What Causes Binge Drinking?
Statistical Techniques for Identifying At-Risk Students

Yuliya Dovzhenko, Brandon Podmayersky, James Yang
Princeton University

Abstract

Text of the abstract.

1 Introduction

Alcohol is a depressant drug that, when used in moderation, acts as a relaxant and social lubricant. Responsible use of alcohol for recreation is a fundamental part of American culture and many other cultures worldwide, but heavy use can have serious consequences. Alcohol abuse is a leading factor in violent crime and automotive fatalities, alcohol poisoning can be fatal in a short time period, and chronic alcoholism can lead to long-term liver failure and reduced brain function.

Binge drinking and overconsumption of alcohol are particularly widespread problems on college campuses, where many students are exposed to serious health risks from drinking. Efforts targeting the elimination of alcohol consumption completely have historically failed and represent a grievous threat to individual liberties. Instead, trends have pointed to education and healthy drinking practices to reduce the danger to students.

One thing that seems clear, though, is that not all individuals are at an equal risk of problem drinking. If we could find some way to identify students that are most likely to engage in risky drinking behavior, alcohol education programs and other initiatives could be targeted specifically at these groups. In this work we use survey data from the 2001 Harvard School of Public Health College Alcohol Study to show that information about a student's background, campus activities, and personal experiences and attitudes can be used to build a predictive classifier to identify students at risk of binge drinking. By dividing the survey data into five categories of data, organized by how difficult they are to collect, we also explore how well our classifiers perform using limited information.

We begin by introducing the dataset and our methodologies for processing the data and evaluating predictive classifiers. From there we move on to detailed explorations of the techniques we have used, including Generalized Linear Models, clustering techniques like k-means, and PCA. Finally, we use the insights gained from using these techniques to provide an overall evaluation of our work and how it impacts the problem of binge drinking on college campuses.

2 Dataset

Our analysis is based on the data from the 2001 Harvard School of Public Health College Alcohol Study [**?**]. This study surveyed 10,904 students from 119 different four-year universities about their drinking patterns, attitudes towards alcohol and other drugs, and participation in various aspects of campus life, among other things. The features used in our analysis come from the pool of 342 survey questions, ranging from basic demographic information ("How old are you?") to detailed information about a student's perception of their own campus ("If a student is caught on your campus using a fake ID to get alcohol, what is likely to happen to the student?").

We are interested in predicting whether a student is likely to engage in binge drinking. In accordance with the survey, we will define binge drinking as consuming five or more drinks in a row. We thus wish to predict the student's response to survey question C1: "Think back over the last two weeks. How many times have you had five or more drinks in a row?", with possible responses "none", "once", "twice", "3 to 5 times", "6 to 9 times", and "10+ times". The other survey responses will serve as features for our classifier.

3 Methodology

We now give a precise description of the methods we use to prepare the dataset for our use and evaluate predictive classifiers.

Response	% of Participants
none	61.22%
once	12.62%
twice	9.67%
3 to 5 times	12.10%
6 to 9 times	3.54%
10 + times	0.86%

Table 1: Responses to survey question C1: "Think back over the last two weeks. How many times have you had five or more drinks in a row?"

Class	Features	Description
1	30	Demographic, Background
2	77	Student Life, Personal Habits
3	82	Attitudes about Alcohol Policy and Student Drinking
4	9	Alcohol-Related Personal History
5	64	Alcohol Consumption

Table 2: The features in the alcohol data set were split into 5 classes based on semantic meaning and ease of collection.

3.1 Data Preprocessing

As is often the case in the real world, the data is not quite as neat and perfect as we would like it to be. Thus, we apply a series of preprocessing steps to prepare the data for our algorithms.

We begin by shuffling the participants randomly to ensure no systematic bias in the ordering of the data. The last 1,905 participants are set aside as a final validation set, and the remaining 8,999 will be used for training and testing when we perform evaluation using cross-validation. The features were all then converted from categorical formats to numeric formats, with integral values representing the possible answers to each questions. Modelling the features as numeric instead of categorical had no impact on the results except for drastically reducing runtime; all experiments were tried both ways but detailed results are omitted for brevity.

A substantial number of features had missing responses for many participants. In some cases this was because a question only needed to be answered if a previous question was answered a certain way (e.g. "If you transferred from another school, was your previous school located in the USA?"), and in others the data was simply missing. Nonetheless, this is problematic as our algorithms do not gracefully handle missing features. 34 participants did not provide a response to question C1 about binge drinking, so we discarded them completely, since there is no way to proceed without a value for the response variable. This left the breakdown of responses in Table 1. For the other features, we could not afford to completely throw away participants since almost every participant was missing some features. To cope with this we used a two-tiered strategy:

- Survey questions with less than 7,000 responses were eliminated from the data completely.
- Survey questions with 7,000 or more responses had their missing values replaced by the mean of the non-missing responses to the same question (rounded to an integer).

Intuitively, those features in the second category are preserved because there are few missing values and they are still likely to have significant predictive power. After this filtering process, 262 features other than the response remained, and these formed the basis for our predictions.

3.2 Evaluation Methodology

We evaluate our algorithms using the standard 5-fold cross-validation technique. The data was divided into 5 folds up front so that every algorithm would use exactly the same folds, and random fold generation would not cause differences in our predictive measurement accuracy. An algorithm is trained on each set of 4 folds and used to predict the responses in the remaining fold; this gives a predicted response for every survey participant that can be compared to the true response.

3.2.1 Evaluation Metrics

We are attempting to predict the number of times within the past two weeks that a student has engaged in binge drinking. This is an ordered categorical variable: the responses were binned (i.e. a student who engaged in binge drinking 3 times and one who did so 5 times would both be in the "3 to 5 times" category), but they are not independent classes - there is a clear numerical structure. For this reason we define two different evaluation metrics:

- The *percentage accuracy* is the percentage of students whose binge drinking category was predicted exactly. Under this metric, if the true answer is "none" then predicting "10+ times" and "once" are equally good.
- The *mean error* is the average of the number of categories between the true response and the predicted response. Under this metric, if the true answer is "none" then a prediction of "once" will receive an error of 1, while a prediction of "10+ times" will receive an error of 5 since the prediction is 5 categories too high.

Percentage accuracy shows us when the algorithm is getting answers perfectly correct, but it completely leaves out the numerical structure of the response. Mean deviation takes this into account and penalizes an algorithm less if its answer is close to the right answer.

3.2.2 Feature Classes

The participants in this study took a comprehensive survey detailing their lives and attitudes towards campus life and alcohol, but in other applications we may wish to predict binge drinking without necessarily having all this information. To quantify the impact of various information, we divide the available features into five classes based on how difficult they are to obtain and how closely (in the intuitive sense) they are related to alcohol consumption. Some information is easy to collect about any student but may yield little predictive power, while others may require a detailed survey but are quite predictive. The classes are as follows and are summarized in Table 2:

1. *Class 1* features correspond to easily-accessible demographic and background information such as age, gender, and religious upbringing.
2. *Class 2* features include student activities such as participation in extracurricular organizations and personal habits not related to alcohol directly, like cigarette smoking.
3. *Class 3* features identify attitudes about alcohol policy and student drinking, such as thoughts on the legal drinking age and perception of how much classmates drink.
4. *Class 4* features correspond to personal history directly related to alcohol, such as drinking behavior in high school and DUI charges.
5. *Class 5* features are direct information about current alcohol consumption, such as the average number of drinks per week. This class mainly serves as an upper bound for predictive accuracy, since if we have this we also probably know the student's binge drinking habits.

When evaluating algorithms we start by using only feature class 1, and then we add each class in turn - next using classes 1 and 2, then classes 1 through 3, and so on. This shows how well the algorithm can perform at each level of available data.

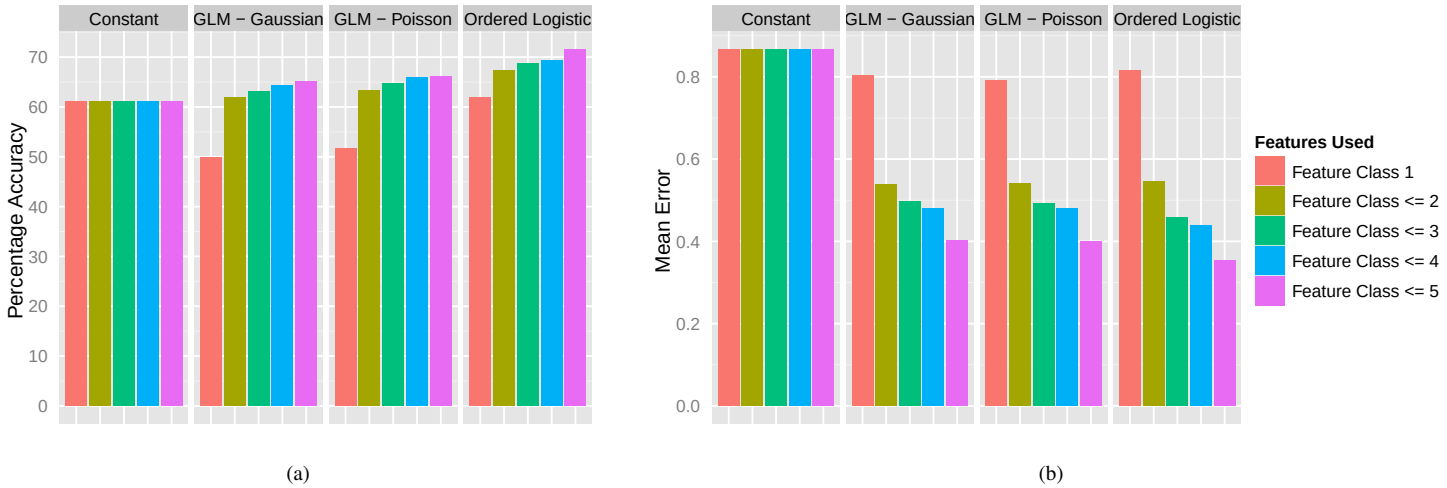
4 Generalized Linear Models

Our first attempts to identify at-risk students come from applying several variants of Generalized Linear Models. Given some subset of the feature classes, we build a linear model of how binge drinking varies with the input variables and use that to form predictions for novel data.

4.1 Algorithms

In this section we explore four different algorithms based on Generalized Linear Models. Here we use a single linear term corresponding to each input feature; see section 4.3 for a discussion of incorporating interaction terms. The four algorithms are as follows:

1. *Constant* - This extremely simple algorithm does not take the input features into account at all, and instead always predicts the most common class (no binge drinking in the past 2 weeks) as shown in Table 1. It serves as a lower bound on classification performance.
2. *GLM - Gaussian* - We use R's built-in glm() function [**?**] to model the category of binge drinking as a linear function of the input features, assuming a Gaussian error distribution on the response variable. We model our categorical response as numerical by giving each category a number from 1 to 6 in increasing order, so 1 is "none" and 6 is "10+ times". This predicts a real-valued response; we achieve better accuracy with some postprocessing. Namely, the real-valued predictions are clamped to the range 1 to 6 and rounded to an integer, since the correct response will always be integral. Empirically we found that the best rounding method is almost always to round down - this encodes our knowledge that most people are not likely to engage in binge drinking.
3. *GLM - Poisson* - This is the same as the *GLM - Gaussian* algorithm except that the response variable is modelled with a Poisson error distribution. The same postprocessing techniques are applied here, but using a Poisson distribution intuitively models the survey response data better than a Gaussian.
4. *Ordered Logistic Regression* - Finally, we use the lmr() function from R package *rms* [**?**] to perform ordered logistic regression, often also known as ordered logit. This captures the fact that the response variable is ordinal - the responses are divided into discrete categories, but the categories are not wholly independent as there is a clear ordering relationship between them. Given a new student's features, we find the probability that they fall into each category of binge drinking and predict the one with the highest probability.



(a) (b)

Figure 1: Predictive accuracy and mean error of each algorithm using various feature classes.

4.2 Results

Figure 1 summarizes the performance of the four algorithms using 5-fold cross validation on our training/testing dataset. As defined earlier, Figure 1 (a) shows the percent accuracy metric and Figure 1 (b) shows the mean error. The Constant algorithm which does not take the data into account at all achieves an accuracy of 61.2% and a mean error of 0.867. This gives an upper bound on the mean error we should expect to see, as all other algorithms achieve a strictly better mean error regardless of what features are available. In terms of predictive accuracy, though, both versions of GLM actually do worse than Constant when only class 1 features are used. The best performance comes from Ordered Logistic Regression when all features are available, with a percentage accuracy of 71.6% and a mean error of 0.354.

The two versions of GLM perform extremely similarly on the whole, with Poisson slightly outperforming Gaussian in most cases - this is not unexpected given the true format of the response variable. Interestingly, when only feature classes 1 and 2 are available, Ordered Logistic Regression actually has a slightly higher mean error than both versions of GLM. Its percentage accuracy, on the other hand, is still much better in these cases, reminding us that the two evaluation metrics capture subtle differences in algorithmic performance. Once class 3 or above features are available, though, Ordered Logistic Regression becomes the clear winner with the highest percentage accuracy and lowest mean error.

4.3 Extensions

There are several possible extensions to the algorithms used here that may lead to better predictive performance. Surprisingly, however, the ones that we tried did not lead to any improvements. Incorporating first-order interaction terms complicated the model and greatly increased the runtime of our experiments but did not increase the percentage accuracy or decrease the mean error. To make the linear model simpler, we tried reducing the number of terms by building the model incrementally, using the Akaike information criterion [] to decide which linear and interaction terms to add. Again, this was much more computationally expensive than the basic models but performed no better. Finally, we also tried regularized regression in the form of LASSO [**?**]. Full results are omitted for brevity, but while LASSO's basic predictions were better than our GLM algorithms at some values of the regularization parameter, it seemed to interact poorly with the rounding/flooring technique we use to convert the predictions to a categorical variable. Further investigations on this front may yield improvements, however.

5 Clustering

FILL ME IN!

6 PCA, ICA, SVD

What is these even?

7 Final Evaluation

FILL ME IN!

8 Conclusions

Conclude meh! Add references too!