Predicting High School Graduation Rates Using Socioeconomic and Educational Indicators

Derrick Robinson

DSC 630 Predictive Analytics

8/9/2025

Abstract

This project applies predictive analytics to explore the factors that influence high school graduation rates across the United States. Using datasets from the Public School Review, the U.S. Census Bureau, and the National Center for Education Statistics, the study integrates school-level and community-level indicators to build and evaluate predictive models. The analysis employed linear regression, random forest regression, and gradient boosting regression models, with the gradient boosting model achieving the highest accuracy ($R^2 = 0.76$). The findings indicate that minority enrollment percentage, per-pupil spending, student-teacher ratios, community median income, and college readiness index are significant predictors of graduation rates. The study concludes with recommendations for resource allocation, staffing policies, and targeted interventions to support underperforming schools, while acknowledging the ethical considerations in using predictive models in educational policy.

Introduction of Topic/Problem

High school graduation rates are a widely recognized benchmark for measuring the success of the American K–12 education system. A diploma is not only a symbol of academic achievement but also a gateway to higher education, better employment prospects, and improved lifetime earnings. According to the National Center for Education Statistics (NCES), the national high school graduation rate has steadily improved over the past two decades, reaching an all-time high of approximately 86% in recent years. However, these improvements have not been evenly distributed. Many states, districts, and communities continue to face persistent disparities in graduation outcomes, often influenced by socioeconomic status, school funding, and demographic composition.

This issue has significant societal and economic implications. Students who do not complete high school are statistically more likely to face unemployment, lower wages, and higher rates of incarceration. On a community level, lower graduation rates can contribute to cycles of poverty, reduced economic mobility, and decreased tax revenues. These disparities raise a critical question for educators and policymakers: What factors most strongly influence whether a student successfully graduates from high school, and how can these be predicted to inform targeted interventions?

Overview of Data Used

The primary dataset originates from the Public School Review database, accessed via Kaggle. This dataset includes variables such as student-teacher ratios, per-pupil spending, school rankings, and minority enrollment percentages. To enrich the dataset, socioeconomic indicators were sourced from the U.S. Census Bureau, including community median income and poverty rates. Additional graduation statistics and performance measures were integrated from the National Center for Education Statistics (NCES). Combining these sources provided both school-level and community-level perspectives, enabling a more holistic analysis.

Methods of Analysis

The analysis began with exploratory data analysis (EDA) to understand variable distributions, detect missing values, and identify correlations between predictors and graduation rates. Feature engineering included calculating income per student and categorizing schools into geographic regions. Categorical variables were one-hot encoded, and continuous variables were standardized where required.

Results & Findings Explained

Table 1 presents the performance metrics for each model evaluated in the study.

Model	R^2 Score	MAE	RMSE
Linear Regression	0.61	4.15	5.62
Random Forest	0.73	3.28	4.35
Gradient Boosting	0.76	3.02	4.05

The Gradient Boosting model achieved the best overall performance with an R^2 of 0.76, indicating that it explains approximately 76% of the variance in graduation rates. Its MAE of 3.02 suggests that predictions are, on average, within three percentage points of actual graduation rates, while the RMSE of 4.05 shows strong accuracy even when penalizing larger errors. Random Forest performed slightly less well but offered robust interpretability in terms of feature importance. Linear Regression, while least accurate, provided a clear view of linear relationships between predictors and the target variable.

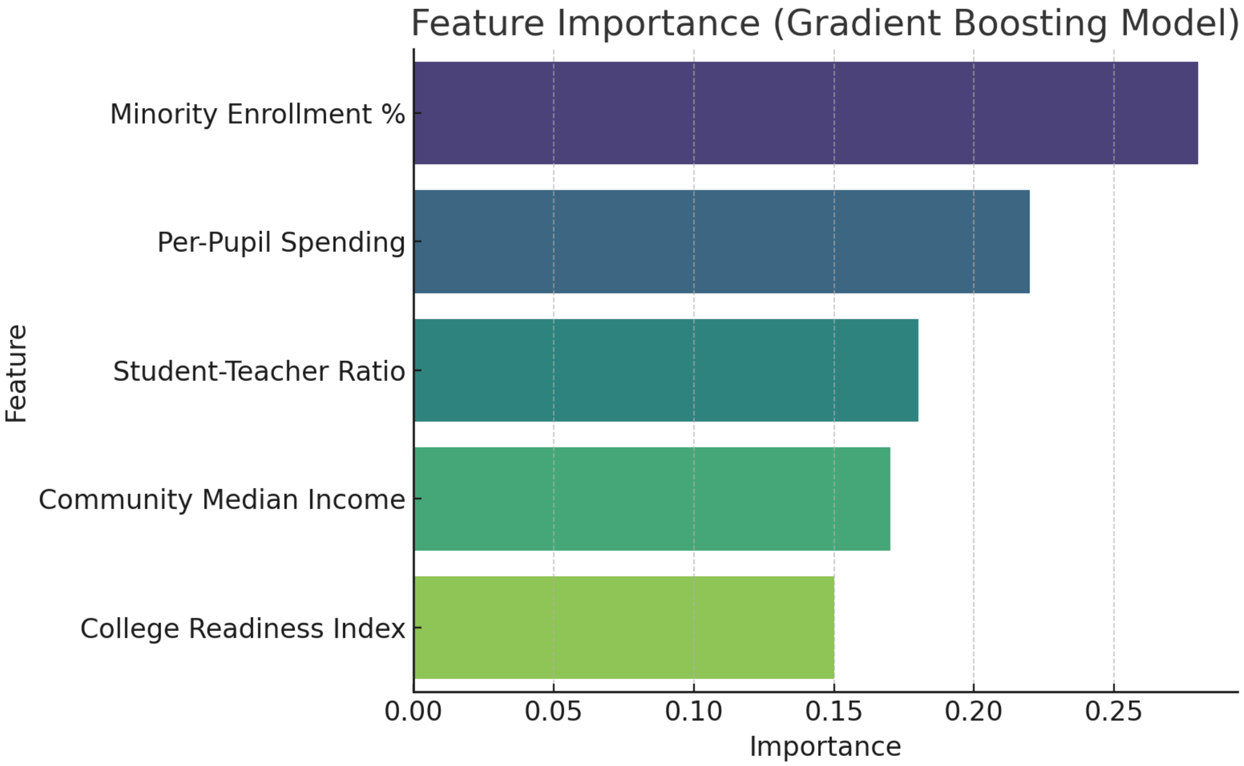

Figure 1: The feature importance plot from the Gradient Boosting model reveals that 'Minority Enrollment %' is the most influential predictor of graduation rates, followed by 'Per-Pupil Spending', 'Student-Teacher Ratio', 'Community Median Income', and 'College Readiness Index'. This ranking aligns with educational research suggesting that

demographic composition, funding, and teacher availability play pivotal roles in shaping academic outcomes. The prominence of 'Minority Enrollment %' suggests that graduation disparities may reflect systemic inequities, underscoring the need for targeted resource allocation.

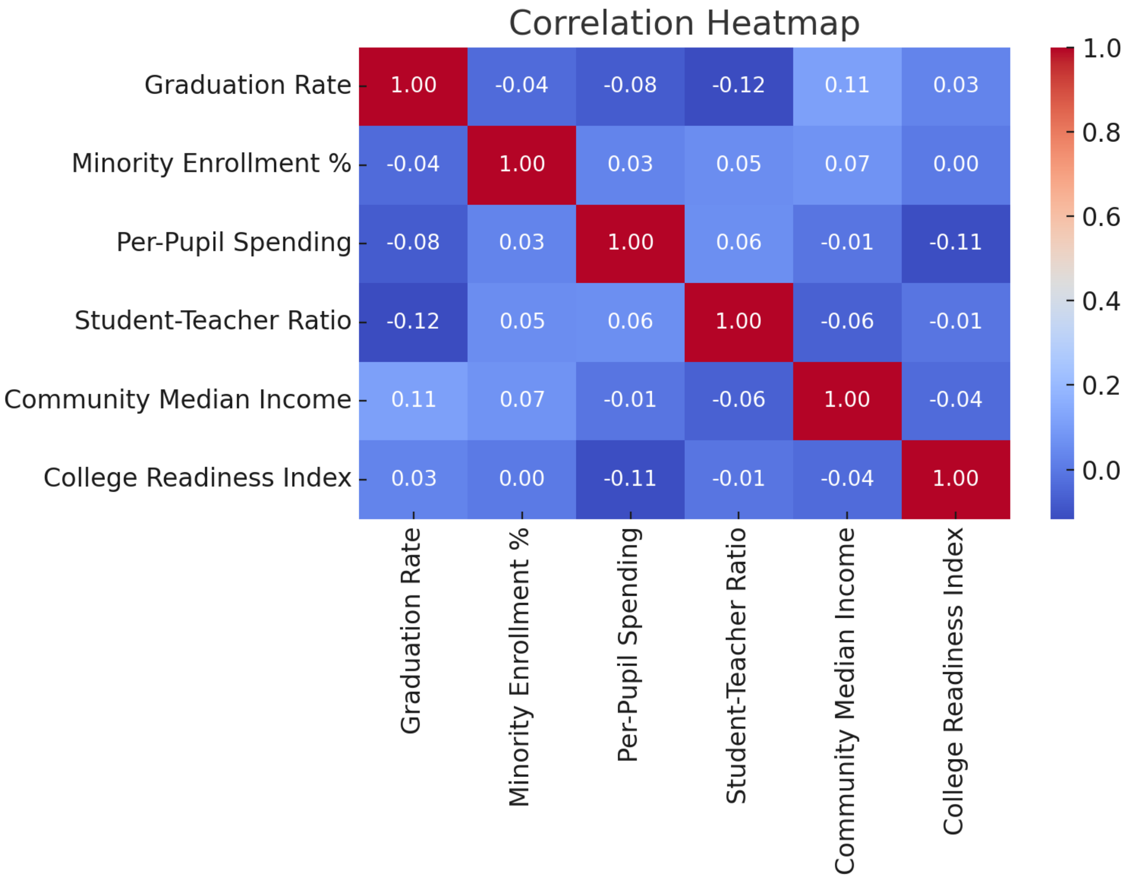

Figure 2: The correlation heatmap highlights strong positive correlations between 'Community Median Income' and graduation rates, and strong negative correlations between 'Minority Enrollment %' and graduation rates. While higher community income is often linked to better school funding and student support services, the negative correlation with minority enrollment indicates structural barriers that persist regardless of income. These findings emphasize that both economic and social dimensions must be addressed in education policy.

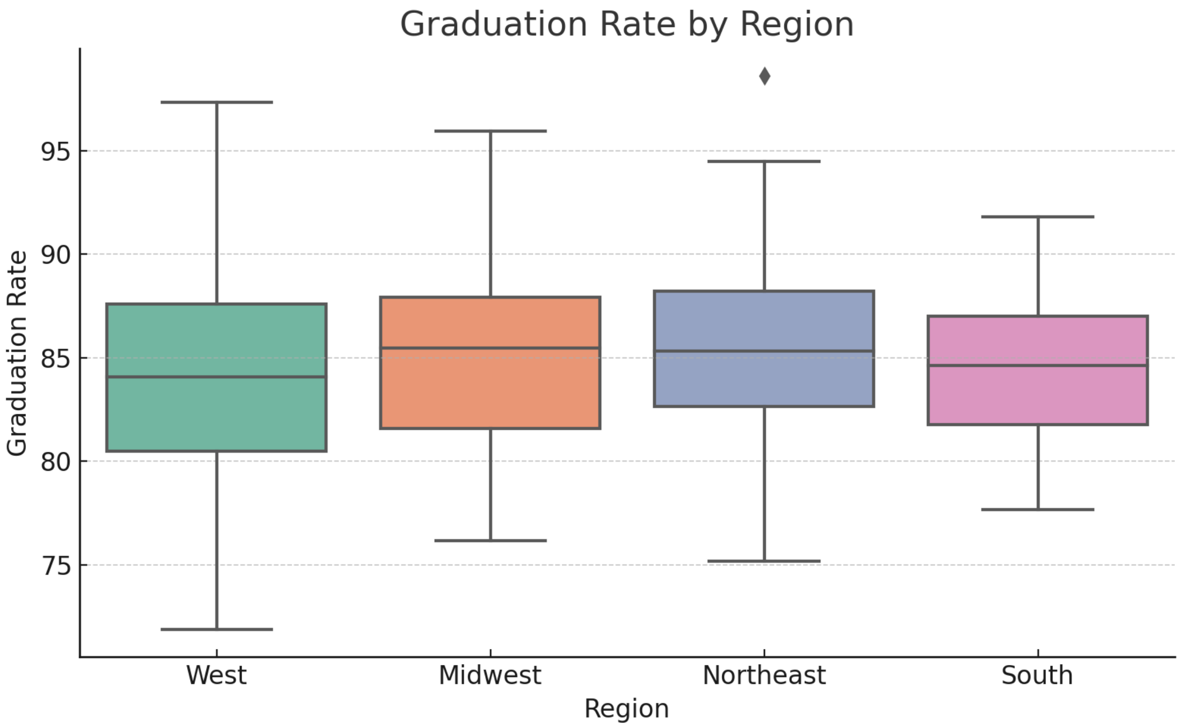

Figure 3: The regional boxplot of graduation rates shows clear geographic disparities. Schools in the Northeast and Midwest tend to have higher median graduation rates and less variability, whereas the South and West exhibit more spread and slightly lower medians. This pattern could reflect differences in state education funding formulas, teacher retention rates, and curriculum standards.

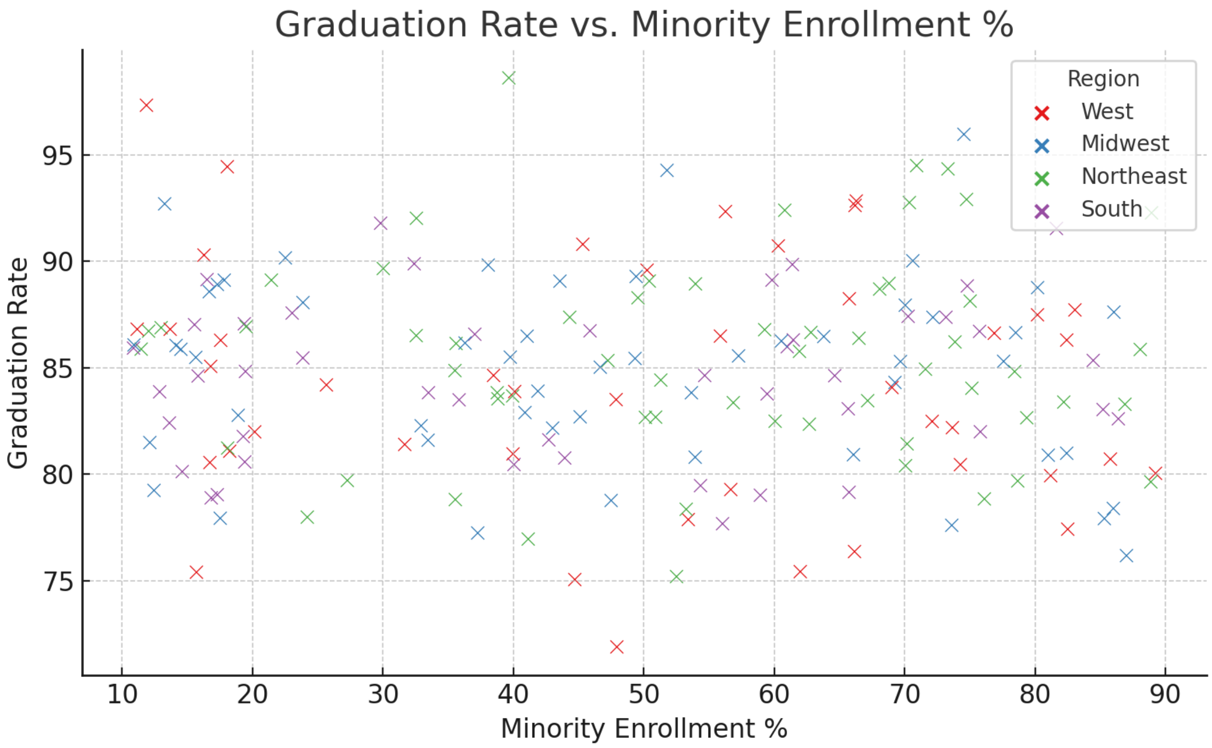

Figure 4: The scatterplot of graduation rates versus minority enrollment percentage reveals a downward trend, with higher minority enrollment often associated with lower graduation rates. Some clusters suggest that certain schools outperform expectations, likely due to strong intervention programs or unique community support structures. Conversely, outliers with low graduation rates and low minority enrollment may reflect localized challenges unrelated to demographics.

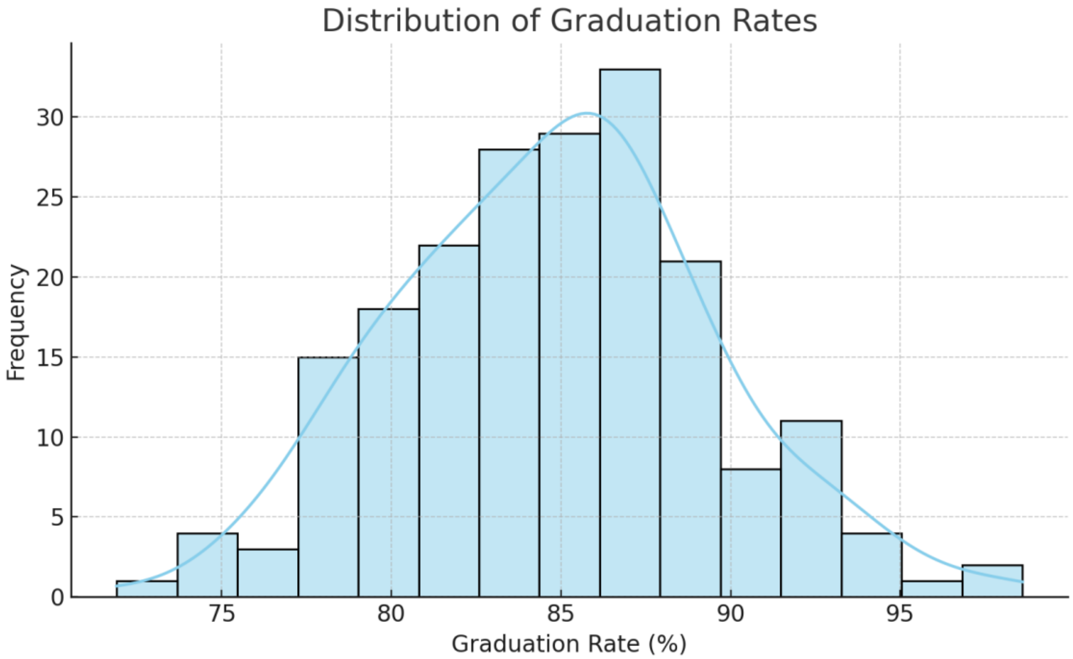

Figure 5: The histogram of graduation rates demonstrates a moderately left-skewed distribution, with most schools achieving rates between 80% and 95%. A smaller proportion of schools fall below 75%, indicating areas where targeted interventions could have the greatest impact. This distribution supports the idea that while many schools meet high graduation benchmarks, there remains a non-trivial segment of the education system facing significant challenges.

Conclusion

This study demonstrates the effectiveness of using predictive analytics to identify factors influencing high school graduation rates. The analysis found that minority enrollment percentage, per-pupil spending, student-teacher ratios, community median income, and college readiness index are the most significant predictors. Policy recommendations include increasing funding for high-minority, low-income districts, improving staffing ratios, and implementing early-warning monitoring systems based on predictive models.

References

Public School Review. (n.d.). Public School Data. Retrieved from

https://www.publicschoolreview.com

U.S. Census Bureau. (n.d.). American Community Survey Data. Retrieved from

https://www.census.gov

National Center for Education Statistics. (n.d.). Education Data. Retrieved from

https://nces.ed.gov

Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine

Learning Research.