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

Team Members: Lydia DiBlasio, Gauri Madhok, Isaac Wahout

Agenda:

- Problem Statement
- Background
- Data Preparation
- Models
- Results
- Optimization
- Next Steps

Background and Problem Statement

Problem: Gap in Graduation Rates / Postsecondary Enrollment

- In 2018, 258 million children worldwide dropped out of school
- 64.4% enrolled in post-secondary education in California
- 84.3% of high school seniors in California graduated in 2020
- High school graduates earn a national average of $8,000 more annually compared to high school dropouts
- On an annual basis, bachelor's degree holders earn about $32,000 more than those whose highest degree is a high school diploma
- Still a very limited understanding about which factors are the main influence for low high school graduation and pursuit of higher education

Past Work:

- Prior research in the area has centered around predicting academic performance of students based on student specific features:
  - 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.
- Other research has looked at features outside of a student's control such as the wages teachers are paid
  - 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%

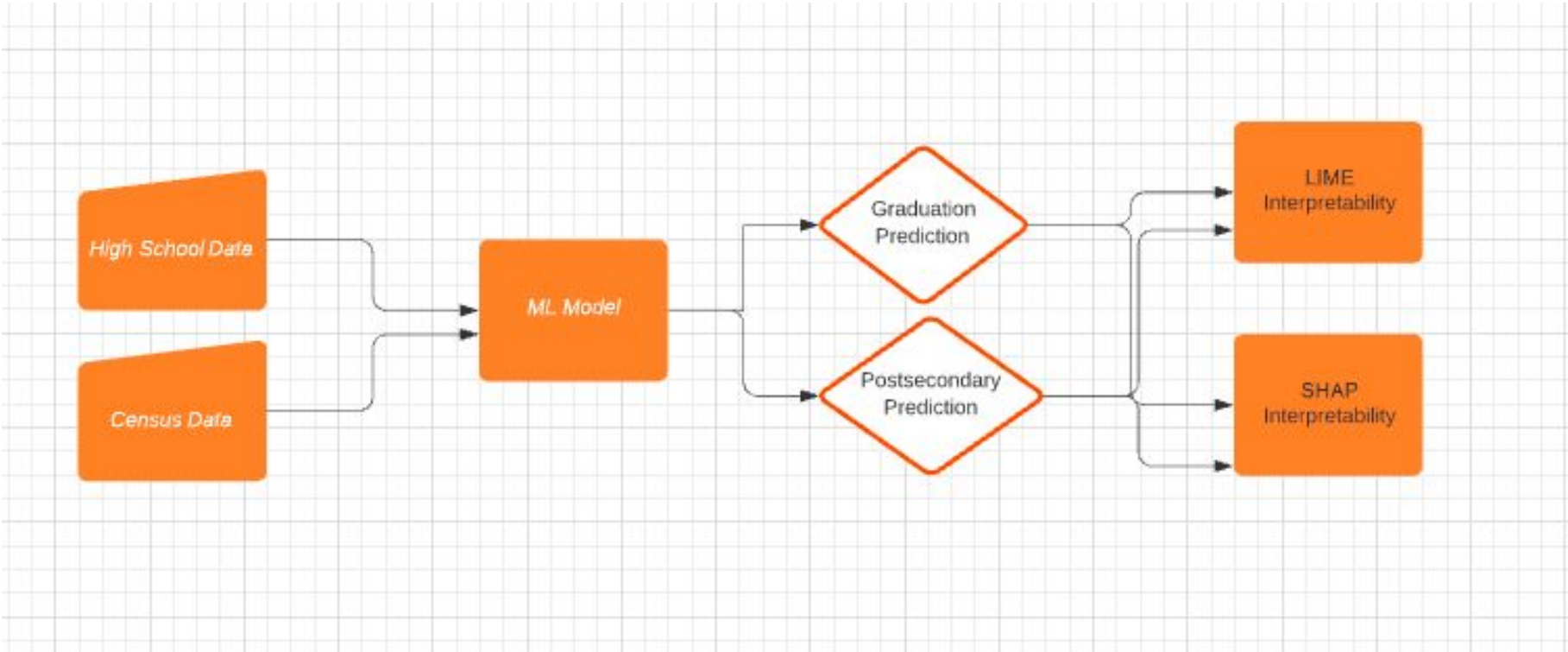
Project Goal

Analyze top features influencing graduation rates and postsecondary enrollment to inform California high schools on how to maximize these rates

Importance:

1. Increase graduation rates in high schools
2. Increase likeliness of going to college after high school
3. Discover correlation between socioeconomic and geographic factors for high schools

Project Diagram



Data Preparation and Processing

Datasets

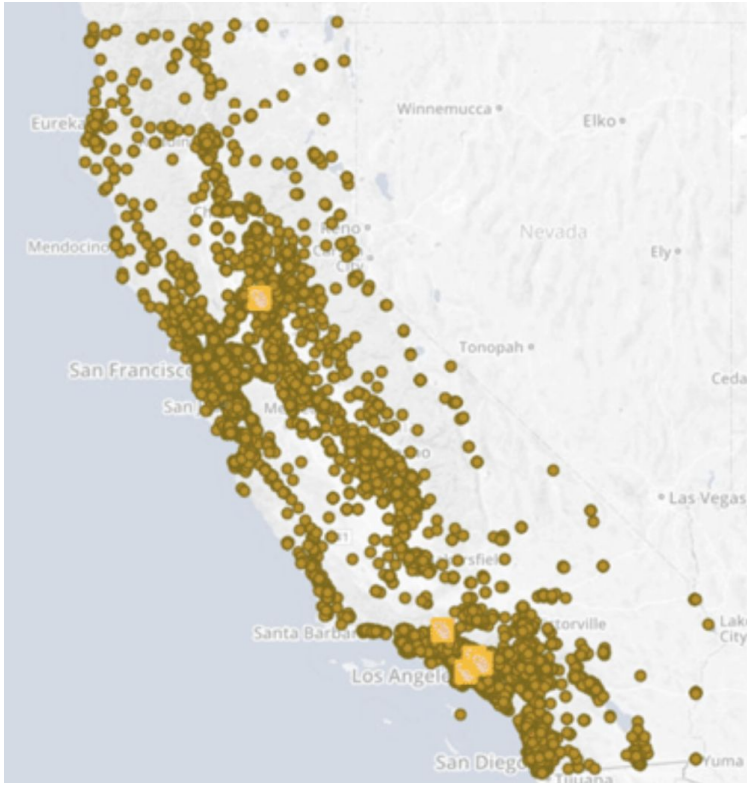
1. California Department of Education
    - ❖ Information at a school level in California, which informs student education
    - ❖ From the school year of 2018-2019

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.

Datasets

1. California Department of Education

   - Data was collected from The California Department of Education for 1940 high schools in California



Datasets

2. U.S. American Community Survey Database

   - ❖ Information on population demographics for 157 school districts in California
   - ❖ Obtained American Community Survey data via Social Explorer for the average of all census tracts in a school district

AMERICAN COMMUNITY SURVEY
U.S. CENSUS BUREAU

Social Explorer®

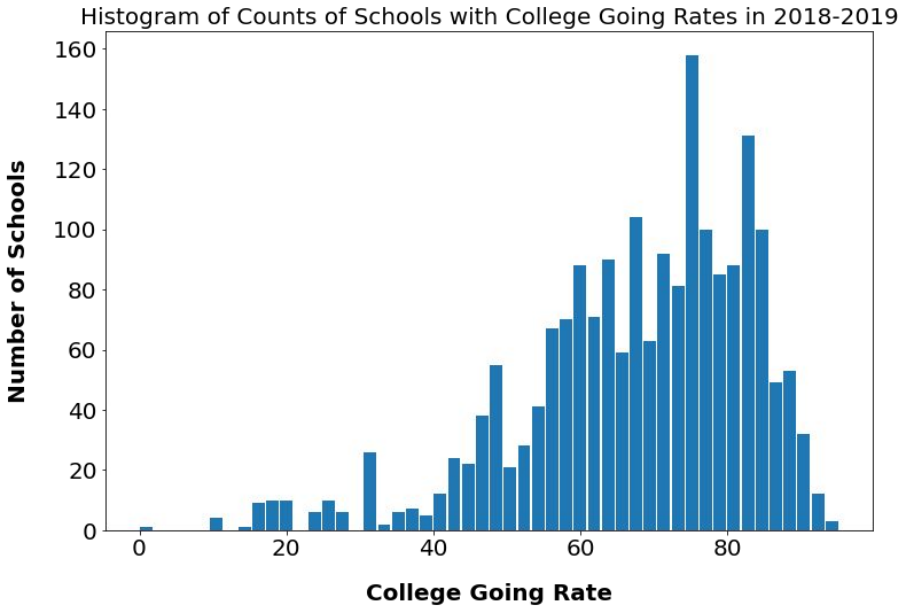
Datasets

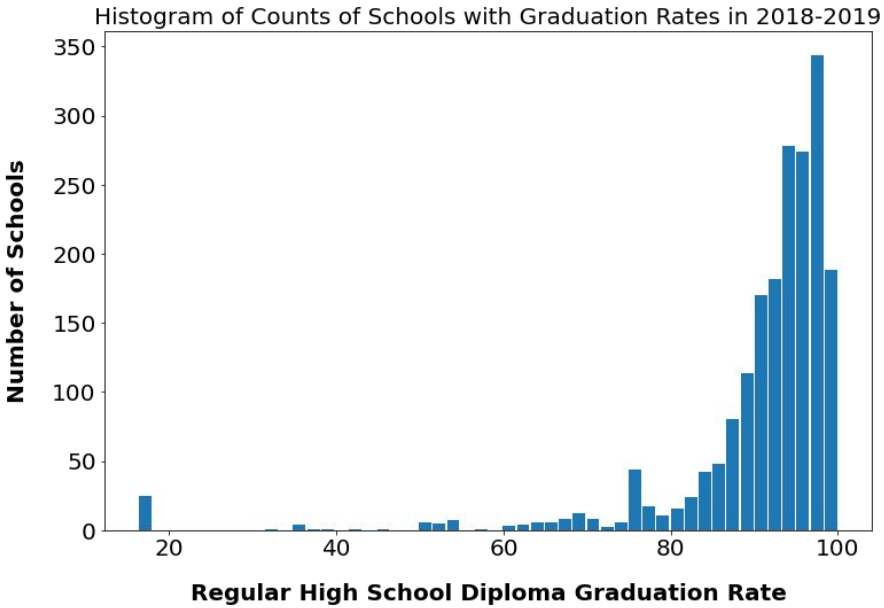
2. U.S. American Community Survey Database

- ❖ Selected 25 features to use in our analysis
- ❖ Types of metrics reported include income, race, employment status, population, education, average home value/rent, and presence of children

_population
_population_density
_percent_white
_percent_black
_percent_native
_percent_asian
_percent_households_with_children
_average_household_size
_percent_adult_high_school_grads
_percent_adult_college_grads
_employment_rate
_unemployment_rate
_median_household_income
_average_household_income
_median_family_income
_average_family_income
_median_nonfamily_income
_average_nonfamily_income
_per_capita_income
_gini_inequality_index
_median_home_value
_median_gross_rent
_median_gross_rent_divided_by_income
_percent_children_in_poverty
_percent_adults_in_poverty

Distribution of Labels





Data Preprocessing

1. Converted counts to rates by dividing by total enrollment
2. Merged school features from various sources based on school code
3. Merged school data with school district-level census features
4. Dropped string features from combined dataset
5. Dropped 394 schools that had NaN values after merging

***Original dataset was 1940 high schools and 75 features**
***Ended up with 1544 high schools and 48 features**

Models

Inputs and Training For All Models

- One year of data used overall (2018-2019)
- Our models were trained on 1544 schools with 48 features per school
- Outcome variable was 'Regular HS Diploma Graduates (Rate)' for graduation rates
- Outcome variable was ‘College Going Rate - Total (12 Months)’ for post-secondary enrollment rate
- Split into train and test with sklearn train_test_split method with a test size of 20% (single split)

Results

Linear Regression Feature Selection on Census Only Data

Graduation Rates:

- 1,544 schools, 25 total features
- Linear Regression r2 score: 0.21
- Top 5 features selected:

Feature	Permutation Importance
_median_household_income	0.043
_median_family_income	0.037
_average_household_size	0.024
_per_capita_income	0.013
_percent_adult_college_grads	0.011

Postsecondary Rates:

- 1,544 schools, 25 features
- Linear Regression r2 score: 0.28
- Top 5 features selected:

Feature	Permutation Importance
_percent_households_with_children	0.014
_population	0.013
__median_family_income	0.013
_average_household_size	0.009
_percent_adult_high_school_grads	0.006

Linear Regression Feature Selection on School Only Data

Graduation Rates:

- 1,544 schools, 23 total features
- Linear Regression r2 score: 0.69
- Top 5 features selected:

Feature	Permutation Importance
ChronicAbsenteeismRate	0.51
PctGE21	0.0037
SPED Certificate (Rate)	0.0025
CHSPE Completer (Rate)	0.0022
PctMathBenchmark	0.0009

College Going Rates:

- 1,544 schools, 23 total features
- Linear Regression r2 score: 0.64
- Top 5 features selected:

Feature	Permutation Importance
ChronicAbsenteeismRate	0.12
Seal of Biliteracy (Rate)	0.019
SPED Certificate (Rate)	0.007
GED Completer (Rate)	0.006
Adult Ed. HS Diploma (Rate)	0.006

Linear Regression Feature Selection on Combined Dataset:

Graduation Rates:

- 1,544 schools, 48 total features
- Linear Regression r2 score: 0.86
- Top 5 features selected:

Feature	Permutation Importance
_population	0.249
_per_capita_income	0.151
_average_family_income	0.146
_population_density	0.136
_median_home_value	0.130

College Going Rates:

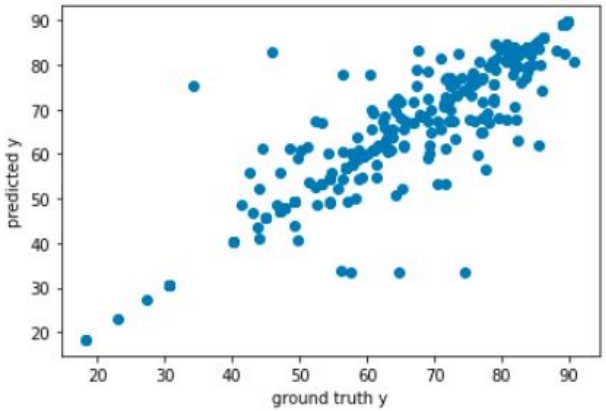
- 1,544 schools, 48 total features
- Linear Regression r2 score: 0.77
- Top 5 features selected:

Feature	Permutation Importance
_population	0.242
_median_home_value	0.182
_average_household_income	0.149
_average_family income	0.148
_average_nonfamily income	0.145

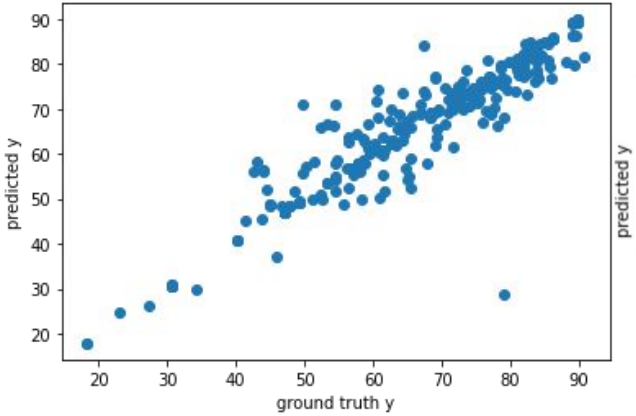
Combined Data Model Performance-- College Going Rate

Model	r2 Score
LinearRegression (default parameters)	.77
Random Forest (hyperparameter grid search)	**.89**
XGBoost (hyperparameter grid search)	.88
SVR (rbf kernel)	.72
KNN (2 neighbors)	.72
MLP (default parameters)	.69

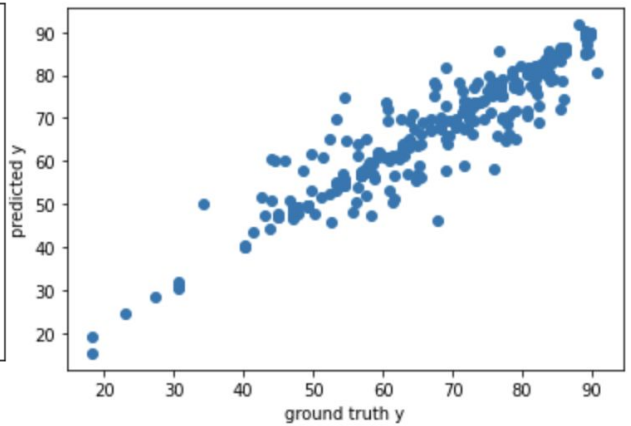
College Going Rates Results - Visualization



Linear Regression



Random Forest

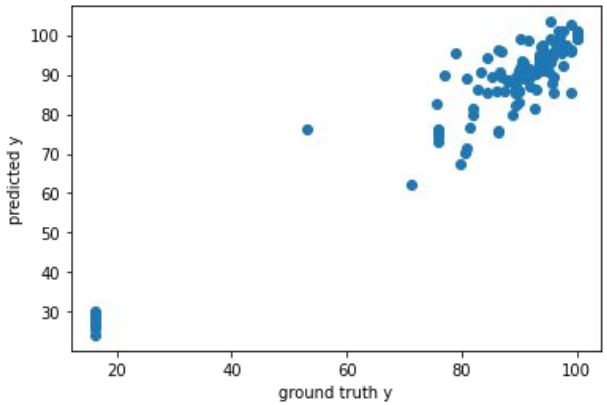


XGBoost

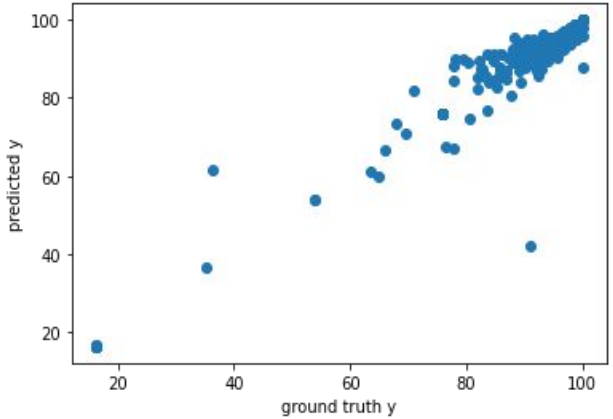
Combined Data Model Performance-- Graduation Rate

Model	r2 Score
LinearRegression (default parameters)	.86
Random Forest (hyperparameter grid search)	**.94**
XGBoost (hyperparameter grid search)	.93
SVR (rbf kernel)	.91
KNN (2 neighbors)	.90
MLP (default parameters)	.77

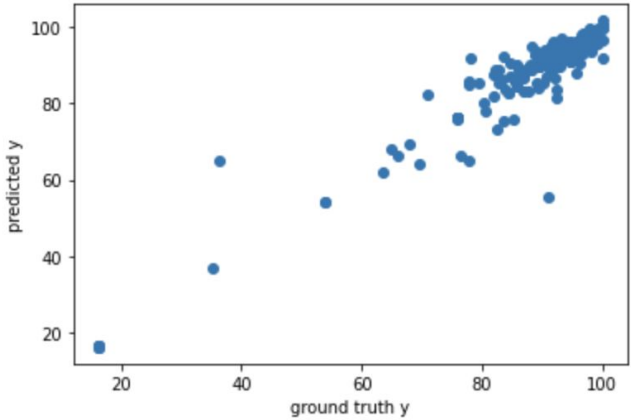
Graduation Rates Results - Visualization



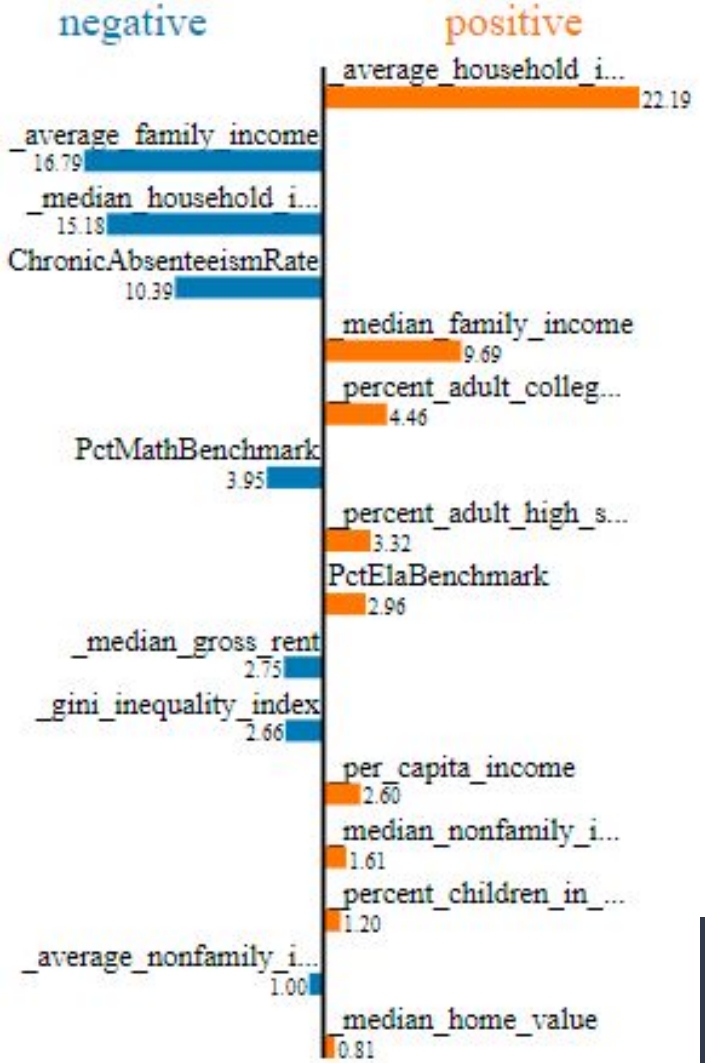
Linear Regression



Random Forest

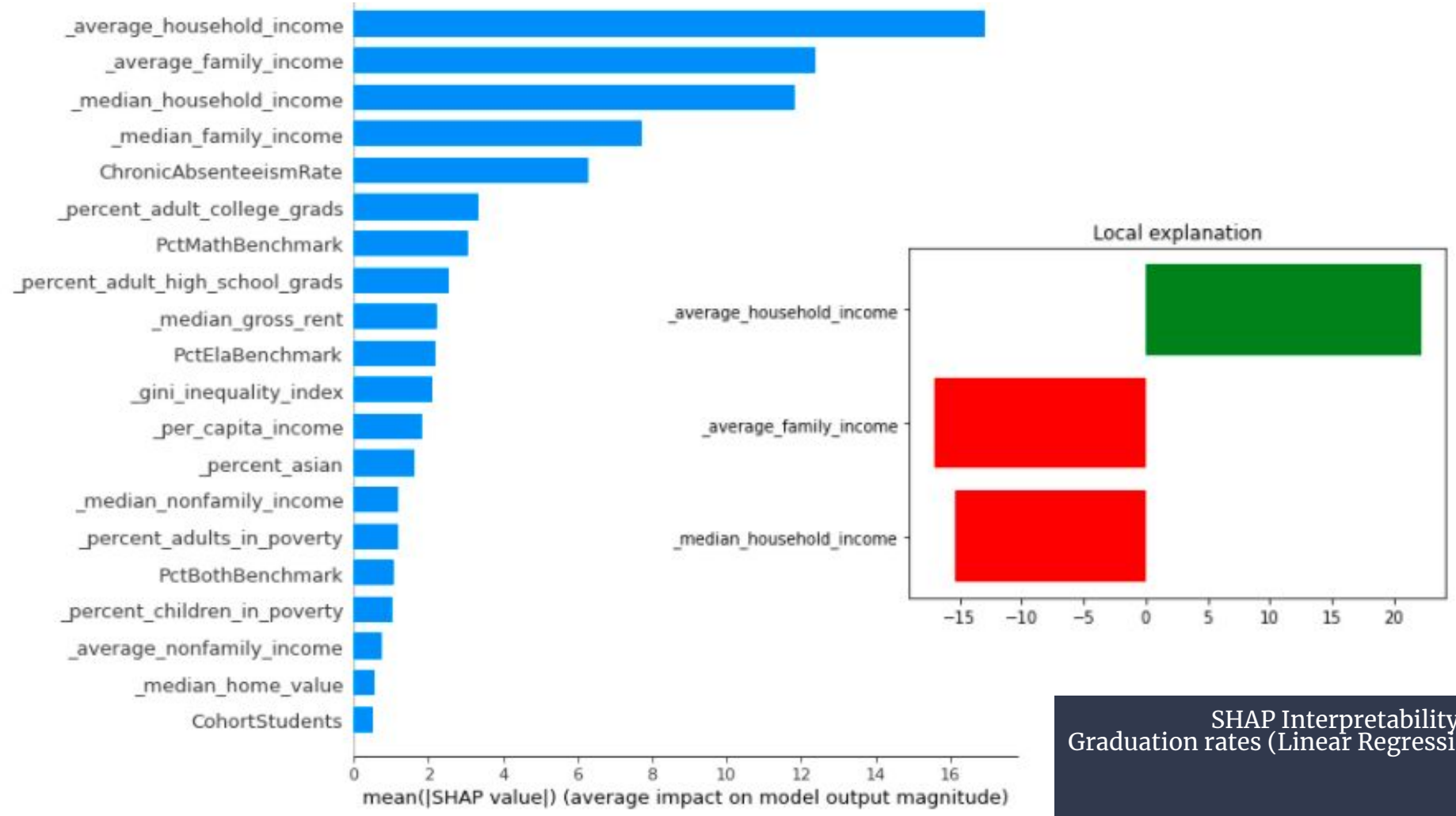


XGBoost



Feature	Value
_average_household_income	154781.63
_average_family_income	174525.22
_median_household_income	111411.00
ChronicAbsenteeismRate	19.60
_median_family_income	136000.00
_percent_adult_college_grads	28.70
PctMathBenchmark	62.96
_percent_adult_high_school_grads	15.18
PctElaBenchmark	90.12
_median_gross_rent	2063.00
_gini_inequality_index	0.48
_per_capita_income	53961.00
_median_nonfamily_income	72244.00
_percent_children_in_poverty	5.19
_average_nonfamily_income	107192.90
_median_home_value	1098700.00

LIME Interpretability on Graduation Rates (Linear Regression)



SHAP Interpretability on Graduation rates (Linear Regression)

Bias Analysis Results – Graduation Rates With Combined Dataset

1. Bias analysis on combined dataset

   - Grad rates below 70% vs. above

     - Random Forest Prediction for schools <= 70:
       - 73 schools
       - **r2 = 0.82 when ran through original RF model**

     - Random Forest Prediction for schools > 70:
       - 1471 schools
       - **r2 = 0.95 when ran through original RF model**

Bias Analysis Results – College Going Rates With Combined Dataset

1. Bias analysis on combined dataset

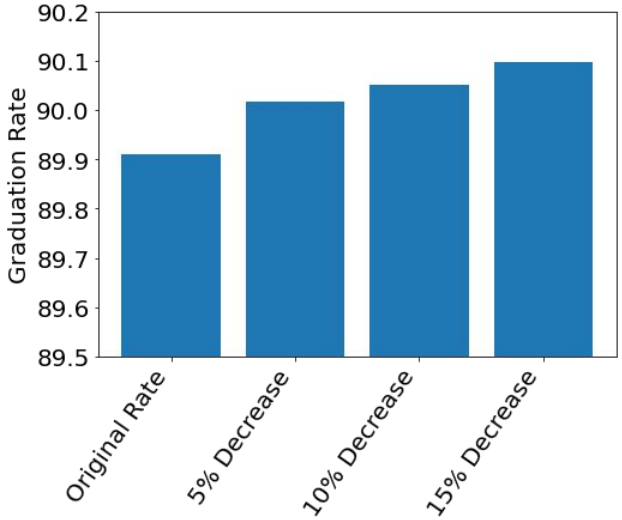
   - College going rates below 70% vs above

     - Random Forest Prediction for schools <= 70:
       - 748 schools
       - **r2 = 0.93 when ran through original RF model**

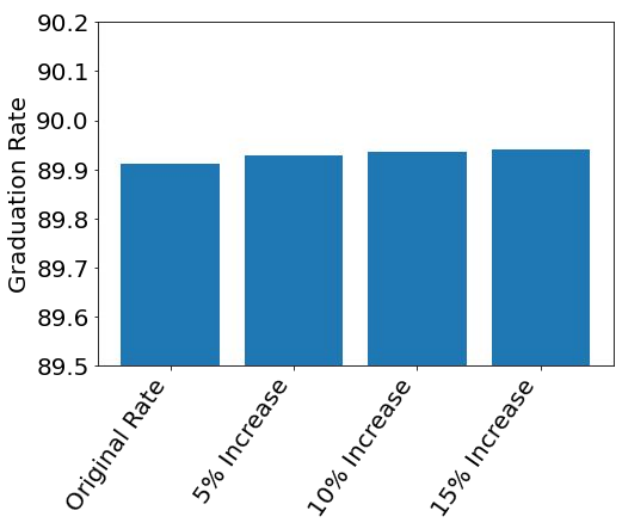
     - Random Forest Prediction for schools > 70:
       - 796 schools
       - **r2 = 0.84 when ran through original RF model**

Analysing Effects of Interventions

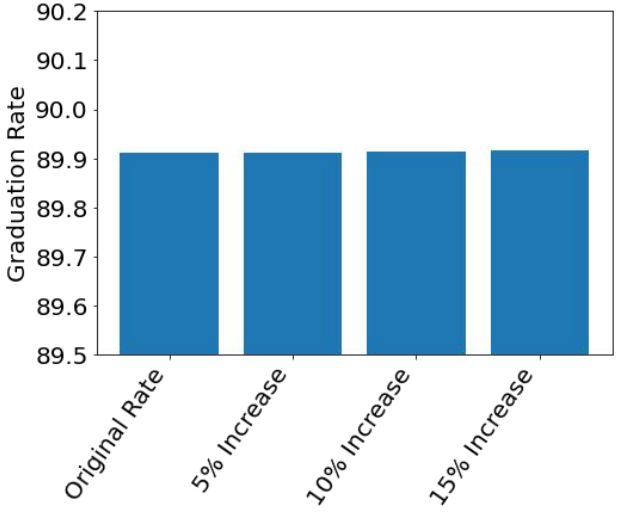
Sensitivity Analysis for Graduation Rate For Actionable Features



Chronic Absenteeism

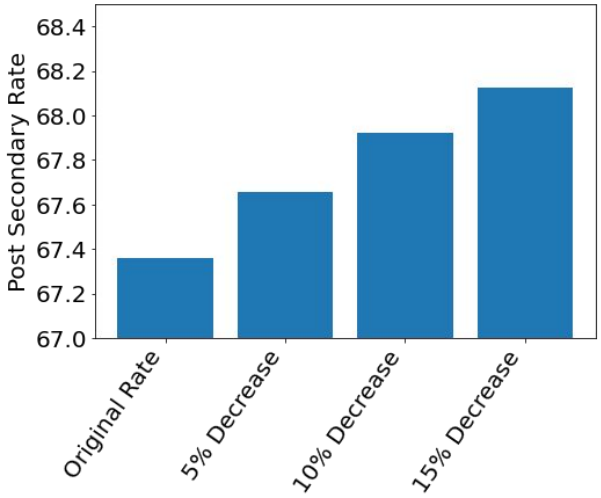


PctElaBenchmark

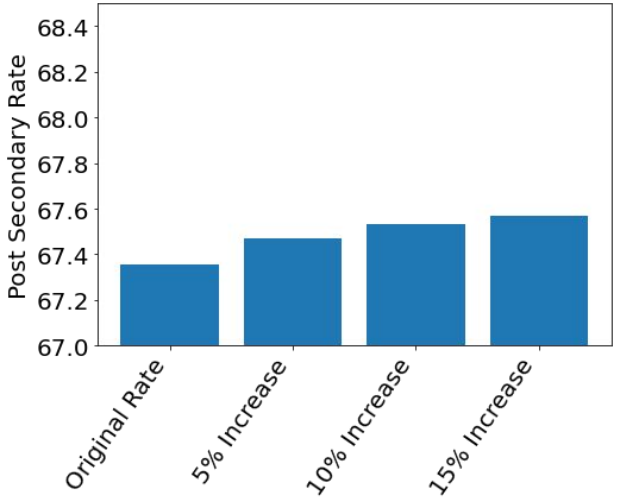


PctBothBenchmark

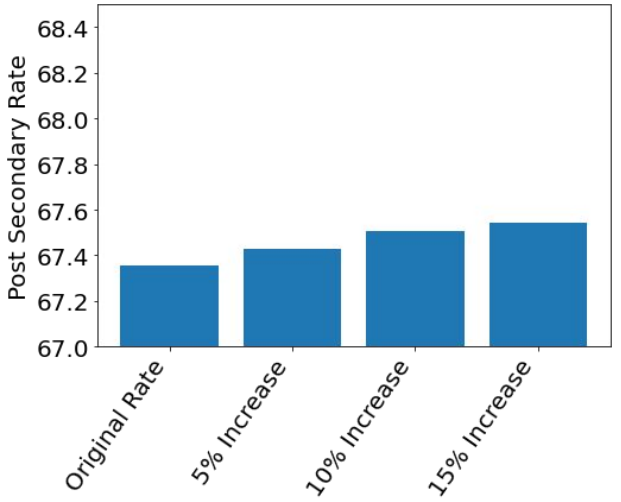
Sensitivity Analysis for Post Secondary Rate For Actionable Features



Chronic Absenteeism



PctElaBenchmark



Seal of Biliteracy

Project Summary

Project Achievements:

- Identified top features that impact student success for graduation rates and postsecondary enrollment
    - Top overall feature (both): **_population**
    - Top changeable feature (both): **Chronic Absenteeism**

- The top features for improving both graduation rates and postsecondary enrollment rates were all community-based features rather than school-based features

- Through analyzing how graduation and post secondary rates changed with increases and decreases to certain features, we are able to inform schools that decreasing the absenteeism rates for students will have the greatest impact on graduation success and pursuit of postsecondary education

Extensions and Next Steps:

1. Neural Network didn't perform well due to lack of data/tuning hyperparameters, see how adding more data or spending more time on hyperparameter tuning could make neural network surpass other models

2. Interesting to have more information on teachers in schools and their relative impact on these graduation/college going rates

3. Bias analysis using race data for geographic location of school and correlation between population of different races and student performance

Questions and Discussion

Thank you!