College Recommendation Engine

Amie Davis

Bellevue University

Spring 2021

https://github.com/amodavis

Abstract

This project is catered to future and current college students and their parents. As a parent of a prospective student, I noticed a lot of data in disparate systems with inconsistent results. There are many resources to the point that it is overwhelming.

The intent of this project is to demonstrate a use case in unsupervised learning. In it, you will see the problem defined and business objective set. I will guide you through my research and methodologies, resulting in a usable recommendation engine.

College Recommendation Engine

One of the greatest challenges faced by high school students and their parents is selecting a college. This is perhaps the first major decision in a young person's life. It can also be the most costly. Like any big decision, it is imperative to perform research and not come to a hasty resolution.

There is a cornucopia of available information on colleges and universities in the Unites States. There are many search engines available online. For example, PrepScholar (n.d.) is a great resource to find out what your chances are at getting accepted into a school. Just enter your standardized test score and grade point average, and voila, your likelihood of acceptance is presented. However, just because you can be accepted into a school, doesn't mean it is the best school for you. Many sites either prompt you for too little or too much information. Demographic information on colleges may be easy to obtain, but it should not be the focus of your college selection.

The College Board provides a student search engine that has a lot of parameters ("Finding Your College Fit", n.d.). Unfortunately, registering for this search engine, or even the SAT, will place you on a lot of mailing lists. Students and parents are bombarded with emails and mailings, which often end up getting tossed and deleted. This excessive information builds both anticipation and anxiety, which leaves everyone overwhelmed.

So, how do you find a college that fits your student? A good source for students to get a peek into academic programs and student life is Fiske's Guide to Colleges (Fiske, 2019). It includes both pros and cons, as well as student interviews. However, after polling dozens who have recently gone through the college selection process, the best method is to visit a campus, meet with a teacher and student, learn about their programs, and get a "feel" for the school. Wow, that's time consuming, and it limits you to a handful of schools that you can see.

What about the thousands of other schools? I propose that if you find one school you really like, there should be a list of similar schools available. This will give you just a handful of colleges to investigate.

Method

Business Problem

Searching for a college is too time consuming and often only includes limited features. A college visit or interview is the only good way to find a college that "fits," but it is impossible to visit all college campuses.

Business Objective: Make it easier to find a college that "fits."

Data Sources

Scorecard data from the US Department of Education ("College Scorecard," n.d.) were utilized for the school year 2018-19. Consolidating different years was attempted, but there was a disconnect between fields for varying years. I avoided the 2020 data, since they were impacted by the pandemic. This dataset includes assessments of colleges to decipher the value of education, such as costs and graduation rates, as well as aggregated demographic information on students, to include standardized test scores. The data is mostly numerical. A detailed data dictionary can be found on the website to decipher the codes.

Exploratory Data Analysis

R was used for most of the data preparation, such as data cleansing and feature reduction. It was the ideal tool for this dataset to allow for numeric analysis of the over 2000 columns. R allowed me to easily toggle between data cleansing and analysis, so I could assess different aspects of the data.

I performed featured reduction using several methods. All columns without values, either all null or all zero, were removed. This left me with just under 1000 features to analyze. Highly correlated features were removed, such as values that attributed to rate calculations. There were separate features for numerators and denominators, as well as percentage and rate values. I retained the percentages and removed the contributing attributes. I also removed unnecessary text fields, such as website addresses. Features missing more than 50% of values were considered and discounted.

With R, I was easily able to analyze feature variance and remove those features having zero or near-zero variance, limiting features to analyze further down to 360, a much more manageable level. Analyzing the remaining dataset enabled me to assess correlation across the entire dataframe. Highly correlated features, those with a Pearson's coefficient of .7 or greater, were removed to avoid redundancy. Null values and outliers were identified and handled on a case-by-case basis.

Modeling Objective

The ultimate question that remains is, which college should we look at. Which colleges would be the best "fit?"

Modeling Objective: Perform unsupervised learning to form school clusters.

I used a k-Means algorithm in Python to create a clustering model to form groups of colleges. The model was trained on the data using multiple values of k. Each value of k was plotted against the computed Sum of Squared Error (SSE) values. I used the elbow method to determine the best value for k. In this case, the curve "elbowed" around k=30.

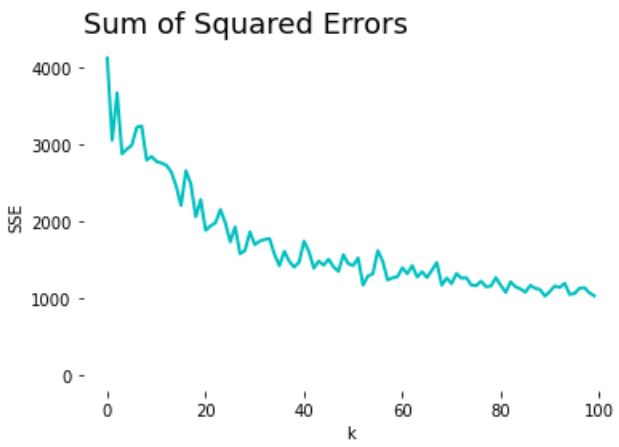

Figure 1 - k-Means Sum of Squared Errors

Model Deployment

In order to deploy the recommendation engine, it was necessary to assign and store designated cluster labels for each school. I created a function to accept a school and return the model's predicted cluster label. That label was then stored into the dataframe, alongside school information.

I used Python to create a small application to prompt users for their preferred school. Unique school identification numbers are required, due to duplicative school names. These unit IDs are available at the College Scorecard website ("College Scorecard," n.d.). Once the user enters a unit ID of a school in which they have already researched and determined will be a good fit, the recommendation engine provides a list of other schools that might also be worth investigating.

Results

Exploratory Data Analysis

After reviewing the initial visualizations of the data, there were some surprises to my education subject matter expert (SME). She was surprised that the admission rate for Vanderbilt was not as stringent as Ivy League schools. There were also a couple of large universities that accepted over 50% of

applicants, specifically Syracuse University and the University of Georgia. With respect to locations, she was curious by the Brigham Young University campus in Hawaii, as well as the vast number of colleges in Puerto Rico. I was intrigued that about 75% of schools have the same in-state and out-of-state tuition.

The dataset contains a lot of financial data. There is information on federal loans, payback rates, and defaults. It also has some aggregated demographic information. The data are sorted into nine categories: academics, admissions, aid, completion, cost, earnings, repayment, school, and student. There were over 2000 features in the dataset, which I was excited about initially. However, that was too much to review to really get to know the data.

My biggest challenge was weeding the number of features down to a manageable level. With over 2000 features, it was difficult to perform validation on the data. Often the correlation reduction process removed rates but kept the correlated numerator and or denominator fields. Clearly all three are correlated, but I wanted to retain the rates field. There were too many of these cases to remove manually, so it was easier just to add them back. This process was very time-consuming.

Research Question Visualizations

PowerBI visualizations were used to review features in which parents and students expressed interest. To provide useful analysis of the business problem, I surveyed local parents and students to find out what their biggest challenges were. Top parental responses were being overwhelmed and tuition concerns. Top student responses were on academic programs and majors. This was a very small sample but gave me ideas for research questions and highlights the differences in how students and parents approach their search.

I also paid particular attention to finances. An article in College Choice (Staff Writers, 2020) addresses the financial aspect, which is most often a concern for adults. Are you getting your money's worth? It indicates that, in addition to the tuition paid, you should consider the money to be earned.

- Given ACT/SAT score, to which schools am I likely to be admitted?

  It may be best to answer this by looking at how difficult it is to get into a school. 1% of schools, labeled as "REACH" schools in the graphic below, accept less than 10% of applicants, whereas over 50% of schools, labeled as "SAFETY" schools below, accept over 70% of applicants.

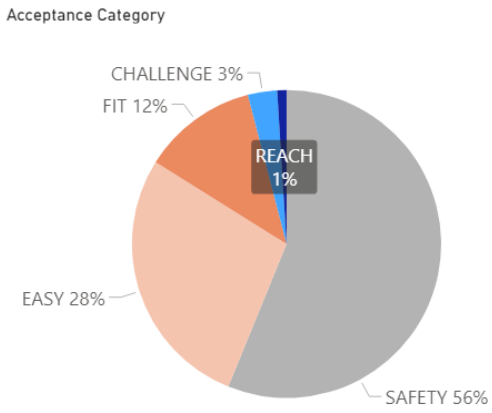

Figure 2 - Breakdown of Schools by Acceptance Category

- Are my admission chances better if ACT or SAT scores are submitted?

  As you can see below, there is a direct correlation between the ACT and SAT scores, so it does not matter which score is submitted. However, if you score better on one test than the other, that is the better score to submit.

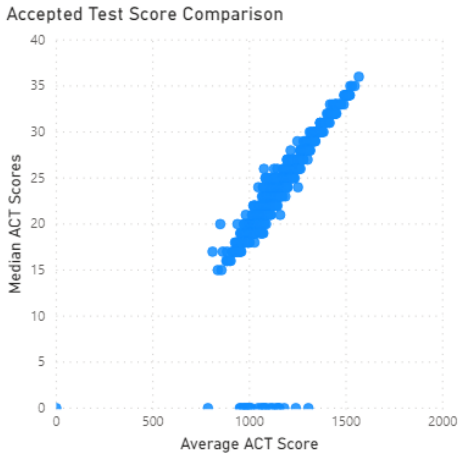

Figure 3 - Comparison of Accepted ACT and SAT Scores

- Which schools offer what fields of study?

  I created an interactive tool in PowerBI to display a list of schools after selecting specific fields of study.

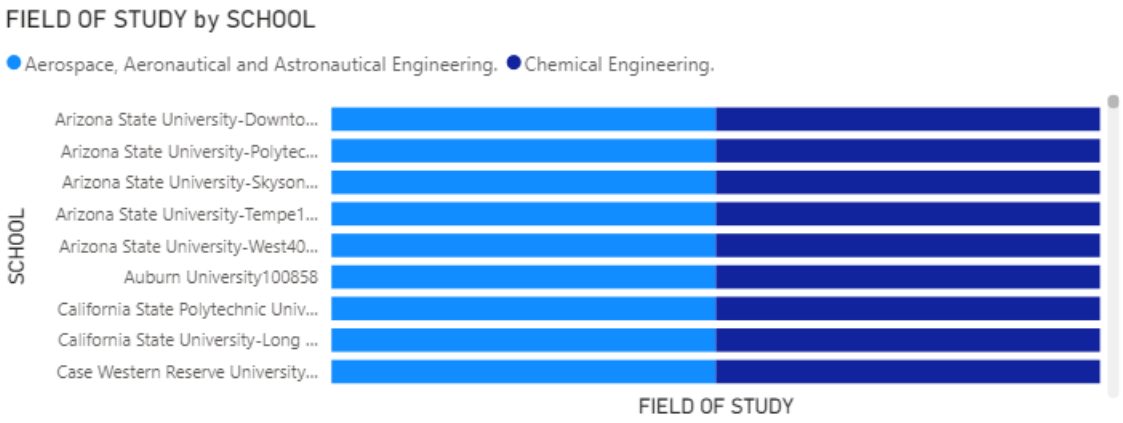

Figure 4 - Interactive Field of Study Tool

- Do the larger universities offer more degree options?

  With the exception of a few universities, this holds true.

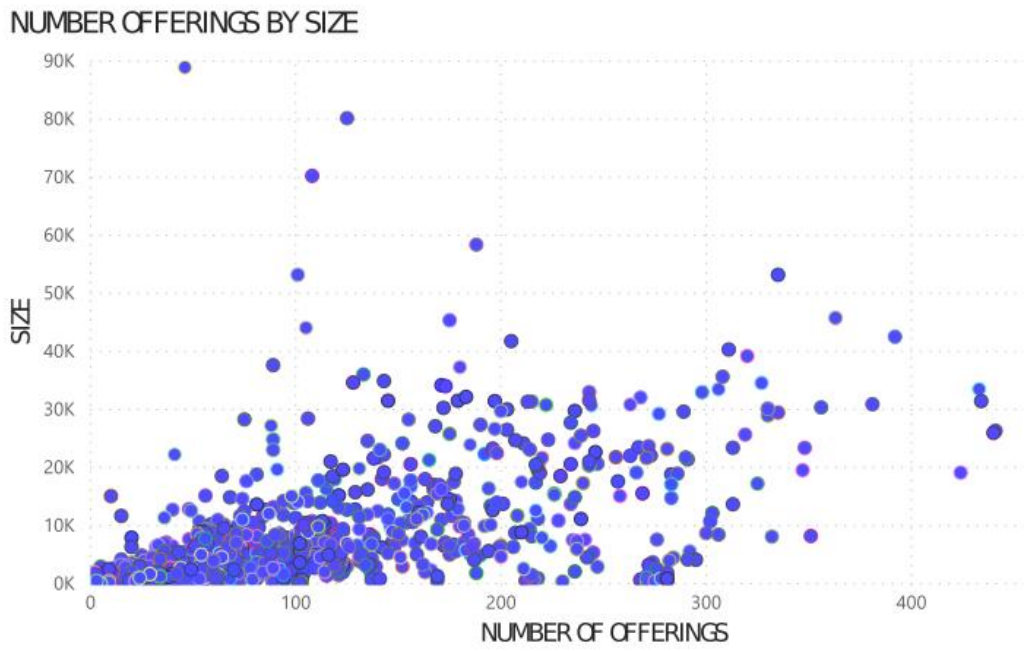

Figure 5 - Amount of Degree Offerings by School Size

- Which schools can I afford?

  Be prepared to pay over $45,000 per year if you attend a REACH school. A FIT school may be a better bargain.

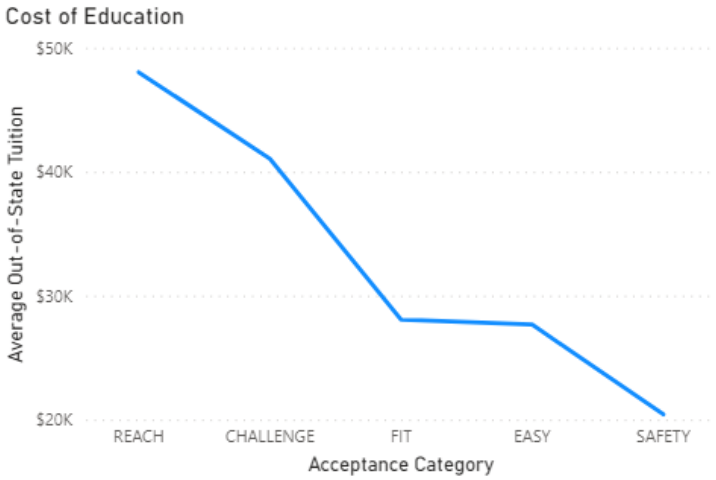

Figure 6 - Tuition Based on Acceptance Rates

- Which out-of-state colleges offer in-state comparable tuition?

  Surprisingly, almost 75% of colleges offer the same tuition for out-of-state students that they offer for in-state students.

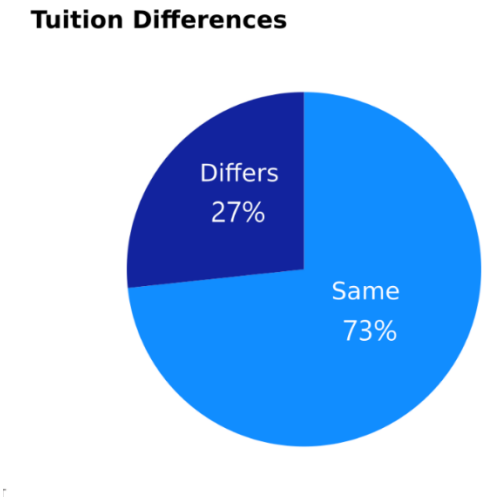

Figure 7 - Most Schools Don't Charge More Out-of-State

- What is the likelihood I can earn enough after I graduate to pay back my student loan?

Almost 50% of undergraduates from a handful of schools were able to pay back their loans in full just two years after graduating and even more had a 30% payoff rate. It is assumed that graduates from those schools earn even to pay back their loans.

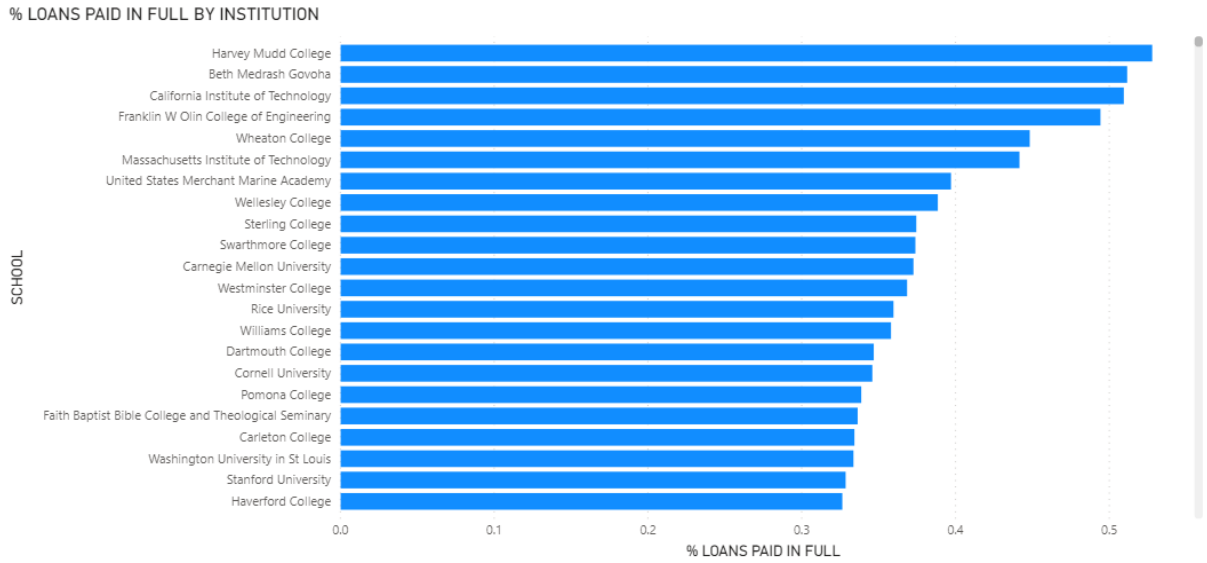

Figure 8 - Schools with High Likelihood of Loan Payback

- Is the tuition related to the size of the university?

  There appears to be no correlation. Small schools can cost just as much as large schools.

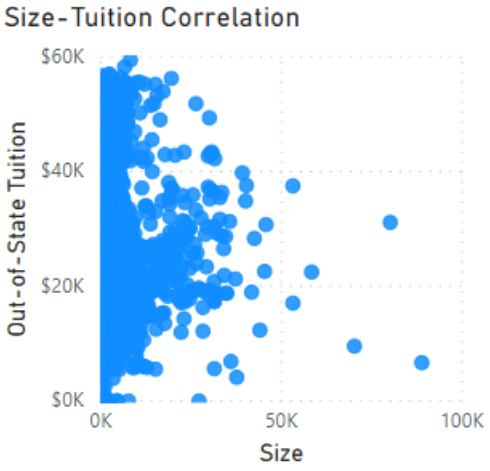

Figure 9 - Tuition Not Based on School Size

Model

The resulting model returns a list of recommended school, given another school of interest.

```
What is the school id of a college which interests you? 214777
Since you like Pennsylvania State University-Main Campus, then you might also
be interested in these other schools.
[20]: 1308 Pennsylvania State University-Penn State Erie-⋯
1309 Pennsylvania State University-Penn State New K⋯
1310 Pennsylvania State University-Penn State Shenango
1311 Pennsylvania State University-Penn State Wilke⋯
1312 Pennsylvania State University-Penn State Scranton
1313 Pennsylvania State University-Penn State Lehig⋯
1314 Pennsylvania State University-Penn State Altoona
1315 Pennsylvania State University-Penn State Beaver
1316 Pennsylvania State University-Penn State Berks
1317 Pennsylvania State University-Penn State Harri⋯
1318 Pennsylvania State University-Penn State Brand⋯
1319 Pennsylvania State University-Penn State Fayet⋯
1320 Pennsylvania State University-Penn State Hazleton
1321 Pennsylvania State University-Main Campus
1322 Pennsylvania State University-Penn State Great⋯
1323 Pennsylvania State University-Penn State Mont ⋯
1324 Pennsylvania State University-Penn State Abington
1325 Pennsylvania State University-Penn State Schuy⋯
1326 Pennsylvania State University-Penn State York
1770 Arizona State University-West
1779 Arizona State University-Polytechnic
1856 Arizona State University-Downtown Phoenix
1944 Pennsylvania State University-World Campus
1984 Arizona State University-Skysong
```

Figure 10 - Sample Recommendations

Discussion

Limitations

There may be data skewed due to the impact the pandemic has had on learning experiences at colleges. I avoided using data from the 2019-2020 school year. There was also inconsistency in formatting for various years. Many values in the older data are left as null.

The model is limited by the feature reduction process. Some near-zero variance features may have skewed the cluster groupings. Also, it is a known limitation that K-Means assumes that data groups in spherical clusters.

Future Recommendations

To avoid problems with availability, I steered away from the Scorecard Application Program Interface (API). Rather, I downloaded the dataset. If this project is deployed for others to use in the future, I would recommend the API, so the model always uses the latest dataset.

Conclusion

Being a current student and the parent of both a current student and a prospective student, I have experienced this business problem first-hand. With so many resources available for future college students and their parents, it is overwhelming. A recommendation engine such as this will assist in the college selection process. I have already shared the interactive PowerBI tools and recommendation results to fellow parents and found even a small project such as this to be quite practical.

References

College Scorecard. (n.d.). Retrieved March 15, 2021, from https://collegescorecard.ed.gov/

Finding Your College Fit. (n.d.). Retrieved March 16, 2021, from https://bigfuture.collegeboard.org/find-colleges/how-find-your-college-fit

Fiske, E. B. (2019). *Fiske Guide to Colleges 2020*. Naperville, IL: Sourcebooks.

PrepScholar. (n.d.). Retrieved March 15, 2021, from https://www.prepscholar.com/

Staff Writers. (2020, June 04). What makes a college a good value? Retrieved March 16, 2021, from https://www.collegechoice.net/what-makes-a-college-a-good-value/

Appendix A

Q&A

Q: What factors were considered when making recommendations?
A: Factors included: location, size, completion rate, acceptance rate, several rates, various loan amounts and repayment rates.

Q: Why are there so many schools with similar names on my recommendation list?
A: Scorecard data tracks each campus separately.

Q: Did you find commonalities within clusters?
A: The k-Means algorithm takes all factors into account and converts them into vectors.

Q: Are the college groupings named?
A: The algorithm labels each cluster with a number. I chose not to name them further.

Q: Does your program handle feedback? For example, was this a good recommendation?
A: Not at this time, but that is a great future recommendation.

Q: What is a unit id and how do I find it?
A: School unit IDs can be found at the College Scorecard website (https://collegescorecard.ed.gov/data).

Q: How do I decide which initial school to use?
A: I recommend starting with the College Board's search engine (https://bigfuture.collegeboard.org), investigate school websites, and plan a campus visit.

Q: How did you handle null values?
A: It depends. Most were set to zero. Median debt values were set to the median of other median values. Private rate values were combined with public rate values.

Q: k-Means does not work well if your classes are imbalanced. Did you experience any imbalanced classes?
A: The largest imbalanced class was a categorical feature denoting online-only schools.

Q: What criteria was used to designate a school as a "reach" school?
A: This can depend on the student. For this study, a "reach" school was defined as one that accepted less than 10% of the students who applied.