The Pursuit of Happiness: CSCI 1951A Capstone

Simran Shankardass, Ahmad Alkhatib, Hannah O'Keeffe, and Thin Su San

May 2023

Goal

Brown is known as the happiest campus, and students say they love it because of how happy it is, not because of its academic ranking or endowment. Therefore, we are interested in investigating which features of a college are predictive of college happiness, and how it relates to retention and other attributes. Happiness is also an important factor in applying to colleges, but not easy to measure – therefore we want to see if we can use these features to provide a recommendation system for colleges similar to an input.

Data

We collected data from two sources. The College Scorecard dataset, produced by the U.S. Department of Education, provided institution-level data for 1000s of colleges, including characteristics, enrollment, costs, and student outcomes. Some attributes had null values for some colleges; we dropped these from the dataset when necessary during analysis. We also scraped student life quality grades and rankings from Niche.com, as a measure of happiness. Our final dataset, joined on college name, contains 1180 data points with 82 attributes after dropping colleges that weren't present in both sets of data. However, the Niche scraped data is not uniformly distributed, so we had many more data points with B+, B, and B- grades than the others.

Model + Evaluation Setup

We want to understand what features of our dataset are predictive of a high student happiness grade, and are interested in retention rate specifically. While we can't necessarily prove causality, by scaling our data and using regularized regression, we can determine which features are most important in predicting grade based on the absolute value of their coefficient from the trained model. We attempted to use a Random Forest model as the scraped happiness grades are categorical, but also converted these to GPA points in order to train a regression model using Lasso. We report different kinds of accuracy - a mean absolute error using the GPA, and percent of accurate predictions using the letter grades. We also extracted the features with the highest Lasso coefficients to inform our K-Means clustering model, which forms groups of similar colleges that we use for the recommendation system. Our optimal K-value is 4, determined from an elbow plot, which gives us a silhouette score average of 0.43. For recommendations, however, we use K=10 so that the number of recommendations provided is manageable.

Results and Analysis

Claim #1: The Lasso regression model performs better than Random Forest and baseline models.
Support for Claim #1: Table 1 shows the MAE and accuracy of the Lasso regression model compared to Random Forest, a baseline model that predicts the most frequent label, and a baseline model that uniformly predicts a label. Lasso has the highest performance in all categories.

	Lasso	**Random Forest**	**Most Frequent**	**Uniform**
MAE	0.403	0.457	0.559	0.960
Correct predictions (%)	27.6	24.4	19.8	6.25
≤1 grade from true grade (%)	67.7	61.5	54.7	30.7

Claim #2: Our model finds retention to be predictive of happiness.
Support for Claim #2: The retention rate attribute has the highest Lasso coefficient by a significant margin, indicating that it is most predictive of happiness. We can conclude this because we standardized the data, allowing us to rank independent variables by coefficient size. (However, our ranked sum hypothesis test showed that there is significant difference between the retention rate and happiness distributions).

Claim #3: The K-means clusters will be moderately compact, with some overlap.
Support for Claim #3: Using the optimal K=4, our clusters' silhouette score average is 0.43, meaning there is some overlap between clusters. If there are features involved in calculating the Niche happiness grades that aren't in our dataset, they might be contributing to this outcome.