Final Report

Evaluating the PSAT by Comparing SAT Score Predictions using PSAT Scores vs Demographic Data

Daniel Ellis Schwartz

Problem Statement

The purpose of this analysis is to evaluate SAT score predictions made using only PSAT scores and predictions made using only demographic data. The relative accuracy of these two methods may allow a recommendation of the value in continued universal administration of the PSAT in Colorado public schools.

Background

The SAT is a standardized test taken by over a million students in the United States every year. While its primary use is for college admissions the SAT has recently been adopted by the state of Colorado as part of a new set of high school graduation requirements. High school students in Colorado take the SAT and Preliminary SAT (PSAT) every year when their schools provide and administer the tests at no charge to students. Only the SAT has any direct impact on a student's college application or high school graduation prospects yet they are also required to take the PSAT.

A key premise of the PSAT is that it gives students, parents, and schools a good idea of how a student will perform on the SAT when they take it the next year. However many studies have shown that some of the best predictors of student scores on standardized tests like the SAT are demographic factors, primarily family income and race. These studies significantly undermine the usefulness of the SAT and indicate major problems with regard to equity and social justice. Beyond the high financial costs to already underfunded schools and the opportunity costs in educational time and resources, these tests are a major source of stress for students and teachers alike. These costs and burdens must be justified for continued universal administration of the PSAT.

The focus here is on the PSAT and the premises upon which it is justified. All of the demographic factors that can be used to predict SAT scores also apply to the SAT. If the PSAT is a good predictor of SAT scores then it should be a better predictor of SAT scores than demographic variables alone. If there is no meaningful difference in predictive value then this would significantly undermine the value of administering the PSAT to all high school sophomores in Colorado.

Datasets

Test score data broken down by demographic categories was pulled from the Colorado Department of Education (CDE) through the Schoolview State Assessment Data Lab online portal. Two data sets were used, one containing PSAT scores from 2017 and one containing SAT scores from 2018. The demographic categories are gender, ethnicity, free & reduced lunch eligible (FRL), English language learners (ELL), and individualized education program status (IEP). Additionally the median family income of each school's zip code was pulled from usa.com.

The data does not contain scores for individual students. It contains the mean scores on the math and English based reading & writing (EBRW) sections of the SAT or PSAT for groups of students that fall into subgroups. For example: a group of male, white students who are not FRL eligible, not ELL, and do not have IEP's at Academy Online high school had a mean scale score of 560.6 on the EBRW section of the SAT in 2018. If there were fewer than 17 students who fell into a particular grouping then no score was reported for that grouping. Consequently many small groupings of students are missing in the data and are not factored into my analysis.

To best approximate tracking individual student scores I used PSAT score data from 2017 and SAT score data from 2018 based on the assumption that most students would stay at the same high school between those schools years.

Median family income data comes from American Community Survey 2014 data accessed via usa.com.

Data Wrangling

The raw data pulled from CDE had 13 columns and approximately 18,000 rows for each of the SAT and PSAT score datasets. Each row represents one permutation of all five categories for each school. Since at least 17 students were required for subgroup's scores to be reported the vast majority of the score entries were missing. After dropping missing entries 726 PSAT scores and 696 SAT scores were available for each test category. The two score datasets were merged based on school and demographic subgroups. PSAT scores that did not have a match SAT score were then dropped since SAT scores were the target.

Median family income data was merged with the score data by pulling school addresses from an online list and matching the school zip codes to median family income by zip code.

The features kept for analysis are:

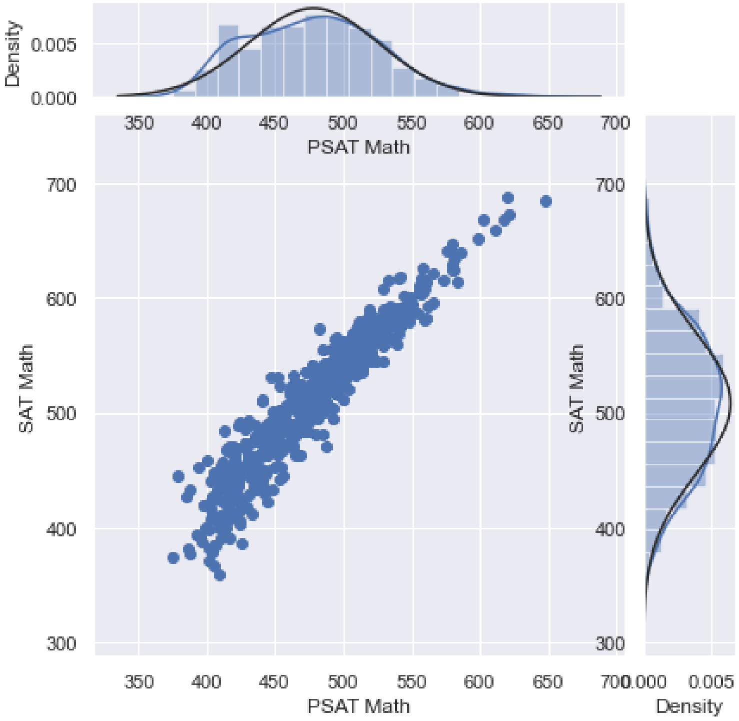
- Gender (2 possible values)
- Ethnicity (7 possible values)
- Free and Reduced Lunch (2 possible values)
- English Language Learners (2 possible values)
- Individualized Educational Program (2 possible values)
- Median Family Income (Continuous variable)
- Mean Score PSAT Math or EBRW (Continuous variable)
- Mean Score SAT Math or EBRW (Continuous variable)

Mean SAT scores for Math and EBRW were the target features.

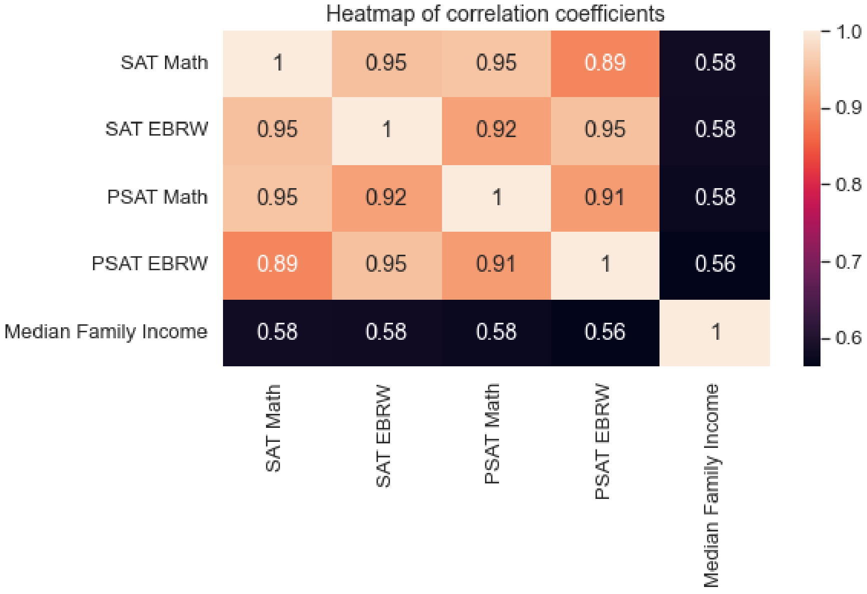
The final dataset used contained 696 rows.

Exploratory Data Analysis

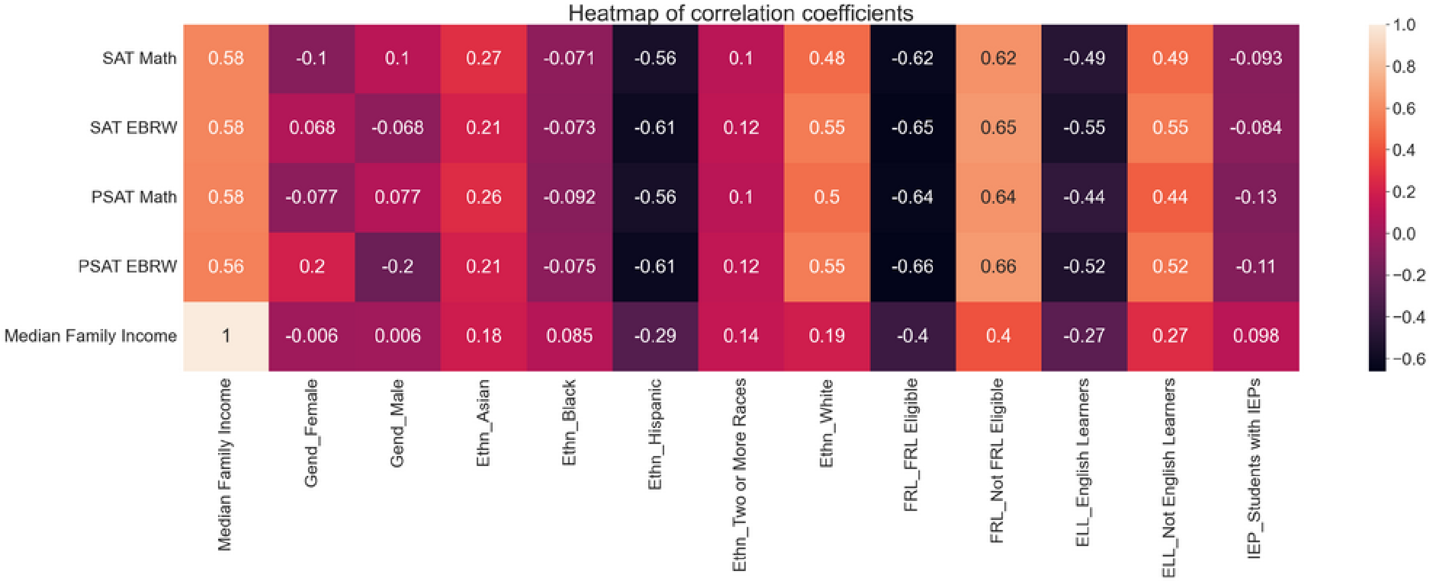
Initially there appears to be a clear linear relationship between PSAT and SAT math scores. Both sets of scores approximate a normal distribution with some variation. The figure below shows the relationship between PSAT and SAT scores as well as the distribution of PSAT math and SAT math scores with normal curve overlaid. The same relationships and patterns were seen in the EBRW scores.



A heatmap of correlation coefficients shows high correlation between scores on different tests or test sections. A coefficient of 0.95 between PSAT Math and SAT Math support the relationship observed in the scatterplot above. This is to be expected. Students who do well in math will do well in the math sections of both tests. The heatmap below also includes the only other continuous variable: median family income. A coefficient of 0.58 (or 0.56) indicates a significant correlation between median family income and all test scores. Additionally, extremely high correlation between SAT Math and SAT EBRW scores indicates that a high score in one subject implies a high score in the other subject.


Heatmap of correlation coefficients

One-hot encoding of the categorical features leads to a dataset with 13 variables, including median family income. The heatmap below shows correlation coefficients with encoded features.



The highest correlations between test scores and features involve FRL eligibility, median family income, and ethnicity. As would be expected FRL eligibility is significantly correlated with median family income.

Modeling

Four sets of models were created and evaluated. Models using PSAT scores and models using demographic variables were each applied to the math and EBRW test subjects. Training & testing sets were split and median family income was scaled according to training data. No other variables needed to be scaled.

Only math scores will be covered in this section. The math score models had the greatest error and the best performing model was the same for both test subjects. EBRW score results will be addressed in the next section.

The metrics used to evaluate all models were mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE).

PSAT based models

Two models were created using PSAT scores, one using linear regression and one using gradient boost regression with cross validation. The models and their metrics evaluated on the test set are summarized in the table below.

Model	**MSE**	**RMSE**	**MAE**	**MAPE**
Linear Regression	308.50	17.56	13.32	2.71%
Gradient Boost Regression	429.20	21.66	15.38	3.14%

The linear regression model performed best with mean absolute error of approximately 13.32 points and 2.86% mean absolute percentage error. Due to the way the SAT is scored the MAE corresponds to roughly 1.3 questions on the math section of the test.

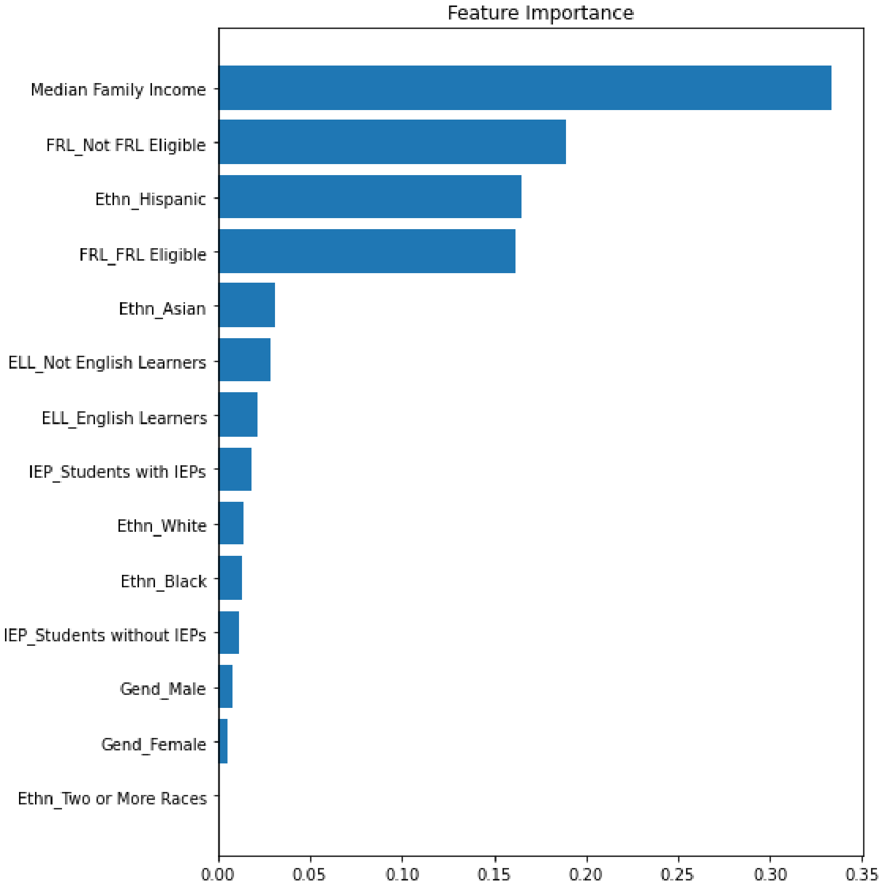
Demographic based models

Three models were created using demographic data: linear regression, gradient boost regression with randomized search cross validation, and random forest regression with randomized search cross validation. Model metrics when applied to the test set are summarized in the table below.

Model	**MSE**	**RMSE**	**MAE**	**MAPE**
Linear Regression	1070.22	32.71	24.82	4.87%
Gradient Boost Regression	857.69	29.29	21.62	4.21%
Random Forest Regression	1158.15	34.03	26.29	5.20%

The gradient boost regression model had the best performance with MAE of 21.62 points and MAPE of 4.21%. This MAE corresponds to approximately 2.1 questions on the math section of the test.

The feature importances in the gradient boost regression model are shown in the figure below. Median family income is by far the most important feature in this model, followed by not FRL eligible, Hispanic ethnicity, and FRL eligible.



The top four important features are interrelated. Free or reduced lunch eligibility is binary so if either FRL eligibility is important then the other will be as well. Median family income of the school's zip code also would correlate with FRL eligibility since the latter is more likely in lower income schools. Finally Latinx people (labeled 'Hispanic' in the dataset) are more likely to have a lower income due to a range of socioeconomic and political factors outside the realm of this analysis.

It is encouraging to note that gender plays a minor role in SAT math scores and an even smaller role in SAT EBRW scores. IEP status also plays a relatively minor role, indicating that the IEP for qualifying students may have the intended equalizing effect.

Findings

The best PSAT based model for predicting SAT math scores has a mean absolute error of 13.32 points and the best demographics based model has a mean absolute error of 21.62 points. The difference between these errors amounts to less than one test question.

For the English based reading and writing (EBRW) test section the errors of the two models were significantly closer. The best PSAT based model for predicting SAT EBRW scores had MAE of 12.88 and the best demographics based model had MAE of 19.69. The difference in these errors is also equivalent to less than one test question.

Best math score model using:	**RMSE**	**MAE**	**MAPE**
PSAT scores	17.56	13.32	2.71%
Demographic variables	29.29	21.62	4.21%

According to these results using PSAT scores to predict SAT scores is statistically superior to using demographic variables alone to predict SAT scores. The question then is whether a subgroup of students' previous test performance provides a meaningful insight into their future test performance as compared to their demographic background.

A reframing is necessary here to make a qualitative evaluation. On the math section of the SAT each question is worth approximately 10.34 points due to the peculiarities of how the test is scored. If we use this to convert the MAE from points to test questions then the PSAT based predictions offer an improvement of approximately 0.80 questions compared to demographic based predictions. Seen in these terms the practical difference between models is diminished.

So what does this all mean? The mean PSAT score of a demographic subgroup of students is a strong predictor of the mean SAT score of that subgroup of students. For that same subgroup of students, their demographic subgroup variables are also a strong predictor of their mean SAT score, though not as strong a predictor as their mean PSAT score. This supports the claim that PSAT scores are a good predictor of future SAT scores. In lieu of having access to PSAT score data, slightly less accurate predictions can be made without using any indicator of academic achievement.

Possible Next Steps

Identifying and accessing more detailed data sources would enable a much more nuanced analysis. Using data on individual students instead of groups of students would include many minority subgroupings that were excluded from CDE's reported data. Even a minor improvement such using the student's zip code instead of their school's zip code would likely increase the accuracy of demographics based predictions. If it becomes possible to predict SAT scores using demographics more accurately than using PSAT scores then the SAT may not be a direct indicator of anything academic.

However, the difference between a predicted mean score and a measured mean score can be used to indicate whether academic factors are overcoming demographic factors. Put simply, if a demographic subgroup of students performs better or worse than expected then it's worth looking into why.

Identifying groups of students who take the SAT but did not take the PSAT could help indicate whether taking the PSAT improves student SAT scores. This was not possible using the data available since it only included mean scores for groups of students.