Predicting Math and RLA Test Based Achievement from Google Satellite Images Using Deep Learning

Anonymous
Anonymous Institution
anonymous@anonymous.edu

ABSTRACT

This paper presents a preliminary study on using machine learning to predict Math and Reading, Language, and Arts (RLA) test-based achievement of public elementary schools in the United States based on Google satellite images and demographic data. Using a dataset of almost 19,000 schools, we trained a regression-based convolutional neural network (CNN) to predict test scores from the Stanford Education Data Archive (SEDA) database. The predictions moderately correlate with actual test scores. We combined the CNN model with traditional demographic data to test how much added information satellite images provide to conventional predictive models. Adjusted R-squared values increased very minimally with the addition of satellite images as predictors, indicating that information contained within the images correlates with other available predictors. Despite this outcome, the study's results suggest that satellite images of elementary schools can be used as a useful predictor of school achievement in Math and RLA, especially when other predictors are unavailable. Overall, this research paper provides a valuable contribution to the field of educational data mining by demonstrating that ability to predict student achievement in Math and RLA with satellite imagery.

Keywords

School achievement, deep networks, machine learning, educational data mining, satellite images

1. INTRODUCTION

The study of demographic and geographical factors' impact on student achievement has been an interesting topic for a long time [11,13]. For elementary students especially, schools are a major environment where children spend at least six hours a day. In this research paper, we explore the potential of using google satellite images as a predictor for math and RLA test-based achievement. This is an innovative approach as it utilizes a unique data source to identify factors that may influence academic performance. Such factors detectable in satellites could include socio-economic levels, amount of green space and parks, density of houses and roads, architectural features such as playgrounds, physical characteristics such as size and condition, among other features. Such factors have been found to influence academic achievement [1,3,4,9,12]

The goal of this study is to investigate the feasibility and effectiveness of using google satellite images in combination with neural network models to predict math and RLA achievement. We thus aim to provide insights on the potential of this approach in identifying areas that may require educational intervention and to provide valuable insights into how the physical characteristics of a school and its neighborhood may impact student performance. This information can then be used to inform educational policies and allocate future resources more effectively.

2. METHODS

Data was acquired from three different publicly available sources. First, school level test-based achievement data, which we refer to simply as 'test scores', were taken from the Stanford Education Data Archive (SEDA) version 4.1 database [10]. These SEDA test scores are calculated from standardized test scores administered from 3rd to 8th grade in math and reading, language, and arts (RLA) over the 2008-2009 through 2017-2018 school years, and normalized across grades, years, and states using the National Assessment of Educational Progress (NAEP). We specifically used the 'cs_mn_avg_ol' variable which is the school's mean achievement Math and RLA score using an ordinary least squares estimate on the cohort scale. These test scores are approximately normally distributed (Figure 1). We also used SEDA 4.1 school covariates to be used in our models using traditional demographic data. Second, we used data from the National Center for Education Statistics (NCES) for poverty levels in schools' neighborhoods [5] and schools' latitude and longitude [6]. Datasets were merged using the NCES school ID. Third, the latitude and longitude were used as input to the Google Static Maps API to download satellite imagery of each school, each of size 227x227x3 pixels. We used a zoom level of 16 which corresponds to roughly 450x450 m2 which includes the school and a small amount of its surrounding neighborhood (Figure 1). We filtered our combined dataset to include schools which had all available data. As the SEDA dataset was missing values for many western states, we chose to analyze schools within the eastern U.S. (Figure 1). We filtered our dataset to include only elementary schools as school layout and architecture visible in the satellite images are presumably different for elementary, middle, and high schools. Lastly, we filtered for public schools to exclude charter and magnet schools, resulting in a final count of 18,939 schools.

We randomly split the dataset into training (75%), validation (15%), and testing (10%) sets. With the satellite image training data, we trained a convolutional neural network using transfer learning based upon the Xception model [2]. We added a layer of 32 hidden units and an output layer of just one unit which was regressed against the SEDA test scores using a mean-square-error loss function. We used the validation data to optimize the number of training epochs and used the testing data to determine our final outcomes. Next, we combined the 32 hidden units of the CNN with traditional variables from the SEDA school covariates and performed linear regression. We also used a reduced model by using forward subset selection in R's leaps package which finds the variables which minimize the Bayesian information criterion (BIC) of the model. Additionally, we performed linear regression using the SEDA school covariates without the 32 hidden units.1

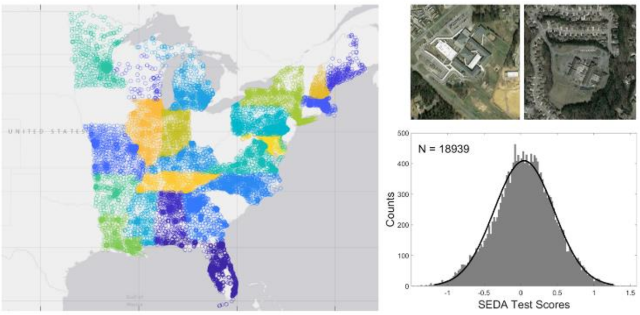

Figure 1. Left: a map depicting the locations of the elementary schools used in the dataset. Upper right: two sample google satellite images. Lower right: a histogram of the test scores fit with a normal distribution.

2.1 Data Availability

Data was acquired from publicly available sources and code for analysis will be made publicly available once results are published. Therefore, all results will be reproducible.

3. RESULTS

Predictions from the convolutional neural network achieved a correlation coefficient of $R = 0.43$ and $R^2 = 0.18$ on a held-out test set (Figure 2). That is 18% of the test score variance could be explained with satellite imagery. By comparison, correlation coefficients of demographic data versus SEDA test scores are listed in Table 1. In particular, the twos strongest correlates are the percent of early childhood development students now at the testing-taking grade levels and the percent of students who are eligible for free or reduced lunch. Other variables with stronger predictive power than satellite images include the income-to-poverty ratio (IPR) of school neighborhoods, the percent of black students, and the percent of white students. However, the satellite images displayed more predictive power than the percent of Asian students, percent of gifted students, percent of limited English proficient students, and the student-to-teacher ratio.

Table 1. List of predictors and their correlations with SEDA test scores, in order or decreasing strength (percent early childhood development students, percent free or reduced lunch, poverty IPR, percent Black students, percent White students, CNN test score predictions, percent Asian students, percent gifted students, percent limited English proficient, student-teacher ratio).

Predictors	% ECD	% FRL	IPR	% Black	% White
R	-0.84	-0.83	0.67	-0.61	0.57
Predictors	CNN Pred	% Asian	% Gifted	% LEP	ST Ratio
R	0.43	0.35	0.30	-0.21	0.03

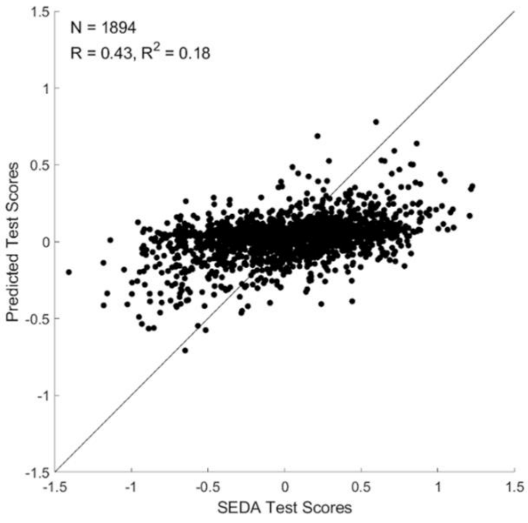

Figure 2. Scatter of SEDA test scores versus those predicted from the CNN using only satellite images in the testing set.

Of the 32 hidden units in the CNN model, only 10 units had nonzero responses. Combining those 10 hidden units with demographic data including early childhood development, free or reduced lunch, poverty, ethnicity, gifted students, limited English proficiency, and student-to-teacher ratio, resulted in predictions with a correlation coefficient of $R = 0.88$ and $R^2 = 0.77$ on a held-out test set. A linear regression model with the demographic data only generated predictions with a correlation coefficient of $R = 0.87$ and $R^2 = 0.76$. We simplified the model using forward subset selection which resulted in 8 variables: IPR, percent Asian students, percent black students, percent ECD students, percent gifted students, and 3 hidden units (numbers 0, 3, 24). This model also generated predictions with a correlation coefficient of $R = 0.87$ and $R^2 = 0.76$. The correlations between these 8 variables are shown in Figure 3.

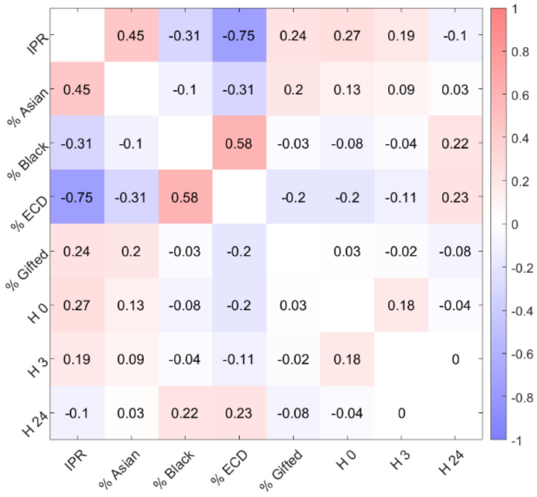

Figure 3. Correlation matrix of variables chosen by subset selection (IPR poverty index, percent Asian students, percent Black students, percent early childhood development students, percent gifted students, CNN hidden unit 0, CNN hidden unit 3, CNN hidden unit 24).

4. DISCUSSION

The results of this study demonstrate that neural networks can moderately predict school achievement scores from a satellite image dataset. However, as in a similar study to predict neighborhood mortality rates from google satellite images [8], the addition of the image features did not significantly improve predictive performance above using demographic data alone. This preliminary research has not yet identified which features in the satellite images are predictive of school SEDA test scores, but this is a major focus of on-going work. Our finding of free or reduced lunch eligibility strongly predicting school performance is consistent with previous studies [7]. We note a couple limitations with this study. First, we only used public elementary schools located in the Eastern part of the United States, which reduces the generalizability of our results. Second, we could have used other satellite information, such as the presence of nighttime light to reflect the socio-economic background of certain counties in the United States or google street view of the schools to identify more proximal predictive features of test achievement. We did train a CNN with satellite images of zoom level 14 which covers a larger portion of the schools' surrounding neighborhoods, but the main results were comparable. To conclude, this research paper provides a valuable contribution to the field of educational data mining and the use of satellite imagery in predicting student achievement in math and RLA.

5. REFERENCES

[1] Browning, M. H., & Rigolon, A. 2019. School green space and its impact on academic performance: A systematic literature review. *International journal of environmental research and public health*, 16(3), 429.

[2] Chollet, F. 2017. Xception: Deep learning with depthwise separable convolutions. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251-1258).

[3] Ferguson, H. B., Bovaird, S., & Mueller, M. P. 2007. The impact of poverty on educational outcomes for children. *Paediatrics & child health*, 12(8), 701-706.

[4] Gershenson, S., & Langbein, L. 2015. The effect of primary school size on academic achievement. *Educational Evaluation and Policy Analysis*, 37(1_suppl), 135S-155S.

[5] Geverdt, D. 2018. *Education Demographic and Geographic Estimates Program (EDGE): School Neighborhood Poverty Estimates* – Documentation. U.S. Department of Education. Washington, DC: National Center for Education Statistics.

[6] Geverdt, D. 2018. *Education Demographic and Geographic Estimates (EDGE) Geocodes: Public Schools and Local Education Agencies*, 2016-2017 (NCES 2018-080). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

[7] Lanham III, J. W. 1999. *Relating building and classroom conditions to student achievement in Virginia's elementary schools* (Doctoral dissertation, Virginia Polytechnic Institute and State University).

[8] Levy, J. J., Lebeaux, R. M., Hoen, A. G., Christensen, B. C., Vaickus, L. J., & MacKenzie, T. A. 2021. Using Satellite Images and Deep Learning to Identify Associations Between County-Level Mortality and Residential Neighborhood Features Proximal to Schools: A Cross-Sectional Study. *Frontiers in public health*, 9.

[9] Misty, L., & Laura, D. T. 2011. The effects of poverty on academic achievement. *Educational Research and Reviews*, 6(7), 522-527.

[10] Reardon, S. F., Fahle, E. M., Ho, A. D., Shear, B. R., Kalogrides, D., Saliba, J., & Kane, T.J. 2022. Stanford Education Data Archive (Version SEDA 2022).

[11] Sutton, A., & Soderstrom, I. 1999. Predicting elementary and secondary school achievement with school-related and demographic factors. *The Journal of Educational Research*, 92(6), 330-338.

[12] Tanner, C. K. 2000. The influence of school architecture on academic achievement. *Journal of educational administration.*

[13] White, J. N. 2001. *Socioeconomic, demographic, attitudinal, and involvement factors associated with math achievement in elementary school*. East Tennessee State University.