Aishwarya Saibewar
DATA 5322
April 13, 2023

Best Factors in Predicting Alcohol Abuse in Teenagers Using Tree-based Models

Abstract

In today's world, alcohol abuse is a serious concern among teenagers and can have adverse effects on their lives if consumed excessively. This could mentally disturb not only the consumer but also the people surrounding them. Heavy consumption of alcohol could lead to impulsive reactions, violent behavior, financial instability, poor decision-making, severe medical disorders, etc. Thus, developing an algorithm that could accurately determine the factors that influence teen alcohol abuse becomes imperative. This research uses data from National Survey on Drug Use and Health (NSDUH) to perform analysis and examine the accuracy of Tree-based models in making predictions. Three models were developed to understand the influential factors. The first model using a decision tree predicts whether a teen has ever consumed alcohol with an error rate of 14.2%. The second model using the Random Forest algorithm classifies the alcohol users into the following categories: seldom, sometimes, frequent, and never used with an error rate of 13.5%. The third model using Boosting predicts the frequency of alcohol usage by a teenager in a year with a mean square error of 120. The most influential factors observed are whether the teen has consumed tobacco or marijuana, followed by education grade and size of the metro.

Introduction

In the United States, teen alcohol abuse is a major public health concern. This can start at a young age when it is very easy to develop an addiction. While the amount consumed in the beginning is limited, this can very soon turn into binge drinking eventually leading to alcoholism. Therefore, it is important to determine the factors that influence teen alcohol consumption.

This research used National Survey on Drug Use and Health (NSDUH) data to train algorithms to make predictions. The SAMHSA (Substance Abuse and Mental Health Services Administration) conducts NSDUH surveys every year across the country and collects data for use of tobacco, alcohol, and illicit drugs. The data is publicly accessible allowing researchers to ingest this data and determine substance abuse patterns and draw related insights. Using this data, we will determine the most influential factors of alcohol abuse in teens using Tree based algorithms.

Theoretical Background

Tree-based models are the most commonly used supervised learning methods as they are easy to interpret and highly stable. They could be used to solve both classification and regression problems. The algorithm includes segmenting the predictor space into several sections. To generate a forecast for a given observation, the mean (regression) or mode (classification) response values are utilized for the training observations in the corresponding region(s). Such

approaches are referred to as decision tree methods, as the set of division rules used to segment the predictor space can be summarized in a tree. The tree can be explained by two entities: internal nodes and terminal nodes (or leaves). The data is split at the internal nodes and the decisions or final outcomes are shown by the leaves at the bottom which indicate the mean value of the region. The downside of decision trees is that they are prone to overfitting the training data, which would ultimately result in inaccurate forecasts of test data. However, the overfitting concern of the training data can be prevented by pruning. It creates a subtree to balance variation and bias and reduces the size of the decision tree. A cross-validation technique is used to improve the model accuracy to select the optimal size of the tree with the lowest deviance.

While pruning can help a decision tree model make better predictions, a single decision tree model won't be able to make accurate predictions on its own. Therefore, Ensemble methods are used to boost the model's predictive power by creating multiple trees and integrating the predictions. Ensemble methods incorporate various base models to create a single "best-fit predictive model." An ensemble method for reducing the variance and improving the performance of decision trees is bootstrap aggregation, also known as bagging. In bagging, all the data is bootstrapped into various samples with replacement i.e., randomly selecting the data with replacement, and a full set of predictors is considered at each split (m=p). Thus, a decision tree is built for each of these samples by considering all of the predictors. Each sample's results are averaged to make a final prediction. The predictions from bagged trees are strongly correlated. Therefore, the Random Forest algorithm (RF) could be used to solve the problem as it considers a subset of the predictors and decorrelates the tree. Usually, in the case of a Random Forest, the square root of the total number of predictors (p) is considered at each split i.e., m=sqrt(p).

Out-Of-Bag error is another measure for estimating the prediction error in ML models that use bootstrapping methods. For each bagged tree, the unused data (usually $1/3^{rd}$ of the training data) is used as validation data. The mean of the predictions from the unused data is used to calculate the error. The overall mean square error (for regression) for all the observations is the OOB error.

Boosting is a sequential ensemble strategy that uses data from several weaker models with low accuracy, that have been previously developed and assigns weights to their output. The errors from the previous trees are used as input for the subsequent trees. This process is repeated until the final model predicts the outcome accurately. The commonly used boosting techniques are the gradient boosting model (GBM) and AdaBoost. This paper uses the Gradient Boosting model as the boosting technique. GBM constantly learns from its mistakes and improves its predictive ability. The tuning parameters required for the GBM model are the number of trees (B), interaction depth (d), the learning rate or shrinkage parameter (λ), and the distribution. The interaction depth represents the number of splits that must be performed on a tree. As the new trees are built to correct the residual error from the previous trees, the model might overfit the training data. As a result, when new trees are added to the model, a weight factor is added to the corrections to slow down learning. The weighting factor is known as the learning rate. The distribution is specified as “Gaussian” for a regression model and as “Bernoulli” for a classification model.

Methodology

Data Preparation

There are 2890 variables in the NSDUH dataset, but only 12 potential variables were taken into account for the analysis of alcohol abuse in teenagers. To prepare the data for use in the model, the missing values of the categorical variables were imputed using the mode of the non-missing values of that variable. Also, a few of the categorical variables were converted to factors. When the data were being prepared for developing a regression model, the response variable had the value "never used alcohol" in many records. For the purpose of creating a reliable model, these records were given a value of zero.

Models

First, the dataset was split into training and testing where 70% of the data was used as a training dataset to train the models and the remainder was reserved for testing purposes. The models were trained using the training dataset with imputed missing values. For binary classification, a decision tree was developed to predict whether a teen has ever consumed alcohol. The decision tree was interpreted to understand the significant variables.

Two models were developed for multi-class classification to predict the teen's drinking patterns, i.e. if the teen is a seldom, sometimes, never, or frequent alcohol user. First, the bagging model was developed by considering 100 trees and a full set of 'p' predictors at each split. Second, the random forest model was developed where the number of trees considered was 100 and the number of predictors was equivalent to the square root of 'p'. The predictions were made on the test data and the model that minimized the error rate was considered as the best model.

The Gradient boosting model was then developed to predict how frequently teenagers consume alcohol in a year. This model was tuned to various interaction depths by keeping the learning rate constant at 0.01. The distribution was Gaussian because it is a regression model. The best model was deemed to have an interaction depth at which the test data's mean square error was minimized. Table 1 provides descriptions of various variables used in the analysis.

Variable	Description
CONSUMED_MARIJUANA	Whether the teen has ever consumed marijuana (Yes/No)
EDUCATION_STATUS	What grade the teen is in now/will be in
CONSUMED_TOBACCO	Whether the teen has ever consumed tobacco (Yes/No)
METRO_STATUS	Metro size status (large metro, small metro, non-metro)
TOTAL_INCOME	Total family income

YOUTH_FIGHTS	Teen had serious fights at school/work (one or more times/Never)
GOVT_ASSISTANCE	Got government assistance (Yes / No)
GENDER	Gender (Male / Female)
HEALTH_STATUS	Overall health (Excellent / Very Good / Good / Fair or Poor)
ALCOHOL_FREQUENCY	Alcohol Frequency past year (1-365)
ALCOHOL_DAYS	Number of days of alcohol in past month (Seldom, Sometimes, Frequent or Never used)

Table 1: Variables used for analysis and their descriptions

Computational Results

Binary Classification

A binary classification model was developed using the decision tree approach of tree-based algorithms to predict whether the teenager has or has not consumed alcohol. Figure 1 below is the resulting classification tree, and the tree interpretation is as follows.
The most important variable for predicting whether a teen has ever consumed alcohol or not is CONSUMED_MARIJUANA since the first split criterion starts at the top of the tree. It is subdivided into two branches: if the teen has not consumed marijuana (Left-hand branch) and if the teen has consumed marijuana (Right-hand branch).

I. Students who have not smoked tobacco and are under the ninth grade are unlikely to consume alcohol.
II. Similarly, Students who have smoked tobacco and are under the ninth grade are more likely to consume alcohol.
III. Teens who have consumed marijuana are likely to consume alcohol.

Therefore, a teen who consumes marijuana or tobacco is more likely to also consume alcohol.
The decision tree model has been predicted with an error rate of 14.29%

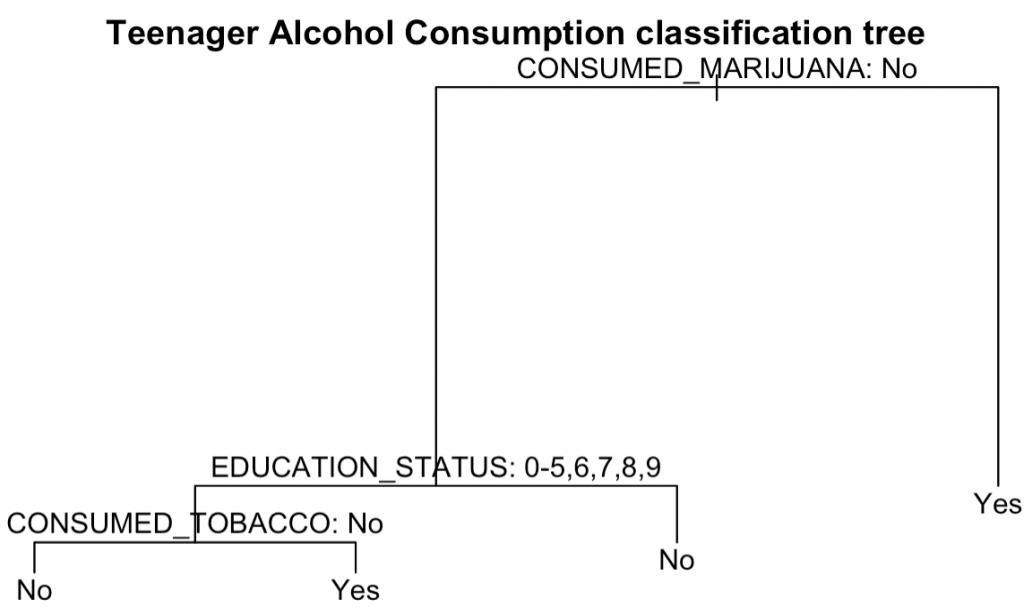

Figure 1: Teenager Alcohol Consumption Classification Tree

Multi-class Classification

A multi-class classification model was developed to predict teen drinking patterns in a year. A bagging model with a full set of predictors (m=p) and a random forest with a square root of predictors (m = sqrt(p)) was developed. As shown in Figure 2, the increasing number of trees is plotted against the Out of Bag error for each model. It can be observed that with the increase in the number of trees, the OOB error decreases. Also, the random forest model with a square root of predictors performs better than the bagging model as the OOB error rate tends to decrease for this model. On comparing the accuracy of the two models on the test data, it can be seen that the error rate of the random forest model is around 13.5%, whereas the bagging model has an error rate of 15.6%. Clearly, the random forest model is more accurate in classifying teen drinking patterns into different classes.

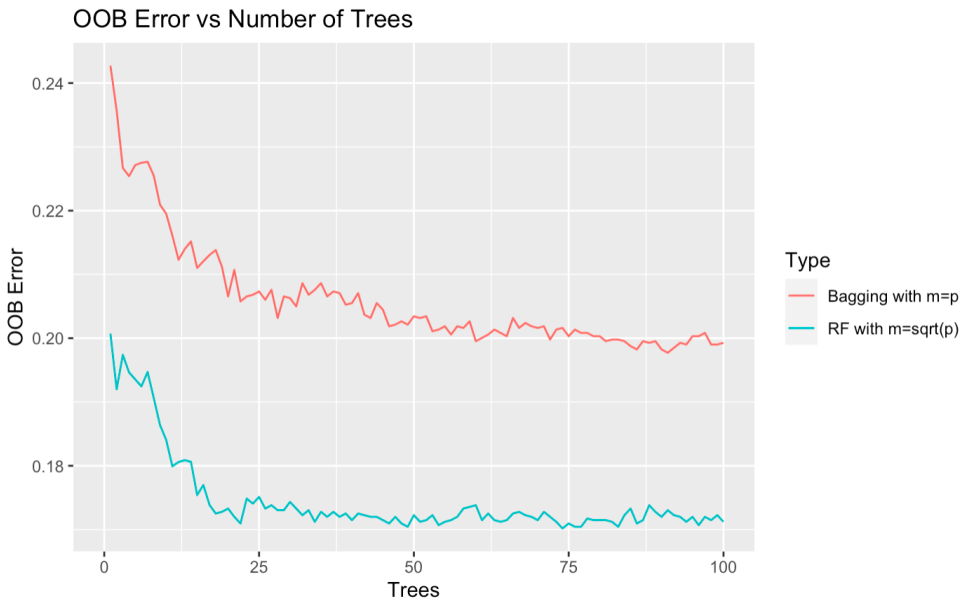

Figure 2. OOB Error vs Number of Trees for two models

Figure 3. shows the importance of each variable in classifying teen drinking patterns into various categories (seldom, sometimes, frequent, or never used). Here, the mean decrease accuracy is plotted against the variables used in the random forest model. This plot describes the significance of each variable by indicating the decrease in accuracy of the model when a particular variable is excluded. It can be observed that CONSUMED_MARIJUANA is the most important variable, followed by CONSUMED_TOBACCO, EDUCATION_STATUS, and METRO_STATUS.

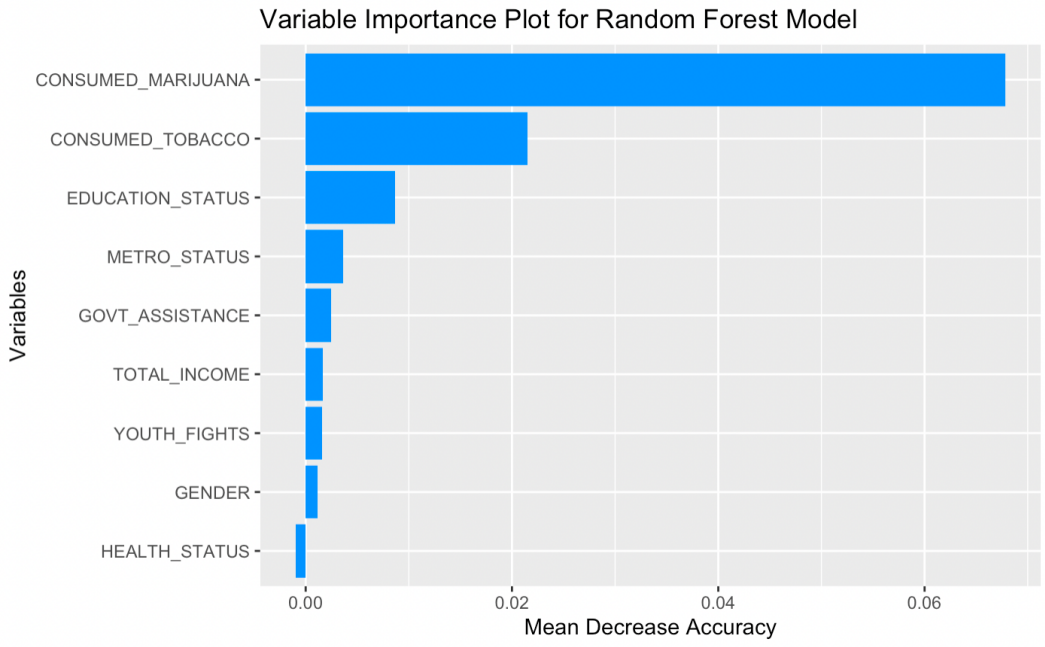

Figure 3. Variable Importance Plot for Multi-class classification model

As the Random Forest model performs better than all the other models the confusion matrix of this model is shown in Figure 4. The majority of the observations in the dataset could be attributed to users who have never consumed alcohol. The "Seldom" category of users is the next most common in the dataset, and the model correctly predicts 92 out of 256 seldom users. However, the model was unable to accurately predict the categories "Frequent" and "Sometimes" alcohol users. This could be because the dataset only contains a small number of records for these categories.

	Predicted			
Actual	Frequent	Never used	Seldom	Sometimes
Frequent	0	4	5	0
Never used	0	1336	35	0
Seldom	2	162	92	0
Sometimes	0	6	9	0

Figure 4. Confusion Matrix for Random Forest Algorithm

Regression

A gradient-boosting model was developed to predict the frequency of alcohol usage by a teen in a year. The distribution for this model is Gaussian as this is a regression model. The training error is depicted in Figure 5 against an increasing number of trees at various interaction depths and a constant learning rate of 0.01. The training errors when interaction depth is 1, 2, and 3 are 423, 409, and 384 respectively. It can be observed that the training error is the least when the gradient boosting model has an interaction depth of 3. Similarly, the mean square error on the test data at interaction depths 1, 2, and 3 are 120.30, 120.39, and 126.89 respectively. Although on the training data, the least mean square error is at depth 3, on the actual test data, the lowest mean square error is observed at depth 1. Thus, the best model for the test data has an interaction depth of 1 and a shrinkage parameter of 0.01.

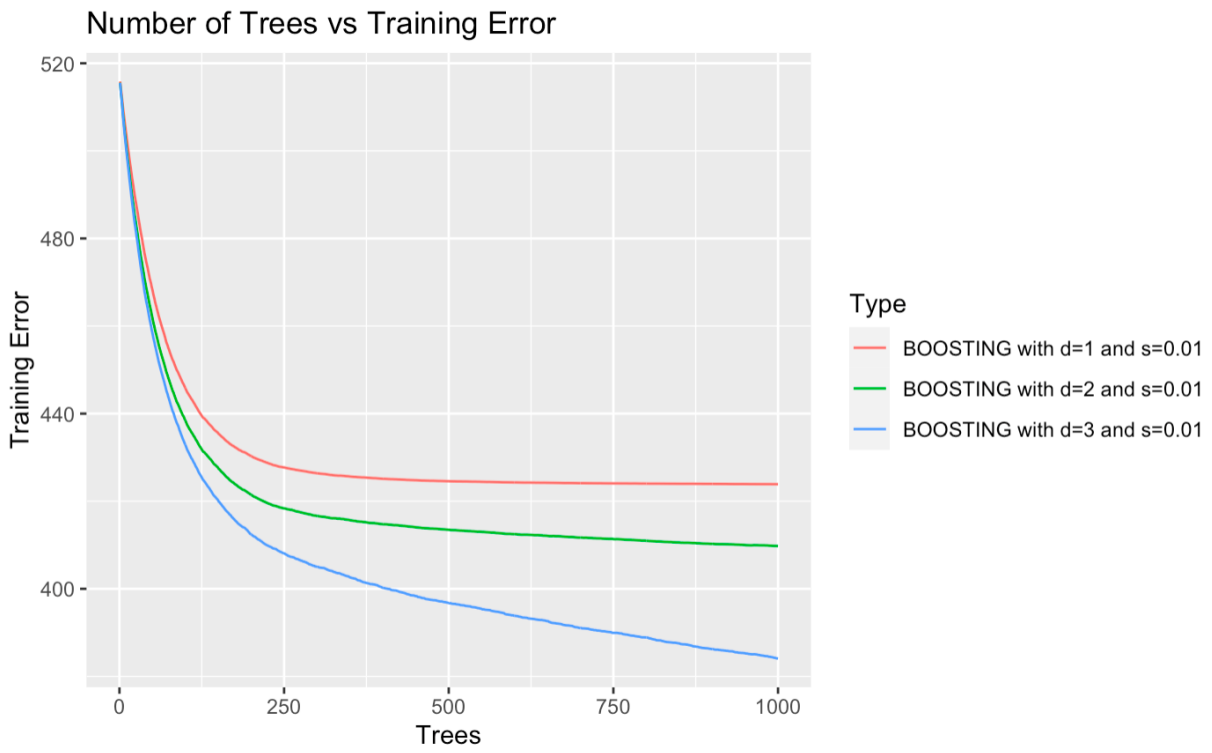

Figure 5. Number of trees vs Training error for different interaction depths

Discussion

According to the findings of this research, Tree-based models can be used to identify the factors that influence teen alcohol abuse. Teenage alcohol abuse is certainly influenced by a variety of factors, but for the NSDUH data, the most significant factors are whether or not the teen has ever used marijuana or tobacco, the teen's education level, and the size of the metro.

This research produced three models, the first of which was a binary classification model based on whether the teen has consumed alcohol or not. A decision tree model was created, and it performed with an error rate of 14.29% and the time taken to train the model was 0.02 sec. The second model was designed for multi-class classification. It could identify teens' drinking patterns and classify them into various categories, such as seldom, sometimes, frequent, and never used. The Random Forest model with a square root of predictors was chosen as the best one for this classification because it had the least error rate (13.5%) and took the least amount of time to train (0.744 seconds) the model. The third model was developed to predict the frequency of alcohol usage by a teenager in a year. The best-boosting model has an interaction depth of 1 and a shrinkage parameter of 0.01 with a mean square error of 120 and a training time of 0.53 seconds.

The majority (88%) of survey records for teenage alcohol abuse are categorized as "never used alcohol" in the NSDUH data. As opposed to the thought that the majority of youth consume alcohol, the data suggests otherwise. However, the inferences are highly reliable on the accuracy of the survey. As per the analysis, teenagers below 9th grade, who consume tobacco, and marijuana are more likely to consume alcohol. As a result, it is of the utmost importance that stringent regulations be enacted to prevent youth access to such substances. Future research would include collecting more survey data and considering more variables that could help increase the models' accuracy.

References

“National Survey on Drug Use and Health.” SAMHSA.gov, https://www.samhsa.gov/data/data-we-collect/nsduh-national-survey-drug-use-and-health.

“National Survey on Drug Use and Health 2020 (NSDUH-2020-DS0001).” Samhsa.gov, https://www.datafiles.samhsa.gov/dataset/national-survey-drug-use-and-health-2020-nsduh-2020-ds0001. Accessed 13 Apr. 2023.