# Predicting California High School Graduation and Post-Secondary School Enrollment Based on Socioeconomic and Geographic Factors

December 16, 2021

**Lydia DiBlasio**
University of Southern California
School of Engineering
diblasio@usc.edu

**Gauri Madhok**
University of Southern California
School of Engineering
gmadhok@usc.edu

**Isaac Wahout**
University of Southern California
School of Engineering
wahout@usc.edu

## Abstract

California is a large state, and thus has many high schools in areas that are rural, suburban, and urban, and have widely varying levels of poverty, population density, and income. These factors all contribute to the gap in high school graduation rates and pursuit of post-secondary enrollment. Along with these features, there are student-based factors that contribute to graduation and post-secondary enrollment rates such as test scores and absenteeism rates. The purpose of this project is to analyze the most important features contributing to student success in completing high school and their pursuit of higher education. Our project aims to provide school administration with actionable and effective ways to improve graduation and post-secondary enrollment rates based on their school's specific characteristics.

## 1 Introduction

One of the United Nations Sustainable Development Goals is to ensure quality education for all. In 2018, around 258 million children around the world were out of school [1]. For children who are in school, they may drop out prior to completing high school. In California, 84.3% of students who started high school in 2016 graduated in 2020 [2]. Of the students who graduated in California, graduation rates varied by economic status, housing situation, and several other factors. Of the students who graduated from high school in California, 64.4% enrolled in post-secondary education [2]. While these are the graduation and post-secondary enrollment rates at the state level for California, each school district has its own rate based on the characteristics of the schools.

From the data, it is clear that students who do not graduate or move on to post-secondary education are not getting the academic support they need. However, there is still a very limited understanding about which factors are the main influence for low high school graduation and pursuit of higher education. A portion of student success stems from personal aspects, such as location, absenteeism, socioeconomic status, and family stability [3]. While these circumstances are influential for high school completion, they may not holistically represent the entirety of student success. Other external factors may be just as prevalent for encouraging pursuit of post-secondary education. The factors we will look at in this project analyze school access to quality resources, such as education levels of staff. We analyzed student data such as test scores and absentee rates and related these features to graduation and post-secondary enrollment. We also quantified the impact of local features around the schools from U.S. census data, such as income level or population density.

The primary goal of our analysis is to inform schools on actionable features that they can improve within the school level to increase these graduation and postsecondary enrollment rates. This is important because it allows for school districts to best allocate the resources that they have given the specific static characteristics of the school and surrounding area. Additionally, we hope our analysis can highlight the factors that most contribute to increasing student graduation and post-secondary rates. We hope that the results of this analysis will contribute to our overarching goal of improving the equity and overall quality of education in the state of California.

# 2 Background and Related Work

In the past, research has examined factors that lead to students not graduating high school on time [3, 4, 5, 6, 7]. The factors analyzed in the prior research to predict student outcomes and graduation rates include academic performance, student demographics, economic status, and teacher wages. To identify these factors, several machine learning models were used including Random Forest, Adaboost, Logistic Regression, Support Vector Machines, and Decision Trees [7, 5, 4].

Prior research in the area has centered around predicting academic performance of students based on student specific features. Lakkaraju et al predicted whether students would graduate from high school on time or not based on student attributes including gender, age, ethnicity, GPA, standardized test scores, absence rates, and similar factors [7]. The machine learning models used by this study included Random Forest, Adaboost, Logistic Regression, Support Vector Machines, and Decision Trees. While the Random Forest model performed the best on the different school districts the study used, the medium size school district had significant variation in performance between models. The study found that GPA in 8th grade was the most significant predictor of whether a student would graduate high school on time [7]. Other research has looked at features outside of a student's control such as the wages teachers are paid. Loeb et al analyzed whether increasing teacher wages would have an impact on student outcomes and found that by increasing teacher wages by 10%, the dropout rate for students would decrease between 3 to 6% [6].

Given the previous research in this space, we have not found a study that combined student level, school level, district level, and region features to predict how all these factors impact graduation rates of students. Furthermore, we have not found research to identify which of all these factors has the greatest impact on graduation rates and how an increase in funding to a school could change student outcomes.

# 3 Datasets

Our final machine learning model will combine census data with school-level data collected by the California Department of Education. Our census data comes from averaging census tract-level features for all census tracts inside of the attendance zone of a particular district, based on American Community Survey data.

## 3.1 Processing Census Data

The census data came from American Community Survey data that was obtained via Social Explorer. The census data was processed using a Shapefile that grouped geographic regions together based on district codes. The data estimates a variety of features for 2019 at the census tract level given data from the previous 5 years. After discussing which features might be relevant for our purposes, we settled on 25 features to use in our analysis.

## 3.2 Processing School Data

The school data used in this project consists of school-level and district-level statistics from the California Department of Education. This data was collected and parsed to include high-school data from the school year of 2018-2019. The features utilized combined data on student achievement and performance, district funding amounts, population demographics, and socioeconomic status of the school. We merged these individual features on school code and district code and ended up with 1,940 rows and 23 features from this data collection.

## 3.3 Combining Datasets Together

For data preprocessing, we had a few steps. First, for the school data, several features were counts which we could not compare directly between schools, so we had to convert the counts to rates by dividing by the total enrollment. Next, there were several different datasets we processed at the school level, so we merged them together based on school code. Next, we merged the school data with the census features at the district level by merging on the district code. Once we had the combined dataset, we dropped features that were strings since they would not go through the model. We also dropped schools that had NaN values after merging which were 394 schools. After this, we ended up with 1544 high schools and 48 features.

# 4 Preliminary Exploratory Data Analysis

## 4.1 Correlated Feature Analysis

In order to determine any features in the dataset that are too highly correlated, we created a correlation matrix that provided the scores of how correlated any two features are. We did this individually for the census features and the school features.

For the school features, we created the following correlation matrix:

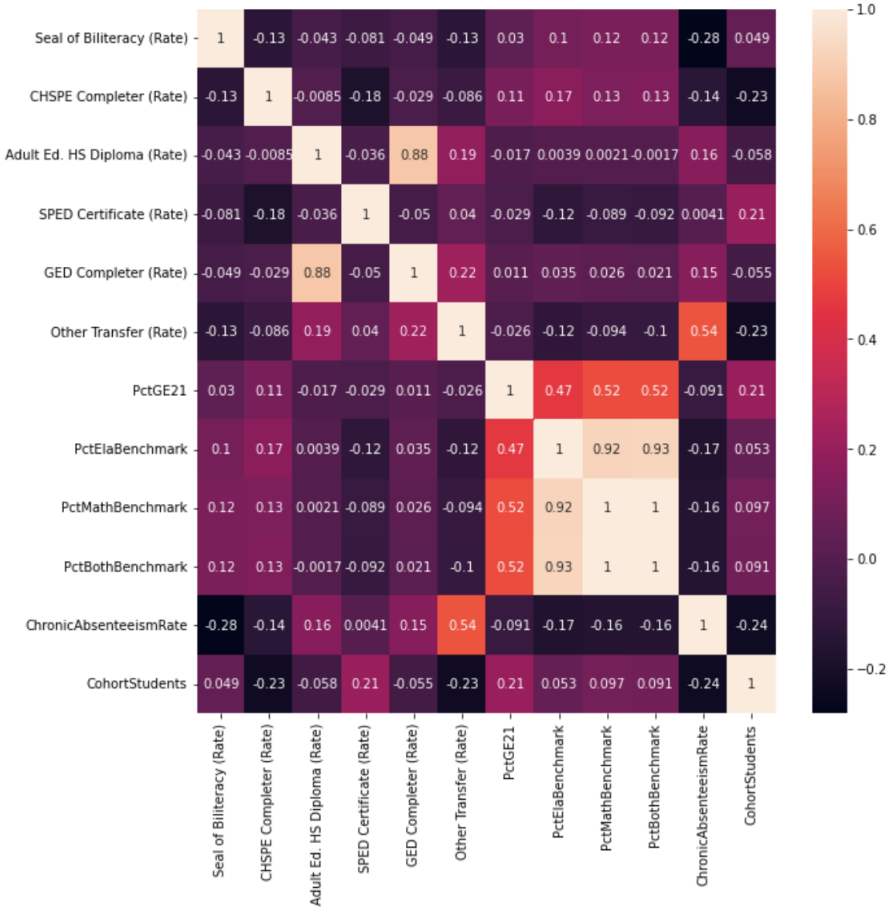


Figure 1: This figure shows the correlation matrix for the school features.

Based on these values, we decided to drop the PctMathBenchmark and the PctElaBenchmark because their correlation value was 0.92 which is above our threshold of 0.9.

For the census features, we also created a correlation matrix, but the matrix is too large to display on the page. The correlated features included the income features such as median family income, average family income, average household income, and median household income. Median household income and median family income had a correlation value of 0.98, and average household income and average family income had a correlation value of 0.99. These features are all correlated with per capita income as well, so we decided to remove median family income, average family income, average household income, median household income, average nonfamily income, and median nonfamily income, and keep per capita income instead. Employment and unemployment rate were also correlated with a value of -1, so we removed unemployment rate as a feature. We also removed median home value because its correlation value with per capita income was 0.91 which was above our 0.9 threshold.

After removing these highly correlated features, our dataset ended up having 34 features and 1564 high schools. 18 of the features are census features and 16 are school features.

## 4.2 Data Graph Output

We conducted a preliminary analysis of our datasets in order to find patterns in our data and see the distribution of the outcome variables of college going and graduation rates. Figures 2 and 3 show the distributions of college going rates and graduation rates respectively. For college going rates, there is significant variation in the rates. While there are more schools with high rates, there are a lot of schools with rates under 60 and they are very spread out. For graduation rates, the schools are much more clustered towards higher rates. There are fewer schools with lower graduation rates below 80, but there are around 25 schools with rates below 20.

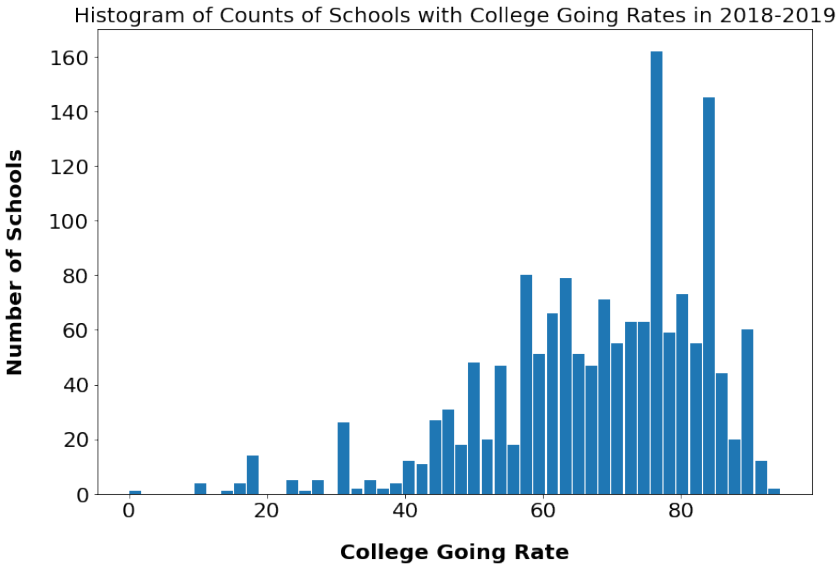


Figure 2: This figure shows the distribution of college going rates in 2018-2019 for all high schools in California.

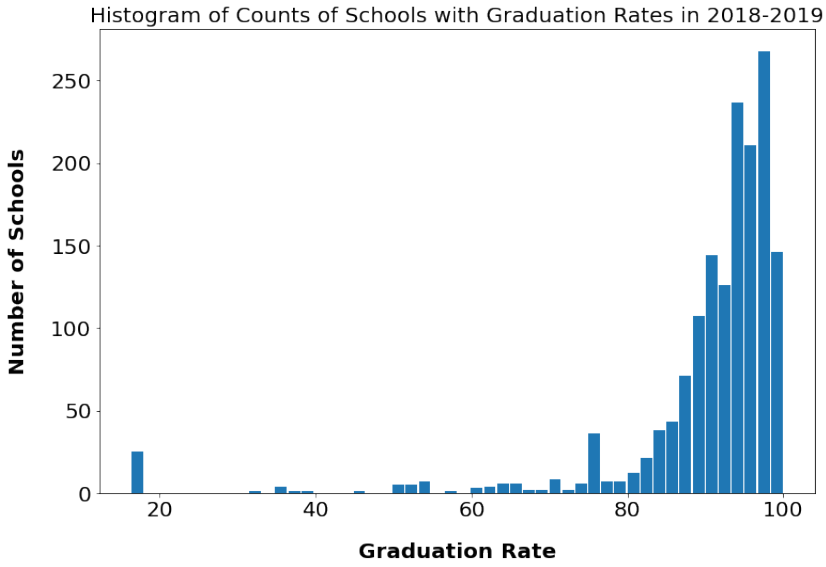


Figure 3: This figure shows the distribution of graduation rates in 2018-2019 for all high schools in California.

# 5 Methodology

For our methodology, we decided to run models on only the school dataset, the census dataset, and then the combined dataset. We did this to understand which set of data performs the best and how the predictions vary together and separately. For our models, we started by doing linear regression, random forest, and XGBoost on the school data and census data separately, and then we did it on the combined dataset. Linear regression was implemented using default parameters. while random forest and XGBoost were run with a hyperparameter grid search to find the most optimal parameters to use on our dataset. Based on this initial modeling, we found that random forest performs better than the other two models, so we used random forest for the rest of our calculations. We did forward feature selection and permutation importance on the separate and combined datasets for the random forest model. We also wanted to try another model, so we used KNN with 2 neighbors on the combined dataset as well. For each of the models, we ran the model 5 times and then calculated the average for r2, standard deviation, mean squared error, and mean absolute error.

# 6 Preliminary Modeling Results

We performed forward feature selection and permutation importance using random forest on census and school level data both individually and combined. We took the target from the column labeled, 'College Going Rate - Total (12 Months)'

for our analysis to predict the rate of postsecondary enrollment. We also analyzed the graduation rate using the column, 'Regular HS Diploma Graduates (Rate)'. We normalized all of the data using sklearn standard scalar. After cleaning our data and dropping rows that contained NaN values, we had data for 1544 schools.

## 6.1 Census Level Data Analysis

We limited our dataset to test the prediction accuracy with census features only. We ran this dataset on three models for regression analysis, which were Linear Regression, Random Forest, and XGBoost. Each model was run 5 times and the metrics reported here are the average of the runs.

**Census Features Only Predictions**

| Prediction Outcome Variable | Census Only Dataset | | | |
|---|---|---|---|---|
| | **Model** | **Mean R^2** | **Average MSE** | **Average MAE** |
| **High School Graduation rates** | Random Forest | 0.57 | 79.5 | 4.87 |
| | XGBoost | 0.56 | 81.48 | 5.01 |
| | Linear Regression | 0.12 | 163.63 | 7.28 |
| **College Going Rates** | Random Forest | 0.43 | 150.79 | 7.56 |
| | XGBoost | 0.41 | 156.06 | 7.72 |
| | Linear Regression | 0.24 | 201.98 | 10.07 |

Figure 4: This figure shows the metrics for both high school graduation and postsecondary enrollment predictions for three regression models for census data only.

## 6.2 School Level Data Analysis

We limited our dataset to test the prediction accuracy with school features only. We ran this dataset on three models for regression analysis, which were Linear Regression, Random Forest, and XGBoost. Each model was run 5 times and the metrics reported here are the average of the runs.

**School Features Only Predictions**

| Prediction Outcome Variable | School Only Dataset | | | |
|---|---|---|---|---|
| | **Model** | **Mean R^2** | **Average MSE** | **Average MAE** |
| **High School Graduation Rates** | Random Forest | 0.81 | 51.22 | 4.61 |
| | XGBoost | 0.80 | 50.44 | 4.10 |
| | Linear Regression | 0.78 | 40.5 | 3.8 |
| **College Going Rates** | Random Forest | 0.82 | 46.97 | 4.19 |
| | XGBoost | 0.80 | 53.02 | 4.41 |
| | Linear Regression | 0.59 | 108.58 | 8.13 |

Figure 5: This figure shows the metrics for both high school graduation and postsecondary enrollment predictions for three regression models for school data only.

## 6.3 Combined School and Census Level Data

We combine our census data with our school-level data to obtain 34 total features. After running each of our models on the combined dataset, we got that random forest outperformed all other models. The averaged metrics from 5 runs are included in the tables below.

| Combined Dataset - High School Graduation | | | | |
|---|---|---|---|---|
| **Model** | **Mean R^2** | **Standard Deviation** | **Average MSE** | **Average MAE** |
| Random Forest | 0.92 | 13.84 | 14.56 | 1.76 |
| XGBoost | 0.89 | 14.27 | 20.1 | 1.87 |
| Linear Regression | 0.80 | 12.82 | 36.80 | 3.81 |
| KNN (2 neighbors) | 0.86 | 13.56 | 24.65 | 2.17 |

Figure 6: This figure shows the metrics for high school graduation predictions for all models on combined dataset.

| Combined Dataset | | | | |
|---|---|---|---|---|
| **Model** | **Mean R^2** | **Standard Deviation** | **Average MSE** | **Average MAE** |
| Random Forest | 0.88 | 14.68 | 31.79 | 3.02 |
| XGBoost | 0.87 | 15.4 | 32.17 | 3.12 |
| Linear Regression | 0.72 | 14.4 | 75.37 | 6.5 |
| KNN (2 neighbors) | 0.71 | 14.94 | 64.71 | 3.68 |

Figure 7: This figure shows the metrics for postsecondary enrollment predictions for all models on combined dataset.

In the figures above, it is evident that random forest is the best performing model compared to other regression models. XGBoost also performed very highly and had accuracy just slightly under Random Forest.

### 6.3.1 Permutation Importance

We took random forest as the highest performing model and ran permutation importance using the scikit-learn implementation. This was done first for graduation rates and then again for college going rates.

| Random Forest - High School Graduation Rates | |
|---|---|
| **Feature Name** | **Permutation Importance** |
| ChronicAbsenteeismRate | 0.03261 |
| _percent_adult_high_school_grads | 0.00998 |
| Seal of Biliteracy (Rate) | 0.00931 |
| GED Completer (Rate) | 0.00861 |
| _percent_black | 0.00785 |
| _employment_rate | 0.00758 |
| Other Transfer (Rate) | 0.00712 |
| _gini_inequality_index | 0.00651 |
| _percent_white | 0.00575 |
| _percent_adults_in_poverty | 0.00543 |

Figure 8: Top ten features and their permutation scores for graduation rates predictions

| Random Forest - College Going Rates | |
|---|---|
| **Feature Name** | **Permutation Importance** |
| ChronicAbsenteeismRate | 0.09081 |
| _median_gross_rent_divided_by_income | 0.01072 |
| _percent_black | 0.01061 |
| Adult Ed. HS Diploma (Rate) | 0.01032 |
| _employment_rate | 0.00958 |
| PctGE21 | 0.00942 |
| Other Transfer (Rate) | 0.00873 |
| CohortStudents | 0.00849 |
| GED Completer (Rate) | 0.00834 |
| _percent_adults_in_poverty | 0.00809 |

Figure 9: Top ten features and their permutation scores for college going rates predictions

For both graduation rates and postsecondary enrollment, the top performing feature was ChronicAbsenteeismRate with a significant permutation importance lead from the other features. However, after this feature there isn't a huge correlation between the top ten features listed for grad rates and postsecondary rates.

### 6.3.2 Forward Feature Selection

We additionally ran forward feature selection using Scikit-Learn's implementation with random forest to compare the top features impacting predictions. The features that were selected by the FFS algorithm are listed below.
**High school graduation rate result:**
'ChronicAbsenteeismRate', '_percent_asian', '_percent_native', 'Other Transfer (Rate)', '_population', '_per_capita_income', '_gini_inequality_index', 'CHSPE Completer (Rate)', 'GED Completer (Rate)', '_percent_children_in_poverty'
**College going rate result:**
'ChronicAbsenteeismRate', '_percent_adult_high_school_grads', '_population', 'Other Transfer (Rate)', '_population_density', 'CohortStudents', '_percent_adult_college_grads', '_gini_inequality_index', '_employment_rate', '_percent_white'

## 6.4 LIME Interpretability for Random Forest Model

In order to gain a better understanding on how the features are impacting the random forest model predictions for high accuracy, we ran the model through the Local Interpretable Model-Agnostic Explanations, also known as LIME Interpretability. Our study employed submodular pick with LIME, which allowed us to visualize the top important features and their correlation to the predictions. Submodular Pick for Explaining Models will pick the top schools that maximize coverage of dataset with respect to the most variating prediction instances [8].

### 6.4.1 High School Graduation Predictions

We used high school graduation rate as the outcome prediction. We took the top 8 most covered points of the dataset as determined by the explainer function. From these top 8 points, the explainer decided the top ten most impactful features on the prediction and their positive or negative correlation. The explainer describes how the random forest predictions were made for these schools.

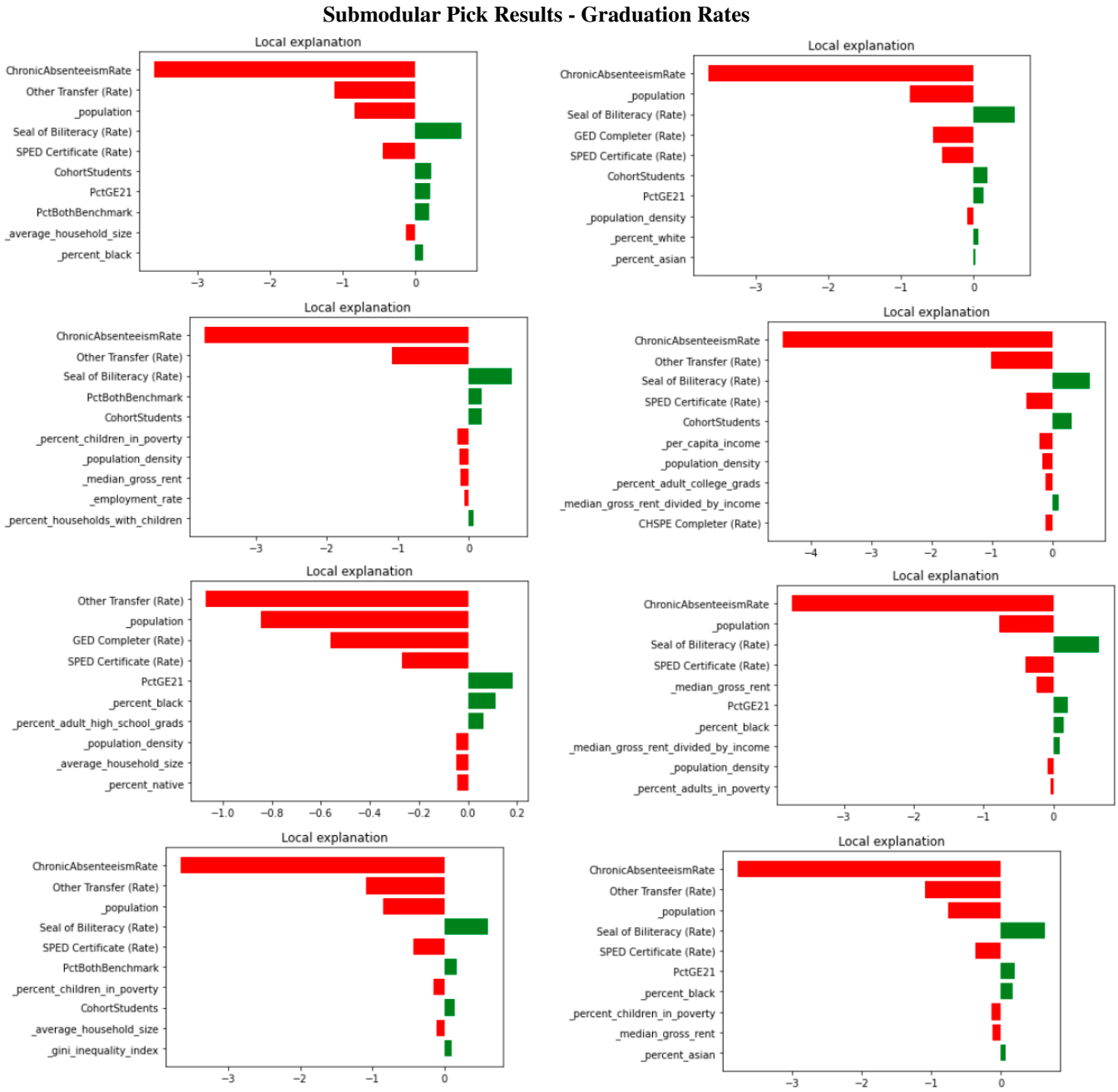


Figure 10: This figure shows the top 8 most representative points and their top features

The model determined that for high school graduation, the top ten most significant features were as follows: 'ChronicAbsenteeismRate', 'CohortStudents', 'GED Completer (Rate)', 'Other Transfer (Rate)', 'PctBothBenchmark', 'PctGE21',

'SPED Certificate (Rate)', 'Seal of Biliteracy (Rate)', '_average_household_size', '_employment_rate', '_gini_inequality_index'

These features can be compared to that of forward feature selection using random forest. Both LIME and forward feature selection chose 'ChronicAbsenteeismRate' as their first selected feature. 'Other Transfer (Rate)' was in the top five for both feature selections. Otherwise, while some of the same features appear in the top ten, they have a highly variating ordering and do not correlate beyond this analysis.

In addition to analyzing the overall features selected by LIME and standard libraries, we also contrasted schools that had high and low prediction outcome rates for high school graduations. We did this by selecting a school at index 290 with a high graduation graduation rate of 97.3 percent, since this was a common rate among a few schools in the dataset. We also selected a school at index 410 with a low graduation rate of 35.3 percent. This rate was also a common outcome value for high school graduation rates, and while it is not the lowest rate in the dataset, it provides a value difference of over 60 percent, which is enough contrast between different school predictions.

We ran these two selected schools using the LIME explainer for a single instance. This explainer shows the top features for each prediction and whether the feature is positively or negatively correlated.

**High Graduation Rate vs. Low Graduation Rate**

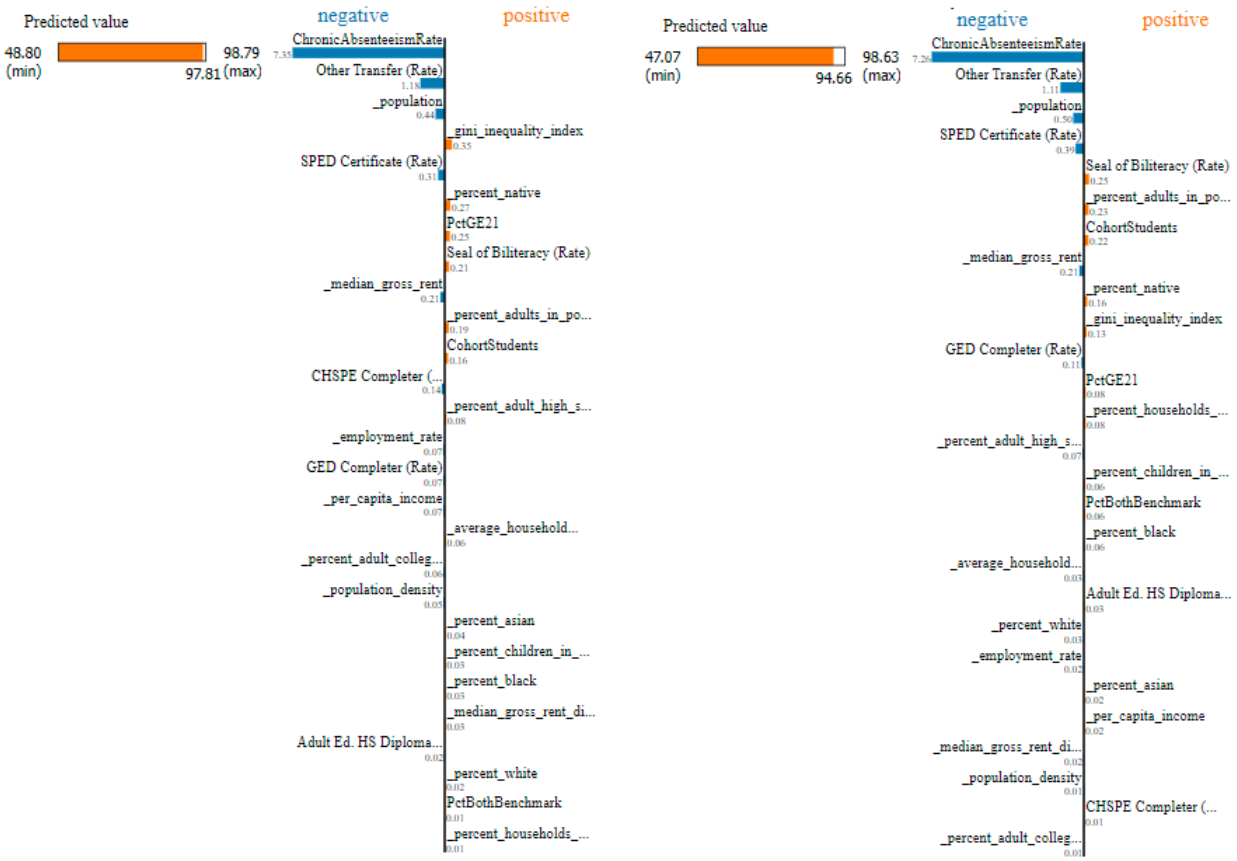


Figure 11: Left figure represents the high graduation rate of true value 97.3 percent, and the right represents the low graduation rate of true value 35.3 percent

The orange bar to the left of each image represents the prediction value for each school. The left school had a prediction of 97.81 and the right school was predicted as 94.66. From these predictions, it is clear that LIME has an easier time predicting higher graduation rates verses lower ones. This is partly affected by our heavily skewed data for graduation rates. From the figures we can compare the top feature impact for a high graduation rate verses a low graduation rate. The blue features are negatively correlated to outcome and the orange are positively correlated. It appears that the top three features ('ChronicAbsenteeismRate', 'Other Transfer (Rate)', '_population') all apprear in the same importance order for both instances. From here, the features deviate slightly, as well as the magnitude of impact value that appears with each feature.

### 6.4.2 College Going Predictions

We used the college going rate as the outcome prediction. We took the top 8 most covered points of the dataset as determined by the explainer function. From these top 8 points, the explainer decided the top ten most impactful features on the prediction and their positive or negative correlation.

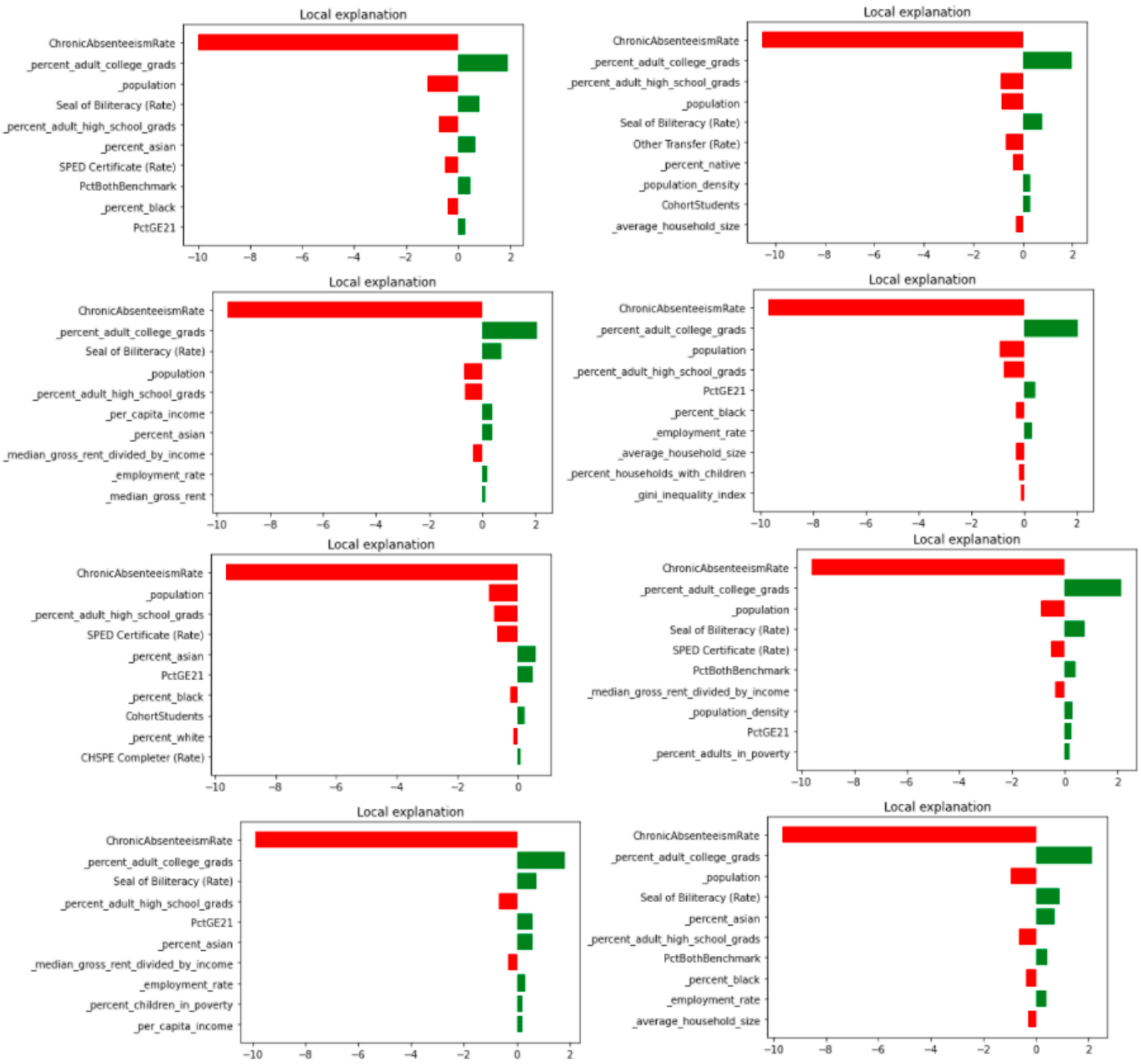


Figure 12: This figure shows the top 8 most representative points and their top features

The model also determined that for postsecondary enrollment, the top ten most significant features were as follows: 'ChronicAbsenteeismRate', 'CohortStudents', 'Other Transfer (Rate)', 'PctBothBenchmark', 'PctGE21', 'SPED Certificate (Rate)', 'Seal of Biliteracy (Rate)', '_average_household_size', '_employment_rate', '_gini_inequality_index', '_median_gross_rent'.

These features can be compared to that of forward feature selection using random forest. Both LIME and forward feature selection chose 'ChronicAbsenteeismRate' as their first selected feature. 'CohortStudents', 'Other Transfer (Rate)', '_employment_rate', and '_gini_inequality_index', were mutually selected by both processes.

In addition to analyzing the overall features selected by LIME and standard libraries, we also contrasted schools that had high and low postsecondary enrollment rates, similar to comparing the graduation rates instances above. We selected the same two schools at indices 290 and 410 in the dataset. They had college going rates of 83.5 and 27.3, respectively. These two percentages have a gap of 56.2 percent between them, which we decided was a large enough margin to compare with the LIME instances.

We ran these two selected schools using the LIME explainer for a single instance. This explainer shows the top features for each prediction and whether the feature is positively or negatively correlated.

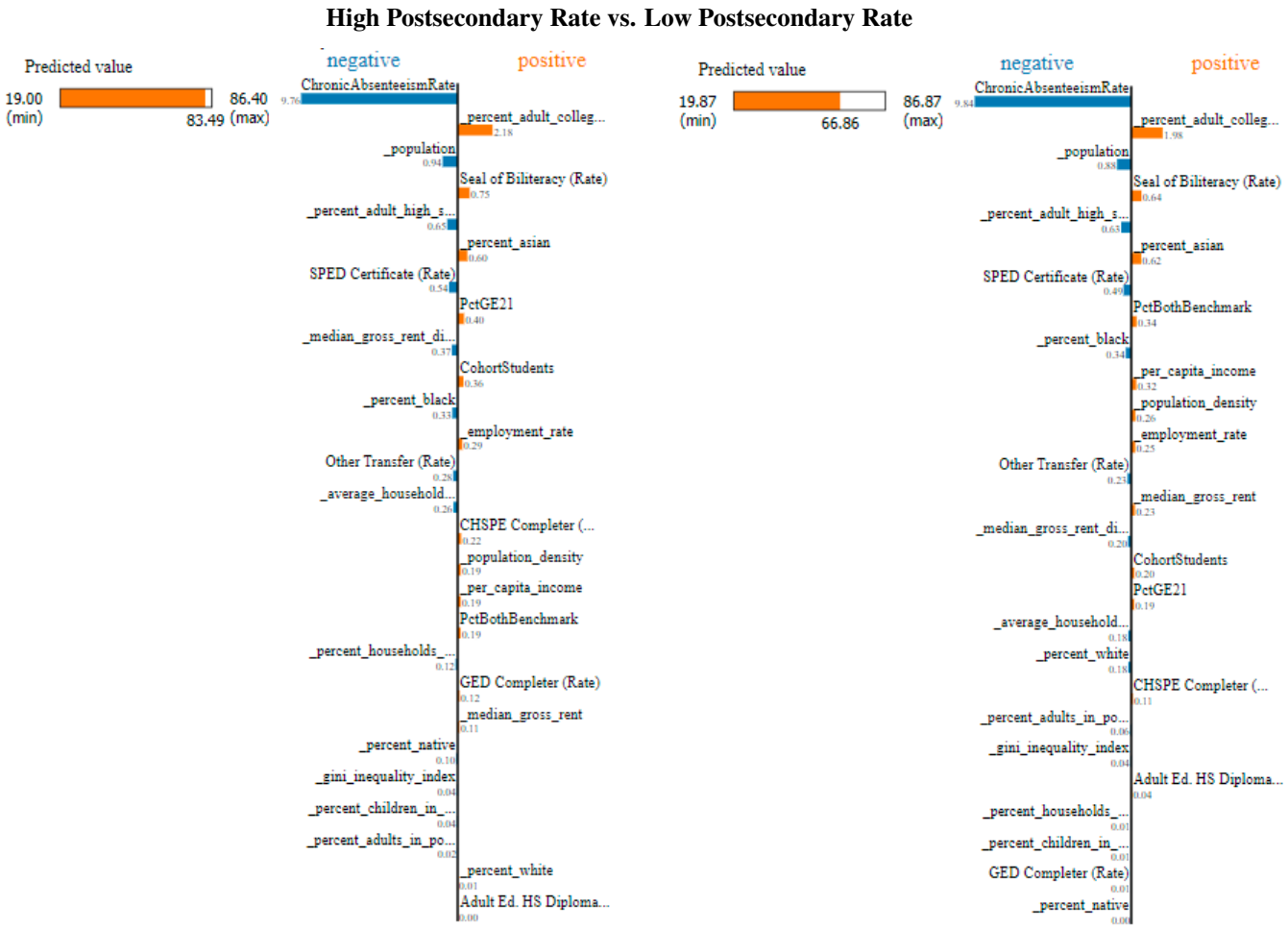


Figure 13: Left figure represents the high graduation rate of true value 83.5 percent, and the right represents the low graduation rate of true value 27.3 percent

The orange bar to the left of each image represents the prediction value for each school. The left school had a prediction of 83.49 and the right school was predicted as 66.86. From the figures we can compare the top feature impact for a high graduation rate verses a low graduation rate. The blue features are negatively correlated to outcome and the orange are positively correlated. It appears that 'ChronicAbsenteeismRate', '_percent_adult_college_grads', '_population', 'Seal of Biliteracy (Rate)', '_percent_adult_high_school_grads', '_percent_asian', and 'SPED Certificate (Rate)' were the top 8 features that appeared in the same importance order for both instances. From here, the features deviate slightly, as well as the magnitude of impact value that appears with each feature.

## 6.5 Bias Analysis Based on District Size

One possible critique of our analysis is that we obtained our data on the district level as opposed to the school level. Thus, it is reasonable to suspect that our predictions may not be as accurate for schools that are part of large districts, since these features are averaged over a larger area. To address this, we computed the average MAE for schools in each district to determine the relationship between the number of schools in a district and this error measure. Below are graphs of district MAE vs number of schools in that district for our 2 success metrics. Note that there are 2 outliers that we removed from the graphs for readability. The first is Los Angeles Unified School District, which has 166 high schools and has a MAE of 2.05 for graduation rate and 3.61 for college going rate. The second is New Haven Unified School District, which has 1 high school and has a MAE of 37.82 for graduation rate and 55.30 for college going rate.

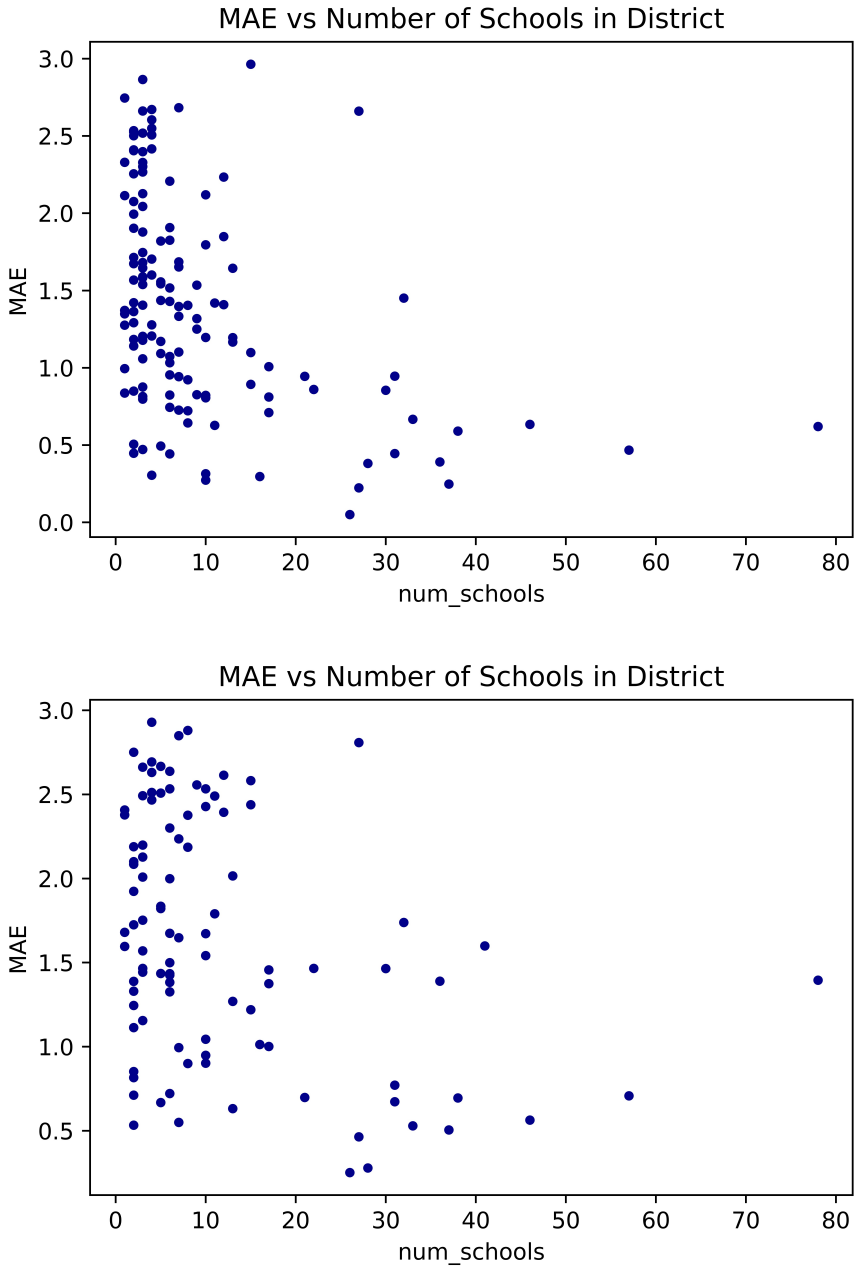


Figure 14: MAE vs Number of Schools for Graduation Rates (Top) and College Going Rates (Bottom)

The correlation between MAE and number of schools was -.110 for graduation rates and -.139 for college going rates. Since these values are close to 0, we do not believe that the low resolution of our census data is a major source of error for our model.

To further show that there is not a large discrepancy between the accuracy of our model between schools from small and large districts, below are plots of predicted vs ground truth outcomes for Los Angeles Unified School District (Blue), which has 166 high schools, and Hanford Joint Union High School District (Green), which has 3 high schools. One can see that there is not a noticeable difference in predictive power for these 2 districts.

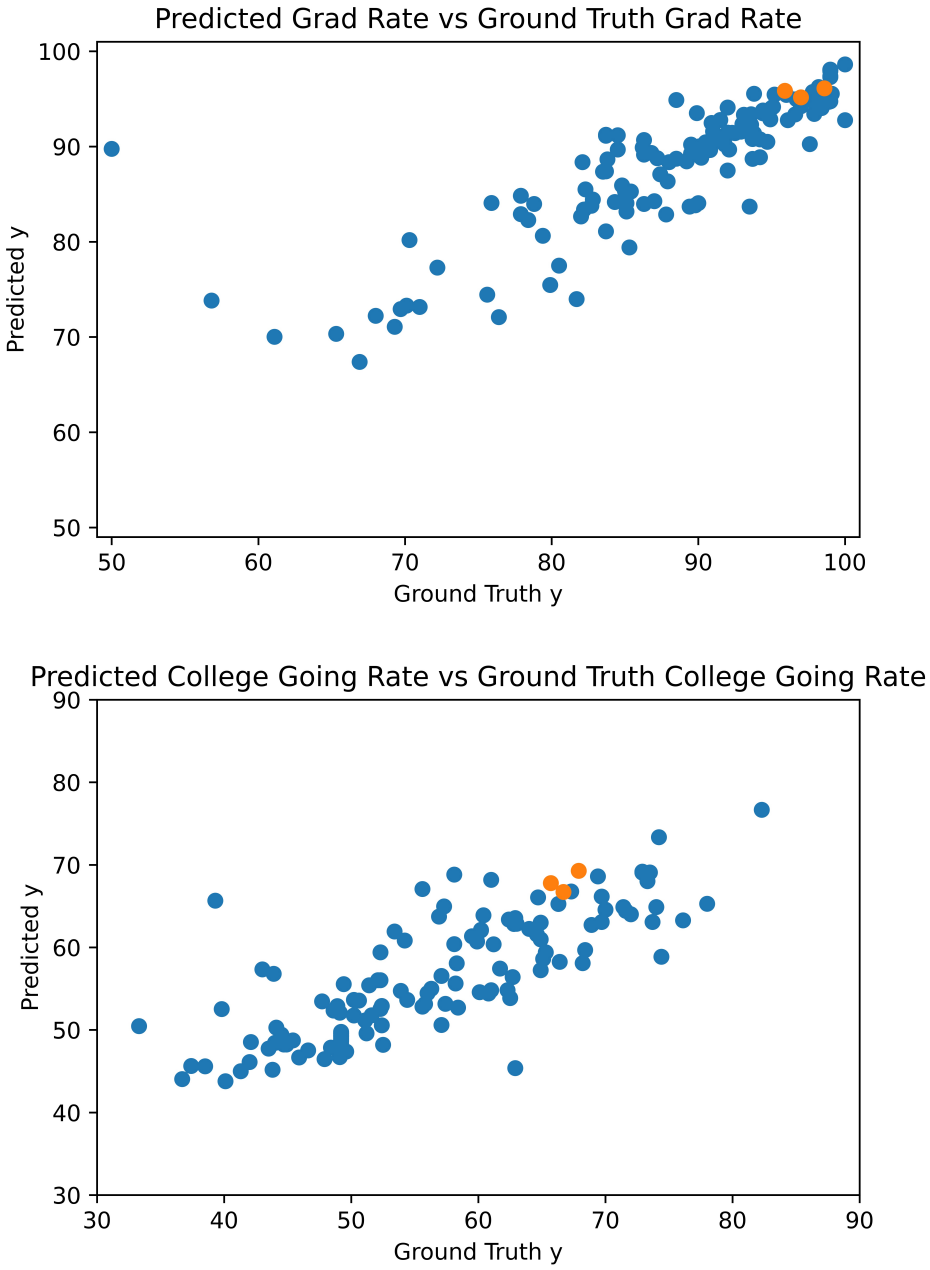


Figure 15: Predicted vs Ground Truth Predictions for Graduation Rates (Top) and College Going Rates (Bottom) for Los Angeles Unified School District (Blue) and Hanford Joint Union High School District (Green)

The MAE for Los Angeles Unified School District was 2.05 for graduation rate and 3.61 for college going rate. The MAE for Hanford Joint Union High School District was 1.93 for graduation rate and 4.29 for college going rate.

# 7 Optimization Steps - Sensitivity Analysis

For optimization, we wanted to do a sensitivity analysis to see how altering certain features would impact graduation and college going rates. To do this, we started by identifying the top features that were actionable in the model. For graduation rates, the top features that schools could make changes towards are 'ChronicAbsenteeismRate', 'PctBothBenchmark', and 'Seal of Biliteracy (Rate)'. For college going rates, the top features that schools could make changes towards are 'ChronicAbsenteeismRate', 'PctGE21', and 'Cohort Students'. For each feature, we either increased or decreased it by 5, 10, or 15%. For chronic absenteeism, we decreased it by 5, 10, and 15%, and for all other features, we increased them by 5, 10, and 15%.

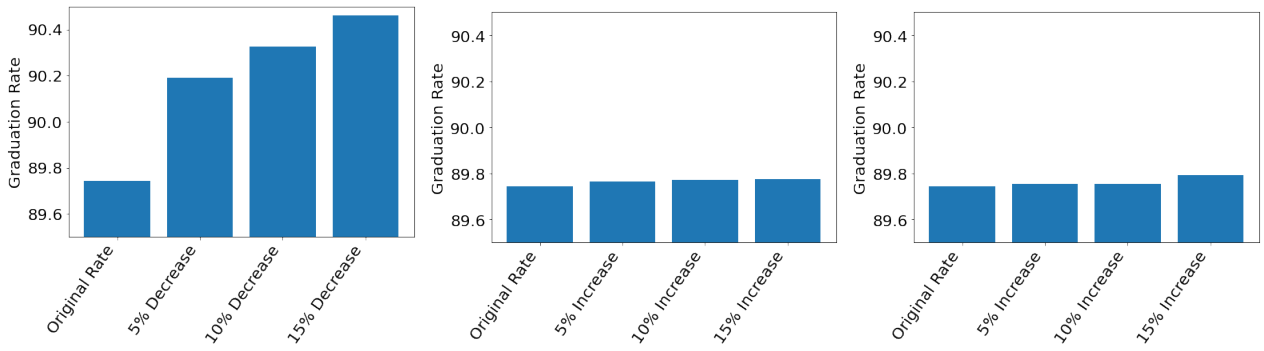


Figure 16: Sensitivity Analysis for Graduation Rates

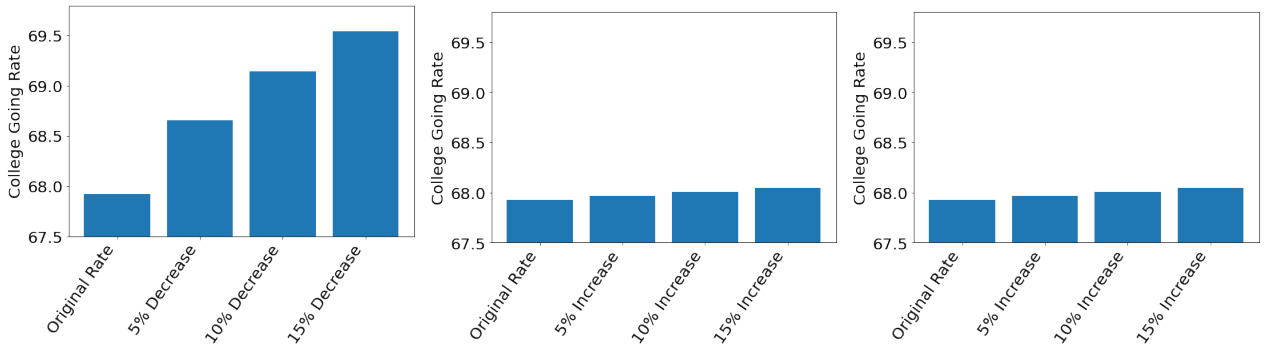


Figure 17: Sensitivity Analysis for College Going Rates

As the figure above shows, graduation rate increased by almost a percent with a 15% decrease to absenteeism rate. For college going rates, decreasing absenteeism by 15% had a larger impact as college going rate went up by 1.5%. For the other features, increasing them did increase the graduation and college going rates, but the impact was not as much as chronic absenteeism. Based on this analysis, chronic absenteeism seems to be the most impactful feature for schools to act on. The test score benchmarks are also important as well as having more students achieve the seal of biliteracy. Cohort students was also an interesting feature to change because increasing the number of students increased the college going rate by a small amount, so larger schools may have higher success rates with their students going to college.

# 8 Conclusion and Extensions

## 8.1 Extensions

### 8.1.1 Increase Dataset Size

While our model performs well on a single year of California high school data, there was still about 20% of our original dataset that had rows with NaN values. It would be interesting to incorporate multiple years of data, perhaps more with lower graduation rates and postsecondary enrollment in order to diversify our school set.

### 8.1.2 School-Level Census Data

Although our results from section 6.5 suggest that using district-level census data was likely not a major source of error in our data, it would still be interesting to run our analysis using school-level census data should it become available.

### 8.1.3 Deeper Analysis with Forward Feature Selection

Our study did not have the capacity to fully delve into forward feaure selection with respect to training the model on top selected features and analyzing these results. It would be interesting to see the degree of improvement for the model when using the FFS techniques. It would also be interesting to run the model on LIME's selected features and compare the results to that of a standard forward feature selection algorithm.

## 8.2 Conclusion

Our work has identified the top features that impact graduation and postsecondary rates at the school and geographic level. When looking at both school and geographic factors, our analysis found that chronic absenteeism was the most impactful feature. However, geographic features have a higher representation in the overall top 10 features. Many economic and geographical features, such as income of an area, population, and percent of racial diversity, played a large role in both graduation rates and postsecondary enrollment. It is also important to note that graduation rates and pursuit of postsecondary education are not equivalent to each other, meaning that certain features that may improve grad rates will not necessarily improve college enrollment. Particularly, our optimization section showed that increasing seal of biliteracy rate in schools can raise graduation rates at a higher impact, while focusing on increasing test scores will improve postsecondary enrollment. However, both aspects are significantly impacted by the school absenteeism rate, which should show California schools that this feature should be highly prioritized.

Since census features are not factors a school can change, we based our sensitivity analysis on features that schools can act on to make a difference to their graduation and post secondary enrollment rates. These actionable features include chronic absenteeism rate, the benchmark test scores, the seal of biliteracy rate, and the number of students in a school. For these features, it was surprising that the test scores were not the most important actionable features since schools tend to focus money on improving test scores. Our analysis can benefit these schools by showing them that there are other factors that can have a greater influence than traditional academic standards. Schools will be able to look at this information and understand where they can allocate their resources to improve these features and achieve the highest outcomes for students.

Ultimately, this analysis informs the complexity of economic and geographic features and shows the massive impact of these features on student success in high school. We hope that further analysis will be conducted to investigate more nontraditional factors so schools can understand these factors and find ways to improve them. Overall, this will lead to greater education quality and equity in America.

# References


[1] "Out-of-School Children and Youth," Nov. 2016.

[2] "Data & Statistics (CA Dept of Education)."

[3] A. Lyttle-Burns, "Factors That Contribute to Student Graduation and Dropout Rates: An In-Depth Study of a Rural Appalachian School District," *Online Theses and Dissertations*, Jan. 2011.

[4] M. N. Yakubu and A. M. Abubakar, "Applying machine learning approach to predict students' performance in higher educational institutions," *Kybernetes*, vol. ahead-of-print, Jan. 2021.

[5] E. Aguiar, H. Lakkaraju, N. Bhanpuri, D. Miller, B. Yuhas, and K. L. Addison, "Who, when, and why: a machine learning approach to prioritizing students at risk of not graduating high school on time," in *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, LAK '15, (Poughkeepsie, New York), pp. 93–102, Association for Computing Machinery, Mar. 2015.

[6] S. Loeb and M. Page, "EXAMINING THE LINK BETWEEN TEACHER WAGES AND STUDENT outcomes: THE IMPORTANCE OF ALTERNATIVE LABOR MARKET opportunities AND NON-PECUNIARY VARIATION," *The Review of Economics and Statistics*, vol. 82, pp. 393–408, Aug. 2000.

[7] H. Lakkaraju, E. Aguiar, C. Shan, D. Miller, N. Bhanpuri, R. Ghani, and K. L. Addison, "A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, (Sydney, NSW, Australia), pp. 1909–1918, Association for Computing Machinery, Aug. 2015.

[8] "Interpretability part 3: opening the black box with LIME and SHAP," Dec. 2019.


# 9 Appendix

## 9.1 Features Definitions for Dataset

Below is the list of all features used for the models and their respective definitions.

| School Features | |
|---|---|
| **Feature Name** | **definition** |
| Seal of Biliteracy (Rate) | rate of students who receive seal of biliteracy |
| CHSPE Completer (Rate) | rate of students who complete the chspe |
| Adult Ed. HS Diploma (Rate) | percent of students who get their high school diploma as adults |
| SPED Certificate (Rate) | special education certificate rate |
| GED Completer (Rate) | percent students who complete their ged |
| Other Transfer (Rate) | rate of student transfers across other schools within the state |
| Dropout (Rate) | percent of students who dropout of school |
| Still Enrolled (Rate) | percent students still enrolled in high school |
| PctGE21 | percent test takers whose act composite scores are greater or equal to 21 |
| PctElaBenchmark | The percent of students who met or exceeded the benchmark for ELA test |
| PctERWBenchmark12 | percent of students who met or exceeded the english and reading benchmark in 12th grade |
| PctMathBenchmark12 | percent of students who met or exceeded the math benchmark in 12th grade |
| PctERWBenchmark11 | percent of students who met or exceeded english and reading benchmarks in 11th grade |
| PctMathBenchmark11 | percent of students who met or exceeded math benchmarks in 11th grade |
| PctBothBenchmark12 | percent of students who met or exceeded math and english benchmarks in 12th grade |
| PctBothBenchmark11 | percent of students who met or exceeded math and english benchmarks in 11th grade |
| ChronicAbsenteeismRate | cumulative rate of absenteeism across the school for the year. |

Figure 18: This figure describes the school features and their definitions as defined by the CDE dataset.

| Feature Name | definition |
|---|---|
| _population | Total population within district attendance zone |
| _population_density | Average population density |
| _percent_households_with_children | Percent of households containing children |
| _average_household_size | Average number of people living in a household |
| _percent_adult_high_school_grads | Percent of adults who graduated high school |
| _percent_adult_college_grads | Percent of adults who graduated college |
| _employment_rate | Percentage of eligible people who are employed |
| _unemployment_rate | Percentage of eligible people who are unemployed |
| _median_household_income | Median income of a household |
| _average_household_income | Average income of a household |
| _median_family_income | Median income of a family |
| _average_family_income | Average income of a family |
| _median_nonfamily_income | Median income of a single individual |
| _average_nonfamily_income | Average income of a single individual |
| _per_capita_income | Average income per person |
| _gini_inequality_index | Metric of economic inequality |
| _median_home_value | Median value of a home |
| _median_gross_rent | Median price of rent |
| _percent_white | Percent of people that are white |
| _percent_black | Percent of people that are black |
| _percent_native | Percent of people that are native |
| _percent_asian | Percent of people that are asian |
| _percent_adults_in_poverty | Percent of adults living in poverty |
| _percent_children_in_poverty | Percent of children living in poverty |
| _median_gross_rent_divided_by_income | Median gross rent divided by income for each person |

Figure 19: This figure describes the census features and their definitions.