Examine Educational Opportunity and Inequality Using Machine Learning Methods with U.S. National Data

Yan Liu[1][0000−0003−3778−1325], Lok Heng Chau[2], and Qiang Hao[3][0000−0001−6361−5035]

[1] Carleton University, Ottawa ON, Canada
yanz.liu@carleton.ca
[2] Paragon Testing Enterprises Inc, Vancouver BC, Canada
lokhengchau@gmail.com
[3] Western Washington University, Bellingham, USA
qiangh@wwu.edu

Abstract. Education opportunity and inequality have been serious concerns in the history of US education. Social economic, ethnic, and racial disparities in academic achievement has been frequently shown in the literature. However, in existing literature researchers investigated these issues in disjointed contexts when using large scale national assessment data because the conventional statistics they adopted only focuses on making inferences from a small number of predictors. Using a large-scale national data, this study aims to predict students' mathematics achievement across 50 states in the US with a total of 74 predictors and over 11,000 school districts. Three machine learning methods, i.e., Random Forests, Lasso Regression, and Genetic Algorithm, were adopted in this study. The results suggest that racial and ethnic proportions in the school district, school related factors (e.g., pupil-teacher ratio, free lunch provided in the school), family socioeconomic status and parent education (e.g., poverty, occupation) are important factors regarding students' achievement in mathematics across the nation.

Keywords: achievement gap, SEDA, machine learning, genetic algorithm, lasso regression, random forests

1 Introduction

The recent research using data from the Programme for International Student Assessment (PISA) suggests that American education is still characterized by racial and socioeconomic inequalities in academic outcomes [1]. Noguera [2] points out that disparities in income and wealth have grown wider and more pronounced than ever before, while the racial and ethnic diversity of American society has been also increasing over the past several decades. About half of the students in American public schools currently are Latino, African American, and Asian [2].

Educational inequality in American public schools becomes an even more prominent issue than ever before given the increasing proportion of racial groups and increasing socioeconomic gaps. There is a great need to provide more empirical evidence to guide educators and policy makers to provide sufficient support to public schools as well as to guide parents to get involved in their children's education efficiently, which helps to ameliorate the learning gaps due to education inequality.

2 Theoretical Framework

2.1 Learning Gaps

Racial and ethnic disparities in children's academic performance have been a long-standing issue in U.S. schooling. The best evidence on racial and ethnic achievement gaps in the US is from the National Assessment of Educational Progress (NAEP) assessments, which have been administered to large, nationally-representative samples of students since 1971. The NAEP assessments have shown the white-black as well as white-Hispanic gaps are still large over the last forty years [3].

Many studies suggest that the minority achievement gaps exist across all socioeconomic levels. Brown-Jeffy [4] reported that a high concentration of Black and Hispanic students led to low mathematics achievement performance across 177 high schools. Olszewski-Kubilius and Clarenbach [5] found that African Americans, Latinos, Native Americans, and English Language Learners had low academic performance at every level of the education system from kindergarten through graduate and professional school. Contrary to people's common belief, the National Center for Education Statistics [3] found that the minority achievement gaps were even larger in higher socioeconomic levels: Black–White and Hispanic–White mathematics achievement score gaps among higher income students increased from 2003 to 2009, while it decreased for lower income students who were eligible for the National School Lunch Program (NSLP).

2.2 Causes of Learning Gaps

Researchers have not reached consensus about the causes of the academic achievement gaps. Some research suggests that students who lack middle-class cultural capital or have limited parental involvement are likely to have lower academic achievement than their peers with better resources [6]. Other studies suggest that race and socioeconomic status are intertwined and shape educational resources and opportunities that affect children's academic achievement [3]. Based on statistics from the National Center for Educational Statistic [7], around half of the African American male students grew up in single-parent households associated with higher poverty and having fewer educational resources. In addition, some other research suggests that engagement in school is more important than other factors to predict students' academic performance [8].

In the past decades, many researchers have tried to identify the independent effects of families, schools, and communities on children's academic performance [9]. However, this body of research has generated a debate on the effects of schools vs. families or other contexts. The debate creates artificial disconnections of these important components, i.e., families, neighborhoods, and schools, because each of these components has both independent and overlapping effects on student achievement. Families choose neighborhoods with a consideration of school choice, whereas schools reflect the composition of the neighborhood [4, 10]. Jackson and Moffitt [9] indicate that it is important for future research to more closely examine the joint effects of families, schools, communities, and other settings on children's academic performance. To ameliorate the gaps in the literature, the present study aims to investigate the joint effects of school and outside-of-school factors on children's achievement performance.

2.3 Educational Research with Big Data

Many international and national assessments have been developed since 1960s. Among of them, PISA, Trends in International Mathematics and Science Study (TIMSS), and Progress in International Reading Literacy Study (PIRLS) are administered to students at either fourth or eighth grades or around this age range across the world. NEAP and Stanford Education Data Archive (SEDA) are national assessments for U.S. students in elementary and junior high schools.

However, most publications or reports that utilized these population data sets adopted inferential statistics approach. First of all, the purpose of inferential statistics is to conduct hypothesis testing to make inferences about a population using sample data with a small number of predictors, which does not make use all the information from the population data. It would be more interesting to use analytical methods, such as machine learning methods, which do not aim to conduct hypothesis testing to make inferences. Second, the results obtained from inferential statistical models may not be reliable because the power gained from the large population size may lead too many covariates to be statistically significant. For large population data, the p-value of 0.05 criteria is not very helpful and thus the findings are less informative. Moreover, the previous studies only examined the disjointed context effects, either schools or families and neighborhoods. That body of research does not provide a full picture to show which factors are important to children's academic performance given that so many factors (e.g., cultural, socioeconomic, school segregation, education resources) are related to children's education.

3 Study Purpose

The purpose of the present study is to provide empirical evidence for addressing two unclear issues in the literature, i.e., the uncertainty of important factors contributing to children's academic performance and the controversial issue around

school segregation. More specifically, this study is to address two research questions: (a) What school and outside of school factors are related to education opportunity and inequality for Grade 3-8 students' mathematics achievement performance across all the school districts in the U.S.? (b) Which variables are more important to predict students' mathematics achievement performance across all the school districts in the U.S.?

This study focuses on variable selection and variable importance ordering instead of conducting inferential statistics. Thus, the study can make use of this population database more efficiently, avoid the limitations described in 2.3 section, and provide a broader view comparing to the existing research by including a great number of diverse predictors in the analyses.

4 Methods

4.1 Data Source

The data were retrieved from Stanford Education Data Archive version 2.1 released in June 2018, which include test scores on mathematics and survey variables for grades 3-8 in almost every public school district in the United States during 2009 - 2015. SEDA is an initiative under the Stanford Center for Education Policy Analysis, which aims to gather data at the national level to help scholars, policymakers, educators, and parents learn how to improve educational opportunities for all children.

About 25 covariates related to ethnic groups were removed because 10 of them have 40%-72% of missingness and 15 of them have 28%-32% of missingness. We did not impute the missing values because these covaiates were related to survey questions designed for subpopulations and some school districts did not have any students from those subpopulations.The missing values are not a serious problem for the remaining covariates: the proportions of missingness across all covariates range from 0% to 4.3%, and half of the covariates only have less than 0.9% of missingness. The missingness can be regarded as missing at random (MAR) in such a case. We imputed missing values via random forests using missRanger R package [11].

In the final analyses, the mathematics scores across grades 3-8 over 6 years (2008/2009 – 2014/2015) for 11,497 school districts in the US were used as the outcome variable. The mathematics scores estimated by ordinary least squares regression are provided by SEDA. A total number of 74 covariates were included in the final analyses. The data includes a wide range of covariates, such as, demographic variables, poverty, race proportions in the school district, and education resources. The details about the test scores and covariates can be found in the SEDA Documentation version 2.1.

4.2 Data Analysis

Three machine learning feature selection methods, genetic algorithm (GA), lasso regression and random forests, were utilized to address our research questions.

Genetic Algorithm Genetic algorithm (GA), a class of evolutionary algorithm, is used to generate high quality solutions to optimization and feature selection inspired by the Darwinian principle of natural (genetic) selection. The algorithm relies on the bio-inspired operators, such as generation, mutation, crossover, and selection [12]. Chatterjee et al. [13] shows that GA can be applied to many statistical problems.

Lasso Regression Lasso regression is a generalization of regression analysis by adding regularization to perform feature or variable selection. Lasso stands for least absolute shrinkage and selection operator. Similar to OLS regression, the lasso regression algorithm tries to find the optimal regression coefficients that can minimize the residual sum squares (RSS) together with a penalty term.

Random Forests Random forests are popular tree-based machine learning methods used for prediction and feature selection. Many studies have suggested that random forests are capable of dealing with missing values, multicollinearity, and high dimensional data, and detecting complex interaction effects [14–18]. To address our research questions, random forests were conducted and a variable importance measure was also computed for identifying which covariates are important for predicting student mathematics performance. We adopted the random forests that was developed by Breiman and his colleagues [4]. The importance measure is based on mean square error with a permutation method.

5 Results

5.1 Genetic Algorithm

In the present study, the GA R package [19] was used for the genetic algorithm (GA) analyses. The analyses were conducted for ten simulated data files, separately. For each imputed data set, 10-fold cross-validation was used to evaluate GA models. The GA models were set up as follows: (1) The initial population size: 1000 (i.e., 1000 of combinations of covariates were randomly generated), (2) Fitness function: the minimum of residual sum squares, (3) Crossover rate: 0.6, (4) Mutation rate: 0.1, and (5) Stopping Criteria: 50 generations.

The final results were combined across ten GA analyses. A covariate is retained if seven out ten votes were obtained across 10-fold cross validation and seven out of ten votes were obtained across ten simulated data files. The left side of Table 1 presents the results obtained from GA method. A total of 26 covariates were selected. The results indicate that racial and ethnic proportions in the school district are important factors to student mathematics achievement performance, including the percentages of Black (perblk), Asian (perasn), and White (perwht). Additionally, percentages of all students in the school district that are in special education programs (perspeced) is also identified, which may represent a minority group of students who need special help.

Several school related factors and education resources were identified as important to students' achievement, including the current expenditures and instructions (ppexp_inst), the average pupil-teacher ratio in the school district (stutch_all), the percent of free lunch in the grade (perfrl), and percent of reduced lunch (perrl).

In addition, several socioeconomic factors were also identified by the GA method, i.e., income at 50th percentile (inc50all) and a few other factors related to income, percent of adults with BA or higher (baplus_all), several variables related to the percent in different occupations (e.g., occserv_all), and a Gini Coefficient for Hispanic population (gini_hsp) that measures the inequality of incomes across individuals.

Table 1: Predictors selected by genetic algorithm method and lasso regression

Lasso Regression		**Genetic Algorithm**	
Predictor	**Coeff**	**Predictor**	**Votes**
perfrl	-	**perfrl**	10
perrl	-	**perrl**	10
perblk	-	**inc50all**	10
ppexp_inst	-	incrat9050all	9
occserv_mal	-	**perwht**	9
perind	-	**ppexp_inst**	9
stutch_hsp	-	**baplus_all**	8
stutch_all	-	ginihsp	8
unemp_mal	-	inc50fem	8
rent_all	-	incrat5010all	8
occsales_fem	-	incrat9010all	8
perspeced	-	**perasn**	8
unemp_all	-	perhsp	8
occtrade_all	-	baplus_hsp	7
unemp_hsp	-	baplus_wht	7
occtrade_fem	-	**educVmalfem**	7
occhealth_fem	+	occeduc_mal	7
ginihsp	+	occsales_fem	7
inlf_all	+	occsales_mal	7
educVmalfem	+	occserv_all	7
occhealth_all	+	occserv_fem	7
occbus_mal	+	**perspeced**	7
inlf_mal	+	pprev_tot	7
inc50all	+	**rent_all**	7
perasn	+	**stutch_all**	7
perwht	+	stutch_wht	7
baplus_all	+		

5.2 Lasso Regression

The glmnet R package [13] was used for lasso regression analysis. The cross-validation is available in this R package by using the function, cv.glmnet. We conducted 10-fold cross-validation lasso regression analyses. A total of 27 covariates were identified by lasso regression. Similar to the results obtained from the GA method, lasso regression identified three groups of covariates, racial and ethnic proportions, school related factors, and socioeconomic related factors. The sign of beta weights for all selected covariates is provided in the last column of Table 1.

Eleven covariates on the bottom have positive relationships and all others have negative relationships with the outcome variable. The percentages of White (perwht) and Asian (perasn) students are positively related to the outcome variable, but the percentages of African Americans (perblk), Native Americans (perind) and students in Special Education programs (perspeced) are negatively related to the outcome. The variables related to the free lunch in the school district are negatively related to the outcome, including percentage of free lunch in the grade (perfrl) and percent of reduced lunch (perrl).

Different from the GA method, lasso regression selected different variables related to school factors, occupation and unemployment status, such as, the average pupil-teacher ratio in the school district for Hispanic population (stutch_hsp), percentage of all students in the district that are in Special Education programs (perspeced), percentage of males in management, business and financial occupation (occbus_mal), percentage of females in health practitioners and technical occupation (occhealth_fem), and unemployment in the district (unemp_all and unemp_mal).

5.3 Random Forests

Different from the other two machine learning methods, random forests analyses can be used for prediction, but also can be used for ranking the relative importance of predictors. This study used randomForest R Package [20]. In order to visualize the results, we plot the importance values using ggplot function from ggplot2 R package [21]. The top 27 covariates were selected to compare with the variables identified Lasso Regression and GA methods. Figure 1 shows the ranking of 27 variables.

Similar to the results of GA and lasso regression, random forests identified three groups of covariates, racial and ethnic proportions, school related factors, and socioeconomic related factors. The percentage of free lunch in the grade (perfrl) the current expenditures and instructions (ppexp_inst), and the ratio of Revenue per pupil to the total Revenue (pprev_tot) are ranked as the top three important covariates. Several variables related to racial and ethnic proportions (perwhite, perblk, perhsp, and perind) were also highly ranked based on their relative importance measures.

Random forests identified similar variables related to occupation and unemployment, such as percentage in poverty (pov_all), percentage of 25-64 year old

females in labor force (infl_fem), percentage of unemployed males (unemp_mal), percentage of adults in poverty (pov_all), and percentage of receiving snap benefits (snap_all).

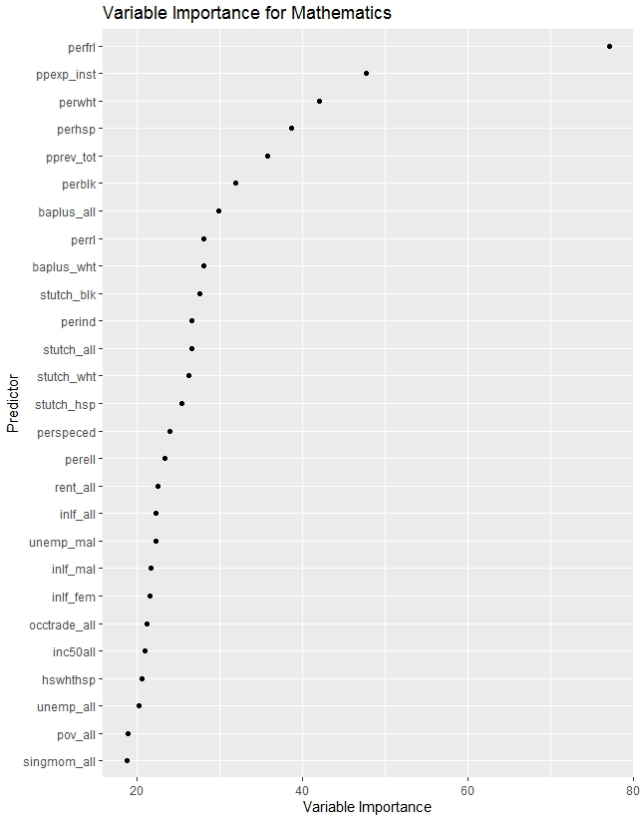

Fig. 1: Variable importance measure for random forests with top 27 covariates selected

6 Discussion and Conclusion

The purpose of this study is to identify what factors are important to predict students' mathematics achievement across all school districts in the US, and to address the controversial issue of school segregation using the recent released educational archive data provided by SEAD. The common findings are that three categories of covariates were identified by all machine learning methods: school related factors and education resources, socioeconomic related factors, and racial and ethnic proportions.

Four factors related to education resources were identified as important to students' academic performance: the percentages of free lunch, reduced lunch, pupil-teacher ratio, and expenditure and instruction index. Although they are school related factors, the percentages of free lunch and reduced lunch are also indicators of the percentage of students from poor families. The results suggest that the higher pupil-teacher ratio may lead to lower achievement performance. However, expenditure and instruction index is found to be negatively related to students' achievement in lasso regression analysis, but the Pearson correlation shows that it has a positive but weak relation ($r = .20$) with the outcome variable. It is not clear how we can interpret this negative relationship. More research is encouraged to look into this issue.

Additionally, the results suggest that socioeconomic factors play an important role in students' achievement performance, which echoed the findings in the existing literature. We found that poverty, education levels, unemployment, and occupations of the population in the school district are highly related to students' school performance, e.g., percentage in poverty, income at 50th percentile, percentage of adults with BA or higher, unemployment in the district.

One of the important findings is that the relationship of the outcome variable and racial and ethnic proportions in the school district are consistent across all machine learning methods. More specifically, the higher percentage of White results in better achievement performance, whereas the higher percentage of Black or Hispanic students is related to lower achievement. The percentage of Asian is positively related to the achievement performance. These findings cannot address whether school segregation is harmful to children's achievement performance directly, but they have shown that the higher proportions of underrepresented subpopulations, the worse the achievement performance the school district has. Therefore, policy makers should keep this in mind when considering the school segregation issue.

One limitation of this study is that we removed some covariates having large proportions of missingness, but these covariates are related to segregation and racial socioeconomic factors. Unfortunately, we cannot include these variables because they are only valid for particular subpopulations. It would be interesting to include those covariates and conduct a separate analysis for each subpopulation.

The contribution of the present study is that the inclusion of a great number of covariates enable us to examine the joint effects of school and outside-of-school factors on children's achievement performance, which provide a full picture to depict what factors play an important role in children's academic performance instead of piecemeal analyses provided in existing research. Additionally, this study introduces machine learning to K-12 educational research. Although machine learning has been widely used in different disciplines, many researchers in K-12 educational research are still not familiar with this approach and may need to understand how this methodology works. This study provides a good example to researchers who are interested in analyzing population data or big educational data.

References

1. Ismael Peña-López et al. *PISA 2015 results (Volume I). Excellence and equity in education*, 2016.
2. Derek Neal. Why has black–white skill convergence stopped? *Handbook of the Economics of Education*, 1:511–576, 2006.
3. Kecia L McCoy. *A study of African American males and their response to online learning*. PhD thesis, Capella University, 2012.
4. Shelly Brown-Jeffy. School effects: Examining the race gap in mathematics achievement. *Journal of African American Studies*, 13(4):388, 2009.
5. Paula Olszewski-Kubilius and Jane Clarenbach. *Unlocking emergent talent: Supporting high achievement of low-income, high-ability students*, 2012.
6. Annette Lareau. Social class differences in family-school relationships: The importance of cultural capital. *Sociology of education*, pages 73–85, 1987.
7. Susan Aud, William Hussar, Grace Kena, Kevin Bianco, Lauren Frohlich, Jana Kemp, and Kim Tahan. The condition of education 2011. nces 2011-033. *National Center for Education Statistics*, 2011.
8. Elizabeth Ackert. Segregation paradox? school racial/ethnic and socioeconomic composition and racial/ethnic differences in engagement. *Social science research*, 70:144–162, 2018.
9. Margot I Jackson and Susan L Moffitt. The state of unequal educational opportunity: Conclusion to the special issue on the coleman report 50 years later. *The Annals of the American Academy of Political and Social Science*, 674(1):281–285, 2017.
10. Robert J Sampson and Patrick Sharkey. Neighborhood selection and the social reproduction of concentrated racial inequality. *Demography*, 45(1):1–29, 2008.
11. Michael Mayer. *Package 'missRanger'*, 2019. R package version 2.1.0.
12. David E Goldberg. Genetic algorithms in search. *Optimization, and Machine-Learning*, 1989.
13. Sangit Chatterjee, Matthew Laudato, and Lucy A Lynch. Genetic algorithms and their statistical applications: an introduction. *Computational Statistics & Data Analysis*, 22(6):633–651, 1996.
14. Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.
15. D Richard Cutler, Thomas C Edwards Jr, Karen H Beard, Adele Cutler, Kyle T Hess, Jacob Gibson, and Joshua J Lawler. Random forests for classification in ecology. *Ecology*, 88(11):2783–2792, 2007.
16. Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3):651–674, 2006.
17. Kathryn L Lunetta, L Brooke Hayward, Jonathan Segal, and Paul Van Eerdewegh. Screening large-scale association study data: exploiting interactions using random forests. *BMC genetics*, 5(1):32, 2004.
18. Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):25, 2007.
19. Luca Scrucca et al. GA: a package for genetic algorithms in R. *Journal of Statistical Software*, 53(4):1–37, 2013.
20. Leo Breiman, Adele Cutler, Andy Liaw, and Matthew Wiener. *Package 'randomForest'*, 2018. R package version 4.6-14.
21. Hadley Wickham. *ggplot2: elegant graphics for data analysis*. springer, 2016.