School-Specific Estimates of Returns to Increased Education Spending in Massachusetts

Isaac Kasevich, Zane Kashner, and Ethan Oro

Motivation

- At the state level and the district level concern is given to the optimal funding of public education
- Social Science studies examine returns to increased spending using naïve methods
- We seek to model school performance using spending, school-level data and zip-code-level demographic data using machine learning methods learned in CS229

Problem Definition

- Attempt to model scholastic achievement using both exogenous and unchangeable input features
- Back-out the effects of spending changes on performance using Causal Inference

Data

Sources

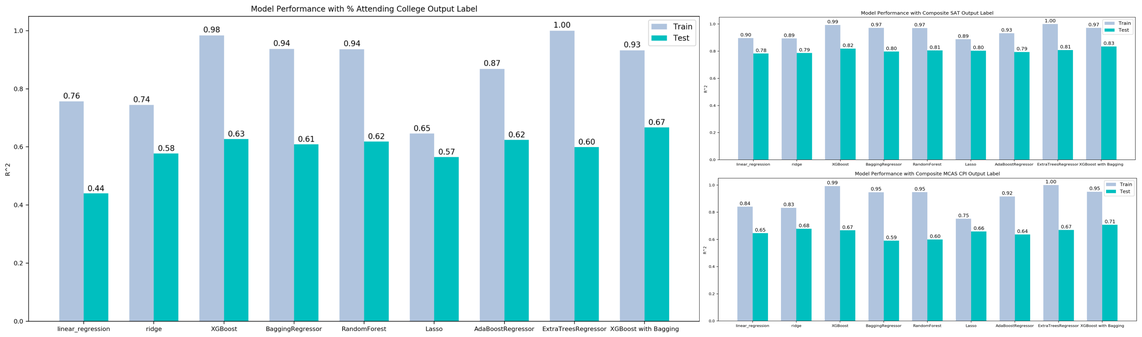
- Massachusetts individual school level data was from a Kaggle dataset from the Mass. Department of Education
- The zip code level demographic, education, and employment data was scraped from towncharts.com which aggregates data from the Census, American Community Survey, Bureau of Labor Statistics, US Geological Survey, Medicare and Medicaid, Common Core of Data and more

Processing

- Preprocessed features, zero-centered mean and normalized variance to one

Modeling Scholastic Achievement; Methods

- Experimented using 9 different machine learning models
- Modeled metrics of scholastic achievement, including % Graduated, % Attending College and Composite SAT

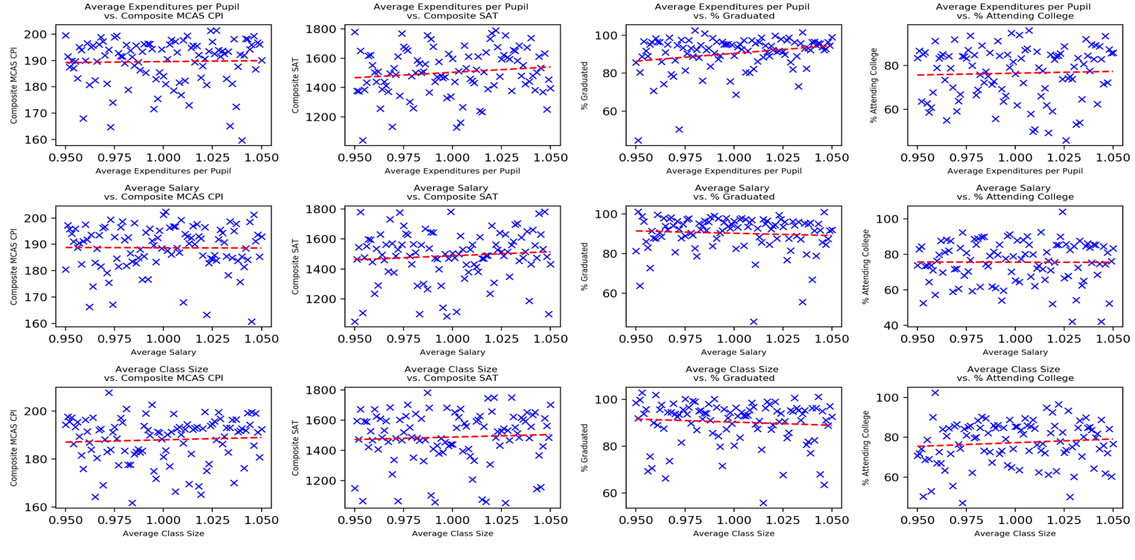


Causal Inference

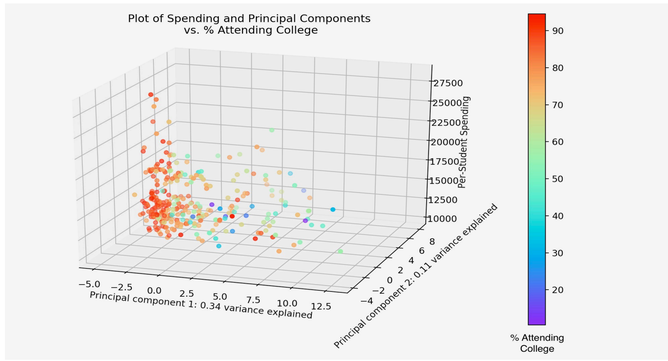
- Estimate the effects of small changes to input features on output metric
- Approximate the gradient of this change using linear regression

$$p(y|X) \approx \hat{b_0} + \hat{b_1} y \qquad \lim_{a \to 0} \frac{p(s+a|X) - p(s|X)}{a} \approx \hat{b_1}$$

Analyzing 12 combinations of changed feature/output metric to visualize correlations



Data Visualization



Principal Components Analysis

- Considering only non-exogenous features (e.g. demographic data), derive top two principal components
- For each example, plot in 3D Avg. Spending and Principal Components
- Color each point by output metric

Challenges

- Difficulty finding a suitable performance metric
- Inexperience and lack of exposure to causal inference methods
- High variance and over-fitting due to limited number of data examples

Future Work

- Numerically solve the constrained optimization problem of optimally allocating a state's education budget to maximize achievement
- Extension of this analysis to elementary and middle schools