Breiman's Two Cultures Response

Andrew Bates

11/27/2018

Prompt

You are confronted with the problem of studying student success in an introductory, freshman level Statistics course. You have been given a wealth of data:

- Student information database: All the demographic information you can imagine! E.g., gender, socioeconomic status, performance in and location of high school, characteristics (ethnicity, veteran, first-generation college student, etc.), commuter status, work status, etc.
- Learning management system database (LMS; e.g., Blackboard): performance on course assessments (homeworks, quizzes, tests, in-class participation), time stamp and time-on-task for online quizzes, clicks and time-stamp on LMS items, participation in discussion boards.
- Publisher learning management system database: time stamp, time-on-task, and exact interactions with the e-textbook, online homeworks (each item on a homework assignment), and online data analysis tool (e.g., Pearson's MySTATLab).
- Interventions: student participation at a campus tutoring center, office hours, recitation sections, campus programs to help at-risk students (time stamp and time spent)

Approach this problem from Breiman's two cultures framework. Briefly discuss

- How you would take the algorithmic modeling approach to study student success with this wealth of data?
- How you would take the data modeling approach to study student success with this wealth of data?
- How can the algorithmic modeling and data modeling approaches compliment each other as you report your findings to key campus stakeholders? In particular, communicating with the course instructor for potential instructional reforms; with the campus administration for a summary of student success in the course and to inform decisions on resources devoted to the course.

Response

A few comments before responding to the questions directly. One question that would need to be settled before embarking on this project is how student success will be measured. Will it be their final grade in the course, whether they pass or fail, or maybe something more ambitious like the grade in the next statistics course they take? For now, we will assume that success means passing the course. However, I don't think this is necessarily the 'correct' criteria. Second, with this much data there are surely several types of analyses that can be done. Assuming access to the data is recurring and not a one-time opportunity, recurring analyses can be done to analyze the effect of altering the course in various ways and would allow for adjustments based on a changing population. We will assume here that access to the data will be forevermore and we wish to do an initial analysis to suggest strategies for improving student success.

Algorithmic Modeling Approach

Taking this approach with access to so much information would be like a 5 year old on a new playground. To start, I would consult with instructors (or someone that knows more about education than me) as to how to

best measure model performance. Are false positives more important than false negatives? By how much? I would then try several different methods to determine which one can best meet this criteria. All methods would need to have the ability to measure variable importance. The importance can be used as a way to determine what has the most impact on student success. Once a final method is settled on, it can be used in subsequent courses to predict whether each student will pass. At the end of the semester, the accuracy (or false positive rate, whatever is most important) can be computed. If it's unsatisfactory, we can go back to the drawing board with this new data in hand. This process could then be repeated until performance is up to par.

Data Modeling Approach

With this method of course, we would need to decide on a class of models. Logistic regression is a natural candidate, but maybe we want to do a mixed effects logistic regression where we control for things like the instructor. I would also probably consult with an education expert but in a different way. With so much data, I would want their input as to what variables they think are important to help narrow the search down. Some sort of model selection procedure would be necessary here as there would just be too many variables to do this by hand. Using the education expert suggestions as a starting point, we would want to limit the number of variables because realistically, interpretability would be lost if there are 100 variables and this is part of the point with this approach. Supposing we use a logistic regression, we can determine important factors maybe by the odds ratio. This can then suggest areas that should be considered for change.

Both Approaches

One way both approaches can be combined is by using something like a lasso or elastic net with logistic regression. These methods sort of sit at the intersection of both approaches. Another way we might do this is to use the algorithmic modeling approach to determine the top X variables as measured by variable importance. We could then take the data modeling approach on this subset of variables.

Communicating the results would be quite a bit different between the instructor and administration. For administration, we would want to simplify things as much as possible. For example, suppose rate of attendance tutoring center are deemed most important. We would say something like "students who attend the tutoring center more than once a week are 5 times as likely to pass the course" and suggest that longer hours or more staff at the center would vastly improve student success. For the instructor, we can get a bit more into the details. We can explain why we determined that tutoring center attendance should be their primary concern. Is this based on a coefficient from a logistic regression model, or on variable importance measure from a random forest that has 98% accuracy? Maybe one of the other top variables is attendance at office hours with students who attend more often having higher proportion of passing. We could then suggest they ask their students why they don't attend more. Maybe office hours each semester are scheduled at the same time as Intro to Business and most of the students simply can't make it to office hours.