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

The impact on academic performance of students' social & economic status is a deeply-studied area of research, with many esteemed institutions performing research on the topic. Academic performance in schools with fewer resources – for instance, funding – or higher needs – such as free and reduced lunch programs – is almost always lower in scientific studies that examine this relationship [Sirin Citation]. In this project, we intend to improve on previous research through the thoughtful application of machine learning and data visualization to the question of how socioeconomic status (SES) is related to academic performance. By itself, the application of machine learning will likely not improve on state of the art of this field of study – likewise with data visualization alone. However, at the nexus of machine learning and data visualization, there is ample opportunity to improve the insight into the relationship between SES and academic performance.

2 OBJECTIVES

We intend to model the impact of SES on academic performance using a novel approach and to produce a delightful visual interface to interactively explore the relationship between these two things. Along the way, we intend to examine what SES factors have the most influence on academic performance. We also intend to examine novel measures for academic performance. We feel that the state of the art is somewhat lacking and that we can provide a better understanding of these factors through applied machine learning and thoughtful interactive data visualization.

3 CURRENT PRACTICES

Per [Okioga, Farooq, Hearn] the current models are statistics-based, with ANOVA and T-Test applied to academic performance output. Typically, [Sirin] 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 [Sirin]. Many of the models used are statistical in nature – featuring ANOVA, T-Test, and classic regression with fixed effect models [insert Ben's papers here.]. We intend to explore the relationship between other, more novel, aggregate measures of both SES and academic performance. We also intend to try multiple statistical and machine learning models to see if the model results themselves can be improved upon. Finally, we intend to provide an intuitive, pleasant interactive graphic interface with which to explore the results.

4 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 merge whatever data sets we deem to be reliable and useful and isolate the most important factors that affect academic performance at the school, neighborhood, district, and state level. We will also examine measures for academic performance including, but not limited to, aggregate test scores and aggregate graduation rates. Then using modern visualization techniques, we will illustrate the difference and allow the stakeholder to explore this relationship using our interface. We will explore various machine learning and statistical methods to find the best model. For SES, we are using NRC (at the school / district level) and Perhaps County Historical Employment and Wages Data (at the county aggregate level). We will also explore using fixed effect models [insert Ben papers here]

5 STAKEHOLDERS

This research should apply to everyone in America who cares about the economy [citation], health [J.O. Lee citation], and the impacts of income disparity on an increasingly failing education system [Domanico]. Particularly, this research should be of interest to parents, teachers, school administrators, and state representatives who make financial decisions regarding which institutions get how much budget on any given year. If our work helps inform academic policy, then it could impact future students' ability to graduate and/or be successful in academic pursuits.

6 RISKS

Since our data reflect aggregate SES and Graduation Rate statistics, a big inference risk is that of 'ecological fallacy' [Sirin - Review of Educational Research] – a misinterpretation of the results that apply aggregate results to individual outcomes. Also, whenever talking about SES, it is important to be thoughtful in our approach to race, since ML Models that associate SES and academic performance are often biased and racist, not

taking into account confounding variables - reference 0.5 article on racist AI. There is also the interpretability risk of so-called 'black box' ML models being difficult to explain clearly. Last, confounding variables as a result of suppressed data (s).

7 COST

Based on How to Get a Free Lunch: A Simple Cost Model for Machine Learning Applications the cost to data ratio is 0.

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

NPV is the net present value.

C0 is the initial cost of equipment, since we will be working with systems we already own and data is freely available, there is no initial cost.

Ct is the decision cost, which is also zero since we won't be making changes to the system or data we are using.

8 TIME

From Lessons from My First Two Years of AI Research, time will depend on many factors. This is a breakdown of a sample timeline:

* Reading/Research: 2 weeks - reading and understanding past research will equip us with a better understanding of the subject. Understanding past research shortcoming and limitations can help us find novel approaches.

* Conversation/Videos/Conference talks 1 week - Any knowledge gaps should be supplemented by having a conversation or other forms of interactions with a subject matter expert.

* Data Analysis 2 weeks - After knowing all the models before and their limitations, we can start modeling our improved version.

* Visualization 1 weeks - Once Analysis is complete, come up with visualizations that allow users to quickly understand our findings

* Final Paper 2 weeks - Wrap everything up into a final paper

We should keep detailed notes so that we are able to quickly reference back. Always track measurable progress for our progress report. Timeline for each stage adjusted to fit our schedule and deliverables.

9 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.

10 CITATIONS AND BIBLIOGRAPHIES

11 APPENDICES