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

Olivier Binette and Raphael Morsomme

1. Introduction

The 2001 College Alcohol Survey is the last element of a multi-round survey carried out at three previous occasions starting in 1993. It investigated multiple aspects of student life and background information as relevant to alcohol consumption and its consequences.

Using this survey, 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.

1.3 Outline

2. Material and methods

2.1 Response variable

We define the student "ressource need" variable as

$$\text{ressource need} = (\text{2-awareness})(\text{consumption risk} + \text{behavioural risk} + \text{situational risk})$$

where student awareness of alcohol risks score and other risk scores are determined from answers on the College Alcohol Survey.

To provide more details, the complement of the awareness score is the mean of a school policy knowledge score (proportion of "Don't know school policy" answers to questions B3 and B5), of a school provided information score (proportion of "no provided information" on questions B6A-B6G), and of an educational material score (proportion of "no educational material or programs" on questions B7A-B7E).

For the alcohol-related risks, we looked at individual survey questions and associated to each possible answer a number between 0 and 1. Each of these numbers represents the marginal probability that a student providing this answer has alcohol-related issues. While it is not obvious how to aggregate these marginal probabilities without specifying a covariance structure between the questions, we opted for the complement of the geometric average of the complementary probabilities. That is, if the marginal probabilities are $p_1, \ldots, p_k$, then the resulting score is $1 - \prod_i (1 - p_i)^{1/k}$. Missing values were omitted. In this way, having large marginal probabilities of alcohol issues on a few questions is given more weight than a low probability of alcohol issue on many questions.

The consumption risk score involves the binge drinking indicator, the frequent binge drinking indicator, as well as question C7 (self-description of alcohol consumption). The behavioural risk score involves the drunk driving indicator, the binge drunk driving indicator, as well as questions C17A-C17K and questions C18E, C18F (consequences of drinking). Finally, the situational risk involves questions D1A-D1C and D1H (consequences of drinking of other students).

The measure was constructed for illustration purposes and should be refined using expert knowledge. Ideally, a model relating the different risks to the survey answers would be specified and trained using examples analyzed by experts.

2.2 Predictive models

As described in the introduction, we build two predictive random forest models [TODO: SAY MORE ABOUT RANDOM FORESTS].

The baseline model only takes for input demographic information about students (questions A1-A3, G1-G4), information about living accomodations (questions A6, A7, B9) and other information that is readily available to schools such as participation in greek life and student GPA (questions A4, A5, F5, B2, B9).

The augmented data model also incorporates information about students preferences (question A8A-A8I), about student activities (question F6A-F6I) and about student satisfaction with life and education (questions F1-F4).

Predictors with more than 25% missing values were removed from the analysis and we then proceeded with a complete case analysis.

The quality of the fit of the models is assessed over the training sample using the percentage of "variation explained". Furthermore, the generalization performance is assessed by looking at the predictive error distribution through the conformal prediction framework detailed in the next section.

2.3 Conformal predictions

3. Results

4. Discussion

Appendix

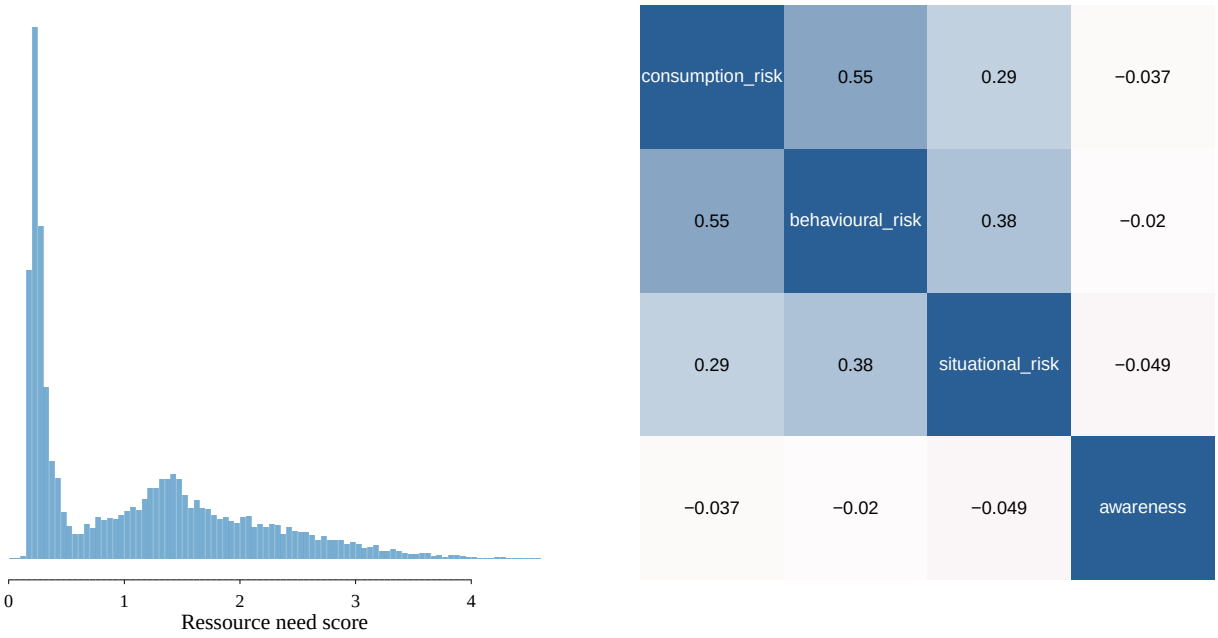

Figure 1: Histogram of the "ressource need" response variable (left) and representation of the correlation between the different components of this variable (right).