Educational Equity in New Jersey:

Analyzing Relationships Between Resources and Outcomes Across School Districts

Christy Hernandez
Pace University - CS 668 - Analytics Capstone - Fall 2024

Abstract / Introduction

New Jersey, a state renowned for its diverse population and mix of urban, suburban, and rural communities, faces persistent disparities in educational resources and student outcomes. This presents a significant challenge to achieving educational equity. Such disparities directly influence student performance and opportunities, perpetuating systemic inequities in underserved communities. Addressing these challenges is essential to creating a fair and effective educational system that benefits all students, regardless of their socioeconomic background or geographical location.

This project seeked to analyze educational equity across New Jersey's school districts, focusing on the relationships between key resources and student outcomes. By utilizing a publicly available dataset, this study examined critical variables such as teacher education level, graduation rate, and student demographics to identify the most impactful disparities. Employing machine learning models, the project aimed to find which factors had the most impact on student success. The findings aim to provide actionable insights for policymakers, enabling targeted interventions and data-driven decisions to promote equitable resource distribution and close achievement gaps. When all is said and done, hopefully this project can contribute to a more inclusive and just educational system in New Jersey.

Data Description / Literature Review

The analysis of educational equity in New Jersey required detailed and comprehensive data that could provide insights into both resource allocations and student outcomes across the state's school districts. For this purpose, a single dataset was sourced from the New Jersey Department of Education's website: the *2022-2023 Performance Reports Database*. This dataset was selected because it offers extensive district- and state-level information necessary for the project, encompassing both input variables like teacher-student ratios, as well as outcome metrics such as test scores and graduation rates.

The dataset, named *Database_DistrictStateDetail.xlsx*, is a 26,249 KB Excel file containing 72 individual sheets of information. Each sheet focuses on a specific aspect of district or state-level performance, providing a granular view of educational data for the 2022-2023 academic year. These sheets include details on district demographics, standardized test scores, graduation rates, financial allocations, and other critical indicators. The richness of this dataset ensures that the analysis covers a wide range of factors impacting educational equity.

A dataset from the New Jersey Department of Education (NJDOE) was chosen because it is a comprehensive, authoritative, and reliable source for educational data within the state. As an official government entity, the NJDOE collects, verifies, and disseminates detailed information about schools, students, and educators to support transparency, policy-making, and research. The data is gathered directly from school districts and undergoes stringent quality control measures to ensure its accuracy and reliability. Additionally, the NJDOE's adherence to state and federal regulations, such as the Every Student Succeeds Act (ESSA), further reinforces its credibility. ESSA, a federal law enacted in 2015, emphasizes accountability, equity, and quality in education by requiring states to monitor and report student performance data, particularly for disadvantaged subgroups. This legislation ensures that data collection processes meet stringent standards and focus on meaningful measures of educational success.

Further supporting the dataset's value, the NJDOE provides extensive resources to assist researchers and policymakers in understanding and using the data. On the NJDOE's website, two key documents can be found: a detailed 104-page reference guide and an 8-page FAQ document, both of which provide insights into the methodology, definitions, and context for the data used in this project. These resources not only underscore the thoroughness of the dataset but also serve as a testament to the NJDOE's commitment to transparency and accessibility.

Using this dataset not only ensures that the analysis is grounded in trustworthy information but also allows the findings to align closely with real-world educational trends and policy contexts in New Jersey. The official nature of this source, coupled with its commitment to providing up-to-date and standardized data, makes it a strong foundation for this project's analysis.

Methodology

To address the objectives, this project employed a multi-faceted approach that began with data exploration and preprocessing, followed by regression modeling and feature importance analysis. The process was structured to ensure that insights were data-driven and actionable. It began by loading the dataset into Python using the pandas library, which as previously discussed consisted of 72 sheets in an Excel file. After the careful manual review of each sheet, nine sheets were selected that were most relevant to the project objectives. These sheets contained diverse metrics, including enrollment trends, standardized test participation, advanced coursework participation, graduation profiles, absenteeism, school incidents, teacher experience, staff-to-student ratios, and educator qualifications. Each sheet includes columns for key variables, providing critical information that allowed the models to comprehensively assess the factors influencing four-year graduation rates.

To designate the target variable, the 'Graduates' column in the '4YrGraduationCohortProfile' sheet was identified, isolating entries where the 'StudentGroup' column was 'Districtwide'. This yielded the districtwide four-year graduation rates, which were used as the dependent variable in subsequent analyses. To facilitate visualization and further data exploration, the project utilized the Google Maps API on Google Cloud to create a geographic heatmap overlay, normalizing graduation rates against state averages to highlight district-level trends.

Predictors were categorized into three groups: family-level, school-level, and student-level factors. Each group was analyzed independently before merging all features into a comprehensive dataset. Family-level factors included the percentages of homeless students, students in foster care, military-connected students, and economically disadvantaged students, sourced from the 'EnrollmentTrendsByStudentGroup' sheet. The processed data for this factor category and every one after was stored in pandas DataFrames for streamlined manipulation and further analysis. The resulting DataFrames host a wealth of information, with our chosen key variables as the columns and each row representing a different school district in New Jersey. Here within the family-level factors, a function called *add_graduation_rate* was also developed to merge features with the graduation rates, handle missing values, and enable reuse in subsequent analyses. For the school-level category, incorporated factors included average teacher experience, percentage of out-of-field teachers, student-to-teacher ratios, and percentages of teachers with advanced degrees. Features like ratios were formatted consistently, such as removing suffixes, and entries were filtered to include only teacher-related data. Student-level factors included standardized test participation rates (SAT, ACT, PSAT), advanced coursework participation (AP/IB courses and exams), chronic absenteeism, and incidents per 100 students enrolled. Features were aggregated and filtered as necessary, such as limiting absenteeism data to 'Districtwide' entries. Each dataset was preprocessed to ensure compatibility for modeling, including converting categorical features to numeric formats where applicable.

The project applied two regression models—Random Forest and XGBoost—to predict four-year graduation rates. Both models were tuned using GridSearchCV to identify optimal hyperparameters. An 80/20 train-test split was employed for all datasets, and model performance was evaluated using mean squared error (MSE) and R-squared (R^2) metrics. Feature importance and permutation importance analyses were conducted for each model to assess the influence of individual predictors on the target variable. Additionally, correlation coefficients were calculated for all features to understand the relationship between predictors and graduation rates. Features with a correlation magnitude of at least ±0.3 were considered to have a meaningful relationship, with those above ±0.5 classified as moderately correlated. This analysis informed the selection of features for additional focused modeling.

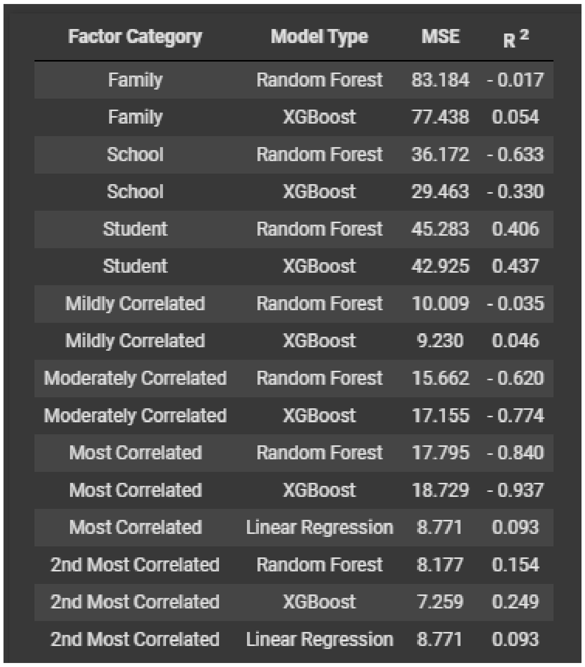
Subsequent modeling focused on subsets of features based on their correlation strength. For mildly correlated features (|±0.3|), the Random Forest and XGBoost models were applied, and their performance metrics were compared. For moderately correlated features (|±0.5|), focused models were trained using the two most correlated features, 'Chronic_Abs_Pct' and 'Economically Disadvantaged Students.' Models trained exclusively on these two features individually were also evaluated to assess their standalone predictive power. Given the strength of these relationships, a Linear Regression model was included for comparison in these two cases. This structured approach enabled a comprehensive understanding of the most influential factors and their predictive power regarding four-year graduation rates.

Analysis / Results

This study aimed to evaluate the impact of various student-, school-, and district-level factors on four-year graduation rates using Random Forest, XGBoost, and Linear Regression models. The analysis tested the following hypotheses: 1) economic disadvantage is negatively correlated with graduation rates, with this correlation being sufficiently strong to influence predictive performance; 2) machine learning models, such as Random Forest and XGBoost, will outperform Linear Regression, with XGBoost being the best-performing model; and 3) chronic absenteeism is the strongest predictor of graduation rates and will have the most significant relationship with graduation outcomes across all models.

The results largely supported these hypotheses. Chronic absenteeism emerged as the most significant predictor of graduation rates across all models, confirming that it has the strongest relationship with graduation outcomes, as anticipated in the third hypothesis. It was consistently identified as the most important feature in terms of predictive value. Economic disadvantage, while showing a negative correlation with graduation rates, was not as strongly correlated as expected. The correlation was moderate, suggesting that while economic disadvantage has an influence, it is not as dominant as chronic absenteeism in predicting graduation rates.

In terms of model performance, machine learning models (Random Forest and XGBoost) did indeed outperform Linear Regression, as hypothesized. XGBoost demonstrated the best performance, with the lowest Mean Squared Error (MSE) for chronic absenteeism (7.259), compared to Random Forest (8.177) and Linear Regression (8.945). However, none of the models achieved high predictive power, with R-squared values remaining low across all models. The highest R-squared value was 0.249 for XGBoost when analyzing chronic absenteeism, suggesting that while the models identified significant predictors, they did not provide strong overall explanatory power.



Factor Category	Model Type	MSE	R^2
Family	Random Forest	83.184	- 0.017
Family	XGBoost	77.438	0.054
School	Random Forest	36.172	- 0.633
School	XGBoost	29.463	- 0.330
Student	Random Forest	45.283	0.406
Student	XGBoost	42.925	0.437
Mildly Correlated	Random Forest	10.009	- 0.035
Mildly Correlated	XGBoost	9.230	0.046
Moderately Correlated	Random Forest	15.662	- 0.620
Moderately Correlated	XGBoost	17.155	- 0.774
Most Correlated	Random Forest	17.795	- 0.840
Most Correlated	XGBoost	18.729	- 0.937
Most Correlated	Linear Regression	8.771	0.093
2nd Most Correlated	Random Forest	8.177	0.154
2nd Most Correlated	XGBoost	7.259	0.249
2nd Most Correlated	Linear Regression	8.771	0.093

Fig 1. Model evaluation table by factor category

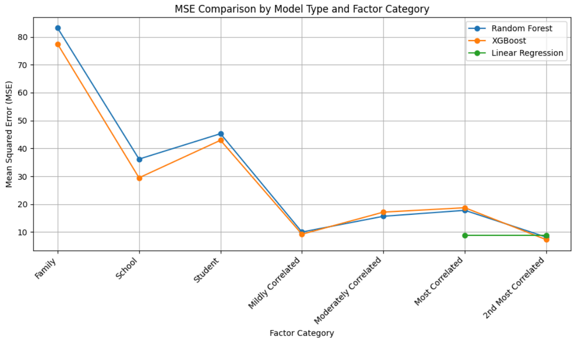


Fig 2. MSE chart by model type and factor category

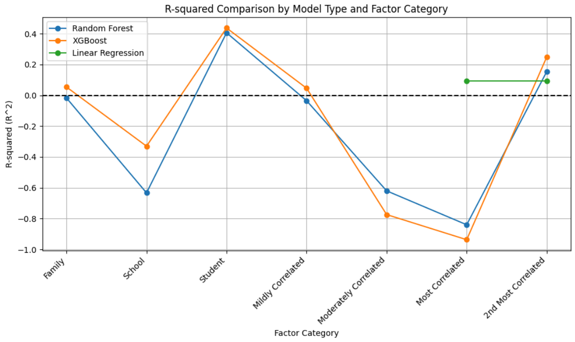


Fig 3. R^2 chart by model type and factor category

Figures 1, 2, and 3 further illustrate the model performance and the importance of chronic absenteeism. Figure 1 shows the Model Evaluation Table, comparing MSE and R-squared values for each model and factor category. Figure 2 displays the slight differences in MSE for each model and factor, with XGBoost performing better in all instances. Figure 3 highlights the low R-squared values across models, emphasizing the limited explanatory power despite identifying significant predictors like chronic absenteeism.

Limitations

Several limitations impacted the scope and outcomes of this analysis. First, the low and often negative R-squared values across all models indicate that the selected factors only partially explain graduation rates. For instance, the highest R-squared score of 0.249, achieved by XGBoost for chronic absenteeism, reflects limited explanatory strength. The analysis was further constrained by the dataset and the chosen predictors, excluding other potentially significant factors such as community resources, school policies, or teacher quality.

Time constraints of less than three months to complete the project significantly limited the complexity and depth of the analysis. This constraint influenced the selection of models and features, potentially reducing the robustness of the findings. Additionally, limited access to data on the long-term impact of educational interventions made it challenging to track life outcomes such as employment or socioeconomic mobility, which could provide deeper insights into the predictive factors of graduation rates. The complexity of analyzing data across diverse districts with varying socioeconomic conditions introduced additional challenges in drawing generalized conclusions.

Another limitation lies in the distinction between correlation and causation. While chronic absenteeism strongly correlated with graduation rates, this relationship does not imply causation. Underlying factors, such as family circumstances or school climate, may drive both absenteeism and graduation outcomes. The short time frame also restricted the exploration of more complex models or advanced techniques that could have improved predictive performance. Despite efforts in hyperparameter tuning and cross-validation, the models struggled to generalize beyond the dataset. Random Forest and XGBoost performed better than Linear Regression, but their predictions remained modest, with high MSE values and low R-squared scores. Feature importance rankings, while consistent, are inherently limited in their ability to identify causal relationships, as they only reflect statistical associations captured by the models.

Potential data bias may have also affected the models' performance. For example, socioeconomic data may not fully capture the nuances of economic disadvantage, and missing information within the dataset could skew results. The limited availability of long-term data on educational interventions further restricted the ability to assess broader life outcomes, such as employment or socioeconomic mobility. The challenges of analyzing data across diverse districts with varying socioeconomic conditions compounded the difficulty of drawing generalized conclusions. While these limitations highlight gaps in predictive performance, they also underscore the need for further research that incorporates additional predictors, explores alternative modeling approaches, and addresses data limitations to enhance the reliability and applicability of findings.

Future Work

Future work could build upon the current study in several meaningful ways. First, expanding the dataset to include variables that capture broader social and environmental influences, such as neighborhood socioeconomic conditions, school funding sources, and access to extracurricular programs, could enhance the predictive power of the models. Additionally, longitudinal data tracking students over multiple years would allow for a more dynamic understanding of how factors evolve and interact to influence graduation rates.

Another promising avenue for future research involves integrating advanced machine learning techniques, such as neural networks or ensemble learning methods, to uncover complex, non-linear relationships between predictors and outcomes. These methods could help address the relatively low performance metrics observed in the current models. Additionally, including feature engineering techniques to generate interaction terms or composite indices—such as combining absenteeism with economic disadvantage—might yield more nuanced insights into the synergies between variables.

From a policy perspective, these findings advocate for targeted interventions to reduce chronic absenteeism and mitigate the impact of economic disadvantage. For instance, initiatives aimed at improving student attendance, such as community-based support programs, family outreach, and technology-driven attendance monitoring, could have a tangible impact on graduation rates. Similarly, providing additional resources to economically disadvantaged students, such as free school meals, tutoring, and mental health support, could help bridge achievement gaps. Collaboration between school districts, policymakers, and local communities will be essential to translating these insights into effective action plans.

Finally, this study contributes to the broader discourse on educational equity by demonstrating the value of combining machine learning techniques with publicly available data. The methodological framework developed here could be adapted to explore similar questions in other states or countries, allowing for cross-regional comparisons and the identification of best practices. Moreover, the visualization tools developed as part of this project, such as the geographic heatmap overlay, offer policymakers an intuitive way to identify disparities and prioritize resources.

Discussion / Conclusion

This project offers valuable insights into the factors influencing educational equity and student outcomes across New Jersey school districts, particularly with respect to four-year graduation rates. By employing data-driven methodologies, including Random Forest and XGBoost models, the study highlights the pivotal role of chronic absenteeism and economic disadvantage in shaping graduation rates. These findings underscore the complex interplay between student demographics, school resources, and educational outcomes, providing a foundation for further research and policy development.

One notable outcome is the confirmation that chronic absenteeism exerts the strongest influence on graduation rates among the variables studied. Economic disadvantage also demonstrated a negative correlation with graduation rates, but its predictive strength was less pronounced. Other features, such as teacher qualifications and standardized test participation rates, showed limited individual impact, suggesting that a more nuanced combination of factors may be required to fully understand educational disparities.

Despite these findings, the relatively low R-squared values across all models indicate that graduation rates are influenced by additional unmeasured or latent factors not included in the dataset. This highlights the limitations of relying solely on quantitative data to capture the multifaceted nature of educational equity. The inclusion of qualitative data, such as student and teacher interviews, community feedback, and policy evaluations, could provide a richer context for interpreting these results.

In conclusion, while this study sheds light on some of the key factors influencing graduation rates in New Jersey, it also reveals the inherent complexity of educational equity. No single factor or model can fully explain the disparities observed, emphasizing the need for a holistic approach that considers both quantitative and qualitative dimensions. By continuing to refine analytical techniques, expand datasets, and engage stakeholders, future research can help drive meaningful progress toward a more equitable education system. Ultimately, addressing these challenges is not just a matter of academic interest but a societal imperative to ensure that all students, regardless of their background, have the opportunity to succeed.

Sources

The State of New Jersey. (2019, September 26). *NJ SCHOOL PERFORMANCE REPORT*. New Jersey Department of Education. https://rc.doe.state.nj.us/