Student Trajectories and School Choice in the New York City Public School System

Anandini Chawla
New York University
akc451@nyu.edu

David Futran
Queens College
David.futran@gmail.com

Ro Liriano
Lehman College
roseencarna@gmail.com

Keri Mallari
Lehman College
kmallari15@gmail.com

Francois Mertil
City Tech
francoismertil@gmail.com

Ilana Radinsky
Yeshiva University
ilanaradinsky@gmail.com

Rivka Schuster
Touro College
collegers96@gmail.com

Thoa Ta
St. John's University
thoa.ta.ds3@outlook.com

I. INTRODUCTION

New York City (NYC) has the largest public school system in the country and serves over one million students each year across 1,800 elementary, middle, and high schools. While there are many datasets publicly available about NYC schools, there are few that describe the landscape at the student level. This summer, the NYC Department of Education (DOE) granted us access to student-level data, which allowed us to explore student trajectories through the public school system and the recently implemented high school choice system. These explorations gave us insights into how early test performance correlates with later success, which students are more likely to leave the public school system, what kinds of high schools students tend to apply to, and the chance a student is accepted to his/her top choice high school.

II. DATA AND METHODS

Our confidential datasets included 11 academic years of data from 2005-06 to 2015-16. We mainly worked with four sets: Students' June Biography, Students' Test Scores, High School Application Process data, and Zoned School data.

In addition, we also used publicly available data from NYC Open Data [1] and Census data for supplementary purposes, e.g. mapping high school program codes to high schools and zones, mapping zip codes to school zones.

III. RESULTS

A. Student Trajectories

To analyze student trajectories through the school system, we first determined whether current academic performance is predictive of future academic performance.

Students are required to take ELA and Math assessments in grades 3 through 8, Science and Social Studies examinations in grades 4 and 8, as well as Regents exams in high school. We created a performance index for each student by year. To do this, we used the Test Scores dataset which contains students scores on all citywide exams, such as ELA, Math, Social Studies and Science exams in elementary school and Regents exams in high school. We took each citywide exam and set the student's score to a percentile in each year for every grade. Then, we calculated the average of all percentiled scores of each student in a given year. This average percentile was then set as the student's performance. If the student did not have data on a single exam in a given year, he/she was given a performance of missing.

Correlation tests across years made it clear that a student's performance is correlated with his/her performance in other years. More specifically, a student's performance in a year is highly predictive of their performance in the subsequent year, as opposed to their performance several years forward. For example, students with a performance of greater than 90 in 3rd grade perform around 90 in 8th grade, 80 in 11th grade, and 65 in 12th grade. Similarly, students with a performance of less than 10 in 3rd grade perform around 10 in 8th grade, 20 in 11th grade and 45 in 12th grade. (Note that these numbers are at aggregate level, not for any specific individual student.) Given the high number of students dropping out of the system, we attributed part of the weakening in correlation to a selection effect, where high performing students are leaving the system either to attend private schools or to leave the state, and poor performing students are leaving the system either to drop out or to leave the state. We concluded a student's future performance can be predicted, but with limited accuracy.

Further, to gain a more nuanced understanding of student trajectories, we studied student dropouts across all grades. Given the data available to us, we defined student dropout as students leaving the public school system, irrespective of their reasons for leaving. The results were insightful. We found a significant peak in the average dropout rate in 10th grade - 41%, nearly three times greater than the average dropout rate (across all grades) of 15%. Additionally, from 2005 to 2012, the percent of incoming freshmen (i.e. 9th graders) who actually graduated from high school increased by ten percent, from 61% to 66%, which speaks positively of the DOE's efforts to improve graduation rates. Despite this, Black and Hispanic 9th graders on average are significantly less likely to graduate, with a 59% graduation rate each, as compared to their White (78%) and Asian peers (81%).

With this heightened understanding of student trajectories though the school system, we aimed to construct a predictive model that could produce an estimated likelihood that a given student would leave the school system in the coming year. Our model features included student features, such as his/her grade level, attendance rate, academic performance, gender, ethnicity, and whether or not he/she was eligible for free lunch (a proxy for poverty). Also included were features of the student's current school, such as attendance rate, graduation rate, and the average student test scores within the school on various standardized exams - these features served as a proxy for how ”good” the school was. With features ready, we fit a logistic regression:

$$\begin{aligned}\log\frac{p}{1-p} \;=\;& \beta_0 + \beta_1 grade + \beta_2 sex + \beta_3 ethnicity + \\ & \beta_4 poverty + \beta_5 attendance + \beta_6 GPA \\ & \beta_9 schoolScores + \beta_7 schoolAttendance + \\ & \beta_8 schoolGraduation + e_i\end{aligned}$$

The performance of our model is as follows: 88% accuracy, 76% precision, 32% false positive rate, and 0.79 AUC. While our model's accuracy is quite high, it could be misleading due to the low baseline dropout rate.

Furthermore, since the model assigns a probability closer to 1 to students whom it is very confident will drop out, students can then be ranked by highest likelihood of dropout. With this knowledge, schools can use the model to identify students that are most at risk of dropout and intervene. To put it into perspective, the public school system at large educates around one million students. The DOE may only have the budget to help 1% of the student body, approximately 10,000 students, to prevent truancy. In such a case, as shown in Fig. 1, almost 100% of that 1% group identified by the model as the most at-risk would actually be at risk of dropout. The DOE could then be confident that they intervene with the students who are in need of help.

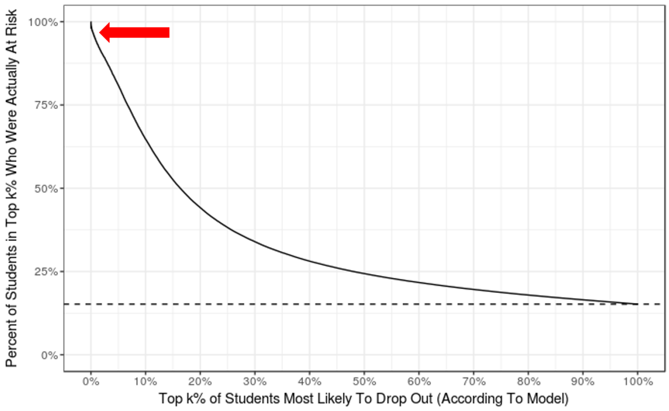

Fig. 1. Top-K Curve shows if the DOE would intervene with 1% of students most at risk of drop-out, they would be accurate that these students would have dropped out with almost 100% accuracy

Our model has some limitations. The data only tell us whether a student left the public school system, but not why he/she left. Therefore, our predicted 'dropout' metric is an aggregation of students leaving for negative reasons, such as a poor performance or difficulties at home, as well as neutral reasons, such as switching to a private school or moving out of NYC. Additionally, given over eight million rows of data to work with, we did not have the memory capacity to include interactions between features in our regression. For example, we would have liked to find out if the effect of attendance or GPA on predicting student dropouts varied by grade; however, we were unable to do so.

B. *School Choice*

As students progress through elementary, middle, and high school, school choice becomes more flexible. In elementary and middle schools, students are zoned to schools within a reasonable distance from their homes, with middle school zones being broader than elementary school zones. For high school, middle-schoolers are eligible to apply to *anywhere* in the five boroughs through a high school choice system implemented in 2003.[1] The only exception is application to specialized high schools, such as Stuyvesant and Bronx High School of Science, where applicants have to take the Specialized High Schools Admissions Test (SHSAT) or attend an audition. We used information from the High School Applications dataset which contains per-application details, including all choices each student ranks in every round, admission methods of those ranked programs, opt out status and reason, and the program that students get finalized to, whether through matching or manual placement. In addition, the data from 2009-10 to 2015-16 has information about whether the applicant applied to a specialized high school, as well as if the applicant received an offer from any specialized high school program.[2]

While students have the freedom to choose from over 400 high schools (with more than 700 programs associated), we found that students tend to list their local schools. Consequently, they mostly get matched to local high schools. A school dynamics map indicates most high schools draw students from nearby neighborhoods, and from nearby middle schools. The only exception is specialized high schools which draw students from all over NYC.

To broaden school choice for students, we developed a Similar School Network, which recommends schools a student can put on his/her list of 12 schools, based on a school the student provides. Two schools are defined as similar if they both exist on any student's top three choices. If they do, a connection is drawn between them, and this connection will become thicker as the two schools become more similar. We transfered this network into a Shiny app, where a student can choose (from a drop-down list) a school that he/she is interested in, and a network will appear informing him/her of other schools that he/she might also care about, since historically many students have applied to them together. The app also displays basic information about the school which a student can use to make better choices.

[1] Prior to 2003, while students could rank schools as they wished, school placement was mostly managed by high school superintendents. In 2003, the DOE implemented a matching algorithm that decreased the number of unmatched high school applicants from 30,000 in 2002 to 3,000 in 2003 [2]. In addition, students can rank up to 12, instead of five, high school programs on their applications.

[2] For details about specialized high schools, go to http://schools.nyc.gov/ChoicesEnrollment/High/specialized/default.htm

Students' first choices on the high school application can reflect how ambitious they are in choosing schools. School quality is an aggregate of its students' performance on standardized tests. Popularity is not an indicator of quality as we saw some schools have mid-range students' test scores but attract a lot of applicants. For instance, the Food and Finance High school has a mid-range rank of 212 but is 8th highest based on number of applications received. We learned that Bronx applicants' top choices, on average, have a wider range of quality compared to top choices by applicants from other boroughs. In other words, while the majority of students' top choices concentrate in good to very good schools (top 40%), the Bronx applicants opt for lower ranked schools as well (significant distribution spreads between top 70% to top 25%). Similar disparity exists between Black and Hispanic applicants (opt for school as low as top 70%) compared to White and Asian applicants (the majority of distribution lies within top 40%).

When plotting applicant's probability of satisfaction, by which we mean getting their top choice, against their percentiled performance, we found that the large majority of applicants are mid-performers, and compared to top and bottom performers, this group have a much lower chance of being matched to their first choices (see Fig. 2). Plotting applicants' top choices against their percentiled performance showed that lower-performing students apply to lower-performing schools, and it seems these aspirations are likely to be satisfied.

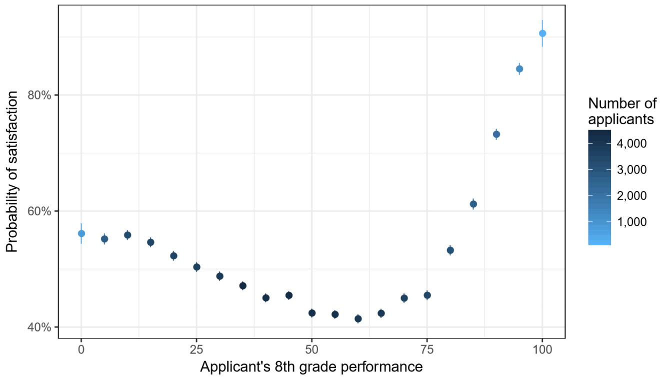

Fig. 2. Top and bottom performers have a higher chance of being matched to their top choice high schools

Since we wanted to predict an applicant's satisfaction at individual level, simply computing and plotting the averages from the data was more exploratory than predictive. Representing the averages, the plots gave us insights into the aggregate trends but did not enable us to predict an applicant's satisfaction, given his/her academic performance, ethnicity, home borough, or the school he/she ranks. Therefore, we ran a logistic regression using information about the applicant, his/her middle school, and the high school he/she lists, to predict whether or not that applicant will be finalized to his/her top choice school. Our model evaluation metrics are: 0.78 AUC, 71% accuracy, 71% precision, 69% recall or true positive rate, and 28% false positive rate. Compared to the baseline scenario when one predicts every applicant to be satisfied, our model has an increase of 21 percentage points in accuracy. The improved performance of the model gives credibility to use the included features to predict the probability an applicant will be satisfied.

IV. CONCLUSIONS AND DISCUSSION

The NYC public school system faces unique challenges because it serves over one million students every year. With the dropout predictive model, we are able to suggest a select percentage of students in need of intervention, so that the public school system can continue serving all students effectively. To boost diversity in the high school applications, the Similar School Network Shiny App recommends relevant schools, which ideally lessens the burden on students and parents navigating over 400 high schools choices, and may introduce students to school choices they are not aware of. The applicant's satisfaction predictive model helps students better manage their expectations when applying to high schools, and gives them the confidence to apply to other schools where they may be accepted.

There is a lot to learn from this data, and this research encourages future work. It would be great to find out the reason(s) behind the disparity in quality of top choice school. We also might want to take into consideration (i.e. plot) the traveling distance between students' home and high school, faceted by boroughs, to see if school's approximate distance from home is a factor that skews the distribution of top choice quality among boroughs. For example, it could be case that the Bronx does not have enough high-performing high schools to serve its local students who do not wish to travel far for high school.

V. ACKNOWLEDGEMENTS

We thank our mentors Jake Hofman, Siddhartha Sen, Dan Goldstein, Brian Quistorff, Jacob LaRiviere, Matt Goldman, and William Cai for their guidance on this project as part of the 2017 Microsoft Research Data Science Summer School. We also wish to extend a special thanks to Josh Smith and the NYC DOE for granting us the access to the datasets.

REFERENCES

[1] ”2013-2014 School Zones — NYC Open Data”, *NYC Open Data*, 2013. [Online]. Available at: https://data.cityofnewyork.us/Education/2013-2014-School-Zones/pp5b-95kq. [Accessed: Aug. 2017].

[2] Abdulkadirolu, A., Pathak, P. and Roth, A. (2005). *The New York City High School Match*, pp. 364-367 [Online]. Available at: http://www.u.arizona.edu/~mwalker/501BReadings/FourRothMatchingMechanisms.pdf. [Accessed Aug. 2017].