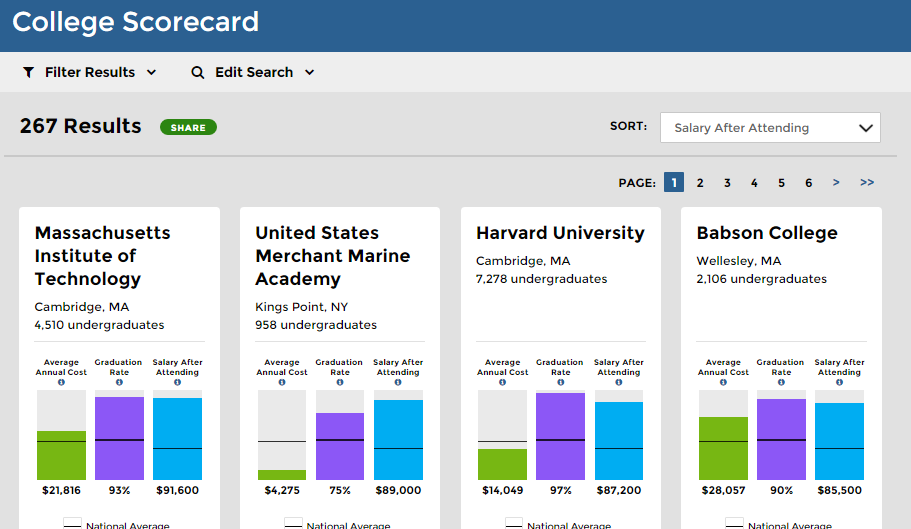
**Tim Yang**

**11-19-2015**

**DC-DAT 9**

**College “Under-Matching” and Data Science**

**Overview**

In September 2015, President Barack Obama launched a consumer-ready College Scorecard on the White House website. The White House has prioritized higher education access for more Americans, and has scrutinized the for-profit university sector. This tool was designed to educate Americans on which colleges provide the most “bang-for-your-buck”, and led to high graduation and career outcomes. 

One specific aspect of President Obama’s initiative focuses on providing higher quality education and economic mobility to low-income Americans. In addition to financial, academic and social hurdles, low-income high school students also face significant obstacles gaining admission to highly selective universities. The term “Undermatching” was coined to describe the phenomenon of well-qualified, low-income students who do not apply and are not matched with competitive colleges. The objective of this project is to produce a data-driven analysis of higher education costs and outcomes specific to the under $30k household market, to better inform and educate families on college-selecting decisions. More specifically, our analysis will center around:

**Project Question**

*“For low-income families, what university characteristics predict college graduation?”*

**Dataset**

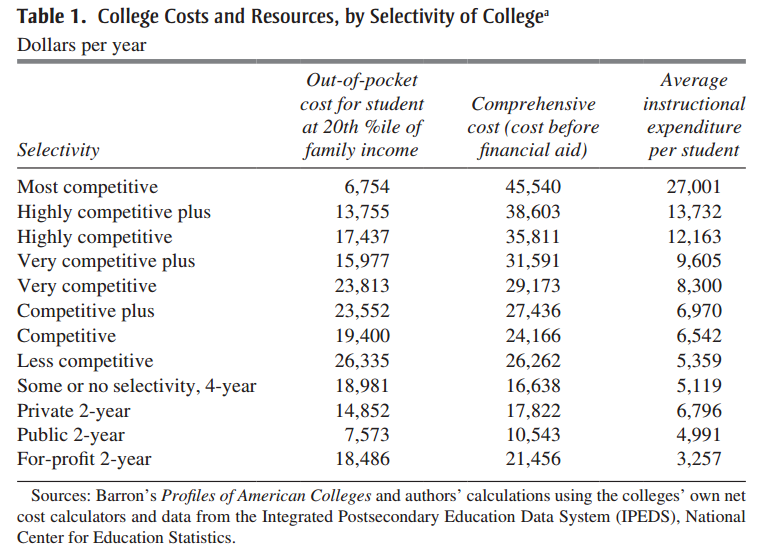
The dataset was extracted from the US Department of Education ScoreCard Data at <https://collegescorecard.ed.gov/data/>. The data was retrieved and aggregated from raw data sourced from the National Center for Education Statistics - Institute of Education Sciences (<https://nces.ed.gov/ipeds/datacenter/>), the National Student Loan Data System, and the Federal Student Loan Office. The White House college scorecard dataset contains institutional information on: School Background, student demographics and income, net tuition cost and expenses, completion, transfer and retention rate, financial aid awards, repayment trends and lifetime earnings.

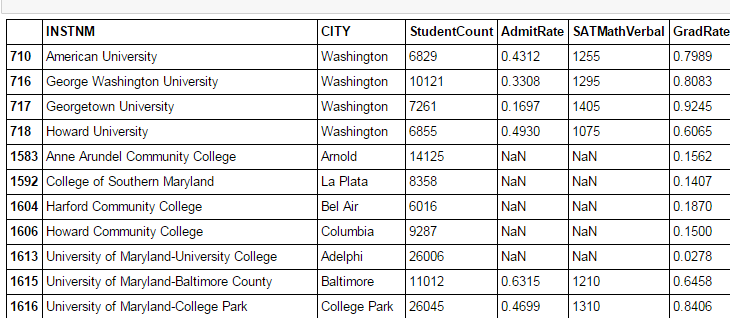
**Data Exploration**

For data exploration, I created a subset of colleges in Maryland, Virginia, and Washington DC to determine which feature variables should be included in the model construction. First, I retrieved the Data Dictionary provided by the White House and grouped the 1954 variables into categories. Then, I selected key variables relevant to low-income households such as “Average Total Cost for Families Net Income < $30,000”. I used mean, median and standard deviation calculations to observe the difference in price and graduation rates by private/non-profit school type, admission rate, and school size.

An analysis of the university information clear shows several distinct categories of schools: most-competitive private, highly competitive private/public, flagship competitive public, and less competitive regional public/for-profit universities.

**I NEED TO GENERATE MY OWN VERSION OF THIS DATA:**

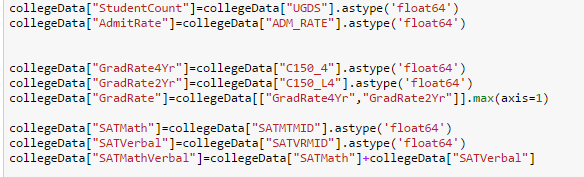
****

****

Initial data preparation has shown that graduation rates are clearly linked to student selectivity, and families must be informed of the difference in student outcomes between a highly selective 4-year university and a not-selective for-profit or regional option.

**Data Preprocessing**

The dataset contained variables for each type of school (public, for-profit, private non-profit) and NULL values for any variables that did not apply to the selected institution. To clean and cast my data, I used the Read\_CSV function and na\_values=”NULL” option. The data read in a strings, and were cast to floats by the following code:



I created model features that aggregated variables across different school types. For example, the Graduation Rates for 4 year Public, Private, and For-profit were stored in separate variables. To combine them into one variable, I used the max function for the three float columns.

5. Model creation - which models did you use why? What were your results? How does this compare to the null model? Which parameters did you change (e.g. k for knn)? did this improve your model?

IN PROGRESS: Private/Public/ForProfit, 2Year/4Year, Admission Rate, SAT Median Score,

**Feature Variables:**

Student Demographics: Student Population, Admission Rate, SAT Scores, % Pell Grant

Tuition Cost: Financial Aid Awards, In State Tuition, Out of State Tuition, Net Cost for Households<$30K

Faculty Funding: Average Faculty Salary, % Faculty Full Time, Tuition Received/FTE,

**Outcome Variable**:

4- Year Completion Rate

Median Income after 6 years since University Admission

6. What is your final model or final insight? What features did you use?

The final model will attempt to create a “True Cost” of college which factors in graduation rates, job placement, and gives a net cost of each school to help low-income families determine school choice.

Similar to “Invoice Price”, “Sticker Price” and true cost-to-own calculations for the auto industry, my model will better inform low-income users to access better educational products in the higher education market.