Chronic Absenteeism in New York City Schools

By Nathaniel Green (New York's Premier Education Consultant)

Table of Contents

Objective

The Problem

Since the onset of the pandemic, chronic absenteeism in New York Public Schools has increased by 16%. The New York Department of Education (DOE) is currently looking for ways to improve attendance.

My Solution

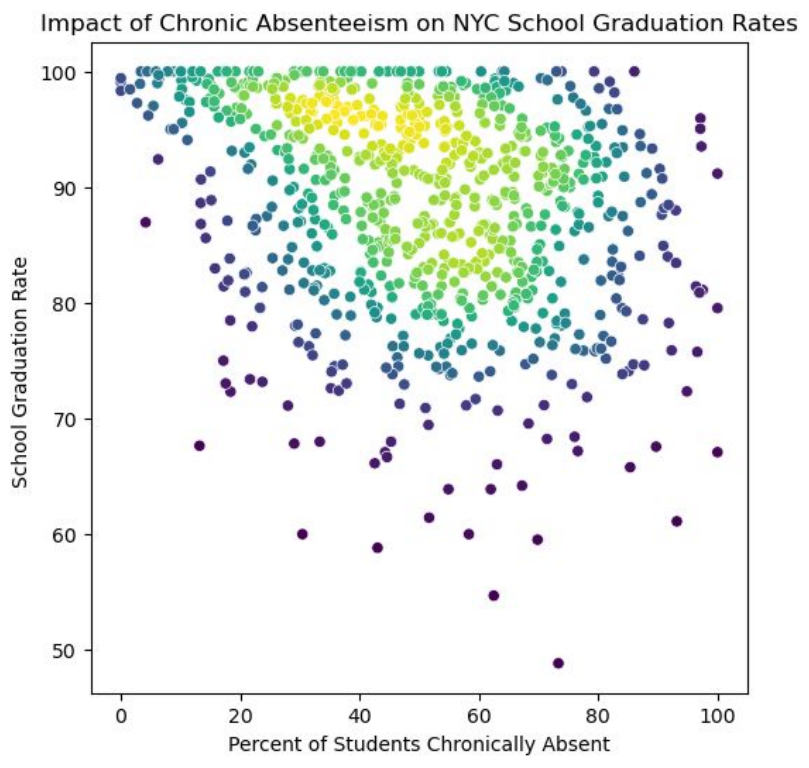
1. Create a model to accurately predict the schools that are most likely to suffer from chronic absenteeism using available data from the DOE.

2. Use the most important features from the model to make recommendations on how the DOE should allocate money to these schools to help lower absenteeism

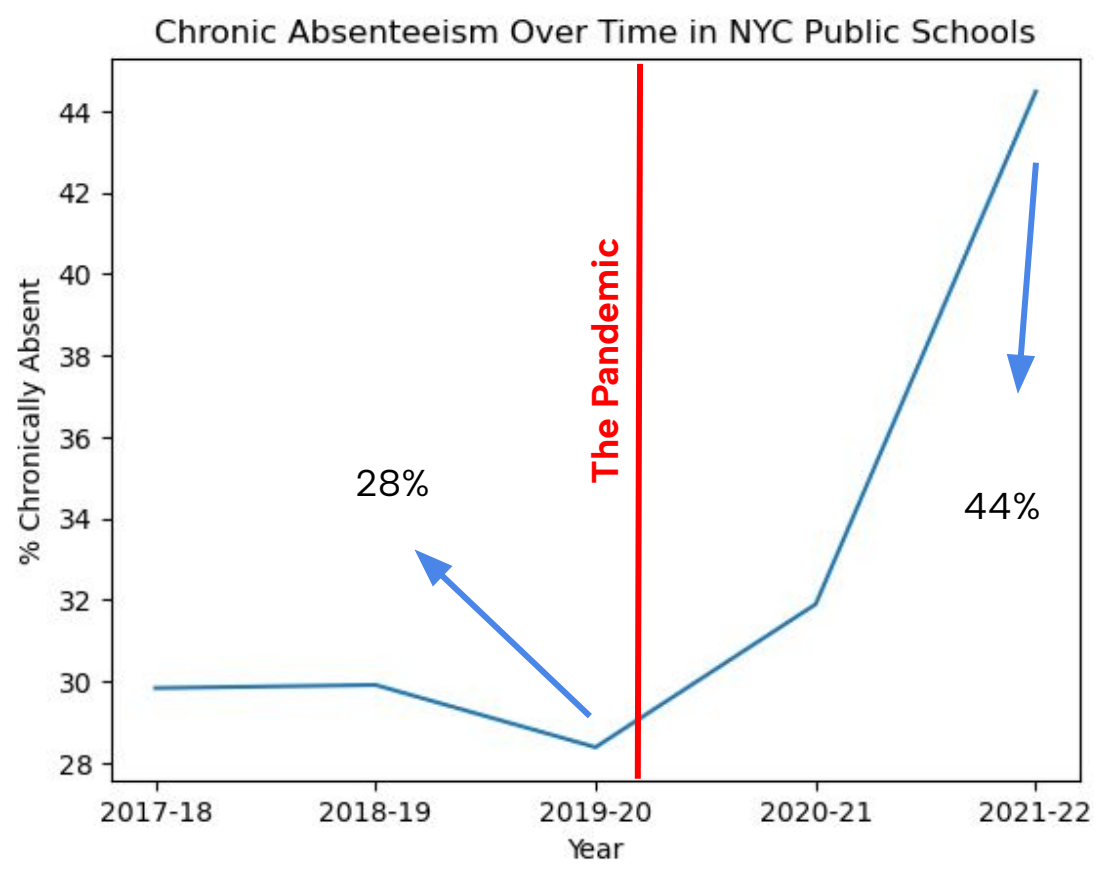
NYC™
Department of Education

What is Chronic Absenteeism (aka my target feature)?

In New York Public Schools, chronic absenteeism is defined as a student missing 10% or more of the school days in a year. Studies have shown that students who are chronically absent have reduced reading proficiency in Elementary School and are less likely to graduate High School



Absenteeism in NYC



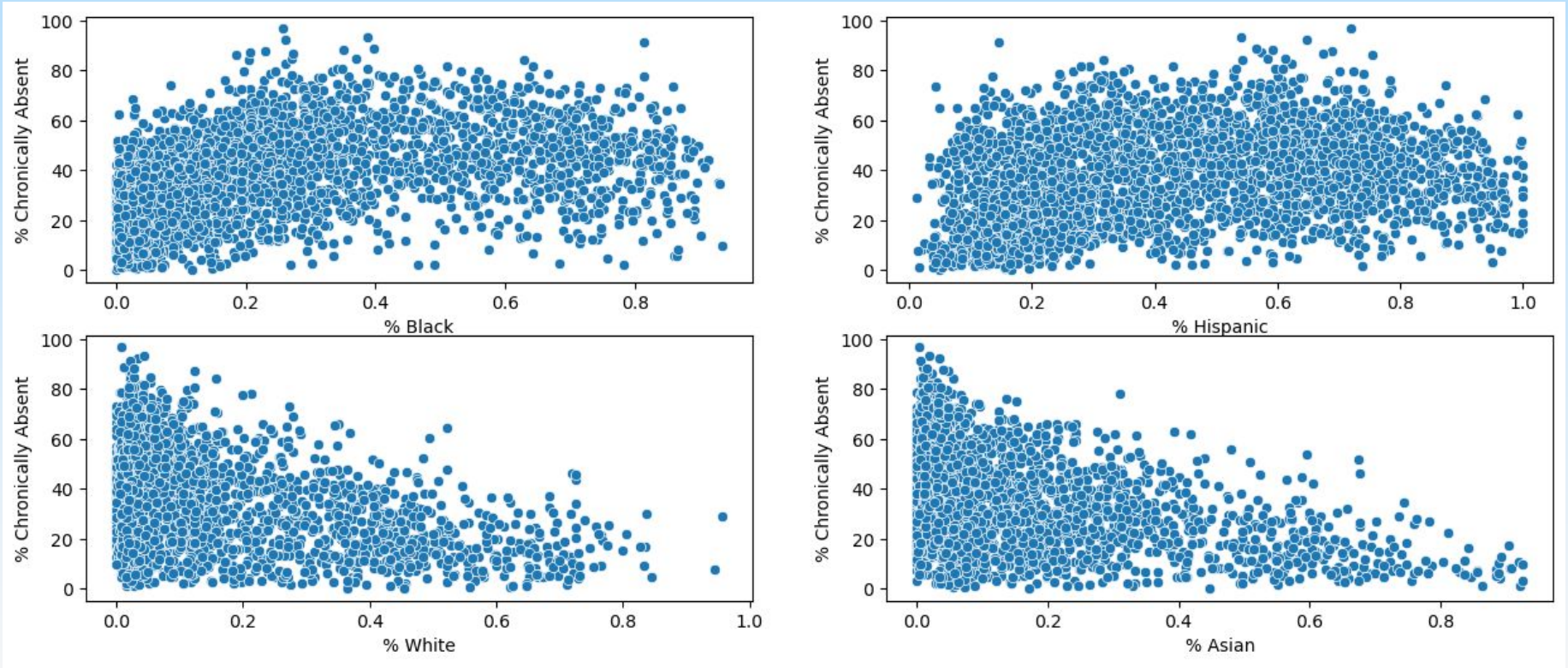
My Data:

I only used data from the 2020 – 2021 and 2021 – 2022 school years and all the data was retrieved from the Open Data NYC or the DOE's Website. I used features from the following data sets to create my final data frame:

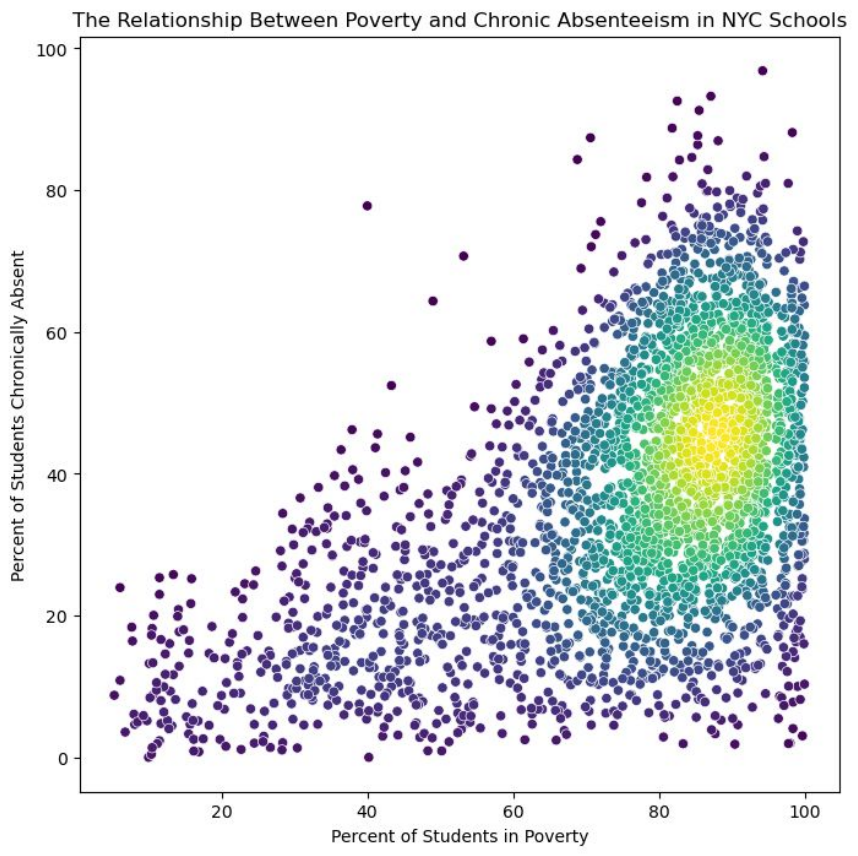
1. DOE Demographic Data
2. Attendance Data
3. Guidance Counselor and Social Worker school Data
4. Indoor/outdoor space in NYC public Schools
5. Graduation Rates
6. Family Survey Data
7. Teacher Survey Data

IMPORTANT NOTE:
Each row in my final data frame represents one school in NYC

Chronic Absenteeism Disproportionately impacts Schools in Black and Hispanic Communities

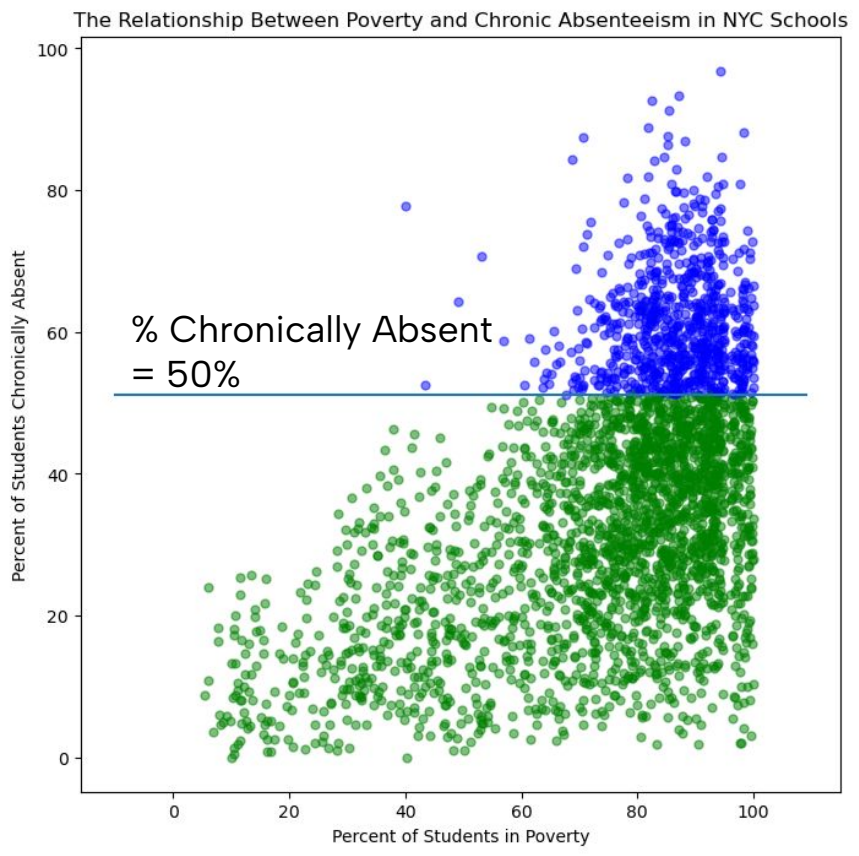


Chronic Absenteeism Disproportionately impacts Schools with higher poverty rates



This makes chronic absenteeism not only an education problem but an equity problem!

Selecting My Target Feature



1 = Schools in the top 25% of Chronic Absenteeism

0 = Schools in the bottom 75% of Chronic Absenteeism

Because my stakeholder really wants to have few false negatives, I will use recall to score my models

Model Assessment For Recall

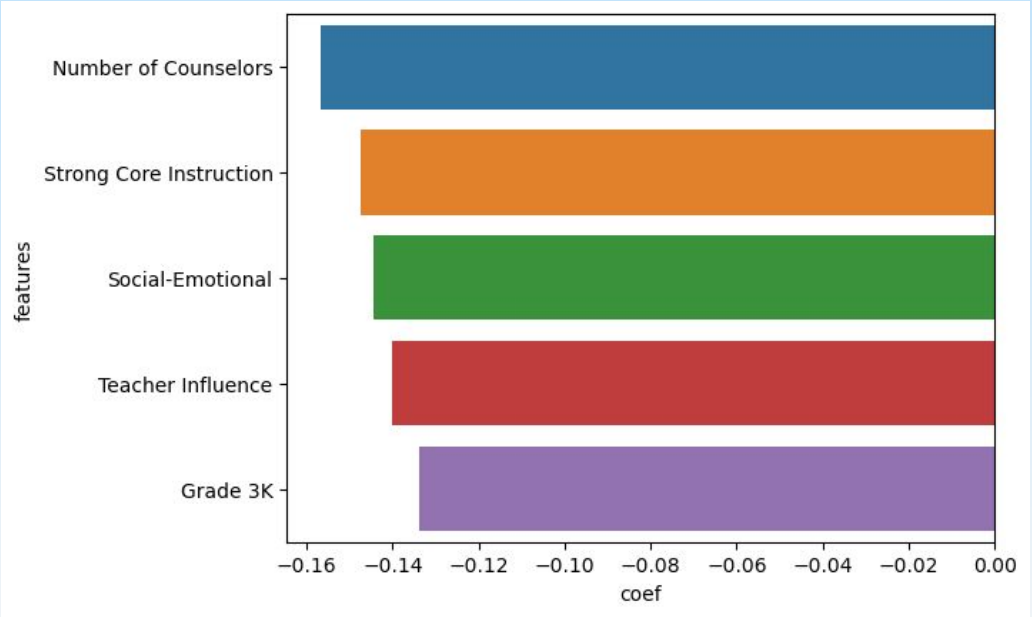
	Test Score (Recall)	Train Score (Recall)	Accuracy	Precision
Baseline (Logistic)	0.83	0.83	0.81	0.62
Logistic with Tuning	0.91	0.89	0.76	0.54
Decision Tree with Tuning	0.91	0.90	0.63	0.42

Despite the recall for the logistic and decision tree model being the same, the accuracy and precision for the logistic regression model is higher. This means there are fewer false positives, which could save my stakeholder money.

Final Model (With Feature Importance listed)

The four most important features of my model:

1. Number of Counselors at a school
2. Strong Core Instruction – interactive and project based learning
3. Social-emotional well being of the student body
4. Teachers having a strong influence on how the school is run



Recommendations

Based on my model, I believe that the Department of Education should flag schools for high rates of absenteeism and provide additional funding to those schools for the following resources:

1. Additional guidance counselors and social works at the school

2. Additional staff that can lead small group instruction and independent projects for students who have fallen behind in reading and math during the pandemic

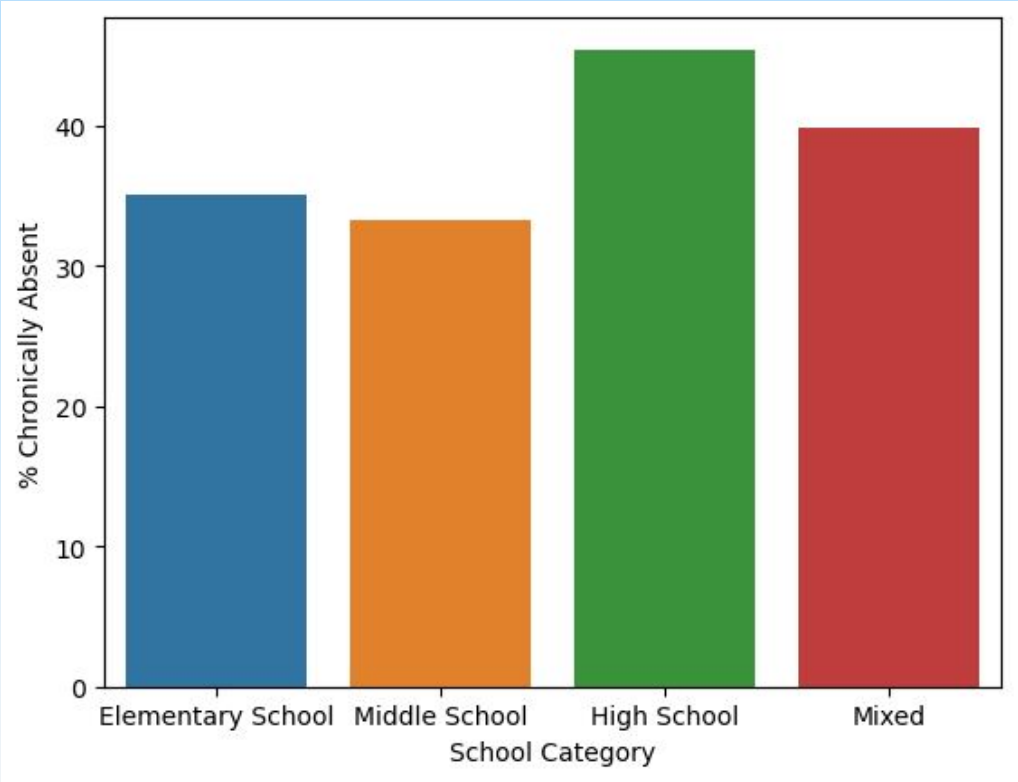
3. Further funds for field trips and community partnerships

Next Steps

- Run more models

- In order to improve my model I would want to scrape and find data on the following:

1. Public transportation offered by each school.
2. Student employment
3. School Start Times
4. The number of tests given by a school
5. Suspension data for schools
6. Lateness Rates for schools



Questions?