Case study 3: Predictive modelling of alcohol-associated risks in College students

Olivier Binette and Raphael Morsomme

1. Introduction

[TODO: Background info about the College alcohol study.]

The goal of our case study was to develop a predictive model of alcohol related risks in college students using information readily available to schools, in order to help:

1. identify students at risk and allocate support ressources as effectively as possible; and
2. determine other pieces of information that a school might additionally gather identify students at risk.

To address the first question, we develop a base predictive model which takes for input a student's demographic information, information about their living accomodation on or off campus, and their GPA, in order to predict a "ressource need" score variable. This score variable is composed of a student awareness score and three interpretable risk scores (for consumption risks, behavioural risks, and situational risks). Responses from the College Alcohol Survey were used to score individuals in these categories and train the predictive model, and the Conformal Prediction framework is used to provide prediction uncertainty quantification.

For the second question, we studied the gain in predictive performance that can be obtained using additional predictors related to student well-being and interests. These predictors are not directly related to alcohol consumption (although one of they include a survey question about the importance of partying) and could reasonably be probed for in order to help determine a student's risk.

1.2 Important considerations for predictive modelling

Predictive modelling comes with particular challenges and considerations which should be addressed in the context of a real-world application. In this case study, we adress the following two points:

- **Meaningfulness:** We construct an interpretable and meaningful response variable disaggregated across student awareness and across three kinds of alcohol-related risks. While we do not have the subject-matter expertise necessary to properly weight the different risks, this opens up our modelling approach to scrutiny and improvement.
- **Reliability and out of sample performance:** We provide uncertainty quantification for the predictions with exact frequentist coverage under weak assumptions. That is, we quantify the accuracy of our model through a quantity Δ such that, for any prediction p, $p \pm \Delta$ is a 95% confidence interval for the predicted value. This Δ is obtained through the conformal prediction framework by computing the marginal distribution of the out of sample prediction error.

Additionally, the following should be considered if a predictive model like the one we proposed were to be used in practice. We do not address these in this case study.

- **Fairness:** The use of age, race, gender, religion and other variables as predictors is problematic for schools under Title IX. Data quality and reliability among these groups, as well as the meaningfulness of the response variable we define for them and the practical implications of the use of such a predictive model should be carefully considered prior to any implementation.
- **Data representativeness:** We only used data from the 2001 College Alcohol Survey. The information it contains may be outdated and is certainly unrepresentative of the student population at any given school. Post-stratification and other adjustments could be carried out in applications.

2. Material and methods

2.1 Response variable