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 (SHS) 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.

Chart, bar chart

Description automatically generated

Of major concern is the racial & ethnic breakdown of the students receiving SHS offers. Black & Latinx students are 68% of the NYC high school population, but only 9% of the SHS offers[[1]](#footnote-1).

## Problem Statement

Which factors that predict success on the SHSAT can increase the number of SHS admissions offers received by Black & 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 by school, 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 receiving SHS offers.

An initial assumption is that those students who do well on English Language Arts (ELA) & Math standardized tests, will similarly perform well on the SHSAT. We'll investigate this and extrapolate as to whether this is the case across all schools/students.

Because my response variable is continuous, I’ll be using linear regression models to determine how many SHS offers schools that fit a certain testing/aptitude standard could be getting based on their standardized test (ELA + Math) 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. In addition for models to perform best, improvements to data dimensionality were implemented, and several summary features were created.

For example, the *2016 School Explorer* has 20 variables with ELA & Math testing information on just the 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

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

These columns were summarized into several other features that I’ve detailed in the Feature Engineering section below.

We can also 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 |
|  |  |

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.

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.

Graphical user interface, text

Description automatically generated

In the 2017-2018 SHSAT Admissions Test Offers By Sending School (Offers) we see that **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.

**Feature Engineering**

In order to better summarize the schools/students into ranges & to allow the models to predict low-dimension data, I've added the following summary 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

**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

The initial test of our assumption that high performers on the ELA & Math tests translates to SHSAT success used a heatmap:

Graphical user interface

Description automatically generated

Our assumption is proven true (AVgScore4 vs NumSpecializedOffers), preliminarily, but even more interesting is that the number of SHSAT test takers (NumTestTakers) is also strongly correlated with the SHS offers!

Let’s look at another plot to get a better idea of where the values are falling:

A picture containing diagram

Description automatically generated

We can see again the positive relationship between ELA & Math test scores (PctScore4ELA & PctScore4Math) and the number of SHSAT test takers (NumTestTakers).

Here is a zoomed in version:

Chart, scatter chart

Description automatically generated

If the assumption holds, that those students/schools that do well on ELA & Math (level 4 scores) will also perform well on the SHSAT, then are Black & Latinx students not in possession of those kinds of scores? Does that hint as to the fewer SHS offers?

### ELA & Math Testing Scores

To investigate this, the following plots dig into the racial and ethnic breakdowns of ELA & Math testing scores, in addition to the number of SHSAT test takers.

This Math testing score breakdown shows significantly less Black & Latinx students scoring 4s (best) on this standardized test.

This ELA breakdown also shows significantly less Black & Latinx students scoring 4s (best).

Here are the ELA & Math test scores side by side.

Chart, bar chart

Description automatically generated

With the strength of the relationship between ELA & Math 4 scores and SHS offers, we can potentially see why Black or Latinx students are generally receiving less SHS offers.

### SHS Offers

Looking at the number of SHS offers we can further depict the disparity of Black & Latinx students

Among the schools that received the **highest** number of SHS offers, nearly all are composed of fewer than 30% Black or Latinx students.

Among schools with the **highest** percentage of their test takers/students receiving SHS offers, **all** are composed of less than 30% Black or Latinx students.

Among the schools where the **least** percentage of their students received SHS offers, **nearly all** are composed of 60% or more Black or Latinx students.

### SHSAT Test Takers

Looking at relationship between SHS offers and the number of SHSAT test takers, we can explore the relationship between the Black & Latinx students and their representation at the SHSAT test.

Most of the schools with the **highest** number of SHSAT test takers are composed of less than 30% Black or Latinx students.

Nearly all schools with the **least** number of SHSAT test takers are composed of 80% or more Black or Latinx students.

## Models & Evaluation

My intent is to use regression-based models because my response variables are continuous 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 averaged several base estimators.

Each of the base estimators is using the first 25 schools, as my training & testing sample, to make predictions on the number of SHS offers a school received.

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 estimators and an ensemble to gain a mere **3%** in explained behavior/R2 by choosing to do a **linear regression** going forward.

## Predictions & Recommendations

As we saw earlier, there was a strong correlation between the number of SHSAT test takers and the number of SHS offers. By that logic, a school should just send *all* of their eligible students (8th graders), right? That probably wouldn’t be a good idea.

The best idea is to send more of your students that’re level 4 scorers on the ELA & Math tests. I explored the relationship between the level 4 students and the percentage of them that took the SHSAT.

Chart, scatter chart

Description automatically generated

A higher testing score means a higher percentage of students (in a school) took the SHSAT.

Based on the ELA & Math test scores, which schools could’ve sent more students to take the SHSAT?

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.

**Table

Description automatically generated**

Here we see a relatively strong relationship (R2 = 0.820) between the percent of the school that took the SHSAT & the ELA/Math scores.

### **Recommendation #1**: Top 25 schools that can send more students (level 4) 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 to improving the number Black or Latinx students receiving SHS offers is to look at the likelihood of SHS offers given a high performing student (ELA & Math 4 scorers).

I’m predicting (successful) performance on the SHSAT by looking at the percent of offers among test takers (PctOffersByStudent) against the percent of level-4 students in 7th grade (PctScore4).

Chart, scatter chart

Description automatically generated

Here we see that the higher the percentage of ELA/Math 4 scorers you send to take SHSAT, the more of them receive SHS offers.

Based on the ELA & Math testing scores, which schools could’ve seen more SHS offers?

Table

Description automatically generated

Here we can see a very strong relationship (R2 = 0.914) between those high performing students and SHS offers.

### **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 **NOT** 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 by sending their level 4 students, 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 SHS. Their high percentage of Black or Latinx students would've improved the rate at which those ethnicities receive offers.

## 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/)

1. Exam Schools: Inside America’s Most Selective Public High Schools by Finn and Hockett [↑](#footnote-ref-1)