

Springboard Data Science Career Track
Capstone Project

‘Right Place Right Time’
PASSNYC: Data Science for Good Challenge
Manju Nair (2019)

Contents

Introduction	3
Overview of Data Set	4
Data Wrangling	4
Exploratory Data Analysis	5
1. SHSAT offer Ratio of Black- Hispanic Population	6
2. SHSAT offer Distribution of White - Asian Population	6
3. SHSAT offer Distribution of ELA 4s Students.....	7
4. SHSAT offer Distribution of Math 4s Students.....	8
5. ENI Distribution Feeder and Non-feeder schools	9
6. Black-Hispanic Distribution in Feeder and Non-feeder schools	9
7. Scatterplot of ENI and Bl- Hi Distribution	10
Machine Learning.....	11
Approach for Data Modeling	11
Data Modeling.....	12
Feature Importance	12
Results and Recommendations.....	13
Training Required or not List	13
School list with option to display Training on OpenStreetMap.....	13
Awareness Sessions required List	14
School list with option to display Training on OpenStreetMap.....	14
Future Work	15

Introduction

To gain entry to New York City's eight ultracompetitive specialized public high schools (SPHS), eighth graders must take a common entrance exam (SHSAT). Out of about 600 public middle schools, just 10 schools account for more than 25 percent of the seats filled in, according to Education Department data. Though the larger body of NYC's school system is two-thirds Hispanic and Black, it is interesting to note that these 10 middle schools are disproportionately Asian and White. Also worth noting is the fact that at the top 10 schools, 70% of students take SHSAT**, whereas on an average in any other middle school in NYC, 35% of students take SHSAT**. There has been demands from all walks of society to build pressure on the city's governing council to take steps to increase the diversity of students gaining admission to these high schools.

PASSNYC is a not for profit organization that aims to increase the diversity of students taking and qualifying in SHSAT. With its 'Data Science for Good Challenge', PASSNYC aims to use the power of data science to increase the effectiveness of its current '**Outreach**' program, which identifies suitable schools and coaches its eligible students from economically and socially backward backgrounds, enabling them to qualify in SHSAT. It also focuses efforts in popularizing 'SHSAT' and 'SPHS' in schools located in under-performing areas that are historically underrepresented in SHSAT registration.

Right Place Right Time' project solves this challenge by

1. Analyzing the existing publicly available relevant data sets to observe significant patterns and correlations amongst the various features,
2. Using ensemble methods to find the most important features influencing the selection criteria for candidates in SHSAT the most and
3. Designing a robust machine learning model to successfully predict separate lists of schools requiring training and awareness creation sessions, that PASSNYC can focus on for maximum impact.

Overview of Data Set

- PASSNYC has shared 2016 School Explorer data which has all the demographic and other relevant data about the NYC schools, contained in 1272 Rows and x 161columns
- 2013_-_2018_Demographic_Snapshot_School which give demographic details from the past 5 years
- SHSAT_Admissions_Test_Offers_By_Sending_School for three academic years that contains details like number of test takers, no of offers received and number of students in class for each of the schools

Data Wrangling

In order to aid PASSNYC, the data sets have to be explored in detail to check whether any trends or patterns can be observed that would help identify the schools that need help

1. Cleaning and Consolidating the Data

In order to consolidate PASSNYC data from different years (and different file formats), column headers were standardized. Some unnecessary attributes were dropped to reduce dimensionality. A couple of new columns were introduced, to calculate scores based on values of some feature variables, which would eventually become the labels for the Machine Learning Algorithm. For some columns, the value was categorization with words like 'Meeting Target', Approaching Target, and 'Exceeded Target'. That was converted to numeric ranking '1','2','3'.

2. Missing Values

Some columns in the dataset had missing values and there were a few inconsistencies in notation that were adjusted for ease of future analysis.

Dropna and fillna were used to drop if the number of rows were insignificant to the data and to replace using mean value as applicable for each of the specific case.

3. Outliers

There were not much significant outliers to be worked upon

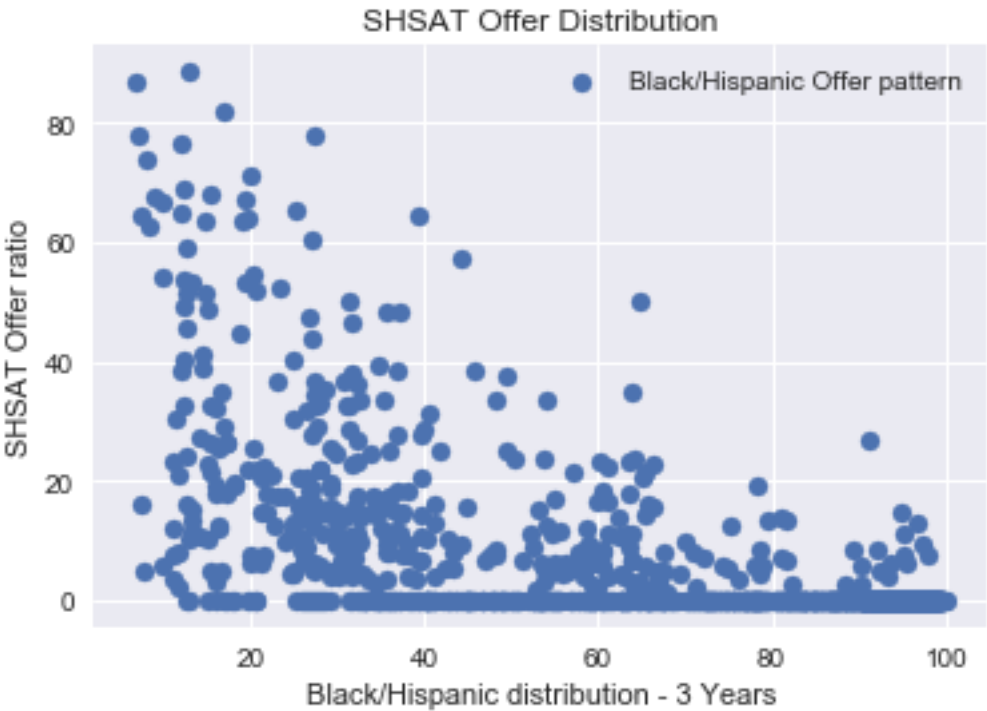
Exploratory Data Analysis

'School Explorer' data frame has a huge set of feature variables and doing a thorough EDA is key to finding out , which of this have the maximum impact on the middle school's getting their students admitted to the Specialized High Schools(SPHS). The task is in finding out which feature variables are key in helping shift focus to the right set of schools.

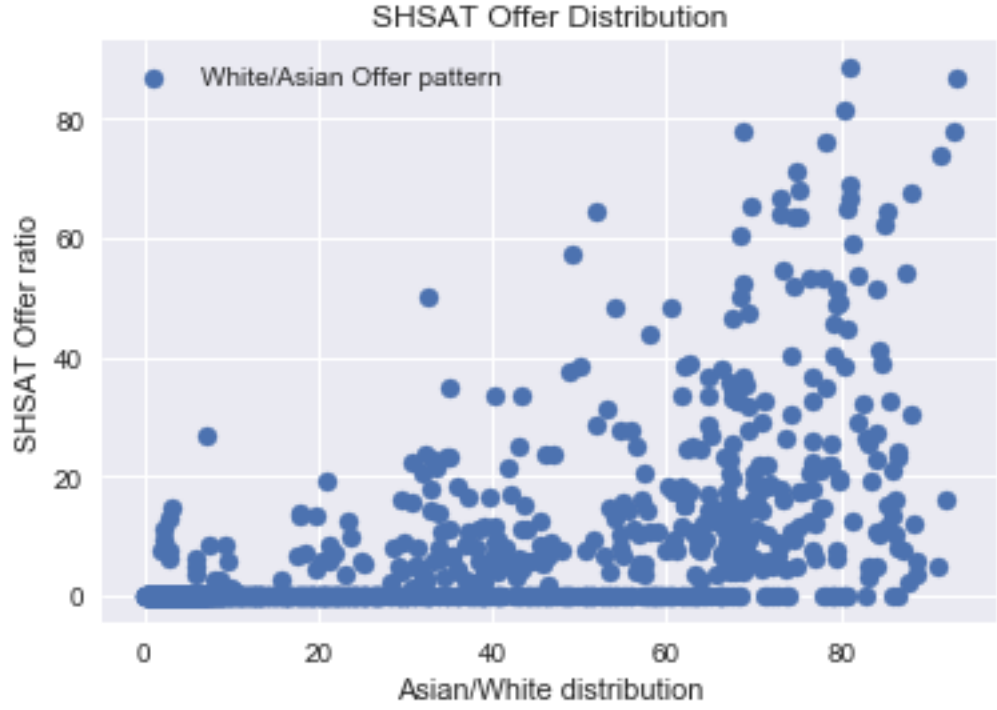
It makes sense to plot a scatter plot between the SHSAT offer ratio against some of the feature variables over the three years to see if there is any co-relation

Given below are some of the scatter plots that I generated. (Please refer the ipynb file for EDA for a comprehensive set)

1. SHSAT offer Ratio of Black- Hispanic Population

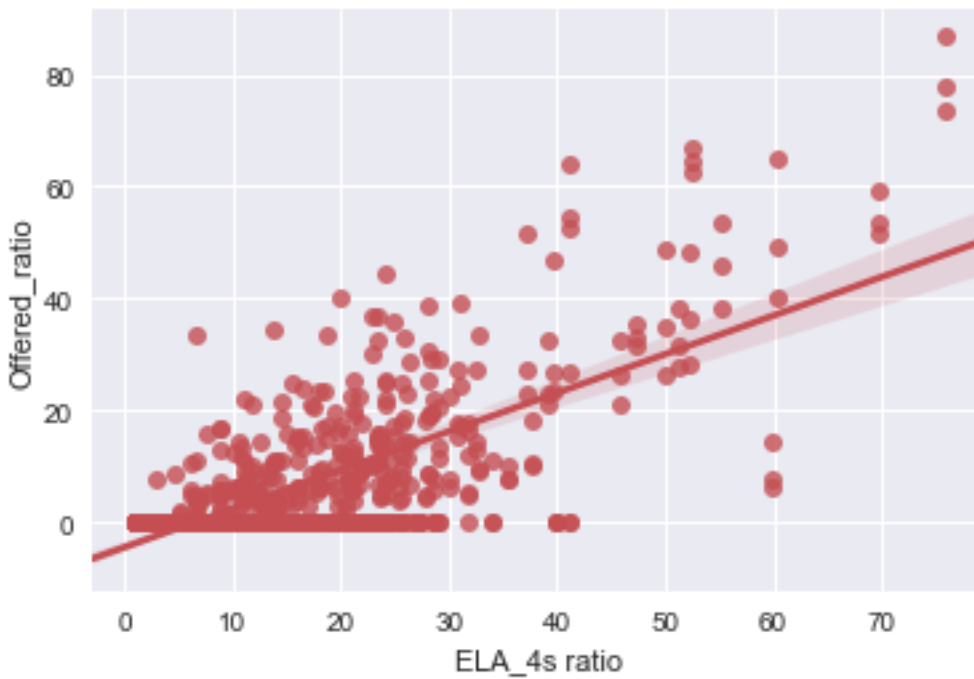


2. SHSAT offer Distribution of White - Asian Population

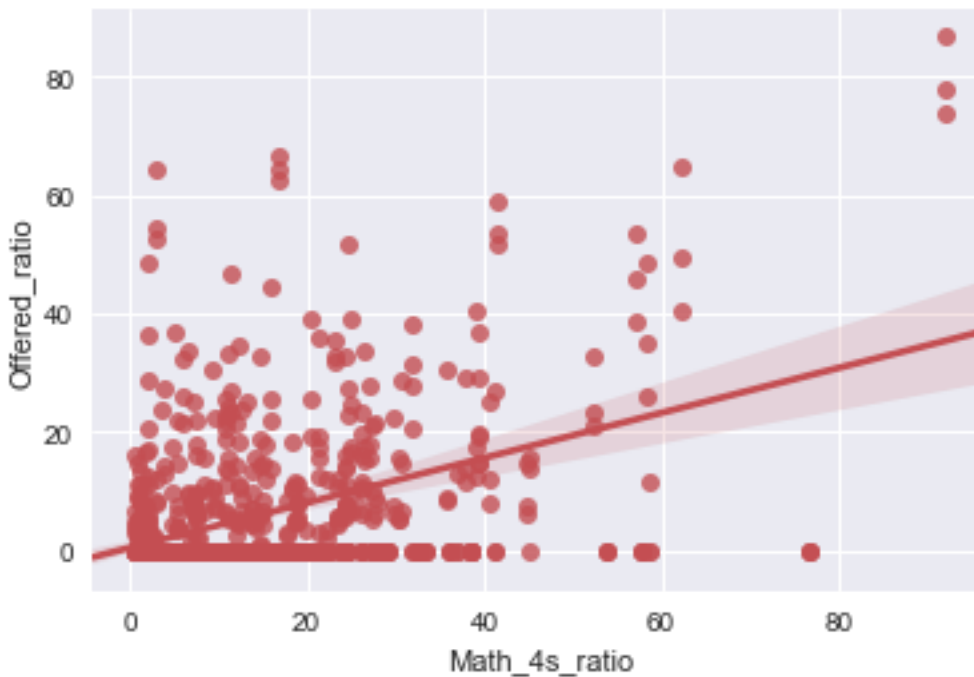


- The Offer ratio plotted against both the population sets also have a correlation coefficient of around 60% only.
- However it is interesting to note that the offer ratio goes up as the White- Asian population in schools go up, whereas, for the Black Hispanic population, with the increase in population, the offer rate was coming down

3. SHSAT offer Distribution of ELA 4s Students



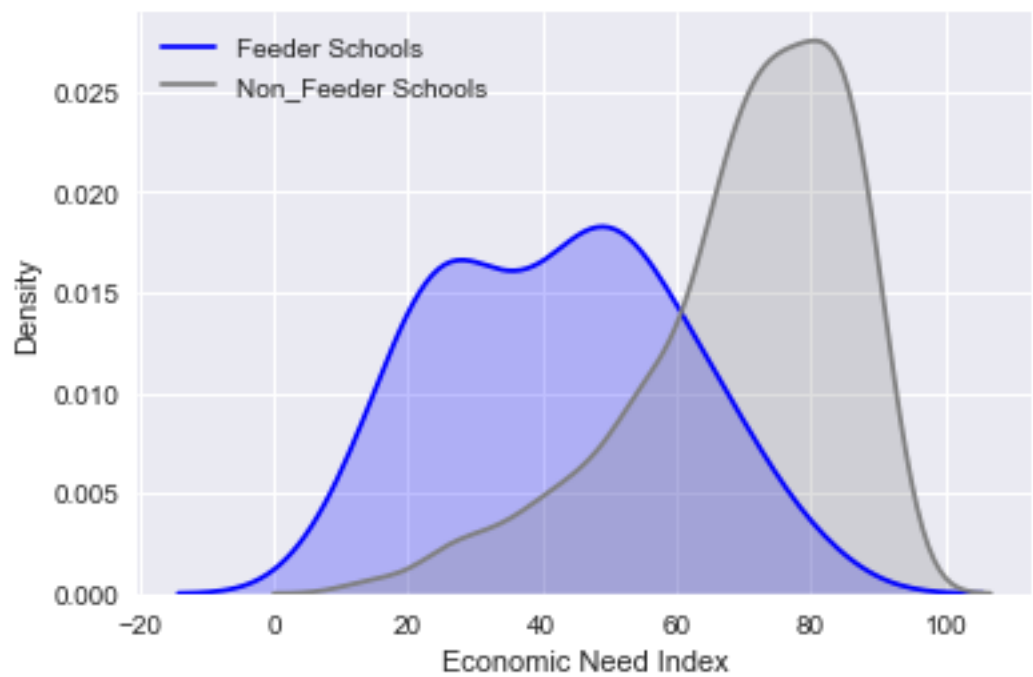
4. SHSAT offer Distribution of Math 4s Students



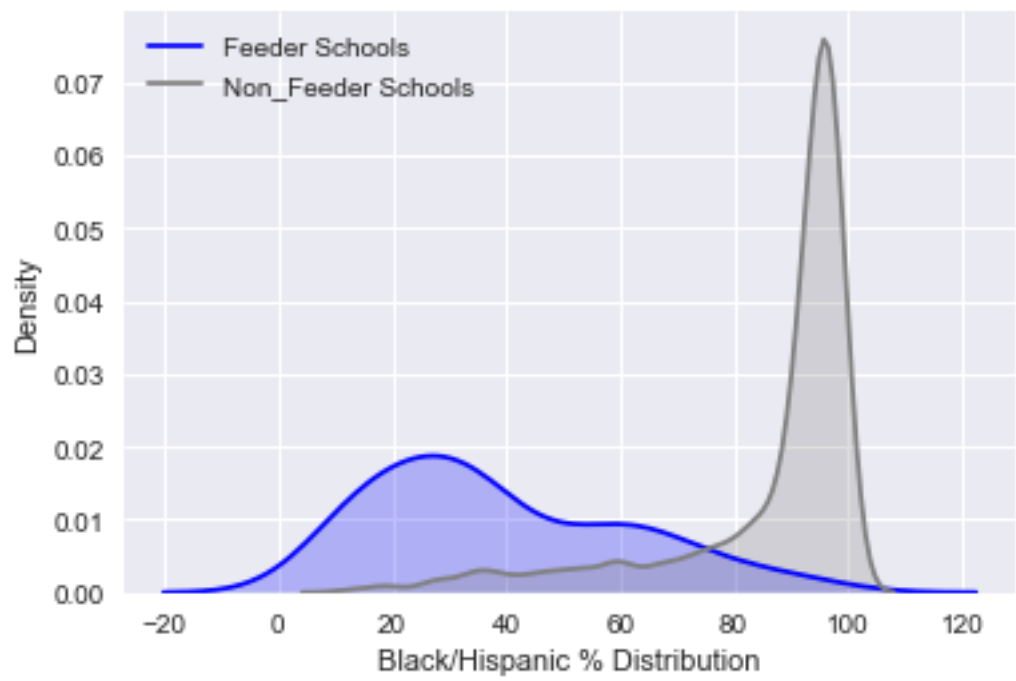
- The Offer ratio plotted against both the ELA 4s scores and Math 4s scores of all the middle schools
- It can be noted from the scatterplot, the ELA 4s and Math 4s scores show a somewhat substantial correlation with the Offers received from SPHS.

Schools are categorized into Feeder Schools (schools that have students who qualify the SHSAT exams) and Non Feeder Schools (school where students do not qualify in the test) and plot density graphs with a couple of feature demographics, to see if there is any correlation.

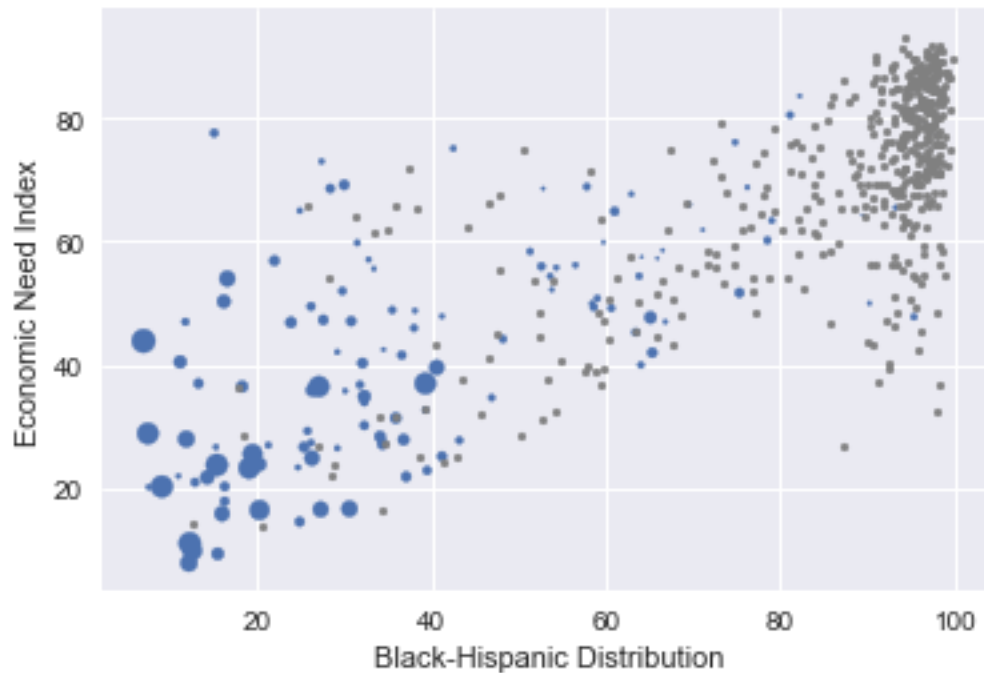
5. ENI Distribution Feeder and Non-feeder schools



6. Black-Hispanic Distribution in Feeder and Non-feeder schools



7. Scatterplot of ENI and Bl- Hi Distribution



It can be seen that

- Feeder and non-feeder schools are strikingly different in terms of their racial composition and economic need level
- The scatterplot above visualizes the positive correlation ($r = 0.78$) between ENI and % Black and Hispanic students.
- Feeder schools (larger-sized blue points) tend to have low-to-medium ENI and lower proportion of Black or Hispanic students, while non-feeder schools (gray points) cluster around the upper right corner of the plot

Machine Learning

Approach for Data Modeling

Main objective of the program is to identify schools

1. with academically qualified students, which would benefit from training programs. Currently, at least some students in these schools are taking SHSAT but without good results.(list1)
2. where the awareness of SHSAT is less and there are not many test takers. In such schools PASSNYC can do road shows and immersion programs to popularize the program and its benefits.(list2)

To rank the schools, I calculate two scores for all schools,

$$\text{Need_Training_Score} = \frac{\text{Number of offers received in the school}}{\text{Total number of students in the school}},$$

with a lower score indicating a training requirement for the school.

$$\text{Need_Awareness_Score} = \frac{\text{Number of test takers in the school}}{\text{Total number of students in the sch}}$$

Need_Training_Score and Need_Awareness_Score are later classified into '0' or '1' based on score value to make the target variable a classifier. Their classification is done based on the values of a set of feature variables in the data set.

This dataset has close to 50 feature variables and plotting a correlation for each of these variables would be impractical.

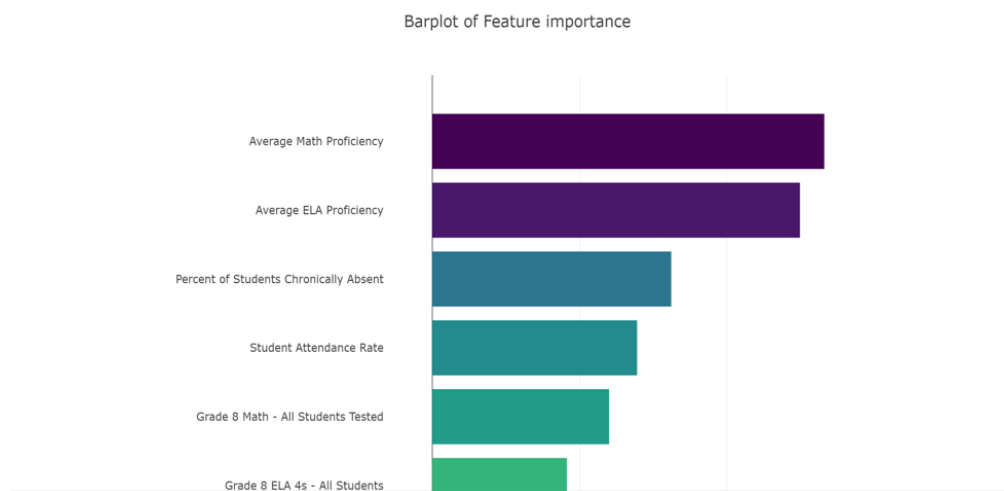
This is where an ensemble method like Random Forest can help. Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is very simple to use but at the same time very effective.

Data Modeling

Feature Importance

With random forest algorithm, it is very easy to measure the relative importance of each feature on the prediction. It measures feature importance by looking at how much the tree nodes, which use that feature, reduce impurity across all trees in the forest. It computes this score automatically for each feature after training and scales the results, so that the sum of all importance is equal to 1.

From the feature importance score, we can decide which features to drop, and which ones to keep.



I used Random Forest Classifier which is a very popular algorithm to solve classification problems to solve this problem. I reserve a test data size of 25% and train the model on 75%. It gives a Mean Absolute Error of 0.08 which is very minimal and the cross validation score is 0.9, which ensures a good prediction, as the model ensures a 91% accuracy in prediction in the list of schools requiring training.

For the awareness requirement list, Mean Absolute Error was 0.13 which is very minimal and the cross validation score is again 0.91, which ensures a good prediction, as the model ensures a 91% accuracy in prediction.

Results and Recommendations

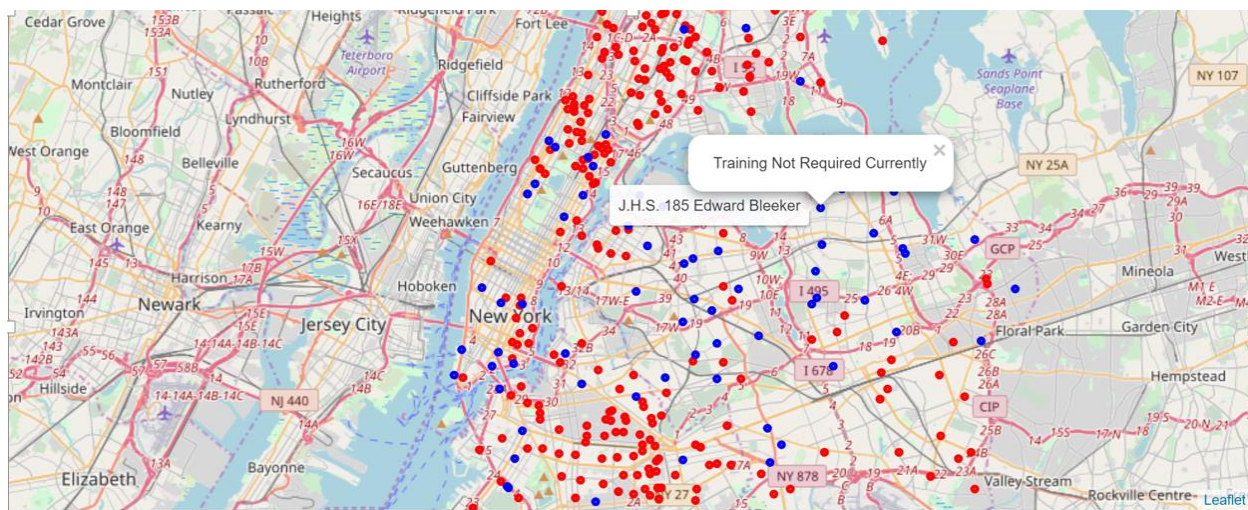
Using this prediction model, we can predict whether any public middle school in NYC would make it to the list of PASSNYC's training_required or awareness_required lists, if we have schools' students' performance data and other feature data.

Below, I have taken a new data set (schools_testdata_df) that has not been used in training, for prediction.

Training required or not list

	School_List	Latitude	Longitude	Training_Aid_Reqd	comments
404	J.H.S. 226 Virgil I. Grissom	40.675063	-73.816692	0	Training Required
405	P.S. 232 Lindenwood	40.665850	-73.851027	0	Training Required
406	Channel View School for Research	0.000000	0.000000	0	Training Required
434	Pathways College Preparatory School: A College...	0.000000	0.000000	0	Training Required
407	Knowledge and Power Preparatory Academy VI	40.601532	-73.763979	0	Training Required
409	Academy of Medical Technology: A College Board...	0.000000	0.000000	0	Training Required
410	Waterside School For Leadership	40.579602	-73.831079	0	Training Required
411	Village Academy	40.603951	-73.749868	0	Training Required
412	Scholars' Academy	0.000000	0.000000	0	Training Required
...
363	P.S. 102 Bayview	40.733420	-73.877750	1	Training Not Required Currently
365	I.S. 119 The Glendale	40.705198	-73.875033	1	Training Not Required Currently
366	I.S. 125 Thom J. McCann Woodside	40.741091	-73.918649	1	Training Not Required Currently
356	I.S. 5 - The Walter Crowley Intermediate School	40.737757	-73.887429	1	Training Not Required Currently
418	J.H.S. 190 Russell Sage	40.723014	-73.851845	1	Training Not Required Currently

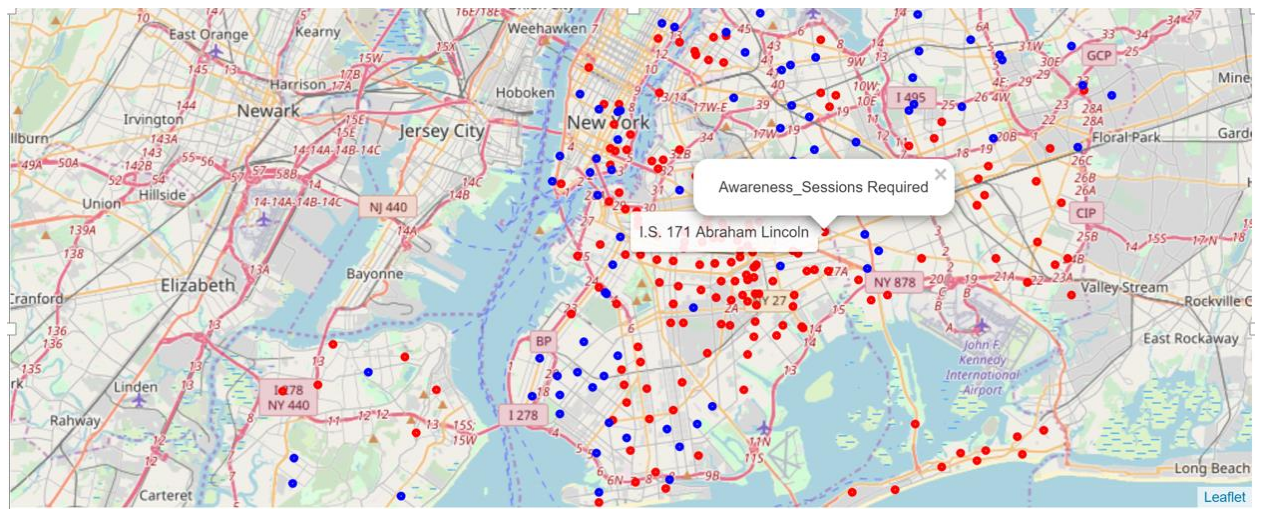
School list with option to display option on OpenStreetMap



Awareness sessions required or not list

	School_List	Latitude	Longitude	Awareness_Sessions_Reqd	Comments
0	P.S. 034 Franklin D. Roosevelt	40.726147	-73.975043	0	Awareness_Sessions Required
434	Pathways College Preparatory School: A College...	0.000000	0.000000	0	Awareness_Sessions Required
433	I.S. 238 - Susan B. Anthony Academy	40.712772	-73.780173	0	Awareness_Sessions Required
432	P.S./I.S. 208	40.744184	-73.727581	0	Awareness_Sessions Required
243	New Voices School of Academic & Creative Arts	40.660969	-73.989018	1	Awareness_Sessions Not Required Currently
244	The Math & Science Exploratory School	40.683835	-73.980355	1	Awareness_Sessions Not Required Currently
419	J.H.S. 217 Robert A. Van Wyck	40.710509	-73.811875	1	Awareness_Sessions Not Required Currently
398	M.S. 137 America's School of Heroes	40.678070	-73.839652	1	Awareness_Sessions Not Required Currently
386	M.S. 158 Marie Curie	40.756375	-73.772320	1	Awareness_Sessions Not Required Currently

School list with option to display option on OpenStreetMap



Future Work

The list of schools can be further categorized, based on

- geography/ Burroughs of NYC
- based on the average school income
- ELA and math scores and ranking given or grouped up ,

so that training plans can be tailored according to the needs of each of these groups.

APPENDIX

List of feature variables:

Economic Need Index	float64
School Income Estimate	object
Percent ELL	object
Percent Asian	object
Percent Black	object
Percent Hispanic	object
Percent Black / Hispanic	object
Percent White	object
Student Attendance Rate	object
Percent of Students Chronically Absent	object
Rigorous Instruction %	object
Rigorous Instruction Rating	object
Collaborative Teachers %	object
Collaborative Teachers Rating	object
Supportive Environment %	object
Supportive Environment Rating	object
Effective School Leadership %	object
Effective School Leadership Rating	object
Strong Family-Community Ties %	object
Strong Family-Community Ties Rating	object
Trust %	object
Trust Rating	object
Student Achievement Rating	object
Average ELA Proficiency	float64
Average Math Proficiency	float64
Grade 3 ELA - All Students Tested	int64
Grade 3 ELA 4s - All Students	int64
Grade 3 ELA 4s - American Indian or Alaska Native	int64
Grade 3 ELA 4s - Black or African American	int64
Grade 3 ELA 4s - Hispanic or Latino	int64
Grade 3 ELA 4s - Asian or Pacific Islander	int64
Grade 3 ELA 4s - White	int64
Grade 3 ELA 4s - Multiracial	int64
Grade 3 ELA 4s - Limited English Proficient	int64
Grade 3 ELA 4s - Economically Disadvantaged	int64
Grade 3 Math - All Students tested	int64
Grade 3 Math 4s - All Students	int64
Grade 3 Math 4s - American Indian or Alaska Native	int64
Grade 3 Math 4s - Black or African American	int64
Grade 3 Math 4s - Hispanic or Latino	int64
Grade 3 Math 4s - Asian or Pacific Islander	int64

Grade 3 Math 4s - White	int64
Grade 3 Math 4s - Multiracial	int64
Grade 3 Math 4s - Limited English Proficient	int64
Grade 3 Math 4s - Economically Disadvantaged	int64
Grade 4 ELA - All Students Tested	int64
Grade 4 ELA 4s - All Students	int64
Grade 4 ELA 4s - American Indian or Alaska Native	int64
Grade 4 ELA 4s - Black or African American	int64
Grade 4 ELA 4s - Hispanic or Latino	int64
Grade 4 ELA 4s - Asian or Pacific Islander	int64
Grade 4 ELA 4s - White	int64
Grade 4 ELA 4s - Multiracial	int64
Grade 4 ELA 4s - Limited English Proficient	int64
Grade 4 ELA 4s - Economically Disadvantaged	int64
Grade 4 Math - All Students Tested	int64
Grade 4 Math 4s - All Students	int64
Grade 4 Math 4s - American Indian or Alaska Native	int64
Grade 4 Math 4s - Black or African American	int64
<i>Grade 4 Math 4s – Hispanic or Latino</i>	int64
Grade 4 Math 4s - Asian or Pacific Islander	int64
Grade 4 Math 4s - White	int64
Grade 4 Math 4s - Multiracial	int64
Grade 4 Math 4s - Limited English Proficient	int64
Grade 4 Math 4s - Economically Disadvantaged	int64
Grade 5 ELA - All Students Tested	int64
Grade 5 ELA 4s - All Students	int64
Grade 5 ELA 4s - American Indian or Alaska Native	int64
Grade 5 ELA 4s - Black or African American	int64
Grade 5 ELA 4s - Hispanic or Latino	int64
Grade 5 ELA 4s - Asian or Pacific Islander	int64
Grade 5 ELA 4s - White	int64
Grade 5 ELA 4s - Multiracial	int64
Grade 5 ELA 4s - Limited English Proficient	int64
Grade 5 ELA 4s - Economically Disadvantaged	int64
Grade 5 Math - All Students Tested	int64
Grade 5 Math 4s - All Students	int64
Grade 5 Math 4s - American Indian or Alaska Native	int64
Grade 5 Math 4s - Black or African American	int64
Grade 5 Math 4s - Hispanic or Latino	int64
Grade 5 Math 4s - Asian or Pacific Islander	int64
Grade 5 Math 4s - White	int64
Grade 5 Math 4s - Multiracial	int64
Grade 5 Math 4s - Limited English Proficient	int64
Grade 5 Math 4s - Economically Disadvantaged	int64
Grade 6 ELA - All Students Tested	int64
Grade 6 ELA 4s - All Students	int64
Grade 6 ELA 4s - American Indian or Alaska Native	int64

Grade 6 ELA 4s - Black or African American	int64
Grade 6 ELA 4s - Hispanic or Latino	int64
Grade 6 ELA 4s - Asian or Pacific Islander	int64
Grade 6 ELA 4s - White	int64
Grade 6 ELA 4s - Multiracial	int64
Grade 6 ELA 4s - Limited English Proficient	int64
Grade 6 ELA 4s - Economically Disadvantaged	int64
Grade 6 Math - All Students Tested	int64
Grade 6 Math 4s - All Students	int64
Grade 6 Math 4s - American Indian or Alaska Native	int64
Grade 6 Math 4s - Black or African American	int64
Grade 6 Math 4s - Hispanic or Latino	int64
Grade 6 Math 4s - Asian or Pacific Islander	int64
Grade 6 Math 4s - White	int64
Grade 6 Math 4s - Multiracial	int64
Grade 6 Math 4s - Limited English Proficient	int64
Grade 6 Math 4s - Economically Disadvantaged	int64
Grade 7 ELA - All Students Tested	int64
Grade 7 ELA 4s - All Students	int64
Grade 7 ELA 4s - American Indian or Alaska Native	int64
Grade 7 ELA 4s - Black or African American	int64
Grade 7 ELA 4s - Hispanic or Latino	int64
Grade 7 ELA 4s - Asian or Pacific Islander	int64
Grade 7 ELA 4s - White	int64
Grade 7 ELA 4s - Multiracial	int64
Grade 7 ELA 4s - Limited English Proficient	int64
Grade 7 ELA 4s - Economically Disadvantaged	int64
Grade 7 Math - All Students Tested	int64
Grade 7 Math 4s - All Students	int64
Grade 7 Math 4s - American Indian or Alaska Native	int64
Grade 7 Math 4s - Black or African American	int64
Grade 7 Math 4s - Hispanic or Latino	int64
Grade 7 Math 4s - Asian or Pacific Islander	int64
Grade 7 Math 4s - White	int64
Grade 7 Math 4s - Multiracial	int64
Grade 7 Math 4s - Limited English Proficient	int64
Grade 7 Math 4s - Economically Disadvantaged	int64
Grade 8 ELA - All Students Tested	int64
Grade 8 ELA 4s - All Students	int64
Grade 8 ELA 4s - American Indian or Alaska Native	int64
Grade 8 ELA 4s - Black or African American	int64
Grade 8 ELA 4s - Hispanic or Latino	int64
Grade 8 ELA 4s - Asian or Pacific Islander	int64
Grade 8 ELA 4s - White	int64
Grade 8 ELA 4s - Multiracial	int64
Grade 8 ELA 4s - Limited English Proficient	int64
Grade 8 ELA 4s - Economically Disadvantaged	int64

Grade 8 Math - All Students Tested	int64
Grade 8 Math 4s - All Students	int64
Grade 8 Math 4s - American Indian or Alaska Native	int64
Grade 8 Math 4s - Black or African American	int64
Grade 8 Math 4s - Hispanic or Latino	int64
Grade 8 Math 4s - Asian or Pacific Islander	int64
Grade 8 Math 4s - White	int64
Grade 8 Math 4s - Multiracial	int64
Grade 8 Math 4s - Limited English Proficient	int64
Grade 8 Math 4s - Economically Disadvantaged	int64