CS 109: (Optional) Contest Submission
Prepared by Andrea Pasinetti
Stanford email: apasinet@stanford.edu
Submitted: Friday, June 12th, 2020

YouTube link: https://youtu.be/xZqnr7iD2iU

Using Data on Student-Teacher Ratios and School Capital Investments to Predict Educational Outcomes for Middle and High School Children in New York State

Two of the most hotly contested policy questions surrounding k-12 education revolve around the impact of small classrooms (i.e. low student-teacher ratios) and the importance of well-equipped school facilities on learning outcomes for students. Many contradictory studies have been published on both sides of the argument, and much philanthropic capital has been deployed in the hopes that smaller, state-of-the-art classrooms might be a silver-bullet solution for communities beset by chronically poor academic outcomes.

For more background reading on the topic:

- Brookings Institute
- Goldman School of Public Policy, UC Berkeley
- Class Size Matters: a non-profit organization focused on researching this very issue

The goal of this project was to see how effective Naïve Bayes might be in making predictions about educational outcomes given knowledge about student teacher ratios and the quality of school facilities.

Unlike many of the assignments submitted for class, the primary challenge with this undertaking was identifying and preparing data culled from publicly available sources. I began this process operating under the assumption that data available through the OECD would be suited to the task, but discovered that many of their datasets are spotty and outdated. My own lack of familiarity with handling different data formats also proved an obstacle.

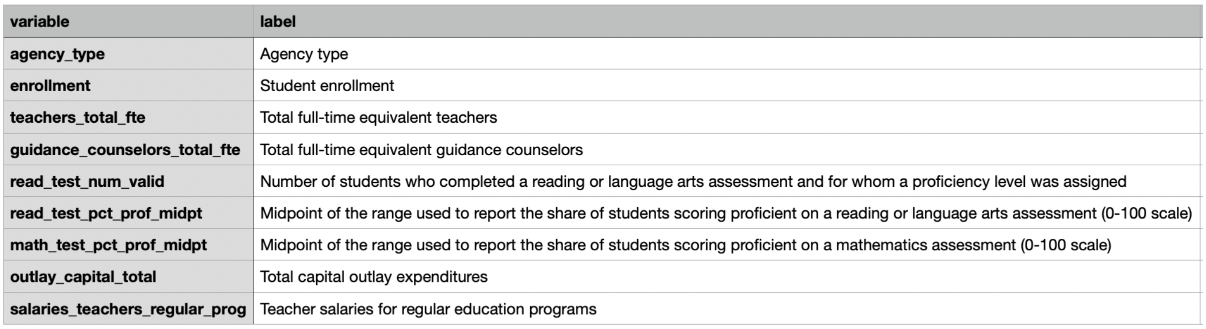
Upon further research, I discovered the Urban Institute's (TUI) repository of education related data for the United States, as well as its easy to use csv generator, which returns a customised dataset populated with categories and timeframes specified by the user.

By way of background, the Urban Institute was founded by Lyndon Johnson as a non-partisan, American think tank with the mandate to evaluate the impact of the President's policies and programs under the Great Society umbrella; it has evolved into an organization that performs "social policy research to 'open minds, shape decisions, and offer solutions.'"[1]

TUI's data tools are incredibly convenient and powerful and I was able to download a dataset with information about New York State schools, recent as of 2017. The website also

[1] https://en.wikipedia.org/wiki/Urban_Institute

produced a helpful *data dictionary* with information about the data categories and variables included in the dataset. I have produced these below for convenience:



variable	label
agency_type	Agency type
enrollment	Student enrollment
teachers_total_fte	Total full-time equivalent teachers
guidance_counselors_total_fte	Total full-time equivalent guidance counselors
read_test_num_valid	Number of students who completed a reading or language arts assessment and for whom a proficiency level was assigned
read_test_pct_prof_midpt	Midpoint of the range used to report the share of students scoring proficient on a reading or language arts assessment (0-100 scale)
math_test_pct_prof_midpt	Midpoint of the range used to report the share of students scoring proficient on a mathematics assessment (0-100 scale)
outlay_capital_total	Total capital outlay expenditures
salaries_teachers_regular_prog	Teacher salaries for regular education programs

Unfortunately, many of these categories included entries labelled 'missing' or 'Not applicable,' giving little indication as to what the missing data might suggest about the institution it corresponds to. Specifically, I wasn't able to ascertain whether the schools had been closed, or consolidated, or whether certain districts simply didn't publish information in particular categories.

As a result, while the data available was plentiful, I decided to reduce it to about 780 entries that were intact across all the categories that I chose to evaluate.

The specific question I sought to address, as noted in the introduction to this brief, was the extent to which the quality of school facilities and teacher-student ratios have an impact on educational outcomes. While student-teacher ratios are a relatively objective measure, the quality of school facilities and educational outcomes require the use of proxy metrics.

To evaluate educational outcomes, I therefore decided to evaluate the mid-point of the percentage range of students in a school that received a proficient score on the core curriculum reading test.

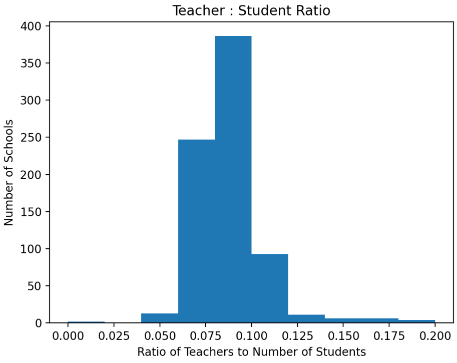
To evaluate the quality of facilities I decided to look at total capital outlay expenditures for a school – a proxy often used in evaluating a business's investment in upgrading and growing its own capabilities.

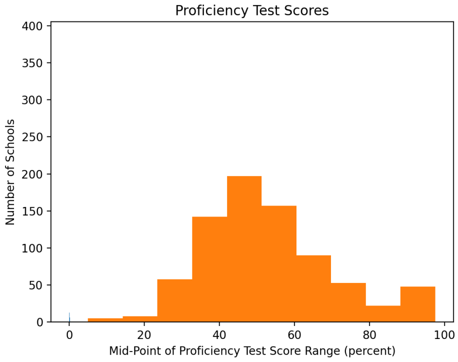
Naïve Bayes Model

Given my lack of familiarity with creating classes in Python, I decided to execute my calculations in a single Pycharm .py file, without creating standalone functions. This means that my calculations for this project cannot easily be extended to other data files / variables without extensive adjustments to the code.

A significant portion of the coding required for this project was data cleaning, and the conversion of continuous data to binary data suited to Naïve Bayes prediction models. Specifically, I generated an array of tuples containing binary data for student-teacher ratios and total capital outlay expenditures per school.

The process of determining demarcations for the binary cut-off was a bit arbitrary; I therefore used numpy's histogram feature to generate histograms for student teacher ratios as well as reading proficiency levels by school. I am reproducing these below for convenience:





I decided to set student-teacher ratio values equal to 1 if they exceeded 0.085, or rather one teacher per ~11.8 students. This is consistent with what appears to a delineator in the data; interestingly 1:12 is identified in many studies as a tipping point beyond which student learning and classroom management begins to deteriorate.

I inspected the data for ranges of capital expenditure by school, and set $1.5M as a demarcation point for a label of 1, corresponding to 'significant investment' in facilities, vs 0, for 'under-investment' in facilities.

Finally, I used 50% to demarcate reading proficiency, with values greater than or equal to 50% corresponding to 1, and all other values corresponding to 0. This approach harmonized well with the data, which looks consistent with a normal distribution.

Approach

The code used to fit the data to a prediction model, as well as to make predictions about my test data was similar, if not identical to the code I wrote for assignment 6; I decided to use a maximum a posteriori estimate to account for cases where data may have been missing.

Of my 780 data points, I used 500 data points to train my model, and 280 for testing purposes.

Conclusion

As a conclusion for this little research project I decided to return a precision score for the Naïve Bayes model, as an indicator of how well my prediction model performed in forecasting educational outcomes (in the very unscientific way they are described above), given available data on capital investment and student teacher ratios.

To my surprise the model was successful ~54% of the time, which exceeded my expectations! Of course, the sample size concerned here is quite small. Materially, it's also unclear what data I ended up omitting by removing schools with no reported values in these categories from my analysis. Nonetheless, this relatively high rate of predictive success suggests that classroom size and facilities seem to have some relationship to student outcomes. Sadly, they probably also coincide with the relative income levels of the communities in which these schools are situated – a factor which is perhaps the clearest determinant of educational outcomes for children.

There remains much work left to be done to ensure that all children have access to a quality education.