We're Data Scientists

Jessica Dai, YeonJun Kim, Amy Pu, Rebecca Zuo

What can we infer about a public school given knowledge about the opportunities it offers?

Datasets

We investigate *public high schools* in the United States. All of these are data collected by the federal government and accessible via data.gov.

1) Arts education
2) CRDC (Civil Rights Data Collection) -- demographics & school resources
3) Edtech resources -- technology available to classrooms (e.g. laptops)
4) Career & technical education programs

Hypothesis:

Based on our knowledge about public school funding and American history, we thought there might be some correlation between demographics of a school -- in particular, the percentages of low-income and black/latinx students -- and the types and quality of extracurricular opportunities it provides.

However, we found it difficult to join the data tables in order to have the anticipated independent and dependent variables due to the limited and varying joinable attributes among the different data sets.

This led to our revised hypothesis: **Based on the attributes from each dataset, we can predict the geographical region and/or community type with accuracy better than random guessing.**

Challenges

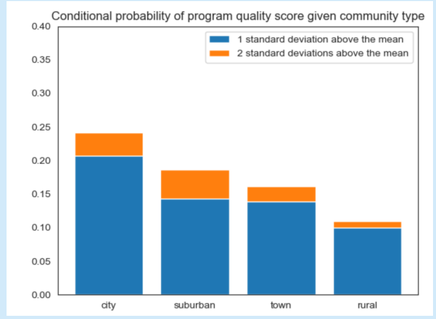
1) How do we join these datasets? -- As it turns out, each of these datasets were the result of separate government surveys -- the only common attributes were school *region* and *community type* (urban/rural/etc).

2) Which variables or features should we treat as predictive? demographics -> extracurriculars, or vice versa?

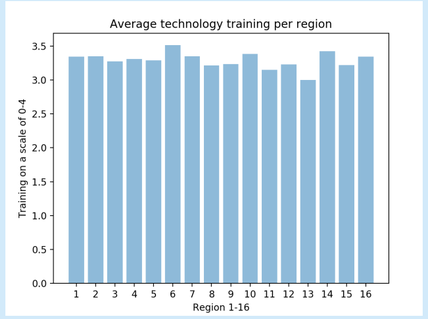
3) There didn't seem to be correlations! Preliminary SQL queries and graphs showed very little statistical significance...

Results

CRDC dataset



CTE dataset



EdTech dataset

Methodology

Given the limitations of our datasets, we decide to predict **school region** (northeast, southeast, central, and west) and **community type** (urban, suburban, town, rural), instead of creating "fake"/ untrue data by joining entries from all the datasets.

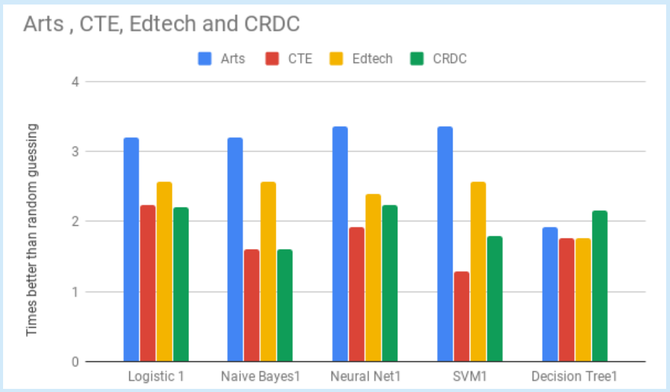
1) For each dataset, build predictive models for **region and community** (16 total possible classifications)

2) The final prediction presumes we have information about the school's demographics, arts, CTE, and edtech programs; given this, **can we predict region and community type**? The answer is based on the votes of the models for each dataset, weighted by their individual accuracy.

Classifiers & Evaluation

CRDC: Attributes: race, gender, and types of programs offered. Labels: location (We were unable to predict the community type, as this information is not provided).
CTE: Attributes: program quality, barriers to providing, barriers to participation, how resources affect adding, and how resources affecting removing programs.
Edtech: Attributes: number of computers, the level of technology integration, and level of technology training.
Arts: Attributes: programs availability/resources and participation rates.
*Labels for CTE, Edtech and Arts: school location and community type.

Accuracies for each dataset's predictive models



Known limitations

1) We recognize a glaring issue is the inability to test the final predictor for how well it does (for the very reason we chose to shift our strategy in the first place).

2) Our model accuracies aren't great -- the best accuracy is around 20%, far from any usability in the real world. That being said, the models are still meaningfully better than random guessing -- the Arts dataset consistently yielded an accuracy around 3x better than random guessing, and each dataset had at least one model perform at least twice as well as random guessing.