Predictive modelling of alcohol-associated risks in college students

Olivier Binette and Raphael Morsomme

February 18, 2020

Goals

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;
2. determine if additional information could help better identify students at risk.

Assumption: alcohol-related risks are an important issue that schools want to address by offering support to students in need.

Challenges

What we deal with:

1. **Meaningfulness.** We predict a "student need" score which is a function of student awareness and alcohol-related risks.
2. **Reliability.** We provide interval predictions with exact frequentist coverage. This communicates uncertainty in the prediction and could help mitigate issues related to over-confidence in the model.

Challenges

Things we don't deal with (but we should):

1. **Interpretability.** It is difficult to summarize the model and explain the predictions.
2. **Fairness.** Non-discrimination (title IX). Issues using race, gender, age as predictors. Suitability of the "student need" response across these groups and quality of the data among them.
3. **Data representativeness.** The data may not represent a given school's student population and post-stratification would be necessary.

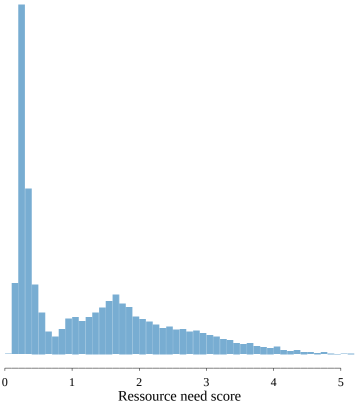
Response variable

- Student awareness score in $[0,1]$: school policy awareness and information received at school.
- Risk scores in $[0,1]$:
  - **Consumption risk**: "binge" drinking and self description.
  - **Behavioural risk**: drunk driving, missing classes, hangover, regret, medical issues, trouble with police, etc.
  - **Situational risk**: insulted, assaulted, damaged property, etc.

$$\text{need score} = (\text{2-awareness})(\text{consumption} + \text{behaviour} + \text{situational})$$

Ideally, use expert advice... this is only a coarse approximation.

Response variable



consumption_risk	0.55	0.29	−0.13
0.55	behavioural_risk	0.38	−0.086
0.29	0.38	situational_risk	−0.17
−0.13	−0.086	−0.17	awareness

Predictors

Base model ($p = 15$):

- Demographic information (age, gender, year in program, race)
- Living accommodation (in dorm, alone, with roommates, spouse or parents; type of dorm, in a fraternity or sorority).
- GPA.

Augmented data model ($p = 39$):

- Predictors of base model
- Ratings of importance of different aspects of student life (athletics, arts, partying, etc)
- Time doing various activities (tv, study, work etc)
- Satisfaction with education and life; friendships and mentorship.

Conformal Prediction

Conformal prediction (to be defined) allows us to:

1. Associate a measure of certainty to any prediction.
   - Regression setting: accompany a point prediction with a prediction interval
2. Compare the fit of the two models by looking at the tightness of the prediction intervals.

Conformal Prediction

Conformal prediction generates prediction *intervals* that are

- valid at a given significance level for *finite* sample (Vovk, 2005)
- distribution-free
- universal
- individualized (Papadopoulos, 2009)
- only assume exchangeability
- cheap (Papadopoulos, 2002), unlike bootstrap

Inductive Conformal Prediction

Given a labeled training set $\{z_i = (x_i, y_i)\}_{i=1}^n$ and an unlabeled test observation x_{n+1},

1. partition training set into a *proper training* set $\{z_j\}_{j=1}^l$ and a *calibration* set $\{z_k\}_{k=l+1}^n$
2. fit predictive model on proper training set
3. compute predictions $\hat{y}_k$ on calibration set and anomaly scores

$$a(z_k) = |\hat{y}_k - y_k|, \quad k = l+1, \ldots, n$$

4. identify a_ϵ, the ϵ^{th} percentile of the $\{a\}_{k=l+1}^n$
5. compute prediction on test observation and set the prediction interval to be

$$\{y : |\hat{y}_{n+1} - y| < a_\epsilon\}$$

Set up

- Predictive model: random forest ($1,500$ trees, $m = p/3$, little pruning).
- Test set is 10% of data set.
- Calibration set is 30% of training set.
- Repeat 10 times to obtain the expected width of prediction intervals for each model at various significance levels.

Results - Coverage

	Significance	Set of Predictors	Mean Width	Coverage
1	0.500	Extensive	1.134	0.499
2	0.500	Restricted	1.367	0.505
3	0.750	Extensive	1.806	0.753
4	0.750	Restricted	2.112	0.754
5	0.900	Extensive	2.511	0.906
6	0.900	Restricted	2.808	0.900
7	0.950	Extensive	2.902	0.954
8	0.950	Restricted	3.218	0.951

Table: Coverage and Mean Width of Prediction Intervals

Results - Width

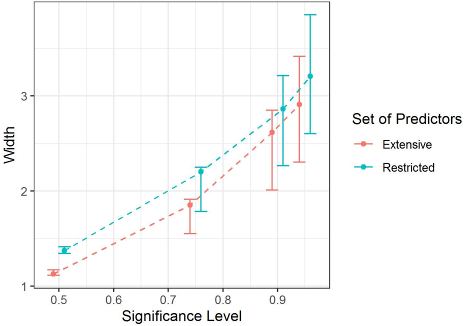


Figure: Median and inter-decile interval width across significance levels.

Results

Base model:

- 16% of variance explained
- Most important predictors: race, part of fraternity or sorority, having roommates or not, etc.

Augmented model:

- 39% of variance explained
- Most important predictor: how much the student likes partying.
- Tighter prediction intervals.

Conclusions

- Baseline student information provides some but limited information about the student "ressource need" variable.
- The random forest model does not perform much better than a linear regression in terms of R^2 value (16% in this case; 39% for the augmented model). Interpretable models would be more appropriate.
- Asking students how they spend their time, what they value the most at college, and how satisfied they are with their education considerably improves predictive accuracy.
- Can use mondrian conformal prediction to make the model fair (valid intervals *conditioned* on, say, gender)

Results - Variable Importance (Base Model)

	Variables	Importance
1	race	123.5
2	roommates	91.2
3	greek_life	90.2
4	marital_status	72.4
5	religion	67.6
6	age	67.1
7	location	65.1
8	live_parents	64.9
9	hispanic	42.9
10	transfer	38.9

Table: Variable importance for predictive model with restricted set of predictors

Results - Variable Importance (Augmented Model)

	Variables	Importance
1	parties	268.0
2	religion	77.0
3	race	72.2
4	roommates	62.1
5	greek_life	46.4
6	marital_status	40.9
7	socialize	40.3
8	live_parents	38.5
9	friends	29.3
10	location	28.1

Table: Variable importance for predictive model with extensive set of predictors

References

Papadopoulos, H., Proedrou, K., Vovk, V., & Gammerman, A (2002)
Inductive confidence machines for regression
European Conference on Machine Learning, pp. 345-356

Papadopoulos, H., Vovk, V., & Gammerman, A. (2011)
Regression conformal prediction with nearest neighbours
Journal of Artificial Intelligence Research, pp. 815-840

Vovk, V., Gammerman, A., & Shafer, G. (2005)
Algorithmic learning in a random world
Springer Science & Business Media.