‘Right Place Right Time’

CAPSTONE PROJECT : MANJU NAIR

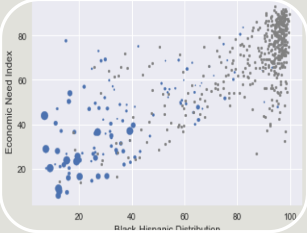
Table of Contents

- Summary
- Context
- PASSNYC’s ‘Data Science For Good Challenge’
- ‘ Right Place Right Time ‘
- Datasets And Data Wrangling
- Exploratory Data Analysis – Findings
- Approach For Data Modeling
- Data Modeling
- Analysis Of Results
- Future Work

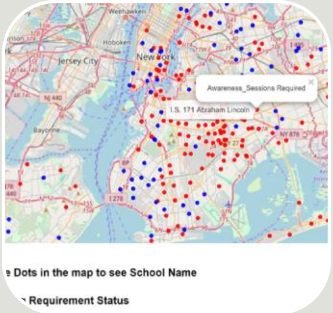
Summary

PassNYC’s ‘Data science for Good challenge’ seeks to use the power of data science to improve its “Outreach " program which enables students from weaker socio-economic backgrounds qualify in SHSAT(exam) and receive placements in specialized high schools(SPHS).

‘Right Place Right Time’ project aims to solve this challenge by deploying in-depth analysis using machine learning techniques , to predict the list of schools that PASSNYC can focus on , for maximum impact with its ‘Outreach’ program.



Findings : Middle schools sending qualifying students to Specialized High Schools(SPHS) are strikingly different in socio-economic distribution , from schools having non-qualifying students



Results:

1.) Furnishes list of middle schools having eligible candidates who will benefit from PASSNYC’s training, as easily locatable plot on a map for ease

2.)Provides list of middle schools where SPHS program and the entrance exam SHSAT can be popularized, to increase the number of test takers from schools, as easily locatable map plots with information display

Future Work : Further categorize list of schools , based on geography/ Burroughs of NYC, average school income, ELA and math scores, to custom-tailor training based on the requirements of each of these groups.

Context

- Entry to New York City’s eight ultracompetitive specialized public high schools(SPHS) for 8th graders is through a common entrance exam(SHSAT).
- Out of 600 public middle schools, just 10 account for more than 25 percent**, of the offers to attend one of these elite schools.
- 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

**Source : Education Department data

PASSNYC‘s ‘Data Science for Good Challenge’

- 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

  - 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

‘Right Place Right Time’ project solves this challenge by

- Analyzing the existing publicly available relevant data sets to observe significant patterns and correlations amongst the various features
- Using ensemble methods to find the most important features influencing the selection criteria for candidates in SHSAT the most
- 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.

Datasets and Data Wrangling

Data sets

- 2016 School Explorer data which has all the demographic 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

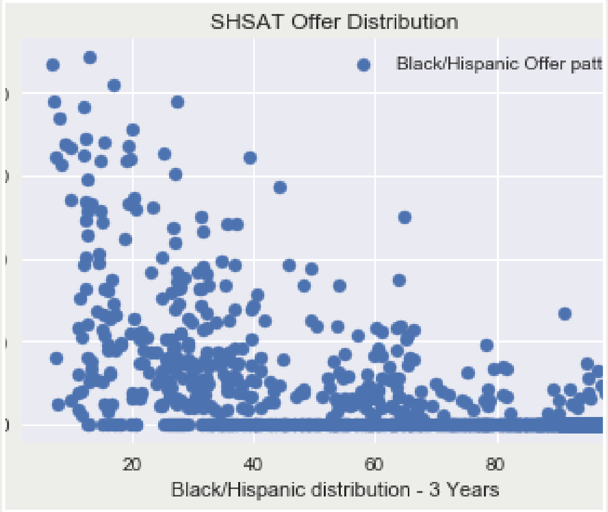
Data Wrangling

- Some unnecessary attributes were dropped to reduce dimensionality and a couple of new columns introduced to calculate scores
- For some columns, the value that were categorizations like ‘ Meeting ’, Approaching and ‘ Exceeded’ ‘Target’ were converted to numeric ranking ‘1’,’2’,’3’.
- Function to strip ‘$’ and ‘%’ in columns

  Fill missing values with mean value where applicable

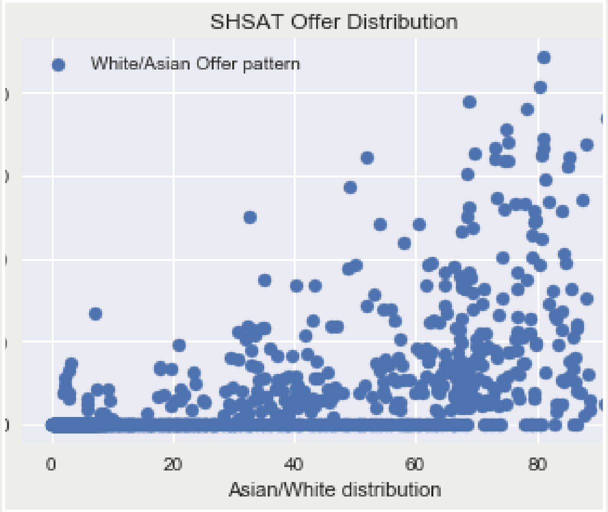
  Dropping NaN values where it was significantly small ore replacing with 0 as applicable

Exploratory Data Analysis – Findings (1/3)



SHSAT offer Ratio

Black- Hispanic Population

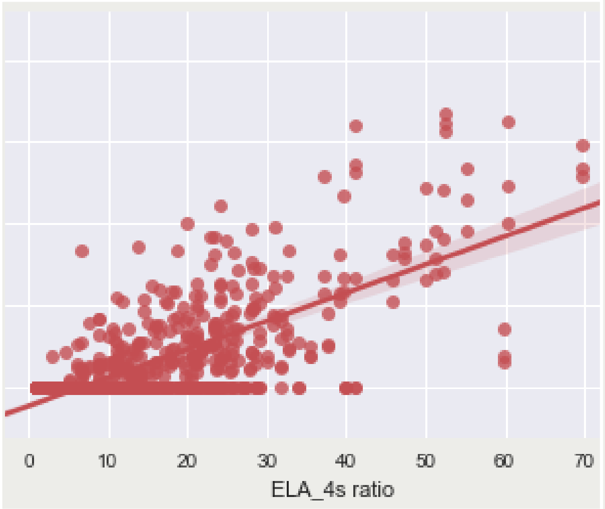


SHSAT offer Distribution

White - Asian Population

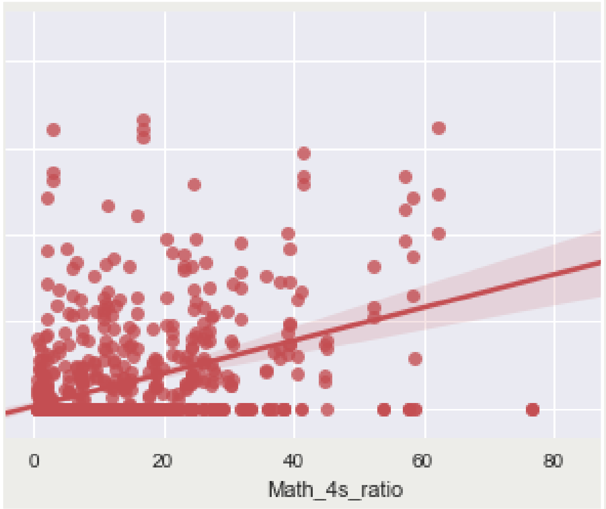
- The Offer ratio plotted against both the population sets also have a correlation coefficient of around 60% only.
- However it is interesting to note that the offer ratio goes up as the White- Asian population in schools go up, whereas, for the Black Hispanic population, with the increase in population,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

Exploratory Data Analysis – Findings (2/3)



SHSAT offer Distribution

ELA 4s Students

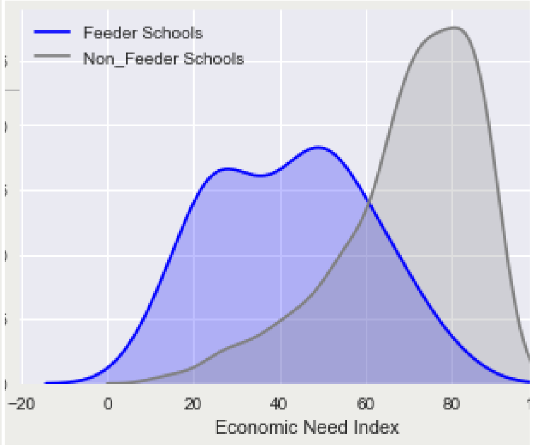


SHSAT offer Distribution

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.

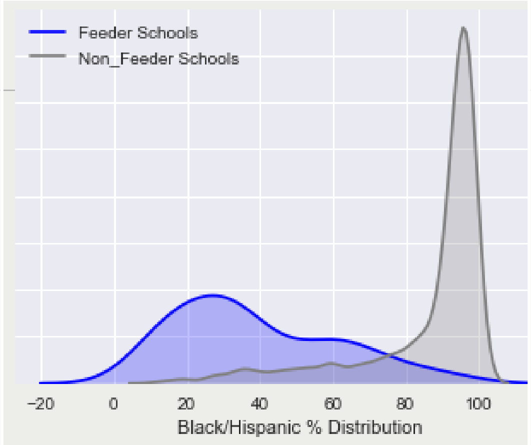
Exploratory Data Analysis – Findings (3/3)



ENI Distribution Feeder** and Non–feeder schools **

Mean ENI of Feeder schools - 42.44

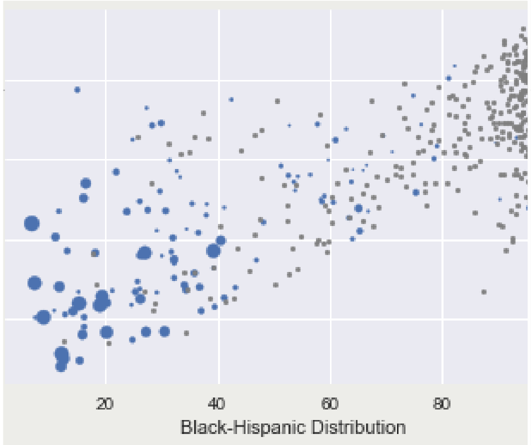
Mean ENI of Non Feeder schools - 69.87



Black-Hispanic Distribution in Feeder and Non-feeder schools

Mean Black-Hispanic Population at Feeder schools - 38.94

Mean Black-Hispanic Population at Non Feeder schools - 86.35



Scatterplot of ENI and Bl- Hi Distribution

Feeder schools have low ENI & Bi-Hi %

Non-feeder schools have high ENI & Bi- Hi %

- 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

** Feeder schools –are the ones with 5 or more students qualifying in SHSAT exams

** Non Feeder schools - and Non-Feeder schools are ones with 0-5 students qualifying in SHSAT

Approach for Data Modeling

Two steps in the program are

1. To identify schools(currently non-feeder schools) with academically qualified students, which would benefit from training programs. Currently, atleast some students in these schools are taking SHSAT but without good results.
2. To identify schools 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.

For all schools,

- Need_Training_Score = 'number of offers received in a school' ***/*** ‘total number of students in the school',

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

- Need_Awareness_Score = 'number of test takers in the school' ***/*** ' total number of students in the school' .

Training_Need_Score and Awareness_Need_Score are later classified into 0 or 1 based on score value to make the target variable a classifier.

Data Modeling

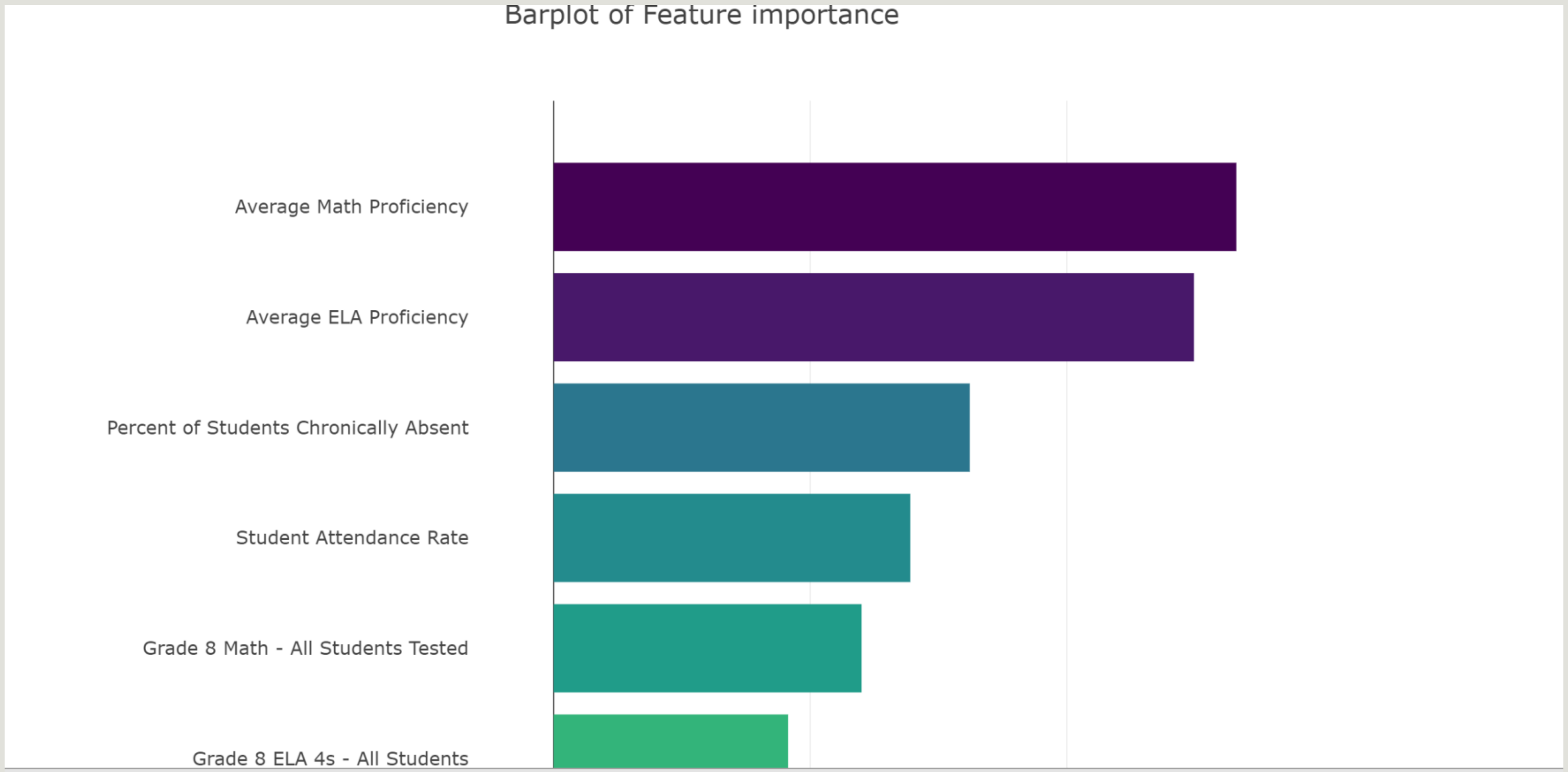
- Using machine learning algorithm to create a model to classify the list of schools into
    1. ones requiring training or not and
    2. ones requiring awareness creation sessions or not

- Random Forest ensemble is used to do Feature Selection -- there are close to 50 feature variables and feature importance can be done and features prioritized

- Random Forest Classifier is used for machine learning and modeling the training data
- The model is then used to predict the result on test data

Data Modeling

Bar plot of feature importance



Data Modeling

Random Forest Classifier

Using the Random Forest Classifier 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.

We can use this Classifier to create two separate list of schools that need help with

1.) training and

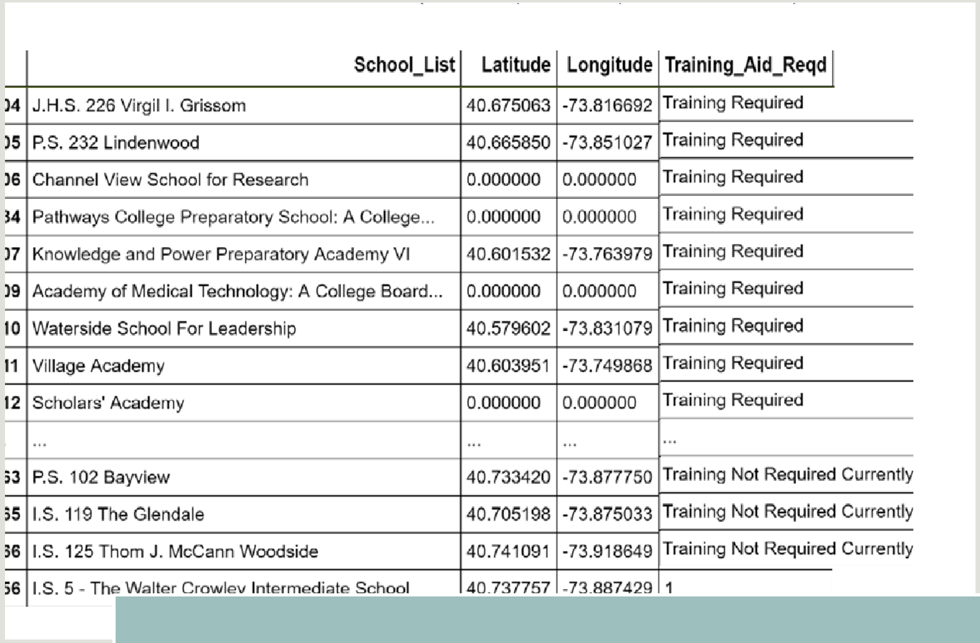
2.) awareness creation

respectively , from the 'School Test' data

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.

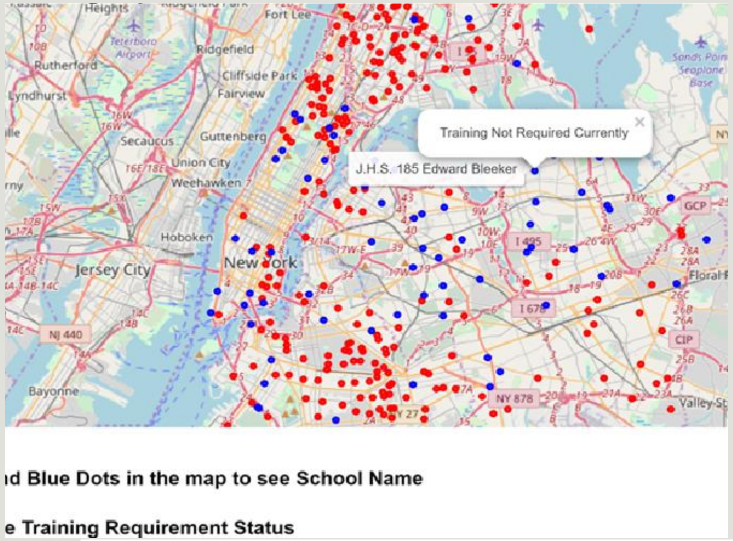
Analysis of Results

Results from prediction on test data (1/2)



	School_List	Latitude	Longitude	Training_Aid_Reqd
04	J.H.S. 226 Virgil I. Grissom	40.675063	-73.816692	Training Required
05	P.S. 232 Lindenwood	40.665850	-73.851027	Training Required
06	Channel View School for Research	0.000000	0.000000	Training Required
34	Pathways College Preparatory School: A College...	0.000000	0.000000	Training Required
07	Knowledge and Power Preparatory Academy VI	40.601532	-73.763979	Training Required
09	Academy of Medical Technology: A College Board...	0.000000	0.000000	Training Required
10	Waterside School For Leadership	40.579602	-73.831079	Training Required
11	Village Academy	40.603951	-73.749868	Training Required
12	Scholars' Academy	0.000000	0.000000	Training Required
	...	...	...	...
63	P.S. 102 Bayview	40.733420	-73.877750	Training Not Required Currently
65	I.S. 119 The Glendale	40.705198	-73.875033	Training Not Required Currently
66	I.S. 125 Thom J. McCann Woodside	40.741091	-73.918649	Training Not Required Currently
56	I.S. 5 - The Walter Crowley Intermediate School	40.737757	-73.887429	1

Training Required or not List



Training Not Required Currently

J.H.S. 185 Edward Bleeker

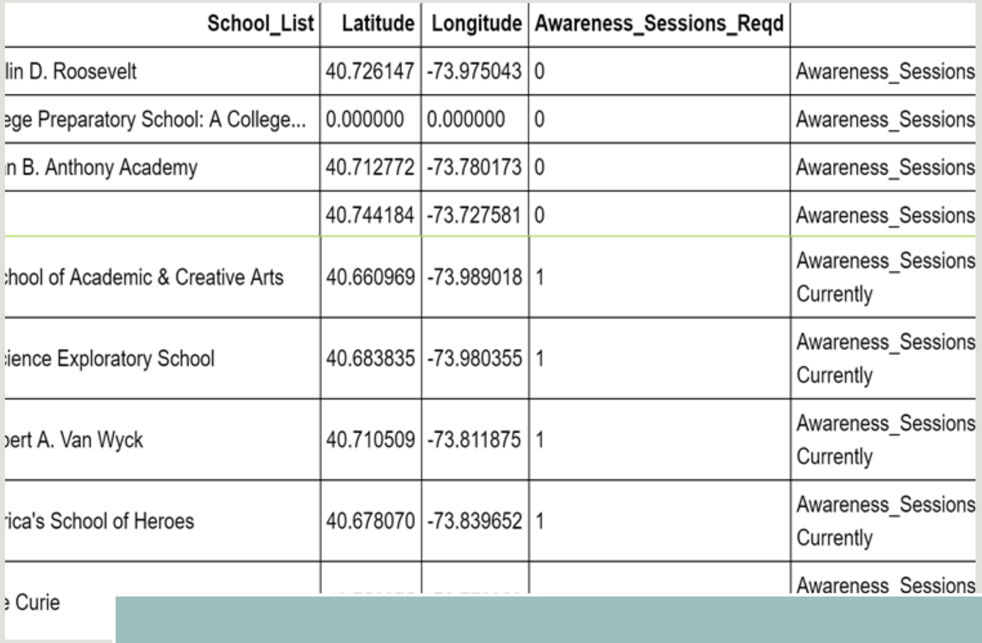
ıd Blue Dots in the map to see School Name

e Training Requirement Status

Map with schools as icons-
information display on hover

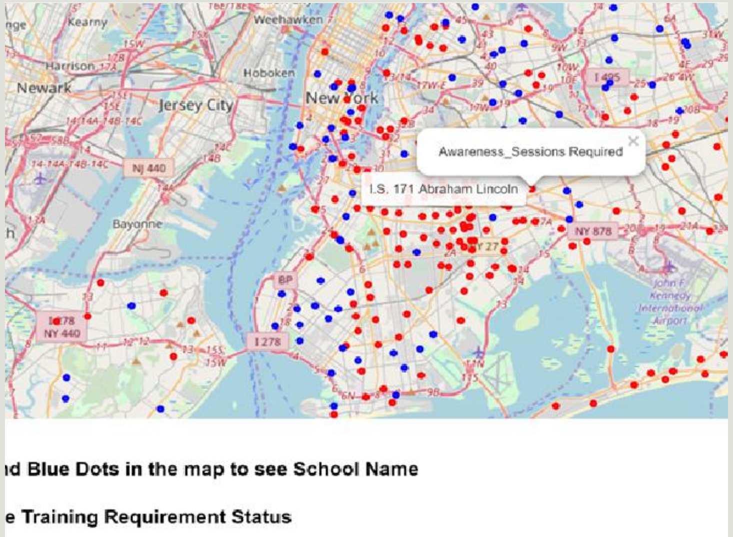
Analysis of Results

Results from prediction on test data (2/2)



School_List	Latitude	Longitude	Awareness_Sessions_Reqd	
lin D. Roosevelt	40.726147	-73.975043	0	Awareness_Sessions
ege Preparatory School: A College...	0.000000	0.000000	0	Awareness_Sessions
n B. Anthony Academy	40.712772	-73.780173	0	Awareness_Sessions
	40.744184	-73.727581	0	Awareness_Sessions
chool of Academic & Creative Arts	40.660969	-73.989018	1	Awareness_Sessions Currently
ience Exploratory School	40.683835	-73.980355	1	Awareness_Sessions Currently
ert A. Van Wyck	40.710509	-73.811875	1	Awareness_Sessions Currently
rica's School of Heroes	40.678070	-73.839652	1	Awareness_Sessions Currently
e Curie	[illegible]	[illegible]	[illegible]	Awareness_Sessions

Awareness session required or not - list



d Blue Dots in the map to see School Name

e Training Requirement Status

Map with schools as icons- information display on hover

Future Work

The list of schools can be further categorized , based on

1. geography/ Burroughs of NYC
2. based on the average school income
3. 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.

Thank You

Acknowledgments

Srdjan Santic
Andrea Constantinof