The Impact of Socio-Economic Status on Academic Performance in New York State

Vuong Tran, Ben Spivey, Kshitij Srivistava, Nick Orangio, & Dave Dyer

New York State Education Department
Knowledge › Skill › Opportunity

Abstract

Through interactive data visualization and machine learning, we have identified proxies for socioeconomic status (SES), identified SES variables that correlate with academic performance, assessed the geographic distribution of SES variables, and fit machine learning models to identify the most significant SES factors associated with graduation rates in New York State.

Introduction

Academic performance is often positively correlated with higher resources/funding and negatively correlated with higher needs, such as free and reduced lunch programs. We've modeled the impact of SES on AP using New York State school data, identified significant features, and produced an interactive visualization.

Data 1 - NYSED

The New York State Education Department (NYSED) data was downloaded from https://data.nysed.gov/downloads.php. It has the following properties.

- Data were primarily in MS Access format.
- There are 423,178 school records.
- Developed custom data extraction code.
- These data power the choropleth map.
- Direct SES indicators were limited but had associated variables.
- Economically disadvantaged, homelessness, and disability were the most salient features in early exploratory analysis.

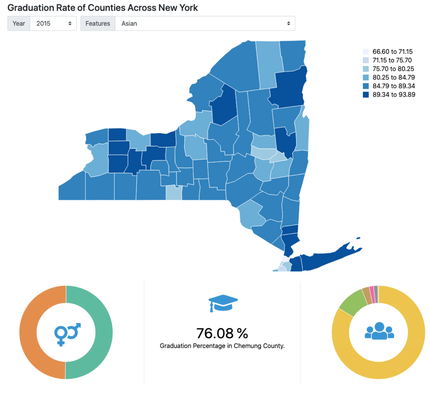

Figure 3: NYSED Choropleth

Data 2 - Cornell

The Cornell data set was sought after the challenges of finding good SES proxies in NYSED. It was downloaded from https://www.nyeducationdata.org/.

- Data were primarily in .csv format
- Cornell is a wide and tall data set with 82 mostly-economic features
- There are 32,632 records by 63 features
- Cornell required many joins
- These data power the PCA Chart
- SES proxies were plentiful
- Aid, Expenditures per Student, and Wealth per Pupil were the most salient features

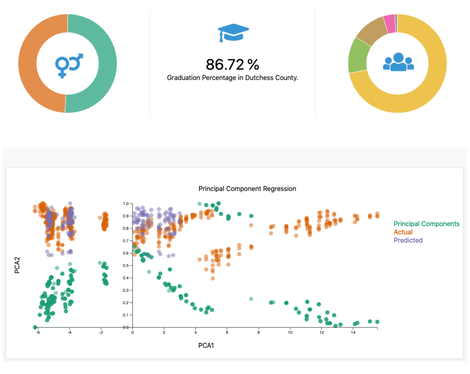

Figure 4: PCA Structure in Dutchess County

Analysis & Modeling

The analysis for each dataset was performed as follows:

1. Identify proxies for SES and Academic Performance (AP).
2. Examine which SES variables correlate with academic performance on a univariate basis.
3. Assess how SES variables are geographically distributed across the state with an interactive visualization.
4. Build multivariate machine learning models to identify significant SES factors associated with AP.

The multivariate analysis was performed using a random forest to identify the most significant SES features associated with academic performance. The Cornell dataset was used for this analysis as it contained many (>60) features measuring different aspects of SES, including many financial metrics. The top features include gross income total wealth per pupil, the amount of state aid, and the alternate pupil wealth ratio.

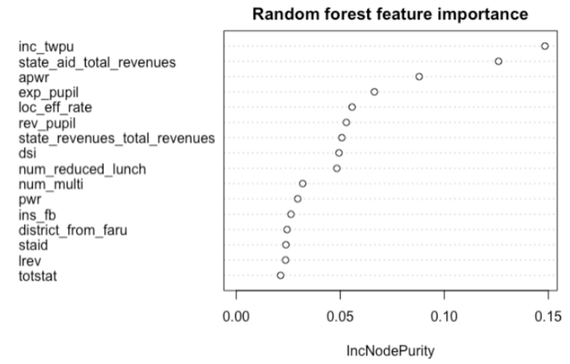

Figure 1: Random forest feature importance with Cornell dataset

Conclusions

- The distribution of graduation rates for economically disadvantaged cohorts is lower than non-disadvantaged with statistical significance. (p<0.001)

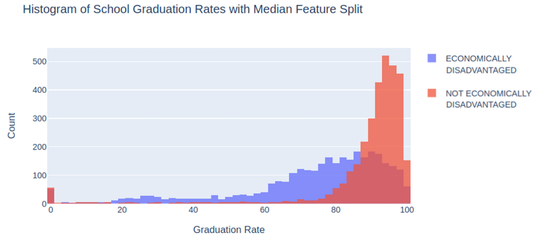

Figure 2: NYSED graduation rates by economic status

- The distribution of graduation rates for economically disadvantaged cohorts in charter schools is higher than in public schools with statistical significance. (p<0.001)
- Random forest feature selection on NYSED Data identified disabled status, economically disadvantaged status, and limited English proficiency as significant factors
- Random Forest Feature Selection on Cornell Data identified gross income total wealth per pupil, the amount of state aid, and alternate pupil wealth ratio as significant factors
- The most important feature contributing to the Principal Components Analysis (PCA) varied by county, but was dominated by economic investment measures
- Interactive exploration of these results are available in the D3 visualization, including visualization of the principal component analysis breakdown and feature analytics

Choropleth Instructions

1. cd to `CODE/d3-bootstrap-viz`
2. run `python -m http.server`
3. point your browser to `localhost:8000`

Contact Information

Ben ben.spivey@gatech.edu
Nick nickorangio@gatech.edu
Vuong vuong.tran@gatech.edu
Kshitij ksrivastava34@gatech.edu
Dave dave.dyer@gatech.edu