New York State Public Schools Graduation Rates: Predicting Success

Stefano Bastianelli
Goergen Institute of Data Science
University of Rochester
Rochester, NY
sbastia2@reslife.rochester.edu

Brooke Brehm
Goergen Institute of Data Science
University of Rochester
Rochester, NY
bbrehm2@u.rochester.edu

Derek Caramella
Goergen Institute of Data Science
University of Rochester
Rochester, NY
dcaramel@ur.rochester.edu

Miguel Novo Villar
Goergen Institute of Data Science
University of Rochester
Rochester, NY
mnovovil@ur.rochester.edu

Kenzie Potter
Goergen Institute of Data Science
University of Rochester
Rochester, NY
kpotter6@u.rochester.edu

I. Introduction

The aim of this project is to model the way government expenditures and labor appropriation impacts secondary education graduation rates in New York State Public Schools. We hypothesize municipality expenditures demonstrate diminishing returns; thus, an optimal expenditure exists to increase graduation rates without expending excessive funds. We contribute to the existing literature that advises local municipalities to develop an optimal taxation scheme via resource allocation to create a more efficient educational landscape. Following rigorous preprocessing, we make observations on the dataset's correlations, distribution, and features - original and engineered - with different analytical and visualization techniques. We explore the effects on graduation rates from a variety of factors, such as funding per student, institutional status, and teacher qualifications. The machine learning algorithms implemented are Elastic Net, SVR, Bayesian Ridge, AdaBoost Regressor, Random Forest Regressor, and neural network regression. Our experiments show that diminishing returns are not present in funding, rather the educational staff's quality affects graduation rates.

II. Literature Review

It might seem controversial that teacher quality plays an important role in student performance, but it is a persistent finding that qualified teachers directly impact educational outcomes [3]. Different paths into teaching and changing requirements shaped the educational landscape and directly affected the learning conditions in New York students. In New York, there has been little difference in student academic achievement found between certified, uncertified, and alternatively certified teachers [7]. However, school districts that are able "to attract and retain high quality teachers (or screen-out less effective teachers) have potentially large benefits for student achievement" [7]. Differences in teacher effectiveness could indicate that schools should "use performance on the job, rather than initial certification status to improve average teacher effectiveness" and potentially make improvements with budgeting and graduation rates [7].

In a journal article from 2019, researchers found that charter schools can impact the costs and efficiencies of other nearby public schools. By increasing the proportion of students with greater need, including those with disabilities and receiving free lunch, New York State charter schools have "increased the cost of education in traditional school districts" [4]. Charter schools could potentially impact spending efficiency in surrounding districts, depending on charter schools' enrollments and the public school district's reaction to the change in student population.

Could additional federal or state funding help level the playing field at all for students in students in disadvantaged socioeconomic situations? Using national graduation rate and poverty data, one study found a "statistically significant relationship between poverty and graduation rate" and a "moderate and negative relationship between poverty and graduation rates" [1]. In the United states, "graduation rate increases as poverty rate decreases" [1]. In a study of New York City's "small schools of choice" (SSCs), which are located in historically disadvantaged communities" and "intended to be viable alternatives to the neighborhood high schools that were closing," it was found that they "markedly improve graduation prospects for many disadvantaged students" [2]. The smaller, more focused SSCs were able to increase overall graduation rates by 6.8% [2].

Allocating equally does not imply allocating equitably. In 2007, the Department of Education in the City of New York created a project to allocate resources based on the student body characteristics. "Why Do Some Schools Get More and Others Less? An Examination of School-Level Funding in New York City" was written at the time the measure was implemented and can be used as a benchmark against our analysis from our 2020 educational data in the New York state

[8]. The study found that funding does not correspond strictly to educational characteristics. While there is a difference in resource allocation across schools, there is a large variation of resource allocation based on outside factors that generate differences across the student body.

There is more to student success than just increasing spending. After reviewing the data from 187 other studies, Hanushek, from the University of Rochester, had more nuanced findings on the relationship between school expenditure and achievement. While most of the data showed a strong positive affiliation between achievement and spending, detailed research provided "strong and consistent evidence that expenditures are not systematically linked to student achievement" [6]. However, they were unable to make suggestions as to how schools may be more efficient with their budgets.

The Center for Policy Research at Syracuse University had several findings regarding the costs of education and adequacy. Syracuse University uncovered that raising costs were linked to rising salaries required to attract high-quality instructors to meet student needs, especially in the larger cities of New York. Students experiencing poverty "are estimated to require almost twice the resources as the average student" [5]. In the state's five largest cities, concentrated poverty rates raise the cost of educating students by 20 to 30 percent [5]. It was found that "enrollment size of a district has relatively little impact on cost" for schools with 1,000 or greater students [5]. Districts with fewer than 1,000 students have an estimated 10 percent higher cost than districts with 1,000 to 15,000 students [5]. Syracuse University estimated the cost for academic adequacy using test scores for state Regents Examinations. The estimates involve higher spending within schools in each of the "Big Five" large cities in the state to reach thresholds for sufficient average Regents Examinations scores. However, they did not use the percentage of students graduating from schools as an "adequacy" performance metric. This research recommends "a better-designed aid system" for students, "a required minimum level of local contribution [to receive state funding, and] improved use of existing resources" [5]. While the other two recommendations appear almost common-sense, the required local funding suggestion could be contrary to the other findings about impoverished districts. Analyzing, predicting, and grouping different patterns across schools, our work could potentially help policy makers update instructor requirements and allocate resources more efficiently to create a more balanced and stronger educational system in New York state.

III. Data

The dataset chosen for this project is the 2020 Report Card Database from the New York State Department of Education [9]. The data contains six tables: teaching certification, school groupings, experience of teachers and principles, expenditures per pupil, accountability status, and high school graduation rates; each dataset contains between 5,542 and 158,920 records for the 2019-2020 school year in New York public schools. Each of the tables provides features describing teachers, students, schools, and districts by utilizing foreign key relationships. The teaching certification features include amounts and percentages of teachers in each school, poverty levels, experienced and inexperienced teachers and principles, expenditures from various government municipalities. Moreover, the data describes each school with an accountability status: good standing, potential target district, potential targeted support and improvement (TSI), potential comprehensive support and improvement (CSI), TSI, CSI, and closing school. Graduation rates for each school are grouped by all students, race, year cohort (4, 5, 6, and combined years to graduate), income level, and English as a second language for comparison, though not all schools contain the entire set of groupings.

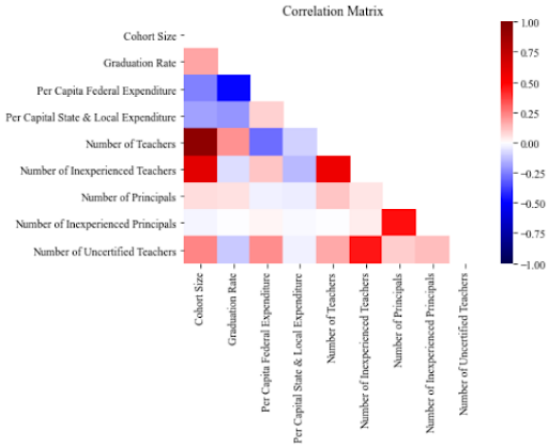

Fig. 1. Correlation Matrix of the NYSED Dataset Dimensions

In our exploratory data analysis, we uncovered several correlations between the attributes using a correlation matrix (Figure 1). As expected, "number of teachers and cohort size", "number of inexperienced teachers and number of teachers" and "number of inexperienced principals and number of principals" were correlated. Also, we can see some interesting correlations: "number of inexperienced teachers and cohort size" and the negative correlations of "per capita federal expenditure" with "cohort size" and "number of teachers." These correlations are a signal that schools tend to be parsimonious with the funds allocation when the number of students is higher; especially if we consider that school budgets are drafted and approved months before the following school year.

Evaluating funding, in Figure 2, we compare federal and state/local fund allocation per student, with federal funds in the range of a few thousand. It would be interesting to see the effect in terms of improvements for New York State schools if the Federal funds were slightly higher. Since much state and local funding comes from property taxes, students at schools with more state/local funding are likely from socioeconomic backgrounds that already put them at an advantage [6]. When we look at Figure 3, with all expenditures combined and compared with graduation rate, there are clear groupings of schools. The majority of the schools in the dataset have allotted $10,000 to $30,000 per student, and within that expenditure

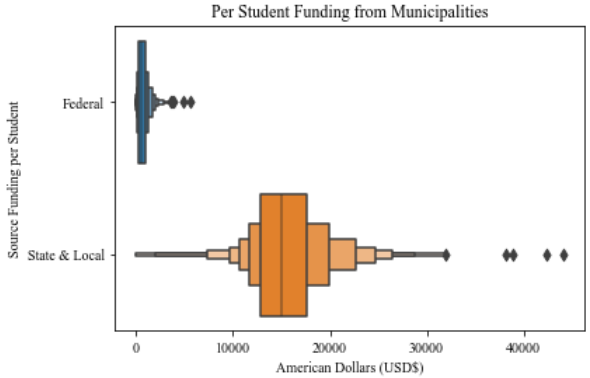

Fig. 2. Boxenplot of per Student Funding from Municipalities

domain the schools are condensed primarily at 60% to 100% graduation rates. Interestingly, schools spending $0 to $9,999 per student have a much higher range of graduation rates than the full dataset, with graduation rates falling roughly between 80% and 100%. There are only a handful of schools with expenditures of more than $30,000 per student. These high expenditure schools could include outlier data points. When we observe the institutional standing of the schools in Figure 4, nearly 89% are in the highest category "good standing," about 11% are currently receiving support and improvement, and there are no schools in potential need of support and improvement or "closing schools." Some of the data points with graduation rates in the range of 0% to 60% in Figure 3 could include those receiving support. Overall, this dataset does not show a strong positive correlation between funding and graduation rates.

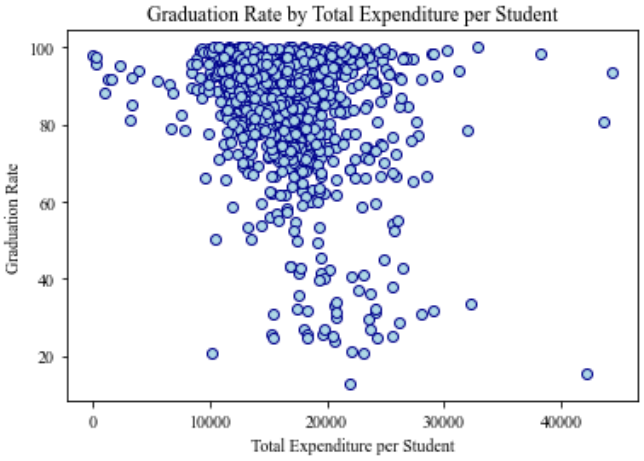

Fig. 3. Plot of Graduation Rate by Total Expenditure per Student

Other features in the dataset that are potentially influencing graduation include teachers. Figure 5 is a histogram of the number of students per teacher in a school, which is the total number of students in a school (cohort size) divided by the total number of teachers. For this histogram, the probability distribution is fairly normal, but slightly right-skewed. The

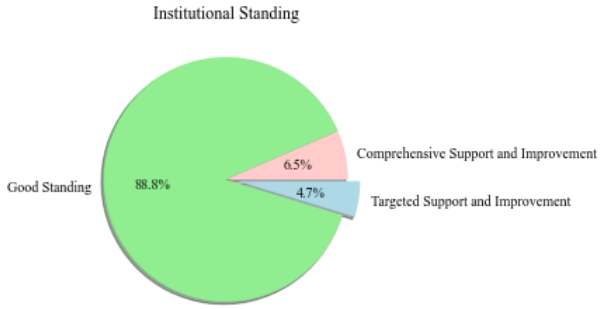

Fig. 4. Pie Chart of Institutional Standing

minimum number of students per teacher is only three, and the maximum is twenty-two. The mode number of students per teacher is actually lower than expected, sitting around ten. However, since this is the total students divided by total number of teachers in each school, this does not represent real class sizes, which may be larger.

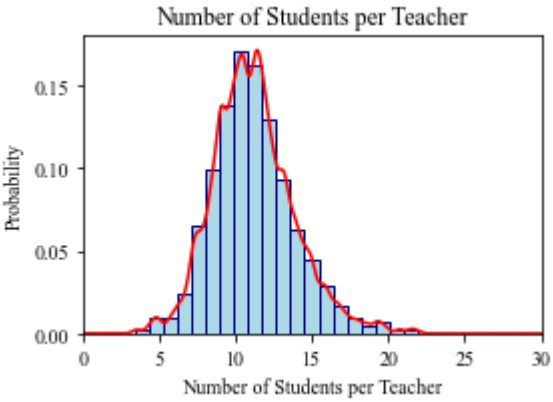

Fig. 5. Histogram Showing the Distribution of the Number of Students Per Teacher in NYS Schools

Continuing our observations on teachers, Figure 6 is a histogram that provides a similar visualization on uncertified teachers (teachers that are teaching outside of their certification). The histogram shows the percentage of uncertified teachers in schools and their probability of occurring. Most commonly, 0% to 20% of teachers in schools were teaching outside of their certification, with 0% being the mode for uncertified teachers. Very few schools have more than 60% of their teachers that are uncertified. This conducive to expectation because a certification is required within a five year period from the start date of teaching. Thus, newer teachers may be uncertified but will become certified shortly after they start teaching. Also, some teachers are possibly certified but in a different area of study than the area they are currently teaching. The majority of teachers should be certified, because certification is a requirement to be a permanent teacher in New York state.

As seen in Figure 6, a kernel density plot of graduation, graduation rates are heavily left (negative) skewed. In this

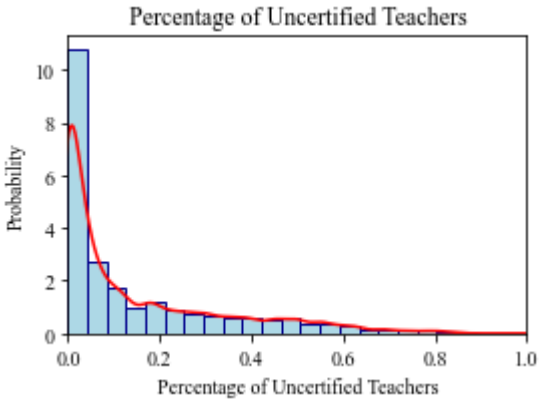

Fig. 6. Histogram Showing the Percentage of Uncertified Teachers

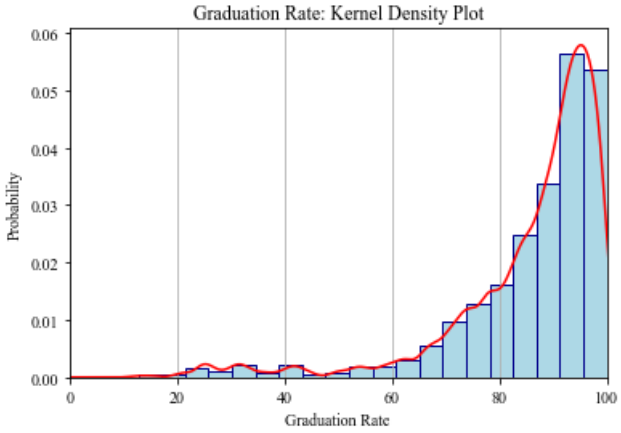

Fig. 8. Graduation Density Kernel Plot

plot, the mode for graduation rates is approximately 95% to 100%. In an effort to see effects of all of the data attributes on graduation rates, we did principle component analysis (PCA) to reduce the feature space to two dimensions. The results of the PCA dimensions are visualized on a 3D-grid with graduation rate on Figure 7. All of the points are clustered in one corner at roughly 60% to 80% graduation rates, but there is not much to be learned from the graph. Only 55.5% of the total variance is explained by the two PCA dimensions, and the rest of the variance is compressed.

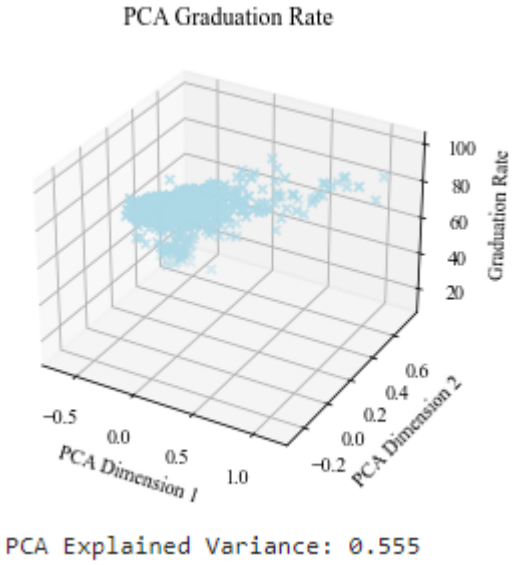

Fig. 7. Grid of Two-Dimensional PCA Compared with Graduation Rate

In the model pipeline, we conducted feature engineering to calculate select feature proportions (Table I). In addition, we encoded the Overall Status of schools and used min-max (0-1) scaling on Cohort Size, Federal Expenditure per Student, State and Local Expenditure per Student, Teachers per Pupil, Inexperienced Teachers per Pupil, Inexperienced Teachers per Teacher, Principals per Pupil, Inexperienced Principals per Pupil, and Inexperienced Principals per Principal.

Feature	Function
Teachers per Pupil	$\frac{\text{Number of Teachers}}{\text{Cohort Size}}$
Inexperienced Teachers per Pupil	$\frac{\text{Number of Inexperienced Teachers}}{\text{Cohort Size}}$
Inexperienced Teachers per Teacher	$\frac{\text{Number of Inexperienced Teachers}}{\text{Number of Teachers}}$
Principals Per Pupil	$\frac{\text{Number of Principals}}{\text{Cohort Size}}$
Inexperienced Principals per Pupil	$\frac{\text{Number of Inexperienced Principals}}{\text{Cohort Size}}$
Inexperienced Principals per Principal	$\frac{\text{Number of Inexperienced Principals}}{\text{Number of Principals}}$

TABLE I
TABLE OF FEATURE ENGINEERING FOR THE MODEL PIPELINE

IV. HYPOTHESIS AND GOALS

We hypothesize municipality expenditures exhibit diminishing returns on graduation rates following a threshold. We aim to answer: Does the number of teachers instructing outside of their area of certification negatively impact graduation rates? Is there a threshold of those teachers outside of certification impacting the number of students graduating? Is there a change in marginal graduate rate gain from incremental increases in government funding? How can educational resources be better allocated?

V. RESULTS

Our initial expectation was that schools with higher expenditures per student will have higher graduation rates. Moreover, we expected the number of teachers instructing outside of their areas to negatively impact the graduation rates of schools. Through our data pre-processing and model testing, we observed that these initial expectations did not appear to hold.

Based on the correlation matrix, Figure 1, there does not seem to be a positive correlation between increase in

expenditure and graduation rate, in fact, there is a negative correlation between the two variables (both state/local and federal). Additionally, a slightly negative correlation exists between graduation rate and the number of uncertified teachers and inexperienced teachers.

From model testing, we found several optimal combinations of features to produce higher predicted graduation rates. Table II displays the the best performing compositions for the following features: Total Funding per Student, Teachers per Pupil, Teacher per Inexperienced Teacher, Number of Principals, Overall Status, and Predicted Graduation Rate. Three of the attributes remain constant in the optimal combinations, with five teachers per pupil, one principle per school, and an overall status of "Good Standing." The other three features in the table have a small range. The optimal Total Funding per Student falls in the range of $13,580 and $17,340, and the number of Teacher[s] per Inexperienced Teacher lies between six and ten for a Predicted Graduation Rate of 96.5% to 97.0%. Within the range shown on this table (Table II), as Total Funding per Student increases so does the Predicted Graduation Rate.

Total Funding per Student	Teachers per Pupil	Teacher per Inexperienced Teacher	Number of Principals	Overall Status	Predicted Graduation Rate
$13,580	5	6	1	Good Standing	96.5
$15,600	5	8	1	Good Standing	97.2
$16,810	5	10	1	Good Standing	97.0
$17,280	5	7	1	Good Standing	97.0
$17,340	5	6	1	Good Standing	97.0

TABLE II
TABLE OF OPTIMAL COMBINATIONS OF FEATURES FOR GRADUATION PREDICTIONS

We trained multiple algorithms - Elastic Net, SVR, Bayesian Ridge, AdaBoost Regressor, Random Forest Regressor, and Neural Network models, with an exhaustive grid search, then recorded maximum error, mean absolute error, mean squared error, and median squared error for each (Table III). From the results in Table III, it was decided that SVR produced the best output, median squared error (3.863). Though the Random Forest Regressor exhibited the smallest mean absolute error (6.41), and SVR demonstrated the second smallest mean absolute error (6.571), we decided the median squared error was more pertinent for our analysis.

Metric	Elastic Net	SVR	Bayesian Ridge	AdaBoost Regressor	Random Forest Regressor	Neural Network
Maximum Error	48.865	66.081	45.522	65.827	58.585	60.66
Mean Absolute Error	7.422	6.571	7.383	7.582	6.41	7.393
Mean Squared Error	119.415	115.972	118.08	118.254	94.777	126.727
Median Squared Error	5.147	3.863	5.072	5.971	4.12	4.619
Models Subject to Exhaustive Grid Search						

TABLE III
TABLE OF MODELS AND THEIR PERFORMANCE METRICS

Coinciding with the patterns we observed in our data analysis, potential outlier schools exists in the validation

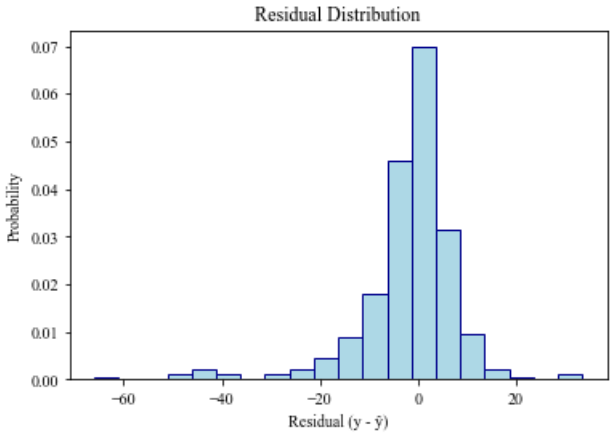

Fig. 9. Histogram Showing Residual Distribution

partition (Figure 3), our model lacks the capacity to accurately forecast these extremely low performing schools relative to the average school. Figure 9 depicts the histogram of the residual distribution, the distribution is left-skewed; hence, the SVR model does not accurately predict under-performing schools with graduation rates lower than approximately 50

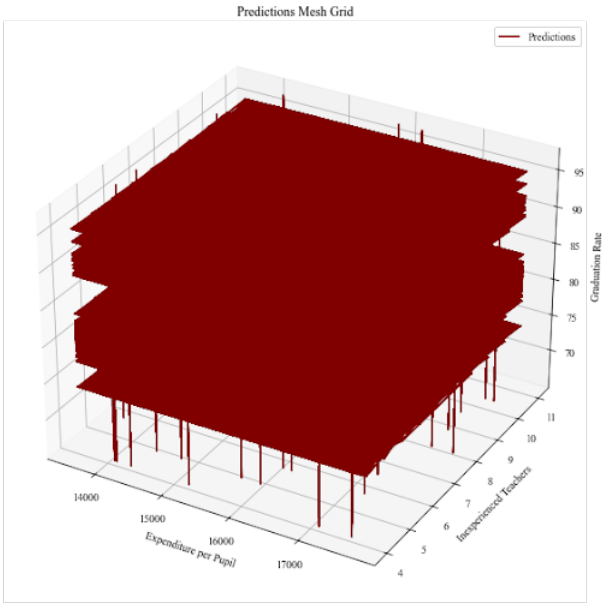

Fig. 10. Mesh Grid of Predictions for Graduation Rate Based on the Expenditure per Pupil and Inexperienced Teachers Features

We sampled 20,000 records from the feature space domain to get a clearer picture of the distribution of our model; Figures 10 through 13 each use the 20,000 samples. In Figure 10, a 3D-mesh grid displays graduation rate based on Expenditure per Pupil and Inexperienced Teachers. The x-axis is Expenditure per Pupil, the y-axis is Inexperienced Teachers, and the z-axis is graduation rate. This figure shows distinct stacked hyper-planes. More information is needed for better analysis because no clear or obvious relationship exists between these features.

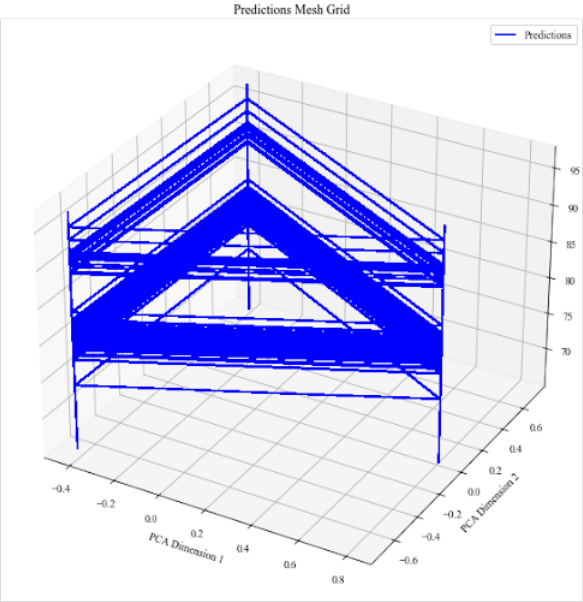

Fig. 11. Mesh Grid of Predictions for Graduation Rate Based on Based on Two-Dimensional PCA

We employed PCA to collapse the feature space into two dimensions, we plotted the mesh grid results in Figure 11. There is a triangular, stacked structure in this mesh grid. This figure shows that the explained variance is approximately 93.5%. The outer pillars point to optimal conditions to obtain higher graduation rates. There are some unobserved features that are influencing outcomes, such features could be district employment rates, district GDP, district crime rate, and more.

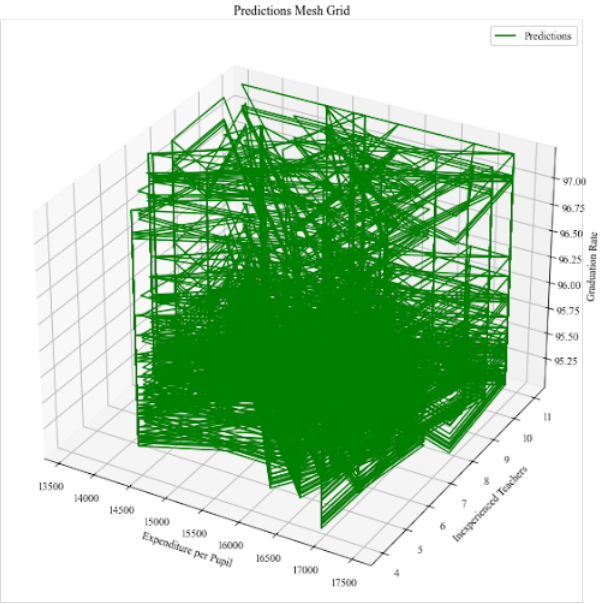

Fig. 12. Mesh Grid of Predictions for Graduation Rate Based on a Cross Section of the Expenditure per Pupil and Inexperienced Teachers Features

Figure 12 displays a cross-section of the data for Expenditure per Pupil and Inexperienced Teachers with Graduation Rate. A couple optimal points protrude around $14,000 spent per student and five to seven inexperienced teachers, this figure indicates no clear relationship between the Expenditure per Pupil and Inexperienced Teacher features.

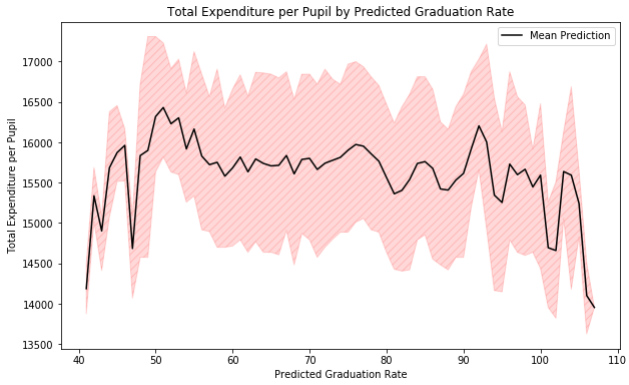

Fig. 13. Graph of Total Expenditure per Pupil and the Predicted Graduation Rate

Figure 13 shows the Total Expenditure per Pupil and the predicted graduation rate. The The black line represents the mean Total Expenditure per Pupil. The red shading on this figure represents the Total Expenditure per Pupil interquartile range. This figure tells us that the Total Expenditure per Pupil does not exhibit diminishing returns.

VI. Conclusion

We found that federal, state, and local funding does not demonstrate diminishing returns, but the quality of teachers and principals impacts graduation rates. Thus, if schools are given adequate funding and employ quality teachers, they are doing what is needed to optimize their graduation rates. In addition, we found some optimal combinations of funding and teacher qualifications for public schools in New York State, shown in Table II. This work can be continued as more data is acquired over time. The analysis could be expanded to other states to improve their graduation rates, and that would permit state interstate comparisons. We hope these results could be used to inform schools on the impacts of funding, teacher experience, and teacher certification on overall graduation rates.

REFERENCES

[1] M. M. Baydu, O. Kaplan, and A. Bayar, "Facing The Influence of Poverty on Graduation Rates in Public High Schools," Procedia - Social and Behavioral Sciences, vol. 84, pp. 233–237, Jul. 2013, doi: 10.1016/j.sbspro.2013.06.541.

[2] H. S. Bloom, S. L. Thompson, and R. Unterman, "Transforming the High School Experience: How New York City's New Small Schools are Boosting Student Achievement and Graduation Rates," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper 1786966, Jun. 2010. Accessed: Apr. 27, 2022. [Online]. Available: https://papers.ssrn.com/abstract=1786966

[3] D. Boyd, H. Lankford, S. Loeb, J. Rockoff, and J. Wyckoff, "The narrowing gap in New York City teacher qualifications and its implications for student achievement in high-poverty schools," Journal of Policy Analysis and Management, vol. 27, no. 4, pp. 793–818, 2008.

[4] C. Buerger and R. Bifulco, "The effect of charter schools on districts' student composition, costs, and efficiency: The case of New York state," Economics of Education Review, vol. 69, pp. 61–72, Apr. 2019, doi: 10.1016/j.econedurev.2019.01.003.

[5] W. D. Duncombe, "Estimating the Cost of an Adequate Education in New York," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper 1808956, Feb. 2002. doi: 10.2139/ssrn.1808956.

[6] A. Hanushek, "The Impact of Differential Expenditures on School Performance," Educational Researcher, vol. 18, no. 4, pp. 45–62, May 1989, doi: 10.3102/0013189X018004045.

[7] T. J. Kane, J. E. Rockoff, and D. O. Staiger, "What does certification tell us about teacher effectiveness? Evidence from New York City," Economics of Education Review, vol. 27, no. 6, pp. 615–631, Dec. 2008, doi: 10.1016/j.econedurev.2007.05.005.

[8] A. E. Schwartz, R. Rubenstein, and L. Stiefel, "Why Do Some Schools Get More and Others Less? An Examination of School-Level Funding in New York City," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper 1508434, Nov. 2009. Accessed: Apr. 24, 2022. [Online]. Available: https://papers.ssrn.com/abstract=1508434

[9] "Teaching and Educational Leadership Standards", New York State Education Department. [Online]. Available: http://www.nysed.gov/.