FOUNDATION OF DATA SCIENCE
PROJECT PROPOSAL

PREDICTIVE MODELLING & ANALYSIS OF NYC SCHOOLS

PRATIK KAMATH
AKANKSHA SHAKTIA

Project Proposal

Introduction

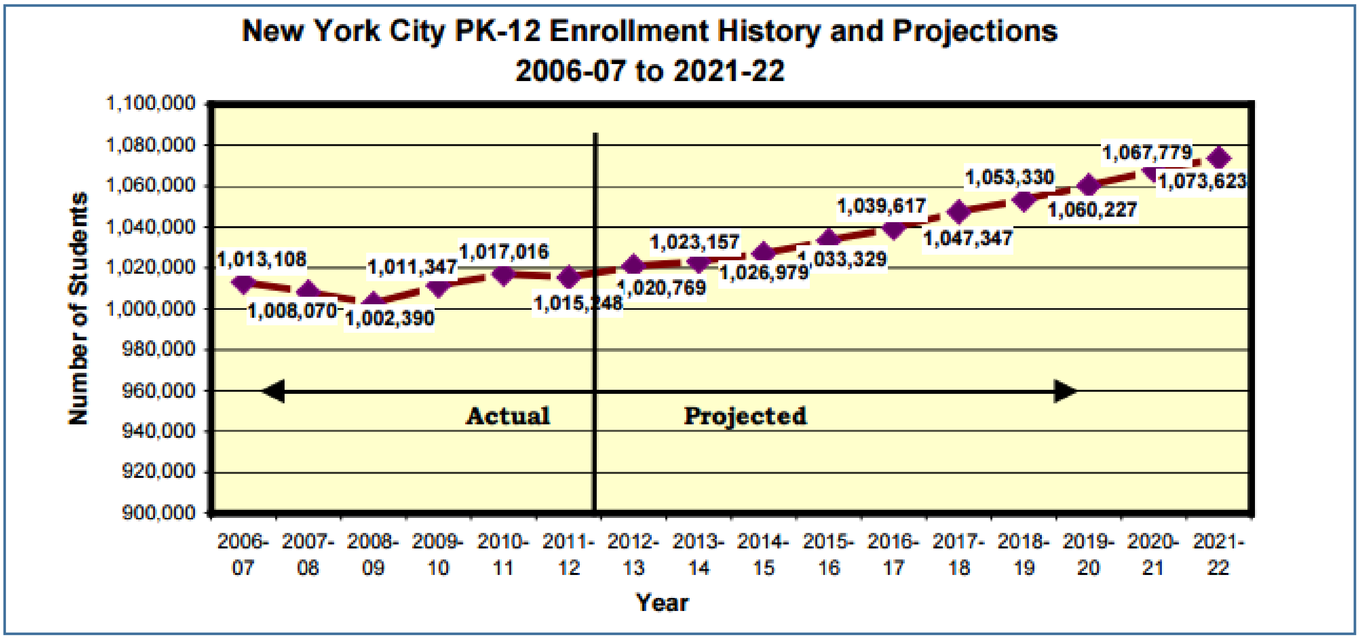
Every year during the time of admissions, parents pour over brochures showing pictures of students in front of lush trees. High schoolers ensure that they put forward activities that are the “right” ones to enroll in. The problem mainly persists due to the admission trends observed over the last couple of years. Over the last 15 years studied, enrollment in U.S. institutions of higher education at all levels rose from 14.5 million students in fall 1994 to 20.7 million in fall 2009, with most of the growth occurring in the last 10 years. Increased enrollment in higher education at all levels is projected to come mainly from minority groups, particularly Hispanics. Enrollment of all racial/ethnic groups is projected to increase, but the percentage for whites is projected to decrease from 63% in 2008 to 58% in 2019, whereas the percentages for blacks and Hispanics are projected to increase from 14% and 12% respectively, to 15% for both groups. In the 2009-10 academic year, the number of foreign students enrolled in bachelor's degree programs in U.S. academic institutions rose 5% from the previous year, to approximately 206,000 [1].

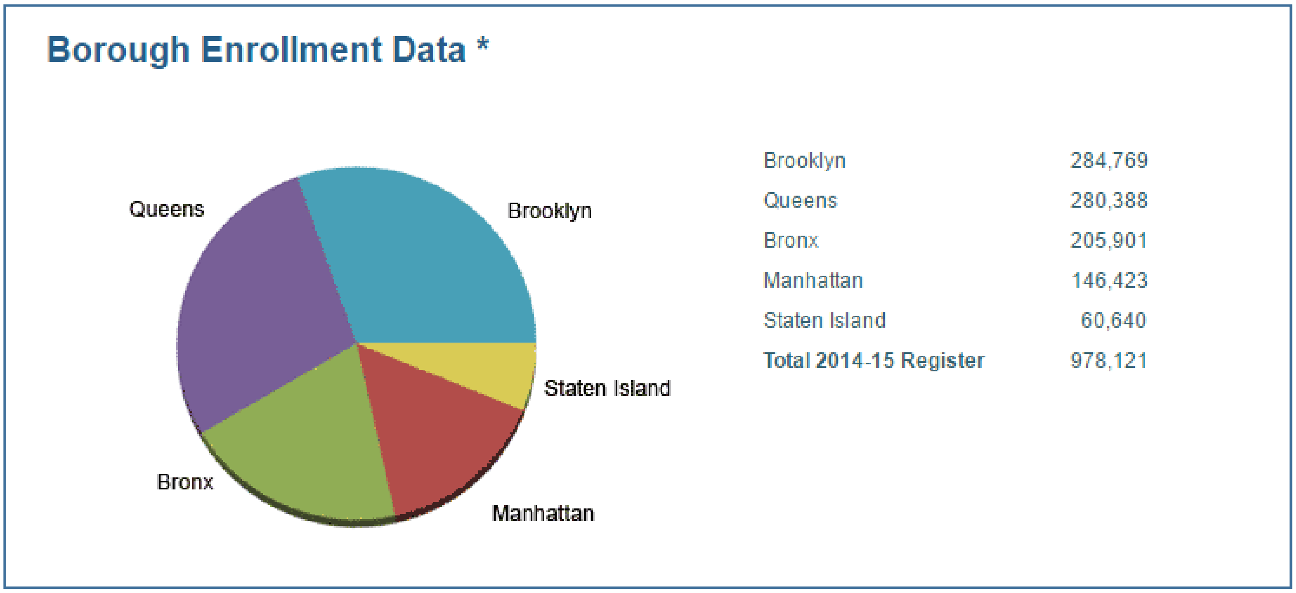
Background

Admissions to undergraduate schools are getting competitive each year. Therefore, the parents are posed with the difficult problem of choosing an appropriate school for their child. Each school promises an array of benefits but the quality of education imparted may paint a different picture. In such situations it becomes crucial for parents to make an informed decision.

From the data of February 2014, there has been 18.7% growth in the number of students as against to only 11.5% growth in schools. [2]

From the graph below we can see the estimated rise in the admissions by 2020-21[3]





Borough Enrollment Data *

Brooklyn	284,769
Queens	280,388
Bronx	205,901
Manhattan	146,423
Staten Island	60,640
Total 2014-15 Register	978,121

Student enrollment data [4]

From the statistics we can see that the number of students is considerably larger compared to the number of schools present to accommodate them.

Data and Analysis

In order to conduct our analysis, we have used the NYC high school education report [5]. The dataset consists of the data for the year 2014-15 as recorded on January 7th 2016. It majorly comprises of ordinal and numeric data. Predictors that rate the school on various parameters consist of ordinal data. We intend to use this data by associating numeric values to it (e.g. on a scale of 1-5, not meeting target will be given the value 1 and exceeding value will be given the value 5). Some key features in the dataset have the numerical datatype. The dataset has percentage values for various features. Since R does not have a data type for percent it will take it as 'character' by default and therefore we will have to convert into numeric data to work with it.

With the data that we have procured, we intend to analyze the performance of different high schools and deduce conclusions for questions like-

1) Rank the schools based on the student achievement and the surveys conducted.
2) Predict the majors of the student based on SAT/ACT scores.
3) Choice of schools for students with special needs
4) Quality of education imparted based upon the performance of teachers.
5) Grading of schools in terms of HRA being offered.
6) Welfare provided to the students such as safety, coping with emotional difficulty, additional guidance for studies.

These are some of the major conclusions that can be drawn from the data.

Model

We want to build a predictive model which will find out the predictors for a particular prediction problem. So the first step for the model will be cleaning the data so that the data is easily usable and

is available for statistical analysis. Then we will do exploratory data analysis on the cleaned data to find descriptive statistics of the variables.
The model will learn from the past data and make predictions for the future. The model will be able to predict the likelihood of a particular event e.g. Predicting the majors of a student depending on past information. We wish to create this model by employing regression or kNN Classifier analysis etc. on the predictors depending on the fit of the problem.

Assumptions

For each conclusion, we will consider a set of features that support the decision being drawn. During the course of analysis, we will change the data type of these features to make them easier to use.

Following is a summary of the conclusions we wish to draw from the dataset, the features that will support the conclusion and the reason for selecting those features.

Decision	Features	Reason
School Rank	Trust Rating Student Achievement Rating Rigorous Instruction - Percent Positive Quality Review - How interesting and challenging is the curriculum? Quality Review - How effective is the teaching and learning? Student Survey Response Rate Teacher Survey Response Rate Parent Survey Response Rate	School rank is determined based upon the feedback given by the teachers, parents and the students. The kind of knowledge imparted also impacts the quality of the institution.
Predict the majors	Average Score Algebra Average Score Geometry Average Score English Average Score US History Average Score Chemistry Average Score Physics Average Score Earth Science Average Score Environment Average Score SAT Math Average Score SAT Critical Reading Average Score ACT Math Average Score ACT English	Based upon the average scores of the students in each of the subject, we can predict the probability that a student from a particular school is most likely to take up which major.
Choice of schools for students with special needs	Percentage of students who say that teachers notice when they are upset or having emotional difficulty Supportive Environment - Percent Positive Responses on School Survey	Since children with special needs require special attention, it is important to consider how tolerant the schools are in providing those.

	Strong Family-Community Ties Peer support for academic work - Combined - Score Personal attention and support - Students - Score	
Quality of education imparted	Course clarity – Score Program coherence - Teachers - Score Teacher influence - Teachers - Score Press toward academic achievement	In order to determine the quality of education imparted, we must consider quality of the program and course being taught and the support provided by the school to the students.
Best schools for teachers to work in	Percent in Temp Housing Percent HRA Eligible Quality Review - How well do teachers work with each other? Percentage of teachers who say that they have opportunities to work productively with colleagues in their school Percentage of teachers who say that they work together to design instructional programs Teacher-teacher trust - Teachers - Score Teacher-principal trust - Teachers - Score Social-emotional measure - Teachers - Score	To determine how the schools fare at providing a healthy work environment to the faculty, we considered the benefits that the existing teachers receive, their level of comfort working for the school and with each other.
Welfare provided to the students	Supportive Environment Rating Trust Rating Supportive Environment - Element Score Percentage of students who feel safe in the hallways, bathrooms, locker room, and cafeteria Percentage of students who say that teachers notice when they are upset or having emotional difficulty Percentage of students who say that teachers treat them with respect Student-teacher trust - Students - Score	Along with the quality of education, it is important to consider how secure a child feels being in school. To evaluate this parameter, we will consider the safety parameters in the dataset.

Since many features listed above have a non-numeric data type, in order to compute the score for each decision, we will convert the data into numeric type as follows-

0	1	2	3
Not meeting target	Approaching target	Meeting target	Exceeding target
Not Developed	Developing	Well Developed	Proficient

References

1. http://www.nsf.gov/statistics/seind12/c2/c2s2.htm
2. http://www.publiccharters.org/wp-content/uploads/2014/02/New-and-Closed-Report-February-20141.pdf
3. http://www.nycsca.org/Community/CapitalPlanManagementReportsData/Demographics/2012-2021StatisticalForecastingReport.pdf
4. http://schools.nyc.gov/AboutUs/schools/data/stats/default.htm
5. https://data.cityofnewyork.us/Education/2014-2015-School-Quality-Reports-Results-For-High-/vrfr-9k4d