Using Bio Signals to Predict Smoker Status*

Uri Guerra †
University of Oregon, Sociology
mguerra2@uoregon.edu

Abstract. An Abstract

Keywords. health; data science; public health; tobbacco use

Introduction

Using various health indicator information classification models are utilized to predict smoker status. As it stands more than 16 million american are currently living with a disease that can be linked to smoking "Health Effects of Smoking and Tobacco Use" (2022). The use of a model like this is useful in public health especially when identifying links between certain bio signals and smoking status. It could lead to a better understanding of individual effects smoking has on the body due to identifying which health factors are most closely associated with smoking status.

Description of Data

The data for this project was collected from an online competition hosted on the Kaggle website called, "Binary Prediction of Smoker Status using Bio-Signals" *Binary Prediction of Smoker Status Using Bio-Signals* (n.d.). The training data was generated from a deep learning model that was generated from the "Smoker Status Prediction using Bio-Signals" data set *Smoker Status Prediction Using Bio-Signals* (n.d.). The data set consist of 22 different predictors that range from age and height to health indicators like cholesterol level, hemoglobin and systolic pressure.

Model Description and Fit

Using the training data set three different models were created which include a generalized linear mode, classification model with ridge penalty, and a classification model with lass penalty. Despite the use of different models there did not appear to be siginifcant difference in model performance. The way the models were evaluated was through the use of the Log Loss and AROC to evaluate the models, with the results that can be seen below.

---

*Thanks to everyone for checking this out.
†Corresponding author.