US Department of Education College Scorecard Data Analysis Research Project

1. Project Scenario

Generally during a student’s junior year of high school, it is usually recommended that parents and students start to do preliminary research about the college admissions process, which includes determining which higher education institutions the student would most like to attend. This preliminary research is often confusing and burdensome for students and parents because there are so many post-secondary schools to choose from, and they all have many differences that need to be taken into consideration, which can create an overwhelming situation. This overwhelming situation is compounded when students have parents that did not attend college themselves, and are also new to researching which schools to potentially attend. There are numerous documented benefits to attending an undergraduate that is a good personal fit, so choosing the right one is a very important decision.

In light of the commonly found overwhelming situation stated above, the Prince George's County Board of Education has asked our data team (Team TAZY) to prepare a presentation for junior year high school students and their parents in order to help them in their decision on what colleges or universities they should consider attending, taking into consideration what would be considered a “typical” Prince George’s County high school student.

1. Research Question Formation  
    Which universities or colleges should “typical” junior year high school students from Prince George’s County public schools consider attending, and what institutional characteristics should they take into consideration when choosing?
2. Overview of the College Scorecard Dataset

The dataset that our group has chosen to use for our team project is the US Department of Education’s College Scorecard dataset, which is freely accessible online from the department’s Scorecard website (https://collegescorecard.ed.gov/data/). Specifically, this data contains records on all Title IV aid receiving higher education institutions in the US, as indicated by the Integrated Postsecondary Education Data System (IPEDS) federal reporting surveys. The purpose of the US Department of Education providing these data freely online is so that prospective high school students and their parents could have a way to compare US higher education institutions across a wide range of data points dealing with college costs and student success outcomes, in order to allow them to make the most data driven and informed opinion as possible on where potential institutions to attend. The data are compiled by the US Department of Education using federal reporting from institutions, data on federal financial aid, and tax information.

There are several datasets made available for consumers to the US Dept of Education College Scorecard website, but we have chosen to use the “Most recent data” dataset provided on the College Scorecards main website that was linked in the previous paragraph. The exact URL to download our chosen data is (<https://ed-public-download.app.cloud.gov/downloads/Most-Recent-Cohorts-All-Data-Elements.csv>).

1. Method to Answer the Research Question  
    We first assume that higher education institutions are able to be rated according to two main factors, which the vast majority of prospective students take into consideration when trying to select which school to matriculate to.   
     
   Those two main general factors would be:
   1. Financial stability of the current/past students
   2. Academic success of current/past students

We also realize that all prospective students are different in terms of their unique preferences for what characteristics they desire in a potential higher education institution. These characteristics could range from the level of degree desired, the state the school is located in, the locale of the school (rural vs. urban), whether the school is public vs private, the undergraduate enrollment size, whether the school is specifically minority-serving certified or not, whether their test scores fall into the range of average admitted freshman class, whether the average age of the entering student body is similar or not, whether the school has open admissions or not, to name a few.

In order to fully answer our research question of, “Which universities or colleges should 11th grade students from Prince George’s County public schools consider attending, and what characteristics should they take into consideration when choosing?”, we will first analyze the data set to see what data are available and what data cleaning and preparation will be required . After data cleaning and preparation, we will identify what variables are related to student financial stability and student academic success. Once variables from the dataset related to student financial stability and student academic success are identified, we will devise a way to “combine” them in a way that made comparisons between institutions within the dataset possible. We will describe in more detail later regarding how we “combine” the variables related to student financial stability and student academic success.

The next step to help answer the research question will be to create four personas that would seem typical of a current 11th grade Prince George’s County public high school student or parent of a high school student. Within the creation of these hypothetical personas, it will become apparent what unique preferences the prospective student desires in their potential higher education institution (many of those characteristics were listed previously).   
  
 Using both the unique prospective student preferences, the two general ratings/rankings (student financial stability and student academic success), and some output data, we will created a College Selection Model. The model will allow students to view the top ten institutions within the US. This would provide the prospective student with the ten potential institutions that would increase the probability of financial stability and academic success. Each student would be able to conduct research into those specific institutions, and see the requirements to attend these institutions.

a. The student would enter their persona data into the model and the model would generate the pool of colleges that matches their persona, and projected financial stability and academic.

b. If the colleges identified by the persona are different from the top 10 institutions, and the student wishes to attend one of the top ten colleges, a comparison of both outcome would allow the student/parent to identified what areas need improvement in order to qualify for one of the top 10 colleges.

1. Dataset Variables to Consider For Student Financial Stability and Student Academic Success Ratings  
      
    In review of the College Scorecard dataset, we realized there is so much data that it was overwhelming, and that we needed to trim it down. The original dataset contained 1,777 data points/variables (columns) for 7,593 institutions (rows). After carefully scanning all of the columns of data available to us, we chose 173 of those 1,777 data points that could potentially be helpful to us in our research process and reduced the number of institutions to 3,824. Those 173 columns are outlined in an excel file on our GitHub repository below. Since there were so many data points used, it was easier to organize the variables chosen for analysis this way than to list them all in this document. <https://github.com/brooksrelyt/infm600/blob/master/College%20Scorecard%20Analysis%20Project/Columns%20and%20Variables%20Used%20From%20Raw%20Data%20File.xlsx>  
     
    The above excel sheet outlines the column number the variable is found in the original College Scorecard raw data set, the variable name (column header), a description/explanation of the variable, and the source of the data. Also available in our GitHub repository is a data dictionary provided by College Scorecard in order to help understand the dataset.  
   <https://github.com/brooksrelyt/infm600/blob/master/College%20Scorecard%20Analysis%20Project/College_Scorecard_Data_Dictionary.xlsx>

This data dictionary contains much of the same information in the excel above, but also contains all of the expected potential values in the various columns. Some variable values are also numerically coded, so this document provides a way to interpret those codes values. For example, the “HIGHDEG” variable (highest degree awarded at the institution) contains five potential values (0, 1, 2, 3, 4), where 0 = non-degree granting, 1 = predominantly certificate-degree granting, etc…  
<https://github.com/brooksrelyt/infm600/blob/master/College%20Scorecard%20Analysis%20Project/College_Scorecard_Data_Dictionary.xlsx>

After we trimmed the College Scorecard dataset down to 173 columns, we decided to choose which variables best represented indicators of student financial stability and student academic success. The variables that we thought did this representation best out of the variables available are below:  
  
**Student Financial Stability**  
1. The percentage of students who defaulted or failed to meet specified conditions within two (2) year of receiving their loan (CDR2)

2. The percentage of students who defaulted or failed to meet specified conditions within three(3) year of receiving their federal loan(CDR3)

3. Average earnings of students 10 years after graduating (MN\_EARN\_WNE\_P10)

4. Median debt of students who completed their degree (GRAD\_DEBT\_MDN\_SUPP)

5. In state tuition costs and fees (TUITIONFEE\_IN)

6. Out of state tuition costs and fees (TUITIONFEE\_OUT)  
  
**Student Academic Success**

1. Retention rate of full-time first students attending four-year or less than four-year institutions (RET\_FT4 and RET\_FTL4) . Combined into one variables for analysis (RET\_FT\_COMBINE).

2. Completion rate for first-time, full-time students attending four-year or less than four-year institutions (C150\_4, C150\_L4). Combined into one (1) variables for analysis (C150\_COMBINE).

3. Proportion of faculty that is full-time (PFTFAC).

4. Percent withdrawn from original institution within 3 years (WDRAW\_ORIG\_YR3\_RT)

**Correlation between variables**

To confirm that we had chosen the best indicators to calculate financial stability and academic success, we will had to determine if there were any relationship between the variables. We will use the plot function to verify the relationship between the variables.

1. For the financial stability variables we will plot

1. Tuition vs Default Rate
2. Graduate Debt vs Default Rate
3. Out-of-State Tuition vs In-State Tuition

2. For the academic success variables we will plot

1. Student Retention Rate vs Student Completion Rate
2. Student Completion Rate vs Withdrawal Rate
3. Calculation of the above “Super Score” Ratings

After we have chosen our variables to describe our general ratings above (we will now refer to these ratings as “Super Scores”), we need to be able to use the variables to form comprehensive ratings which take each respective variable into equal consideration. The approach we decided to use is to calculate the Z-Score for each variable for each institution. The Z-Score will tell us how many standard deviations above or below the mean an institution’s observed variable value is in relation to all of the schools in the dataset. Then, we add those Z-Scores together for each Super Score variable group to form the overall rating.

Each institution’s Z-Score for each variable is calculated as follows:  
Z-Score = (Institution variable observed value - variables mean) /

variable standard deviation  
  
 There is one problem with simply adding the various Super Score variable Z-Scores together to form an overall rating, though. For some variables a negative Z-Score is better, because you want the variable value to be as low as possible (for example, default rate). For other variables, a high positive Z-Score is better (for example, mean yearly earnings ten years after graduation). In order to add the Z-Scores together in a way that would allow a “better” rating to alway be greater (positive), we multiply any Z-Score by - 1 that falls into the negative = better category, so that a positive rating is always better, and allows the addition of Z-Scores to make sense.

1. Reduce Dataset to Only Include Schools Appropropriate For Analysis  
    We decided that while scanning our dataset, that some institutions needed to be removed in order for the analysis to be as relevant to our audience as possible and to be consistent. In light of this, we decided to only keep schools (rows) in our dataset that meet the following dataset criteria:
   1. Highest degree awarded as Certificate Degree, Associate's Degree, Bachelor's Degree, or Graduate Degree
   2. Institutions currently operating
   3. Institutions not on heightened cash monitoring status by the Department of Education
   4. Institutions not exclusively Graduate/Professional or Unreported/Undefined
2. Initial Dataset Findings
   1. The institutions that rank highest according to our Super Score data ratings
   2. The mean and standard deviation for several variables that prospective student would most likely be interested in (in-state tuition costs, retention rates, completion rates, average admitted SAT and ACT test scores).
   3. Discuss histograms graphs for any of the variables mentioned above
   4. Discuss correlation between any dependent variables that we find and   
      Potential dependent variables could be:  
      -Income Level

-Whether they will seek Federal Loan

-Percentile of SAT

-Percentile of ACT

-Type of degree to be attained (Associate or Bachelor)

-State

-Locale (City, Urban, Town, Rural)

-School Preference (Public or Private)

* 1. Other findings?

1. Matching personas with recommended Institutions based on their unique preferences and identifying their financial stability and academic success