Predicting Test Performance in California Public Schools

Gabrielle Wald
Springboard School of Data
July 7th, 2021

Problem

There is a strong need to find more informed and nuanced factors that impact student performance.

1

HIGH POVERTY AREAS DISPROPORTIONALLY EDUCATE CHILDREN OF COLOR.

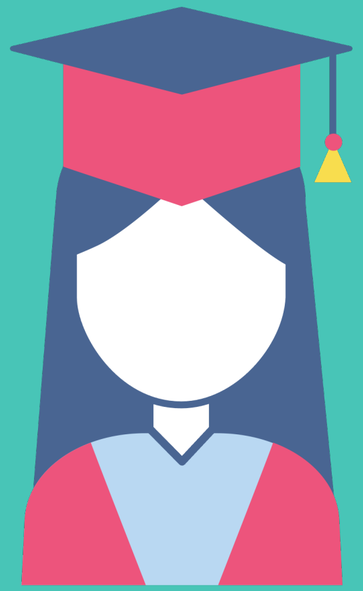
2

IT COSTS MORE TO EDUCATE LOW-INCOME STUDENTS.

GOALS

1. Understand the current demographics of K-12 public schools, learn what factors are most correlated with **student performance (pass rate).**

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



BENEFICIARIES
Who might care?

1

At government level:
- Policy makers
- Department of education
- Federal government

2

At school level:
- School administrators
- Teachers
- Parents

Data Understanding and Cleaning

- Data Acquisition
- Data Manipulation
- Data Cleaning

What factors may affect student performance?

California state, K-12 Public Schools, academic year 2018-2019.

1

DATA SOURCES:

- California Assessment of Student and Progress (CAASPP)
- California Department of Education (CDE)
- National Center of Education Statistics (NCES)

2

TARGET & FEATURES:

- **Target**: Percentage of students passing test standards per school.
- **Features**: Demographic data (ethnicity, sex, parent education, English fluency status, socioeconomic status), revenue and expenditure per pupil per school.

Data Manipulation

1. CAASPP assessment file was merged with 2 other files (Entity table, and Subgroup ID table).
2. Original dataset contained 3,013,079 rows and 32 columns.
3. Data is extracted at school level.

Data Merge

1. Demographic info was remapped using functions.
2. One observation per row and one variable per column.
3. The assessment file is merged with other files via district code, district name or zip code.

```
<class 'pandas.core.frame.DataFrame'>
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)
```

The result is a dataset with 10,435 rows and 39 columns

Data Cleaning

1. Imputed demographic missing values with zeroes.
2. Monetary variables were imputed with the median.
3. Columns with number of students turned percentage.
4. Added dummy columns to distinguish original vs imputed data to reduce bias

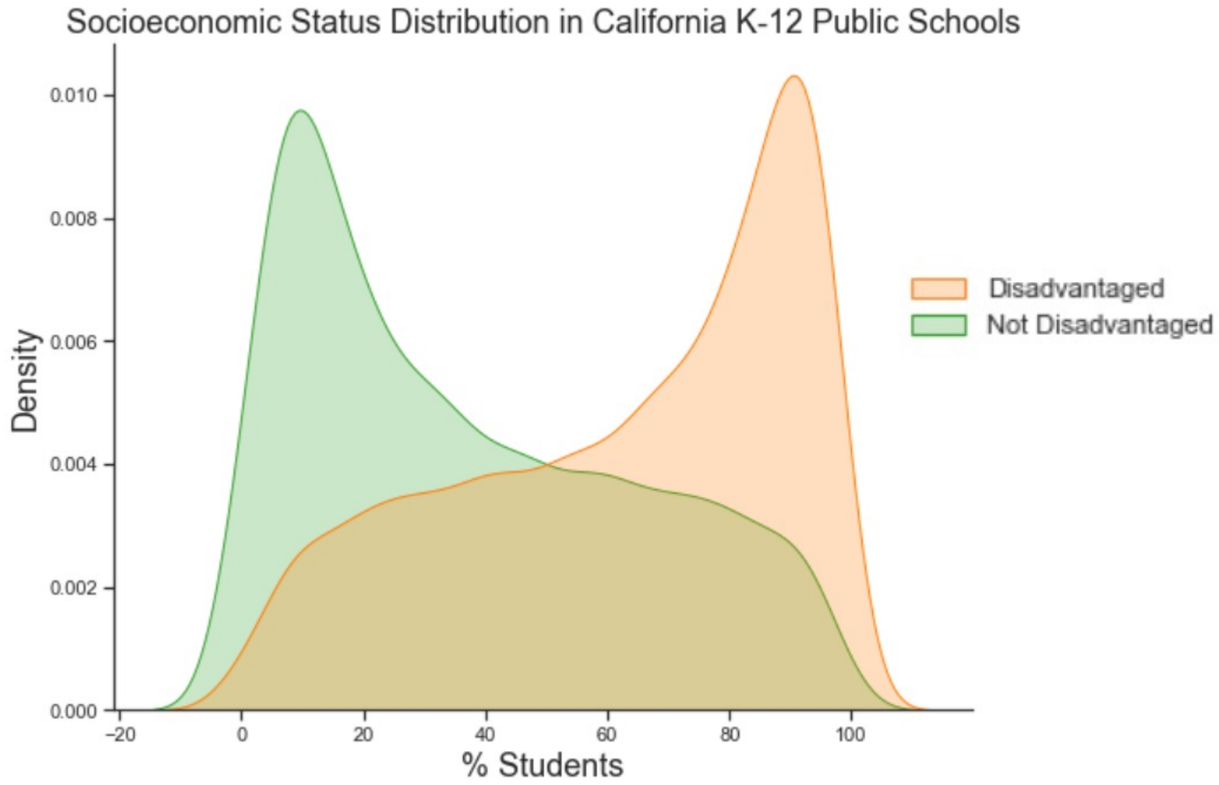
Exploratory Data Analysis

Libraries:

1. **Matplotlib**
2. **Seaborn**
3. **Pandas**

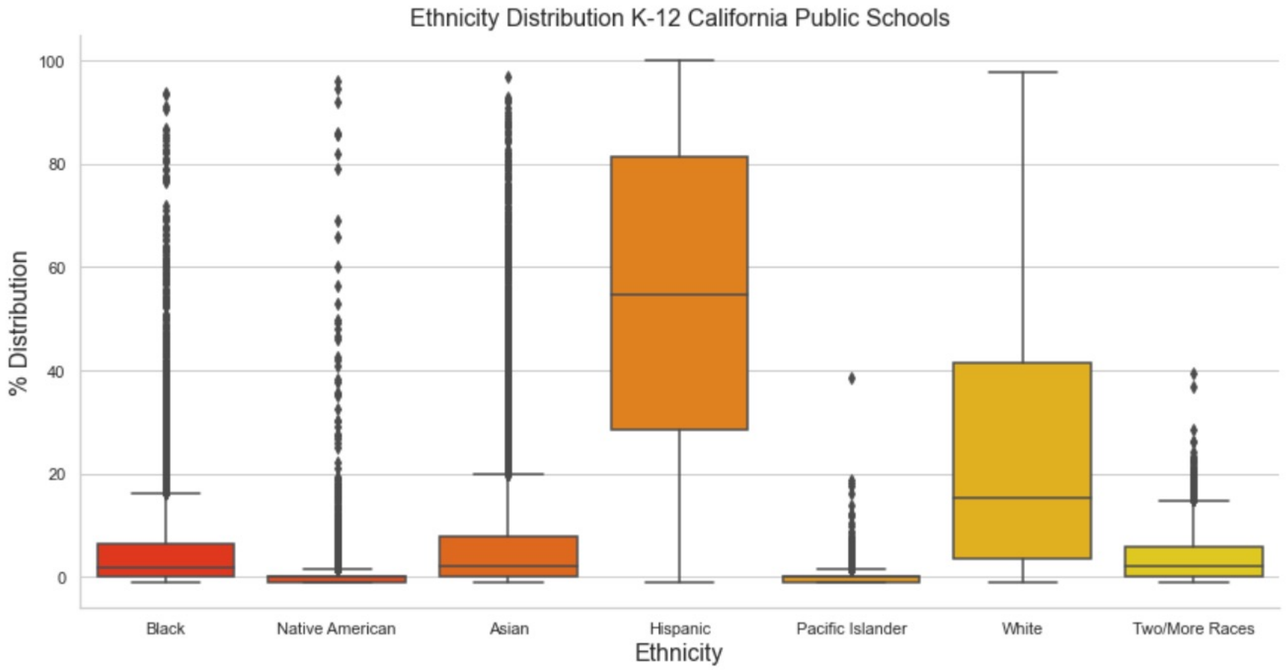
The data is explored to find trends, insights, and potential outliers based on visualization and hypothesis testing.

Higher proportion of economically ‘disadvantaged’ students across K-12 public schools.



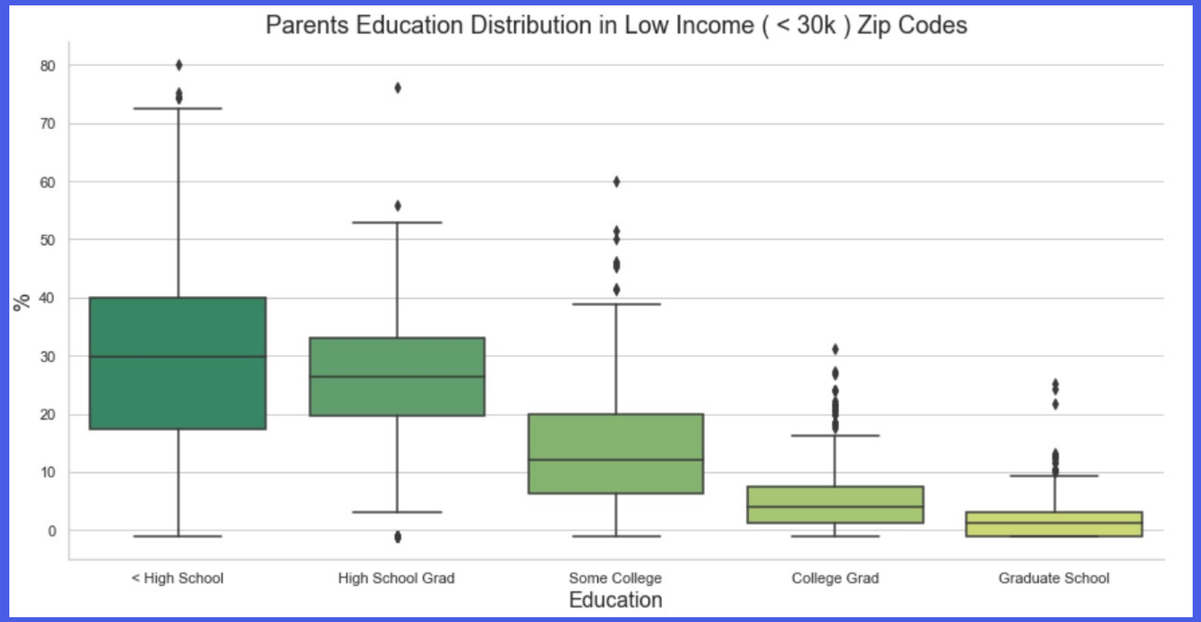
California is a minority-majority state.

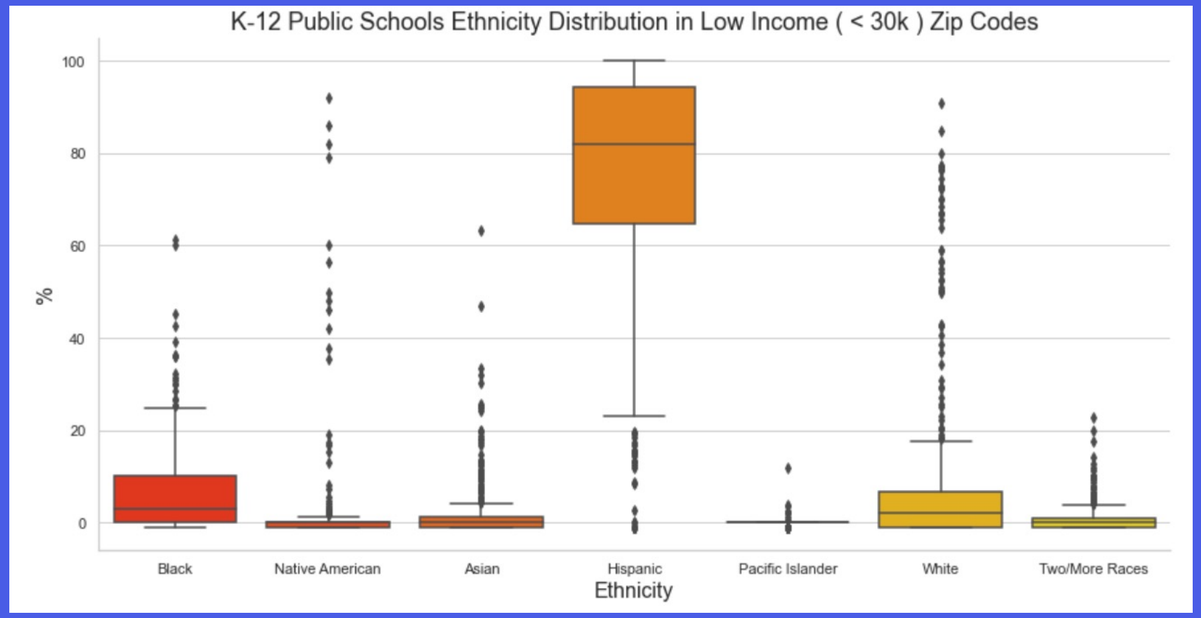
MORE THAN 50% OF STUDENTS SELF-IDENTIFY AS HISPANIC.



Income, Education & Ethnicity

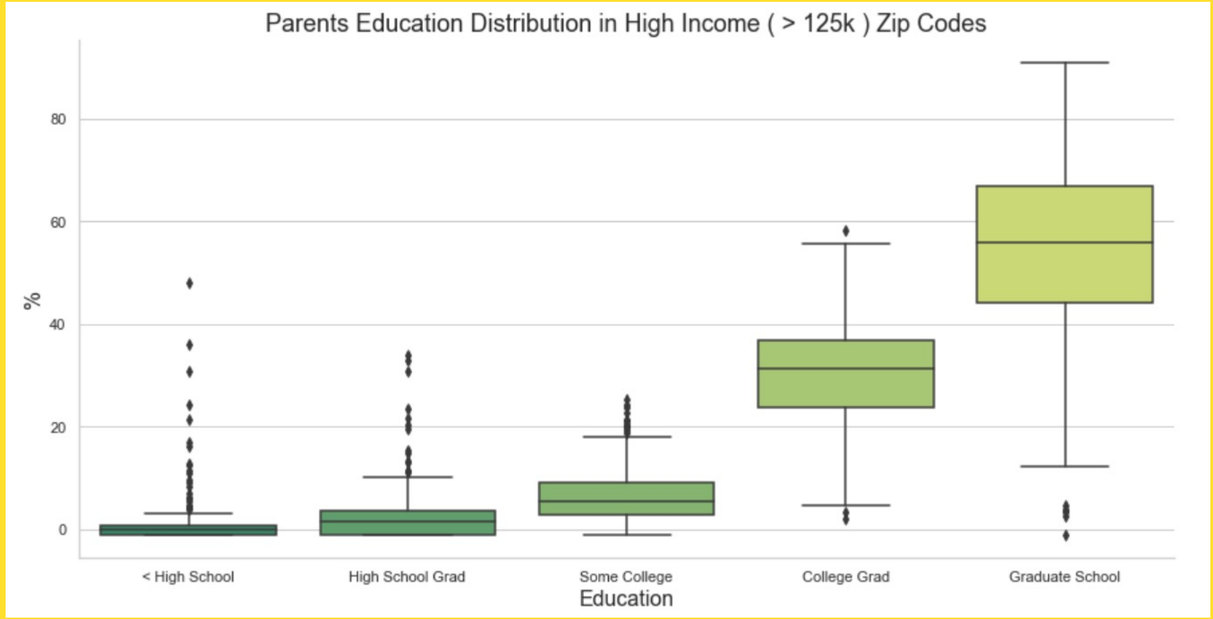
- Correlations and Graphs
- Hypothesis Testing

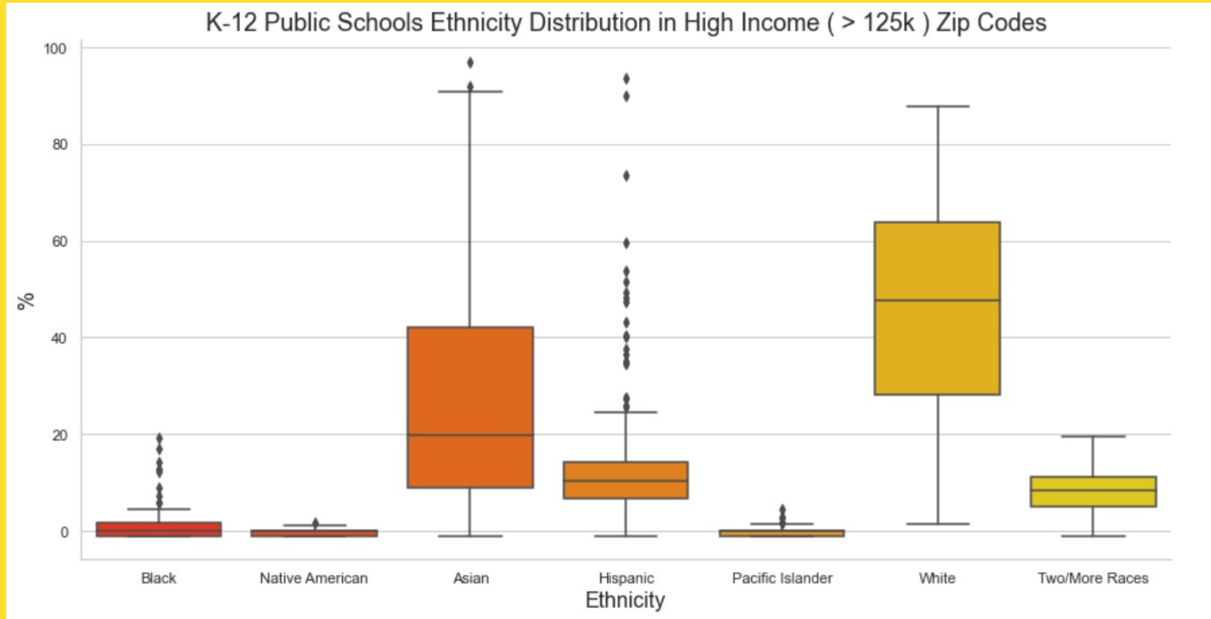




Parent education is correlated with median household income.

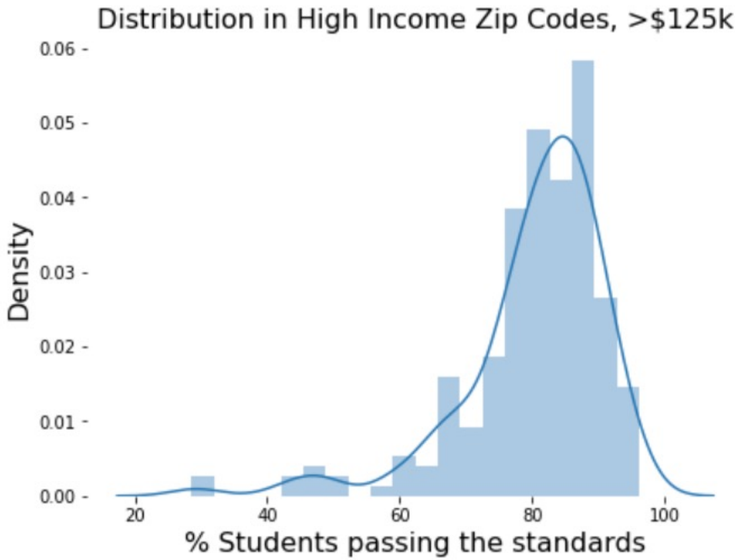
1. Lower percentage of students passing the standards (right skewed distribution) in low-income zip codes.
2. Less than high school is the most common school level.
3. Hispanics are majority.

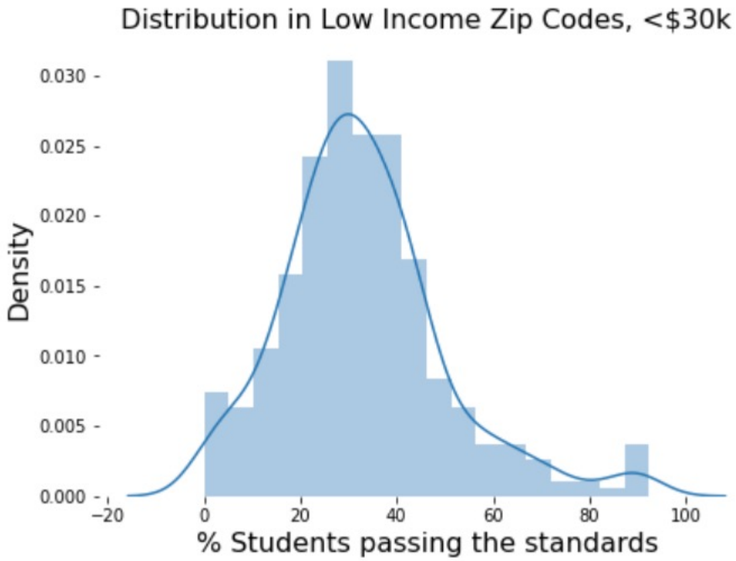




Income is correlated with student passing rate.

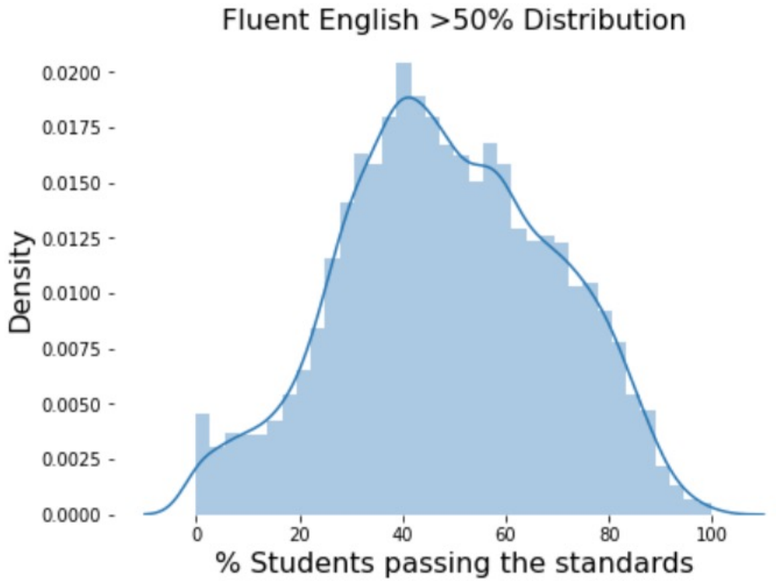
1. Higher percentage of students passing the standards (left skewed distribution) in high income zip codes.
2. Graduate school is the most common school level.
3. Asians and Whites are majority.

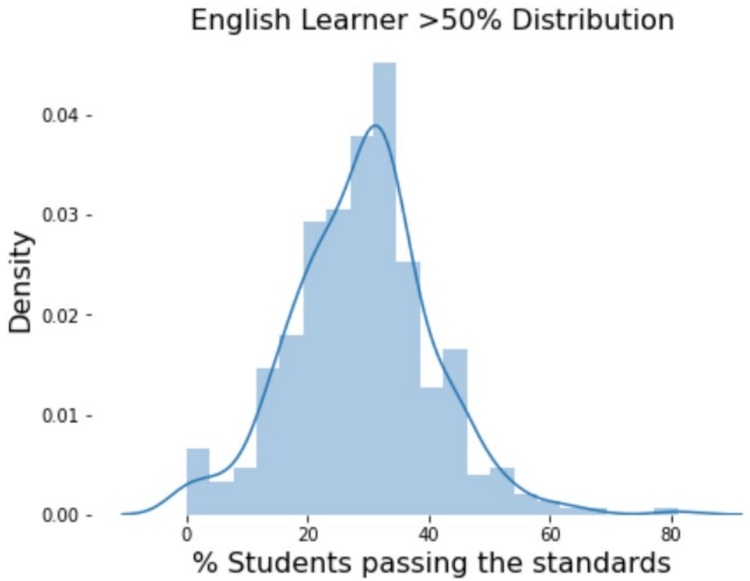


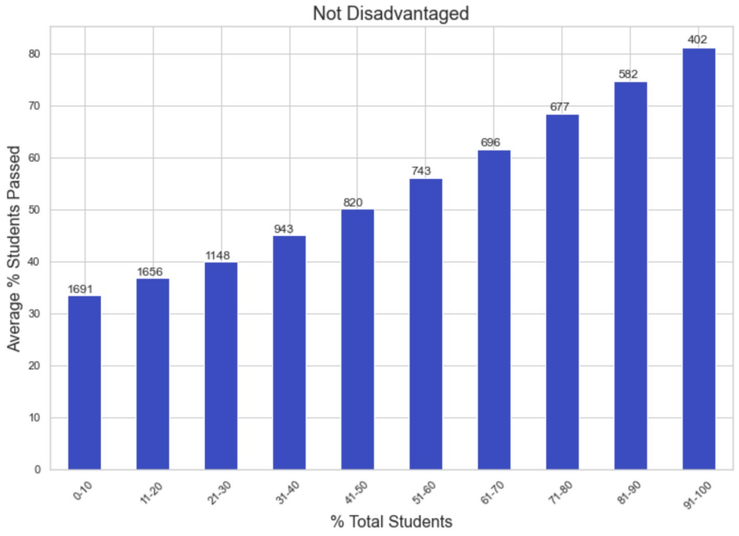


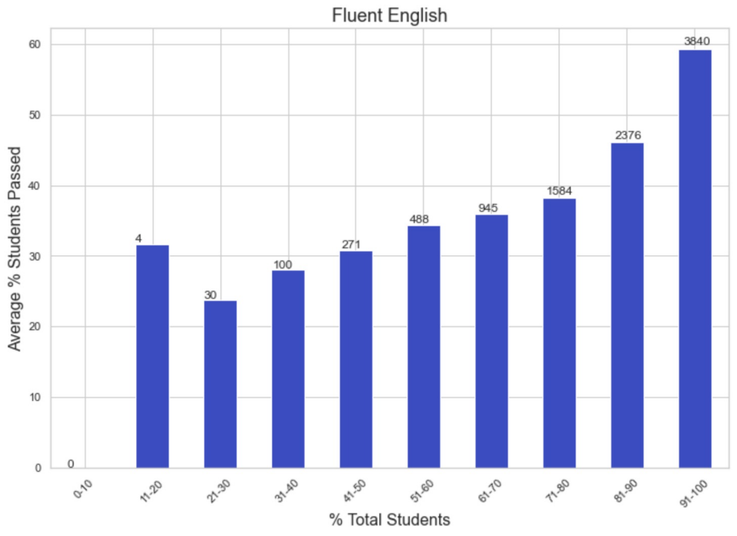
Students in higher income zip codes score higher at standard tests.

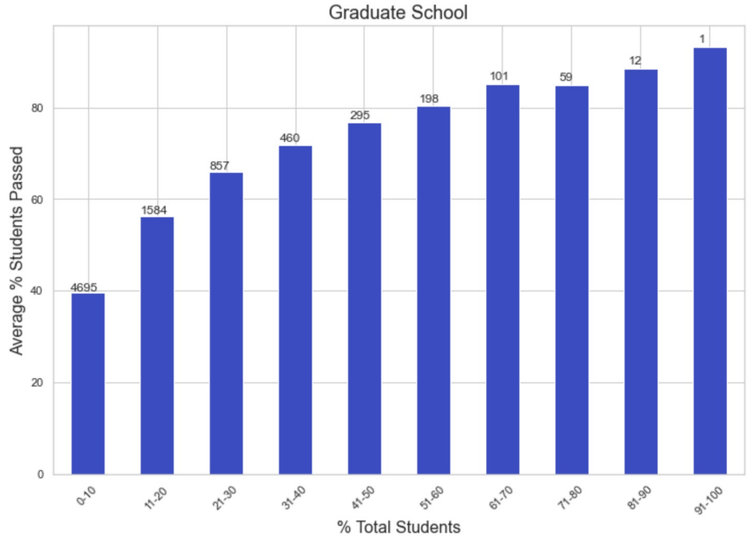
English learners score lower than students fluent in English.

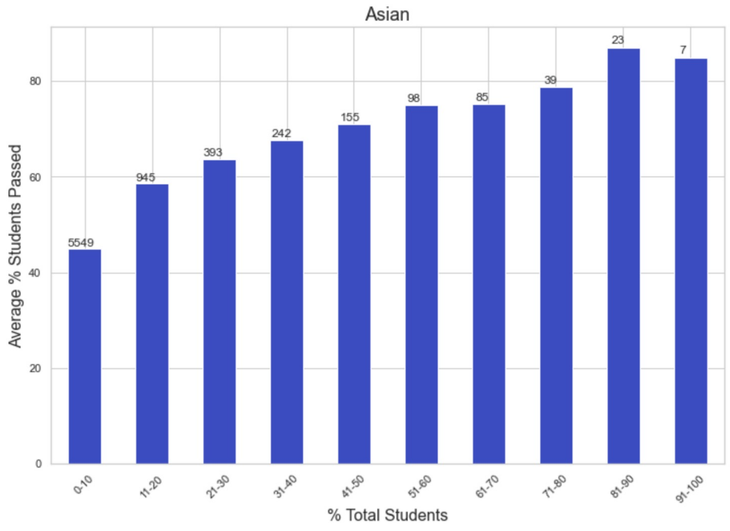


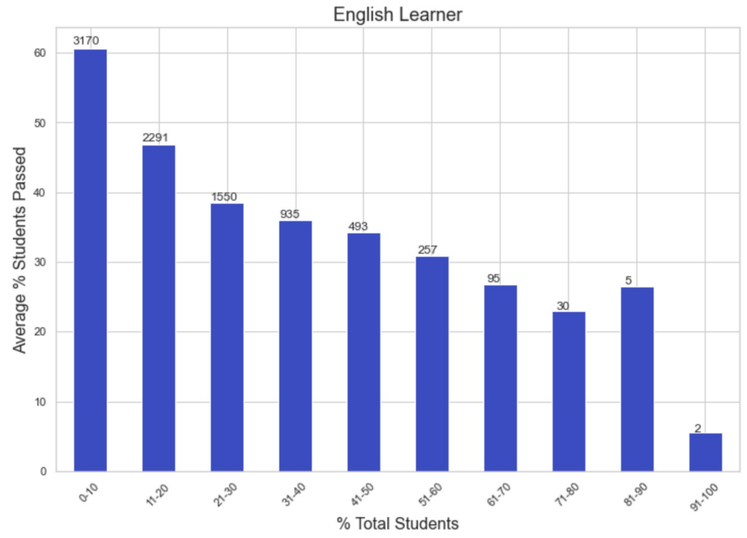


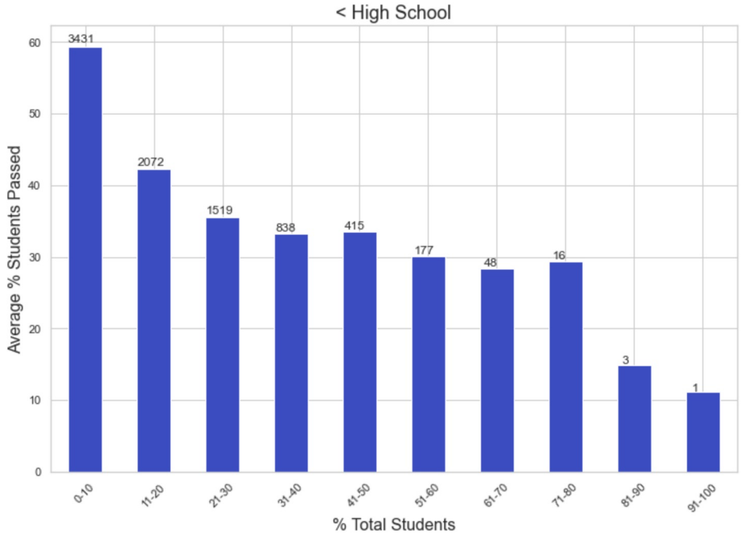

Not Disadvantaged
Average % Students Passed
% Total Students
0-10
11-20
21-30
31-40
41-50
51-60
61-70
71-80
81-90
91-100
1691
1656
1148
943
820
743
696
677
582
402

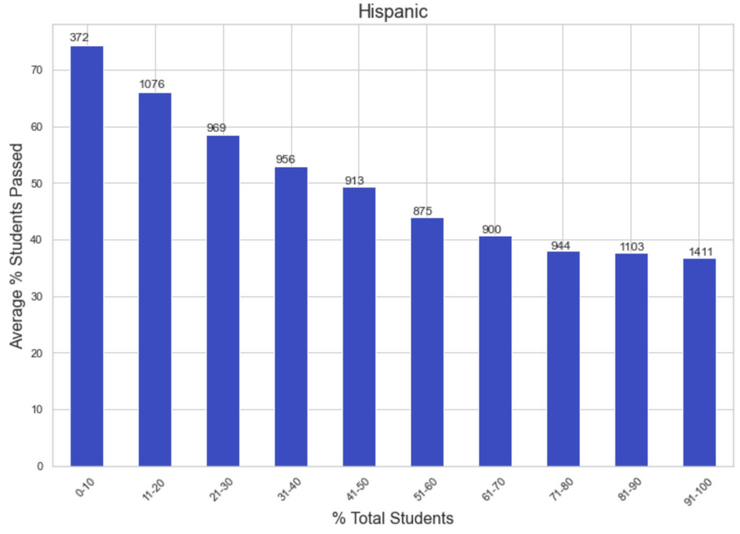

Fluent English
Average % Students Passed
% Total Students
0-10
11-20
21-30
31-40
41-50
51-60
61-70
71-80
81-90
91-100
0
4
30
100
271
488
945
1584
2376
3840

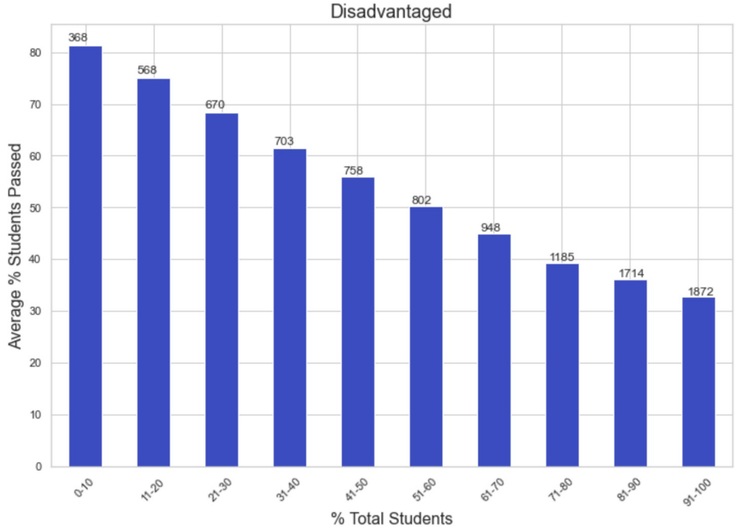

Graduate School
Average % Students Passed
% Total Students
0-10
11-20
21-30
31-40
41-50
51-60
61-70
71-80
81-90
91-100
4695
1584
857
460
295
198
101
59
12
1


Asian
Average % Students Passed
% Total Students
0-10
11-20
21-30
31-40
41-50
51-60
61-70
71-80
81-90
91-100
5549
945
393
242
155
98
85
39
23
7


English Learner
Average % Students Passed
% Total Students
0-10
11-20
21-30
31-40
41-50
51-60
61-70
71-80
81-90
91-100
3170
2291
1550
935
493
257
95
30
5
2


< High School
Average % Students Passed
% Total Students
0-10
11-20
21-30
31-40
41-50
51-60
61-70
71-80
81-90
91-100
3431
2072
1519
838
415
177
48
16
3
1


Hispanic
Average % Students Passed
% Total Students
0-10
11-20
21-30
31-40
41-50
51-60
61-70
71-80
81-90
91-100
372
1076
969
956
913
875
900
944
1103
1411


Disadvantaged
Average % Students Passed
% Total Students
0-10
11-20
21-30
31-40
41-50
51-60
61-70
71-80
81-90
91-100
368
568
670
703
758
802
948
1185
1714
1872

Advanced Analytics & Insights

Evaluation Metrics: MAE, RMSE, and R2

Libraries:

1. **Scikit-learn**
2. **Statsmodel**

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.

Models applied:

1. Linear Regression
2. Lasso
3. Decision Tree
4. Random Forest
5. Gradient Boosting

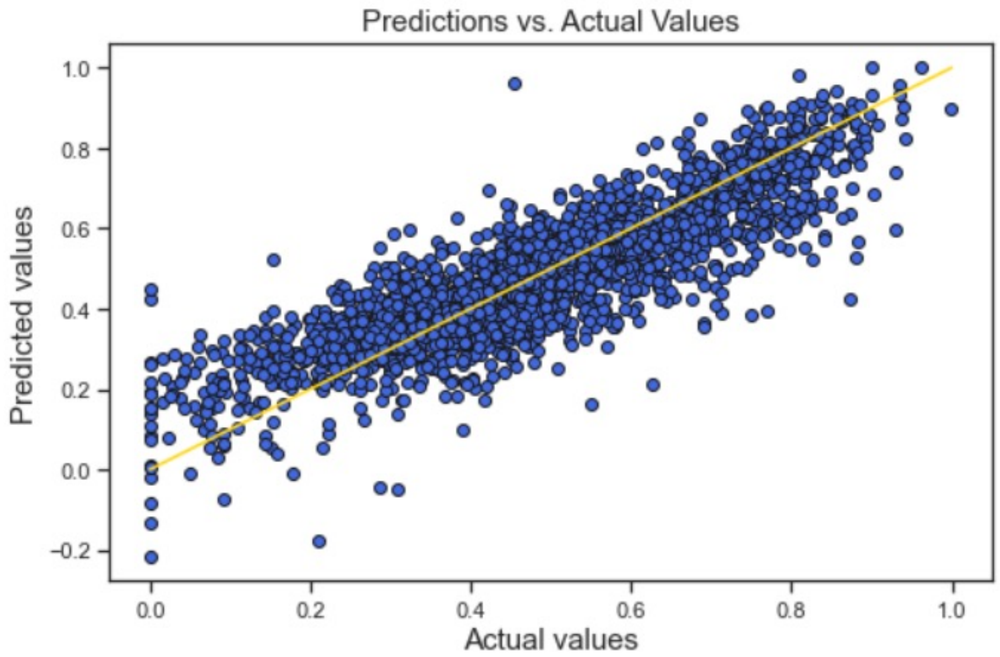
Linear Regression

1

MODEL PERFORMANCE

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.

Issue: Multicollinearity.
Possible solution: Lasso or tree-based models.



Lasso

Lasso is a regularization method with the capability of "selecting" variables by penalizing the high value coefficients.

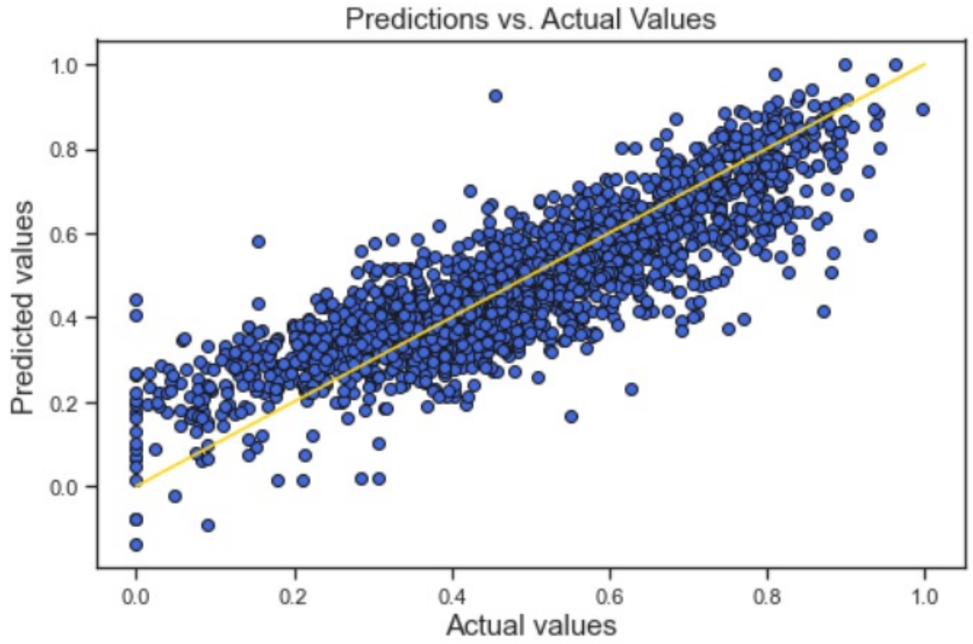
2

MODEL PERFORMANCE

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.

DID IT IMPROVE THE LINEAR REGRESSION MODEL?

Lasso didn’t improve the model.



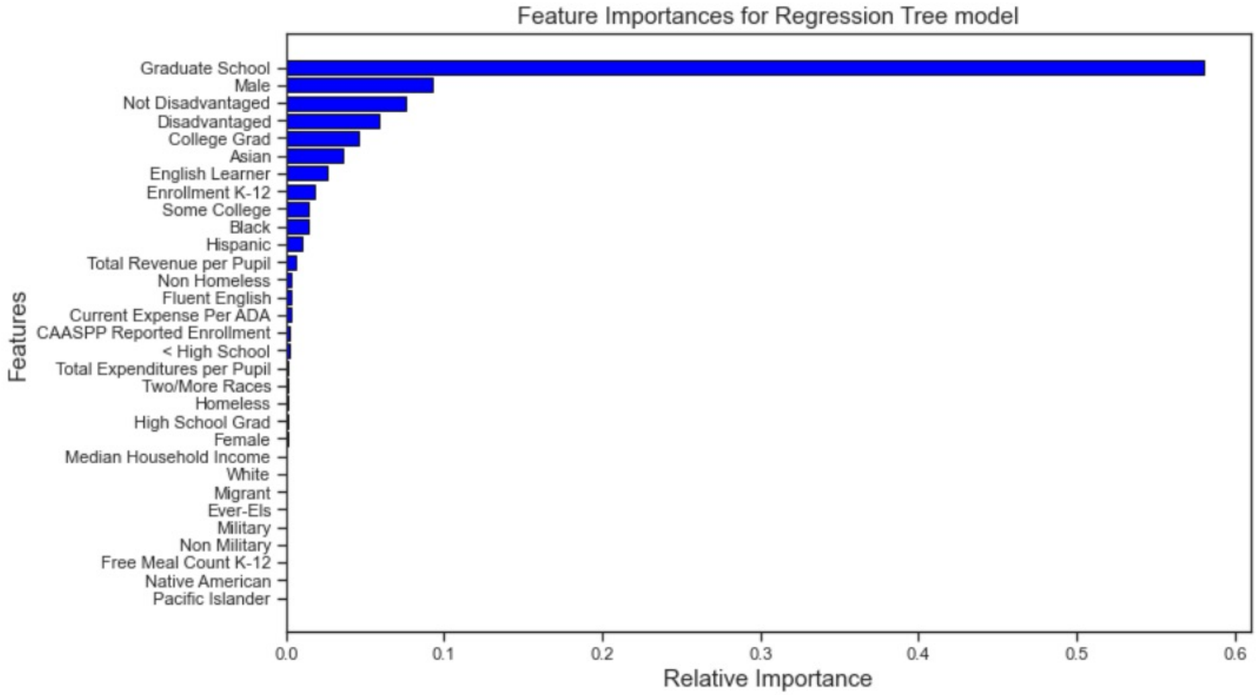
Decision Tree

Decision trees regression normally uses mean squared error (MSE) to decide to split a node in two or more sub-nodes.

3

MODEL PERFORMANCE

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.

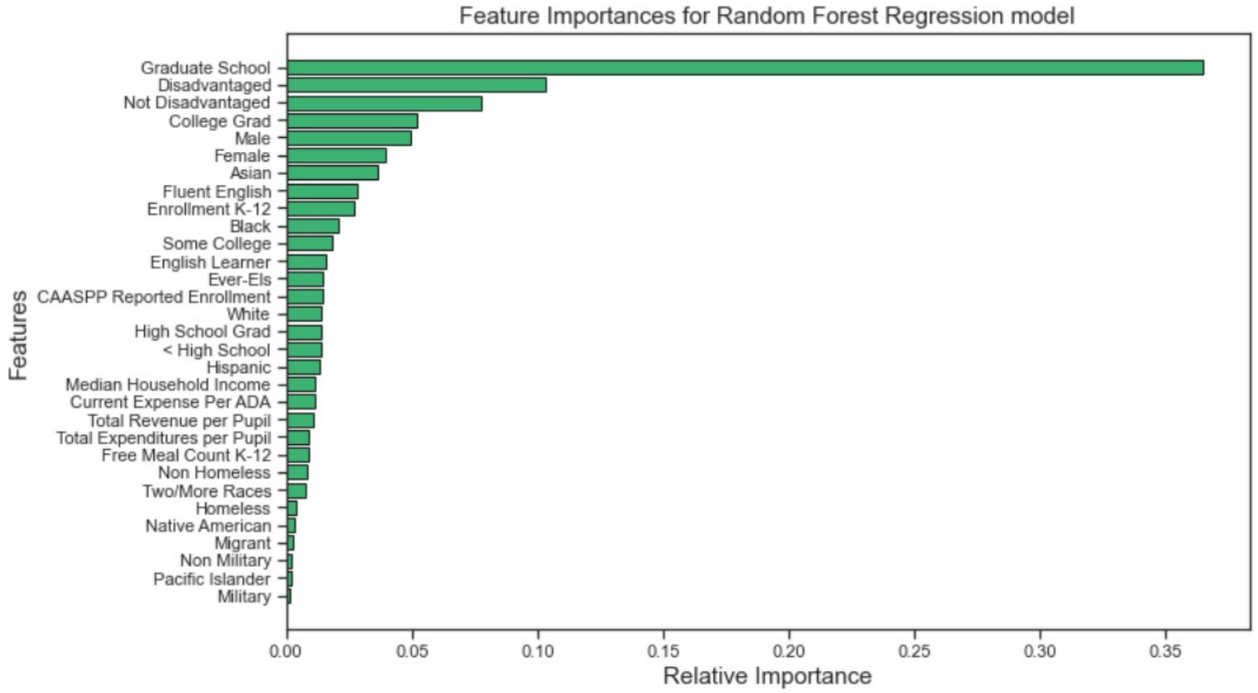


Random Forest

4

MODEL PERFORMANCE

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.

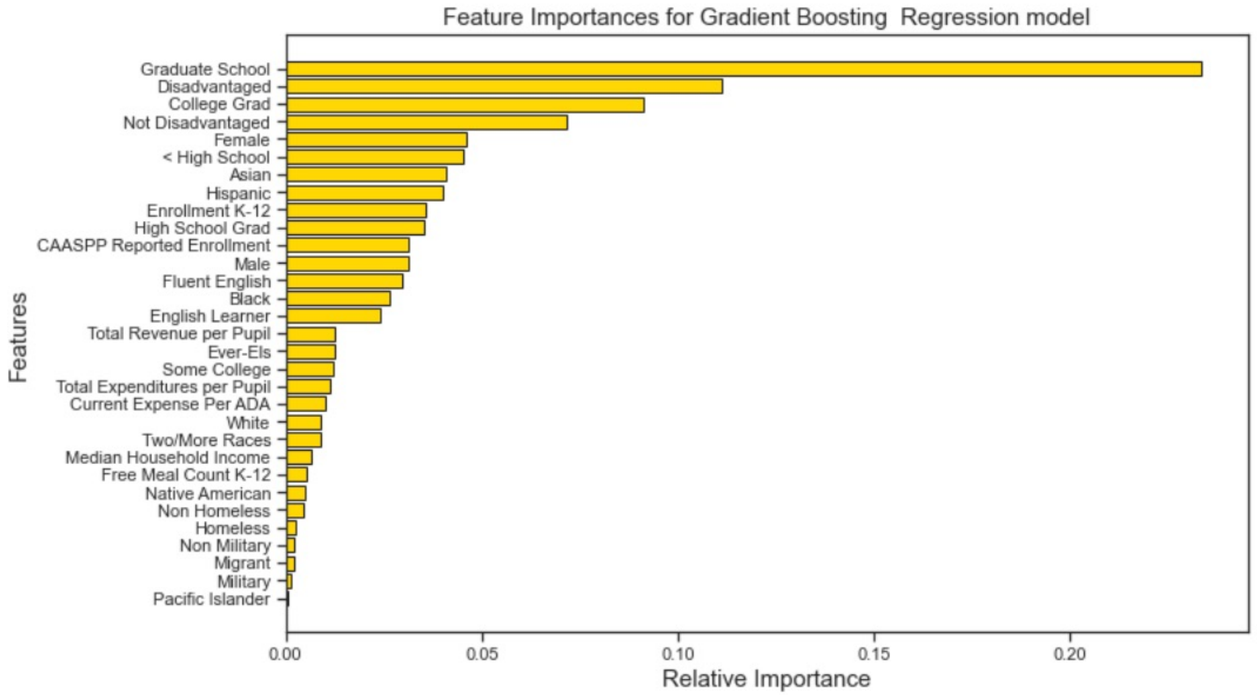


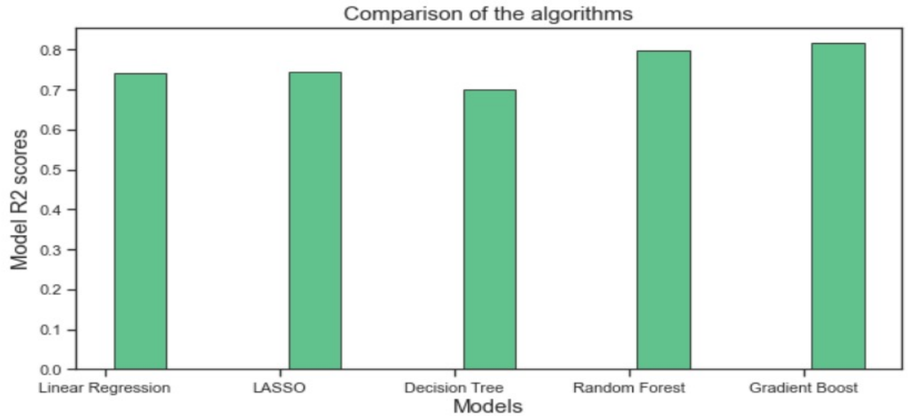
Gradient Boosting is the best performing model!

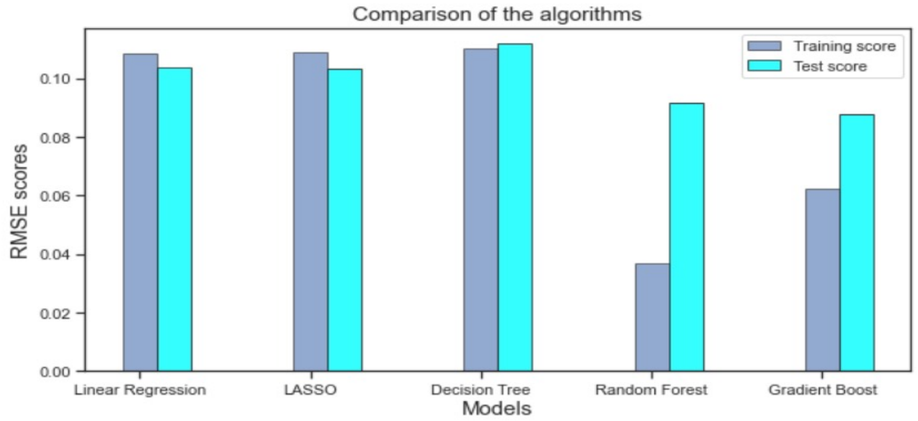
5

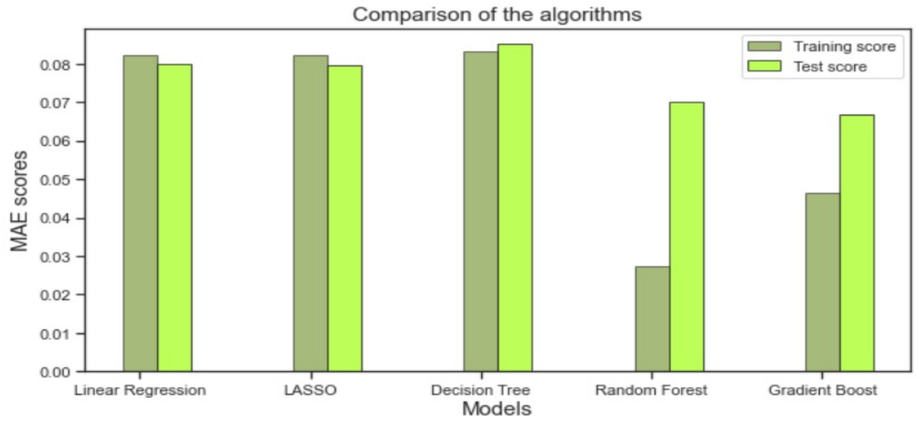
MODEL PERFORMANCE

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.









Model Comparison

Findings & recommendations

- Conclusion
- Recommendations

Conclusion

1. The higher the level of parental education, the higher the achievement of students.
2. Economically disadvantaged students have more difficulties than not-economically disadvantaged students.
3. Hispanic and ELS students score lower.
4. Asian, Whites and English fluent students scores higher.

Recommendations

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

Questions?

Thank you!