

Data Mining final project

John Bowman, Amal Kadri and Alice Kemp

Spring 2022

Abstract

Introduction

The subject of school performance has been heavily researched in the past with most studies coming to the conclusion that household income and racial/ethnic demographics are the most predictive factors of school performance. Students who come from households with predominantly high socioeconomic status tend to perform better than their peers who come from lower socioeconomic circumstances - this trend further aggregates to the school and district level with schools located in neighborhoods of higher socioeconomic status typically outperforming those in poorer areas, as measured by metrics such as standardized test scores, graduation rates, and college acceptance rates. However, in a state as racially diverse as Texas, how well do these trends explain over versus under-performance at the district level? In this report, we will analyze district-level data gathered from the Texas Education Agency during the 2019 to 2020 school year covering student and faculty demographics, SAT/ACT test scores, median household income, enrollment, and graduation rates. From this data, we will identify districts that over or under perform their predicted outcome score and use machine learning techniques to analyze the correlated variables responsible. By doing so, we hope to uncover the key factors that make a district out or under perform other districts with similar demographic makeup. By doing so, we will hopefully deepen our understanding of school performance and use our findings to narrow the achievement gap between districts in Texas and beyond.

Methods

Results

Conclusion

Appendix