# Predicting Test Performance in California Public Schools

Gabrielle Wald

Springboard School of Data

## Contents

## 1. Introduction

### 1.1. Problem Statement

Research shows that high-poverty areas disproportionally educate children of color. A student’s race/ethnicity and social class are highly predictive of whether they will attend a high-poverty or high-minority school. For instance, African American and Hispanic students—even if they are not poor—are much more likely than White or Asian students to be in high-poverty schools. There is a growing body of evidence that shows that increased investment in education leads to better outcomes and that the positive effects are even larger among low-income students. On the other hand, it costs more to educate low-income students and provide them with a robust education capable of overcoming their initial disadvantages. There is a strong need to find more informed and granular causes that impact test achievement in schools. This investigation seeks to understand what other factors are associated with student test performance and ultimately predict the proportion of students in each school that pass standardized tests.

### 1.2. Objectives

I. Understand the current demographics of wealthy to high-poverty schools across the state of California.

II. Identify how much funding is available per pupil in wealthy vs high-poverty areas.

III. Learn what factors are most correlated with student performance (pass rate).

IV. Create a predictive model to find the proportion of students passing standard tests per school.

### 1.3. Beneficiaries

I. Department of education/ federal government to gain insights and potentially innovate adult (parents) education.

II. School administrators and policy makers to effectively identify schools that need more support and allocate resources to address tutoring needs, mentoring, and extracurricular activities.

III. Teachers can be oriented to provide more support for under-performing groups and reduce achievement gaps.

IV. From a parents perspective, the results can be used to select a high performing school that meets academic standards.

## 2. Data Wrangling

Here data was transformed and mapped from “raw” form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes.

### 2.1. Data Acquisition

The dataset for this project is unique. It has been constructed using several data files, referring to the academic year of 2018-2019 and publicly available, from California Assessment of Student and Progress (CAASPP), California Department of Education (CDE), and the National Center of Education Statistics (NCES). The assessment data refers to the Smarter Balanced Summative Assessment, a standardized test applied yearly to K-12 students of public schools in California. It contains student demographic information, and the response variable of interest (number of students who meet the standards per school). The other datasets include variables such as current expense per daily attendance, number of students per school receiving free or reduced price meals, total revenue and expenditure per pupil, and median household income per zip code.

### 2.2. Data Manipulation

To make sense of the assessment data file from CAASPP, there are 2 other files, Entity table, and Subgroup ID table. These files were merged with the assessment data file to join demographic, school and district names with the appropriate score data. The assessment data was extracted at school level keeping ‘demographic information’ and ‘the proportion of students meeting or exceeding the standards per school’. Two datasets were created, one for language arts and one for mathematics. The language arts dataset was used for this project. Demographic information originally contained in one column was remapped using a function to become several new columns (i.e. one column per demographic). Each demographic column has as value the number of students belonging to that demographic per school. The goal is to have one observation per row and one feature per column. More details:

- The assessment data file is provided in a ‘csv file’ format. For the record of 2018-2019, the original dataset contained 3,013,079 rows and 32 columns. Several columns were dropped such as ‘Filler’, ‘Test year’, ‘Test Type’, ‘District Code’, ‘Test ID’, ‘Grade’, and a few others that were not relevant to answer the questions proposed here.

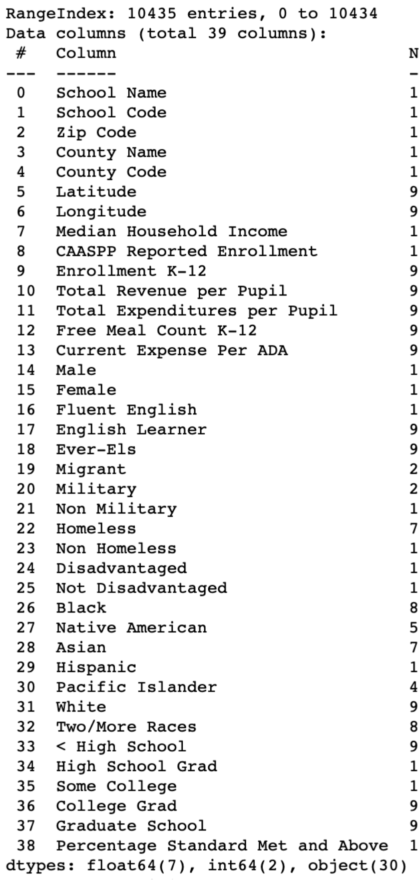
- The assessment data consists of state, counties, districts, schools, test scores, and demographic information on parent education, races, disabilities, gender, and English-Language fluency.
- The Subgroup ID lists the codes with the groups (e.g., gender, English-language fluency, economic status, ethnicity, (ethnicity for economically disadvantaged, ethnicity for not economically disadvantaged), disability status, parent education, migrant).

All data files were imported into pandas DataFrame. Additional datasets were merged via district code, district name, or zip code. The final result is a dataset with one row per school (10,435 rows) and 39 columns. More details:

- The median household income datafile contains four original columns, Rank, population, median household income, and zip code. The first two were dropped, leaving the last two columns. This was merged with the assessment data via zip code.
- The current expense per average daily attendance refers to the cost of education. The file contains district code, district name, current expense ADA and current expense per ADA. It was merged with the assessment data via district code.
    - **Average Daily Attendance (ADA):** Total ADA is defined as the total days of student attendance divided by the total days of instruction. The type of ADA used is annual district ADA (for the same year as the expenditures).

- **Cost Per ADA:** By district, the adjusted expenditures are divided by the total ADA to arrive at the Current Expense (or Cost) of Education per ADA.

- The Total revenue and expenditure data per pupil contains latitude, longitude, total revenue per pupil, total expenditure per pupil and district name. It was merged to the assessment data file via district name. String manipulation was used to match the district name of both files.
- Lastly, the file on free or reduced price meals is an indicator of low income family/ poverty. The file contains a number of free meal counts (k-12) per school and school code. It was merged via school code with the assessment data file.



```
RangeIndex: 10435 entries, 0 to 10434
Data columns (total 39 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   School Name                         10435 non-null  object
 1   School Code                         10435 non-null  int64
 2   Zip Code                            10435 non-null  int64
 3   County Name                         10435 non-null  object
 4   County Code                         10434 non-null  float64
 5   Latitude                            9310 non-null   float64
 6   Longitude                           9310 non-null   float64
 7   Median Household Income             10367 non-null  float64
 8   CAASPP Reported Enrollment          10434 non-null  object
 9   Enrollment K-12                     9992 non-null   float64
 10  Total Revenue per Pupil             9310 non-null   object
 11  Total Expenditures per Pupil        9310 non-null   object
 12  Free Meal Count K-12                9992 non-null   float64
 13  Current Expense Per ADA             9930 non-null   float64
 14  Male                                10384 non-null  object
 15  Female                              10182 non-null  object
 16  Fluent English                      10412 non-null  object
 17  English Learner                     9636 non-null   object
 18  Ever-Els                            9941 non-null   object
 19  Migrant                             2576 non-null   object
 20  Military                            2591 non-null   object
 21  Non Military                        10434 non-null  object
 22  Homeless                            7548 non-null   object
 23  Non Homeless                        10430 non-null  object
 24  Disadvantaged                       10286 non-null  object
 25  Not Disadvantaged                   10238 non-null  object
 26  Black                               8512 non-null   object
 27  Native American                     5195 non-null   object
 28  Asian                               7977 non-null   object
 29  Hispanic                            10221 non-null  object
 30  Pacific Islander                    4646 non-null   object
 31  White                               9868 non-null   object
 32  Two/More Races                      8326 non-null   object
 33  < High School                       9487 non-null   object
 34  High School Grad                    10043 non-null  object
 35  Some College                        10068 non-null  object
 36  College Grad                        9898 non-null   object
 37  Graduate School                     9451 non-null   object
 38  Percentage Standard Met and Above   10434 non-null  object
dtypes: float64(7), int64(2), object(30)
```

### 2.3. Data Cleaning

Addressing missing data is essential since datasets containing values such as blanks, NaNs or other placeholders are incompatible with scikit-learn estimators which assume that all values in an array are numerical. There are two strategies for solving this issue, one is by dropping the missing data which comes at the price of losing data that may be valuable (even though incomplete). The other strategy is imputation, missing values can be inferred from the known data. Here, the rows with missing data for the response variable were dropped, and in order to keep the highest number of observations imputation is used for the features.

I. There are reasons to believe that NaN means zero for the demographic data. So, NaN is imputed with zero; for other missing values (such as *)  -1 was imputed.

II. Dummy columns were created for features with imputed values signaling originally missing data.

III. Number of students was transformed to the percentage of students in several columns to keep the same format as the response variable.

IV. The median was imputed for a few features with missing values, as appropriate.

V. Data types were updated to integer, string or float as appropriate.

## 3. Exploratory Data Analysis

The data is explored to find trends, insights, and potential outliers based on visualization and hypothesis testing. These graphs and figures are important communication tools for collaborating in data science teams or presenting to business-oriented customers. The libraries used were matplotlib, seaborn, and pandas.

### 3.1. Overall Distribution Highlights

I. There is a higher proportion of disadvantaged students across K-12 public schools in CA (left skewed distribution).

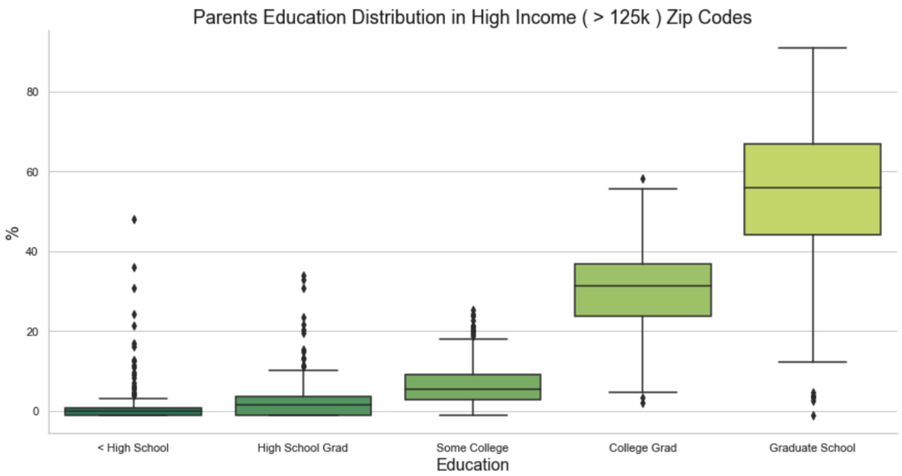
II. CA is a minority-majority state with a higher percent of students self-identifying as Hispanic.

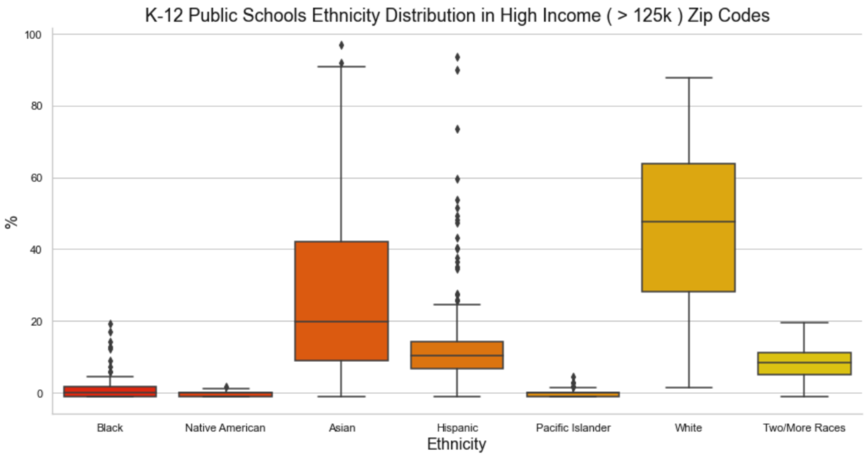
III. The second most common ethnicity is White, followed by Black and Asian, two or more races, Native American, and Pacific Islander.

### 3.2. Income, Education & Ethnicity

A. High Income Zip Codes (> $125k)

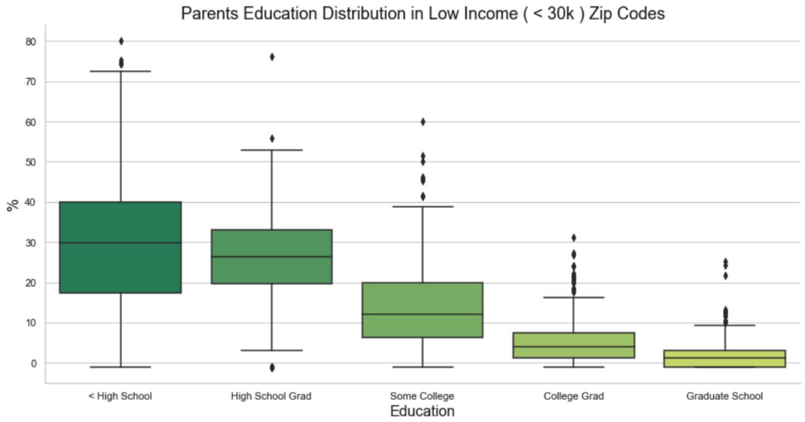
In zip codes where the median household income is higher than $125,000, there is a higher percentage of students passing the standards (left skewed distribution), and a larger number of parents have completed a four year college degree or higher. In fact the most common degree achieved by parents in this income bracket is graduate school, and the most common ethnicities here are White and Asian.

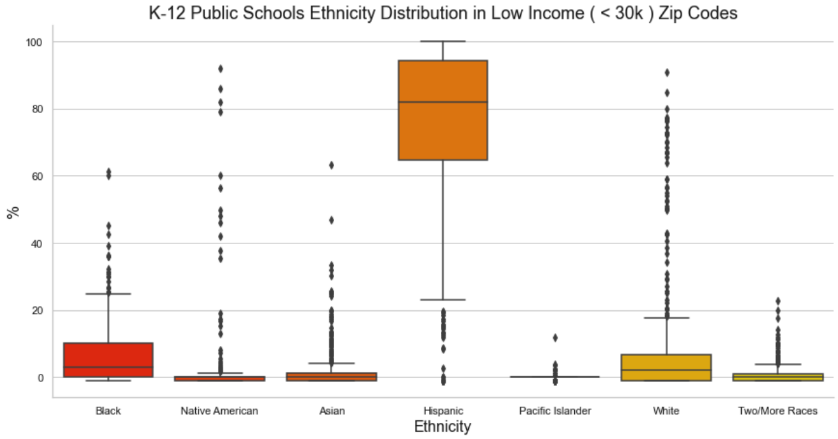




B. Low Income Zip Codes (< $30K)

In zip codes where the household income is less than $30,000, there is a lower percentage of students passing the standards (slightly right skewed distribution). Parents are more likely to have a high school diploma or be drop-outs. In fact, the most common level of education in this income bracket is less than high school. And, by far the most common ethnicity is hispanic.

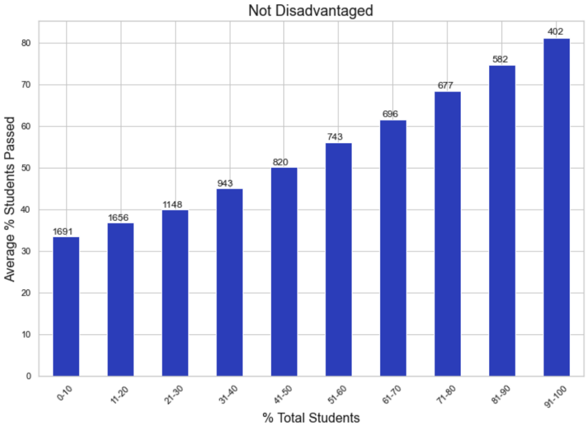




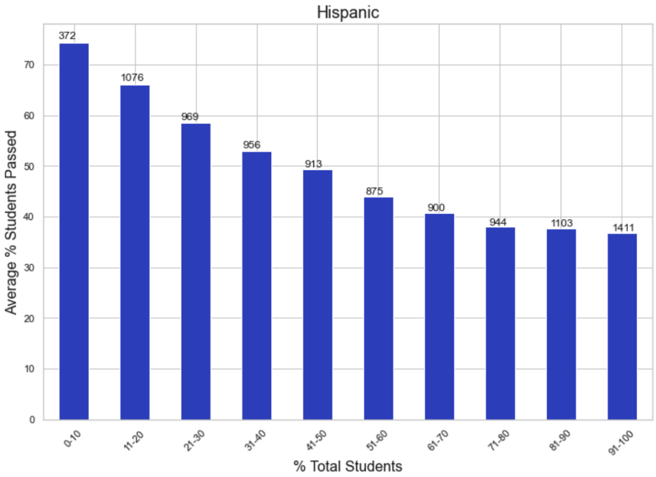
### 3.3. Pairwise Relationship

Pairwise comparison shows the relationship between two variables. In this case, the objective is to visualize the relationship between the response variable (average % Students Passed) and other independent variables. Linear and approximately uniform relationships are identified.

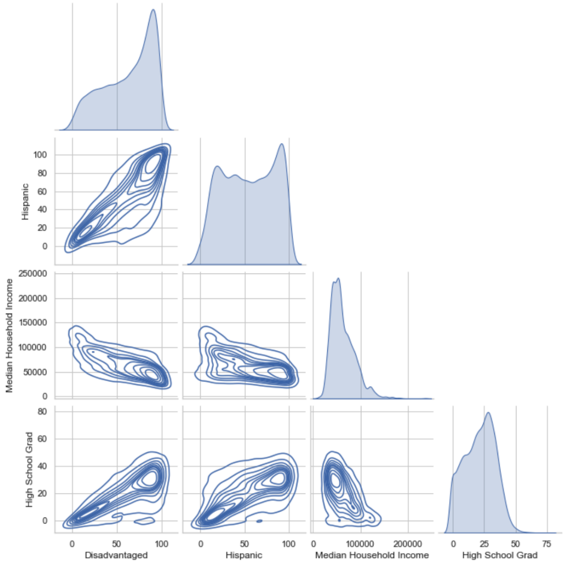
- Disadvantaged/ Not Disadvantaged: There is a strong linear relationship between the average percentage of students passing and these socioeconomic variables. As the number of students fitting the Not Disadvantaged category increases, so does the percentage of students passing the test standard. The opposite is true for Disadvantaged students.

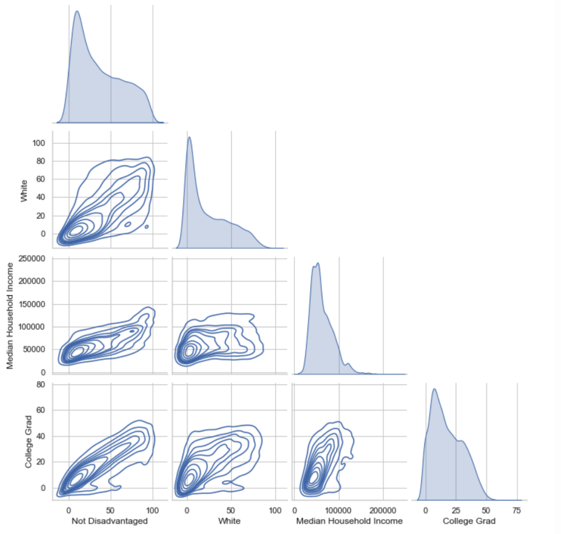


- Male/Female: As the percentage of female students increases, so does the percentage of students passing per school. That is, schools with a higher percentage of female students are performing better in test standards. The opposite is true for male.
- Ethnicity: Asian show a positive linear relationship with the response variable, while Hispanic show a negative linear relationship with the response variable.



The plots below show the relationship between some of the most correlated independent variables. For instance, there’s a high positive correlation between Hispanic and Disadvantaged, and Hispanic and High School Graduate. On the other hand, there is a high positive correlation between Asian and Not Disadvantaged, and Asian and Graduate Level education. Higher education is positively correlated with median household income. Knowing that Asian students in general outperform Hispanic students in test performance, it seems that having parents with higher education and therefore high-paying jobs are relevant factors.

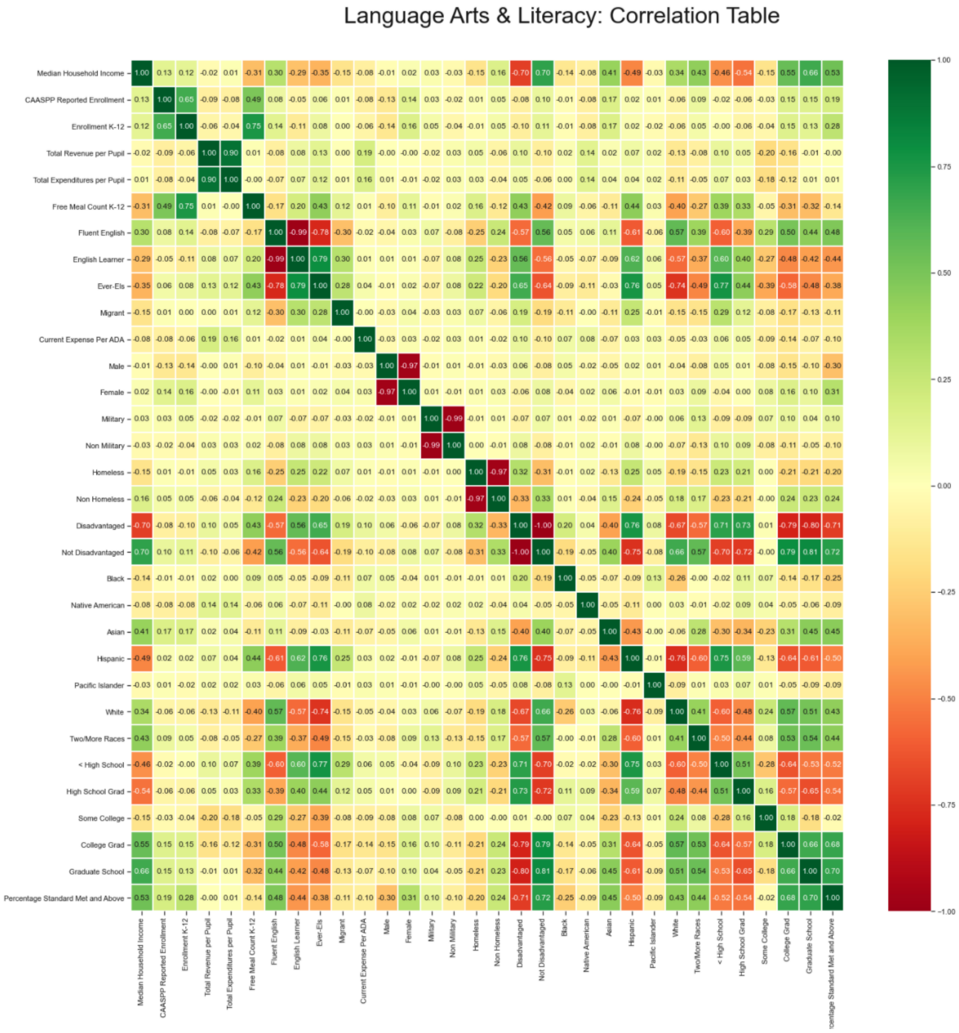




## 3.4. Correlation

### I. Pearson Correlation Table

Measure the strength and direction of the linear relationship between the two variables. The correlation coefficient can range from -1 to +1, with -1 indicating a perfect negative correlation, +1 indicating a perfect positive correlation, and 0 indicating no correlation at all.



Language Arts & Literacy: Correlation Table

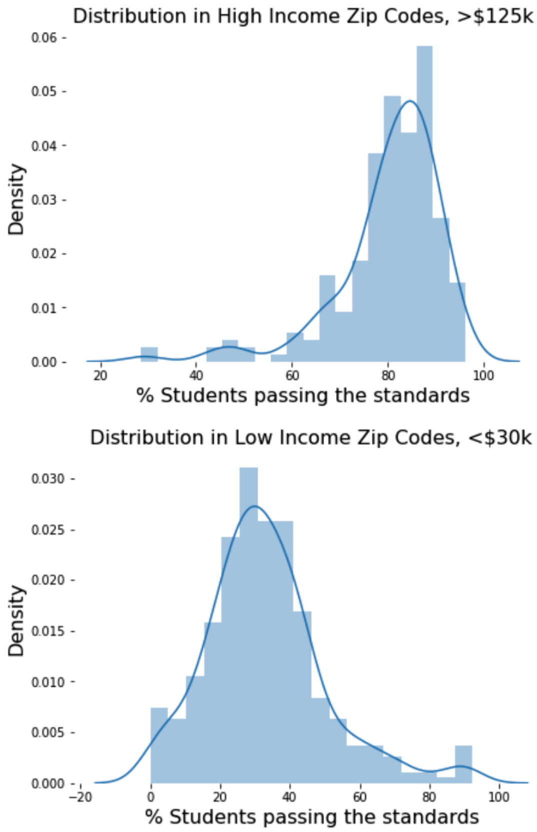
| | Median Household Income | CAASPP Reported Enrollment | Enrollment K-12 | Total Revenue per Pupil | Total Expenditures per Pupil | Free Meal Count K-12 | Fluent English | English Learner | Ever-Els | Migrant | Current Expense Per ADA | Male | Female | Military | Non Military | Homeless | Non Homeless | Disadvantaged | Not Disadvantaged | Black | Native American | Asian | Hispanic | Pacific Islander | White | Two/More Races | < High School | High School Grad | Some College | College Grad | Graduate School | centage Standard Met and Above |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Median Household Income | 1.00 | 0.13 | 0.12 | -0.02 | 0.01 | -0.31 | 0.30 | -0.29 | -0.35 | -0.15 | -0.08 | -0.01 | 0.02 | 0.03 | -0.03 | -0.15 | 0.16 | -0.70 | 0.70 | -0.14 | -0.08 | 0.41 | -0.49 | -0.03 | 0.34 | 0.43 | -0.46 | -0.54 | -0.15 | 0.55 | 0.66 | 0.53 |
| CAASPP Reported Enrollment | 0.13 | 1.00 | 0.65 | -0.09 | -0.08 | 0.49 | 0.08 | -0.05 | 0.06 | 0.01 | -0.08 | -0.13 | 0.14 | 0.03 | -0.02 | 0.01 | 0.05 | -0.08 | 0.10 | -0.01 | -0.08 | 0.17 | 0.02 | 0.01 | -0.06 | 0.09 | -0.02 | -0.06 | -0.03 | 0.15 | 0.15 | 0.19 |
| Enrollment K-12 | 0.12 | 0.65 | 1.00 | -0.06 | -0.04 | 0.75 | 0.14 | -0.11 | 0.08 | 0.00 | -0.06 | -0.14 | 0.16 | 0.05 | -0.04 | -0.01 | 0.05 | -0.10 | 0.11 | -0.01 | -0.08 | 0.17 | 0.02 | -0.02 | -0.06 | 0.05 | -0.00 | -0.06 | -0.04 | 0.15 | 0.13 | 0.28 |
| Total Revenue per Pupil | -0.02 | -0.09 | -0.06 | 1.00 | 0.90 | 0.01 | -0.08 | 0.08 | 0.13 | 0.00 | 0.19 | -0.00 | -0.00 | -0.02 | 0.03 | 0.05 | -0.06 | 0.10 | -0.10 | 0.02 | 0.14 | 0.02 | 0.07 | 0.02 | -0.13 | -0.08 | 0.10 | 0.05 | -0.20 | -0.16 | -0.01 | -0.00 |
| Total Expenditures per Pupil | 0.01 | -0.08 | -0.04 | 0.90 | 1.00 | -0.00 | -0.07 | 0.07 | 0.12 | 0.01 | 0.16 | 0.01 | -0.01 | -0.02 | 0.03 | 0.03 | -0.04 | 0.05 | -0.06 | 0.00 | 0.14 | 0.04 | 0.04 | 0.02 | -0.11 | -0.05 | 0.07 | 0.03 | -0.18 | -0.12 | 0.01 | 0.01 |
| Free Meal Count K-12 | -0.31 | 0.49 | 0.75 | 0.01 | -0.00 | 1.00 | -0.17 | 0.20 | 0.43 | 0.12 | 0.01 | -0.10 | 0.11 | -0.01 | 0.02 | 0.16 | -0.12 | 0.43 | -0.42 | 0.09 | -0.06 | -0.11 | 0.44 | 0.03 | -0.40 | -0.27 | 0.39 | 0.33 | -0.05 | -0.31 | -0.32 | -0.14 |
| Fluent English | 0.30 | 0.08 | 0.14 | -0.08 | -0.07 | -0.17 | 1.00 | -0.99 | -0.78 | -0.30 | -0.02 | -0.04 | 0.03 | 0.07 | -0.08 | -0.25 | 0.24 | -0.57 | 0.56 | 0.05 | 0.06 | 0.11 | -0.61 | -0.06 | 0.57 | 0.39 | -0.60 | -0.39 | 0.29 | 0.50 | 0.44 | 0.48 |
| English Learner | -0.29 | -0.05 | -0.11 | 0.08 | 0.07 | 0.20 | -0.99 | 1.00 | 0.79 | 0.30 | 0.01 | 0.01 | 0.01 | -0.07 | 0.08 | 0.25 | -0.23 | 0.56 | -0.56 | -0.05 | -0.07 | -0.09 | 0.62 | 0.06 | -0.57 | -0.37 | 0.60 | 0.40 | -0.27 | -0.48 | -0.42 | -0.44 |
| Ever-Els | -0.35 | 0.06 | 0.08 | 0.13 | 0.12 | 0.43 | -0.78 | 0.79 | 1.00 | 0.28 | 0.04 | -0.01 | 0.02 | -0.07 | 0.08 | 0.22 | -0.20 | 0.65 | -0.64 | -0.09 | -0.11 | -0.03 | 0.76 | 0.05 | -0.74 | -0.49 | 0.77 | 0.44 | -0.39 | -0.58 | -0.48 | -0.38 |
| Migrant | -0.15 | 0.01 | 0.00 | 0.00 | 0.01 | 0.12 | -0.30 | 0.30 | 0.28 | 1.00 | -0.00 | -0.03 | 0.04 | -0.03 | 0.03 | 0.07 | -0.06 | 0.19 | -0.19 | -0.11 | -0.00 | -0.11 | 0.25 | -0.01 | -0.15 | -0.15 | 0.29 | 0.12 | -0.08 | -0.17 | -0.13 | -0.11 |
| Current Expense Per ADA | -0.08 | -0.08 | -0.06 | 0.19 | 0.16 | 0.01 | -0.02 | 0.01 | 0.04 | -0.00 | 1.00 | -0.03 | 0.03 | -0.02 | 0.03 | 0.01 | -0.02 | 0.10 | -0.10 | 0.07 | 0.08 | -0.07 | 0.03 | 0.03 | -0.05 | -0.03 | 0.06 | 0.05 | -0.09 | -0.14 | -0.07 | -0.10 |
| Male | -0.01 | -0.13 | -0.14 | -0.00 | 0.01 | -0.10 | -0.04 | 0.01 | -0.01 | -0.03 | -0.03 | 1.00 | -0.97 | -0.01 | 0.01 | -0.01 | -0.03 | 0.06 | -0.08 | 0.05 | -0.02 | -0.05 | 0.02 | 0.01 | -0.04 | -0.08 | 0.05 | 0.01 | -0.08 | -0.15 | -0.10 | -0.30 |
| Female | 0.02 | 0.14 | 0.16 | -0.00 | -0.01 | 0.11 | 0.03 | 0.01 | 0.02 | 0.04 | 0.03 | -0.97 | 1.00 | 0.01 | -0.01 | 0.01 | 0.03 | -0.06 | 0.08 | -0.04 | 0.02 | 0.06 | -0.01 | -0.01 | 0.03 | 0.09 | -0.04 | 0.00 | 0.08 | 0.16 | 0.10 | 0.31 |
| Military | 0.03 | 0.03 | 0.05 | -0.02 | -0.02 | -0.01 | 0.07 | -0.07 | -0.07 | -0.03 | -0.02 | -0.01 | 0.01 | 1.00 | -0.99 | -0.01 | 0.01 | -0.07 | 0.07 | 0.01 | -0.02 | 0.01 | -0.07 | -0.00 | 0.06 | 0.13 | -0.09 | -0.09 | 0.07 | 0.10 | 0.04 | 0.10 |
| Non Military | -0.03 | -0.02 | -0.04 | 0.03 | 0.03 | 0.02 | -0.08 | 0.08 | 0.08 | 0.03 | 0.03 | 0.01 | -0.01 | -0.99 | 1.00 | 0.00 | -0.01 | 0.08 | -0.08 | -0.01 | 0.02 | -0.01 | 0.08 | -0.00 | -0.07 | -0.13 | 0.10 | 0.09 | -0.08 | -0.11 | -0.05 | -0.10 |
| Homeless | -0.15 | 0.01 | -0.01 | 0.05 | 0.03 | 0.16 | -0.25 | 0.25 | 0.22 | 0.07 | 0.01 | -0.01 | 0.01 | -0.01 | 0.00 | 1.00 | -0.97 | 0.32 | -0.31 | -0.01 | 0.02 | -0.13 | 0.25 | 0.05 | -0.19 | -0.15 | 0.23 | 0.21 | 0.00 | -0.21 | -0.21 | -0.20 |
| Non Homeless | 0.16 | 0.05 | 0.05 | -0.06 | -0.04 | -0.12 | 0.24 | -0.23 | -0.20 | -0.06 | -0.02 | -0.03 | 0.03 | 0.01 | -0.01 | -0.97 | 1.00 | -0.33 | 0.33 | 0.01 | -0.04 | 0.15 | -0.24 | -0.05 | 0.18 | 0.17 | -0.23 | -0.21 | -0.00 | 0.24 | 0.23 | 0.24 |
| Disadvantaged | -0.70 | -0.08 | -0.10 | 0.10 | 0.05 | 0.43 | -0.57 | 0.56 | 0.65 | 0.19 | 0.10 | 0.06 | -0.06 | -0.07 | 0.08 | 0.32 | -0.33 | 1.00 | -1.00 | 0.20 | 0.04 | -0.40 | 0.76 | 0.08 | -0.67 | -0.57 | 0.71 | 0.73 | 0.01 | -0.79 | -0.80 | -0.71 |
| Not Disadvantaged | 0.70 | 0.10 | 0.11 | -0.10 | -0.06 | -0.42 | 0.56 | -0.56 | -0.64 | -0.19 | -0.10 | -0.08 | 0.08 | 0.07 | -0.08 | -0.31 | 0.33 | -1.00 | 1.00 | -0.19 | -0.05 | 0.40 | -0.75 | -0.08 | 0.66 | 0.57 | -0.70 | -0.72 | -0.00 | 0.79 | 0.81 | 0.72 |
| Black | -0.14 | -0.01 | -0.01 | 0.02 | 0.00 | 0.09 | 0.05 | -0.05 | -0.09 | -0.11 | 0.07 | 0.05 | -0.04 | 0.01 | -0.01 | -0.01 | 0.01 | 0.20 | -0.19 | 1.00 | -0.05 | -0.07 | -0.09 | 0.13 | -0.26 | -0.00 | -0.02 | 0.11 | 0.07 | -0.14 | -0.17 | -0.25 |
| Native American | -0.08 | -0.08 | -0.08 | 0.14 | 0.14 | -0.06 | 0.06 | -0.07 | -0.11 | -0.00 | 0.08 | -0.02 | 0.02 | -0.02 | 0.02 | 0.02 | -0.04 | 0.04 | -0.05 | -0.05 | 1.00 | -0.05 | -0.11 | 0.00 | 0.03 | -0.01 | -0.02 | 0.09 | 0.04 | -0.05 | -0.06 | -0.09 |
| Asian | 0.41 | 0.17 | 0.17 | 0.02 | 0.04 | -0.11 | 0.11 | -0.09 | -0.03 | -0.11 | -0.07 | -0.05 | 0.06 | 0.01 | -0.01 | -0.13 | 0.15 | -0.40 | 0.40 | -0.07 | -0.05 | 1.00 | -0.43 | -0.00 | -0.06 | 0.28 | -0.30 | -0.34 | -0.23 | 0.31 | 0.45 | 0.45 |
| Hispanic | -0.49 | 0.02 | 0.02 | 0.07 | 0.04 | 0.44 | -0.61 | 0.62 | 0.76 | 0.25 | 0.03 | 0.02 | -0.01 | -0.07 | 0.08 | 0.25 | -0.24 | 0.76 | -0.75 | -0.09 | -0.11 | -0.43 | 1.00 | -0.01 | -0.76 | -0.60 | 0.75 | 0.59 | -0.13 | -0.64 | -0.61 | -0.50 |
| Pacific Islander | -0.03 | 0.01 | -0.02 | 0.02 | 0.02 | 0.03 | -0.06 | 0.06 | 0.05 | -0.01 | 0.03 | 0.01 | -0.01 | -0.00 | -0.00 | 0.05 | -0.05 | 0.08 | -0.08 | 0.13 | 0.00 | -0.00 | -0.01 | 1.00 | -0.09 | 0.01 | 0.03 | 0.07 | 0.01 | -0.05 | -0.09 | -0.09 |
| White | 0.34 | -0.06 | -0.06 | -0.13 | -0.11 | -0.40 | 0.57 | -0.57 | -0.74 | -0.15 | -0.05 | -0.04 | 0.03 | 0.06 | -0.07 | -0.19 | 0.18 | -0.67 | 0.66 | -0.26 | 0.03 | -0.06 | -0.76 | -0.09 | 1.00 | 0.41 | -0.60 | -0.48 | 0.24 | 0.57 | 0.51 | 0.43 |
| Two/More Races | 0.43 | 0.09 | 0.05 | -0.08 | -0.05 | -0.27 | 0.39 | -0.37 | -0.49 | -0.15 | -0.03 | -0.08 | 0.09 | 0.13 | -0.13 | -0.15 | 0.17 | -0.57 | 0.57 | -0.00 | -0.01 | 0.28 | -0.60 | 0.01 | 0.41 | 1.00 | -0.50 | -0.44 | 0.08 | 0.53 | 0.54 | 0.44 |
| < High School | -0.46 | -0.02 | -0.00 | 0.10 | 0.07 | 0.39 | -0.60 | 0.60 | 0.77 | 0.29 | 0.06 | 0.05 | -0.04 | -0.09 | 0.10 | 0.23 | -0.23 | 0.71 | -0.70 | -0.02 | -0.02 | -0.30 | 0.75 | 0.03 | -0.60 | -0.50 | 1.00 | 0.51 | -0.28 | -0.64 | -0.53 | -0.52 |
| High School Grad | -0.54 | -0.06 | -0.06 | 0.05 | 0.03 | 0.33 | -0.39 | 0.40 | 0.44 | 0.12 | 0.05 | 0.01 | 0.00 | -0.09 | 0.09 | 0.21 | -0.21 | 0.73 | -0.72 | 0.11 | 0.09 | -0.34 | 0.59 | 0.07 | -0.48 | -0.44 | 0.51 | 1.00 | 0.16 | -0.57 | -0.65 | -0.54 |
| Some College | -0.15 | -0.03 | -0.04 | -0.20 | -0.18 | -0.05 | 0.29 | -0.27 | -0.39 | -0.08 | -0.09 | -0.08 | 0.08 | 0.07 | -0.08 | 0.00 | -0.00 | 0.01 | -0.00 | 0.07 | 0.04 | -0.23 | -0.13 | 0.01 | 0.24 | 0.08 | -0.28 | 0.16 | 1.00 | 0.18 | -0.18 | -0.02 |
| College Grad | 0.55 | 0.15 | 0.15 | -0.16 | -0.12 | -0.31 | 0.50 | -0.48 | -0.58 | -0.17 | -0.14 | -0.15 | 0.16 | 0.10 | -0.11 | -0.21 | 0.24 | -0.79 | 0.79 | -0.14 | -0.05 | 0.31 | -0.64 | -0.05 | 0.57 | 0.53 | -0.64 | -0.57 | 0.18 | 1.00 | 0.66 | 0.68 |
| Graduate School | 0.66 | 0.15 | 0.13 | -0.01 | 0.01 | -0.32 | 0.44 | -0.42 | -0.48 | -0.13 | -0.07 | -0.10 | 0.10 | 0.04 | -0.05 | -0.21 | 0.23 | -0.80 | 0.81 | -0.17 | -0.06 | 0.45 | -0.61 | -0.09 | 0.51 | 0.54 | -0.53 | -0.65 | -0.18 | 0.66 | 1.00 | 0.70 |
| Percentage Standard Met and Above | 0.53 | 0.19 | 0.28 | -0.00 | 0.01 | -0.14 | 0.48 | -0.44 | -0.38 | -0.11 | -0.10 | -0.30 | 0.31 | 0.10 | -0.10 | -0.20 | 0.24 | -0.71 | 0.72 | -0.25 | -0.09 | 0.45 | -0.50 | -0.09 | 0.43 | 0.44 | -0.52 | -0.54 | -0.02 | 0.68 | 0.70 | 1.00 |

### 3.5. Hypothesis Testing

Hypothesis testing is a test of significance used to assess the plausibility of a hypothesis. The test provides evidence concerning the plausibility of the hypothesis, given the data. A two-sided t-test was used for all pairs of student groups assuming independence.

I. **Students living in higher income zip codes pass standard test scores at a higher rate than students living in lower income zip codes.**

A. We reject the null. In order words, the difference between student scores in higher vs lower income zip codes is statistically significant. The distribution can be seen in the figure below.



Distribution in High Income Zip Codes, >$125k

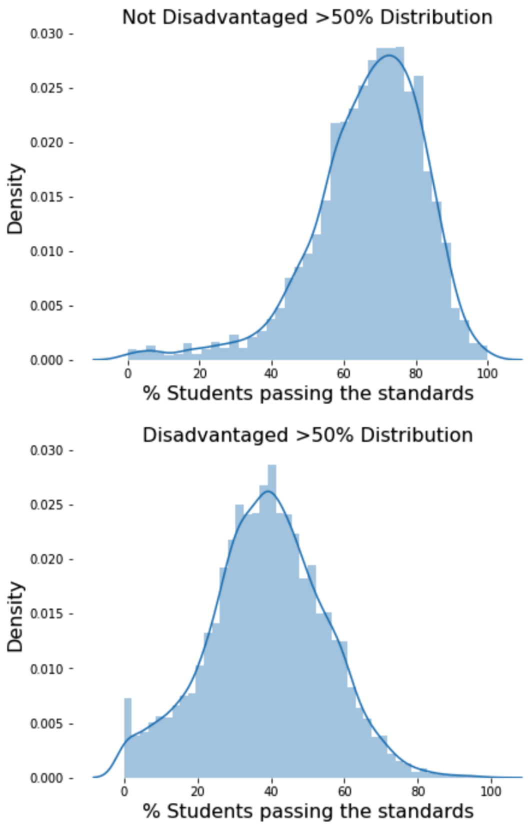
Distribution in Low Income Zip Codes, <$30k

II. **Schools with more students fluent in English pass standard test scores at a higher rate than schools with more English learners.**

A. We reject the null. There is a statistically significant difference in the proportion of students passing the test standard in school with a higher number of fluent English students.

### III. Schools with more ‘not disadvantaged’ students pass standard test scores at a higher rate than schools with more disadvantaged students.

A. We reject the null. There’s evidence to suggest that schools with a higher proportion of not disadvantaged students pass the test standard at higher rates. The distribution can be seen in the figure below.



## 4. Modeling

### 4.1. Evaluation Metrics: MAE, RMSE, and R2

- Mean Absolute Error (MAE): MAE is the mean of the absolute value of the errors.
- Root Mean Squared Error (RMSE): RMSE is the square root of the mean of the squared errors.
- R2: R2 is the number that indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. Basically, R2 represents how accurate the model is. R2 shows how well data points fit a curve or line.

### 4.2. Regression

Regression analysis is a subfield of supervised machine learning. It aims to model the relationship between a certain number of features and a continuous target variable. The response or target variable here is the ‘Percentage Standard Met and Above’. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.

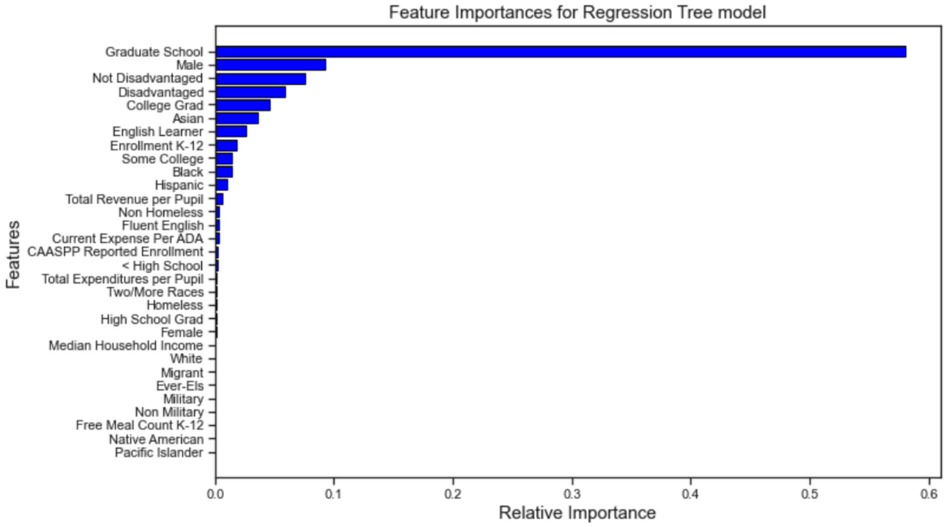
The model is fit using Scikit-learn and statsmodel. The major affecting features to predict ‘Percentage Standard Met and Above’ are number of ‘Fluent English’ speakers and ‘Female’ students. The result of train and test split for the Linear Regression model are as follows: MAE: 0.08, RMSE: 0.104 and R2 score: 0.7403.

### 4.3. Lasso

Lasso is a regularization method with the capability of "selecting" variables by penalizing the high value coefficients. In other words, lasso performs feature selection by shrinking some coefficients to zero, allowing the model to select a small number of variables as the final predictors. It uses the L1 norm. Lasso did not improve the model much even after hyperparameter tuning. The result of train and test split for the Lasso Regression model are as follows: MAE: 0.08, RMSE: 0.107 and R2 score: 0.7422.

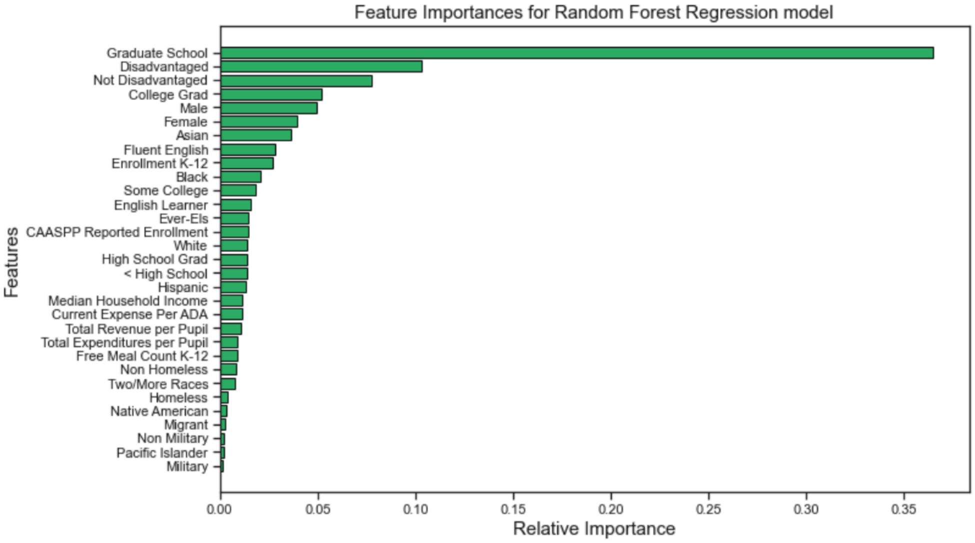
### 4.4. Decision Tree

Decision trees regression normally uses mean squared error (MSE) to decide to split a node in two or more sub-nodes. The major affecting features to predict ‘Percentage Standard Met and Above’ are the number of students whose parents have ‘Graduate level’ education, number of ‘male’ students, and socioeconomic status ‘Not Disadvantaged’. The result of train and test split for the Decision Tree model are as follows: MAE: 0.085, RMSE: 0.111 and R2 score: 0.699.



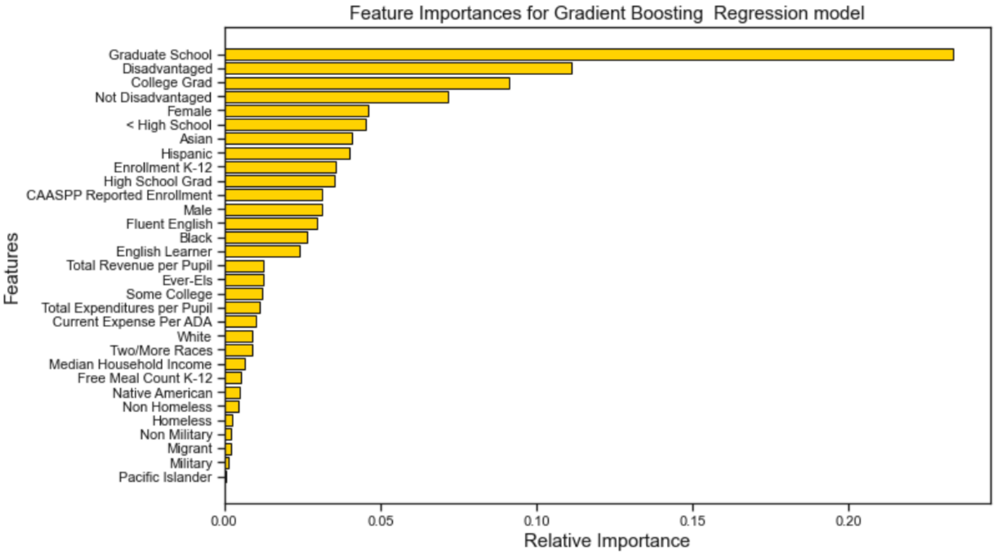
### 4.5. Random Forest

A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement if bootstrap = True (default). The major affecting feature to predict ‘Percentage Standard Met and Above’ is again the number of students whose parents have a ‘Graduate Level’ degree, followed by the number of students belonging to socioeconomic status ‘Disadvantaged’ and ‘Not Disadvantaged’. The result of train and test split for the Random Forest Regression model are as follows: MAE: 0.07, RMSE: 0.092 and R2 score: 0.797.



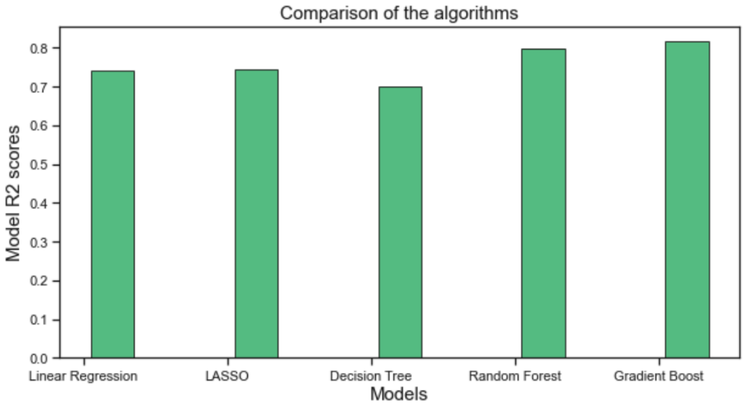
### 4.6. Gradient Boosting

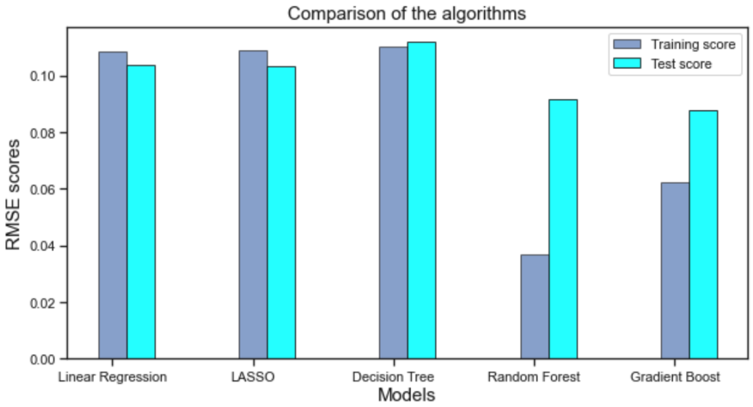
Gradient Boosting builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage a regression tree is fit on the negative gradient of the given loss function. The major affecting feature to predict ‘Percentage Standard Met and Above’ is once again the number of students whose parents have a ‘Graduate Level’ degree, followed by the number of students belonging to socioeconomic status ‘Disadvantaged’ and ‘College Grad’. The result of train and test split for the Gradient Boosting Regression model are as follows: MAE: 0.067, RMSE: 0.087 and R2 score: 0.815.



## 4.7. Models Comparison

In order to predict the proportion of students passing standards in California K-12 public schools 31 features were considered, either directly from the dataset or engineered from the data. The most relevant features in terms of relative importances are parents' level of education, socioeconomic status, and students' ethnicity followed by other demographics. To measure model accuracy R2 was plotted for test data, and MAE and RMSE were plotted for training and test data. The best performing model with R2 = 0.815 is the Gradient Boosting.





## 5. Conclusion and Recommendations

I. The higher the level of parental education, the higher the achievement of students.

II. Female students exceed male students in English language arts test performance.

III. Economically disadvantaged students have more difficulties than not-economically disadvantaged students.

IV. Students fluent in English achieve higher performance, while English learners score lower.

V. Asian students achieve higher performance, while Hispanic students score lower.

It is clear that high scores are strongly correlated with students raised by highly educated families. And, there is also a strong correlation between education level and higher household income. As a result, students in highly educated families are exposed to various learning opportunities including sports, traveling, musical instruments, arts, tutoring or other enriching activities. Their parents are also better equipped academically to help with homework and school assignments.

To improve academic achievement the following suggestions address factors that may be undermining performance:

- Creation of tutoring or mentoring programs where students can have 1 to 1 interaction time.
- After school programs dedicated to English learners to help with school assignments.
    - Play time that provides language instruction through story-telling, craft and fun activities.

## 6. Limitations and Future Research

There is no ideal model to predict proportions. It’s problematic to predict a numerical variable that’s bounded between 0 and 1. Nonetheless, the predictions for this project fall well into the range 0 to 1, and the metrics show well performing models. A potential solution to try next is to use generalized linear models with logistic regression, which can be done in R. Adding new variables such as number of effective teachers per school, teachers salary, information on extracurricular activities, and so on can improve the models and potential solutions.

Special thanks to my mentor Nadav, who has been a great support and genuinely involved with the outcome of this project.