Classification on Secondhand Smoke Exposure Using Machine Learning Models

AUTHOR Kahang Ngau
COURSE CSCI8901 Research & Evaluation Method
DATE Dec 07, 2021

INTRODUCTION

Racial and socioeconomic disparities play a critical role in stark health disparities. Prior study suggests secondhand smoke (SHS) is one of the greatest factors contributing to diseases and harming human bodies. This project will investigate the abilities of different logistic models on classifying whether people are under SHS exposure or not, using demographic predictors. SHS is measured by serum cotinine (by blood examination) in this project.

OBJECTIVE

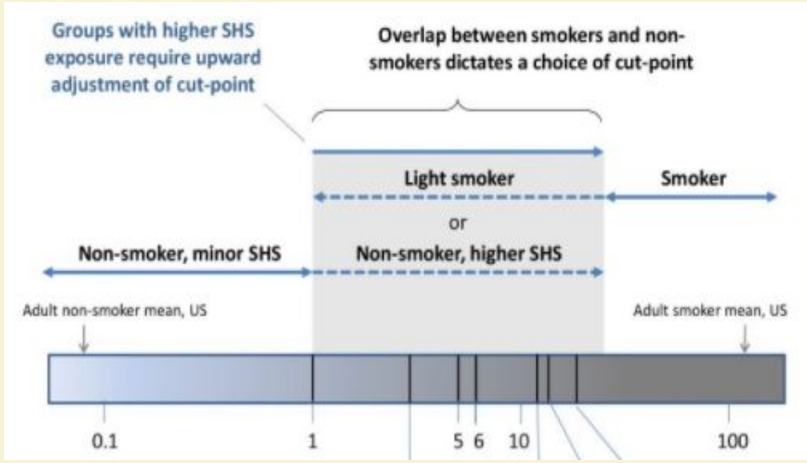
The purpose of this project is to use different statistical and machine learning classification models to predict whether non-smokers are heavily exposed by SHS or not. And to choose the best model with lowest **RMSE** and highest **AUROC score**.

METHODOLOGY

Data Collection - The data was collected from National Health and Nutrition Examination Survey (NHANES), covering the period of 1999-2016.

Data Cleaning - **Non-Smokers** were those who have not smoked **100** cigarettes in their lives.
The independent predictors were to keep only for those that have **70%** or more valid values.
NAs were dropped which leads to the final data set with **22,533** observations and **9** variables.

Statistical & ML Models
General Linear Model, Random Forest , and XGBoost



INDEPENDENT VARIABLES

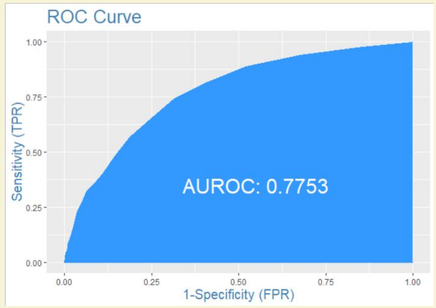
DMDCITZN	100.00%	Citizenship of the United States
DMDHHSIZ	100.00%	Number of people in the household
DMDHREDU	97.00%	Highest education level received
DMDMARTL	97.10%	Marital Status
HOD050	98.90%	Number of rooms in home
HOQ065	98.90%	Home owned, bought, rented, other
INDFMPIR	91.40%	Poverty income ratio (PIR)
RIAGENDR	100.00%	Gender
RIDAGEYR	100.00%	Age
RIDRETH1	100.00%	Race and ethnicity
LBXCOT	100.00%	Cotinine (ng/mL) - Dependent Variable

ANALYSIS

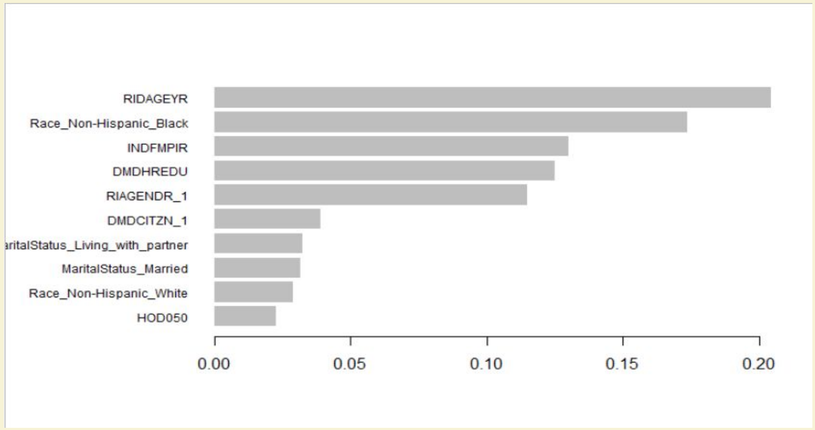
Built models in R, with hyperparameter tuning in XGBoost. GLM and ML logistic models prediction performance are following:

	RMSE	AUROC	SPEED
GLM	0.2685	0.7743	Fast
RF	0.2701	0. 7473	Slow
XGBoost	0.2680	0.7753	Fast

ROC Curve
It is drawn by the **False positive** rate on X-axis and the **True Positive** rate on Y-axis.



FEATURE IMPORTANCE



CONCLUSION

The final selected model is the XGBoost machine learning model, as it has the lowest RMSE and highest AUROC. Future work will focus on discovering other methods to improve the model accuracy and exploring more different models (like SVM, AdaBoost, and K-neighbors Classifier) to compare with results.