Using Bio Signals to Predict Smoker Status*

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

Abstract. This project focuses on the use of classification models to predict the smoking status of individuals based on diverse health indicators. With over 16 million Americans currently afflicted by diseases associated with smoking, there is an increased interest in the ability to use data science and machine learning models in public health. The application of this model in public health is particularly valuable for identifying correlations between specific biosignals and smoking status. Such insights contribute to a deeper understanding of the individualized impacts of smoking on the body by pinpointing the health factors most closely linked to smoking status.

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

Github Repository

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. The predictors are also numeric so while preparing the data for processing variables with zero variance were removed and were standardized, although the data overall was well prepared and easy to work with.

---

*Thanks to everyone for checking this out.

†Corresponding author.

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 significant 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.

Table 1. Model Performance Evaluation

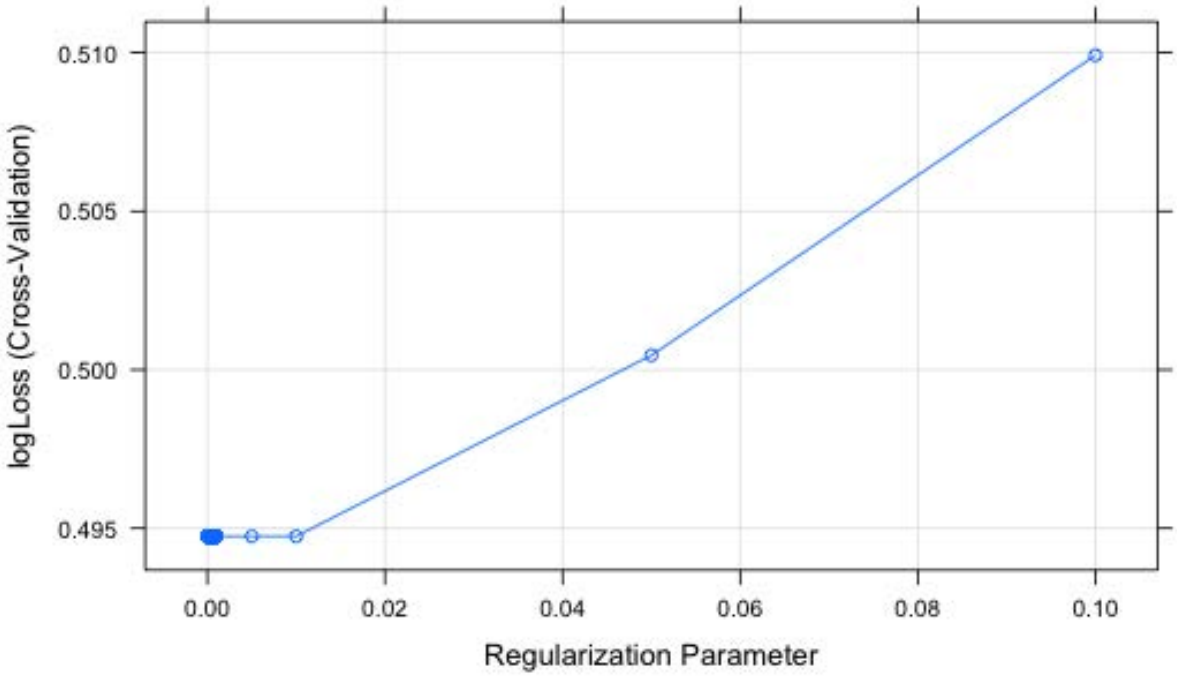
Model	Log Loss	AROC
GLM	0.4912959	0.8421145
Classification Model with Ridge Penalty	0.4945809	0.8396281
Classification Model with Lasso Penalty	0.4913238	0.8420534

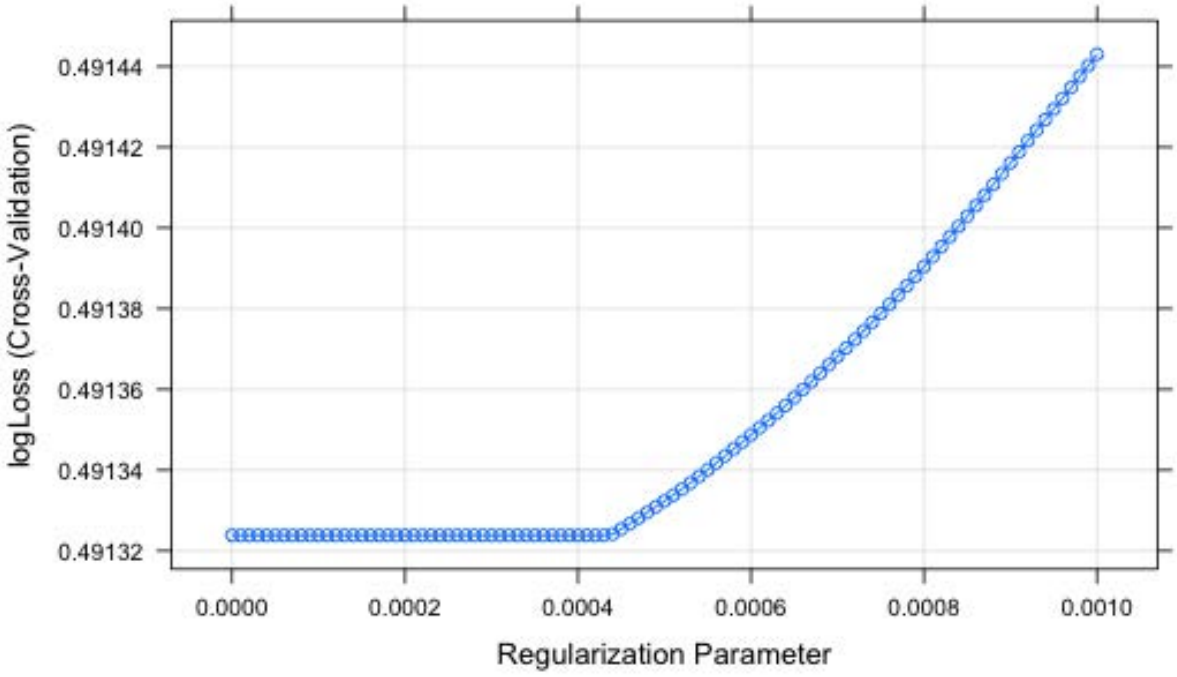
Log Loss was chosen since it highly penalizes incorrect predictions, as the models were created to classify whether predictors indicated a non-Smoker vs. a Smoker. AROC however, was chosen because 1) the Kaggle competition the data set came from utilized AROC to evaluate performance 2) in addition to its particular success at evaluation binary classification models since it measures the probability distribution of false and true positives.

Since there was no significant difference in model performance the ridge penalty model was chosen when doing further evaluation of performance and when determining which variables were most influential in outcome. After creating the Classification Models with Ridge and Lasso Penalties, respectively they were plotted to asses the results of the models, in addition to seeing model fit.

Conclusion

The three most important variables in predicting smoking status was hemoglobin, triglyceride, and y-GTP. Hemoglobin is a protein contained within red blood cells and is associated with proper oxygenation of tissues. Triglycerides are a type of fat (or lipid) found within our blood and is often associated with obesity, stroke, heart disease and other serious illnesses *Triglycerides* (n.d.). Y-GTP however, is associated with liver disease at higher levels especially when alcohol or other drugs are involved. Considering the affects smoking has on your health the connection between these predictors and smoking status are not surprising and fall within what I expected. The practical findings of this project are minimal and more than likely would require more data and a more focused analysis of health indicators and their relationship with smoking status. However, it is clear that being able to predict cerain health conditions or behaviors is useful in public health. I do not do a lot of work in public health or medical sociology but this project has made more curious about other projects like this especially when considering the ethics of implementing something like this into the healthcare system.





References

Binary Prediction of Smoker Status using Bio-Signals. (n.d.). Retrieved December 8, 2023, from https://kaggle.com/competitions/playground-series-s3e24

Health Effects of Smoking and Tobacco Use. (2022). In *Centers for Disease Control and Prevention*. https://www.cdc.gov/tobacco/basic_information/health_effects/index.htm

Smoker Status Prediction using Bio-Signals. (n.d.). Retrieved December 8, 2023, from https://www.kaggle.com/datasets/gauravduttakiit/smoker-status-prediction-using-biosignals

Triglycerides: Why do they matter? - Mayo Clinic. (n.d.). Retrieved December 8, 2023, from https://www.mayoclinic.org/diseases-conditions/high-blood-cholesterol/in-depth/triglycerides/art-20048186