Proposal for Capstone Project 1

Ahrim Han

2018/11/7

Title: Prediction of academic performance of public schools in California

(or Identification of the schools that need the most supports)

Problem statement: The California Assessment of Student Performance and Progress (CAASPP) is the assessment system to measure how well students are mastering California's academic standards in English language arts/literacy and mathematics. These results can be used to monitor the student progress and to give feedback the teachers need to change teaching methods more effectively. However, the test scores vary widely from school to school. There is a common belief that affects high test scores such as "schools with many Asian students" or "schools with high-income families". There is a strong need to find more informed and granular causes that impact the test achievements and school ratings of schools.

Who might be the beneficiaries? To broadening educational opportunities, administrators of the school districts/department of education or other organizations can effectively identify the schools that need most supports. Budgets and human resources can be allocated in the order of the needs for tutoring, mentoring, extracurricular programs, educational consultants, and so on. In the equity aspect, schools should strive to create an environment where all students feel valued and all students are learning to high standards. Therefore, using the provided results, teachers can put much more effort into the under-performing groups to reduce the achievement gaps. From a parent's perspective, these results can be an indicator to select a good school that meets the high academic standards.

Data: The test score data of the CAASPP will be acquired from the California Department of Education [1] for years of 2017 and 2018. This data contains the scores of two parts, English language arts/literacy (ELA) and mathematics, for students in grades 3-8 and grade 11. The test data is comprised of state, counties, districts, and schools along with the test scores. The information on parent education, races, disabilities, gender, English-Language fluency can be combined with the test data. For obtaining more accurate and meaningful analysis results or a prediction model, we will acquire the additional data from the GreatSchools [2]. We can also combine the information on teacher demographics [3] or house prices [4].

Solution approach: We categorized the students into four groups (“exceeded standard”, “met standard”, “nearly met standard”, and “did not meet standard”). Then, we will use a supervised classification algorithm to build a predictive model. The classification algorithm not only classifies into four classes but also predict the probability of each class. We will try various machine learning techniques to pick the one which performs best. Based on these results, we aim to identify the schools/districts/groups of students who can effectively raise the test scores. Also, we can suggest the strategies by analyzing the features of the “exceeded standard” group model.

Deliverables:

1) Codes (notebooks) for data cleaning, exploratory data analysis, interactive data visualization, machine learning model development
2) Report on the capstone project
3) Slide on the capstone project

Resources

[1] California Department of Education, 2018

https://caaspp.cde.ca.gov/

https://caaspp.cde.ca.gov/sb2018/research_fixfileformat18

[2] GreatSchools API

https://www.greatschools.org/api/docs/technical-overview/

[3] Civil Rights Data Collection

https://ocrdata.ed.gov/

[4] Zillow research data

https://www.zillow.com/research/data/