Understanding Test Performance in Pennsylvania Schools

Capstone Three Project Proposal
Christopher Chung
January 2022

Business Problem

Pennsylvania's public school system is home to a wide range of schools in terms of performance, demographics, and funding. It's no question that there exist sharp race/ethnic differences in student achievement. There are significant achievement gaps based on family economic status as well. There is growing evidence that increased funding in education leads to better outcomes, particularly for schools in high-poverty locations.

One measure of school outcomes is the Pennsylvania System School Assessment, or the PSSA. The PSSA is a standards-based, criterion-referenced assessment which provides students, parents, educators, and citizens with an understanding of student and school performance related to the attainment of proficiency of academic standards. Every Pennsylvania student in grades 3 through 8 attending district schools, charters, and cyber charters is assessed in English Language Arts and Math.

The goals of this project are:

1. To investigate the magnitude of gaps in school performance across multiple demographic groups.
2. To understand what other factors are associated with school performance (including funding and expenditures per student).
3. To create a model that will establish specific connections between the independent variables and the dependent variable, which is the percentage of students that pass (proficient or above) in each school. Two models will be created, one for Mathematics assessment and the other for English Language Arts (ELA) assessment.

The intended stakeholders are school administrators and policy makers. This problem is relevant to them because PSSA results factor heavily into the School Performance Profile, which is the state's school rating system. Furthermore, school scores are used by districts in decisions about school turnaround interventions and school closings, as well as in the charter school renewal process.

Sources:
https://www.education.pa.gov/K-12/Assessment%20and%20Accountability/PSSA/Pages/default.aspx
https://philadelphia.chalkbeat.org/2015/11/18/22182593/understanding-the-pssa-exams

Datasets

The datasets are all from https://futurereadypa.org, which is a collection of school progress measures related to school and student success.

School and Performance Data for 2018-2019 School Year

- Level of granularity: School
- General and demographic information about each school.
    - Enrollment
    - Percent male/female
    - Race/ethnic composition of school
    - Percent economically disadvantaged

    - Percent of gifted students
    - Percent of students in foster care
    - Percent of students that are homeless
    - **Target variable:** Percent Proficient or Advanced in ELA
    - **Target variable:** Percent Proficient or Advanced in Mathematics
    - etc.

School Fiscal Data for 2018-2019 School Year

- Level of granularity: School
- For each school, we have:
    - **Federal - Non-Personnel:** The federal amount spent per pupil for public elementary and secondary education (pre-K through 12th grade) on non-personnel.
    - **Federal - Personnel:** The federal amount spent per pupil for public elementary and secondary education (pre-K through 12th grade) on personnel.
    - **Local - Non-Personnel:** The local amount spent per pupil for public elementary and secondary education (pre-K through 12th grade) on non-personnel.
    - **Local - Personnel:** The local amount spent per pupil for public elementary and secondary education (pre-K through 12th grade) on personnel.
    - **State - Non-Personnel:** The state amount spent per pupil for public elementary and secondary education (pre-K through 12th grade) on non-personnel.
    - **State - Personnel:** The state amount spent per pupil for public elementary and secondary education (pre-K through 12th grade) on personnel.
- Intention is to join with above dataset on School ID Number

The final dataset will have approximately **2700 rows**, one row for each school including, but not limited to, the following features:

- School ID
- District ID
- Enrollment
- Gender by percentage
- Race by percentage
- Percent of students economically disadvantaged
- Percent of students in foster care
- Percent of students homeless
- **Federal - Non-Personnel:** The federal amount spent per pupil for public elementary and secondary education (pre-K through 12th grade) on non-personnel.
- **Federal - Personnel:** The federal amount spent per pupil for public elementary and secondary education (pre-K through 12th grade) on personnel.
- **Local - Non-Personnel:** The local amount spent per pupil for public elementary and secondary education (pre-K through 12th grade) on non-personnel.
- **Local - Personnel:** The local amount spent per pupil for public elementary and secondary education (pre-K through 12th grade) on personnel.
- **State - Non-Personnel:** The state amount spent per pupil for public elementary and secondary education (pre-K through 12th grade) on non-personnel.
- **State - Personnel:** The state amount spent per pupil for public elementary and secondary education (pre-K through 12th grade) on personnel.
- Percent proficient or advanced in Math
- Percent proficient or advanced in ELA

Anticipated Data Science Approach

I will follow the classical data science project development process: data wrangling, exploratory data analysis, baseline modeling, extended modeling, project report, and project presentation slide deck.

For instance, I anticipate that my exploratory analysis will consider–but will not be limited to–investigating the correlation of school performance (percentage of students proficient in Math and ELA) with various features using python data visualization packages such as matplotlib.pyplot and seaborn.

As for the modeling components of this project, I anticipate that I will use supervised regression algorithms to build and evaluate the performance of various models. I will create two groups of models, one for which the target is the percentage of school proficient (or above) in Math, and another for which the target is the percentage of school proficient (or above) in ELA.

To evaluate my models, I will use mean absolute error, mean squared error, r-squared, and a study of the distribution of residuals to determine the worst-case upper and lower bound of the errors for the predictions associated with the test set.

Once the models are built, they can be used to:

a) Identify the connections between features (such as demographics data, percentage of students economically disadvantaged, expenditures per student, etc) and the targets.
b) Propose possible recommendations for schools with low performance by extracting counterfactual explanations from the models.

Deliverables

All Jupyter notebooks
A written final report
A presentation slide deck