Predicting a potential child smoker using SAS® Enterprise Miner 12.1

Jin Ho Jung
Gaurav Pathak
Dr. Goutam Chakraborty
Over the years, there has been a growing concern about consumption of tobacco amongst youth. But no concrete studies have been done to find out what exactly leads children to start consuming tobacco. This study is an attempt to figure out the potential reasons for the same. Through our analysis we have also tried to build a model to predict whether a child would smoke next year or not.

This study was based on the National Youth Tobacco Survey. It provides national data key to design, implement, and evaluate comprehensive tobacco prevention and control programs. Base SAS® 9.3 and SAS® EM 12.1 are used to predict the best model and find significant factors that affect tobacco usage among young children.

Such findings suggest how public policy makers can effectively campaign for tobacco-free youth.

**Method**

- **Data Preparation**
  - Data preparation is of primary importance when it comes to solving a business problem. The survey-based data initially consisted of 18,867 observations and 197 variables.
  - The target variable is a binary variable which indicates whether a child would smoke next year or not.
  - From a pool of 192 potential input variables, 13 important variables were selected (Fig. 1) using variable selection methods, partial least squares and decision tree models.

- **Predictive Modeling**
  - Once the data was prepared and ready for analysis, it was partitioned into Training, Validation and Testing with a ratio of 50:30:20.
  - As a part of predictive modeling, Decision Trees, Stepwise Logistic Regression Model and Stepwise Logistic Polynomial Regression Model were built based on the previously selected variables. Using Model Comparison Node in SAS® Enterprise Miner 12.1, competing models were diagnosed and compared with each other.
  - Stepwise Logistic Regression Model outperformed other models (See Fig. 3). With misclassification rate as our selection criteria, stepwise regression model produced Validation misclassification of 0.028497.
  - Factors such as company of friends, cigarette brand ads, accessibility to the tobacco products, and passive smoking turned out to be most important predictors in determining a child smoker.

**Introduction**

The purpose of this study is to investigate whether a youth would smoke next year and which factors influence youth to start smoking.

Tobacco companies’ advertising and the number of smoker friends a youth has mostly influence a youth’s willingness to smoke in the next year.

This study gives directions to public policymakers and school teachers to develop and implement tobacco prevention programs. This will help to reduce youths’ tobacco use and ultimately prevent illness and death caused by tobacco use.

We believe that our findings provide insightful information to both public policymakers and school teachers in order to effectively control and campaign youths’ tobacco use.

**References**

- U.S. Department of Health & Human Services (2013), Youth & Tobacco
- Wayne W. LaMorte, MD, PhD, MPH, Boston University School of Public Health (2013).