Capstone 1 PASSNYC Project –Inferential Statistics

PASSNYC uses public data to identify students within New York City’s under-performing school districts and, through consulting and collaboration with partners, aims to increase the diversity of students taking the Specialized High School Admissions Test (SHSAT). By focusing efforts in under-performing areas that are historically underrepresented in SHSAT registration, we will help pave the path to specialized high schools for a more diverse group of students. Datasets that are available for this project are School explorer dataset with demographic and academic details of students in all NYC public schools.

SHSAT data set that has details of the count of students’ schools, the number

that enrolls for tests, count that take the test and the count that gets offer. 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.

'School Explorer' dataframe 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 schools getting their students admitted to the Specialized High Schools(SPHS). Recent studies/ newspaper reports show that the 3 highest-status schools—Stuyvesant, Bronx Science, and Brooklyn Tech in SPHS —have black and Latino student populations of 4, 9, and 13 percent, respectively far below the 70 percent in public schools citywide, which corroborates with PASSNYC's observation . The task is in finding out which feature variables are key in helping shift focus to the right set of schools.

The objective of PASSNYC is to use the data to aid the selection of schools that are underrepresented in these specialized high schools (SPHS). A glance into the demographic distribution of the population in these middle schools show that the underrepresented schools/ hoods have a majority of Black Hispanic Student population, students/ families with a high Economic Need Index, higher crime rate, higher absenteeism etc.

It makes sense to plot a scatter plot between the SHSAT offer ratio and the Black Hispanic Population in middle schools over the three years to see if there is any co-relation.

As part of analysis, multiple scatter plots were done to find the relationship between the above mentioned features and the High School admission offers received for these schools. The correlation coefficient calculation helped in determining which of the feature variables were significant and which could be dropped.

I also used feature importance calculation to order out the features and select the significant features statistically.