Executive Summary

Group: Will Chirciu, Amy Edwards, Julie Simms, Breanna Atwood, Jeff Moy

PASSNYC is an organization headquartered in New York City dedicated to providing more opportunities to underserved students attending schools in underdeveloped communities. They gather data to provide those interested with insight as to how the performance of the school system is affected by or related to certain demographics. The school explorer database gave access to characteristics such as income, economic need index, and ethnic makeup (percent of Asian students, percent of Black students, percent of English Learning Students) to get a better idea of the demographic situation surrounding the schools. The database also reveals data related to the performance of the schools: average test scores, how involved the community is in the classroom, how much the teachers care about providing a supportive, safe environment for the students to grow, and so on.

Using this database, the goal was to extract as much information as possible to better understand the culture of the inner-city school system. A canonical correlation analysis was used to discover the relationship between the demographics and performance of the NYC schools. It was found that the economic need index (which helps determine the economic need of a school) has a strong negative relationship with the average ELA and Math Proficiency test scores. This result makes sense from an outsider standpoint as schools in economic need tend to have fewer resources available to their students. This would most likely lead to a decline in test scores. The CCA also showed that schools with higher percentages of black students have communities and family members that aren’t as involved in the day-to-day and have students that don’t feel safe or supported by their teachers and peers. Additionally, it was discovered that community schools tend to have lower standards in the classroom, have teachers who are not as committed to the success and improvement of their classroom and schools, and have less developed relationships with families, business, and community-based organizations. This is counter-intuitive to the definition of a community school and requires more exploration.

K-means clustering was performed on the data to best define groups of schools based on income and average proficiency test scores. Income was not a major factor in shaping the clusters. The cluster of schools with high proficiency scores consisted of schools with both high and low income estimates and vice versa. This challenges the idea that schools with less money produce less intelligent students. With this, it was conclusively decided that school income estimate was not a good factor in determining if a school needs additional resources. The two clusters could be further broken down using demographics and school relationships. Smaller groups make it easier to test out policy initiatives that will actually target the issues these schools are facing. A school with a large population of English Language Learning students and high test scores will have different needs from a school with a large percent of white students and low test scores.

The student achievement rating tells whether a school is meeting its goals in terms of its state test results, student growth and performance, how the students performed in core courses, and how well students are prepared. The goal was here was to create a predictive model that tells whether a school missed the mark, met, or even surpassed these challenges. Using Linear Discriminant analysis, a linear combination of variables was found that accounts for 98% of the variability in the data. The variables involved are School Income Estimate, Economic Need index, Average Math Proficiency Score, and Average ELA Proficiency score. Leave-one-out cross validation confirmed an accuracy of 64% on the training data. Testing the model on unseen data also gave an accuracy rating of 64%. Ultimately, the variables selected in the PASSNYC dataset do not have enough information in them for an acceptable student achievement prediction model.

Economic need index is used in determining the socioeconomics of the school population. The larger the index, the larger the need. The following factors are used in calculating the index: the percent of students in temporary housing, the percent of employees and employers who are eligible for HRA benefits, and the % of students who are eligible for free lunches. Wanting to know what characteristics provided by the data may drive the economic need index, stepwise regression was performed to determine the significant predictors. Based on several significance tests the following factors were discovered to have the most significant influence on economic need index: school income, strong family community ties, average Math performance rating, % White, % Hispanic students, student attendance rate, and % of students who are English Language Learners. A model was created with 89% accuracy using the above-mentioned predictors. Utilizing the model, Institutions can predict the ENI even if they don’t have all of the information needed to compute it.

The data contains over 100 variables that detail the number of students in grades 3 through 8 that scored at least a 4 on their Proficiency Tests. Wanting to see if they could be grouped with the ethnicity percentages, a Principal Component Analysis was run. PCA can determine if there are any hidden factors in the data. A KMO Adequacy test was used to confirm the factorability of the data. The PCA analysis was performed using students in grades 3-8 with scores of 4 and the different ethnicity categories (White; Multiracial; Asian or Pacific Islander; English Language Limited; Black or African American; American Indian or Alaskan Native; and Hispanic or Latino). Nine components were extracted that are broken up by race and grade level. This gives 9 relationships between the data that were previously unknown and can be used in future regression models.

This data only provides a glimpse of the challenges these students face in the inner NYC school system. The goal was to enable PASSNYC to identify the schools where minority and underserved students stand to gain the most from services like after school programs, test preparation, mentoring, or resources for parents. Two predictive models and 3 grouping models have been created to help identify said schools. From a practical standpoint, the predictive models and clustering analysis can be used to assess the schools in determining whether their need is economic or academic. Alternatively, using the PCA and CCA results can help identify specific people groups that might benefit more from additional resources or financial intervention. Despite all this, it is recommended that a Time Series analysis be done to see how the performance of these schools evolve across multiple years. It would be incredibly useful to have a model to predict future test scores, enabling institutions to react when they see a decline. Additionally, data relating to number of students and faculty within each school would provide more context for certain variables like ‘Income’ to create a more accurate picture of the NYC school system.