# College Admissions and Data Mining

Sam Strasser
February 7, 2008

Ask any senior in high school that is considering college and he will tell you that the college application process is opaque and often entirely arbitrary. Some students get accepted to a school that rejects a seemingly equally qualified student, and the process seems downright unfair. Data mining could provide just the answer to the constant questioning of the students who cannot be involved in the admissions decisions. From the perspective of the educational institutions, the process may seem just as arbitrary. After students do a sizeable amount of work to apply to a college, and then that college accepts the applicant, students turn down that offer. Other students accept the offer but drop out or transfer out of the institution before completing their degree. Obviously data mining will not result in 100 percent yield or 0 percent dropout rates, but understanding factors that predict yield and dropouts will help the institutions make key decisions.

## Student Retention

Several data miners have attacked the problem of predicting which students will drop out or transfer out of an institution. By predicting likely dropouts, the research will allow better intervention and support programs that can be tailored to individuals. The case study also suggests that by convincing students not to transfer or switch majors, the programs can save the students money. This kind of fiscal argument appears over and over again in explaining the rationale of doing data mining in the area of college admissions. Both of these given reasons suggest that it is the students' interests that drive data mining initiatives. It is not too far of a leap though to assume that colleges might be less likely to accept someone they thought might drop out.

The study of retention uses some fairly intuitive academic factors to predict whether a student will graduate, such as high school GPA, AP scores, grade on the state English exam and the number of times the major was changed. GPA turns out to be vastly more important than any other predictor, as might be expected. The study also uses demographic factors such as age, gender, minority status (binary value representing either minority or non-minority) and family income. Somewhat surprisingly, the second most important factor in predicting whether a student will drop out is age. The last category that the study uses in performance in college, including college GPA and the number of credits attempted. The regression model using these factors predicted the data in the test set with between 85% and 95% accuracy (three different methods were used).

The results, though fairly interesting, are beside the point. The only factor with a positive effect on dropout rates was minority statues. In other words, minority students were more likely to drop out than non-minority students, holding all other factors equal, bringing complex issues like race to light. To some it might even suggest that being a minority makes you less likely to complete school and therefore less attractive as an applicant. The study again leads to the proposition that data mining be used in the admissions process. The two most important predictors, age and high school GPA, are both known at the time of application, and could therefore be used to help determine an applicant's likelihood of completing the degree.

## Admissions Yield

Administrators of educational institutions worry about more than student retention, and focus much of their attention on yield during the admissions process. They hope to find out which students will matriculate if they are accepted. Several specific institutions have conducted studies to examine their own yield, and what can be changed. In one such study, an unnamed institution is studied. The factors are split into demographic, academic and recruiting communication and are intuitive. Gender, ethnicity, age, high school size, GPA, SAT scores and high school rank are all used. The study also analyzes the number of times the school communicated with the prospective student, and in what form (e-mail, phone, letter etc.). The study ran on two sets of data: the current year (splitting half into training data and half into test data) and the following year. For the current year, the algorithm correctly predicted the applicant's choice 75% of the time, but for the following year, it was only accurate 64% of the time. The relatively poor results did not provide much by themselves but did point out that financial aid offerings, omitted from the data, played a much larger part than administrators had thought. By doing the study, they gained a mildly effective algorithm but a vastly smarter approach to analyzing yield.

Another study, done at Willamette University in Oregon, did far better. A class of students in the MBA program there was assigned a yearlong project whose goal was to better predict yield for the next year of admitted students. During the first year of the program, the resulting model produced equally good results to the outside consultant the university had been using. By doing the processing themselves, though, the university realized important factors that had not been included, such as financial aid. They found out that by increasing financial aid, they could increase an accepted student's chance of matriculating, which is not a surprising result. That result does have a subtle but crucial implication for the admissions office. By adjusting the potential financial aid offering for a given student, the admissions officer can very quickly see what amount of money would induce the student to

go to their school. Not only could the model help officers determine financial aid offerings, but it allowed the officer to adjust offerings without understanding the complex underlying model.

Although it was significantly more accurate, predicting the yield to within 2.5% instead of 15% as before, the model created a completely unintended side effect. The yield predictions increased dramatically, but the ethnic diversity of the student population dropped. It is not exactly clear why, although it would not be too hard to guess. Still, the important point is to note that data mining techniques chew numbers and without including in the goals all of the goals, like maintaining diversity, the model will "fail" to achieve what we want it to.

## Admission

The college application, after all that mining, remains mysterious. Schools are traditionally not particularly willing to share the details of their admissions process. It is possible that all these methods are available at many institutions already (but not too likely). Using data mining to change college acceptances brings up a lot of issues that need careful debate. As noted, race played an important role in retention rates and it could feasibly have a large role in any such process run on admission data. Consider the case *Gratz v. Bollinger*, in which the University of Michigan was sued for using affirmative action in its admissions policy. The University lost because it assigned too high of a score to minorities and because their use of affirmative action was not narrowly tailored to the goals of the University. Imagine how different that argument might have been if instead of randomly picking a score to assign to minorities (they picked 20), the University used the now familiar methods of data mining to establish "narrowly tailored" goals, and then to algorithmically achieve them. It may be unclear how institutions currently use data mining in their admissions, but it is obvious how they could reshape many of the recurring debates around the issue.

## References

Christopher M. Antons, Elliot N. Maltz. "Expanding the Role of Institutional Research at Small Private Universities: A Case Study in Enrollment Management Using Data Mining"

Lin Chang. "Applying Data Mining to Predict College Admissions Yield: A Case Study"

Serge Herzog. "Estimating Student Retention and Degree-Completion Time: Decision Trees and Neural Networks Vis-a-Vis Regression"

Sutee Sujitparapitaya. "Considering Student Mobility in Retention Outcomes"