College Admission Analysis using Predictive Analytics

Table of Contents

Introduction

Higher education is now a massive industry, with 20 million American students enrolled in colleges. 77% of colleges and universities spend over 100,000 per annum on brand strategy work and among those about a third spend over $200,000 per year. This has become a growing trend in higher education. Although most colleges are non-profit but still they employ marketing professionals to reach out to and engage the millions of American college applicants per year. It was released on Sept of 2015. This dataset has been widely used in various news analysis and academic researched. As I was going over those articles, I became interested to see what type of information I can gather from it which could help High School age students looking for colleges and admission.

Business Problem / Questions

It is a trend that almost all colleges employ predictive analytics and data analysis as an integral part of the admissions process over the past few years. Using tis college scorecard data I am trying to answer the following questions:

- Based on the data on test scores, cost to attend, size of school and salary which school offers the best education
- Are there certain schools or states where student debt and loan default rates are high compared to earnings?
- Considering diversity, is there a good ratio of male/female, ethnic and racial mix?
- What can we find about the income distribution patterns of the students getting admission?

DataSet

The Dataset was downloaded from the [U.S. Department Of Education][1] website. The [College Scorecard][1] was created in 2013 under President Barack Obama's administration to make data about colleges more accessible to consumers in a centralized, interactive tool

The dataset consists of Instituition-level data files for 1996-97 through 2019-20 containing aggregate data for eac institution. It includes information on institutional characteristics, enrollment, student-aid, cost and student outcomes.

> **Files used**
>
> - CollegeScorecardDataDictionary.xlsx: This excel sheet contains the data dictionary of the various files supplied. I looked at the FieldOfStudy_Data_Dictionary tab to glean through all the available features and chose my required features. The data element descriptions for the different Variable Names used within the data is described in this file. I found the following 4 columns very useful: * Name of Data Element * Variable Name * Value * Label
> - Most-Recent-Cohorts-Field-of-Study.csv Data files with data about specific fields of study within institutions.
> - Most-Recent-Cohorts-Institution.csv Data files with data about institutions as a whole. College Scorecard data files for data elements calculated at the institution-level.

Methods

> **Data Set identification**

I am using Jupyterlab and using pandas to view the various datafiles that I ave downloaded from the website. After looking at the files I will decide wich attributes to use in my analysis.

> **Reading the available articles**

This is a nice article that defines the usefulness of tis dataset [snews]. I went through a lot of online articles about this topic and it fascinates me as much I can understand about the college admission process and the internal working of it. Needless to say that do have a daugter who is a Sophomore in high school. So this project is very close to my heart.

> **Tools used for this analysis**
>
> - Tableau for visualization I am planning to use tableau to visualize the data and also for Exploratory data analysis.
> - Jupyterlab for analysis

Ethical Consideration

College Scorecard still contains flaws that undermine its overall effectiveness. Among the concerns is the Scorecard’s focus on limited variables, particularly monetary measures of value. Reducing the measure of an institution’s worth to two or three numeric factors does not present a full or accurate representation of an institution’s mission and character. The College Scorecard must continue to evolve if it is to be helpful to families, and fair to all institutions, particularly those that serve a high proportion of low-income students. The Scorecard also faced criticism in 2017, when it was revealed that it had been publishing inaccurate loan repayment rates for most colleges. Blaming the blunder on a “coding error,” the Department was quick to fix the mistakes.

Challenges / Issues

- Data quality issues e.g., some sectors had high shares of students with an unknown completion status
- Data suppression in order to protect student privacy
- Data hidden and not available for analysis

- The missing elements were as a result of no compulsory or mandatory clause for reporting and as a result inaccuracy of the report cannot be ruled out.

References / Acknowledgement

Data Sources used

1. https://collegescorecard.ed.gov/data/
2. https://www.usnews.com/education/best-colleges/articles/how-us-news-calculated-the-rankings
3. https://www.usnews.com/education/best-colleges/paying-for-college/articles/how-students-should-use-new-college-scorecard-data
4. https://community.tibco.com/wiki/college-scorecard-analysis
5. https://www.kaggle.com/datasets/kaggle/college-scorecard
6. https://deepnote.com/@omar-khan/Analysis-of-COLLEGESCORECARD-dataset-VG7JYdzUTDOOFv003pRAVw