

NYC PUBLIC SCHOOLS: INFLATED GRADUATION RATES, BUT A LOW COLLEGE READINESS INDEX

Michelle Cronin

GA - DAT30

Quick Draft Presentation – March 2, 2016

THE ISSUE

- Several articles have come out recently pointing out that although graduation rates in NYC public schools have risen over the years, the rate of students prepared for college has remained low
- Thus, NYC public schools are graduating many students each year who are not considered college-ready
- Can we predict which schools are graduating high levels of college-ready students at a rate near-consistent with its graduation rate?

The screenshot shows a news website layout with a dark header and a light-colored main content area.

DAILY NEWS (in white on black) is on the left, followed by a red bar containing **NEW YORK** and other navigation links: NEWS | POLITICS | SPORTS | ENTERTAINMENT.

The New York Times logo is in the center of the header.

N.Y. / Region is on the right side of the header.

NYC CRIME | BRONX | BROOKLYN | QUEENS | MANHATTAN | EDUCATION | WEATHER | OBITUARIES | NEW YORK WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS are listed below the header.

School grades: Department of Education data show less than a third of New York City high school students are college-ready (headline in large bold letters) is the first article. It includes a sub-headline: "City Education Department releases A's to F's for city high schools; numbers are up to 29% from 25% last year". Below it are author information ("By BEN CHAPMAN, VERA CHINESE, RACHEL MONAHAN / NEW YORK DAILY NEWS / Updated: Tuesday, November 27, 2012, 1:19 AM") and social sharing icons.

Most New York Students Are Not College-Ready (headline in large bold letters) is the second article. It includes author information ("By SHARON OTTERMAN Published: February 7, 2011"), a search bar, and social sharing icons.

NEW YORK POST logo is on the right.

METRO (in red) is a category label.

Most NYC high school graduates at CUNY community colleges get remedial help (headline in large bold letters) is another article. It includes author information ("By Aaron Short") and a timestamp ("July 5, 2015 | 9:54am").

WHAT DATA HAVE I GATHERED, AND HOW DID I GATHER IT?

- Mostly measurable characteristics of NYC public high schools (performance, demographics, location, etc.) from NYC Open Data and the Department of Education:
 - 2013_2014_HS_SQR_Results_2015_03_02.xlsx
 - 2014_2015_HS_SQR_Results_2016_01_07.xlsx
 - DemographicSnapshot201011to201415Public_FINAL-2.xlsx
 - 2015_Graduation_Rates_Public_School-2.xlsx- DOE_High_School_Directory_2013-2014.csv
 - DOE_High_School_Directory_2014-2015.csv- 2014 Public Data File SUPPRESSED.xlsx
 - 2015_Public_Data_File.xlsx- Location_Information_Report.csv
 - Zip_MedianValuePerSqft_AllHomes.csv

`data.shape`

(975, 46)

male_percent	poverty_percent	avg_home_value_sqft	district_admin_code	grade8_english	grade8_math	regents_algebra	regents_english	regent
0.623529	0.866667	1321.166667	1	2.18	2.06	64	66	53
0.621711	0.891447	1321.166667	1	2.27	2.37	64	69	67
0.530030	0.735736	1324.500000	1	2.66	2.63	65	75	64
0.432507	0.845730	1321.166667	1	2.28	2.09	61	69	63
0.484726	0.273199	1321.166667	1	3.50	3.53	80	91	87

WHICH AREAS OF THE DATA HAVE I CLEANED, AND WHICH AREAS STILL NEED CLEANING?

- Done:
 - Dropped and imputed missing values
 - Feature Selection – Necessary due to overlapping features in sources as well as presence of irrelevant features
 - Some Feature Engineering:
 - Length of School Day (End of day – Start of day)
 - Diversity Score (Sum of the squared reported race/ethnicity percents [black, white, Hispanic, Asian])
 - Overall SAT score (Reading SAT + Writing SAT + Math SAT average scores)
 - Overall Regents score (Algebra + English + Living Environment + Global History Regents average scores)
 - Grade 8 Proficiency score (Grade 8 Math + Grade 8 English scores)
 - Tidied the names
- To Do:
 - Dummy categorization (Zip codes and/or Administrative districts)
 - Possibly more feature selection

WHAT STEPS HAVE I TAKEN TO EXPLORE THE DATA?

- Created visuals to look at correlation (heatmap) and missing values
 - Very unclean data → so most EDA thus far has been targeted at cleaning it
 - Explored feature variance
 - Looked for outliers

WHAT INSIGHTS HAVE I GAINED FROM MY EXPLORATION?

- High school's graduation rate and college readiness index share high correlation with the following features (in varying strengths):
 - Post-Secondary enrollment rate
 - Student Attendance Rate
 - Percentage of students who take the SAT
 - Overall Regents score
 - Grade 8 Proficiency Scores
- Average graduation rate for 2014 and 2015 combined: 73.34%
 - Standard Deviation: 0.160871
- Average college readiness index for 2014 and 2015 combined: 29.05%
 - Standard Deviation: 0.232938

HOW MIGHT I USE MODELING TO ANSWER MY QUESTION?

- Random Forest
 - Protect against overfitting (great because I have a considerable amount of features, and they're still growing)
 - Will tell me most important features
- Logistic Regression
 - Lasso/ridge will drop/let me know unimportant features
 - Will allow me to view feature coefficients
- k-means Clustering:
 - I may be able to discover distinguishable groups of schools to find some counterintuitive combinations of characteristics as well as obvious ones



WILL I BE ABLE TO ANSWER MY QUESTION WITH THIS DATA, OR DO I NEED TO GATHER MORE DATA (OR ADJUST MY QUESTION)?

- Yes, I am fairly confident that I will be able to answer my question with this data
 - Potential problem: very unbalanced class

