Funding: *This study was funded by grant U54CA229974 from the National Cancer Institute of the National Institutes of Health and the FDA Center for Tobacco Products.*
Disclosure: *The authors declare that there is no conflict of interest.*

Machine Learning Application for Predicting Smoking Cessation Among US Adults

Mona Issabakhsh,[1*] Luz M Sánchez-Romero,[1] Thuy TT Le,[2] Alex Liber,[1]
Jiale Tan,[2] Yameng Li,[1] Rafael Meza,[3] David Mendez,[2] David Levy[1]

[1]Georgetown Lombardi Comprehensive Cancer Center; [2]University of Michigan School of Public Health;
[3]British Columbia Cancer Research Center

***Corresponding Author**
E-mail:mi416@georgetown.edu

Background

- Identifying significant elements of smoking cessation is critical for developing optimal cessation treatments and interventions.
- Machine learning (ML) is a powerful tool to find the contributing factors for smoking cessation and develop accurate predictive models, specifically in large datasets with a vast number of variables.
- **Objective:** This study aims to find determinants of smoking cessation, and to develop accurate predictive models for smoking cessation among US adults, applying ML algorithms.

Methods

- **Data:** longitudinal data from the PATH study (w1-2, w2-3), a US nationally representative survey is used.
- **Analyses:** predictive models with random forest, gradient boosting machine, generalized linear model, and extreme gradient boosting algorithms are developed.
- Because of the skewed class distribution in the data (7% quit rate), random sampling and ensemble-based techniques for variable selection and predictive model training are applied.

Results

Table 2: Evaluation results of the predictive models.

Sample	Model	Sensitivity	Specificity	Balanced Accuracy	ROC-AUC
No Sampling					
	GBM	0.0135	0.9972	0.5054	0.7696
	XGBoost	0.0676	0.9917	0.5296	0.7574
	GLM	0.0495	0.9929	0.5212	0.7392
	RF	0.0045	0.9992	0.5018	0.7584
Over Sampling					
	GBM	0.6712	0.7732	0.7222	0.7757
	XGBoost	0.3108	0.9094	0.6101	0.7021
	GLM	0.6531	0.7165	0.6848	0.7244
	RF	0.0360	0.9948	0.5154	0.7614
Under Sampling					
	GBM	0.7162	0.7114	0.7138	0.7652
	XGBoost	0.7432	0.6937	0.7185	0.7645
	GLM	0.6667	0.6409	0.6538	0.6991
	RF	0.7432	0.6917	0.7175	0.7652
Bagging					
	GBM	0.6824	0.7445	0.7135	0.7631
	XGBoost	0.7008	0.7019	0.7014	0.7557
	GLM	0.6607	0.6637	0.6622	0.7063
	RF	0.7297	0.7146	0.7221	0.7637

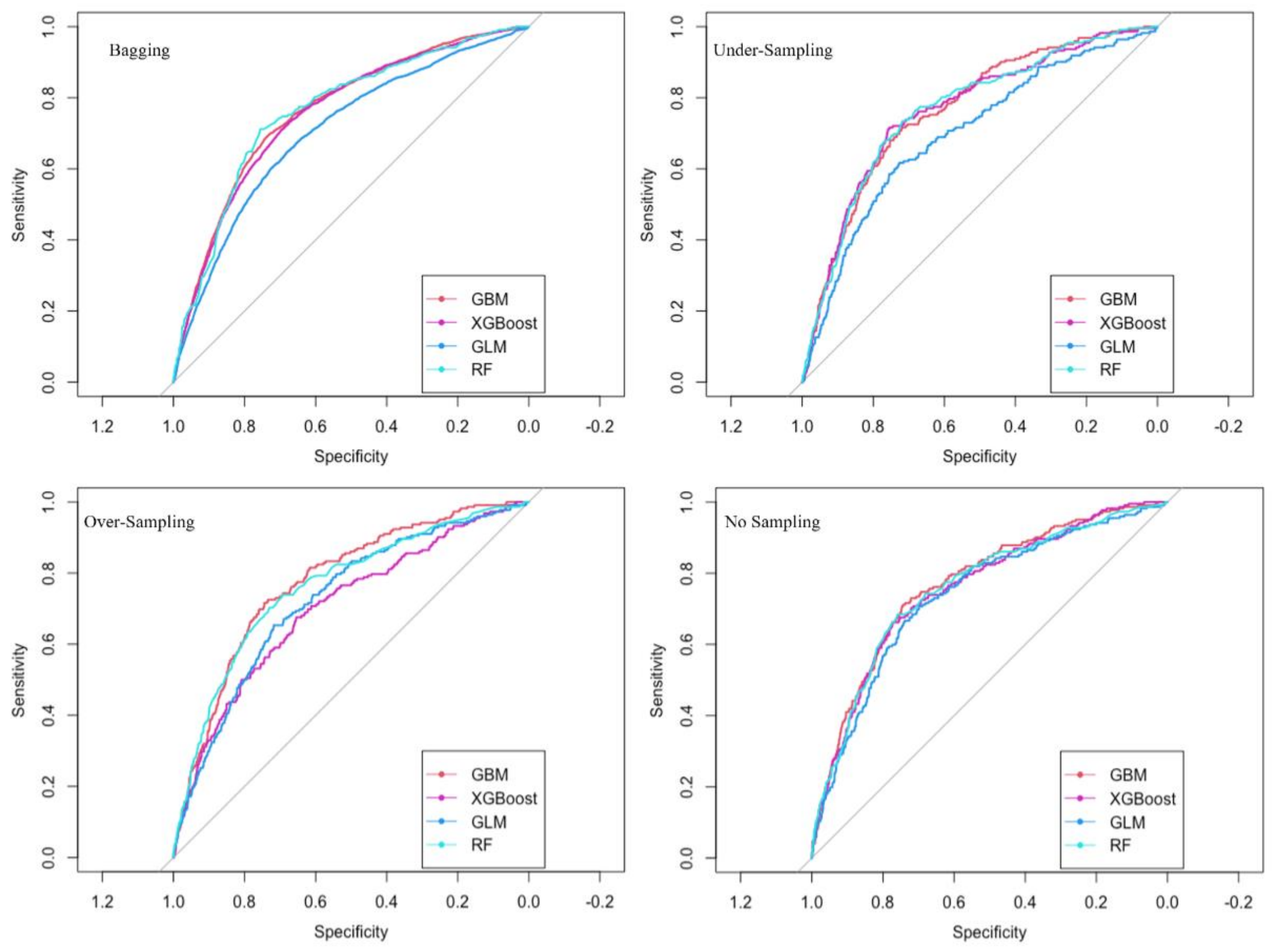

Fig. 4: ROC comparison of the predictive models.

Results

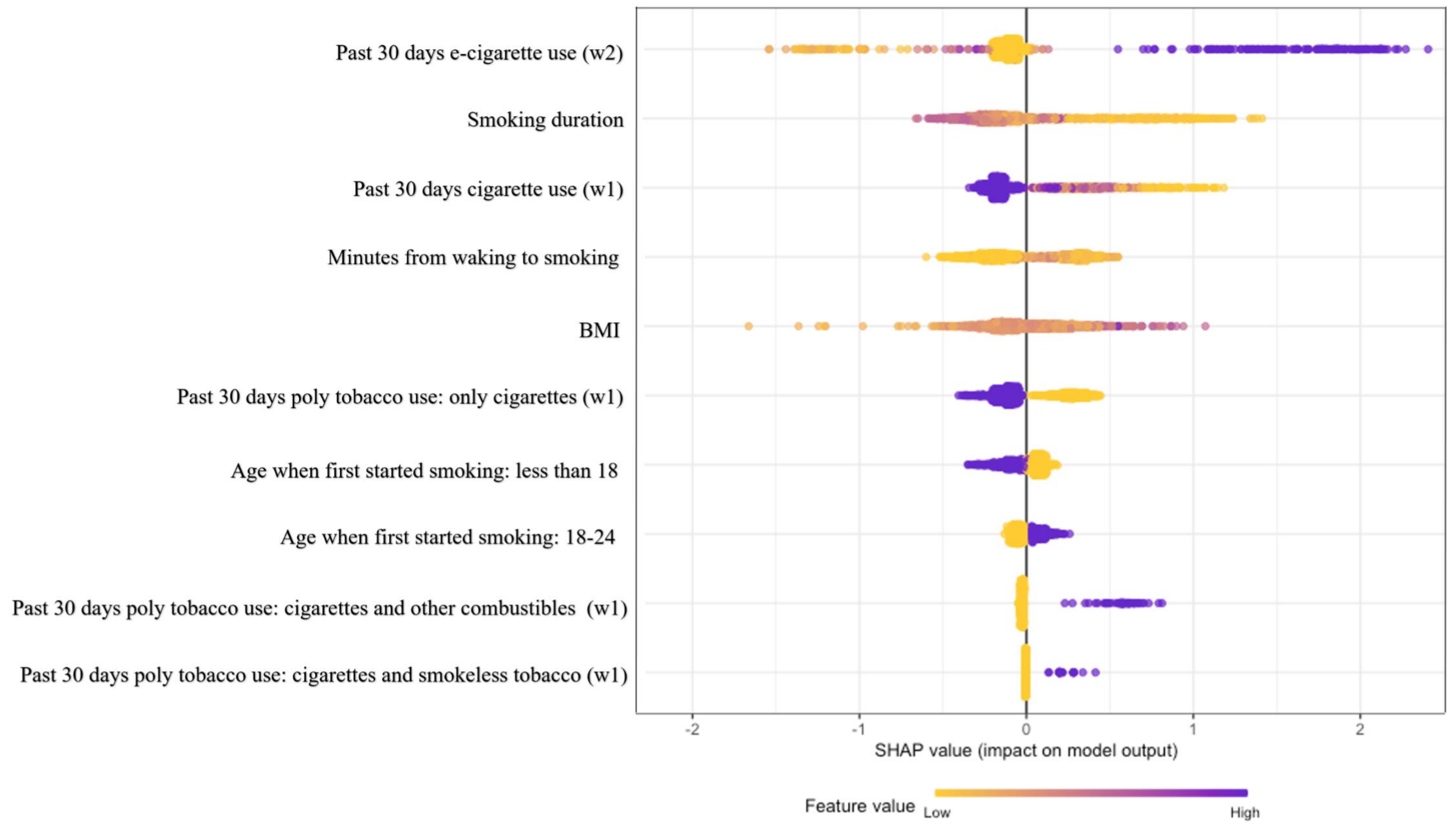

Fig. 3: TreeSHAP summary plot for the combination of the top five variables selected by RF and GBM.

Conclusions

Our analysis shows that the following characteristics among US adults increase their chances of smoking cessation:

- Higher past 30-days e-cigarette use at the time of quitting
- Fewer past 30-days cigarette use before quitting
- Ages 18 or older at smoking initiation
- Fewer years of smoking
- Higher BMI
- Poly tobacco past 30-days use before quitting