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

Michelle Cronin

GA - DAT30

Final Presentation DRAFT – March 7, 2016

THE ISSUE

- Although graduation rates in NYC public schools have risen over the years, the rate of students prepared for college has remained low (73% vs 29% on average, from my data)
- Thus, NYC public schools are graduating many students each year who are not considered college-ready
- Can we predict which schools are graduating college-ready students at a rate above NYC's average graduation rate (within one standard deviation)? ← High-performing schools

HYPOTHESIS

- Major contributing factors may include:
  - Race/ethnicity of the student body
  - Poverty/wealth of the student body
  - Geographical locations of the schools/crime
  - Testing (test scores directly contribute to the college readiness index due to CUNY observations of remedial class rates)

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 website and the NYC Department of Education website:
  - 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
  - seven_major_felony_offenses_by_precinct_2000_2014.xls
  - violation-offenses-by-precinct_2000-2014.xls
  - Zip_MedianValuePerSqft_AllHomes.csv

```
data.shape
```

```
(975, 51)
```

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

PRE-PROCESSING

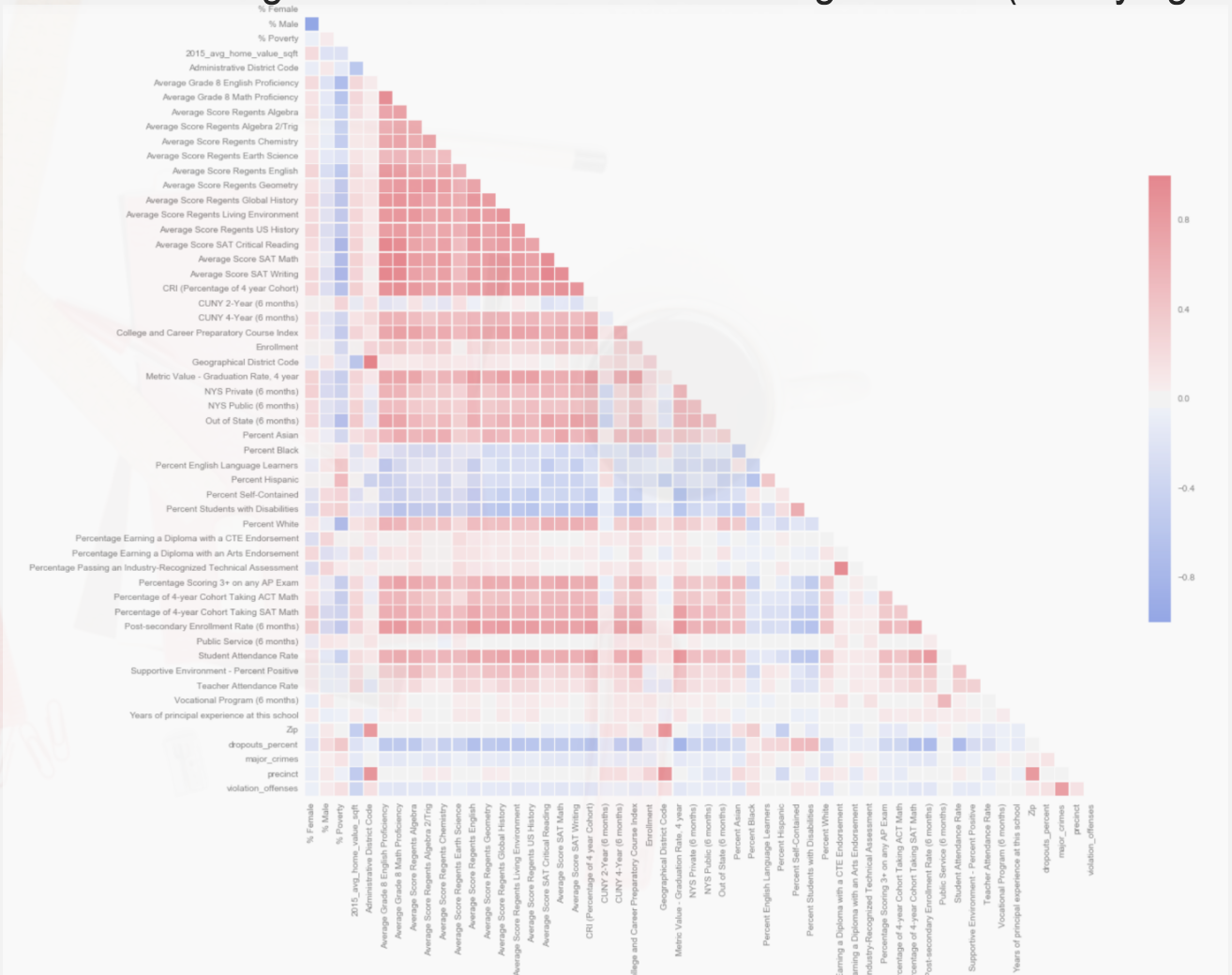
- Dropped and imputed missing values
- Feature Selection – Necessary due to overlapping features in sources as well as presence of irrelevant features
- Feature Engineering:
  - Length of School Day (End of day – Start of day)
  - Diversity Score (Sum of the squared reported race/ethnicity percents [black, 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)
  - All Crime (Major crimes + violation offenses)
- Tidied the names
- Dummy categorization (Zip codes and Administrative districts, separately)

FEATURE SELECTION

- Created visuals to look at correlation (heatmap) → Removed highly correlated features (ex: many Regents exams)
- Created visuals to view missing values → Removed those missing 50%+ (ex: pupil-teacher-ratio)
- Viewed feature variance → removed those with lowest variance (ex: Public Service percentage attendance)
- In the end, removed features directly related to the response variable to see how well other features would fare
  - Graduation Rate
  - College-Readiness Index (CRI)
  - Test scores related to the CRI: average SAT scores, Regents values, percentage of students scoring 3+ on AP exam
- Ran models on four different datasets:
  - One with test scores (only for comparison)
  - One without test scores
  - One with administrative district codes (also without test scores)
  - One with zip codes (also without test scores)

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
  - Poverty and Disability (Negatively correlated)



MODELING

- Minimum Benchmark: 87%
- Random Forest
  - Protect against overfitting (great because I have a considerable amount of features)
  - Will tell me most important features (which I can feed into other models)
  - Using accuracy to validate the models, they all beat the benchmark at ~95%
  - Using F1, only model with zip codes does (barely at 88%)
- 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

	Features	Importance Score
4	grade8_english	0.154855
28	dropouts_percent	0.116998
5	grade8_math	0.108643
24	student_attendance_rate	0.102605
33	grade8_proficiency	0.096839
23	post-secondary_enroll_rate	0.059962
7	cuny_4yr	0.043111
22	sat_percent_took_exam	0.037006
9	college_nys_private	0.025561
12	asian_percent	0.022963
21	act_percent_took_exam	0.022491
32	population_diversity	0.022011
2	poverty_percent	0.019970
18	white_percent	0.013664
11	college_out_of_state	0.013422
16	self-contained_percent	0.012589

CHALLENGES/SUCCESSES

- CHALLENGES:
  - Putting together different datasets into one ate up tons of time, including endlessly searching for additional datasets for information I was lacking but felt could be valuable
  - I have so many features that even my shortest dataset feature-wise wouldn't run Logistic Regression or SVM on it
  - Issue with k-means with inverse_transform to view cluster centers; somehow a different size than it should be??

- SUCCESSES:
  - Interesting takeaways, some supporting hypothesis but some new perspectives too
  - Despite highly unbalanced class, models performed better than I expected

CONCLUSIONS

- Grade 8 Proficiency importance
- FUTURE STEPS:
  - Complete similar work but in middle/elementary schools
  - Longitudal view of the data as more data is released with each schoolyear; track progress