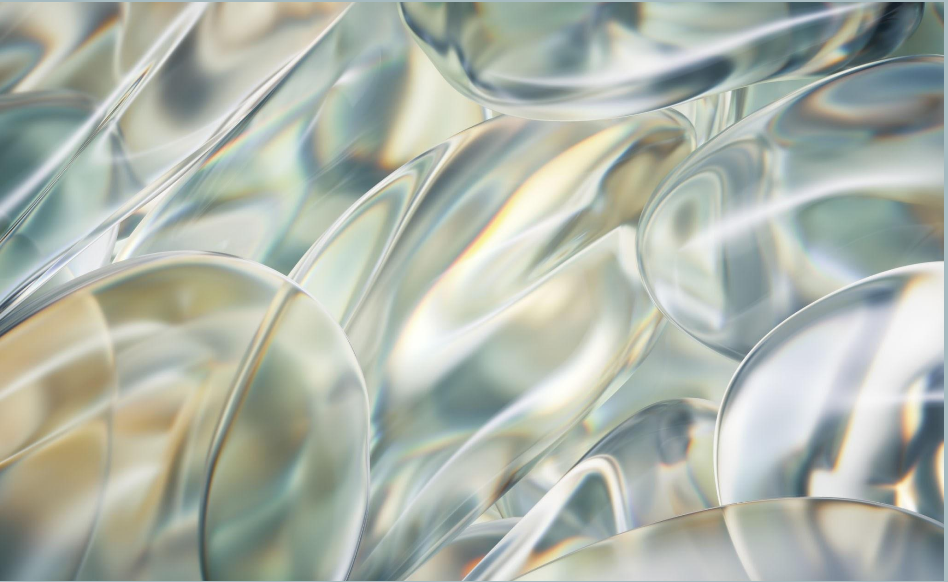
# Predicting Smoking Cessation Success Using Machine Learning



A Multi-Model Approach with PATH Study Data

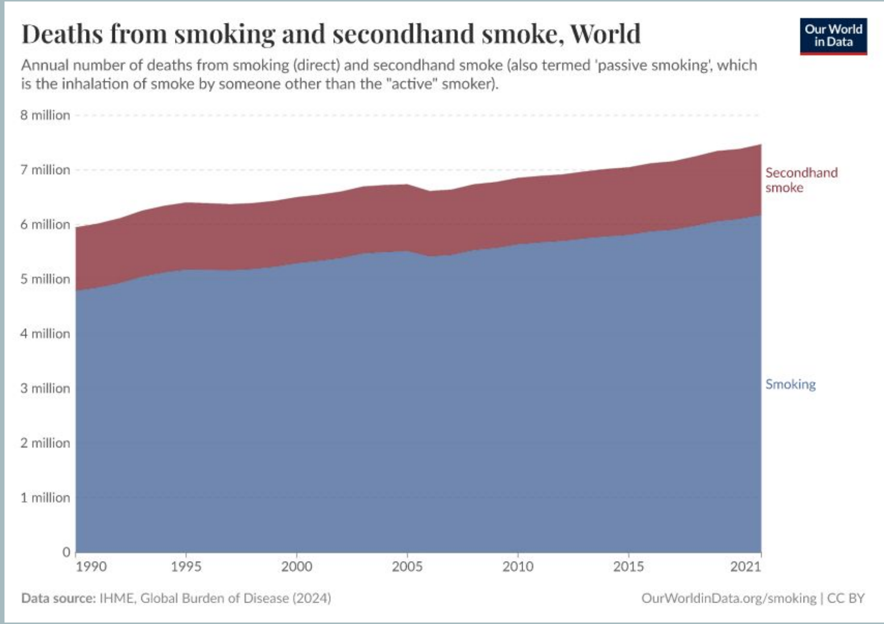
By: Angel Nivar, Ananda Downing

# Why Smoking Cessation Prediction Matters

- Smoking remains the leading cause of preventable death in the US (~480,000 deaths/year)
- Only 7% of quit attempts succeed without support
- Healthcare resources for cessation interventions are limited

**Problem:** How do we identify which smokers will benefit most from intervention?

And how do we know which intervention methods work best in a given scenario?




INTRODUCTION

# Research Question & Data Source

What factors predict successful smoking cessation after a quit attempt?

- **Data Source: Population Assessment of Tobacco and Health (PATH) Study**
  - Longitudinal survey by FDA and NIH
  - Waves 1-5 (2013-2019) → Extended to Wave 7
  - National probability sample of US adults

# Tool & Tech Stack

**Environment**

- **Python 3.10+** with virtual environment isolation
- **Jupyter Notebooks** for exploratory analysis and documentation
- **Git/GitHub** for version control

**Data Source:**

- PATH [Population Assessment for Tobacco & Health] Study (ICPSR) — **STATA** format (.dta)
- 7 waves of longitudinal adult smoking data

**Key Design Choices:**

- XGBoost native NaN handling (no imputation required)
- Class weighting over SMOTE (preserves distribution)
- Person-level train/test split (prevents data leakage)

**VENV**

| | |
|---|---|
| Preprocessing | Pandas<br>NumPY<br>PyReadStat |
| ML | Scikit-Learn<br>XGBoost |
| Interpretability | SHAP |
| Visualization | Mathplotlib<br>Seaborn<br>Plotly |
| Dashboard | Streamlit |
| Imbalanced Data | Imbalance<br>learn |

METHODOLOGY

# Data Lineage

**Stage 1: Raw Data Acquisition**

- **Source:** PATH Study via ICPSR (FDA/NIH longitudinal survey)
- **Format:** STATA (.dta) files, 7 waves (2013-2020)
- **Size:** ~32,000 adults per wave, 1,700+ variables each

**Stage 3: Feature Engineering**

- Map raw PATH variables to 43 features
- **Categories:** Demographics, Dependence, Cessation Methods, Environment, Interactions
- Wave-aware variable naming

  ( R01_AC1002 → R02_AC1002)

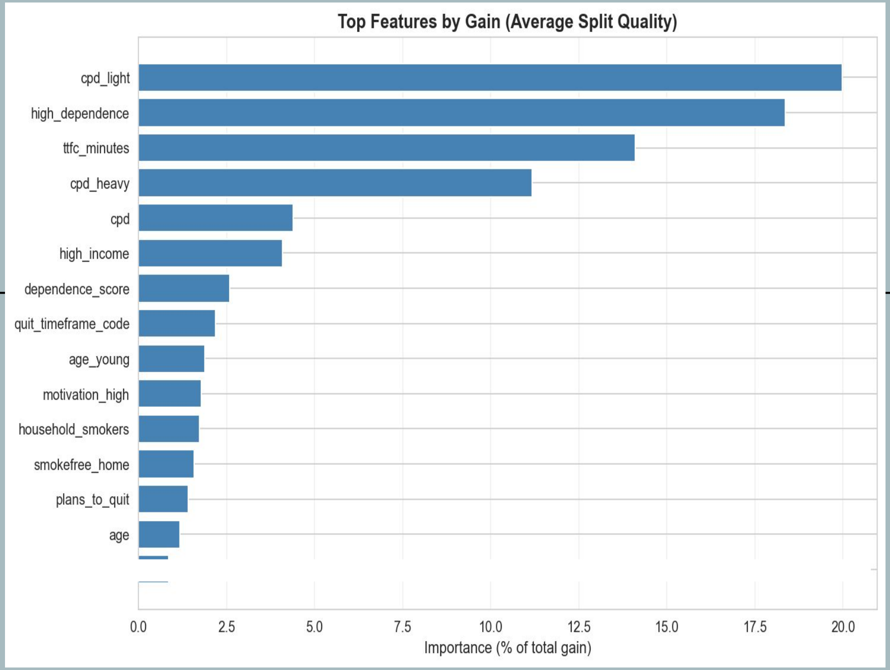
**Stage 2: Person-Period Construction**

- Identify current smokers at each wave (smoking in past 30 days)
- Create transitions: W1→W2, W2→W3, … W6→W7
- Merge baseline features with follow-up smoking status
- Define outcome: quit_success = 1 if not smoking at follow-up
- Handle PATH missing codes (-9, -8, -7, -1 → NaN)
- **Output:** ~60,000 transitions from 24,576 unique persons

**Stage 4: Modeling Splits**

- Split by person_id (not observation) to prevent leakage
- 60% train / 20% validation / 20% test
- Class weighting applied (72% no-quit majority)
- Used SMOTE to address imbalance for early ML testing

# Feature Engineering




**43 Features Across 5 Categories:**

- **Nicotine Dependence:** Time to First Cigarette (TTFC), Cigarettes Per Day (CPD), dependence score
- **Demographics:** Age, sex, race/ethnicity, education, income
- **Cessation Methods:** NRT (patch, gum), Varenicline, Bupropion, Counseling, Quitline
- **Environment:** Smoke-free home policy, household smokers, workplace policies
- **Interactions:** High-dependence × Varenicline, Young × Counseling, NRT + Medication combos

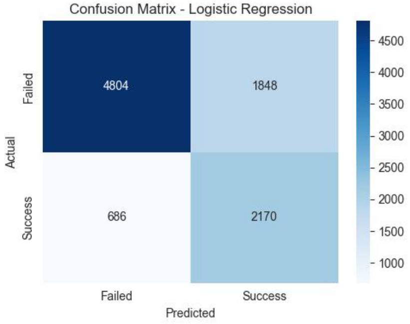
METHODOLOGY

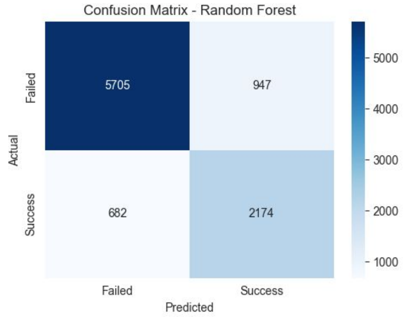
# Modeling Pipeline

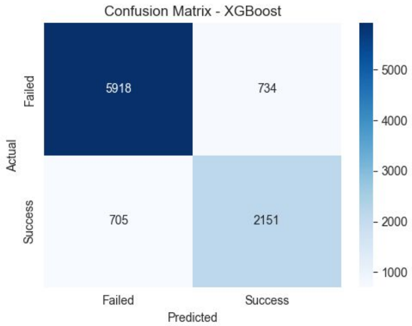
## Model Selection:

- Structured Data

1. **Logistic Regression** — Interpretable baseline
2. **Random Forest** — Handles High Dimensionality
3. **XGBoost** — Efficient predictive model (best performer)



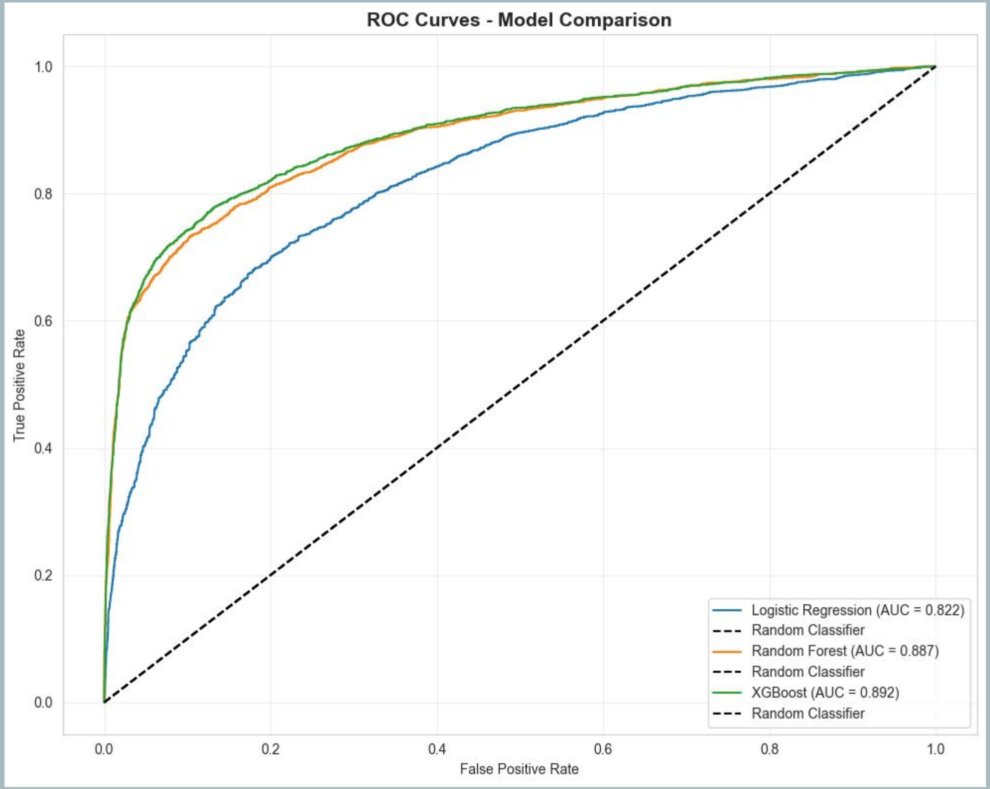




## Critical Design Decisions

- Split by person_id (not observation) to prevent data leakage
- Class weighting enabled in ALL models (72% no-quit majority class)
- 60/20/20 train/validation/test split
- Published benchmark: Issabakhsh et al. (2023) achieved **0.72 AUC** on similar data

# Model Performance Comparison

| Model | ROC-AUC | PR-AUC | F1 |
|---|---|---|---|
| Logistic Regression | .822 | .698 | .631 |
| Random Forest | .887 | .827 | .727 |
| XGBoost | .892 | .834 | .759 |

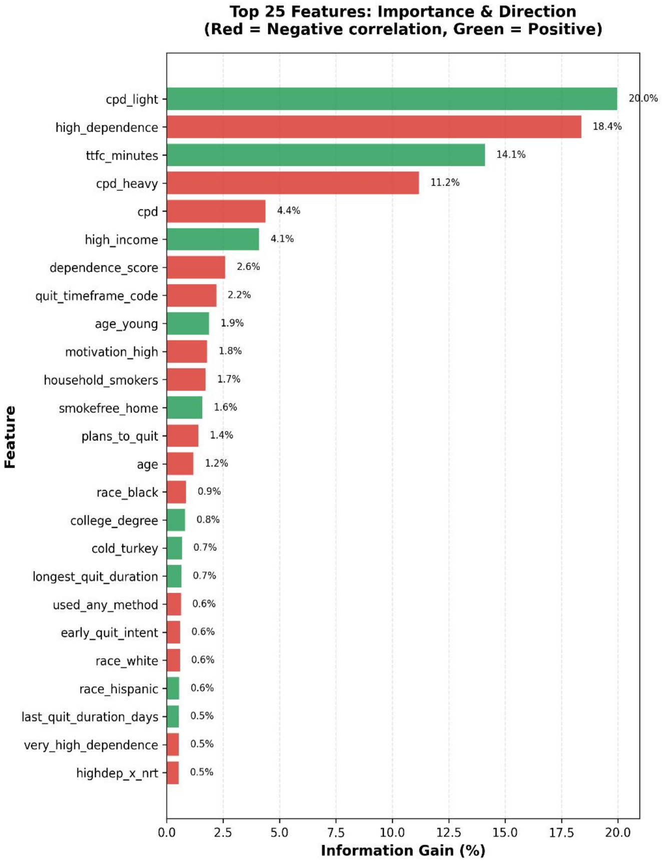



RESULTS

# What Drives Predictions?

**Top 5 Predictors (XGBoost Feature Importance):**

1. **CPD Light (<3 cigs/day):** 20.0% importance — Light smokers much more likely to quit
2. **High Dependence Score:** 18.4% — Strongest barrier to quitting
3. **Time to First Cigarette:** 14.1% — Earlier = harder to quit
4. **CPD Heavy (10+/day):** 11.2% — Strong negative predictor
5. **Cigarettes Per Day:** 4.4% — Dose-response relationship




**Clinical Interpretation:**

- Nicotine dependence measures dominate predictions
- Modifiable factors (medications, counseling) show positive effects when combined

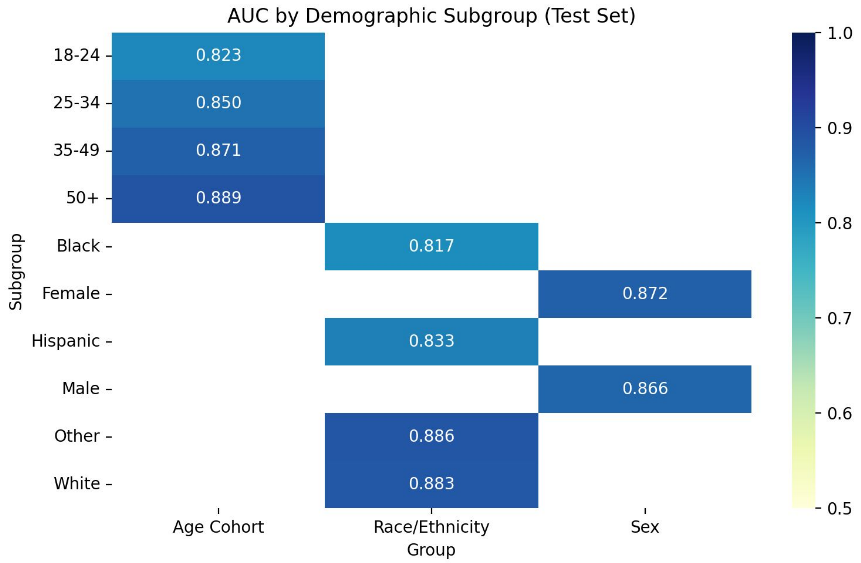
# Fairness Analysis

**Performance Across Demographic Subgroups:**

Analyzed by: Sex, Age cohort, Race/Ethnicity

**Key Finding:** Limited disparity data available in processed dataset, but framework established for:

- AUC comparison across groups
- False positive/negative rate parity
- Base rate differences



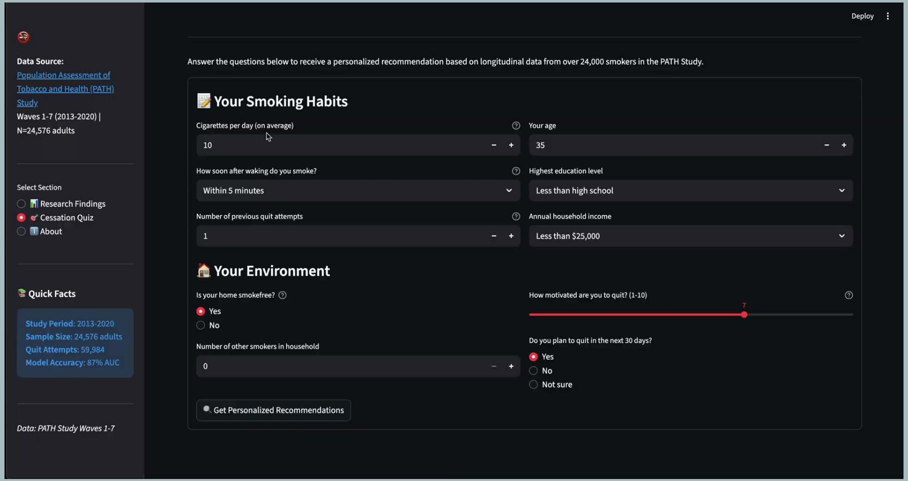

**Implications for Deployment:**

- Transparent reporting of any disparities is critical
- Model should be monitored for performance drift across demographic groups
- Additional data collection may be needed for underrepresented groups

- **High AUC (≈0.80–0.90):** Good discrimination in that subgroup; scores for quitters and non-quitters are well separated.
- **AUC ≈0.50:** Near-random ranking in that subgroup; the model struggles to tell them apart.

# Interactive Prediction Tool

- **Overview:** Project summary and key metrics
- **Model Performance:** ROC curves, confusion matrix
- **Feature Importance:** Interactive SHAP visualizations





- **Prediction Tool:** Input patient characteristics → get quit probability
- **Fairness Assessment:** Performance by demographic group
- **Key Insights:** Actionable clinical recommendations

RESULTS

# Key Takeaways & Recommendations

**Main Findings:**

1. XGBoost achieves 0.83 validation AUC (exceeds published benchmark)
2. Nicotine dependence measures are the strongest predictors
3. Light smokers (<3 CPD) have dramatically higher quit success
4. Combination therapy (medication + counseling) improves outcomes

**Limitations:**

- Self-reported outcomes (no biochemical verification)
- Validation-to-test performance gap needs investigation
- Missing data in key variables (77% CPD missing)
- Temporal limitations (2013-2019 data)

**Future Work:**

- Hyperparameter tuning and cross-validation
- Stratified models by demographic groups
- Prospective validation in clinical settings
- Integration with electronic health records

# Thank You