Which Public School Districts in Texas Have the Lowest Average SAT Scores?

Breonna Moore

Project Design

I used data from educational institutes to predict average SAT scores for public school districts (psds) in Texas (TX). Insights gained from this model can be used by nonprofit SAT Preparatory organizations to identify which psds need their resources the most.

Modeling for this project was performed using Linear Regression (LR), Lasso Regression with cross validation (LassoCV), and Ridge Regression with cross validation (RidgeCV). The latter two models are types of linear regression that use regularization and were cross validated across ten folds. All models R-squared values ranged from 0.17 to 0.42. However, I found that Linear Regression was the best model for predicting SAT scores on my dataset with a R-squared value of 0.42 and a Mean Absolute Error (MAE) of 50.52 SAT points on my test dataset.

The LR model identified features that are the most and least associated with predicting avg. SAT scores by assigning their Beta values from 0 to negative or positive infinity, with zero signifying no correlation and pos. and neg. infinity signifying the greatest correlation to avg. SAT scores in the model. The most important feature, albeit negatively correlated with avg. SAT scores, was the Percentage of Economically Disadvantaged Students, which is a classification for students receiving free or reduced-price lunch. A one percent increase in the percentage of economically disadvantaged students would result in a 0.33 point decrease in the avg. SAT scores of that psd. Consequently, psds with a higher percentage of economically disadvantaged students were the most likely to have lower average SAT scores. The second and third most negatively associated features were Annual Dropout Rate and Percentage of Non-White Students, respectively. There is no surprise for the latter metric having a large impact on SAT scores as the data from the College Board on SAT scores is in agreement. At first glance of my model, the monetary features seemed to have a smaller correlation with avg. SAT scores than I assumed they would. However, after looking into the features and their interactions, I found that median parental income was highly correlated with percentage of parents with graduate and professional degrees, meaning that the higher the percentage of parents with graduate or professional degrees a psd had, the higher the median income was for that psd. Even though it seemed as though median income had very little association with avg. SAT scores in the model, I think that it does impact avg. SAT scores vicariously through the percentage of parents with graduate and professional degrees feature, which is the top most positively associated feature to avg. SAT scores.

Tools

- Python
    - Web Scraping: BeautifulSoup
    - Data Collection: Pandas
    - Data Analysis: Scikit-learn, StatsModels, NumPy
    - Data Visualization: Matplotlib
- Microsoft PowerPoint, Microsoft Word, Google Docs

Data

I collected data from Ballotopedia, the Texas Education Agency, and the National Center for Education Statistics (NCES). I scraped the TX psds, and the remainder of the data was collected from .csv, .xlsx, and .txt files. The avg. SAT scores were from the 2016 - 2017 school year, and the features were from the 2014 - 2015 school year. The majority of the files included a school district identification number (sdi) that was used to identify the school, but the files from the NCES did not include the sdis that TX schools used. I found a data source that I used to relate NCES's local education agency identification numbers (leaid) to sdis, which posed problems when some of the leaids did not match for the same psd. These inaccuracies in the data initially cost me many datapoints, but I was able to manually merge the rows that were labeled incorrectly which allowed me to recover those datapoints.

I used ten features, and started with 1010 rows, which decreased to 700 after cleaning the data. The features used in my dataset and a hyperlink to their sources are listed in Appendix 1. The features are a combination of school data, parental income information, and student data.

Some of the features in my dataset had to be log1p transformed to correct their positively skewed distributions, and to remedy the multicollinearity, I removed some features for my final linear regression model.

What I Would Improve Upon Next Time

I would have liked to gather more actionable features such as which classes students had taken, extracurricular activity participation, teachers' education, and length of teaching career to name a few. I think this would have increased the model's usefulness to schools that could use these highly associated features as a guideline or suggestion for students and teachers. While I know that something like PSAT test scores would probably have a high correlation with SAT scores, I wanted to find other features that influence SAT scores since PSATs are not taken until later in high school and some schools do not offer the test. This research could be combined with other research to the pinpoint the students that need the most help and the best time to help these students, which may be before PSATs are offered.

Even with my model not describing as much of my test data as I would have liked, I learned a lot about how to take a machine learning project from beginning to end on my own whilst using copious online resources to guide my path.

Appendix 1: Features

Variable	Data Type	Description
Average SAT Scores	Numerical	Average SAT scores by School District
Funding per student	Numerical	Annual funding per student by School District
Annual Dropout Rate	Numerical	% of students who dropped out of Grades 7-12
% Economically Disadvantaged Students	Numerical	Economic status of students
Average Daily Attendance	Numerical	Sum of attendance counts ÷ days of instruction
Student Sex percentage	Numerical (multiple cols)	Percent male and female
Student Race percentage	Numerical (multiple cols)	Percentage of students in different racial groups
Median Parental Income	Numerical	Parental income
Parental Education Attainment	Numerical (multiple cols)	Parental education level
Poverty in the last 12 months	Numerical	Percentage of people whose income in the past 12 Months is below poverty level
Student Enrollment	Numerical	Total student enrollment
FTE Teacher Count	Numerical	Total FTE teacher count
Average Class Size	Numerical	Student enrollment ÷ FTE teacher count

Data Sources

1. Scraping Texas School Districts: https://ballotpedia.org/List_of_school_districts_in_the_United_States#Texas

2. Funding: https://rptsvr1.tea.texas.gov/school.finance/forecasting/financial_reports/1415_FinActRep.html

3. Attendance: https://tea.texas.gov/Finance_and_Grants/State_Funding/State_Funding_Reports_and_Data/Average__Daily_Attendance_and_Wealth_per_Average_Daily_Attendance

4. Student Demographic & Parental Data: https://nces.ed.gov/programs/edge/TableViewer/acsProfile/2014

5. Annual Dropout Rate: https://tea.texas.gov/Reports_and_Data/School_Performance/Accountability_Research/Completion%2C_Graduation%2C_and_Dropout/Annual_Dropout_Data%2C_2016-17

6. Economically Disadvantaged: https://rptsvr1.tea.texas.gov/adhocrpt/adstc.html

7. Student Enrollment: https://rptsvr1.tea.texas.gov/adhocrpt/adste.html

8. Teacher Count: https://rptsvr1.tea.texas.gov/adhocrpt/adfte.html

9. SAT scores for Texas: https://tea.texas.gov/Reports_and_Data/School_Performance/Accountability_Research/College_Admissions_Testing__SAT_and_ACT