What College is Best for Me?

CSC 495: Data Driven Decision Making

Sal Camassa
Wyatt Maxey

CSC 591-006: Data Driven Decision Making

Aman Chauhan
Parth Nagori
Vidhyalakshimi Sreenivasan

North Carolina State University
Department of Computer Science

May 1, 2018

Outline

Our group was challenged with creating an application to make the overwhelming search for the right college a more straightforward process. The solution we chose to develop focuses on answering one question: which college is the best fit for me? By answering this question, the application is able to present users with a list of 10 schools that match the criteria of being a best fit for them. This allows users to narrow the focus of their college search and therefore feel less overwhelmed.

The methodology used in developing the solution can be broken into five parts.

1. The first is gathering pertinent information. To accomplish this, our team turned to a handful of data sources providing both qualitative and quantitative data. The bulk of our quantitative data comes from College Scorecard which houses data collected by the United States Department of Education. We collected temperature data from NCDC Climate Data Online. We obtained crime rate data from US Department of Education Campus Safety and Security website. For assessing nightlife, we used the Google Places API to gather information about clubs near the colleges.
2. After gathering all pertinent information on universities, the data needed to be sifted through and cleaned up. Attributes from one source needed to be matched with those from other sources when describing the same entity. Entities with missing attributes needed to be accounted for. Unstructured, qualitative data had to be analyzed and transformed into something math could be performed on in order to allow for computation.
3. With university data in hand and ready for computational operating, some metrics that varied by user needed to be accounted for. Our approach to this was to create a survey that would gather information about the user that we could turn into metrics needed for our computation. We focused on creating a survey which asked questions in a more personal, rather than logical manner. This mindset lead us to reframe certain questions such as “what range of acceptance rate are you looking for?” to “Academically, I would prefer …” with answer choices ranging from “To be rigorously challenged” to “Not touch a book, I’m here to party!”. Asking questions with this style presented the user with a question they are more comfortable answering while allowing us to collect the data we needed.
4. Once the computation has been performed, the results need to be presented to the user. Results are sent to a web interface from the backend of the application as a list of schools. This information is presented to the user on a web page in both a table and a map. The table shows a rank if it has been sorted, the school rame, city the school is in, and the state. The map shows each university as a node as well as the user’s current location.
5. There is a sort button below the table that opens a dialog when clicked asking the user for an importance rating on a scale of one to five, with five being the most, of the weather, crime, and nightlife of the area the potential school would be located in. These importance ratings are submitted to the backend, where a TOPSIS ranking is performed. The metrics for weather are based on the user’s ideal temperature for each of the four seasons submitted as part of the survey versus the actual averages for each area the schools are located. The crime metric is the crime rate of the corresponding zip code of the school’s location. The nightlife metric is calculated by querying the Google places API with the latitude and longitude of the school for all the bars within two and a half kilometres and gathering both the total count of the results as well as the average Google user rating out of five of all of the bars.

Table of Contents

Outline **2**

Table of Contents **3**

Question **4**

- Problem Statement 4
- Problems/Opportunities 4
- Stakeholders 4
- Decision Influencing Facts 4

Description **5**

- Problem Statement Breakdown 5
- Which universities can student afford? 5
- What is the student's profile? 5
- Is the Student likely to be accepted? 5
- What is the school's profile? 6
- Insights from Data 6
  - Top 10 Nearest Neighbors - Location Data 6
  - Top 10 Nearest Neighbors - Weather Only 7
  - Nearest Neighbor - College Scorecard 7

Methodology **8**

Data Cleaning **9**

- University Data 9
- Crime Rate Data 11
- Weather Data 11
- Nightlife Data 11

Survey Construction **11**

Architecture **12**

Validation **13**

- DBSCAN 13

Repository and Sources **14**

- GitHub Repository 14
- Data Sources 14
- Demo 14

Question

Problem Statement

Display a finite collection of colleges for the stakeholders to choose from based on their selection criteria - student profile, financial state, location.

Problems/Opportunities

1. College search is quite overwhelming as there are a lot of options. We want to make this process less overwhelming for the students.
2. Factoring in student profile in the filtering criteria. Many publicly available applications don't take into account user expectations and preferences.

Stakeholders

Assisting the students and parents in narrowing down the choice of university/college.

Decision Influencing Facts

Category	Facts
School	Public vs Private Main vs Branch Campus
Academics	Programs offered based on degree type
Admissions	Acceptance rate SAT and ACT scores
Budget	Tuition fees TAs/RAs Scholarships offered Part-time opportunities In-state vs Out of state
Repayment	Cohort Default rate
Completion	Completion Rates Retention Rates
Earning	Income of former students Percentage of students recruited
Census	Demographic data by location

Description

Problem Statement Breakdown

Questions for Problem Statement	Attributes
Which universities can student afford?	Cost, Aid, Earnings
What is the student's profile?	Academics, School Environment
Is the student likely to be accepted?	Academics, Admissions
What is the school's profile?	Academics, Earnings, Completion, Cost, Admissions, Rankings, Location, Student Body

Which universities can student afford?

Major Attributes	Finer Details
Cost	Average cost of attendance (Living expenses), Tuition and Fees
Aid	Federal Loans, Federal Grants, Typical Monthly Loan Payment, Cohort Default Rate, Repayment Rate
Earnings	Average earnings, Median Earnings, Earnings of former students

What is the student's profile?

Major Attributes	Finer Details
Academics	SAT Scores, ACT score, Degree desired
School Environment	Student Body size, Diversity, Religious affiliation
Budget	Private/public, In-State/Out-State, Income bracket

Is the Student likely to be accepted?

Major Attributes	Finer Details
Academics	Programs offered, Program rankings
Admissions	SAT Scores, ACT Scores, Acceptance Rate, High school GPA

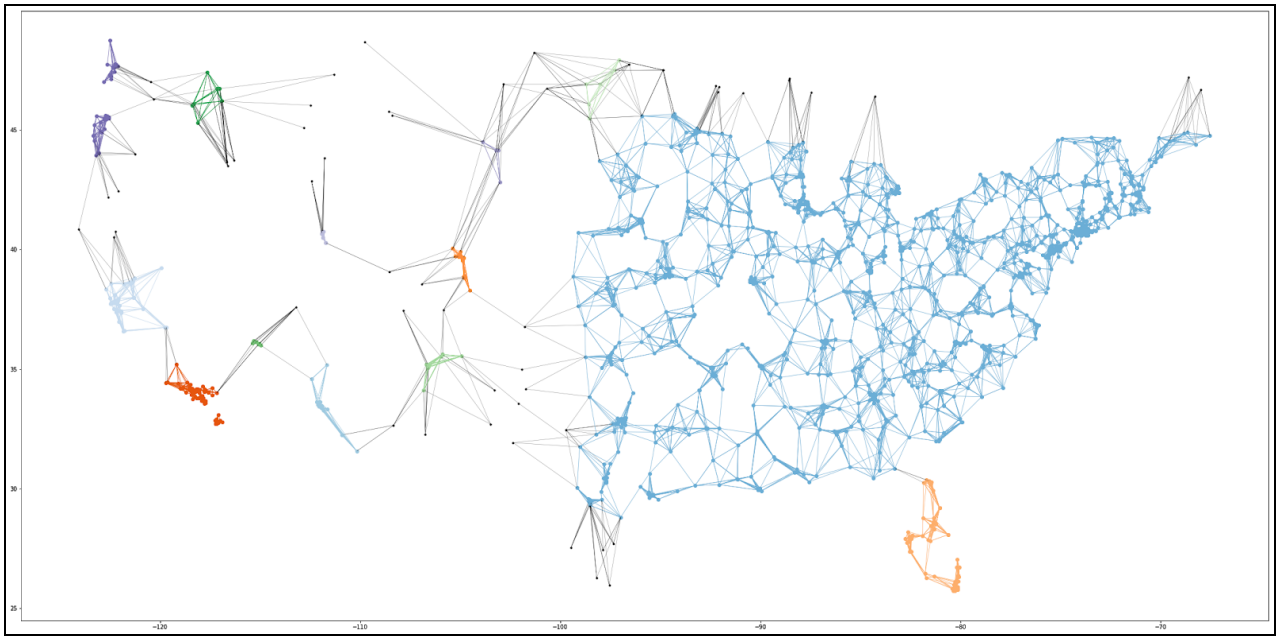
What is the school's profile?

Major Attributes	Finer Details
Academics	Programs offered
Earnings	Average earnings, median earnings
Completion	Completion rates for first time students, completion rates for transfer students
Cost	Cost of attendance, tuition and fees, Average net price
Admissions	Admission rate, Midpoint ACT Score, Midpoint SAT Score
Rankings	Program rankings
Location	Urban/Rural, State, Main/Branch, Online-only
Student Body	Number of undergrad, Demographics, Part time/Full time, % over 25 years old

Insights from Data

Top 10 Nearest Neighbors - Location Data

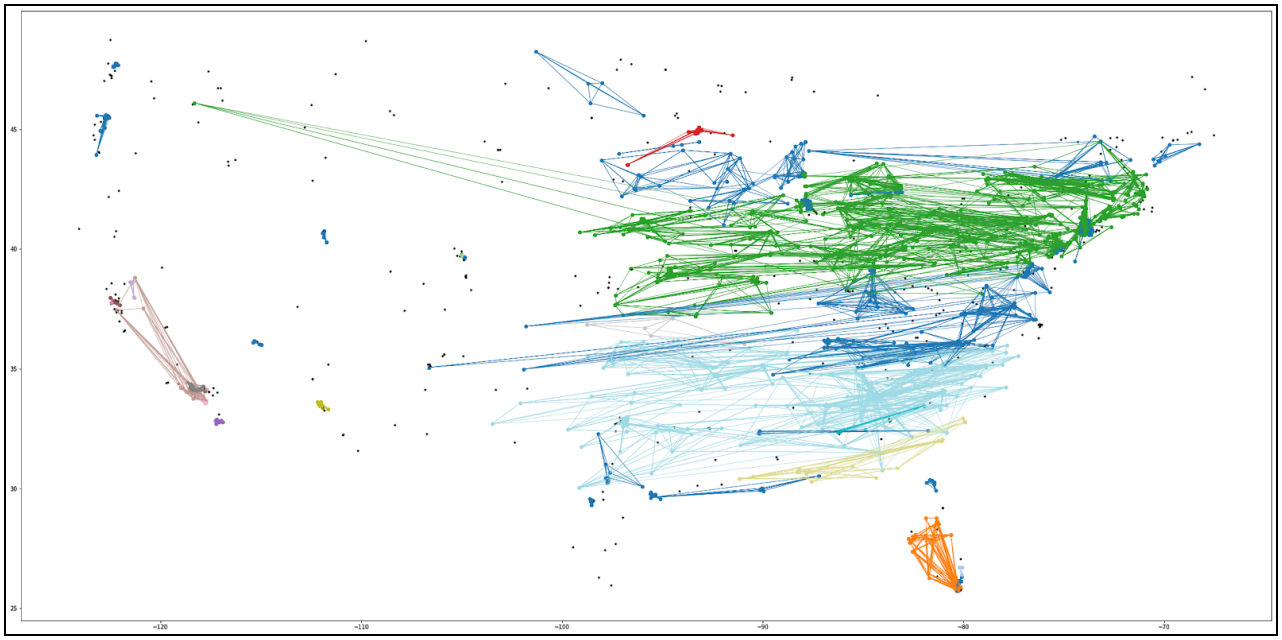
The map below shows the Top 10 Neighboring Colleges for each College in the College Scorecard Dataset based on distance between their geographic coordinates. Zip Codes of each college were mapped to their respective geographic coordinates. Euclidean Distance was used as the metric.



As we can see in the above map, based on location alone, colleges seem to be clustered together in some very well-defined regions. Links and points in the same cluster identified by DBSCAN have the same color. Links between points in two separate clusters are colored black.

Top 10 Nearest Neighbors - Weather Only

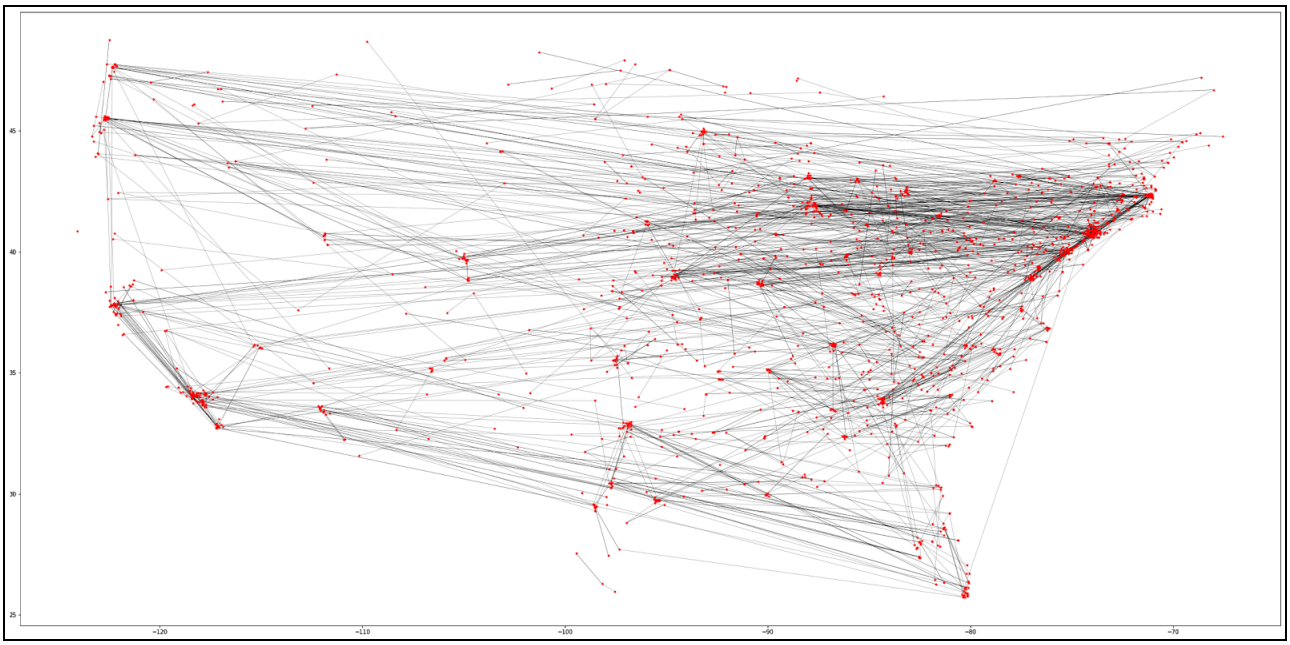
The map below shows the Top 10 Neighboring Colleges for each College in the College Scorecard Dataset based on similarity in temperature. Euclidean Distance was used as the metric.



Clusters were identified by using DBSCAN, and node and links inside each cluster are assigned a specific color. Intercluster links are transparent, in order to highlight the similarity of college location according to the weather in that surrounding region.

Nearest Neighbor - College Scorecard

The map below shows the nearest neighbor to each college if we consider all the attributes of the College Scorecard. Euclidean Distance was used as the metric.

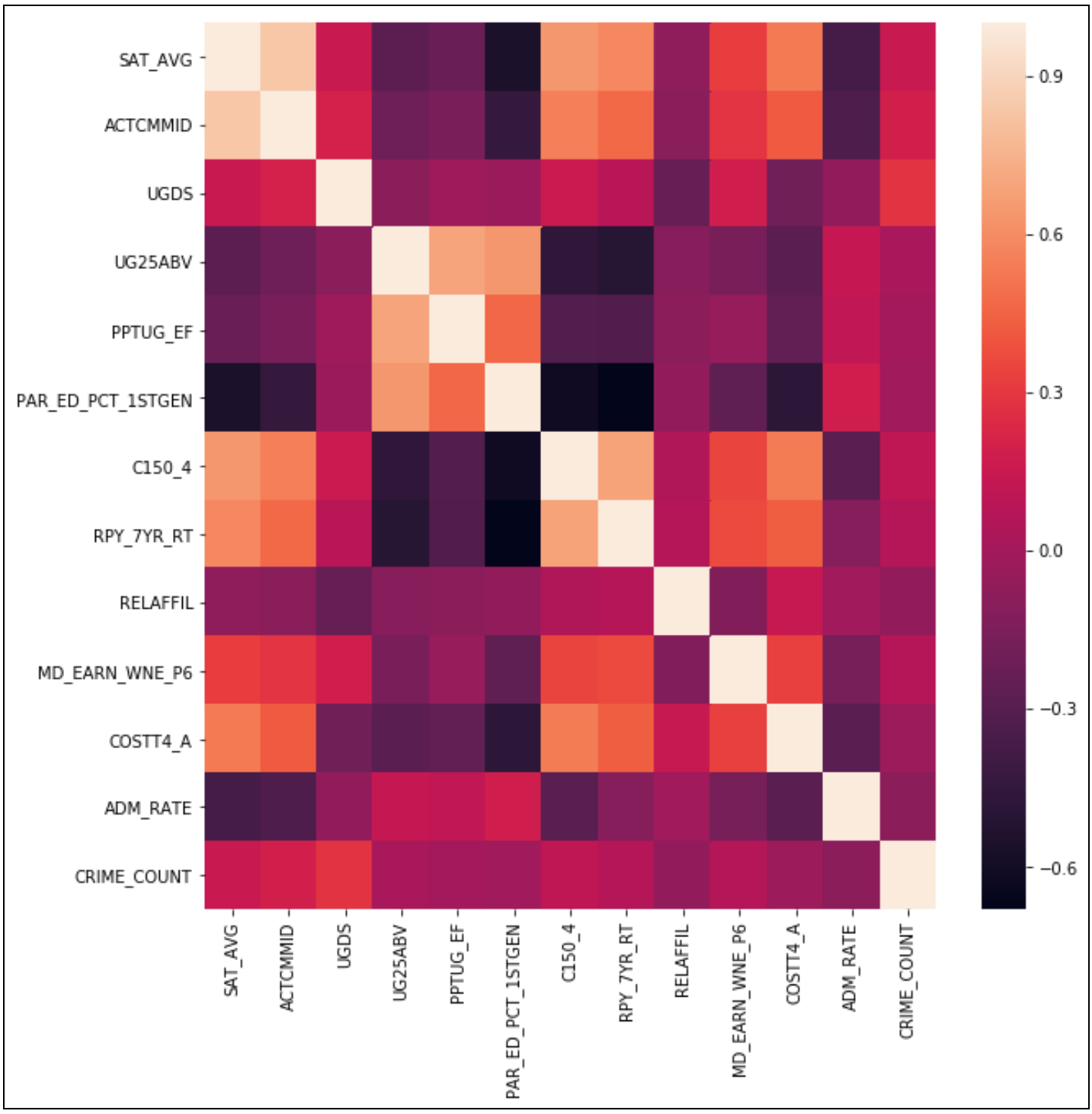


As we can clearly see, due to the high density of colleges on the East Coast, the nearest neighbors on the East Coast are mostly near them.

Methodology

- Gather university data
  - College Scorecard
    - General information
      - Name
      - Location
      - Current operating status
      - Public or private
    - Student body information
      - Size
      - Demographic composition
    - Academic information
      - Acceptance rate
      - Average and/or median ACT/SAT Score
  - Google Places
    - Number of bars around location
    - Average Google User rating of each bar
  - NCDC Climate Data Online for Temperature Stats
  - US Department of Education Campus Safety and Security
- Organize university data
  - Selecting subset of College ScoreCard Data
    - PCA
    - Applying weights to factors
  - Joining temperature and crime rate with main dataset using zipcode.
- Gather User (student) Data through Survey
  - General information
    - Current state
    - Zip code
    - Gender
  - Academic performance
  - Student body preferences
  - Temperature preferences
  - Financial information
- Custom Weighted K-Nearest Neighbors
- Present results to user
  - Table
  - D3.js map
- Allow user to sort results via TOPSIS for best fit
  - User defined weights
  - Three factors:
    - Nightlife
    - Crime
    - Weather

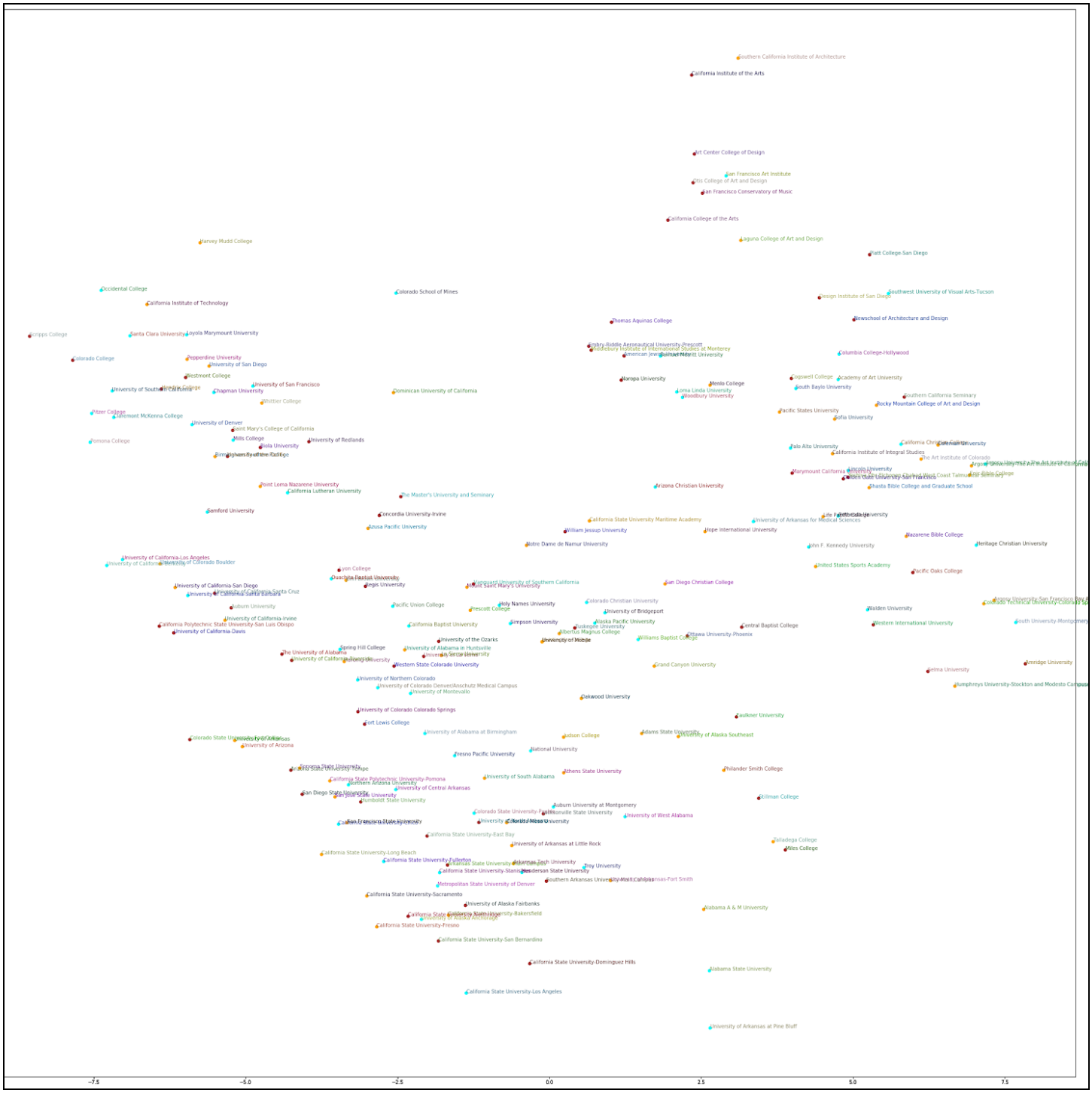
Data Cleaning



University Data

A main source of data was the College Scorecard dataset which is nationally collected data from the U.S. Department of Education. We first narrowed the dataset to only undergrad colleges. We then filled in NaN values with the respective mean for that data attribute. This dataset provided most of the filtering information which built the school profiles. Users would answer questions in the survey which would build a student profile with attributes derived from scorecard data.

On the top is a heat map of the correlation between 13 attributes from the cleaned dataset. We ensured that none of the variables were too correlated with other variables in the dataset.



Above is the scatter plot of the first 200 colleges after applying PCA, with PCA component 1 on the x-axis and PCA component 2 on the y-axis. If we look closely, we can see some obvious grouping of colleges in this scatter plot. With most of the state colleges concentrated in the bottom, seminary schools at the top left, and Colleges from California present majorly in the mid-right section.

Crime Rate Data

We first attempted to obtain crime rate data from the FBI Crime Data Explorer. By using this dataset, we were constrained to state-level crime rates, which only helped in filtering of schools from various states. We then discovered the U.S. Department of Education's Campus Safety and Security data. This data contained crime information in several categories (murder, burglary, etc.) for different zip codes. We obtained this data for each zip code, and for those with missing data we substituted the state average. This datasource allowed for a closer crime comparison between schools in the same state.

Weather Data

NCDC Climate Data Online provides API to collect Temperature data from their Global Summary of The Month dataset.

- For each zip code, average temperature for Spring, Summer, Fall and Winter was fetched from each Weather Station in the nearby region.
- If the zip code became too specific, the FIPS associated with that zip code was used to fetch data from Weather Stations in the FIPS region.
- If no data could be obtained even then, these missing values were filled with mean temperature values from the Weather station in that state.

These columns were then appended to the cleaned dataset.

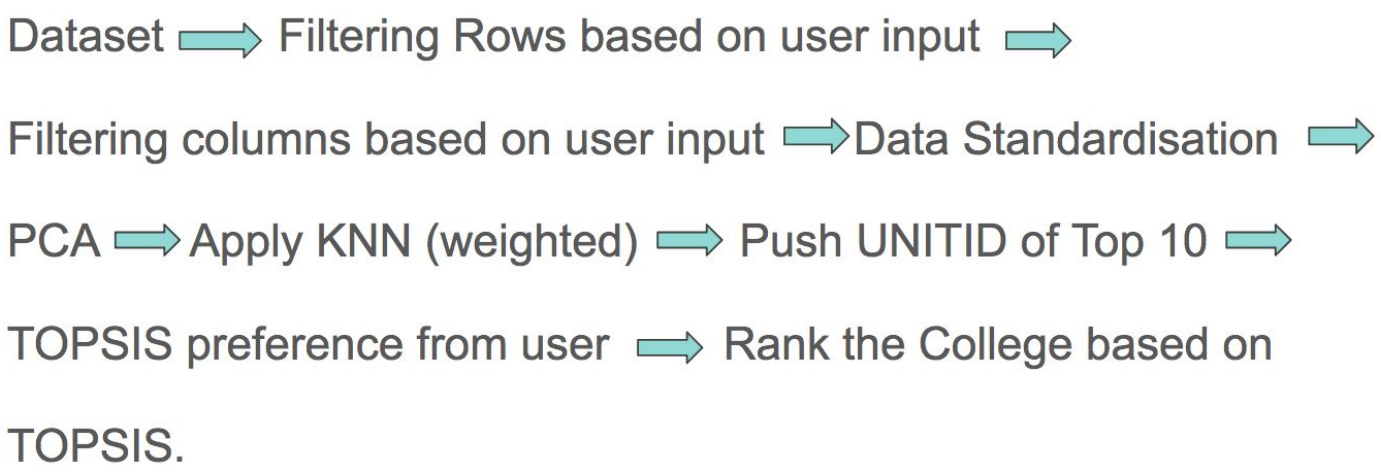
Nightlife Data

Nightlife data was collected for each university that was in the final list of ten following completion of the survey and the KNN best fit model. Nightlife data was collected from the Google Places API by making an HTTP call for each university. For each call, the universities latitude and longitude were passed as parameters along with the "type" of place being set to "bar" and setting a radius of 2.5 kilometers. Data was returned as a JSON list of locations that matched the criteria of the call. We decided to quantify the nightlife by measuring the quality and quantity of the bars returned. Taking the number of bars returned and dividing by the maximum that the Google Places API could return gave us a measure of quantity. We then took the average Google Places user rating, measured out of five stars, of the bars returned. By averaging the quantity score and quality score, we were able to obtain an overall nightlife score for each university in our final ten.

Survey Construction

The first step to building a student profile for the user was to collect their preferences through the web interface. To construct the survey, we used the SurveyJS library. This tool let us easily construct questions of varying types such as radio, dropdown, text input, and sliders. The library maintains the markup for the survey in a JSON file and once a user has completed all of the questions, a JSON object is neatly packed and sent for processing. We also utilized the library's ability to include conditional questions, which was useful in keeping our question set intuitive for the user. We constructed a 28 question to collect data to build student profiles. Even though our question set seems a bit lengthy, we feel that the time and effort spent on the survey provides the most accuracy in results presented to the user. As mentioned earlier, we focused on creating a survey which asked questions in a more personal, rather than logical manner. This mindset lead us to reframe certain questions such as “what range of acceptance rate are you looking for?” to “Academically, I would prefer …” with answer choices ranging from “To be rigorously challenged” to “Not touch a book, I’m here to party!”. Asking questions with this style presented the user with a question they are more comfortable answering while allowing us to collect the data we needed.

Architecture



Once a user submits a query through the Survey form, we reduce the dimensionality of the dataset by filtering the rows on the basis of some upper bounds on the input values. We then drop the columns which are of no interest to the user and should not be influencing the final results.

We then standardize the dataset values and the user input so that all the features have mean 0 and a standard deviation of 1. After standardizing we use an importance metric to assign a weight to each feature based on how much a particular feature should influence the final result. So the feature that should have a higher influence on the final result has a lower weight assigned to it and is kept closer to the other instances whereas a feature that contributes less to the final result is pushed farther away from other instances by assigning it a larger weight.

The dimensionality even after the initial filtering tends to remain high, so to reduce the dimensions even further, we perform a Principal Component Analysis on the remaining features and extract the first 10 PCA components that explain the maximum variance in the dataset.

After all the preprocessing on the filtered dataset, we apply K-Nearest Neighbors algorithm to find a maximum of 10 nearest neighbors to the user input. The UNIT IDs of the resulting colleges are then pushed to the results page.

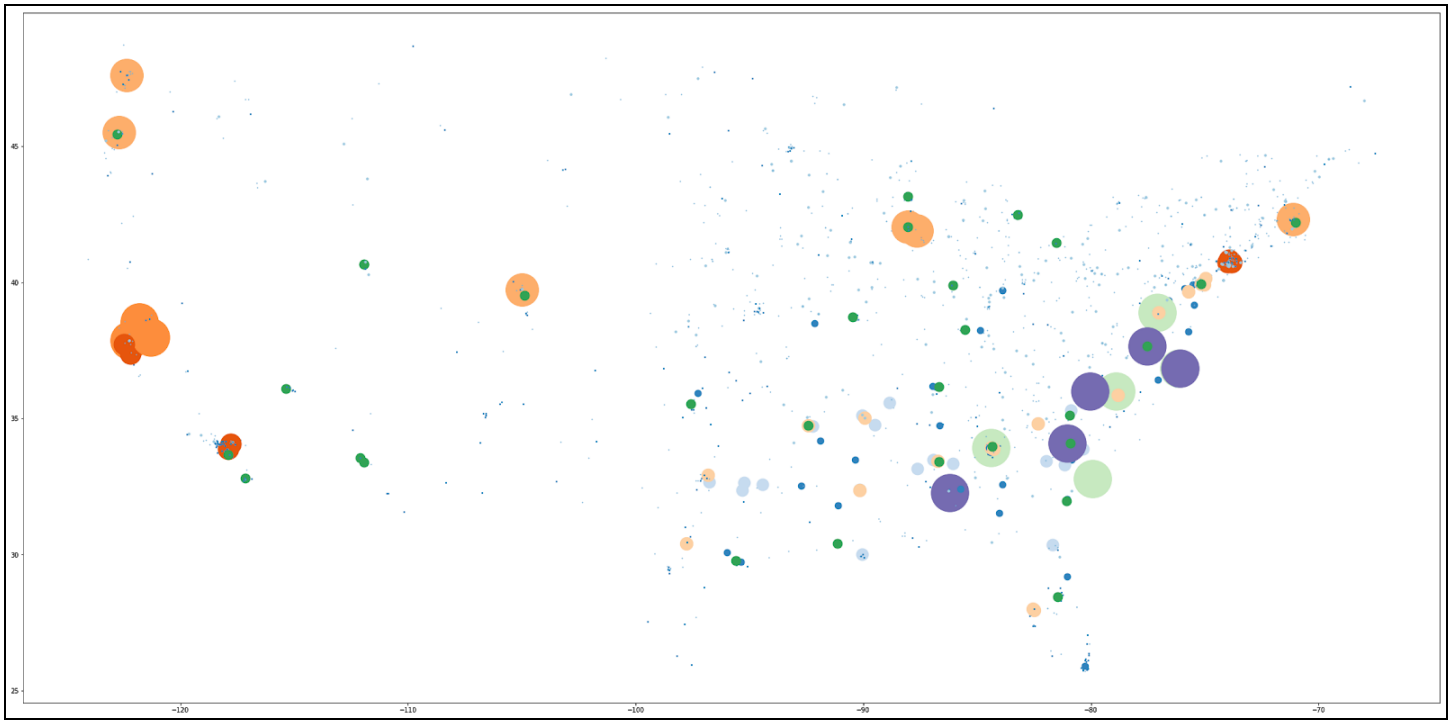
On the results page, we provide a second layer of filtering to further fine tune the results on some secondary parameters like Temperature, Crime Rate and Nightlife. Here the user can assign an importance metric on his own to the secondary features. We run a TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution) analysis based on the weights provided by the user to sort the results coming from the first layer, i.e., KNN.

Validation

As we did not have any source of validated data (a peer reviewed dataset of top 10 colleges for specific profile), we could not validate our predictions offline. As the timeframe of the project was quite small, we could not incorporate user feedback to validate our results on the fly.

DBSCAN

DBSCAN is a clustering algorithm that only takes in as input a threshold of maximum distance to be included in a cluster, and generates the clusters automatically. For our dataset, we used a threshold of 4.5, which generated the following clusters (represented on the coordinates of each college).



Rank	School Name	City	State
1	Georgia Institute of Technology-Main Campus	Atlanta	GA
2	University of Maryland-College Park	College Park	MD
3	University of Delaware	Newark	DE
4	Auburn University	Auburn	AL
5	Virginia Polytechnic Institute and State University	Blacksburg	VA
6	The University of Alabama	Tuscaloosa	AL
7	Clemson University	Clemson	SC
8	Rowan University	Glassboro	NJ
9	SUNY at Binghamton	Vestal	NY
10	James Madison University	Harrisonburg	VA

If a cluster contains a large number of colleges, the colleges of that cluster will be represented by smaller points. If the cluster is small, it indicates uniqueness of that cluster, and hence those colleges have been highlighted by bigger points. Hence for the colleges on the left generated from user input, if the user's input falls inside a unique cluster, the user is most likely to be suggested more colleges from the same cluster. 6/10 colleges from the list on left lie in the 'lilac' and 'light green' colored clusters on the East Coast of United States, thus validating our results.

Repository and Sources

GitHub Repository

https://github.com/aman-chauhan/CSC-591-BestCollege4Me

https://github.com/parthnagori/bestcollege

Data Sources

https://collegescorecard.ed.gov/data/documentation/

https://ope.ed.gov/campussafety/#/datafile/list

https://www.ncdc.noaa.gov/cdo-web/

https://developers.google.com/places/

Demo

https://bestcollege.herokuapp.com/bestcollege/index

https://tinyurl.com/ybatdr