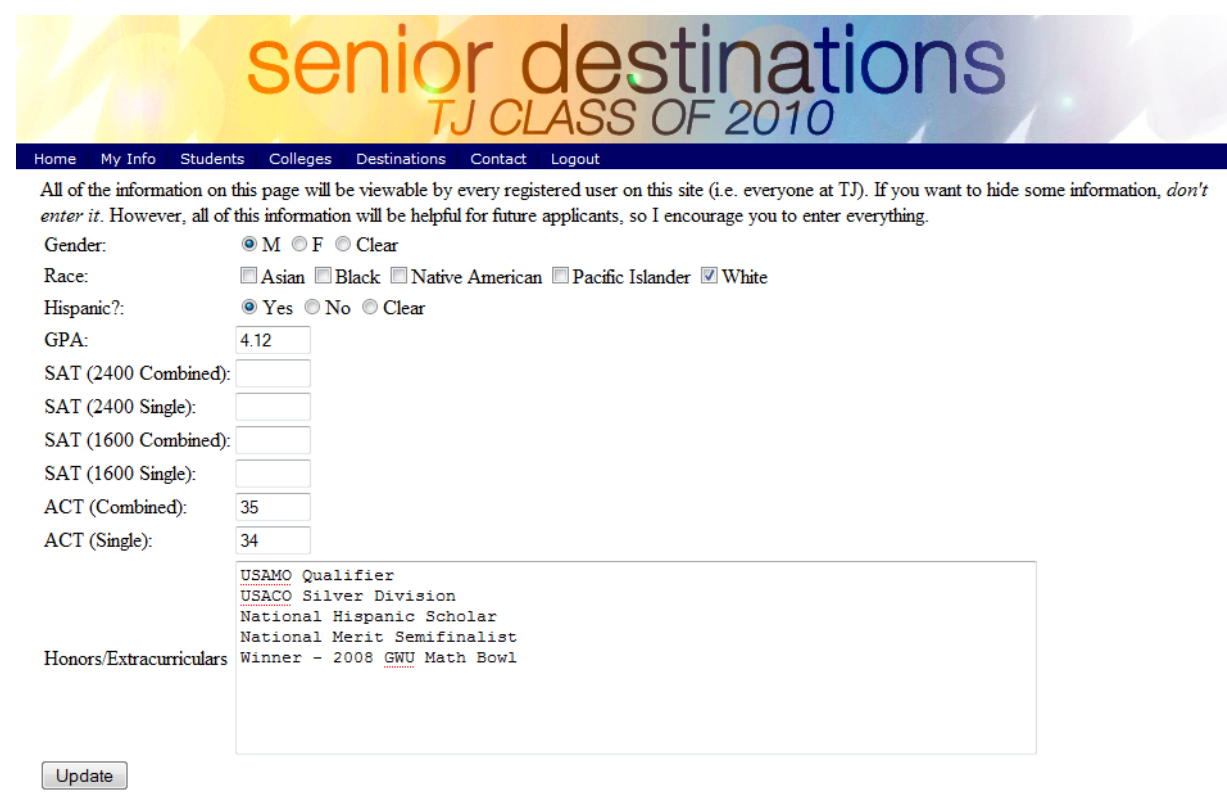


Machine Learning of the College Admissions Process

Sam Rush

Abstract

The goal of this project is to analyze the various biases that exist in the college admissions system by attempting to predict college decisions. This project will attempt to reduce college admissions to pure numbers, excluding data that is inaccessible such as essays and teacher recommendations. Past user-submitted data from the 2007, 2008, and 2009 *Senior Destinations* websites will be used to train an algorithm which will take an application as input data and output a decision. Then, factors such as the gender bias and the race bias will not only be proven to exist but will be quantifiable based on their role in the least squares fit.



senior destinations
TJ CLASS OF 2010

Home My Info Students Colleges Destinations Contact Logout

All of the information on this page will be viewable by every registered user on this site (i.e. everyone at TJ). If you want to hide some information, *don't enter it*. However, all of this information will be helpful for future applicants, so I encourage you to enter everything.

Gender: M F

Race: Asian Black Native American Pacific Islander White

Hispanic?: Yes No

GPA:

SAT (2400 Combined):

SAT (2400 Single):

SAT (1600 Combined):

SAT (1600 Single):

ACT (Combined):

ACT (Single):

USAMO Qualifier
USACO Silver Division
National Hispanic Scholar
National Merit Semifinalist
Honors/Extracurriculars Winner - 2008 GWU Math Bowl

Image 1: An example page of the Senior Destinations site, where students can enter their information.

Introduction

The college application process has become a hypercompetitive environment in which students embark on a four year process of padding their resume to look impressive to an admissions officer. College admissions is often publicized as a wholistic process in which admissions officers look at everything without "weighting" certain aspects of your application such as GPA. Therefore students look to excel in all areas instead of taking the most efficient path, which is not immediately obvious. So, how do we determine what's really important to a college? In this paper I attempt to answer that question.

Procedure

The goal is to solve the system $Ax=B$, where A is a matrix in which the rows are students and the columns are admissions factors (SAT, GPA, etc.) and B is a column vector of the students' decisions (i.e. 1 for accepted and 0 for rejected) I am using a linear least squares fit calculated using the QR decomposition to solve this inconsistent system. Then, I extend this method to nonlinear least squares quickly using the Gauss-Newton method since the QR decomposition has already been obtained.

Results

The computer does a decent job at predicting admissions based only on GPA, SAT scores, and Gender. The algorithm is ready to use all factors at the disposal of the Senior Destinations site next year, which amounts to over 20 factors for each application as opposed to just three. Unfortunately this was not possible this year due to the poor design of previous years' sites.

College	# Correct	Out of	Prediction Rate
Brown University	36	29	80.6%
Cornell University	54	65	83.1%
Duke University	47	56	83.9%
University of Pennsylvania	34	41	82.9%
University of Virginia	121	130	93.1%
Virginia Tech	64	64	100%

To illustrate the regression that the machine currently uses, I have included graphs with only SAT and GPA (obviously with a 3rd parameter, we would not have enough physical dimensions to view the graph) below.

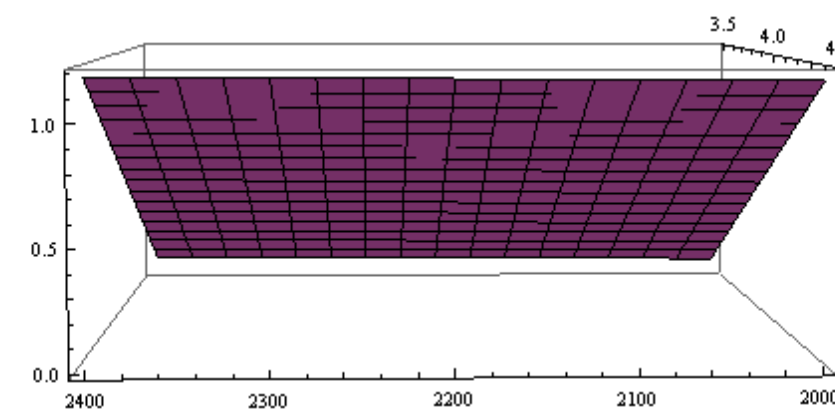


Image 2: The graph of SAT vs. GPA vs. Acceptance (Accept is 0.5 and above) for UVA.

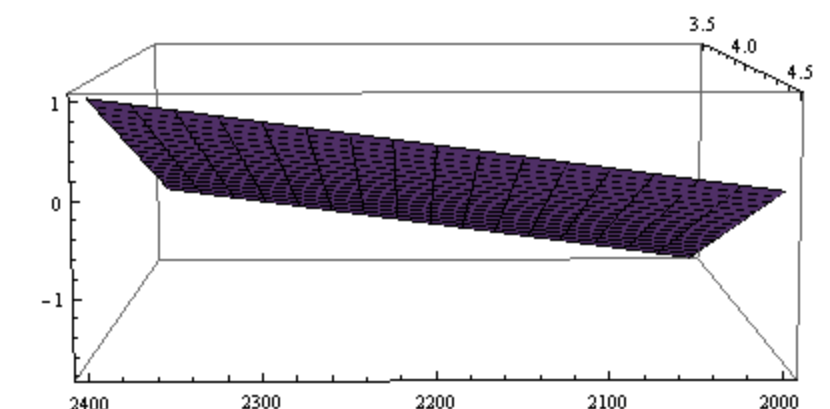


Image 3: The graph of SAT vs. GPA vs. Acceptance (Accept is 0.5 and above) for UPenn.

Discussion

The two graphs illustrate the different methodologies that these two institutions use to select their students. UVA's graph has a steep slope in the GPA direction and an almost unnoticeable slope in the SAT direction, indicating that it cares a lot more about your GPA than your SAT. Penn's graph, on the other hand, has a much larger slope in the SAT direction, but still a greater slope in the GPA direction, indicating that your SAT will be a determining factor in your application, but not as much as your GPA will. Also notice the scales on the axes for the two graphs. A much larger percentage of the UVA graph is in the "accept area" (greater than 0.5 on the z-axis) than the UPenn graph. This should make sense, as it UPenn is generally harder to get into than UVA.

Conclusion

The 80% success rate was hit for 14 of the 15 colleges with over 30 applicants with GPA and SAT data. The one college for which it failed was Harvard University, which was highly unusual in its lack of selection of TJ students for their class of 2014. Therefore the project is a success. The Senior Destinations website, along with a new "Will you get in?" page will be bundled together and given to a member of the TJHSST class of 2011 to be used for next year's destinations site. When the number of factors available to be used increases from 3 to approximately 20 next year, I expect the rate for most colleges with sufficient data to exceed 90%.