Capstone 1 - PASSNYC Project

In-depth Analysis using Machine Learning

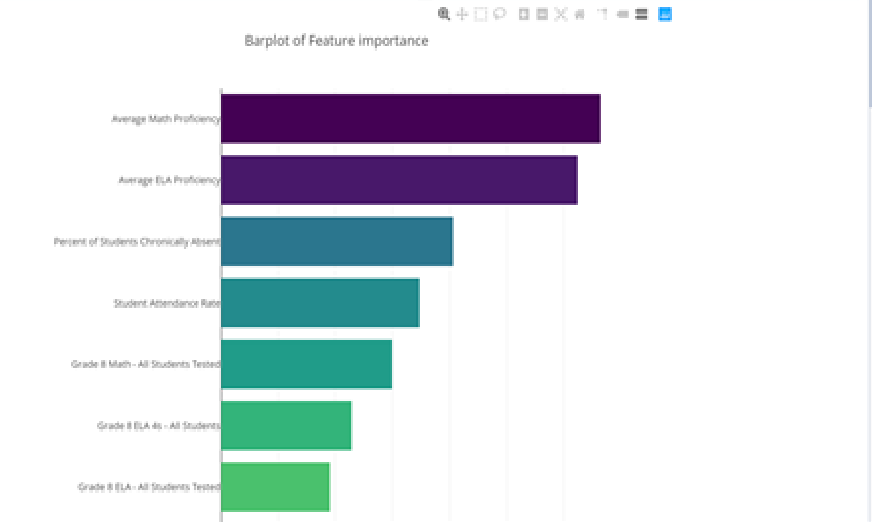
A quick glance at the list of feature variables of PASSNYC dataset shows that there are more than 100 variables. And selecting some of these variables and checking for correlation, showed some patterns , but not significant enough to discard the remaining and focus on these few.

So it is better to rely on Machine Learning techniques to arrive at tht feature variables that are significant and have control in predicting the SHSAT test score for students of these schools.

One option was to use Decision Tree which is a natural way of feature selection. Since tree split is based on maximum gain of gini impurity, trees would automatically split towards more important features.

Another option was to use Random Forest algorithm which is an ensemble method using tree bagging with feature selection at each split.

I used the latter to arrive at the most important features.Below given is my result.



I selected top 20 features from the list I made with feature importance greater than 0.015.

I suggest PASSNYC to take a two pronged approach in achieving their goals.

First would be to identify the list of schools, where there are capable students aware of the SPHS and the SHSAT exam criteria for admission to SPHS. They do take the tests, but they don’t get admissions, as the competition ti very tough and most of the qualifiers benefit from access to training and resources that these students cannot afford. So PASSNYC can focus on training students from such schools.

Second would be to identify schools where there is no awareness amongst student communities about the SPHS and SHSAT. PASSNYC can co-ordinate for running awareness sessions in these schools.