School Specific Estimates of Returns to Increased Education Spending in Massachusetts

Isaac Kasevich[1], Zane Kashner[2] and Ethan Oro[3]
{isaack97, zkashner, eoro}@stanford.edu
https://github.com/ethan-oro/returns-to-variable-shcool-spending

***Abstract*— Currently, most–if not all–of the existing literature estimating the effect of educational investment on a variety of output metrics do so using linear regression with relatively unimpressive explanatory power. From these models researchers tend to find weakly positive causal relationships between increased investment and increased achievement. We seek to improve upon this by using more sophisticated models to model outcomes based on investment and a number of other controls. We find that there are great gains to be made in explanatory power from these new models, but they still support the findings of the existing literature: that the relationships between increased investment in a variety of arenas and many different measures of school success are weakly positive.**

I. INTRODUCTION

At the state and the district level a great deal of debate and resources go towards funding public education. There are a number of studies that estimate the returns of increased school resources to student achievement [1] [2]. We seek to build upon these papers by building a model that estimates different measures of high school performance based on policy-relevant factors, as well as exogenous factors, such as the community context of a given school.

Within the social science literature there are a number of different metrics upon which schools are judged. Some of the most common ways that outcomes are measured include standardized test scores, graduation rates, and the rate at which students progress to college. Given that although all of these are measures of how "good" a school is, the mechanisms by which they are changed are likely different. We try using a number of different models to estimate each of these outputs. Massachusetts has rigorous standardized testing, the MCAS, that all students are required by law to participate in the tenth grade [3]. Since this occurs so early in high school, we also consider composite SAT as another standardized test.

In the analysis of school performance that we found within the literature, school specific models were limited to linear regression. We seek to build upon this by developing more accurate models using a variety of more sophisticated techniques. We seek to determine whether these models agree with the assertions from the social science literature about the effects of changing expenditures, class sizes, and teacher salaries. In order to estimate the effect on the performance of a given school we approximate the limit definition of a derivative near the current levels of any of these explanatory variables.

II. RELATED WORK

There has been a considerable body of research by economists and policymakers looking into the relationship between school spending and outcomes. A recent meta-analysis of 377 different publications investigating the relationship between schools funding and academic outcomes notes that irrespective of the the input feature — class size, teacher quality and expenditure per pupil — between 10 and 20 percent of studies found a statistically significant positive correlation, between 5 and 10 percent of studies found a statistically significant negative correlation and the rest were unclear [8]. However, upon further investigation of many of the studies cited, as well as a host of recent studies, we noticed a few trends: many of the studies used an ordinary least-squares linear regression model to find a relationship between school spending and academic outcomes [9] [10] [11]. It was also made clear through these studies that school inputs and school-specific variables were not going to be sufficient to generate a sufficiently comprehensive picture of student achievement; we needed to collect zip-code level census data in order to account for such factors as median income and parental education, profession and hours spent around the home. [10]

Thus, we sought to apply more advanced machine learning methods that might be able to draw more complex relationships between input features and output markers as discussed in existing social science literature.

III. DATASET AND FEATURES

We combined data from a variety of sources to augment school level data-sets with information about their surrounding areas. In order to estimate public school performance we needed context about the income, cost, demographics, and education levels of the communities they served.

Massachusetts individual school level data was sourced from a Kaggle dataset from the Mass. Department of Education [4]. This data has demographic information about each school and enrollment statistics in addition to standardized testing results, graduation rates, college progression rates and funding levels for each school.

We combined these school level inputs with data scraped from `towncharts.com` using both Beautiful Soup 4's html parser and Selenium Webdrivers [5] [6]. This scraped data is from a website that aggregates data from the Census, American Community Survey, Bureau of Labor Statistics, US

Geological Survey, Medicare and Medicaid, Common Core of Data and more. We linked these two datasources by the zip codes in which the schools were located.

We collapsed many categorical features into single features that seemed more informative. For example, we aggregated two zip-code level features

$$absent\ morning = \int_{12:00AM}^{7:30AM} prop\ start(t)dt$$

$$absent\ evening = \int_{11:00AM}^{12:00PM} prop\ start(t)dt$$

to define the proportion of adults who begin work at times such that we would expect them to be away when students are going to or returning from school [7].

Throughout our analysis we removed a number of features we determined to be largely peripheral to the measures of success. This was done in order to reduce over-fitting. Most features removed were ones that theoretically should have no effect on student performance, but were captured in our initial data process.

We transformed categorical variables into one-hot dummy variables. Once we joined our school-level data source by zip-code with our scraped data source, we split our full school data into Elementary, Middle and High Schools and ran our analysis on the 290 high schools in our sample that are public, non charter.

IV. METHODS

As mentioned previously, social scientists have attempted to statistically observe the correlation between school funding and student performance using unsophisticated methods such as ordinary least-squares linear regression. We hope to model this relationship using more sophisticated machine learning techniques. After developing a suitably accurate closed predictive model, we apply continuous intervention to attempt causal inference to isolate the effects of key input variables.

A. Principal Components Analysis for Data Visualization

Before attempting to fit predictive models to our dataset, we visualized our data to see, observationally, if a correlation could exist. To visualize such a high-dimensional feature space in three dimensions, we reduced our exogenous feature space to $\mathbb{R}^2$ using Principal Components Analysis on only those features unaffectable by policymakers. For example, a feature such as classroom size would not be included in PCA since we presume that classroom size could be affected by spending. Rather, we only considered demographic features perceived to be outside the scope of increased scholastic funding.

We plot the two principal components of the demographic data along with *Average spending per Pupil* against one output metric, *% Attending College*:

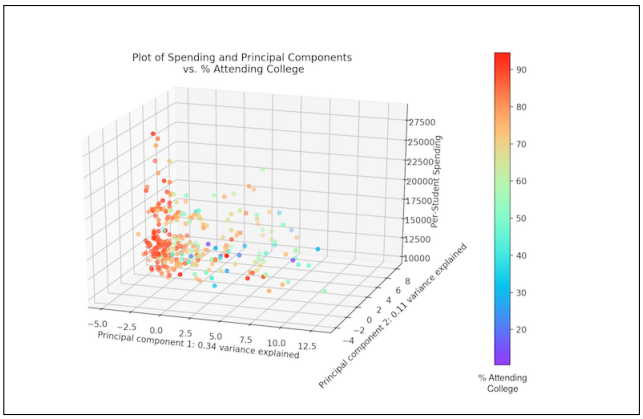

Fig. 1. Principal Components Analysis using % Attending College output metric

From Fig. 1 above we can identify a correlation between spending and the output metric, though we do see significant variance in the principal component space.

B. Closed predictive models to predict student performance

We began by benchmarking performance against the model utilized by many of the existing studies (linear regression) and achieved similar R^2 performance to that of the studies. After observing significant training/validation set performance disparity, we attempted to reduce variance using regularization in ridge and lasso regressions. Though we maintained similar training set performance, we saw marked improvement in validation set performance. Given the high relative dimensionality of the feature space with respect to the number of training examples, we then experimented with several different ensembled tree-based models, on which we elaborate below. We considered and tested more complex models including Support Vector Machines and fully-connected Neural Networks, though abysmal performance on these models indicated an insufficient number of training examples for effective use of SVMs or NNs. Models were implemented in Python using a combination NumPy and SkLearn libraries [12].

Tree-Based Models

Our most accurate models fell under the broader category of tree-based models. We experimented with a number of ensembling methods and combinations thereof, including bagging and boosting. We sought to find robust models and prevent overfitting. Models included XGBoost, Tree Regression, Random Forrest, AdaBoost Tree Regression, Extra Trees Regression, and Bagged XGBoost Tree Regression – which ended up being the most successful model of those discussed.

Hyperparameter tuning increased model performance and reduced overfitting. We capped tree-depth in RandomForest to 4, which prevented overfitting while remaining complex enough to predict accurately. Extra Trees Regressor (Extremely Randomized Trees) performed best with a maximum

tree depth of 5 and 120 estimators. Adaboost, a meta-estimator that uses errors in current predictions to slightly adjust tree weights, performed best using 68 estimators. XGBoost, with a tree depth of 3, achieved exceptional training performance but overfit badly on the validation set. Bagging multiple XGBoost trees reduced overfitting and increased validation set performance. Across all bagged tree models, we found greatest performance using approximately 100 trees.

We delve more thoroughly into the performance differences between various models in the results section.

C. Derivative Estimation and Causal Inference

Using our most successful model, Bagged XGBoost, we used a technique known as Continuous Intervention, Causal Inference [13] to isolate the effect of variables we believed government spending could impact. The goal of causal inference methods is to estimate the derivative of the model's cost function on a single test example x with respect to a specific input feature x_j about the original value of that input feature x_{j_0}. We feed a trained model slight perturbations of the same example, holding all features constant except for the augmented feature x_j. In our implementation, we iteratively augmented the analyzed input feature using 50 values a sequentially selected from the range $a \in [0.950 \times x_{j_0}, 1.050 \times x_{j_0}]$ and fed each augmented input $x_j = x_{j_0} + a$ through our model, denoted $M(x)$. We then used ordinary least-squares linear regression to approximate the numerical interpretation of the derivative:

$$\frac{\partial M(x_0)}{\partial x_j} = \lim_{a \longrightarrow 0} \frac{M(x : x_j = x_{j_0} + a) - M(x_0)}{a} \ .$$

Concretely, we find the slope $\hat{b}_1$ in the linear model $p(y|x) = \hat{b}_0 + \hat{b}_1 x$ which minimizes the cost function

$$\hat{b}(x) = \arg\min_b \|M(x) - bx\|_2^2$$

using data points generated by M in the small interval around x_{j_0}. We interpret this slope as the direct causal relationship between an independent parameter and the output metric — the degree to which the output metric is affected by the input feature in question.

V. RESULTS

We were able to observe significant gains in predictive accuracy using more complex models against the baseline of linear regression. However, even with more accurate predictive models, we still observed weak correlation between government spending and student performance in line with existing studies on the matter.

A. Predictive Model Performance

We chart the performance of each model on its training set and an unseen test set. In general, tree-based regression models significantly outperformed other regressive models in R^2 score, defined as

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2} \ ,$$

where $\hat{y}$ is the model's prediction, y is the ground-truth output and $\bar{y}$ is equal to $\mathbb{E}[y]$. Thus, a model that simply predicts the expected value of y would achieve an R^2 score of 0.0 and a model that predicts perfectly achieves an R^2 of 1.0. Note that, using this version of R^2, $R^2(y, \hat{y}) \in (-\infty, 1]$; a model can be arbitrarily bad, thus generating an $R^2 << 0$. We ran each of our models using four different ground truth metrics, and we present the results for each below:

In TABLE I, using the College Progression output metric, we saw best test-set performance with a Bagged XGBoost Tree Regression model, which achieved an R^2 score of 0.67, a 0.24 improvement on linear regression.

In TABLE II, using the Graduation Rate output metric, we saw best test-set performance with an Extra Trees Regression model, which achieved an R^2 score of 0.65, a 0.19 improvement on linear regression.

In TABLE III, Using the Composite MCAS score metric, we saw best test-set performance with Bagged XGBoost, which achieved an R^2 score of 0.75, a 0.10 improvement on linear regression.

In TABLE IV, using the Composite SAT score metric, we saw best test-set performance with Bagged XGBoost, which achieved an R^2 score of 0.86, a 0.08 improvement on linear regression.

TABLE I
METRIC: % OF STUDENTS PROGRESSING TO COLLEGE

Model	Training Set R^2 Score	Test Set R^2 Score
Least-Squares	0.76	0.44
Ridge	0.74	0.58
Lasso	0.65	0.57
SVM	0.99	-0.03
Random Forest	0.94	0.62
XGBoost	0.98	0.63
AdaBoost	0.87	0.56
Extra Trees Reg.	0.87	0.68
Bagged XGBoost	0.93	0.68

TABLE II
METRIC: GRADUATION RATE

Model	Training Set R^2 Score	Test Set R^2 Score
Least-Squares	0.77	0.46
Ridge	0.76	0.47
Lasso	0.65	0.54
SVM	0.99	-0.03
Random Forest	0.93	0.55
XGBoost	0.98	0.52
AdaBoost	0.87	0.52
Extra Trees Reg.	0.89	0.65
Bagged XGBoost	0.93	0.64

We note that though the R^2 values are much higher across all models for the Composite MCAS and Composite SAT

TABLE III

METRIC: AVERAGE COMPOSITE 10^{th} GRADE MCAS

Model	Training Set R^2 Score	Test Set R^2 Score
Least-Squares	0.84	0.65
Ridge	0.83	0.68
Lasso	0.75	0.66
SVM	0.99	-0.02
Random Forest	0.95	0.60
XGBoost	0.99	0.67
AdaBoost	0.92	0.64
Extra Trees Reg.	0.94	0.73
Bagged XGBoost	0.95	0.75

TABLE IV

METRIC: AVERAGE COMPOSITE SAT SCORE

Model	Training Set R^2 Score	Test Set R^2 Score
Least-Squares	0.90	0.78
Ridge	0.89	0.79
Lasso	0.89	0.80
SVM	0.49	-0.02
Random Forest	0.97	0.81
XGBoost	0.99	0.82
AdaBoost	0.93	0.79
Extra Trees Reg.	0.94	0.85
Bagged XGBoost	0.97	0.86

output metrics, this is likely due to drastic differences in metric scale relative to the Graduation Rate and College Progression metrics, which are measured as percentage values between 0 and 100. Composite SAT scores, by contrast, fall within the $1000 - 2400$ range.

We saw statistically significant performance improvements against linear regression in all output metrics, though the most substantial improvement in R^2 value occurred using the College Progression metric using Bagged XGBoost Tree Regression.

B. Causal Inference Results and Implications

Using the most successful model for each output metric, we proceeded with the previously discussed causal inference model. We conducted this analysis for all combinations of critical input features: {`Average Expenditure per Pupil`, `Average Teacher Salary`, `Average Class Size`} and output metrics: {`% Graduated`, `% Attending College`, `Composite MCAS`, `Composite SAT`}. The approximated derivative for each of these combinations for a single school, *Athol High School, Athol, Mass.*, appears in TABLE V below:

TABLE V

CAUSAL INFERENCE DERIVATIVE APPROXIMATIONS FOR A SINGLE SCHOOL

	% Grad.	% Att. College	MCAS	SAT
Avg. $/pupil	1.94	-2.07	2.65	-27.57
Avg. Salary	2.62	0.44	2.67	-16.52
Avg. Class Size	1.26	1.43	1.69	-2.07

Before analyzing these results, we note that, since our predictive models tended to have high variance and less-than-ideal performance, the causal inference results should be taken as a proof of concept and not as definitive. That said, we do notice some interesting trends that corroborated results of social science research.

Most interestingly, average classroom size generally had weakly positive correlation with output metrics. This runs counter to our initial intuition, as we hypothesized that smaller classrooms would increase student performance. In addition, we notice weakly positive correlations between teacher salary and three of the four output metrics. Coupled with the average class size results, for this school the recommended allocation of funds would be to hire better teachers at higher salaries as opposed to simply hiring more teachers to reduce classroom size.

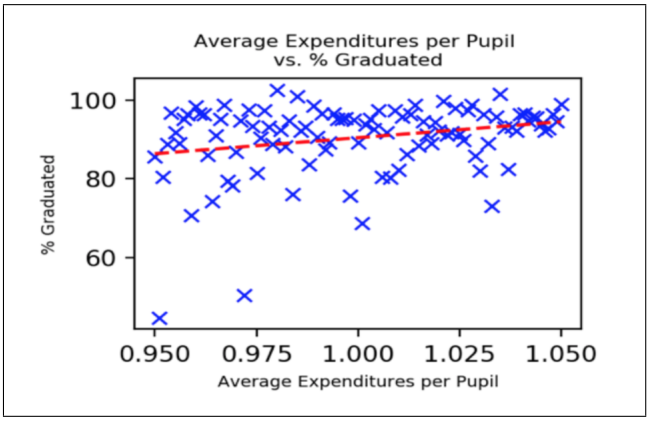

Fig. 2. Causal Inference of Avg. Expenditures per Pupil vs. % Graduated for a single school (Athol High School)

Fig. 2 shows a graphic representation of the causal inference method, where the red line indicates the result of the linear regression. We take the slope of this line as the approximated derivative of the Graduation Rate output metric with respect to Average Expenditures per Pupil about it's initial value for this school.

VI. CONCLUSION

We found that there were substantial gains to the status quo — which used linear regression — to be made in the accuracy of models predicting student outcomes by using more sophisticated models. In particular, tree models — led by XGBoost with Bagging — outperformed all other models. Bagged XGBoost had R^2 values 0.24 higher for predicting college attendance, 0.10 higher for predicting MCAS scores and 0.08 higher for predicting composite SAT scores compared to the linear regression baseline. Extra Trees Regression performed 0.19 better than linear regression in predicting graduation. We hypothesize that these models were higher performing because they better incorporated the interactions between different variables, a key factor in the complicated task of predicting school-wide achievement.

We used these more informative models to perform causal analyses of the effect of slightly changing expenditures per pupil, student teacher ratio, and average teacher salary for all schools in our sample. In a comprehensive literature review performed by Eric Hanushek, 66%, 82% and 73% of studies of these respective effects have statistically insignificant relations [8]. Our analysis supported this near consensus, We also found modest — but statistically insignificant — relationships that were generally in the direction we would expect.

In the future, we would like to incorporate other states' school data into our analysis. Building a framework to normalize test scores and researching laws to ensure that variables were comparable was outside the scope of what we could achieve in this project, but we hypothesize that great gains to accuracy could be made with a larger sample set of schools. We also believe that having year-over-year funding and test result data (which we could not find for the schools in our initial dataset) could provide additional insights into progress made by schools. Further, with more accurate predictive models and more data, we would like to solve the constrained optimization problem of reallocating existing education dollars to achieve whatever education goals the state may have, which we attempted but found intractable given the time constraints of this project. Finally, we could extend this analysis to elementary and middle schools, rather than limiting it to high schools.

CONTRIBUTIONS

Most discussion of project topics, datasets, and goals was between all three of us as a group. However, given each member's unique skills we often worked in complementary manners. In all arenas, different group members assisted their teammates with advice and debugging assistance.

Isaac Kasevich was primarily responsible for creating the data pipeline. He cleaned the inputs, and linked zip code data to specific school information. He created the general framework by which any model we used was trained, tested, and validated. Isaac implemented the PCA analysis and data visualization. He also assisted with using the causal inference framework on a bevy of different models for explanatory-response pairs. Additionally, Isaac explored using convex optimization for budget optimization given our models– which did not make it into our paper.

Zane Kashner created the causal inference framework. He created the process by which we estimate the local effects of small changes of certain inputs on a variety of measures. Zane played a part in the data process, generating new features and with cleaning. Additionally, he found the datasets that we used for both school level and zip code level information. Zane also created the framework for a two staged model linking spending to inputs dependent upon spending — which did not make it into our paper.

Ethan Oro was in charge of implementing a number of models as well as the corresponding hyperparameter optimization. In addition to this role, he was the team member responsible for scraping the zipcode level data. He also headed up the literature review of the existing research in this area. Ethan also investigated integrating Illinois school data into our analysis — which is not yet integrated in our analysis.

REFERENCES

[1] Rob Greenwald, Larry V. Hedges, Richard D. Laine, *The Effect of School Resources on Student Achievement*. University of Chicago, Illinois State Board of Education.

[2] Harold Wenglinsky, *How Money Matters: The Effect of School District Spending on Academic Achievement*. Educational Testing Service.

[3] *Massachusetts Comprehensive Assessment System*. Massachusetts Department of Education.
http://www.doe.mass.edu/mcas/participation.html

[4] *Massachusetts Public Schools Data.*
https://www.kaggle.com/ndalziel/massachusetts-public-schools-data.

[5] *Education data for all Counties in Massachusetts*.
http://www.towncharts.com/Massachusetts/Massachusetts-county-index-Education-data.html.

[6] *Economy data for all Counties in Massachusetts*.
http://www.towncharts.com/Massachusetts/Massachusetts-zipcode-index-Economy-data.html.

[7] James Vaznis, *Students find more awareness with later starts*.
Boston Globe.
https://www.bostonglobe.com/metro/2016/03/09/students-see-benefits-from-later-school-start-times/OOb4vtHm4XZTBLm5X78V9L/story.html.

[8] Eric A. Hanushek, *Assessing the Effects of School Resources on Student Performance: An Update* .
Educational Evaluation and Policy Analysis.
https://journals.sagepub.com/doi/pdf/10.3102/01623737019002141

[9] Iida Hakkinen, Tanja Kirjavainen, Roope Uusitalo *School resources and student achievement revisited: new evidence from panel data*.
https://ac.els-cdn.com/S0272775702000602/1-s2.0-S0272775702000602-main.pdf?_tid=7492da58-3ddb-45de-9644-d3972c8fc657&acdnat=1544678687_ec77e066b3029adfb5f22f0c1d69ffe1

[10] Dennis J. Duggan
Scholastic achievement: its determinants and effects in the education industry.
https://www.nber.org/chapters/c4489.pdf

[11] William Sander
Expenditures and student achievement in Illinois: New evidence
https://www.sciencedirect.com/science/article/pii/0047272793900435

[12] Pedregosa et al.
Scikit-learn: Machine Learning in Python. JMLR 12, pp. 2825-2830, 2011.
https://scikit-learn.org/stable/

[13] Judea Pearl
An Introduction to Causal Inference. Int J Biostat. 2010 Jan 6; 6(2): 7. Published online 2010 Feb 26. doi: 10.2202/1557-4679.1203
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2836213/