A group of people standing in front of a crowd posing for the camera

Description automatically generated

# Specialized High School Scholastic Aptitude Test Admission Performance

Capstone Two: Predictors of test performance for specialized high school admissions offers.

* [Jupyter Notebook](https://github.com/dj-m/CapstoneTwo/blob/master/notebooks/CapstoneTwo_PredictingHighSchoolAdmissionsOffers.ipynb)

## Background

Performance on the Specialized High School Admissions Test (SHSAT) determines eligibility to one of the eight specialized high schools in New York City. It is administered by the New York City Department of Education (DOE) to about a third of the city’s 8th graders, with **5,000** receiving admissions offers. Of major concern is the racial & ethnic breakdown of admitted students, showing significant underrepresentation from Black and Latinx students.

## Data

Two main data sets will be used:

2016 School Explorer (Explorer)

* This dataset consists of 1272 schools in New York city, and 161 variables, provided via Kaggle. Primarily, it’s school descriptors, e.g. grades, race & ethnicity student percentages, high/low performing percentages of students. Data is available as a single csv file.
  + [Kaggle Dataset](https://storage.googleapis.com/kagglesdsdata/datasets%2F33225%2F44131%2F2016%20School%20Explorer.csv?GoogleAccessId=gcp-kaggle-com@kaggle-161607.iam.gserviceaccount.com&Expires=1595273466&Signature=fk6%2BI64ZSKXeenOP24Hsst9gNp2z3gFDGnz1rSXzcrnS874EFfR1VjUPVu0mCoN0bwXxJ7udjKpGlD51QLqiolRTpt9t%2F6ko672nNzd2KU0zJd4xRN8yW4Ouk1XxbCCTN2u6In241T1%2BY1RMpSp5rQgko83zQtwPClQPsl%2BWynlztsHV1aWF2K1J6MUy1SBaXyHvTSXBiMp1G2FvCoVjRVyjkXwV94Xgayi8Zgs3ISjyVUZn3yYzuyarl8NUwSnryWnfCE1debgt5z9AP5aTv7IbUA297hpYAhHZR0NjtKMwoadxypbWbBZ6cUTgI8KT4L4q8LdCgJ6SDJolkJERaQ%3D%3D) (API)

2017-2018 SHSAT Admissions Test Offers By Sending School (Offers)

* This dataset consists of the 2017 SHSAT results, published by NYC in 2018. All test takers are north of 28,000, grade 8 students, Test takers and offers received are grouped by school. Data is available as a single csv file from the NYC Open Data portal.
  + [NYC Dataset](https://data.cityofnewyork.us/resource/vsgi-eeb5.csv) (CSV)

## Approach / Method

The goal of this analysis is to elicit which factors predict performance on the SHSAT. These factors will serve as beacons to direct or draw services, whether education-based or otherwise, towards improving the percentage of Black and Latinx students admitted to the specialized high schools.

This approach aims to quantify which variables lead to admissions offers \_beyond\_ prior proxies: English Language Learners, Students with Disabilities, Students on Free/Reduced Lunch, and Students with Temporary Housing.

Initially we can assume that those students who perform well on typical standardized tests, throughout the school year, would therefore perform well on the SHSAT. We'll investigate this and extrapolate as to whether this is the case across all schools/students that follow this assumption.

As a bit of forecasting, I'll use linear regression models to determine how many admissions offers schools that fit a certain testing/aptitude standard could be getting based on their current testing scores.

## Data Cleaning

To determine what factors are related to receiving admission offers to the specialized high schools, the data feeding into the models need to be not only numeric but free of errors.

We can see that the \_2016 School Explorer\_ data set has three columns almost entirely of null values. These can be filled with an appropriate value for the data type of those columns.

|  |  |
| --- | --- |
|  | Percent |
| Other Location Code in LCGMS | 0.999214 |
| Adjusted Grade | 0.998428 |
| New? | 0.978774 |
| School Income Estimate | 0.311321 |
|  |  |

Given the test-takers in the *2017-2018 SHSAT Admissions Test Offers By Sending School* are a year away from taking the test in *2016 School Explorer* dataset, I'll focus on the 7th graders.

*2016 School Explorer* has 20 variables with information 7th graders. This data is broken up into two kinds of information, ELA (English Language Arts) & Math. Scoring on these tests top out at 4, with 1 representing the worst score.

A picture containing text

Description automatically generated

Therefore, the best students are in the '4s' columns shown above.

Summary of columns:

* All students tested
* All students with 4 scores
* American Indian or Alaska Native with 4 scores
* Black or African American students with 4 scores
* Hispanic or Latino students with 4 scores
* Asian or Pacific Islander students with 4 scores
* White students with 4 scores
* Multiracial students with 4 scores
* Limited English Proficient students with 4 scores
* Economically Disadvantaged with 4 scores

In 2016, the total number of 7th graders in NYC Middle Schools was **69,053**. Of those, **8,320** had ELA scores of 4, and **10,888** had Math scores of 4.

Graphical user interface, text

Description automatically generated

**537** NYC Middle Schools sent at least 6 students to SHSAT for a total of **25,349** 8th graders taking the test. **57** schools send 0-5 8th graders to take the test. **121** NYC Middle Schools saw \_at least\_ 6 of their students receive offers, for a total of **4,018** 8th graders having received an offer. **473** schools saw 0-5 of their 8th graders receive an offer.

**Merging Datasets**

Using the DBN & Location Code I'll merge *Explorer* data for 7th graders to the *Offers* information for SHSAT testers. In the process it looks like **2** schools in *Explorer* didn't have information in the *Offers* dataset.

## Exploratory Data Analysis & Feature Engineering

Looking at the assumption that those students/schools that have the majority of the 4 scores will, in turn, perform well on the aptitude test for the specialized high school, we can see that Black & Latinx students may receive less admittance offers based on this limited criterion.

Chart, bar chart

Description automatically generated

In order to better summarize the schools/students into ranges of test scores, I've added the following features:

* “Percentage of SHSAT takers receiving an offer” (Numbers of SHSAT takers / Number of Offers by school)
* “The total number of Black/Hispanic students in Grade 8” (Number of 8th graders \* Percent Black / Hispanic)
* “Percentage of students who did the SHSAT” (Number of SHSAT takers / Number of 8th graders)
* “Average Mark” (the average of Average ELA Proficiency and Average Math Proficiency)
* “Percent of students with Level 4 ELA in Grade 7 (Grade 7 ELA 4s - All Students / Grade 7 ELA - All Students Tested)
* “Percent of students with Level 4 Math in Grade 7 (Grade 7 Math 4s - All Students / Grade 7 Math - All Students Tested)
* “Percent of students with Level 4 in Grade 7” (average of 4 percentages ELA and Math in Grade 7)
* “Average number of Level 4 students” (Grade 7 ELA 4s - All Students + Grade 7 Math 4s - All Students)/2

**Schools with the highest number of test takers**

What we're seeing in this next plot is that those schools that send the most 8th graders to the SHSAT, have less of their school, percentage-wise, represented by Black or Latinx students.

Interestingly, there is a high percentage of Black/Hispanic students (The William W. Niles (**82%**) school and The Eugenio Maria De Hostos (**78%**) school), near the middle of the pack and the lowest, respectively.

Chart

Description automatically generated

**Schools with the least number of test takers**

*Nearly* all of the schools with the least number of test takers in 2017 (55) had low average marks (average of AvgELA and AvgMath).

Also, most schools had a high percentage of Black or Latinx students.

Chart

Description automatically generated

**Number of offers by school**

The top **25** schools with the most offers received had lower percentages of Black or Hispanic students (highest percentage is at Frank Sansivieri school with **59%** Black or Latinx students).

Chart

Description automatically generated

**Highest percentage of offers for the number of test takers**

Below are the top **20** schools that had the highest percentage of offers for the number of test takers, representing how successful that school was as to the number of students that were admitted to the specialized high school.

The Christa McAuliffe School had the most success with **82%** of 251 students taking the test getting an offer.

The schools scoring best at the percentage of students actually getting an offer are very low in Black or Latinx student percentages (the exception is the small Columbia Secondary School with **64%** Black/Latinx).

A picture containing chart

Description automatically generated

**Least percentage of offers for the number of test takers**

Of those schools which had *at least* 6 offers, the 20 schools with least success are shown below.

Two of the largest schools that are predominantly Black/Latinx and sent many students to the test are J.H.S 118 William W. Niles school @ **424** 8th graders (**82%** Black/Latinx) and I.S. 318 Eugenio Maria De Hostos @ **467** 8th graders (**78%** Black/Latinx).

A picture containing chart

Description automatically generated

## Models & Evaluation

My intent is to use regression-based models because my label/y is numeric in nature and the interest I have is in how many offers a school ought to expect given the features/independent variables one could supply to the model.

Initially I will determine which regressor algorithm performs best then I will use an ensemble meta-estimator that fits several base regressors, each on the whole dataset then averages the individual predictions to form a final prediction.

Each of the Regressors is used to used to make the first **25** predictions.

Chart, scatter chart

Description automatically generated

* Gradient Boosting Regressor R2: 0.930
* Random Forest Regressor R2: 0.946
* Linear Regression R2: 0.930
* Voting Regressor R2: 0.966

Given the R2 scores are so close, I'll lean towards simplicity rather than running several base regressors and an ensemble to gain just a **3%** gain in explained behavior/R<sup>2</sup> by choosing to do a **linear regression** going forward.

## Predictions & Recommendations

There's a strong correlation between the percent of test takers in a school and the number of offers that that high school received.

One approach to achieving more admissions offers is to send more students to take the SHSAT, based on the average number of level 4 scorers in the school.

Chart, scatter chart

Description automatically generated

The model is based on 530 schools (536 schools with at least 6 SHSAT takers, as SHSAT is unknown for category 0-5 takers. For 6 out of those 536 schools the AvgMark is 0 as a result.

**Recommendation #1**: Top 25 schools that can send more students to the take the test

Below, are the predictions by school regarding the percent of students that could have taken the SHSAT (PerModelDidSHSAT). The PotentialTakers takes the difference between the PerModelDidSHSAT and PerDidSHSAT (ModAgainstDidSHSAT) multiplied by the total number of 8th grade students in each school (count\_of\_students\_hs\_admissions).

Table

Description automatically generated

* The above-referenced schools ought to send more students to take the SHSAT as their average marks may translate into more of their students receiving offers to attend the specialized high schools.
* Increasing the number of test-takers from schools with higher percentages of Black &/or Latinx test-takers will help address the deep disparity of offers being received by White and Asian students.

Another approach is to only send test takers that're high performing students, or those who are level 4.

All possible performance related predictors (mark and level4) are very strongly correlated with each other (multicolinear). Another way to look at what can predict (successful) performance on the SHSAT is to look at percent of offers by student (PctOffersByStudent) that took the SHSAT to the percent of level-4 students in 7th grade (PctScore4).

Chart, scatter chart

Description automatically generated

**Recommendation #2**: Top 25 schools that can increase the percent of their students receiving offers

Here I'm making predictions by school regarding the percent of offers per student according to the model (mod\_offers). Real offers (RealOffers) looks at the modeled percent offers (PctModelOffers) multiplied by the number of SHSAT takers. PotentialOffers takes the difference of RealOffers from actual offers (NumSpecializedOffers).

Table

Description automatically generated

The above table is filtered to exclude schools in eight overperforming districts, detailed below. "Overperforming" means total offers (NumSpecializedOffers) divided by total 8th graders (count\_of\_students\_in\_hs\_admissions).

The table displays schools that should have received at least 10 Extra Offers according to the model in under and average performing districts (listed below).

Table

Description automatically generated

In particular, **P.S. 235 Janice Marie Knight School** & **J.H.S. 118 William W. Niles** are great candidates that should've seen their students receive more **21** and **11** more offers, respectively, for admission to the specialized high schools. Their high percentage of Black &/or Latinx students would've improved the rate at which those ethnicities receive offers across NYC.

## Future Analyses

* An aspect that I wasn't able to explore was using GIS to determine if there are any differences in admissions offers to schools/students based on how close the feeder school is to the specialized high school.
* I was only able to use one year of admissions and test data. It would have been interesting to determine if there're any trends in the data across more than one year of data.
* The data only contained the performance on tests that are administered to students during a typical school year.
  + It would be interesting to see how preparatory tests for the SHSAT relate to the number of offers received by schools/students.
  + Also, looking at any after-school prep programs' impact on the number of offers received, would be interesting.

## Credits

* Project structure based on [cookiecutter data science project template](https://drivendata.github.io/cookiecutter-data-science/)