BUS 41204 - MACHINE LEARNING - PROJECT PROPOSAL
WINTER 2017

(Machine) Learning in Higher Education Admission and Enrollment

Sondre Skarsten, Kevin Tse and Uyen Tran
February 13, 2017

This research would examine the determinants of admission and enrollment at a Midwestern University in the United States. The following proposal will give an overview of the questions we will attempt to answer and a detailed description of the data we have acquired.

1 INTRODUCTION

Higher education is an understudied market mostly because of scarity of data. There are a number of outstanding questions that could be explored at the university and student level. For example, financial aid and scholarships have long been used by universities and colleges as an accepted method to price discriminate their student customers. By utilizing these financial offerings, a university can reduce the net price for certain groups of student recipients, e.g. low-income and/or high-performing students, such that enrollment objectives and university revenue are maximized. For some universities in highly competitive tranches of the market, the selection of these offerings becomes a key component in attracting students where education alone can no further be differentiated.[1] Likewise, where a university's education and prestige earn them a place in a less-crowded, luxury tranche of the market, monopolistically competitive forces allow for increases in tuition costs without decreasing enrollment.

In order to effectively use these tools, universities must have adequate and accurate information regarding their student applicants. Universities mitigate asymmetrical information obstacles through rigorous application processes. Using the Free Application for Federal Student Aid (FAFSA)[2], writing samples, test results, resumes, cover letters, interviews and other information gathering techniques, universities obtain pertinent student details such as college placement scores, high school grade point average, family household income, etc.[3]

Tuition for higher education is well suited for the theory of price discrimination for the following reasons:

- Students cannot resell their admission and enrollment on a secondary market, thereby eliminating arbitrage factors.
- Differences in price elasticities among students result in different optimal prices when price discriminating.

[1] Each university is differentiated by quality, major options, resources provided, location, sports teams, etc.
[2] A growing number of universities also require the College Scholarship Service profile (CSS) which is offered by the College Board. It is designed supplement the FAFSA in order to give a closer look into household finances and expenses of a student and family. It is much more detailed than the FAFSA.
[3] Universities have access to financial aid information which reveals some heterogeneity in budgets, though perhaps not in terms of preferences

Ideally, this project would help capture different aspects of admissions decisions, price variation and price discrimination.

2 DATA

The data that we are using is from a school (which at this point requires anonymity). The school is located in the Midwestern region of the United States. The sticker price tuition rate is approximately $40,000 per year, and $50,000 per year when additional costs are included (residence, dining, and student fees). There are between 1,000 and 2,000 students enrolled in the college. Approximately 50% - 60% are female. The major of students enrolled at the college are out of state residence, between 70% and 80%. Like most colleges, this colleges offers a wide range of majors and minors including pre-professional programs designed for graduate school admission.

Our data includes the data from 2014 and 2015 admissions. It includes the students that applied to the college, the ones that were accepted and the ones that ultimately enroll. We have detailed data on each of the following variables at the student level:

- **Demographics**:
  - Birthdate
  - City, State and Zip code of residence when the student applied
  - Gender
  - Ethnicity
  - Dependency - whether the applicant is still a dependent
  - Household Income
- **Student Aptitude Measures**
  - High School Grade Point Average (GPA) [4]
  - ACT and/or SAT scores of all sections (reading, writing, math)
- **Application Details**
  - Date of Application
  - Early decision - whether or not the student applied early decision, signaling that if the applicant was accepted by the college he/she would enroll)
  - Athletic Application - whether the student applied as an athlete
  - Major(s) - what major the student applied to be admitted to
  - Campus visits - whether or not the student came for campus visits, and how many times
  - Part-time or Full-Time application
- **Admission and Enrollment Outcomes**
  - Admit Type - whether they were waitlisted or admitted
  - Enrollment
  - Financial/Merit awards - the college gives a number of different awards to help with the tuition costs

Using this data we will start by looking at the first stage of decision making: whether or not the student was admitted given the information set. Then the second dependent variable we could explore is conditional on acceptance whether or not the student enrolled. These patterns could be further enhanced by looking at the value of price discrimination via the different financial rewards the school offers the admits.

Ultimately, we would like to examine the following three items provided our data's limitations:

1. A model to predict the yield rate given the characteristics of an applicant;
2. The marginal change in yield rate with respect to change in merit scholarship for different subsets of students (perhaps we can find which subset students respond more elastically to price);

[4] This college also reweights the GPAs based on the rigor of the high school

3. Given an annual budget of financial aid, academic and diversity objectives, return is the optimal distribution of merit scholarships to offer to the students.

Using the tools we have already learned we can look at both predictive power by leveraging boosting and bagging. Moreover, we could also use cross-validation to see what variables are most important for these predictions.

3 Supplemental Data

Throughout the exploration process, several other data sources were considered. This section discusses the merits these other data sets and how we might be able to leverage them in this project.

3.1 The National Center for Education Statistics (NCES)

This survey is a nationally representative, longitudinal study of high school sophomores in 2002 and high school seniors in 2004. The students are followed throughout high school and college years. The unique feature of this survey is that it includes surveys of the student, the student's parents, math and English teachers and school administrators. Additionally, along with complete high school transcripts, the data set also includes student assessments in math from 10th and 12th grade as well as English assessments from 10th grade.[5]

What makes this data set relevant to the underlying demand model for this specific college. Some of the responses could be proxies for the student preferences for different types of colleges. Previous versions of the NLS questionnaire (1972) also include which schools the student, as a senior, was accepted to (up to 3 colleges/universities) as well as which school the student accepted offers from.

3.2 College Board Data

The College Board collects data on SAT, AP and PSAT/NMSQT on each of the students that takes these exams. This data includes information on race and ethnicity as well as other time trends across other demographic variables. The data can be broken down by state and grade level. Due to the unique position that the College Board is in, the data they have includes where these exam scores were sent.

4 Conclusion

Using the data we already have, we will examine the patterns surround admission into the college as well as the probability of enrollment conditional on acceptance. These findings could be interesting and useful to the college and other colleges as the admission process is expensive and as a result resources could be reallocated more efficiently.

[5] Note that due to the detailed information, this data set is highly restricted.

Sources:

College Board. (2013). Trends in Higher Education Series: Trends in College Pricing (2013). trends.collegeboard.org. The College Board.

Katherine Baicker, David Cutler and Zurui Song. Workplace Wellness Programs can Generate Savings, Health Affairs, 29, no.2 (2010):304-311 (published online January 14th, 2010; 10.1377/hlthaff.2009.0626).

Wieczner, Jen "Your Company Wants to Make you Healthy", Wall Street Journal, April 8th, 2013. Assessed online: November 1, 2013 - online.wsj.com/news/articles/SB10001424123