Enhancing the CPS with Administrative Tax Data

Machine Learning Meets Microsimulation

Nikhil Woodruff & Max Ghenis

PolicyEngine

National Tax Association Annual Conference
November 14, 2024

POLICY ENGINE

The Challenge

POLICY ENGINE

- Current Population Survey (CPS)
  - Rich demographics, state ID, household structure
  - But underreports income, especially at top
  - Limited tax information

The Challenge

- Current Population Survey (CPS)
  - Rich demographics, state ID, household structure
  - But underreports income, especially at top
  - Limited tax information
- IRS Public Use File (PUF)
  - Accurate tax records from administrative data
  - But no demographics beyond age/sex
  - No state ID
  - Protected by strict confidentiality rules

Key Innovation: Open Dataset with PUF Quality

POLICY ENGINE

- First openly available dataset integrating CPS and PUF
- No confidentiality restrictions
- Enables:
  - Transparent policy analysis
  - Integration with other tools
  - Community contributions and validation
- Powers PolicyEngine's microsimulation platform

Our Approach

POLICY ENGINE

- Generate synthetic tax variables from PUF using:
  - Quantile regression forests to learn distributions
  - Match to CPS demographic variables

Our Approach

POLICY ENGINE

- Generate synthetic tax variables from PUF using:
  - Quantile regression forests to learn distributions
  - Match to CPS demographic variables
- Stack synthetic records with original CPS

Our Approach

POLICY ENGINE

- Generate synthetic tax variables from PUF using:
  - Quantile regression forests to learn distributions
  - Match to CPS demographic variables
- Stack synthetic records with original CPS
- Calculate taxes and benefits via microsimulation

Our Approach

POLICY ENGINE

- Generate synthetic tax variables from PUF using:
  - Quantile regression forests to learn distributions
  - Match to CPS demographic variables
- Stack synthetic records with original CPS
- Calculate taxes and benefits via microsimulation
- Optimize weights against 570 targets:
  - IRS Statistics of Income income bins
  - Program participation totals
  - Single-year age population counts

Validation: Administrative Targets

POLICY ENGINE

- Enhanced CPS outperforms source datasets:
  - 63% of targets vs original CPS
  - 71% of targets vs PUF

Validation: Inequality Measures

POLICY ENGINE

- Tax unit income inequality matches IRS data

Table: Key tax unit-level distributional metrics

Metric	CPS	Enhanced CPS	PUF
Gini coefficient	0.495	0.572	0.570
Top 10% share	0.361	0.425	0.410
Top 1% share	0.085	0.154	0.150

Validation: Top Rate Reform

POLICY ENGINE

- Biden's 2025 budget: raise top rate to 39.6% above $400k

Table: Projected revenue from top rate increase, 2025

Source	Revenue (billions)
Treasury	$75.4
Enhanced CPS	$75.7
Baseline CPS	$28.7

Advanced ML Methods

POLICY ENGINE

- Quantile regression forests (QRF) for:
  - Tax variable imputation from PUF
  - Housing costs from ACS
  - Prior year earnings from ASEC panel

Advanced ML Methods

POLICY ENGINE

- Quantile regression forests (QRF) for:
  - Tax variable imputation from PUF
  - Housing costs from ACS
  - Prior year earnings from ASEC panel
- Dropout-regularized gradient descent for:
  - Weight optimization
  - Preventing overfitting
  - Handling sparse subgroups

The PolicyEngine Platform

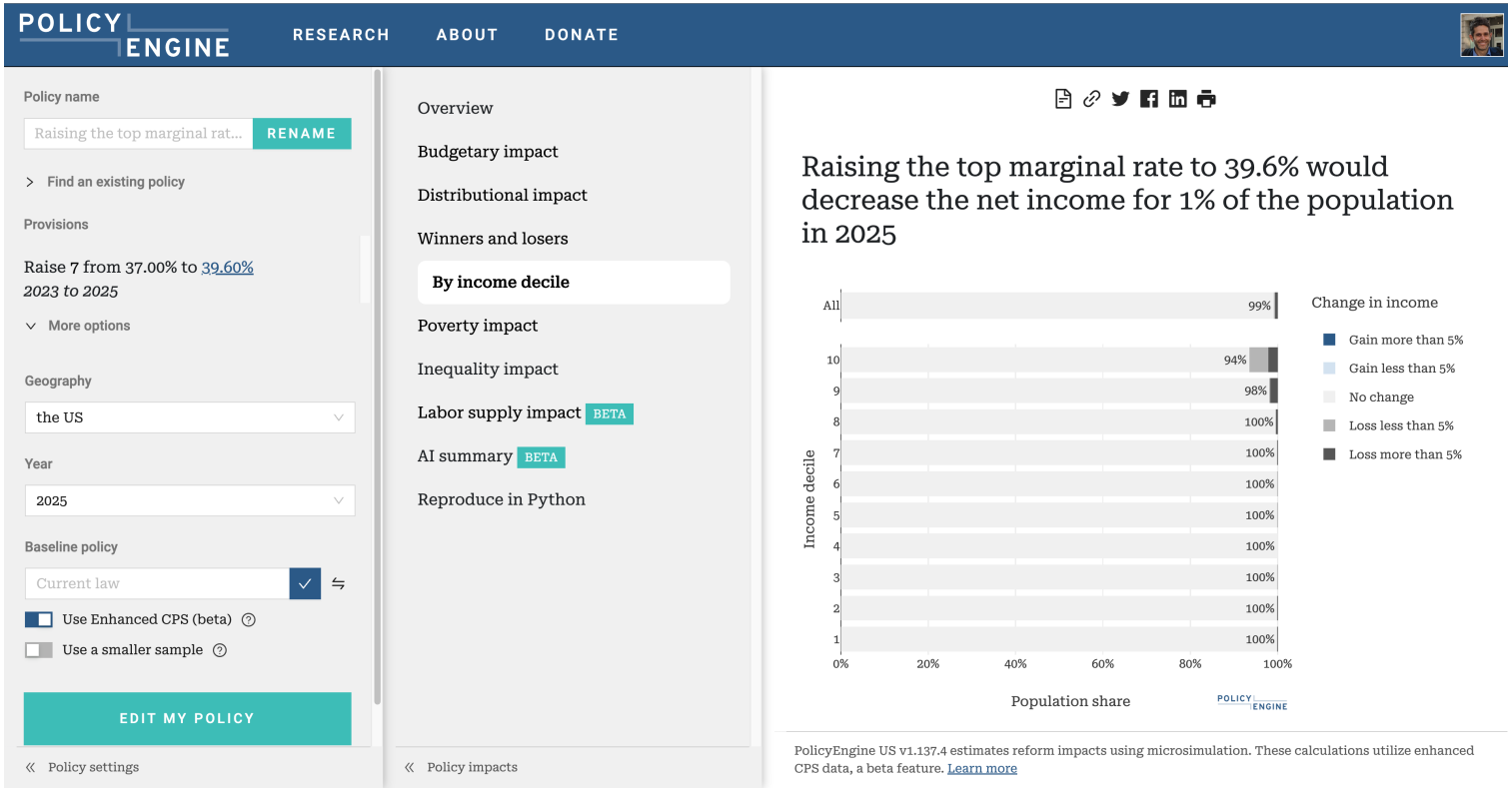

Figure: Screenshot of PolicyEngine's US tax and benefit calculator

- Web interface for policy analysis
- Powered by enhanced CPS
- Instant distributional impacts

Open Access

POLICY ENGINE

- All code open source
  - github.com/PolicyEngine/policyengine-us-data
  - Reproducible enhancement pipeline
  - Automatic validation dashboard

Open Access

POLICY ENGINE

- All code open source
  - github.com/PolicyEngine/policyengine-us-data
  - Reproducible enhancement pipeline
  - Automatic validation dashboard
- Use through Python package or web interface

Open Access

POLICY ENGINE

- All code open source
  - github.com/PolicyEngine/policyengine-us-data
  - Reproducible enhancement pipeline
  - Automatic validation dashboard
- Use through Python package or web interface
- Growing community of users:
  - Academic researchers
  - Think tanks
  - Government agencies

Next Steps

POLICY ENGINE

- Geographic extensions:
  - Congressional district weights
  - State-specific calibration
  - County-level synthetic data

Next Steps

- Geographic extensions:
  - Congressional district weights
  - State-specific calibration
  - County-level synthetic data
- Methodological improvements:
  - Time series validation
  - Uncertainty quantification
  - Alternative ML architectures

Thank You

POLICY ENGINE

- Paper: github.com/PolicyEngine/policyengine-us-data/paper
- Code: github.com/PolicyEngine/policyengine-us-data
- Web app: policyengine.org
- Contact: max@policyengine.org