Risk Factors for Adolescent Nicotine Vaping from 2017 to 2023: A Data-Driven Approach

PennState Institute for Computational and Data Sciences

PennState College of the Liberal Arts

CSUA Consortium on Substance Use and Addiction

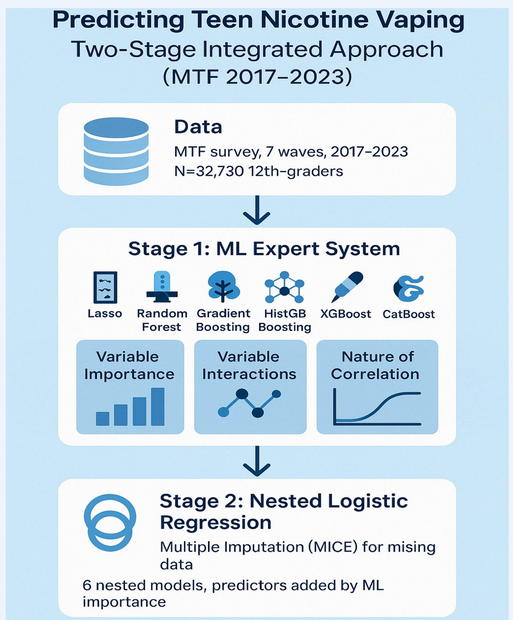
siyangni@psu.edu
https://github.com/siyangni
https://www.linkedin.com/in/siyangni/

Siyang Ni, MPhil, PhD Candidate
Center for Social Data Analytics, Department of Sociology and Criminology, Pennsylvania State University, University Park, PA

Scan Me

PROBLEM

Method

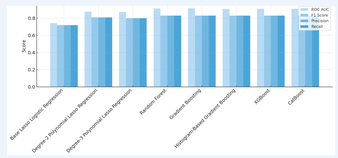


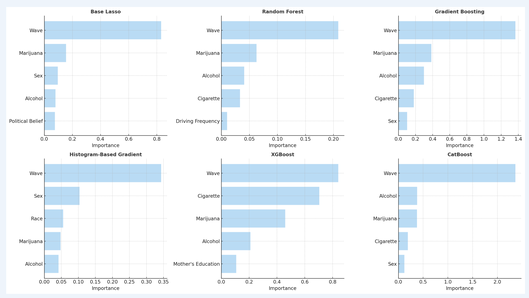
Monitoring the Future (MTF) Survey is an ongoing, nationally representative annual survey tracking substance use, attitudes, and behaviors among U.S. adolescents (8th, 10th, 12th graders) since 1975.

The Expert System

- Trained six supervised ML models (Lasso, RF, GB, HGB, XGBoost, CatBoost) to predict vaping status.
- Used model-agnostic interpretation:
  - **Variable Importance:** Identified key predictors (Coefficients, Tree Importance, SHAP).
  - **Variable Interactions:** Explored joint effects (SHAP Interaction Indices, Polynomial Terms).
  - **Nature of Correlation:** Visualized relationships using Partial Dependence Plots (PDPs).

Tree-based models (esp. GB, CatBoost, XGBoost) significantly outperformed linear models (ROC AUC > 0.90 vs. 0.74 for base Lasso), indicating non-linearities within the dataset.



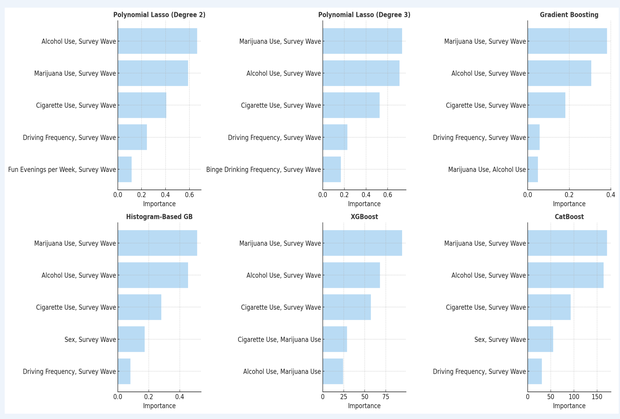


Important Predictors of Nicotine Vaping

- Survey Wave (wave)
- Past-12-Month Marijuana Use (V2116)
- Past-12-Month Alcohol Use (V2105)
- Lifetime Cigarette Use (V2101)
- Demographic Factors (sex, race, Region - V13)
- Political Belief (V2166)
- Driving Frequency (V2196) & Fun Evenings Out (V2194)
- Mother's Education (V2164), Average Grade (V2179), Self-Rated School Ability (V2173)

Survey wave interacts with three main types of predictors:

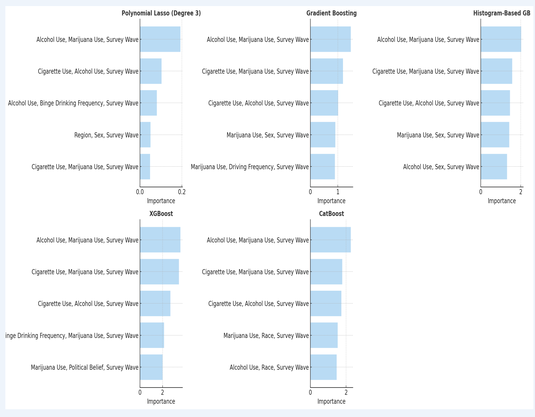
- Substance Use x Wave
- Behavioral/Social Factors x Wave
- Demographic/SES Factors x Wave



Important Two-way Interactions

- (Marijuana Use, Survey Wave)
- (Alcohol Use, Survey Wave)
- (Cigarette Use, Survey Wave)
- (Driving Frequency, Survey Wave)

Synergistic Effects of Polysubstance Use: Different substance use variables also interact with each other, independent of the survey wave.



The most consistently important and highly ranked interactions across multiple models involve **Survey Wave (wave)** interacting with **two other key predictors,** indicating that the combined influence of various factors on vaping likelihood is not static but **dynamically changes over time.**

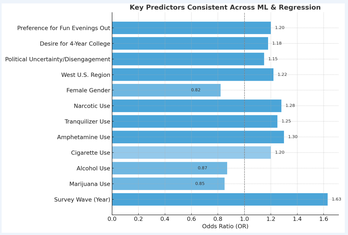
Important Three-way Interactions

- (Alcohol Use, Marijuana Use, Survey Wave)
- (Cigarette Use, Marijuana Use, Survey Wave)
- (Cigarette Use, Alcohol Use, Survey Wave)

While the substance-use interactions with wave were dominant, other patterns **involving demographic or social factors interacting with wave** and another predictor also emerged frequently.

The Logistic Regression

- **6 nested models**, incrementally adding predictors based on their ranking frequency across the ML models (from predictors ranked as top 20 important by all models to those ranked top 20 by at least one model).
- **Use Odds Ratios and p-value** to validate ML findings within a traditional inference framework.



Discussion