Identifying Key Factors in Improving Education Inequality

Christopher Tran
University of Illinois at Chicago
Chicago, Illinois
ctran29@uic.edu

Weixin Liu
University of Illinois at Chicago
Chicago, Illinois
wliu53@uic.edu

ABSTRACT

Studies have shown the importance of education attainment and equal distribution of education in helping to make income distribution more equal. Factors such as socioeconomic status families such as parental education and social class play a big role in the performance of children in primary and secondary education. School resources also have a large effect on education. Naturally, schools with more technology are able to offer more resources to their students which affects their learning capabilities. In this study, we wish to look at school information as well as income data and decide what key factors we can gleam that affect school performance. We will use feature selection methods and treat predicting school performance as a classification problem. In this way we can measure the predictive power of selected features to determine what key factors could be looked at in improving education inequality.

1 INTRODUCTION

Education is important in developing future generations. But, even in today's society, there is large amounts of inequality in education. This inequality is largely related to income inequality, since often times poor families do not have access to good education. We wish to study the link between school performance and income geographically, and we wish to discover properties of well performing schools that may suggest ways to decrease this education gap.

Our goal is to find patterns in school and census data that may suggest ways to decrease inequality in the education system. We will be looking at integrated school data from the National Center for Education Statistics (NCES) Common Core of Data (CCD) and the US Census Bureau American Community Survey (ACS) data. The former data, contain administrative data such as number of students and number of teachers. The Department of Education also has The EDFacts Initiative website that contains student proficiency by schools which will help in evaluating and comparing performance. The ACS data are information on average income level of families in each state by geographic location. We wish to link these two datasets to discover important variables that may show ways to improve education.

To link these datasets, we propose a network to capture geographic properties between schools such as average family income, poverty rates, and fiscal properties of different districts. Using student subject proficiency provided to evaluate school performance, we will study the effects of variables that increase or decrease a schools performance level. In this way, we may be able to infer some important properties which can suggest ways to improve education quality in schools that perform poorly. We suggest that average family income plays an important part, but we wish to discover other key factors that affect education.

For our model, we propose a homogeneous network for measuring school performance, where a node represents a school, and links between schools are based geographic distances between schools. Node attributes will be gathered from the CCD and EDFacts dataset to get important administrative attributes to schools, such as number of teachers and students. We also will use reading and math proficiency scores to evaluate school performance. We propose to bin schools into discrete classes based on school performance. Using the ACS data, we can assign average family income per school zone to account for income inequality between different schools. Since much of school funding comes from the state and local (district) level, we believe schools close to each other (such as in the same district) will have similar performance. Also, due to the nature of school funding, we propose to restrict links to schools that are in the same state.

To discover important features we consider the problem of education gaps as a classification problem which classifies schools into performance bin (i.e. very good, good, bad, etc.) and determine features that are important. We will compare traditional methods in machine learning such as random forest models for feature ranking and classification, and compare to network based feature selection and collective classification algorithms. We propose that treating schools independently of location and connections does not provide good results compared to traditional algorithms.

2 RELATED WORK

Education in general is a well studied problem in the social sciences field. From the literature, child achievement can be attributed to a number of factors such as socioeconomic status or resources available at school. Studies have been done on the effect of parental education as well as family income on educational achievement on children [3]. Davis-Kean found that socioeconomic factors play an indirect role to a child's academic achievement, such as parents' belief and behaviors. Parents' education level was also found to be an important factor when looking at school children. School factors also play a big role in Further studies done in the school performance of Australia [7] attribute 5 main factors which affect school performance: previous student attainment, socioeconomic status of the students, school size[1] (based on number of students), location type (rural or urban), and school sector (public, private, or Catholic). Furthermore, studies have show the importance of higher education attainment and equal distribution of education help to make income distribution more equal [1, 5].

Feature importance and selection algorithms are important part in discovering what factors lead to good predictions. Trees are a very simple and interpretable method for feature selection. Random forest has been shown to be an effective method in ranking the

[1] Reports show that as school size falls below 1,000 students, average student attainment falls too.

importance of variables in a natural way [2], and is a very powerful ensemble learning method. Henderson et al. [6] present a feature extraction algorithm that combines local features with neighborhood features to output regional features. An unsupervised feature selection used on networked data was developed by Li et al. [8] that is robust to noisy links. Much work has been done on classifying networked data, and algorithms have shown in cases when nodes are not independent, collective classification performs better than traditional methods [10].

3 DATASET DESCRIPTION

We explore school level data and census data to quantify and measure factors that affect educational attainment. For now we focus on the 2013-2014 school year, but we hope to have a framework for implementing data from multiple reporting years.

3.1 Data Collection

Our dataset comes from multiple government resources which we will discuss in detail. School level information and statistics is provided by the NCES CCD[2]. The data is provided as a means to provide consistent, reliable data to U.S. Department of Education and researchers in addressing education needs [4]. The public school universe data covers the 50 states, the District of Columbia, and five U.S. Island Areas (American Samoa, Guam, the Commonwealth of the Northern Mariana Islands, Puerto Rico, and the U.S. Virgin Islands). There are numerous variables in the data file, so we remove some extraneous information for our predictive purposes and select some variables that we believed to be important. Extraneous features include things like phone number, four digit zip code, and other unimportant identifying factors of schools. Important features include variables such as urban-locale code, number of students, and full time equivalent teachers. In total there were 102,815 observations in the CCD data.

To add more information to get better overall view on school performance we use the ACS[3] data by the U.S. Census Bureau, which provides county level income and poverty data. Family income plays an important part in educational attainment so we link the CCD and ACS data to provide a general view on income and poverty levels per school district.

To assess the performance of schools we use data collected by the U.S. Department of Education EDFacts initiative [4], and use reading/language arts and mathematics achievement to evaluate school performance. The data measures the percentage of students that scored at or above proficient in the given subject. To protect privacy of students, the Department has blurred data based on a tier for number of students per school. Table 1 shows the ranges used percentage of proficient students for the number of students per school. Schools with less than 6 students were suppressed with ’PS’. In total there were 80,126 observations in this data set.

Table 1: Ranges used for Reporting Percentages

Number of Students Reported	Ranges used for Percent Proficient
6-15	<50%, ≥50%
16-30	≤20%, 21-39%, 40-59%, 60-79%, ≥80%
31-60	≤10%, 11-19%, 20-29%, 30-39%, 40-49%, 50-59%, 60-69%, 70-79%, 80-89%, ≥90%
61-300	≤5%, 6-9%, 10-14%, 15-19%, 20-24%, 24-29%, 30-34%, 35-39%, 40-44%, 45-49%, 50-54%, 60-64%, 65-69%, 70-74%, 75-79%, 80-84%, 85-89%, 90-94%, ≥95%
More than 300	≤1%, 2%, . . . , 98%, ≥99%

3.2 Data Processing

We wish to use the datasets described in 3.1 to identify key factors that contribute to school performance. To quantify school performance, we combine the reading/language arts proficiency file with the mathematics proficiency file, and "bin" schools based on percentage of proficient students. We then label the data based on percentage proficiency bins. We decide on 5 classes of schools: "very bad", "bad", "neutral", "good", "very good". Originally, we wished to include the graduation rate dataset provided by EDFacts as part of our school performance evaluation, but the dataset only includes high schools. We believe it is important to look at all public/private schools. To create the labels, we create 5 intervals divided evenly from 0 to 100. The CCD and EDFacts dataset have unique school IDs (NCESSCH) that we can use to link school information with performance into one file. Since the ACS data has data by county, we find school county information from the CCD data file and join income and poverty information by school based on the county location.

Since there were some variables in the files that were not needed, we manually selected variables we thought would affect school performance, and remove extraneous variables. Variables selected are discussed in Section 3.3. Due to the scale and state self-reporting nature of the data, there are bound to be some missing entries. If there are missing entries in a row, we exclude the entry in further steps.

3.3 Variable Selection

There are many extraneous variables provided in the dataset that may not provide any predictive power. For example, in predicting school performance, we do not need identifying information such as school name or phone number. We try to find variables that provide useful information for the school in the CCD dataset. Table 2 shows variables used and a brief description on the variable. We decide to omit some variables such as the number of first grade vs second grade students since proficiency is measured by grade level. This omitting removes many number of variables since number of students per grade level per race was also measured.

4 EXPERIMENTS

We treat our problem as a classification problem for predicting school performance. In treating our problem as a prediction task,

[2] https://nces.ed.gov/ccd/pubschuniv.asp.
[3] https://www.census.gov/programs-surveys/acs/.
[4] https://www2.ed.gov/about/inits/ed/edfacts/data-files/index.html

Table 2: Variables used in prediction

Variable Name	Description
TYPE	NCES school type code. Code ranges from 1-5
ULOCAL	NCES urban locale code. City Large, City Small, Suburb, etc.
FTE	Total full-time teacher equivalent classroom teachers
TITLEISTAT	Which Title I program a school is eligible for
MAGNET	Whether a school is a magnet school
CHARTR	Whether a school is a charter school
SHARED	A school that offers vocational/technical education and attend the school on a part-day basis
FRELCH	Number of students eligible for free lunch
REDLCH	Number of students eligible for reduced lunch
MEMBER	Total number of students
Ethnicity	7 variables. Total number of students of an ethnicity (AM, ASIAN, BLACK, HISP, PACIFIC, WHITE, TR)
VIRTUALSTAT	Whether a school is a virtual school
Income	Median income of the county of the school
Poverty	Poverty percentage for the county of the school

we can discover and rank important features based on predictive power. For our initial results, we use SVM and Random Forest for prediction task and use Random Forest for feature ranking. We implement methods using the scikit-learn package [9].

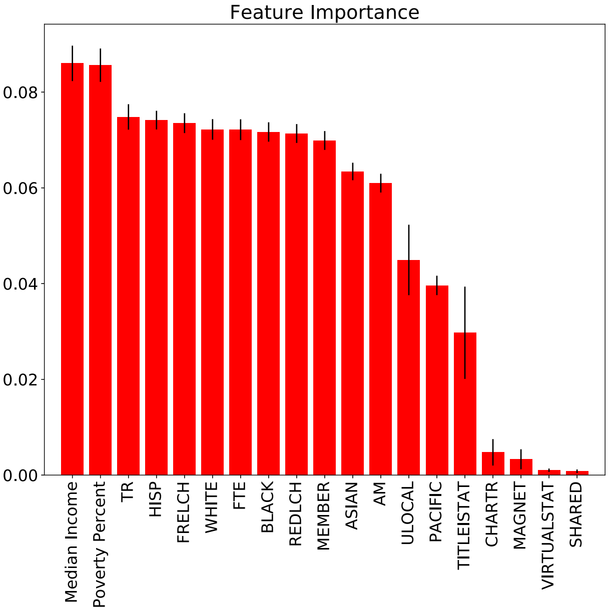
Figure 1 shows the feature ranking using Random Forest. The results show that income and poverty are important features in classification. However, our initial experiments perform poorly in the classification task. For both Random Forest and SVM we only get about 36% accuracy, which too low. We try to adjust the labels, since we believed the binning of the labels is incorrect. Using 3 classes instead of 5, we get some better, but not good results with about 55% accuracy. We discuss the poor results and some ways we propose to alleviate the problems in Section 5.

5 DISCUSSION

Our preliminary results did not show good results and did not give promising direction of the predictive power of any features because of the extremely low accuracy. We discuss potential problems with our steps and model and posit ways to fix these concerns.

- **Feature Selection and Creation**: One problem that may occur is due to the feature selection and creation phase of data processing. We may be missing some important features that have a hand in predicting school performance. Another possible issue is some features are not standardized across different regions. Median income level for example, may not be a good predictive feature since different regions of the country require different levels of income for the same living style. An alleviation to this is to put more emphasis on poverty level and figure out if poverty is a standardized measure across the country.
- **Class Labels**: Another problem is the creation of class labels. As we saw from Table 1, some schools are reported with very low number of students. It might be a good idea to drop these low enrollment schools to lower proficiency ranges to better bin school. Also, it is not clear how to bin schools by performance. One way to counter act this problem is to only focus on "bad" schools and "good" schools and transform this into a binary classification problem. We will perform some analysis on school performance and bin only schools that perform well and perform poorly. We can some statistical measures such as mean and standard deviation to extract these bins, using standard methods such as box-plots to get quartiles, and ignore schools close to the mean/median proficiency levels. This will help alleviate the difficulties in predicting schools on "edge" cases.

Figure 1: Feature importance for the 20 variables



Our next steps are to address the problems stated earlier, and to build a network of schools instead of treating schools independently. In this way we wish to capture geographic properties of related schools in each state. The potential benefits of this is to gain more insight into what kind of factors lead to different performance of schools. Also from this network, we may be able to see some well performing schools located closely in the network with under performing schools. From there we can identify why this particular school is performing well rather than performing similarly to surrounding schools.

REFERENCES

[1] Richard Breen and Inkwan Chung. 2015. Income Inequality and Education. *Sociological Science* 2 (2015).

[2] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[3] Pamela E Davis-Kean. 2005. The influence of parent education and family income on child achievement: the indirect role of parental expectations and the home environment. *Journal of family psychology* 19, 2 (2005), 294.

[4] Mark Glander. [n. d.]. Documentation to the NCES Common Core of Data Public Elementary/Secondary School Universe Survey: School Year 2013âĂŞ14 Provisional Version 2a. ([n. d.]).

[5] Jose De Gregorio and Jong-Wha Lee. 2002. Education and income inequality: new evidence from cross-country data. *Review of income and wealth* 48, 3 (2002), 395–416.

[6] Keith Henderson, Brian Gallagher, Lei Li, Leman Akoglu, Tina Eliassi-Rad, Hanghang Tong, and Christos Faloutsos. 2011. It's who you know: graph mining using recursive structural features. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 663–671.

[7] Stephen Lamb, Russell W Rumberger, Da Jesson, and R Teese. 2004. School performance in Australia: Results from analyses of school effectiveness. Report for the Victorian Department of Premier and Cabinet. *Melbourne, Centre for Post-compulsory Education and Lifelong Learning, University of Melbourne* (2004).

[8] Jundong Li, Xia Hu, Liang Wu, and Huan Liu. 2016. Robust unsupervised feature selection on networked data. In *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, 387–395.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[10] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. 2008. Collective classification in network data. *AI magazine* 29, 3 (2008), 93.