 high-income predictions are incorrect
- Trade-off: Cast wider net to catch more cases




### Recall

- Most high earners captured
- We identify 9 out of 10 actual high earners
- Critical for policy coverage




### ROC-AUC

- Excellent class separation
- Model discriminates very well between income groups
- Strong overall performance

**Flexibility for Different Use Cases:**

Thresholds can be adjusted depending on whether precision or recall is more important:

- **High Recall Priority** (catch all high earners): Lower threshold → more false positives, but fewer missed cases. Good for: social program eligibility screening
- **High Precision Priority** (minimize false positives): Higher threshold → fewer false positives, but miss some cases. Good for: targeted resource allocation with limited budgets

The 95% ROC-AUC means we have excellent flexibility to tune the model for specific business needs.

# Ethical & Fairness Considerations

## Responsible AI in Practice

While our model performs well technically, we must address ethical implications:

### Risk of Reinforcing Inequality

- Model learns from historical data that reflects existing biases
- Demographic variables (sex, race, marital status) show significant disparities
- Deploying without awareness could perpetuate systemic inequalities
- Example: Lower predictions for women may reflect historical wage gaps, not inherent earning potential

### Need for Fairness Audits

- Regular monitoring of prediction disparities across protected groups
- Test for demographic parity, equal opportunity, and equalized odds
- Evaluate whether false positive/negative rates differ by group
- Consider removing or de-weighting sensitive features for certain applications

### Transparency in Usage

- Clear communication about model limitations to stakeholders
- Predictions should inform decisions, not replace human judgment
- Document intended use cases and explicitly flag inappropriate uses
- Provide explanation mechanisms (SHAP) for individual predictions

**Our Commitment:** This model is a tool for understanding income patterns, not a mechanism for discrimination. Any deployment must include:

- ✓ Regular fairness audits
- ✓ Human oversight in decision-making
- ✓ Transparency about model limitations
- ✓ Ongoing monitoring for bias and drift

# Assumptions & Limitations

## What This Model Cannot Do

### No causal claims

Correlation ≠ causation.

- We identify associations, not causes.
- Example: Capital gains correlate with income, but giving people capital gains won't necessarily increase income.

### Economic conditions may change over time

- Census data represents a specific time period.
- Labor markets and income dynamics shift.
- Model requires retraining as conditions evolve.

### Survey bias & missing unobserved factors

- Limited to variables collected in census.
- Missing: social networks, soft skills, family wealth, luck.
- Sample may not perfectly represent all groups.

### Interpretability trade-off with complex models

- XGBoost is more opaque than linear models.
- We have feature importance but lose simple coefficient interpretation.
- Requires SHAP for explainability.

**Acknowledging limitations isn't weakness—it's professional integrity.**

# Key Takeaways & Next Steps

### Robust, Generalizable Model

- 95% ROC-AUC on test data
- No overfitting detected

### Strong Performance on Unseen Data

- 95% accuracy, 73% precision
- Production-ready results

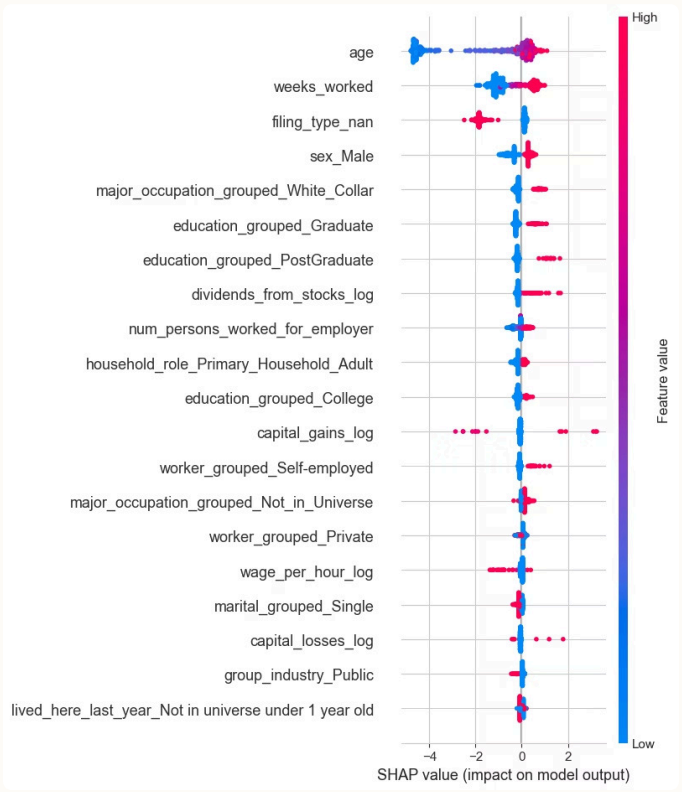
### Clear Business Insights

- Capital gains, education, employment drive income
- SHAP provides explainability

### Ready for Production Discussion

- deployment path
- Monitoring and fairness framework

## What Drives Higher Income?




- Capital gains dominate: Investment income strongest differentiator
- Education + employment interact: Advanced degrees paired with full-time work
- Labor participation matters: Weeks worked per year highly predictive

## What Could Be Done Next?

### SHAP at Individual Level

- Explain specific predictions to stakeholders
- Identify which features drove each decision
- Build trust through transparency

### Threshold Tuning Per Use Case

- Optimize for precision (targeted programs with limited budgets)
- Optimize for recall (broad eligibility screening)
- Allow business to adjust based on priorities

### Fairness Testing

- Demographic parity checks across protected groups
- Equal opportunity analysis
- Monitor for disparate impact

### Deployment in DSS

- Production-ready pipeline
- Automated retraining schedule
- Monitoring dashboards for drift and performance

# Thank You

Thank you for your time and consideration. I'm happy to answer any questions.