Socioconomic Status and the Impact on Academic Performance

Dave Dyer
Georgia Institute of Technology

Nick Orangio
Georgia Institute of Technology

Ben Spivey
Georgia Institute of Technology

Kshitij Srivastava
Georgia Institute of Technology

Vuong Tran
Georgia Institute of Technology

1 INTRODUCTION & OBJECTIVES (HQ1)

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, but the magnitude of the correlation is not widely agreed-upon[2]. 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 a visual interface to explore the model interactively. Like many topics with sufficient data, this area is well studied. We seek to improve on status quo by investigating additional confounding factors (e.g., teacher education), evaluating Machine Learning (ML) models compared to state of the art fixed-effect & classical statistic-based models [3], and creating insightful, interactive data visualizations. Applying both ML and thoughtful visualization design should have an impact on this area of research as we aim to improve on the state of the art.

2 PRIOR WORK & CURRENT PRACTICES (HQ2)

The relationship between SES and academic achievement has been demonstrated earlier in Benner et al. [5], Zwick [6], Farooq [4], 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], parental education level[4], systematic test scores[2], 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. While this is a good practice to understand relative relationships, it fails to capture random effects and confounding factors as in [6, 7]. We learned that linear fixed effects models are commonly used to account for confounding factors for educational studies [1, 13, 14], but these may not account for any non-linear relationships as a machine learning model would. Typically, SES is measured by a few different measures – Free & Reduced Lunch Ratio (FRL)[2], family education levels[4], 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.

3 APPROACH (HQ3)

From the graduation rate dataset (which can be found at at https://data.nysed.gov/), we intend to test multiple co-variates including lurking variables at the school [5], neighborhood, district, and state level, in order to characterize what constitutes SES and academic performance at various levels of aggregation. We confirm that our dataset is large as it has thousands of rows in total as we include multiple years. We'll use a graphing library (D3) to allow the user to explore various segments of data in both a univariate and bivariate manner. We seek to find correlations, not causal relationships, as the data may not be sufficient to include all latent factors. We will explore both machine learning and traditional statistical learning algorithms to characterize relationships between SES and academic performance at differing levels of aggregation. 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 schools effect [1, 13] and teacher qualification effect [14]. We may need to perform feature engineering to define the factors appropriately and must consider nuance as some factors like charter schools may only affect grades levels (e.g., 6th grade) that are overlapping [1].

4 STAKEHOLDERS AND IMPACT (HQ4 & HQ5)

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 (HQ6)

In Sirin, et. al.'s Review of Educational Research, a big 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 clearly indicate, visually, what level we are operating at and reinforce that outcomes may not apply to individuals. Zwick also has a possible scenario of Simpson's paradox which we will try to avoid.[6] Another risk is the 'black box' problem of machine learning models. Gilpin claims that complex machines and algorithms cannot perform interpretation tasks.[9] If our best models happen to be ML models we run the risk of not being able to explain how the model arrives at its decisions.

6 COST HQ7

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. [15] This somewhat applies to our project, because while we hope to use our own equipment, 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 (HQ8)

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 working more heavily on D3 visualization and python modeling, respectively; and Kshitij and Ben 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 (HQ9)

At a project level, our Minimum Viable Product (MVP) is a working, well-researched model and a working interactive data exploratory visualization tool. We intend to evaluate the model using metrics that will vary, depending on what models we choose to compare. At the bare minimum, we will use Area under the ROC Curve or F1 Score for the model [16] . 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] Jinnai, Y. (2014). "Direct and indirect impact of charter schools' entry on traditional public schools: New evidence from North Carolina." *Economic Letters*, 124, 452-456.

[2] Sirin, S. R. (2005). "Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research." *Review of Educational Research*, 75(3), 417–453.

[3] Hearn, J. C. (1988). "Attendance at Higher-Cost Colleges: Ascribed, Socioeconomic, and Academic Influences on Student Enrollment Patterns." *Economics of Education Review*,7(1), 65-76.

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

[5] Benner, A. D., Boyle, A. E., Sadler, S. (2016). "Parental Involvement and Adolescents' Educational Success: The Roles of Prior Achievement and Socioeconomic Status." *Journal of Youth and Adolesence*, 45, 1053-1064.

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

[7] Battle, J., Lewis, M. (2008). "The Increasing Significance of Class: The Relative Effects of Race and Socioeconomic Status on Academic Achievement." *Journal of Poverty*, 6(2), 21-35.

[8] Lee, J. O., Kosterman, R., Jones, T.M., Herrenkohl, T.I., Rhew, I.C., Catalano, R.F., Hawkins, J.D. (2016) "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*, 139, 61-69.

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

[10] Domanico, R. (2020). "NYC Student Achievement: What State and National Test Scores Reveal." *The Manhattan Institute for Policy Research* https://eric.ed.gov/?id=ED604331.

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

[12] Berkowitz, R., Moore, H., Astor, R.A., Benbenishty, R. (2017). "A Research Synthesis of the Associations Between Socioeconomic Background, Inequality, School Climate, and Academic Achievement" *Review of Educational Research*, 87(2), 425–469.

[13] Winters, M. (2012). "Measuring the effect of charter schools on public school student achievement in an urban environment: Evidence from New York City." *Economics of Education Review*, 31, 293-301.

[14] Harris, D.N., Sass, T.R. (2011). "Teacher training, teacher quality, and student achievement." *Journal of Public Economics*, 95(7-8), 798-812.

[15] Domingos, P. (1998). *How to Get a Free Lunch: A Simple Cost Model for Machine Learning Applications.* 15th National Conference Artificial Intelligence (AAAI-98) and International Conference on Machine Learning (ICML-98): Workshop on AI Approaches to Time Series Problems.

[16] James, G., Witten, D., Hastie, T., Tibshirani, R. (2017). *Introduction to Statistical Learning: with Applications in R.* Springer. 147, 354-355.