Indicators of student test performance

Andrew Walther

10/7/2020

Introduction

Data set

The data used in this analysis was acquired from Kaggle and is the StudentsPerformance data set. This data set includes demographic information for 1000 students along with test scores on math, reading, and writing exams. Some of the demographic information collected includes: gender, race/ethnicity, parental level of education, free/reduced lunch, and completion of test preparation course. The data set can be found here: https://www.kaggle.com/spscientist/students-performance-in-exams.

Specific Terminology

We compute a metric in this analysis called `avg.score` that is simply the average score for each student across their math, reading, and writing exam scores.

Exploratory Visualizations

Regression model to distinguish examine scores by gender

Decision tree to separate scores by gender

Correlation between different tests

Conclusions

-conclusions from EDA & Modeling

Future Ideas

The dataset used in this analysis didn't provide much information about students who took the exams. In addition, we found that the 3 exam scores reported for each student were strongly correlated with each other, essentially relegating us to a single numeric response in the average score. Going forward, it would be interesting to take advantage of a much large set of data, like demographic & performance metrics for students who take the ACT & SAT each year. Some additional factors that could make further analysis interesting are: student age, household income, hometown or state of residence, hours spent preparing for a particular exam, and coursework grade point average, among others.

Given more information about a student's background and potential aptitude for an exam, it would be interesting to work on building a prediction model using variables like parental education, household income, GPA in coursework, and hours spent studying to determine if these factors can be used as valid predictors of

a student's outcome on an exam. This could be a simple as a multiple regression model or as complex as a deep learning neural network, depending on how clear of a pattern there might be in the data. Additionally, if data from the SAT or ACT exams is considered, it would be very interesting to record data regarding a student's future in higher education like: institution attended, rank of institution attended, selected academic major, undergraduate GPA, and grades across different disciplines that can be related back to scoring in various portions of the SAT/ACT exam to determine if a test score is a good benchmark for past performance (in high school) and a strong indicator of future aptitude for success in university coursework. For example, does a high math score on the ACT translate into a student electing to be a math or statistics major and have success in that particular area of coursework?