Standardized Testing is it Really Effective?

Rohan Shajil

Inpirit AI

Abstract

How effective is standardized testing in assessing students' knowledge? With how much standardized testing has been reinforced in our education, students should be properly assessed for their knowledge, and not be given tests that interfere with their academic learning. Under the federal, Every Student Succeeds act, schools must provide standardized tests. However, some of these standardized tests might not be doing what their intended purpose is and instead interfering with academic learning. had data from previous students on sat scores and GPAs from college and high school. I can use this data and plug it into different models to determine how well these sat scores determine GPA. I used different models to help determine and plot the results. To evaluate and determine how accurate these results are, I used three different error-calculating methods. When using linear regression to evaluate sat scores and first-year college GPA, there was a 0.532 mean percent error. This error seemed high for these two variables. There isn’t a strong enough correlation between sat scores and college GPA for it to be implemented in schools across the nation.

Introduction

Standardized testing has been a part of American education since the mid-1800s. It is now required for all 50 states for students to take a standardized test under the No Schild Left Behind Act of 2002. Colleges look at the scores on the tests students and factor them into college admissions which can be a factor in student success. For this to be the case a fair standardized test that properly assesses student knowledge must be implemented. The Scholastic Aptitude Test(SAT) is a popular standardized that is taken by students all over the country. Many colleges make it mandatory for students to take this Standardized test as it is thought to be a good indicator of student's academic knowledge, however many have argued otherwise. Students have complained about the time it takes to study for these tests and how it often interferes with school learning. As a big time commitment, this test must be able to fulfill its intended purpose of assessing student knowledge. The question this raises is how well the SAT assesses students' knowledge. This problem will be a supervised and regression problem. Since the data we are working with includes numerical data such as sat scores and GPA the output will be quantities.

Background

Many other researchers have contributed to this ongoing problem. For example, Gerhard Sonnert, Melissa Barnett, and Philip Sadler explored the consequences of standardized testing in AP calculus classes. They hypothesized that standardized tests do not have an impact on student academic performance and instead have a negative impact on academic performance. To explore this issue they viewed standardized test scores and used data from the factors Influencing College Success in Mathematics to study the long and short terms effects of standardized testing. They found that on the AP and SATs, however, this seems to be not the case. Although students were prepared for the AP Calculus test their calculus ability varied between test scores not giving an accurate representation of their knowledge. This article highlights the failure of standardized testing but only shows this through the calculus ap test. In another article, the author references many studies performed by a Michigan professor in 1983. This professor navigated through most of the standardized tests and content in textbooks which is taught and school and found out that 80 percent of the content does not appear in the standardized tests. He concluded that the majority of the material in standardized tests does not reflect what is taught in school. However, the study was from 1983 so standardized testing has changed quite a bit and the material has also changed.

Dataset

To analyze this problem in standardized testing I took a data set of 1000 students with their sat subscores, total sat score, high school GPA, sex, and first-year college GPA. To split this data I used 700 entries for my training set and 300 entries for my testing set. To avoid overfitting for my model I set aside only 300 testing entries at random. There are no null values in the data set I am working with so no categories will need to be cut out or modified. When describing the sex of each respondent I used a one to describe a male and a 2 to describe a female. With the option of having both the verbal and math portion, I have created some scatter plots to show how the categories in the data correspond with each other. The first scatter plot will show the sat sum and the first-year GPA of college. There are two colors in the graph, blue representing the male and light pink representing female. As shown in the graph there is almost no correlation seen in these two categories. This is something we have to keep inmid when choosing a model to asses our data. There also seems to be no correlation between the sexes and there gpa and sat. In the next figure it shows a scatter plot between the highschool gpa and sat sum. There are two colors in
the graph, blue representing male and light pink representing female.

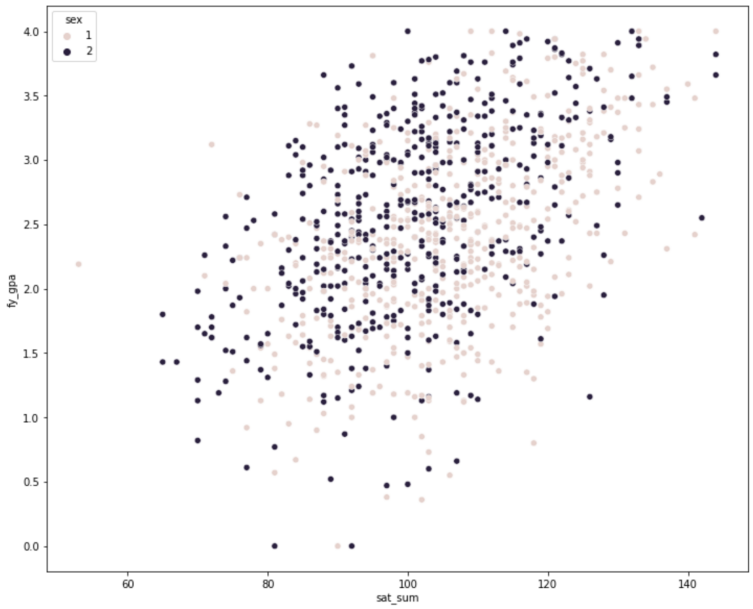

Figure 1 Sat sum vs First year GPA

As shown in the graph there is almost no correlation seen in these two categories. This is something we have to keep inmid when choosing a model to asses our data. There also seems to be no correlation between the sexes and there gpa and sat. In the next figure it shows a scatter plot between the highschool gpa and sat sum.

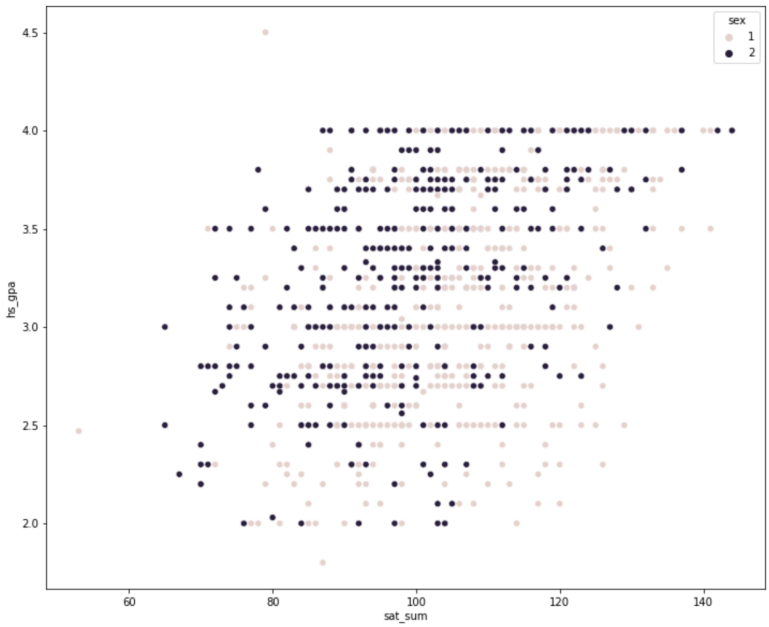

Figure 2 Sat sum vs Highschool GPA

In this scatter plot there is almost no correlation visible to us and and no correlation between the two sexes recorded. In the next two scatter plots figure 3 and figure 4, I split up the sat sum into verbal and math and compared that to the college gpa.

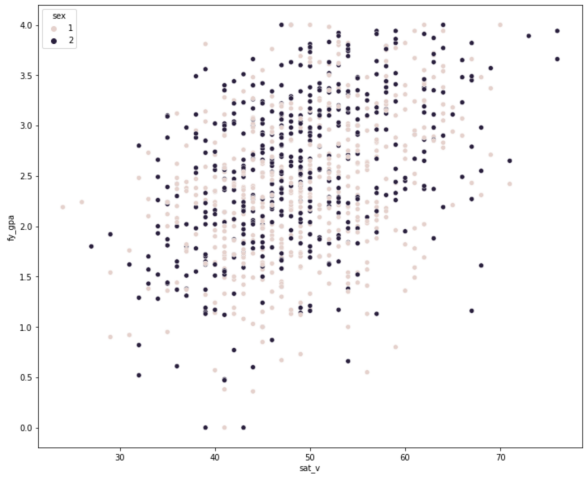

Figure 3 Sat Verbal vs First Year GPA

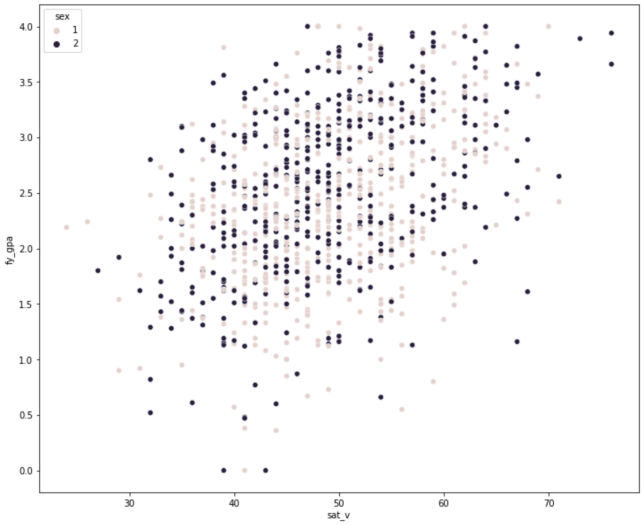

Figure 4 Sat Math vs First Yeat GPA

As shows in the graph there still no correlation between the two variables even after splitting it up into verbal and math. Although there is no connection, figure 4 shows less variation compared to figure 3 showing that math might have more of an indication of college academic success.

Methodology/Models

Given the graphs and data sets it was reasonable to try different models to evaluate the data. The dataset chosen is numerical so I started of with a regression model to evaluate it, more specifically a linear regression model. The basic idea behind linear regression is to find the line that best fits the data, such that the sum of the squared differences between the actual and predicted values (also known as residuals) is minimized. This line represents the relationship between the dependent and independent variables and can be used to make predictions about the dependent variable for given values of the independent variables.

A linear regression model takes the form of an equation, where the dependent variable is represented as a linear combination of the independent variables and parameters:

$y = \beta 0 + \beta 1x1 + \beta 2x2 + ... + \beta nxn$

where:

- y is the dependent variable
- x1, x2, ..., xn are the independent variables

- β0, β1, β2, ..., βn are the parameters of the model, also known as coefficients

The parameters of the model are estimated from the data using a method such as ordinary least squares (OLS), which seeks to minimize the sum of the squared residuals. Once the parameters are estimated, the linear regression model can be used to make predictions about the dependent variable for new values of the independent variables.

I split the data where 700 of the entries are used for training and 300 entries are used for testing. I then used my regression model to train the data i imputed which was the 700 entries. I then made predictions of the first-year college GPA with the testing data and plotted all my results.

After this, I used a different model called the ridge regression model. Ridge regression is a type of regression analysis that is used to analyze the relationship between a dependent variable and one or more independent variables. It is similar to ordinary least squares (OLS) linear regression, but with a slight modification in the way that the parameters are estimated.

In ordinary least squares regression, the goal is to minimize the sum of the squared residuals between the observed and predicted values of the dependent variable. This can lead to overfitting when the number of predictor variables is high relative to the number of observations. Overfitting means that the model fits the training data very well, but may not generalize well to new data.

Ridge regression addresses this issue by adding a penalty term to the objective function that is being minimized. This penalty term, also known as a "shrinkage

parameter," helps to reduce the magnitude of the coefficients, which can prevent overfitting. The equation for ridge regression is:

$y = \beta0 + \beta1x1 + \beta2x2 + ... + \betanxn + \varepsilon$

where y is the dependent variable, x1, x2, ..., xn are the independent variables, β0, β1, β2, ..., βn are the coefficients, and ε is the error term. The objective function to be minimized in ridge regression is the sum of the squared residuals plus a penalty term:

$L = \sum (y - y\_pred)^2 + \lambda \sum \beta^2$

where λ is the shrinkage parameter, and β^2 is the square of the coefficients. A larger value of λ results in more shrinkage and a stronger penalty on the magnitude of the coefficients.

Ridge regression can be a useful tool when there is multicollinearity (high correlation) among the predictor variables, as it helps to avoid the unstable coefficient estimates that can result from multicollinearity.

Using this ridge model imputed my training data which I split the same proportions as the linear regression model. Then I used my testing data to make some predictions about the first-year GPA of college. I then created a scatterplot of my results.

The final model that I utilized in my analysis is the random forest regression model. Random Forest Regression is a type of machine learning algorithm that is used for regression problems. It is an ensemble method, meaning that it combines the results of multiple individual models to make a prediction. The individual models in a Random Forest Regression are decision trees. The idea behind Random Forest Regression is to use many decision trees to create a "forest" of models. Each tree in the forest is grown using a random sample of the data, and each tree is grown using a random subset of

the independent variables. The final prediction is made by averaging the predictions of the individual trees. For this model, it was most optimal to split my data as 800 for training and 200 for testing. I imputed my training values into the random forest model and made predictions on first-year GPA.

Results

When calculating the predictions using the testing data I calculated the error using three methods. These methods included mean absolute error, mean squared error, and r2 score. Mean Absolute Error (MAE) is a commonly used metric for measuring the quality of predictions. It's a measure of how close the predicted values are to the true values. It is a simple and straightforward measure of prediction error that is easy to understand and interpret. This type of error is more sensitive to outliers as it weighs it more in the calculations. The R^2 (R-squared) score is a commonly used metric to evaluate the goodness of fit of a regression model. It measures the proportion of variation in the dependent variable that is explained by the independent variables in the model. An R^2 score of 1 indicates that the model perfectly fits the data, while an R^2 score of 0 indicates that the model does not explain any of the variation in the dependent variable. A negative R^2 score indicates that the model is a poor fit for the data and that the predicted values are worse than simply using the mean of the dependent variable.The mean squared error is a measure of the magnitude of the error, with a higher value indicating a worse fit and a lower value indicating a better fit. It is a

commonly used metric for evaluating the accuracy of regression models, and it can be easily minimized by adjusting the parameters of the model. When calculating these error for my linear regression model I got the mean squared error to be 0.4576, the r squared score to be 0.1788, and the mean absolute error to be 0.5320. For the ridge regression model I got a mean squared error of 0.4849, a r squared score of 0.1296, and a mean absolute error of 0.5605. Based on these results we can see that there can be no correlation concluded between the given sat scores and college gpa.

Acknowledgments

This research paper was not possible without the assistance of Christina Cheng and the guides give by Inpirit AI.

Refrences

.Author Learn more about Mark Dynarski. , M. D., & Learn more about Mark Dynarski. Mark Dynarski Advisor, M. D. (2022, October 26). *When done right, standardized tests reveal a student's knowledge*. George W. Bush Presidential Center. Retrieved February 9, 2023, from https://www.bushcenter.org/publications/when-done-right-standardized-tests-reveal-a-students-knowledge

Martin, H. (2021, August 13). *Why standardized tests fail to assess students' knowledge*. Talented Ladies Club. Retrieved February 9, 2023, from https://www.talentedladiesclub.com/articles/why-standardized-tests-fail-to-assess-students-knowledge/

Moilanen, S. (n.d.). *How well do standardized tests measure a student's ability?* The Adams Kilt. Retrieved February 9, 2023, from https://theadamskilt.com/opinion/how-well-do-standardized-tests-measure-a-students-ability/

Murrell, P., & Murrell, P. (2021, September 6). *What standardized tests do not measure*. Rethinking Schools. Retrieved February 9, 2023, from https://rethinkingschools.org/articles/what-standardized-tests-do-not-measure/