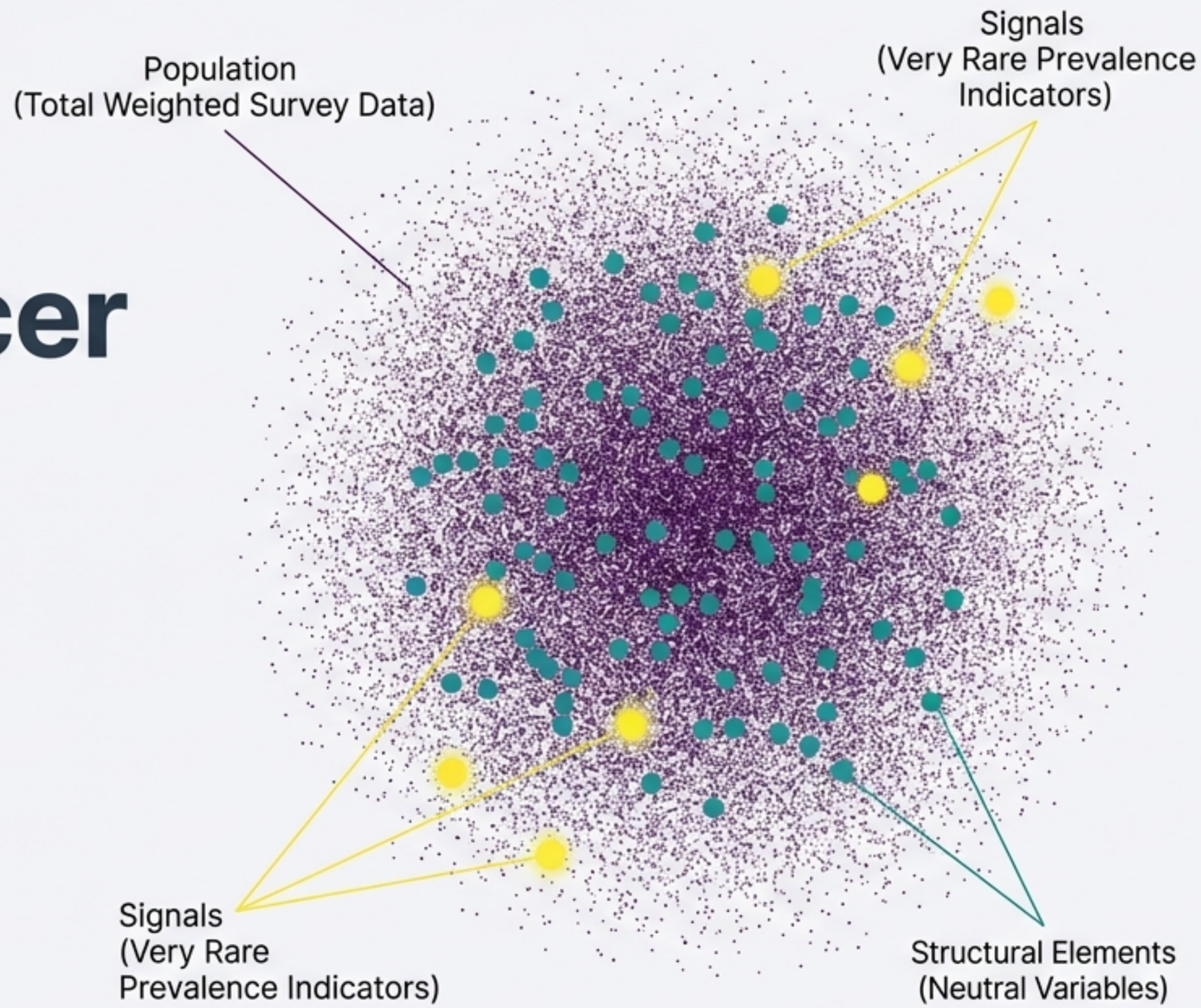
Predicting Cancer Prevalence in BRFSS 2023

A Comparative Analysis of Linear vs. Tree-Based Models on Weighted Survey Data



Based on the 2023 Behavioral Risk Factor Surveillance System (CDC)

The Dataset: 433,000 Human Stories

Volume

433,323 Rows

350 Columns

Target Variable

CHCOCNC1

Self-Reported Cancer Diagnosis

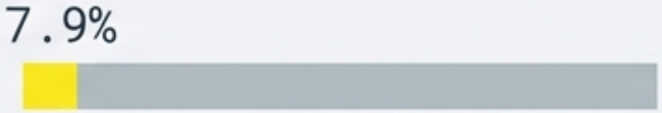
Source

CDC BRFSS

2023 Survey Data

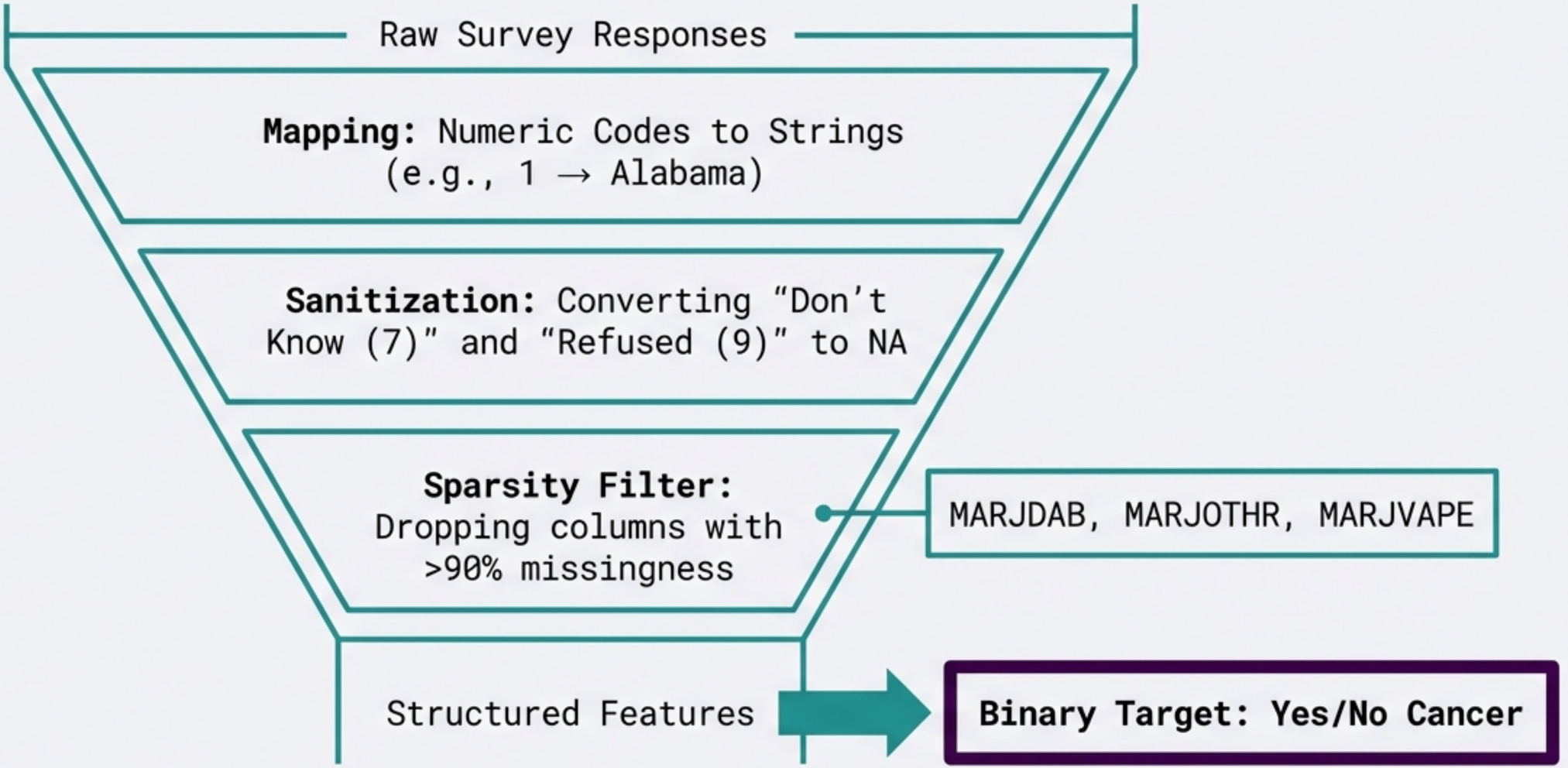
Class Imbalance

7.9%



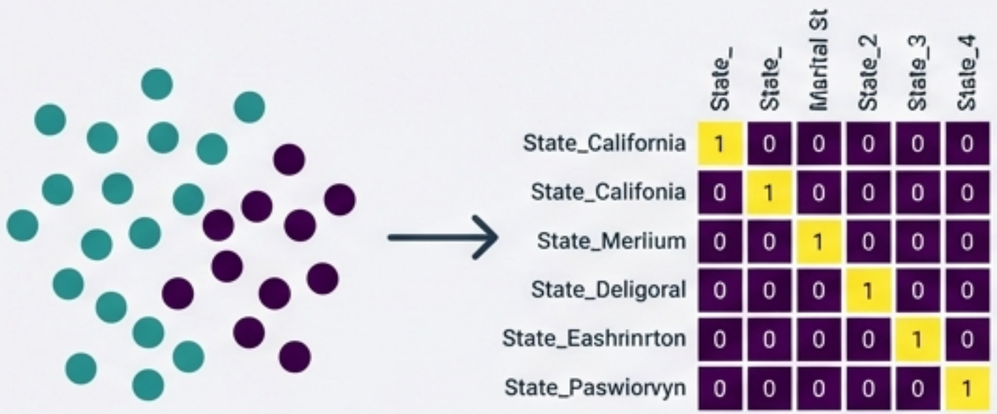
Weighted Prevalence

From Raw Survey Responses to Structured Features



Feature Engineering Strategy

Nominal Features

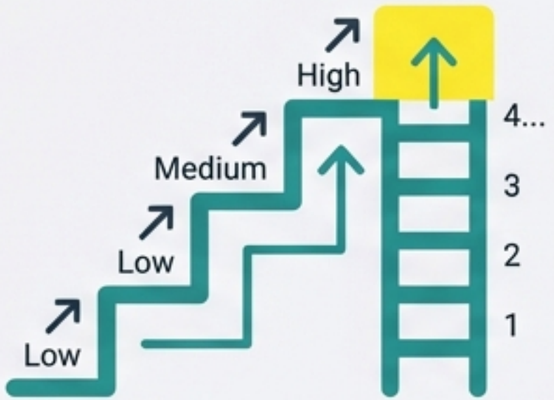


Technique: One-Hot Encoding

Examples: State, Marital Status, Race (_RACEPRV)

Missing Data: Imputed as explicit ‘Unknown’ category to capture signal.

Ordinal & Numeric Features

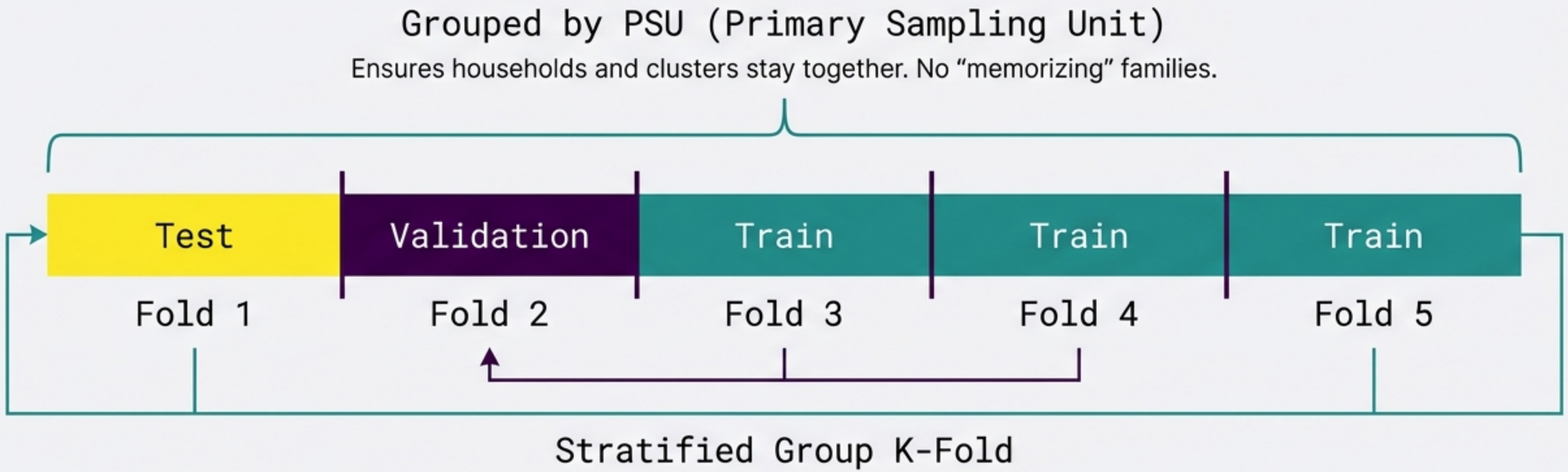


Technique: Preserving Hierarchy

Examples: Education, Income, BMI (_BMI5CAT), General Health

Missing Data: Median imputation for continuous variables.

The Validation Framework: Preventing Leakage



- Outer Loop: 5 Splits
- Inner Loop: 4 Splits
- Result: 5,265 unique groups in Test Set

Modeling the Population, Not the Sample

Variable: _LLCPWT (Final Weight)

Training: Models fit with sample_weight parameters.

Evaluation: AUPRC and ROC-AUC calculated using weighted scoring.

Normalization: Weights normalized to mean=1.0 to stabilize training.

The Contenders: Linear vs. Non-Linear

Logistic Regression

Role: The Baseline

Strength: Interpretable, Fast

Specs: L2 Regularization, OneHotEncoder (drop first)

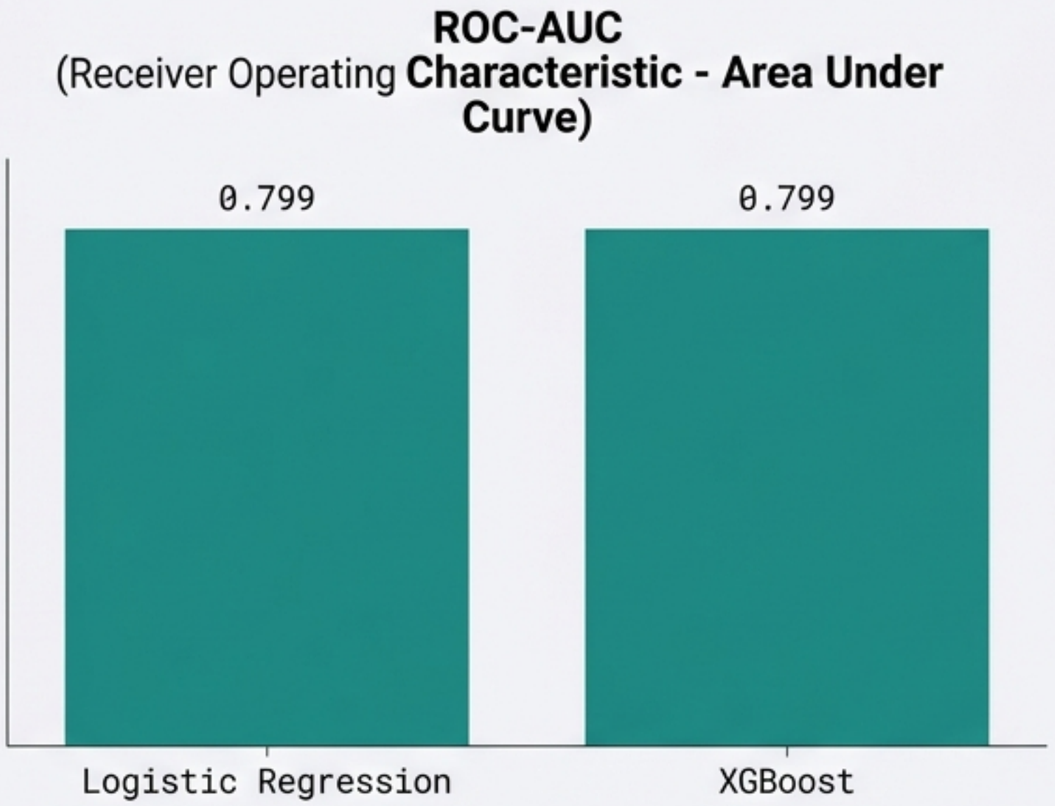
XGBoost (Gradient Boosting)

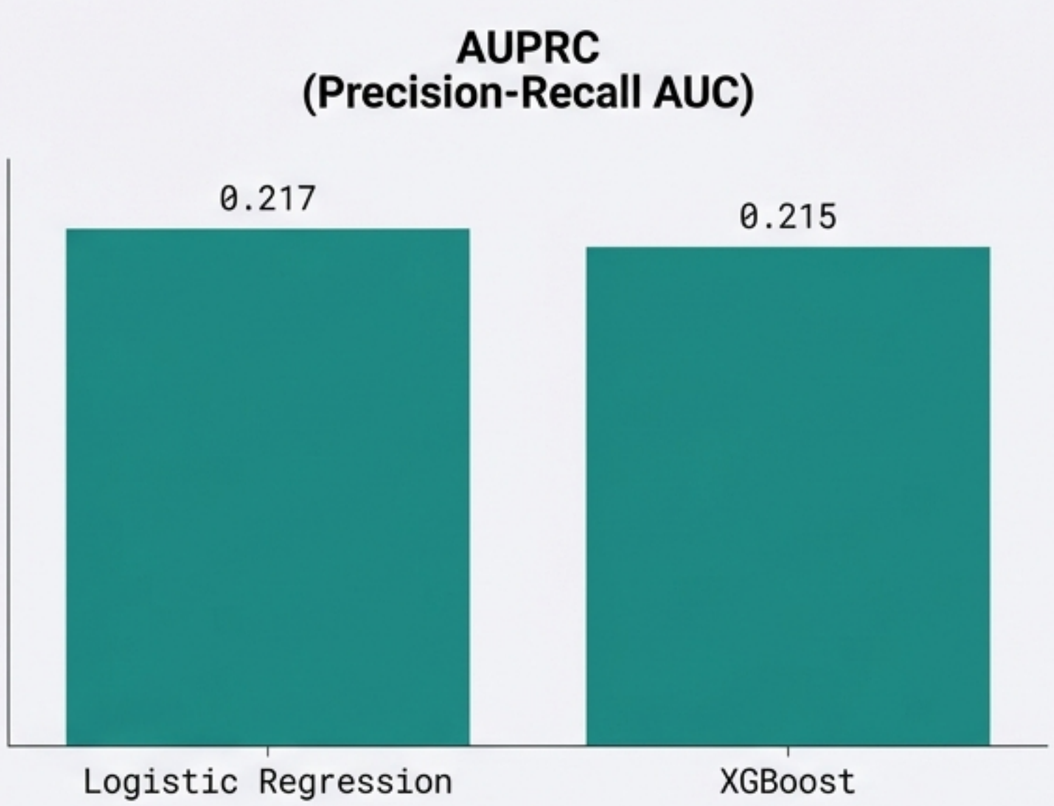
Role: The Challenger

Strength: Complex Non-Linear Interactions

Specs: 600 Estimators, Depth=5, Learning Rate=0.05, Subsample=0.9

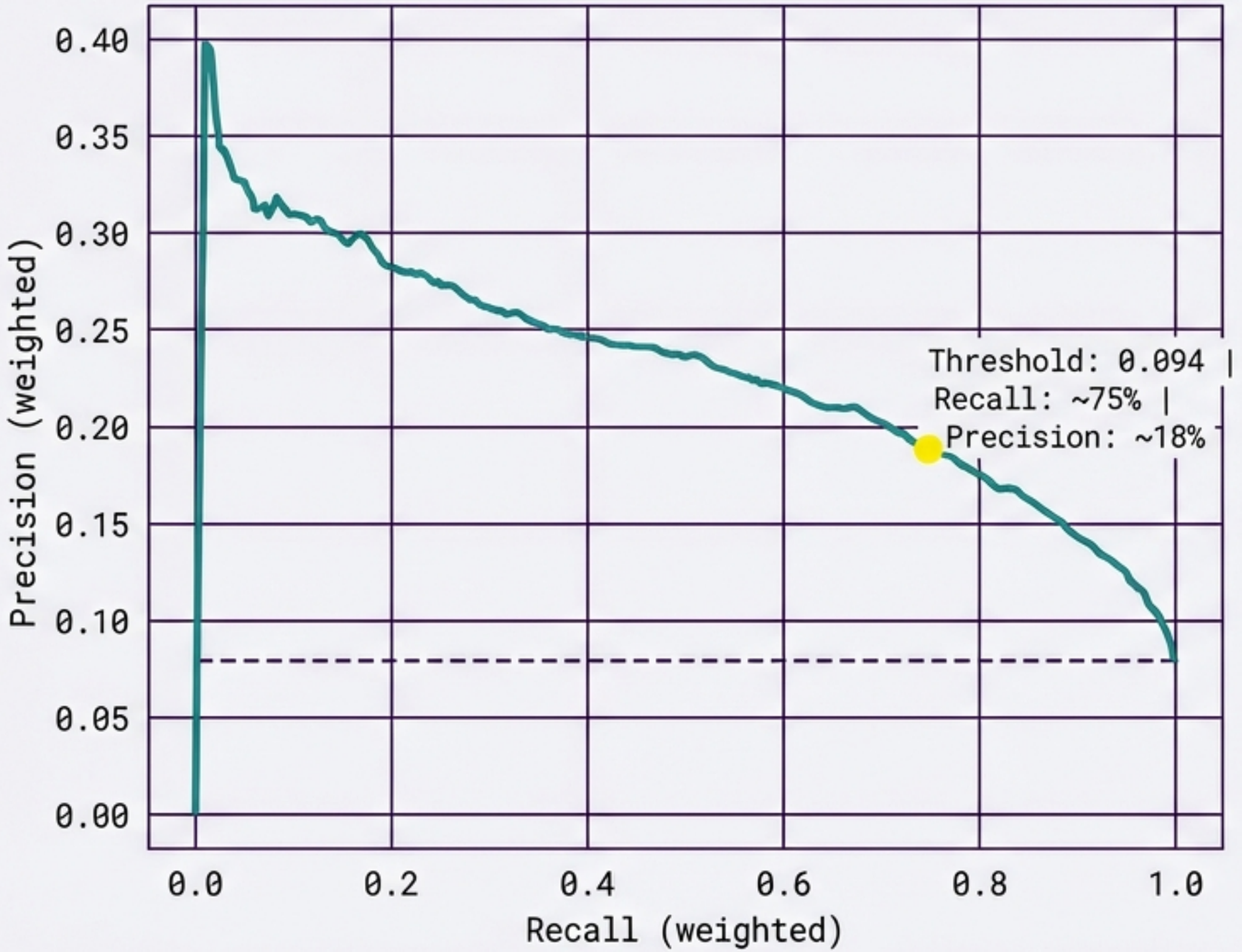
Performance: Complexity Did Not Beat Simplicity





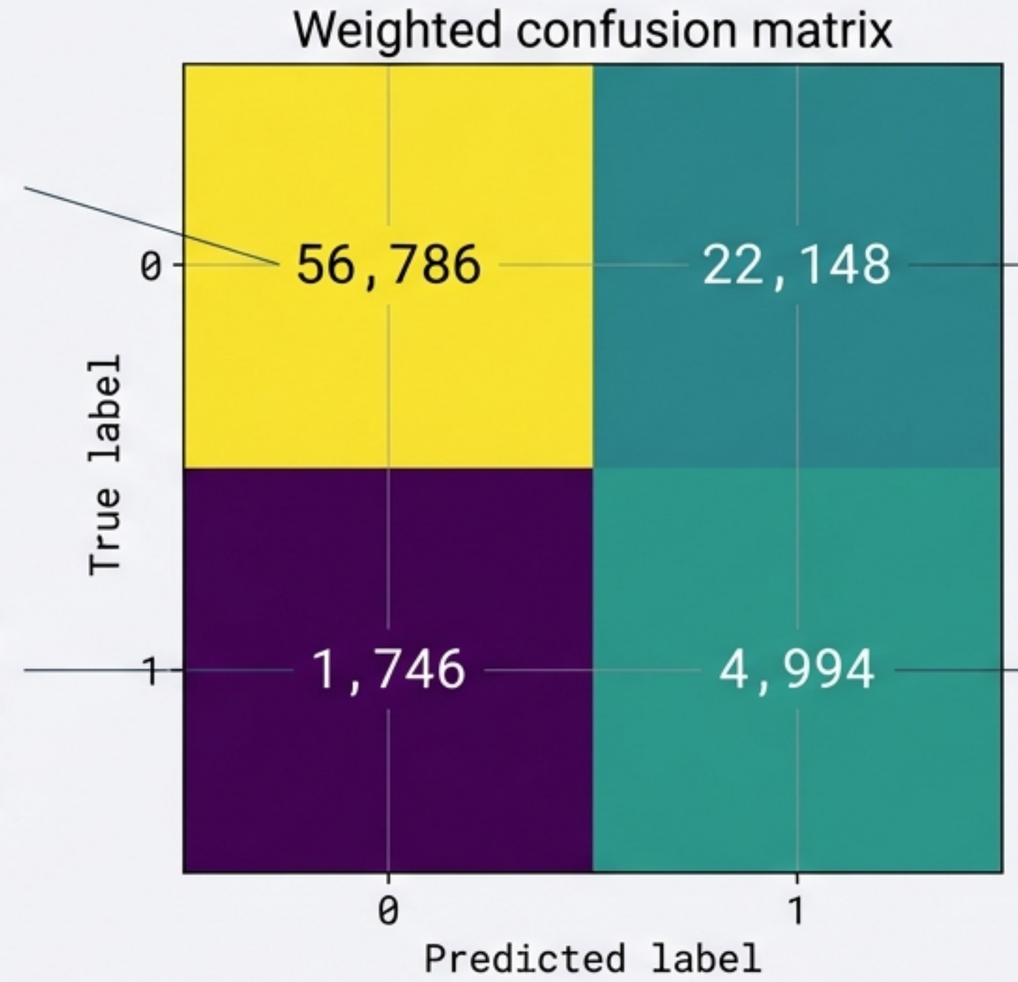
Lift: ~2.75x over baseline prevalence (7.9%). The tree-based model offered no significant advantage over linear regression.

The Precision-Recall Trade-off



The Reality of Rare Events:
To correctly identify 3 out of 4 cancer cases (75% Recall), we must accept a false positive rate where only ~1 in 5 flagged individuals (18% Precision) actually has the disease.

Visualizing Predictions: The Confusion Matrix



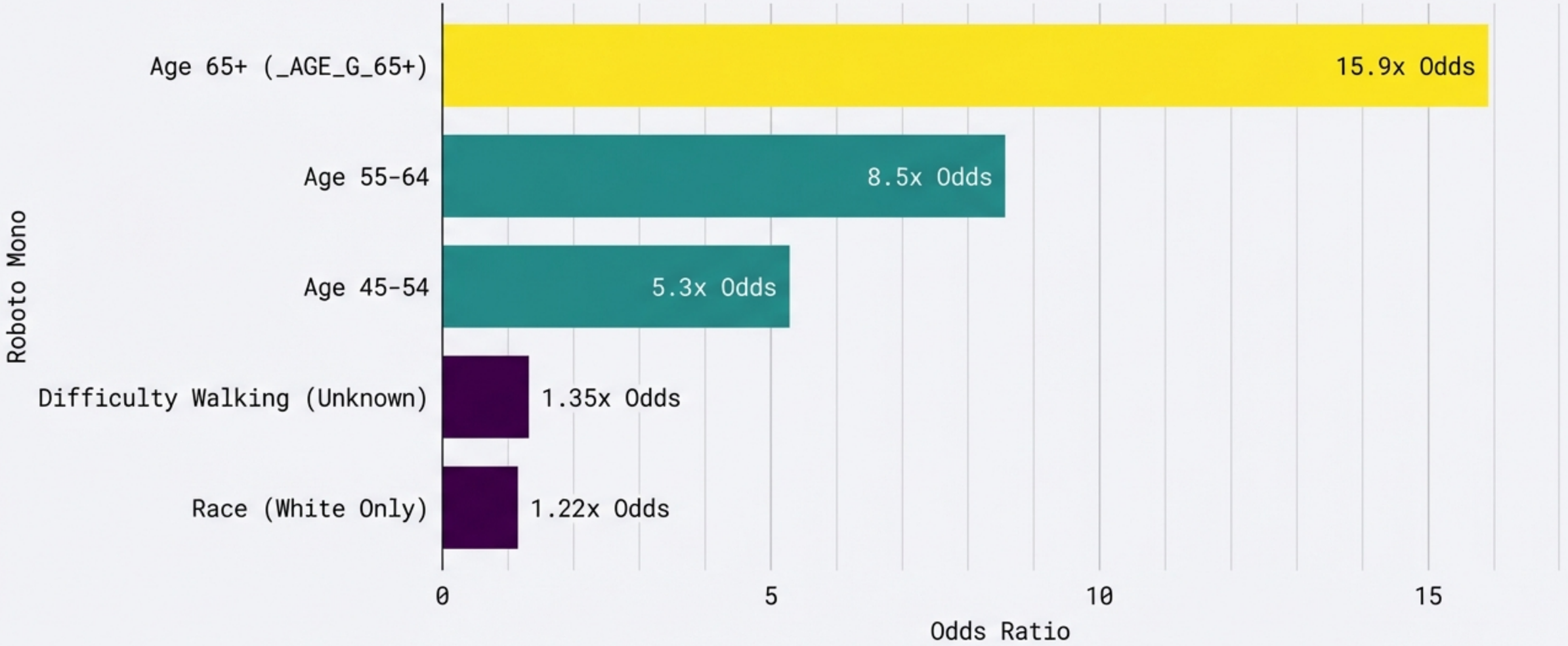
True Negatives

False Positives (The cost of high recall)

False Negatives (Missed Cases - Optimized to be low)

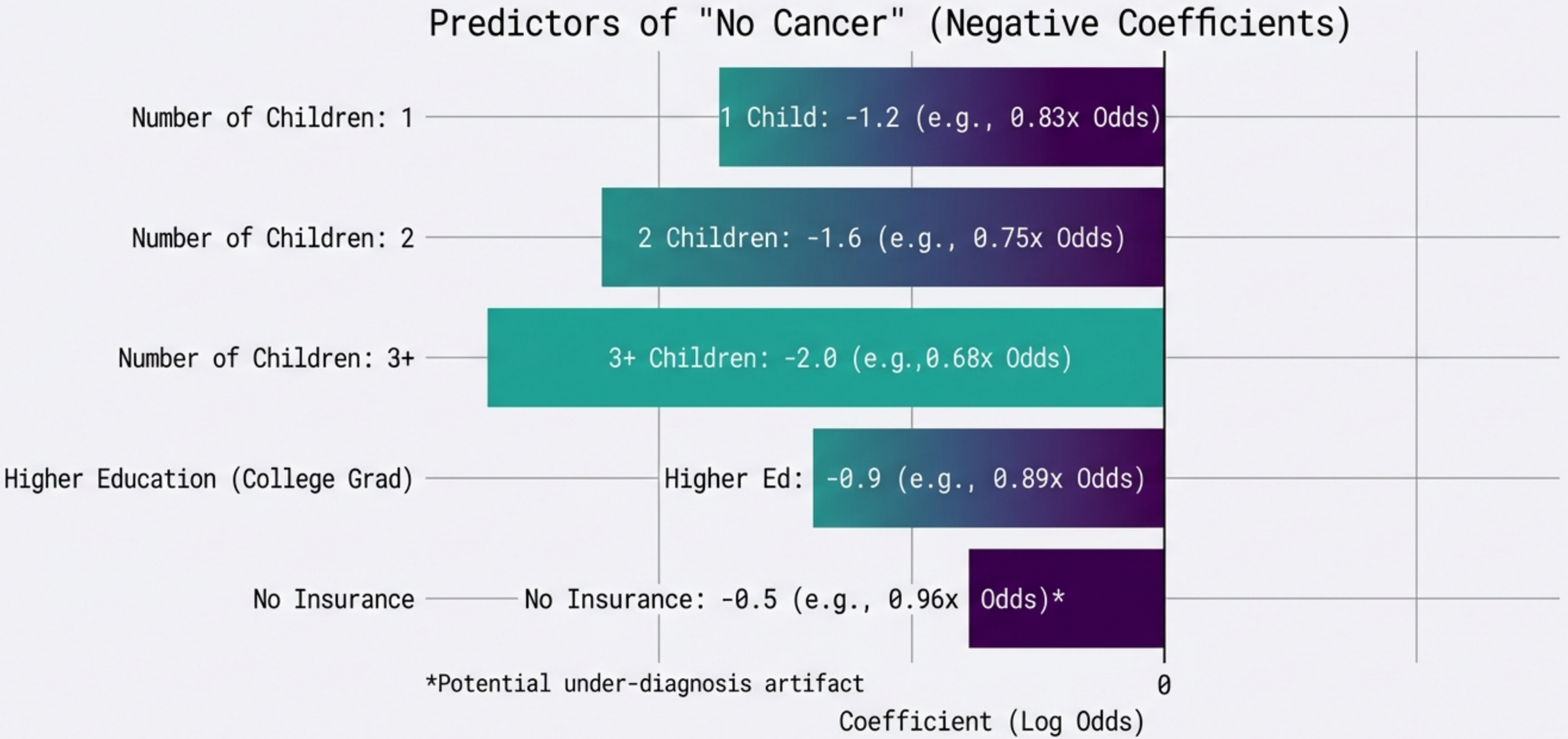
True Positives (Correctly Identified)

Feature Importance: Demographics are Destiny



Age is the dominant signal, overshadowing behavioral factors.

Protective Factors & Negative Correlations

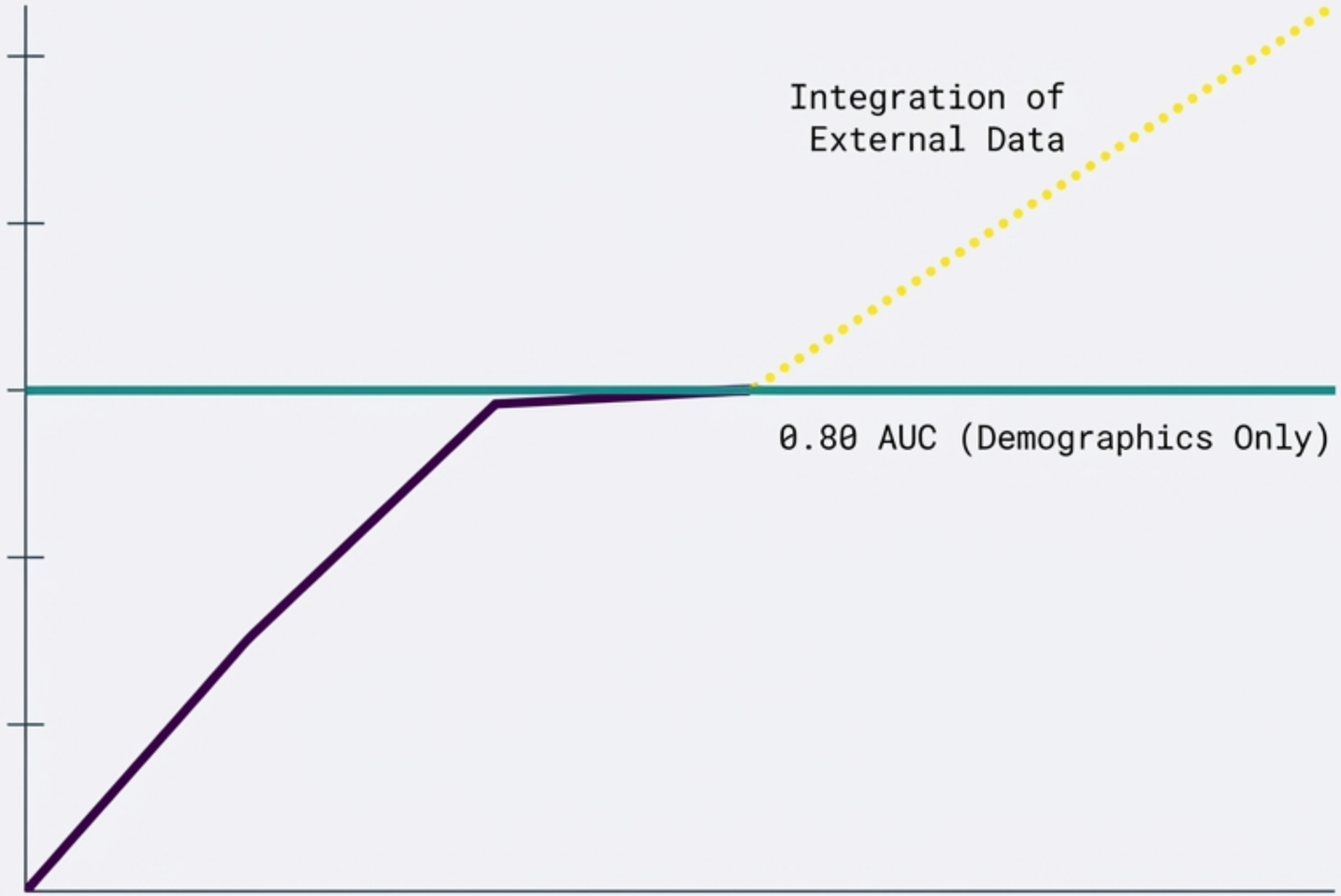


The Signal in the Silence

Feature: nom__DIFFWALK_Unknown
Coefficient: +0.30 (Odds Ratio 1.35x)

Respondents who did not answer the 'Difficulty Walking' question had a significantly higher likelihood of reporting cancer. Missing data is not just noise; it often correlates with health status (e.g., respondents too ill to complete the full survey).

Breaking the 0.80 Ceiling



Next Steps

1. Merge County-Level Air Pollution Data (PM2.5)
2. Analyze Environmental Toxins
3. Correlate with Geographic Risk Factors

Conclusion

2.75x

Lift Over Random Chance

Roboto Mono Regular Deep Slate

Age

Dominant Predictor

Roboto Mono Regular Deep Slate

0.80

AUC Ceiling

Roboto Mono Regular Deep Slate

We have maximized the signal from the survey responses. The next leap in predictive power requires integrating the world outside the survey.