# Predicting K-12 School Performance in California based on School Funding and Regional Socioeconomic Factors

**Andrew Hariri**
University of Southern California
Viterbi School of Engineering
*hariria@usc.edu*

**Leena Mathur**
University of Southern California
Viterbi School of Engineering
*lmathur@usc.edu*

**Nicole Ng**
University of Southern California
Viterbi School of Engineering
*shinyinn@usc.edu*

**Jae Shim**
University of Southern California
Viterbi School of Engineering
*jaeshim@usc.edu*

## Abstract

*Pervasive* and *persistent* achievement gaps exist across K-12 schools in California. School performance data from state-wide standardized tests shows that K-12 achievement gaps are often correlated with demographic, socioeconomic, geographic, and school-related characteristics. Large variations in annual expenses per pupil, adjusted for cost of living, exist across school districts. Property taxes and housing prices within a given district significantly impact the amount of local funding received by K-12 public schools. While some education policy researchers have theorized that school funding functions as a strong determinant of school performance, there is an ongoing debate in the education research community regarding the importance of school funding in relation to school performance. The primary goal of our project was to develop and leverage machine learning (ML) models to identify school-related and district-related financial and socioeconomic factors predictive of K-12 school performance. Our project contributes insights into key factors that are predictive of California K-12 school mathematics and reading proficiency across all schools, as well as for high-schools, middle-schools, and primary schools, respectively. Future work involves leveraging insights from our modeling results to develop improved school performance prediction models, leveraging insights from our feature importance results to optimize the placement of community interventions across census districts that improve important dimensions of a school district to maximize school performance, as well as exploring the possibility of optimizing the allocation of school funding to maximize school performance.

## 1 Introduction

Achievement gaps in California K-12 public schools, identified through the analysis of student performance on state-wide standardized tests, are largely correlated with a variety of demographic, socioeconomic, geographic, and school-related characteristics In 2018, socioeconomically advantaged students in California scored twice as high on math and reading standardized tests compared to socioeconomically disadvantaged students [6]. On California mathematics proficiency exams in 2019, 21% of Black students and 28% of Latino students demonstrated proficiency, in stark comparison to 54% of White students and 74% of Asian students who demonstrated proficiency [9]. Furthermore, achievement gaps in California persist across geographic regions, with rural school students lagging behind suburban and urban peers [6]. These achievement gaps across schools and school districts also exist alongside *funding gaps* — local property values generate local property tax revenue and lead to different amounts of funding from local, state, and federal revenue sources. There is an ongoing debate in the social science and education research communities regarding the relationship between school funding, various socioeconomic factors, and school performance [15]. These complex relationships are worth further exploring, as they have the potential to inform the design of *evidence-based policy and community initiatives* to promote the advancement of a more equitable and effective K-12 education system in California and other regional education systems. *Every student deserves equitable access to a quality education to provide them with the foundation to reach their human potential* [12]. Since artificial intelligence (AI) techniques have the potential to help humans address societal problems by identifying patterns in high-dimensional data [13], we were motivated to leverage AI to identify school-related and district-related financial and socioeconomic factors predictive of school academic performance. These predictive factors could inform the design of future K-12 educational policies and community initiatives to promote school performance and narrow achievement gaps. The following sections provide an overview of the specific problem statement addressed by our project, as well as the goals and approaches that we leveraged to address the problem.

### 1.2 Problem Statement

Our project addresses the specific societal problem of *pervasive* and *persistent* achievement gaps that exist across K-12 schools in California, often correlated with *local characteristics* related to demographic, socioeconomic, and geographic factors and *school-related characteristics* such as school funding received.

### 1.3 Goals and Approaches

The **primary goal** of our project was to develop and leverage machine learning (ML) models to identify school-related and district-related financial and socioeconomic factors predictive of K-12 school performance. While we analyzed the relationship between demographic characteristics (e.g., race) and geographic characteristics (e.g., zip code, area code) and school performance, we chose to use funding and socioeconomic features in our models. Given an ongoing debate among social scientists and education policy researchers regarding the relationship between school funding and school performance at a national scale [15], our project contributes findings regarding the importance of funding features in predicting California's K-12 school performance. Our findings in this paper have the potential to inform the design of state education policies and initiatives to improve school performance.

Our approach to achieving these goals consisted of the following 5 steps:

1. Collecting and combining a rich dataset of K-12 school mathematics and reading proficiency data, school funding data, and census-level socioeconomic and geographic characteristics of all public school districts and schools in California.
2. Analyzing and identifying relationships between school mathematics and reading proficiency and different local and school-related characteristics.
3. Developing and experimenting with 9 different linear and nonlinear regression models to predict the percentage of students in each school who were proficient in mathematics and reading (these were our two labels which represented *school performance* in our project)
4. Using these ML regression models to identify important school-related and district-related financial and socioeconomic factors that were predictive of school mathematics and reading proficiency.
5. Discussing how the insights from our models can inform the design of K-12 education policies and community initiatives to improve and optimize school performance.

This paper is organized with the following structure: Section 2 of this paper describes the background and related works that informed the design of this project. Section 3 details the different data sources that we acquired and combined for our project. Section 4 provides key findings from our exploratory data analysis. Section 5 presents modeling results and discusses the implications of these results. Section 6 discusses future work and concludes the paper.

# 2 Background and Related Work

## 2.1 School Funding Process in the United States Education System

While European and Asian nations largely fund schools equally through centralized funding sources, the wealthiest 10% of school districts in the United States can spend up to 10 times more than the poorest 10% of school districts [3]. In the United States, school funding comes from a mix of federal, state, and local government contributions. Annual expenditures per student, adjusted for cost of living, vary greatly from state to state and from school district to school district. Furthermore, local property taxes are known to be one of the driving forces behind public school funding [7], leading to vast differences in funding available to schools.

While funding models vary from state to state, many states use the **foundation grant model** to help ensure that funding allocation is close to equal per student. In the foundation grant model, the state sets a minimum amount of funding for each student. If districts are unable to meet this minimum threshold, the state fills in the gap by providing districts with an amount equal to the set funding minimum minus the amount the district was able to afford per student. Once the school district has an allocated amount of funding per student of at least the set funding minimum (ie. at least $10,000 per student) districts are able to add a 1-1.4% property tax on properties within that district. Since local funding allocation is based on a property tax percentage and not a fixed amount, *local funding allocation can vary widely depending on property wealth in the school district*, *as visualized in California K-12 funding data from 2019-2020* in **Figure 1** [14]**.** The property tax percentage plays a significant role in the amount of funding received by each student, especially in wealthier districts. For example, a .4% increase in the property tax from 1% to 1.4% in a district with a property value average of $2 million per home district with 5,000 homes and 2,000 students results in a +$10,000 per student [4]. In larger cities, this increase can be even greater. Almost every state, including California, uses the foundation formula to allocate the appropriate resources for school districts [5].

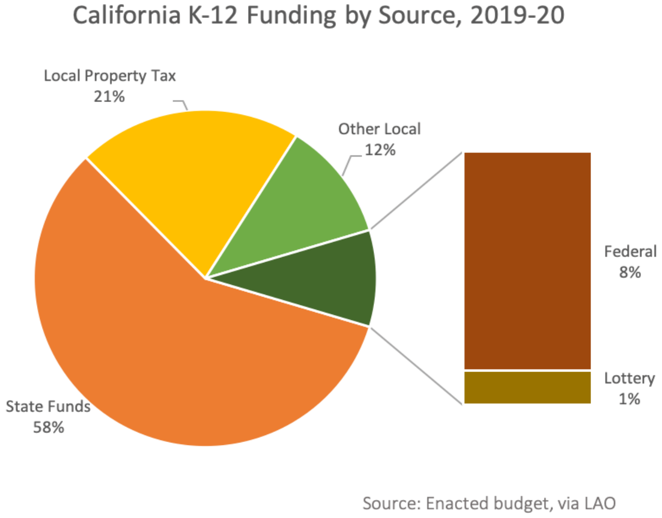


**Figure 1.** This figure visualizes the proportion of California K-12 public school funding received from various local, state, and federal sources in the 2019-2020 academic year [14]. ***Local property taxes* contributed, on average, *21%* of each school's total annual funding received**.

Despite the existence of discrepancies in school funding, the relationship between public school funding and school achievement is debated among social scientists and education policy researchers. One of the seminal works in this area is the *Coleman Report* from 1966 (cited almost 20,000 times), which was the first comprehensive study of public school performance across the United States. Coleman's report was the first to document the existence of an *achievement gap* between groups of students in the United States, in this case between White students and Black students, and the regression analysis in the report found that a student's (1) family background and (2) level of classroom socioeconomic diversity were the highest predictors of academic performance, while school funding amounts played a minimal role as a determinant of academic performance. Coleman's report sparked a debate that still exists in the social science and policy literature today about the extent to which school funding plays a role in predicting academic performance [8, 10]. As noted in [10], the amount of spending per student in the United States has quadrupled since 1960 (controlled for inflation), while academic performance has remained stagnant. Our project explores this debate regarding the relationship between school funding and achievement, and contributes computational findings regarding the extent to which school funding contributes to school performance in California K-12 schools, when considered alone and when used in models along with various school-related and district-related financial and socioeconomic factors.

### 2.2 Prior Machine Learning Approaches for Modeling School Academic Performance

Prior research has examined and identified disparities in academic achievement based on prior academic performance, extracurriculars [1], and environmental factors (e.g, parental education, health, creativity, etc) [2] in order to predict academic performance. A variety of classical machine learning models (e.g., decision trees,  support vector machines, etc) have been leveraged for academic performance predictions in various *micro-level* and *macro-level* contexts.

Existing research has focused on modeling school performance at a *micro level*, for individual students within schools. Shahiri et al. (2015) [1] used data on students in Malaysian schools to create machine learning models that predicted academic success based on past class assessment performance, student demographic, social network interaction, CPGA (cumulative grade point average), course structure, and extracurriculars. The classical machine learning models implemented by [1] included decision trees, ANNs, Naive Bayes, K-Nearest Neighbor, SVM, and a shallow neural network. Their research found that the neural network demonstrated the highest prediction accuracy at 98%, with CGPA being one of the strongest correlators for future success in class. A meta-analysis of 357 papers related to predicting student academic performance found that a variety of regression and classification models have been leveraged, trained primarily on a student's family, demographic, and social network features. Prior research has also focused on modeling school performance at a *macro level*, focused on predicting the performance of an entire school instead of individual students. For example, researchers in [2] leveraged logistic regression and light gradient boosting, trained on community-related features (e.g., availability of clean water, household goods) to predict the performance of South African high schools.

To the best of our knowledge, no prior research has developed computational models to predict the performance of K-12 schools in California based on school-level and district-level funding funding and socioeconomic factors. In addition, to the best of our knowledge, no prior research has developed models to allocate funding to optimize the academic performance of K-12 schools in California.

## 3 Datasets

This section provides an overview of the three data sources that our project leveraged: (1) annual school performance data from standardized tests administered by California's Department of Education, (2) annual school district funding data from the NCES, and (3) regional socioeconomic and geographic factors from Social Explorer. Due to constraints on available school district funding data (described in 3.2), we scraped data from multiple years but focused pre-processing and subsequent modeling work on the 2016-2017 academic school year (2017 fiscal year).

### 3.1 Acquiring California K-12 School Performance Data

Data on California K-12 school performance was scraped from the *Education Data Portal*, which maintains an API to access data from the National Center for Education Statistics' Common Core of Data (CCD). The performance data comes from the annual California Assessment of Student Performance and Progress (CAASPP) standardized test, which is administered in the areas of math and reading (which encompasses different language arts/verbal skills). The datasets obtained for the years 2014-2018 contain 20 characteristics for each K-12 school in California, encompassed by the following categories:

- School identification information (NCES ID, LEA ID, school name)
- School-level characteristics (e.g., enrollment)
- School assistance programs available (e.g., reduced price lunch, free lunch)
- School performance on CAASPP assessments, specifically each school's **percentage of students performing above state proficiency levels in mathematics and reading**. These two characteristics are used as the **labels** predicted in our modeling process and they serve as a proxy for school academic performance.

Out of the 10,406 schools in our 2016-2017 school performance dataset, we found that ~6% of the schools contained null or missing values for school performance labels. Preliminary pre-processing of our dataset included dropping all schools with null performance label values. This resulted in a usable dataset of **9,787** California K-12 schools with data from 2016-2017.

### 3.2 Acquiring California K-12 School District Funding Data

Data on California public pre-K-12 school funding was scraped from the National Center for Education Statistics (NCES). We wanted to use the most recent revenues and expenditures data possible, however, the most recently published data was from the 2016-2017 school year (fiscal year 2017). Therefore, we chose to use this time frame for our dataset, which includes the following key features:

1. TOTALREV: total revenue
2. TFEDREV: total federal revenue
3. TSTREV: total state revenue
4. TLOCREV: total local revenue
5. TOTALEXP: total expenditures
6. TCURELSC: total current spending for elementary-secondary programs
7. TCURINST: total current spending for instruction
8. TCURSSVC: total current spending for support services
9. TCAPOUT: total capital outlay expenditure
10. SCHLEV: school level code (01 = elementary school, 02 = secondary school, etc.)
11. Z34: Total Employee Benefits
12. Other school-related characteristics (e.g., LEA ID)

Each datapoint in our funding dataset represented a school district, resulting in a total of **1171 california school districts**. We note that this dataset includes the *LEA ID* for each school district, which corresponded with the *LEA ID*s of schools in other datasets used in this project.

### 3.3 Acquiring Socioeconomic and Geographic Factors

Data capturing various socioeconomic and geographic factors were scraped from Social Explorer, which maintains tables of location-based data. The yearly datasets obtained for years 2014-2018 were obtained from the following tables: American Community Survey (5-Year Estimates) (ACS), Income Limit and Fair Market Rent, and School Digger. All datasets contained data at the census tract level.

#### *3.3.1 ACS Dataset*

The ACS is a nationwide survey that provides social, economic, housing, and demographic data every year. As a 5-year estimate, it uses data collected over a period of 60 months to predict estimates for the next year. All monetary values were adjusted to the inflation rate of 2017 for consistency. Out of its **2201 characteristics**, **258 were manually selected** through extensive group discussion. Geographic information and tags such as FIPS codes, census tract, and school district were retained, as well as data regarding population, gender, household types, housing unit price and occupant information, education level of population, industry of employment, household income, rent, health insurance, and poverty level.

In addition, we computed the Concentrated Disadvantage Index (CDI) from information contained in the ACS dataset. The CDI was calculated using the following variables:

1. Percent of individuals below the poverty line

2. Percent of individuals on public assistance
3. Percent female headed households
4. Percent unemployed
5. Percent less than age 18

This computed CDI was included as another feature column within the ACS dataset, resulting in a **total of 259 characteristics for 5056 census tracts**.

### *3.3.2 Income Limit and Fair Market Rent*

The Income Limit and Fair Market Rent (FMR) is collected by the Department of Housing and Urban Development for each fiscal year, and this information is used to determine standard payment amounts of federal housing assistance programs. FMR is a gross rent estimate that includes shelter rent and all major utility costs, excluding that of telephone, cable or satellite television, and Internet service.

The following characteristics were scraped from the Income Limit and FMR dataset:

- Fair Market Rent for Efficiency
- Fair Market Rent for One Bedroom
- Fair Market Rent for Two Bedrooms
- Fair Market Rent for Three Bedrooms
- Fair Market Rent for Four Bedrooms

### *3.3.3 School Digger*

Additional data was collected from School Digger, which contains information regarding K-12 school information and performance data. The only dataset that could be accessed was the dataset for 2018. However, we ended up using the data after determining that the characteristics contained in this dataset were unlikely to change across years. Although most of this information was already contained in the files that we scraped from the Education Data Portal (described in Section 3.1), we still collected the following 28 characteristics across various categories as a useful additional, complementary file of information:

- School identification information (NCES ID, LEA ID, school name)
- School population by race
- School type (charter, private, magnet, virtual)
- Number of students receiving free or reduced-priced lunch

Scraping this data from California K-12 schools resulted in **28 characteristics across 2121 schools**. This data source was acquired and explored, but ultimately not included in the dataframe due to redundant features found in other sources.

### 3.4 Combining the School Performance, Funding, and Socioeconomic Factors Datasets

We combined the *school performance* and *school funding* datasets by using the *LEA (local educational agency) ID* contained in both datasets, resulting in a combined school dataset. The *CCD* dataset is based on Federal Information Processing Standards (FIPS) codes, so we converted every latitude and longitude in the combined school dataset to FIPS codes, mapped them to ncessch IDs also in the *CCD* dataset, and merged the FIPS codes with the combined school dataset using this mapping. Due to the FIPS codes extracted from the *CCD* dataset specifically being census tract-level FIPS codes, data from the *ACS* dataset was easily merged as the combined school dataset contained census tract FIPS codes. Lastly, the *housing* dataset contained county-level data, which included FIPS codes. The *housing* dataset was merged with the combined school dataset as the former's FIPS codes were substrings of the census tract-level FIPS codes found in the combined school dataset.

## 4 Exploratory Data Analysis

This section provides key findings from exploratory data analysis of different trends across our datasets, specifically related to (1) achievement gaps across students of different racial backgrounds, (2) relationships between household income and school funding, (3) relationships between local revenue amounts and school performance, and (4) the distribution of school math and reading proficiency levels across schools. All of the significance values reported in the sections below were computed through two-tail independent sample Welch's t-tests with equal variance not assumed.

## 4.1 Achievement Gap

We analyzed the average math and reading proficiency levels of school districts that had student populations with different majority racial backgrounds. As visualized in **Figure 2**, majority Asian schools districts were the only districts to exceed the state average proficiency levels in math and reading, outperforming majority White school districts and majority Black school districts**.** This affirms the **existence of achievement gaps** and motivates our research into how these gaps may persist across other factors, particularly socioeconomic factors.

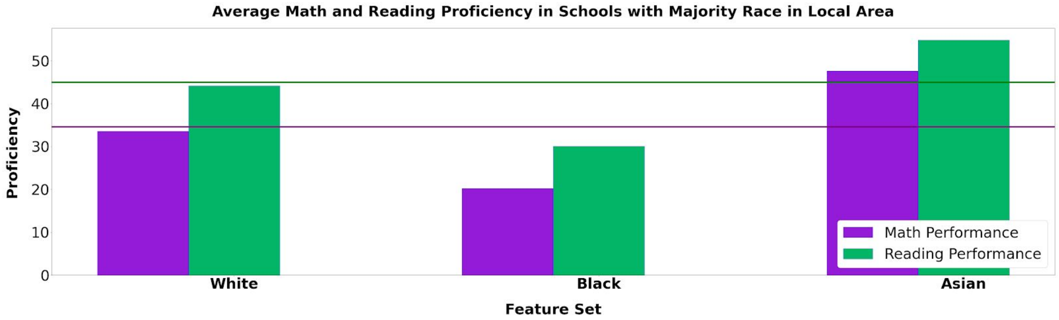


**Figure 2.** This figure visualizes average math and reading proficiency for schools in majority White, Black, and Asian school districts. We used 3 of the 5 racial categories specified by the United States Census; the other two groups (American Indian/Alaska Native and Native Hawaiian/Other Pacific Islander) did not have any school districts in California for which they were the majority.

## 4.2 Foundation Grant System

In order to confirm our understanding of the Foundation Grant System (Section 2.1) and its applicability to our California K-12 school funding data, we visualized the relationship between median household income (which can be viewed as a proxy for property tax) and federal, state, and local school funding, as can be seen in **Figure 3**. As the income quartile of a household increases, the median amount of local revenue received by a corresponding school district increases in a relationship that can be observed as close to linear. Pairwise t-tests comparing these differences in local revenue across different household income quartiles found these differences to be statistically significant ($\alpha = 0.05$, $p << 0.01$). This phenomenon can be explained by the foundation grant system, since areas with a higher median household income are likely to generate higher local revenue from higher property taxes. In contrast, as the income quartile of a household increases, the median amount of state and federal revenue received by a corresponding school district decreases in a relationship that can be observed as close to linear. This phenomenon can also be explained by the foundation grant system, since the state and federal revenue were likely distributed to enable a more equal distribution of total funding across income quartiles (which can be seen in the top left graph of **Figure 3**).

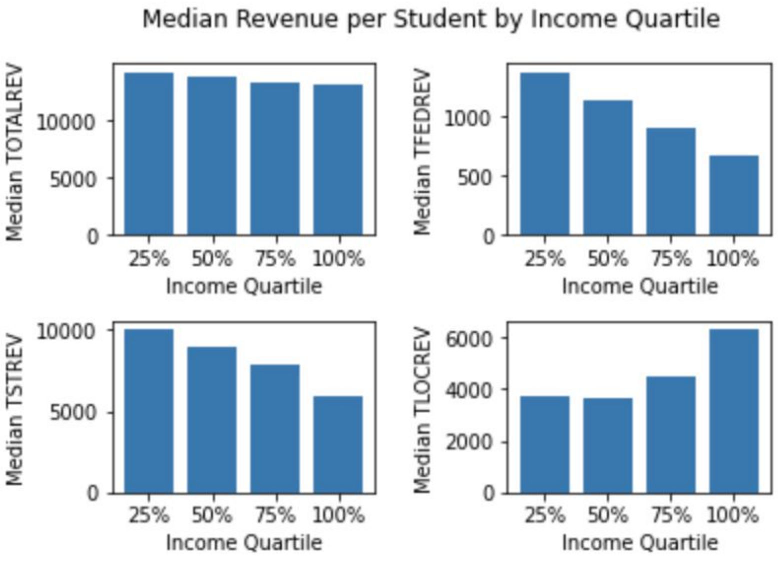


**Figure 3.** This figure visualizes total revenue (TOTALREV), federal revenue(TFEDREV), state revenue (TSTREV), and local revenue (TLOCREV) received by each student across different median household income quartiles.

### 4.3 Importance of Local Funding

Since we established that school districts with higher median household income (more socioeconomically privileged areas) receive more local funding, this section presents an analysis of the relationship between local funding and school performance. Schools in districts receiving higher amounts of local funding have higher median performance scores ($\alpha = 0.05$, $p \ll 0.01$) as seen in **Figure 4.** This relationship is only true for local revenue and not for state, federal, or total revenue, suggesting that the characteristics of a school district that enable it to receive more local revenue (e.g., higher property taxes, which can function as a proxy for socioeconomic privilege) can correlate with the performance of schools in that district.

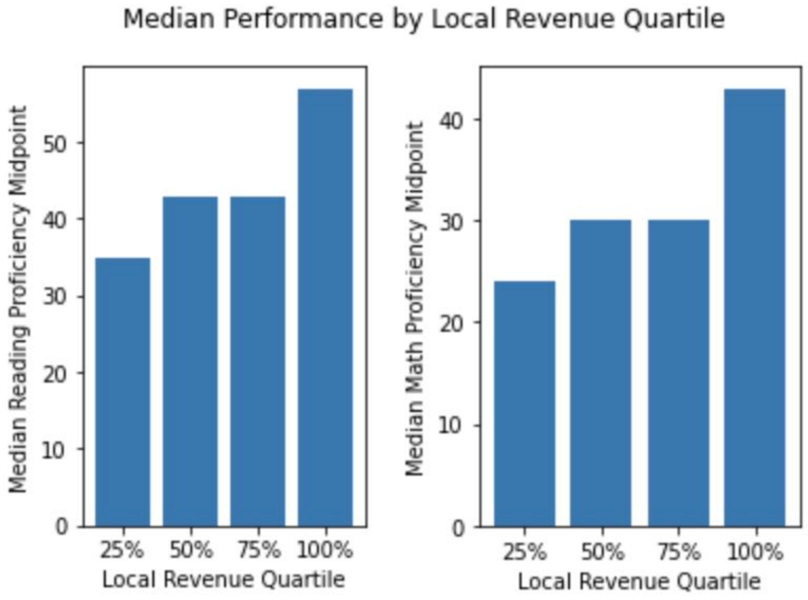


**Figure 4.** California K-12 school performance across local revenue quartiles from 2016-2017

## 4.4 School performance distribution

This section presents an overview of the distribution of each school’s average percentage of students proficient in math and reading, based on 2016-2017 standardized test (CAASPP) data.

The distribution of mathematics and reading proficiency levels across all California K-12 schools is visualized in **Figure 5**. The average percentage of students proficient in math across all K-12 California schools was **34.6%**, with a standard deviation of 20.8% and a median of 30.0%. The average percentage of students proficient in reading across all schools was higher than math, with an average proficiency of **45.0%,** a standard deviation of 21.0%, and a median of 43.0%.. It appears that math proficiency across schools is more *right-skewed* than reading proficiency.

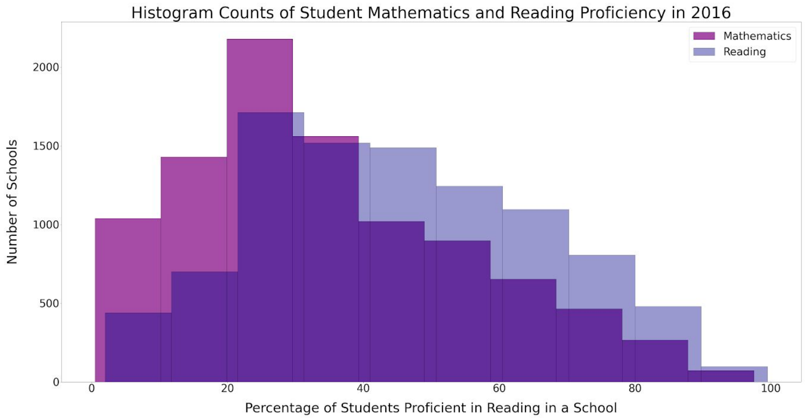


**Figure 5**. Distribution of math and reading proficiency for California K-12 Schools in 2016-2017, based on CAASPP test scores.

The distribution of math and reading proficiency scores differ significantly ($\alpha = 0.05$, $p << 0.01$) among primary schools, middle schools, and high schools, as visualized in **Figure 6.**

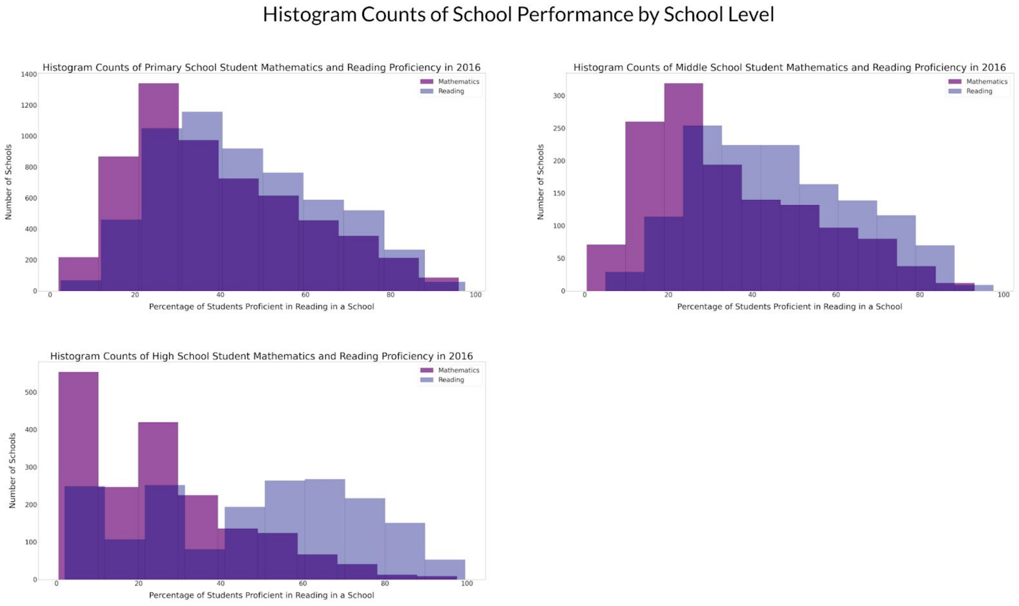


**Figure 6.** Distribution of math and reading proficiency across school level (primary, middle, high) of California K-12 schools in 2016-2017, based on CAASPP test scores

## 5 Modeling Methodology, Results, and Discussion

This section details the methodology for our regression models to predict California K-12 school performance, the modeling and feature importance results, and presents a discussion of results.

### 5.1 Pre-processing Data for Models

Before using our data in models, the following pre-processing steps were conducted on the entire combined dataframe of California K-12 schools, funding, and socioeconomic factors:

- Removed one school with an unexpected values (e.g., a field was "N" instead of numeric)
- Dropped feature columns that contained all NaN values
- Imputed missing values with the column mean if less than 5% of a column was missing (this only applied to continuous, numeric values)
-

### 5.2 Regression Modeling Methodology

We formulated the problem of predicting California K-12 school performance as a machine learning regression task to predict one of two labels: (1) the percentage of students in each school who were proficient in mathematics and (2) the percentage of students in each school who were proficient in reading, both based on standardized test performance data from the California Assessment of Student Performance and Progress in 2016-2017. We chose to develop **separate regression models for predicting mathematics and reading proficiency**, instead of a single regression model with the task of predicting two labels, due to the difference in the distributions of math and reading proficiency levels (Section 4.4). We also experimented with ***generalized models*** that were trained and tested on all schools to predict math and reading proficiency and ***individualized models*** that were trained and tested per school level on high-schools, middle-schools, and primary schools, respectively. This experimental design choice was motivated by the difference in label distributions across different types of schools (Section 4.4).

The general modeling process shared by all models is visualized in **Figure 7** below. **We conducted experiments with 9 different linear and non-linear regression models:** Linear Regression, Ridge Regression, Lasso Regression, ElasticNet Regression, DecisionTree Regressor, XGB Regressor, Support Vector Regression (SVM) with linear and rbf kernels, and a Multi-layer Perceptron (MLP) Regressor. All models were implemented with default scikit-learn hyperparameters, with the exception of the MLP Regressor (our hidden layer size was 10 instead of 100). All experiments were conducted with 5-fold cross-validation and the following primary metrics were computed and averaged across each cross-validation fold: (1) the *$R^2$ score*, which represents the proportion of variance in school proficiency that can be explained by the model's input features and (2) the *mean absolute error (MAE)*, which represents the average absolute value of the differences between the model's prediction of school proficiency and the actual proficiency. While we did also compute the *mean squared error* across each fold, the observed trends in mean absolute error and mean squared error were quite similar, so we chose to focus our paper on the two metrics of *$R^2$ score* and *mean absolute error*. After conducting experiments without feature selection, we implemented sequential forward selection to select the best 10 features; this method was implemented on the training sets of all cross-validation folds to prevent information from leaking into the testing sets. However, we found no substantial increase in model performance through this feature selection method, so all results reported in this paper have been obtained from models without any feature selection. Future work (detailed in Section 6) includes experimenting with other feature selection methods to determine if more effective approaches exist.

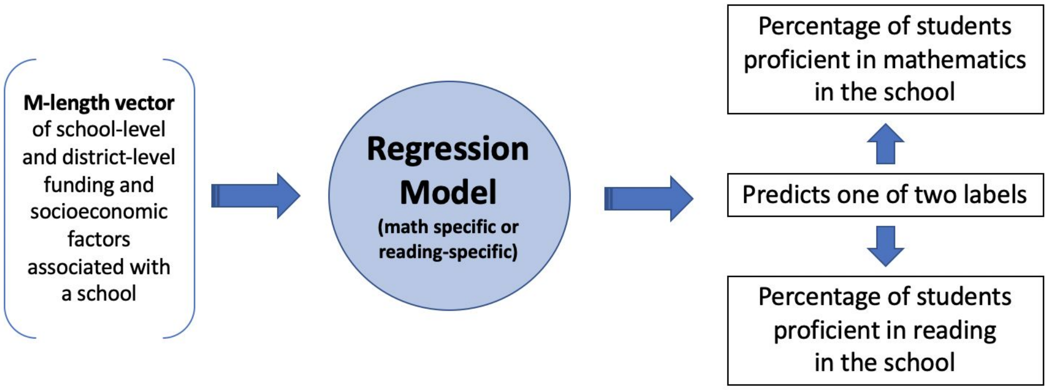
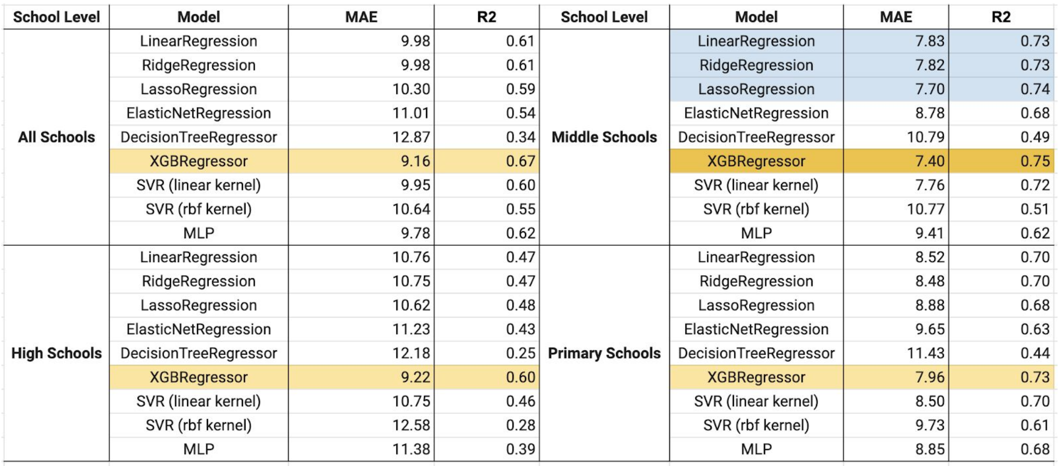

M-length vector of school-level and district-level funding and socioeconomic factors associated with a school
Regression Model
(math specific or reading-specific)
Predicts one of two labels
Percentage of students proficient in mathematics in the school
Percentage of students proficient in reading in the school

**Figure 7**: An illustration of the high-level process shared by each of our regression models to predict mathematics and reading proficiency levels of California K-12 schools in 2016-2017.

### 5.3 Mathematics Proficiency Modeling Results

Results from regression models predicting the mathematics proficiency of schools are presented in **Table 1** below and visualized in **Figure 8.** Across the generalized and individualized models, the **XGBoost Regressor** (performances highlighted in yellow in Table 1) outperformed all other models, indicating the effectiveness of gradient boosted trees as a modeling approach for integrating funding and socioeconomic factors to predict school mathematics performance in K-12 California schools. For the generalized model across all schools, XGBoost achieved an **MAE of 9.16** and **R$^2$ of 0.67**. Given that the mean and standard deviation of mathematics proficiency levels for all schools were 34.6% and 20.8%, a model with a MAE of 9.16 indicates the model's predictions could be useful estimates of a school's level of proficiency. For the individualized models predicting math proficiency across the three school levels, **XGBoost performed better for primary schools (MAE = 7.96, R$^2$ = 0.73) and middle schools (MAE = 7.46, R$^2$ = 0.75), versus high school (MAE = 9.22, R$^2$ = 0.60).** These results indicate the importance of modeling California K-12 school mathematics proficiency through individualized models per school level, versus combined generalized models. High-schools, specifically, appear to have mathematics proficiency with a variance that cannot be as well explained by the funding and socioeconomic features in our dataset, compared to middle schools and primary schools.



| School Level | Model | MAE | R2 | School Level | Model | MAE | R2 |
|---|---|---|---|---|---|---|---|
| All Schools | LinearRegression | 9.98 | 0.61 | Middle Schools | LinearRegression | 7.83 | 0.73 |
| | RidgeRegression | 9.98 | 0.61 | | RidgeRegression | 7.82 | 0.73 |
| | LassoRegression | 10.30 | 0.59 | | LassoRegression | 7.70 | 0.74 |
| | ElasticNetRegression | 11.01 | 0.54 | | ElasticNetRegression | 8.78 | 0.68 |
| | DecisionTreeRegressor | 12.87 | 0.34 | | DecisionTreeRegressor | 10.79 | 0.49 |
| | XGBRegressor | 9.16 | 0.67 | | XGBRegressor | 7.40 | 0.75 |
| | SVR (linear kernel) | 9.95 | 0.60 | | SVR (linear kernel) | 7.76 | 0.72 |
| | SVR (rbf kernel) | 10.64 | 0.55 | | SVR (rbf kernel) | 10.77 | 0.51 |
| | MLP | 9.78 | 0.62 | | MLP | 9.41 | 0.62 |
| High Schools | LinearRegression | 10.76 | 0.47 | Primary Schools | LinearRegression | 8.52 | 0.70 |
| | RidgeRegression | 10.75 | 0.47 | | RidgeRegression | 8.48 | 0.70 |
| | LassoRegression | 10.62 | 0.48 | | LassoRegression | 8.88 | 0.68 |
| | ElasticNetRegression | 11.23 | 0.43 | | ElasticNetRegression | 9.65 | 0.63 |
| | DecisionTreeRegressor | 12.18 | 0.25 | | DecisionTreeRegressor | 11.43 | 0.44 |
| | XGBRegressor | 9.22 | 0.60 | | XGBRegressor | 7.96 | 0.73 |
| | SVR (linear kernel) | 10.75 | 0.46 | | SVR (linear kernel) | 8.50 | 0.70 |
| | SVR (rbf kernel) | 12.58 | 0.28 | | SVR (rbf kernel) | 9.73 | 0.61 |
| | MLP | 11.38 | 0.39 | | MLP | 8.85 | 0.68 |

**Table 1**: This table illustrates the MAE and R$^2$ values for predicting math proficiency levels for California K-12 schools with generalized and individualized models per school level.

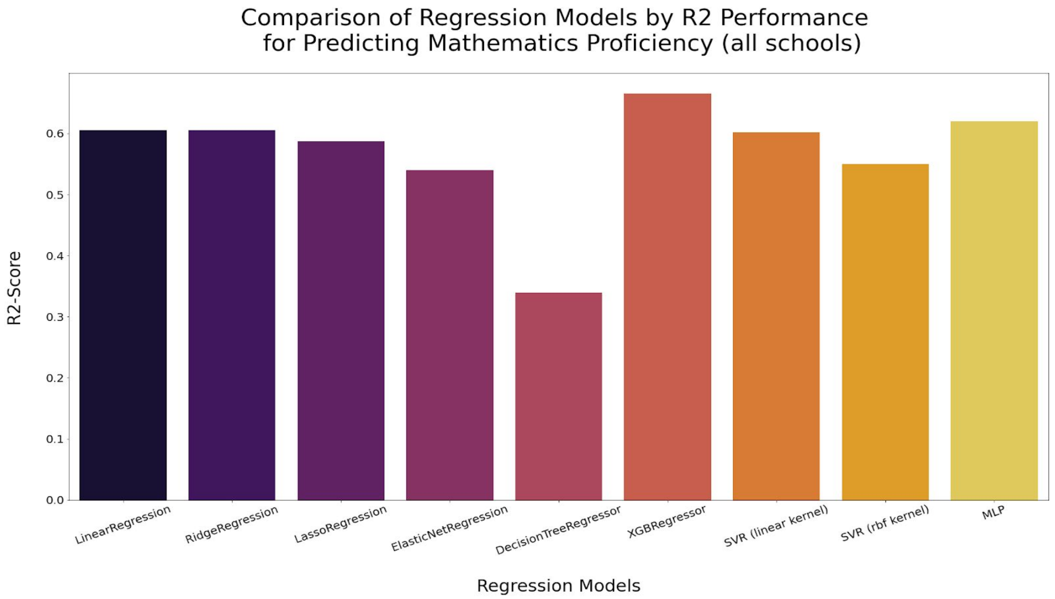
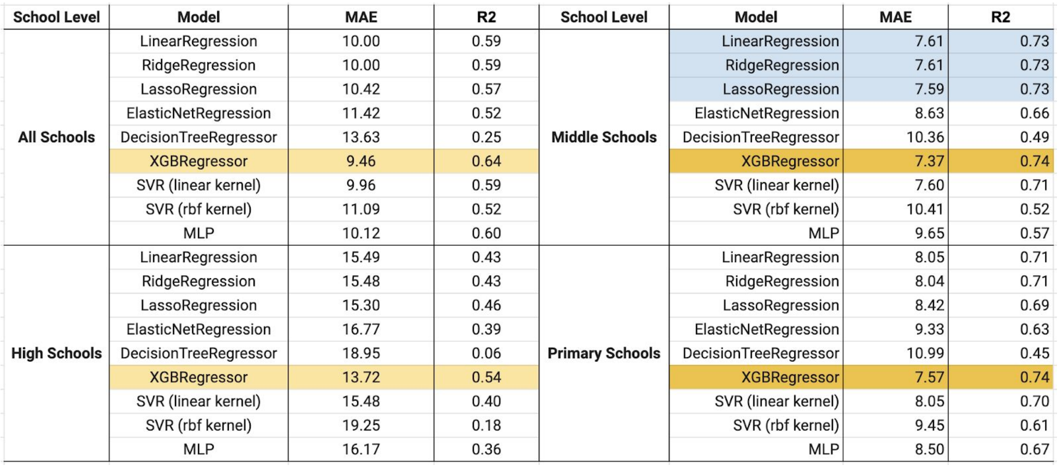


**Figure 8**. This graph visualizes differences in model performance ($R^2$ values) for predicting math proficiency levels for all California K-12 schools with generalized models.

### 5.4 Reading Proficiency Results

Results from regression models predicting the reading proficiency of schools are presented in **Table 2** below and visualized in **Figure 9.** Across the generalized and individualized models, the **XGBoost Regressor** (performances highlighted in yellow in Table 2) outperformed all other models, indicating the effectiveness of gradient boosted trees as a modeling approach for integrating funding and socioeconomic factors to predict school reading performance in K-12 California schools. For the generalized model across all schools, XGBoost achieved an **MAE of 9.46** and **$R^2$ of 0.64**. Given that the mean and standard deviation of reading proficiency levels for all schools were 45.0% and 21.0%, a model with a MAE of 9.46 indicates the model's predictions could be useful estimates of a school's level of proficiency. Similar to trends observed in modeling mathematics proficiency, **XGBoost performed better for primary schools (MAE = 7.57, $R^2$ = 0.74) and middle schools (MAE = 7.37, $R^2$ = 0.74), versus high school (MAE = 13.72, $R^2$ = 0.54).** These results indicate the importance of modeling California K-12 school reading proficiency through individualized models per school level, versus combined generalized models. Similar to our findings from the mathematics proficiency modeling, high schools appear to have reading proficiency with a variance that cannot be as well explained by the funding and socioeconomic features in our dataset, compared to middle schools and primary schools.



| School Level | Model | MAE | R2 | School Level | Model | MAE | R2 |
|---|---|---|---|---|---|---|---|
| All Schools | LinearRegression | 10.00 | 0.59 | Middle Schools | LinearRegression | 7.61 | 0.73 |
| | RidgeRegression | 10.00 | 0.59 | | RidgeRegression | 7.61 | 0.73 |
| | LassoRegression | 10.42 | 0.57 | | LassoRegression | 7.59 | 0.73 |
| | ElasticNetRegression | 11.42 | 0.52 | | ElasticNetRegression | 8.63 | 0.66 |
| | DecisionTreeRegressor | 13.63 | 0.25 | | DecisionTreeRegressor | 10.36 | 0.49 |
| | XGBRegressor | 9.46 | 0.64 | | XGBRegressor | 7.37 | 0.74 |
| | SVR (linear kernel) | 9.96 | 0.59 | | SVR (linear kernel) | 7.60 | 0.71 |
| | SVR (rbf kernel) | 11.09 | 0.52 | | SVR (rbf kernel) | 10.41 | 0.52 |
| | MLP | 10.12 | 0.60 | | MLP | 9.65 | 0.57 |
| High Schools | LinearRegression | 15.49 | 0.43 | Primary Schools | LinearRegression | 8.05 | 0.71 |
| | RidgeRegression | 15.48 | 0.43 | | RidgeRegression | 8.04 | 0.71 |
| | LassoRegression | 15.30 | 0.46 | | LassoRegression | 8.42 | 0.69 |
| | ElasticNetRegression | 16.77 | 0.39 | | ElasticNetRegression | 9.33 | 0.63 |
| | DecisionTreeRegressor | 18.95 | 0.06 | | DecisionTreeRegressor | 10.99 | 0.45 |
| | XGBRegressor | 13.72 | 0.54 | | XGBRegressor | 7.57 | 0.74 |
| | SVR (linear kernel) | 15.48 | 0.40 | | SVR (linear kernel) | 8.05 | 0.70 |
| | SVR (rbf kernel) | 19.25 | 0.18 | | SVR (rbf kernel) | 9.45 | 0.61 |
| | MLP | 16.17 | 0.36 | | MLP | 8.50 | 0.67 |

**Table 2**: This table illustrates the MAE and R$^2$ values for predicting reading proficiency levels for California K-12 schools with generalized and individualized models per school level.

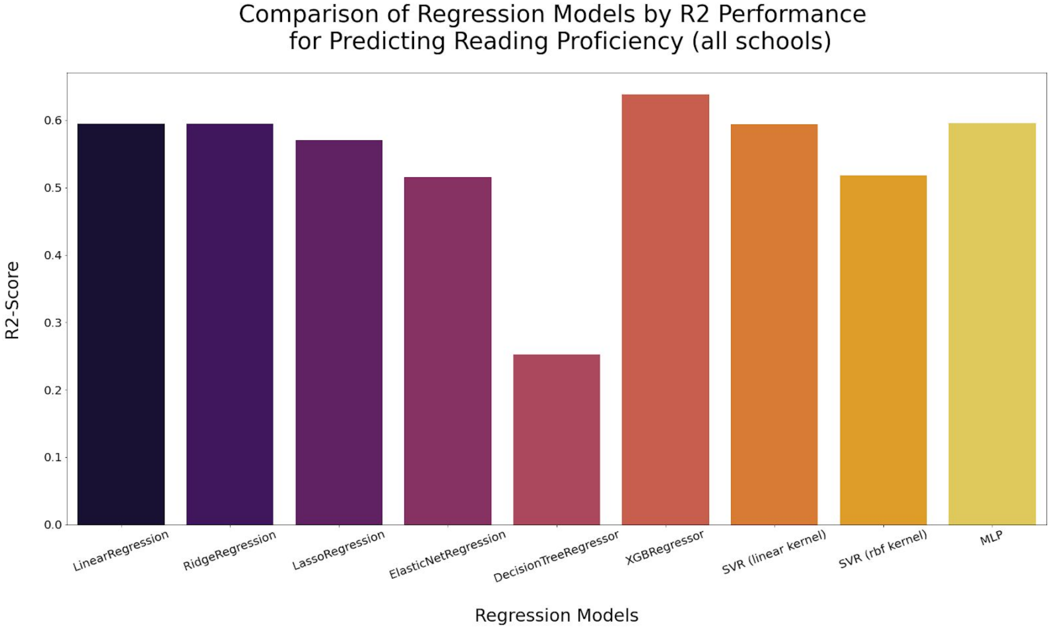


**Figure 9**. This graph visualizes the MAE and R2 values for predicting reading proficiency for all schools across California K-12 schools in 2016-2017, based on CAASPP test scores.

### 5.5 Ablation Analysis

To better determine the contributions of different types of features, we conducted an ablation analysis with an XGBoost model trained on the the following feature sets: *funding features*, *school-related features* (the proportion of students receiving free or reduced price lunch), *housing features* (e.g, features related to the housing environment of the community, such as the prevalence of houses with married adults raising children), and *population features* (e.g., features related to the people in students' community such as the local careers of adults). Results from this ablation study are visualized in **Figure 10**.

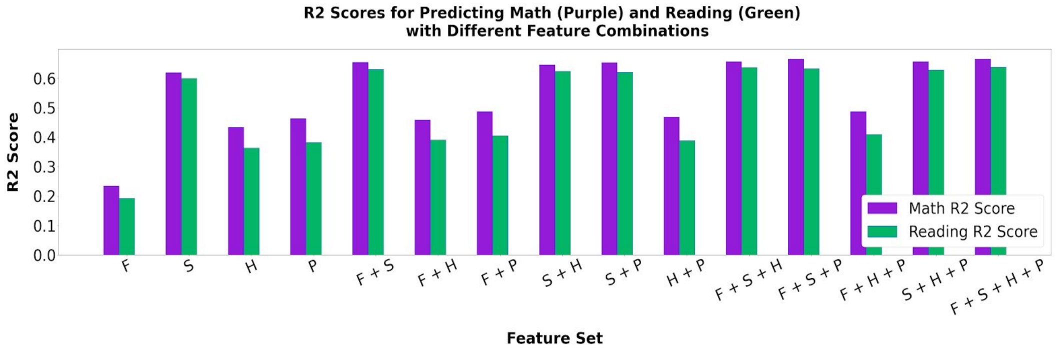


**Figure 10**. This graph visualizes the R$^2$ values for predicting mathematics and reading proficiency with different feature combinations for all K-12 schools across California in 2016-2017, with proficiency determined by CAASPP test scores. "**F**" = funding features, "**S**" = school-related features, "**H**" = housing features, and "**P**" = population features.

Our findings indicate the **importance of school-related features, specifically the proportion of students in a school receiving free or reduced price lunch, for predicting school mathematics and reading performance**. An XGBoost model trained only on the proportion of students receiving free or reduced price lunch achieved a MAE of 9.83 and R$^2$ of 0.62 for predicting mathematics proficiency and a MAE of 10.00 and R$^2$ of 0.60 for predicting reading proficiency in generalized models across all schools. These model performances are comparable to the models that leveraged all feature sets together (Sections 5.2 and 5.3). **This pattern was more prominently observed in the predictive power of these two features for individualized models specific to middle schools and primary schools.** For predicting middle school proficiencies, an XGBoost model trained only on these features achieved an MAE of 7.75 and R$^2$ of **0.73** for predicting mathematics proficiency and a MAE of 7.58 and R$^2$ of **0.73** for predicting reading proficiency. For predicting primary school proficiencies, an XGBoost model trained only on these two school-related features achieved an MAE of 8.70 and R$^2$ of 069 for predicting mathematics proficiency and a MAE of 8.20 and R$^2$ of 0.70 for predicting reading proficiency. **High-schools did not exhibit this predictive relationship,** and an individualized XGBoost model for high schools trained on these features achieved an MAE of 9.43 and R$^2$ of 0.58 for predicting mathematics proficiency and a MAE of 13.92 and R$^2$ of 0.52 for predicting reading proficiency.

Our findings also indicate the **importance of combining funding with socioeconomic factors in models predictive of school performance**. Funding, alone, in this ablation study achieved an R$^2$ of less than 0.30 for both mathematics and reading, but funding contributes to the highest-performing model, along with the other school-related, housing, and population features.

To better visualize the distribution of these key school-related features, we developed a graph with the feature distribution of the proportion of students receiving free or reduced price lunch across each census tract in California, which can be seen in **Figure 11.** As seen in the graph [insert a few sentences analyzing why the graph is interesting]

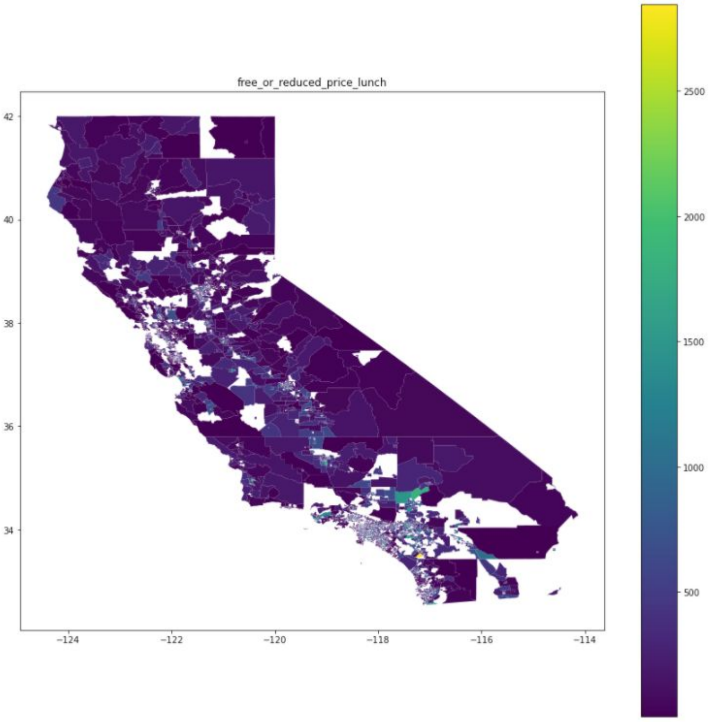


**Figure 11.** This chart visualizes the number of students with free or reduced price lunch for each census tract in K-12 schools across California in 2016-2017. Empty or white census tracts did not have data for free or reduced price lunch.

### 5.6 Feature Importance in XGBoost Regressor

In order to identify specific features within the highest-performing XGBoost model that were important contributors to model performance, we conducted permutation feature importance tests. **Table 3** lists the top 5 most important features for predicting mathematics and reading performance across each of the generalized and individualized models across school types. .

| **School Level** | **Mathematics** | **Reading** |
|---|---|---|
| **All Schools** | ● Proportion of students receiving free lunch<br>● Proportion of students w/ reduced-price lunch<br>● Foreign-born population<br>● Total funding revenue per student<br>● Land area | ● Proportion of students receiving free lunch<br>● Proportion of students w/ reduced-price lunch<br>● Total funding revenue per student<br>● Fair market rent for efficiencies<br>● Total non-elementary school/secondary school expenditures per student |
| **Primary Schools** | ● Proportion of students receiving free lunch<br>● Fair market rent for efficiencies<br>● Proportion of students w/ | ● Proportion of students receiving free lunch<br>● Foreign-born population<br>● Proportion of students w/ reduced-price lunch |

| | | |
|---|---|---|
| | reduced-price lunch<br>● Proportion of foreign-born people in area<br>● Total capital outlay expenditures | ● Population 25+ with a Bachelor's degree<br>● Total capital outlay expenditures |
| **Middle Schools** | ● Proportion of students receiving free lunch<br>● Foreign-born population<br>● Proportion of students w/ reduced-price lunch<br>● Employed population 16+ doing maintenance work<br>● Population 25+ years who are high-school graduates | ● Proportion of students receiving free lunch<br>● Proportion of students w/ reduced-price lunch<br>● Employed population 16+ doing maintenance work<br>● Foreign-born population<br>● Employed population 16+ doing public administration work |
| **High Schools** | ● Proportion of students receiving free lunch<br>● Proportion of students w/ reduced-price lunch<br>● Foreign-born population<br>● Employed population 16+ doing maintenance work<br>● Total funding revenue per student | ● Proportion of students receiving free lunch<br>● Proportion of students w/ reduced-price lunch<br>● Total funding revenue per student<br>● Land area<br>● Total revenue from state sources |

**Table 3**: This table illustrates the top 5 most important features, per permutation feature importance, for predicting mathematics and reading performance across each of the generalized and individualized XGBoost models. The school-related features for proportion of students on free and reduced price lunch are colored teal and funding features are colored purple.

Permutation feature importance **further supports the significance of the proportion of students receiving free lunch and proportion of students receiving reduced-price lunch** as features predictive of school performance, as these features appear within the top 5 features of each of the generalized and individualized XGBoost models across all schools and all school levels. **Different funding features emerged as comparatively more important across different types of schools**.

**Across all schools collectively**, the *total funding revenue per student* was important for predicting math and reading performance, and the *total non-elementary school/secondary school expenditures* (e.g., community service and adult education) was key to predicting reading performance. These two features could serve as a proxy for how supportive the school district community is, since students who receive more funding and live in areas that invest in more community initiatives are likely to have the personal, social, economic, and cultural capital they need to perform well in school [20, 21].

**For primary schools**, the *total capital outlay expenditures* feature was important for predicting both the math and reading performance of primary schools. These expenditures encompass anything related to construction, land and existing structures, or instructional equipment at the school. It is possible that these types of infrastructure (e.g., larger playgrounds requiring more capital outlay expenditures to maintain) are more important for fostering the development and support of children at the primary school age.

**For middle schools**, no funding features emerged as important within the top 5.

**For high schools**, the *total funding revenue per student* was important for predicting math and reading performance, and the *total revenue from state sources* was important for predicting reading performance. Other important features that emerge across all models were various population features which can be seen as indicative of the community's level of social and cultural capital (e.g, the proportion of adults with high-school and college degrees). These types of population features can influence students' experiences and development at the micro-level (e.g., the amount of one-on-one encouragement they receive from adults about pursuing education) and at the macro-level (e.g., norms enforced by the community), which will impact their academic performance [17, 18, 19].

## 6 Conclusions and Future Extensions

Our project developed and leveraged machine learning regression models to identify school-related and district-related financial and socioeconomic factors predictive of California K-12 school performance in mathematics and reading. To the best of our knowledge, this project represents the first attempt at developing computational models to predict the performance of K-12 schools in California based on these factors. Our project *contributed modeling insights* related to this novel context, specifically (1) the demonstrated effectiveness of gradient boosted trees as a regression modeling approach predicting California K-12 school performance and (2) the demonstrated advantage of developing *individualized* models that are trained and tested by school level for high-schools, middle-schools, and primary schools, instead of *generalized* models that apply to all schools collectively. Our project also *identified key features* that emerged as predictive of school performance, and this knowledge can be leveraged in the design of education policies and community interventions to promote school performance.

Future work also includes framing and solving an optimization problem to determine the best census tract districts to place community interventions to maximize mathematics and reading performance across all of California's K-12 schools. This problem could be formulated through a *mixed integer linear program* that would likely be a variant of the *minimum set covering problem* to place community interventions in census tracts, subject to a budget constraint that would be the amount of money the state government could afford to spend on these interventions. In addition, since our project identified key funding features that were important predictors of school performance, it would also be interesting to extend this project by conducting simulation funding experiments with our models (e.g., experimenting with different amounts of artificially-set funding features in our models and observing how school performance predictions change) and developing an optimization problem to allocate specific types of funding (e.g., capital outlay expenditures) to specific types of schools (e.g., middle schools) to optimize math and reading performance.

To disseminate our findings, we recently presented at the Fall 2020 Projects Showcase of USC's Center for Artificial Intelligence in Society's Student Branch on November 14, 2020. Furthermore, we have also created a website to host our team's progress on the project. In the future, we hope to develop and embed interactive maps on this website using Mapbox's API, to illustrate the distribution of data in California for each of the key features that we collected. For example, users would be able to see the distribution of students on free or reduced price lunch by census tract (like **Figure 11)** and develop a more intuitive understanding of the various financial, geographic, demographic and socioeconomic factors that contribute to or correlate with California K-12 school performance.

### References

[1] Shahiri, Amirah Mohamed, et al. "A Review on Predicting Student's Performance Using Data Mining Techniques." *Procedia Computer Science*, vol. 72, 2015, pp. 414–422., doi:10.1016/j.procs.2015.12.157.

[2] H. Wandera, V. Marivate and M. D. Sengeh, "Predicting National School Performance for Policy Making in South Africa," 2019 6th International Conference on Soft Computing & Machine Intelligence (ISCMI), Johannesburg, South Africa, 2019, pp. 23-28, doi: 10.1109/ISCMI47871.2019.9004323.

[3] Smedley BD, Stith AY, Colburn L, et al.; Institute of Medicine (US). The Right Thing to Do, The Smart Thing to Do: Enhancing Diversity in the Health Professions: Summary of the Symposium on Diversity in Health Professions in Honor of Herbert W.Nickens, M.D.. Washington (DC): National Academies Press (US); 2001. Inequality in Teaching and Schooling: How Opportunity Is Rationed to Students of Color in America. Available from: https://www.ncbi.nlm.nih.gov/books/NBK223640/

[4] “How Do School Funding Formulas Work?” *Apps.urban.org*, 29 Nov. 2017, apps.urban.org/features/funding-formulas.

[5] “K-12 Funding: Funding Mechanism.” Caspio, Education Commission of the States, Aug. 2019, c0arw235.caspio.com/dp/b7f930000c8896077c5e400c99cb.

[6] Loeb, Susanna, et al. “Home: Getting Down to Facts II.” *Home | Getting Down to Facts II*, Stanford University, Policy Analysis for California Education (PACE), Sept. 2018, gettingdowntofacts.com/.

[7] Biddle, Bruce J., and David C. Berliner. “A Research Synthesis / Unequal School Funding in the United States.” *Unequal School Funding in the United States - Educational Leadership*, Association for Supervision and Curriculum Development, May 2002, www.ascd.org/publications/educational-leadership/may02/vol59/num08/unequal-school-funding-in-the-united-states.aspx.

[8] Greenwald, Rob, et al. “The Effect of School Resources on Student Achievement.” Review of Educational Research, vol. 66, no. 3, 1996, pp. 361–396. JSTOR, www.jstor.org/stable/1170528. Accessed 26 Oct. 2020.

[9] Cano, Ricardo. “California Kids' Test Scores Again Rise by Inches as Achievement Gap Yawns.” *CalMatters*, 11 Oct. 2019, calmatters.org/education/k-12-education/2019/10/california-schools-test-scores-2019-achievement-gap-caaspp-smarter-balanced/.

[10] Hanushek, Eric A. “The Impact of Differential Expenditures on School Performance.” *Educational Researcher*, vol. 18, no. 4, May 1989, pp. 45–62., doi:10.3102/0013189x018004045.

[11] Arto Hellas, Petri Ihantola, Andrew Petersen, Vangel V. Ajanovski, Mirela Gutica, Timo Hynninen, Antti Knutas, Juho Leinonen, Chris Messom, and Soohyun Nam Liao. 2018. Predicting Academic Performance: A Systematic Literature Review. In Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE ’18 Companion), July 2–4, 2018, Larnaca, Cyprus. ACM, New York, NY, USA, 25 pages. https://doi.org/10.1145/3293881.3295783

[12] *Transforming Our World: The 2030 Agenda for Sustainable Development*. Report from the United Nations Sustainable Development Agenda. Established in 2015, *Link*.

[13] Hager G. D, Drobnis A, Fang F, Ghani R, Greenwald A, Lyons T, Parkes D. C, Schultz J, Saria S, Smith S. F, Tambe M, Artificial Intelligence for Social Good, Workshop Report, Computing Community Consortium, Washington, 2017.

[14] California State Enacted Budget 2019-2020. *K-12 Funding by Source*. Publicly-available official state funding data from Legislative Analyst’s Office, 2020 https://lao.ca.gov/Education/EdBudget/Details/210.

[15] C. Kirabo Jackson & Rucker C. Johnson & Claudia Persico, 2016. "The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms," *The Quarterly Journal of Economics*, vol 131(1), pages 157-218.

[16] Coleman, J. S., United States., & National Center for Education Statistics. (1966). *Equality of educational opportunity*.

[17] Becker, Bronwyn E, and Suniya S Luthar. “Social-Emotional Factors Affecting Achievement Outcomes Among Disadvantaged Students: Closing the Achievement Gap.” Educational psychologist vol. 37,4 (2002): 197-214. doi:10.1207/S15326985EP3704_1

[18] Bronfenbrenner U. “The ecology of human development: Experiments by nature and design.” Cambridge, MA: Harvard University Press; 1979.

[19] Cicchetti D, Toth SL. Transactional ecological systems in developmental psychopathology. In: Luthar SS, Burack JA, Cicchetti D, Weiz JR, editors. *Developmental psychopathology: Perspectives on adjustment, risk, and disorder*. New York: Cambridge University Press; 1997. pp. 317–349.

[20] Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology*,

94, S95-120.

[21] Lareau, A. (1987). Social class differences in family-school relationships: The importance of cultural capital. *Sociology of Education*, 60(2), 73-85.