SESIoN: Socio Economic Status Impact on New York State Academic Performance

Dave Dyer
Georgia Institute of Technology
dave.dyer@gatech.edu

Nick Orangio
Georgia Institute of Technology
norangio3@gatech.edu

Ben Spivey
Georgia Institute of Technology
gte146u@gatech.edu

Kshitij Srivastava
Georgia Institute of Technology
ksrivastava34@gatech.edu

Vuong Tran
Georgia Institute of Technology
vtran62@gatech.edu

KEYWORDS

education, socioeconomic status, machine learning

1 INTRODUCTION & OBJECTIVES

The impact of social & economic status (SES) on students' academic performance (AP) is a well-studied area of research. Academic performance is often positively correlated with higher resources/funding [1]and negatively correlated with higher needs, such as free and reduced lunch programs[2], but the magnitude of the correlation is not widely agreed-upon. In this project, we intend to model the impact of SES on AP using New York State School data. We will produce a working model and an interactive visual interface to explore the model interactively. While this is a deeply-studied area, we believe that we can improve on current state of the art models, which typically use fixed-effect & statistical models [3, 4], yet eschew the use of Machine Learning (ML) models or elegant, interactive data visualization. While ML or data vizualization alone may not improve upon the state of the art, we believe that applying both ML and thoughtful visualization design can have an impact on this area of research.

2 PRIOR WORK & CURRENT PRACTICES

The relationship between SES and academic achievement has been explored earlier in Benner et al. [5], Zwick [6] and Battle et al. [7]. All prior work agree in differing measures that SES is an important predictor of academic achievement in combination with other factors like parental involvement [5], sytematic test scores, ethnic group [6] and race [7]. We've also seen that [4–7], most of prior research relies on parametric models where there is an attempt to quantify said relationships by evaluating relative goodness of fits using ANOVA and T-Tests [Winters, Jinnai]. While this is a good practice to understand relative relationships like in [6, 7], it fails to capture random effects and confounding factors. Our work will use a number of machine learning based models to better capture non-linear relationships. Additionally, [6] also has a possible scenario of Simpson's paradox which we will try to stay clear of. Typically, [2] SES is measured by a few different measures – Free & Reduced Lunch Ratio (FRL), family education levels, Occupation, and Income. Academic performance, meanwhile, is typically measured by GPA or some form of standardized testing [2]. We intend to explore the relationship between other, more novel, aggregate measures of both SES and academic performance. Finally, we intend to provide an intuitive and interactive graphic interface to explore the results.

3 APPROACH

Taking the graduation rate data from https://data.nysed.gov/, we intend to use a multiple models to characterize the relationship between academic performance and SES. We will join any of the available relevant data sets and isolate the most important factors that affect academic performance at the school, neighborhood, district, and state level. We will also compare other factors like school types (e.g. charter schools) to socioeconomic factors .

We plan to use measures for academic performance including, but not limited to, aggregate test scores and aggregate graduation rates. Then using D3 javascript library (D3), we will illustrate differences and allow the

stakeholder to explore relationships using our interface. We seek to make correlation relationships, not causal relationships, as the data may not be sufficient to include all latent factors.

We will explore machine learning and statistical methods to find appropriate models. For SES, we are using NRC (at the school / district level) and Perhaps County Historical Employment and Wages Data (at the county aggregate level) as features. We also plan to investigate confounding factors that could affect conclusions on SES, such as school types and teacher education.

We will also explore using fixed effect models as used in economic studies of charter school factors [Winters, Jinnai]. Fixed effects models are linear regression models combining confounding factors, such as student/school characteristics, with factors of interest such as charter school exposure in Winters or charter school penetration in Jinnai. However, for our project, charter schools would be a confounding factor for SES effects. We may need to perform feature engineering to define the factors appropriately as with the public-to-charter move term calculated by Winters. We must also consider nuance as some factors like charter schools may only affect grades levels (e.g., 6th grade) that are overlapping [Jinnai].

4 STAKEHOLDERS

Some surprise beneficiaries of our research are people interested in healthier lifestyles for students, based on J.O. Lee's research [8], which links poor academic performance with worse health outcomes. Additionally, researchers of the relationship between school climate, SES, and Academic performance would benefit; In Berkowitz's study of school climate impact on AP she noted a lack of consistent SES - AP modeling [12]. Last, this project should be of interest to people invested in improving academic performance overall. While school administrators often point to overall improving academic performance [10], there is broad consensus that the SES - AP correlation is mitigated by well-resourced schools (such as charter schools) [1, 10]. If we are successful, all of the above could use our research to help reinforce their arguments for improving resources for lower SES schools.

5 RISKS

In Sirin, et. al.'s Review of Educational Research, one of the primary risks is 'ecological fallacy' [2] – a misinterpretation of the results where one applies aggregate results to individual outcomes. This very much applies to our project, since we work with aggregate data at the school, district, county, and state level. In order to not run afoul of this risk, we will need to clearly indicate, visually, what level we are operating at and reinforce that outcomes may not apply to individuals.

Another risk that Data Scientists should always consider is the so-called 'black box' problem of machine learning models. Gilpin claims that complex machines and algorithms cannot perform interpretation tasks [9]– tasks better left for a human. If our best models happen to be ML models that have many hidden layers (e.g. Deep Learning Convolutional Neural Networks) we run the risk of not being able to explain how the model arrives at its decisions.

6 COST

Based on How to Get a Free Lunch: A Simple Cost Model for Machine Learning Applications a straightforward cost model for ML applications can be applied. This somewhat applies to our project, because while we hope to not incur additional equipment, storage, and compute costs, it is possible that our scope expands to where only distributed computing will be able to handle the calculations. The cost for an ML project is the following:

$$NPV = C_0 + \Sigma \frac{C_t}{(1+r)^t}$$

Where NPV is the net present value, C_0 is the initial cost of equipment, and C_t is the decision cost. Since we are working with freely available data, and we intend to use compute & storage we already own, the NPV of this project should be $0.

7 TIME

In Kushal Singla, et. al's analysis [11] of software engineering for Agile ML projects, they identify that a slight majority of tasks end up in the backlog for ML projects, and there is a huge time variance on how long a story takes. On average, an ML story takes about 18 hours to complete, with an extremely high standard deviation of 31 hours. For our team of 5, we intend

to divide and conquer many tasks with Nick and Dave working more heavily on the documentation, design, organization, presentation, and oversight; Vuong and Ben working more heavily on D3 visualization and python modeling, respectively; and Shitij working on a pretty even blend of all of the above. We intend to adopt an agile approach, and remain flexible as new information / requirements arise. We estimate 1 week for lit survey, proposal, and presentation development. We estimate 5 stories (roughly 90±155h) in a two week sprint for data cleaning, loading, design, EDA, and preliminary modeling; another 5 (90±155h) stories for Statistics/ ML modeling, and 8 stories (144±248h) for D3 visualization, interactivity, testing and iteration.

8 SUCCESS CRITERIA & EVALUATION

At a project level, our Minimum Viable Product (MVP) success criteria are a working, well-researched model and a working interactive data visualization that displays the results. At a model level, success criteria would be a feature selection algorithm that works, with reasonable, explainable features. We intend to evaluate the model using efficacy statistics that will vary, depending on what models we choose to compare. At the bare minimum, we will use p-values to evaluate the efficacy of a regression model that may or may not include feature selection criteria. At the most ambitions level, we will be able to enrich the NYS data with natural experiment data that would help us prove a causal (if aggregate) relationship. There is, obviously, a lot of room in between those two goals for evaluation, and I suspect we'll fall nearer the former than the latter.

We will know if we are successful if we can get data cleaned, model developed, and a Choropleth map working by the end of October as an MVP.

REFERENCES

[1] Yusuke Jinnai *Direct and indirect impact of charter schools' entry on traditional public schools: New evidence from North Carolina* Economic Letters 124, Volume 124 (2014) 452-456

[2] Selcuk R. Sirin *Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research* . Review of Educational Research, Fall 2005, Vol. 75, No. 3, pp. 417–453.

[3] James C. Hearn *Attendance at Higher-Cost Colleges: Ascribed, Socioeconomic, and Academic Influences on Student Enrollment Patterns* Economics of Education Review, Vol. 7, No. 1, pp. 65-76, 1988.

[4] M.S. Farooq, A.H. Chaudhry, M. Shafiq, G. Berhanu *Factors Affecting Students' Quality of Academic Performance: A Case of Secondary School Level* . Journal of Quality and Technology Management, Volume VII, Issue II, December, 2011, pp. 01-04

[5] Aprile D. Benner, Alaina E. Boyle, Sydney Sadler1 *Parental Involvement and Adolescents' Educational Success: The Roles of Prior Achievement and Socioeconomic Status* . Springer Science+Business Media New York 2016

[6] Rebecca Zwick *The Role of Admissions Test Scores, Socioeconomic Status, and High School Grades in Predicting College Achievement.* Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana 2012, 49(2), 23-30

[7] Juan Battle & Michael Lewis *The Increasing Significance of Class: The Relative Effects of Race and Socioeconomic Status on Academic Achievement* . Journal of Poverty, 6:2, 21-35, DOI: 10.1300/J134v06n02_02

[8] J.O. Lee, R. Kosterman, T.M. Jones, T.I. Herrenkohl, I.C. Rhew, R.F. Catalano, J.D. Hawkins *Mechanisms linking high school graduation to health disparities in young adulthood: a longitudinal analysis of the role of health behaviours, psychosocial stressors, and health insurance* . Public Health Vol. 139 (2016) pp. 61-69.

[9] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter and L. Kagal, *Explaining Explanations: An Overview of Interpretability of Machine Learning*, 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), Turin, Italy, 2018, pp. 80-89, doi: 10.1109/DSAA.2018.00018.

[10] Ray Domanico *NYC Student Achievement: What State and National Test Scores Reveal*, The Manhattan Institute, 2020

[11] K. Singla, J. Bose and C. Naik, *Analysis of Software Engineering for Agile Machine Learning Projects* . 2018 15th IEEE India Council International Conference (INDICON), Coimbatore, India, 2018, pp. 1-5, doi: 10.1109/INDICON45594.2018.8987154.

[12] Ruth Berkowitz, Hadass Moore, Ron Avi Astor, Rami Benbenishty *A Research Synthesis of the Associations Between Socioeconomic Background, Inequality, School Climate, and Academic Achievement* . Review of Educational Research April 2017, Vol. 87, No. 2, pp. 425–469 DOI: 10.3102/0034654316669821