Quality of elementery school education in Chicago and socioeconomic factors

Alexandru Lopotenco

December 20, 2021

Contents

1 Executive Summary

Problem. Understanding the limitations and problems of the public school system is crucial in assessing how we can improve the education quality in a certain region or even at a national level in US. Hence, my final project focuses on analyzing how certain factors impact the quality of education in public schools from the Chicago area. Even though analyzing how some direct educational metrics like grades or post-graduation outcomes will definitely tell how to asses the quality of education in a schools, those statistics are highly correlated and one could achieve trivial results by studying the relationship between results of a school and its quality of education. Hence, I incorporate socioeconomic metrics and explore how those affect the outcomes of schools.

Data. My data set is pulled from one source - the Chicago data portal. A part of the data comes directly from the city education department and it portrays various scores that describe the quality of particular schools, but also some more specific metrics that focus on the outcomes and results of the education process in those schools. To connect this data with socioeconomical factors, I used the census data of Chicago split on community areas and some health data. My socioeconomical data comprises obvious indicators such as yearly income, but also metrics more related to education such as the percentage of populations with certain level of degrees and also some societal metrics such as teenage births. My primary response variable was an aggregate score that is the mean of three other scores which asses safety, instruction and environment quality based on surveys taken by students.

Analysis. The first step in performing the analysis was splitting the data in a testing and training sub-datasets. Given the simplicity of the OLS and the rather inter-connected factors used in prediction, I also opted for more sophisticated regression techniques that would be helpful in determining the more relevant explanatory variables. I built four different cross-validated models: OLS, ridge regression, LASSO regression, decision trees and random forest.

Conclusions. Surprisingly, the hardship index seems to play a small role in predicting the aggregate score as per the results of all the models - only Ridhe identified it as a top factor. Indeed, factors such as misconduct and attendance have been found to be leading forces in predicting aggregate scores, which is not surprising, but shows that those non-socioeconomical factors have a stronger relevance. I hope this analysis could help public schools in enforcing certain policies and authorities to track the principal reasons of inferior education quality.

2 Introduction

Background. Analyzing the relation between elementary school education and other aspects of our society can be fruitful is finding the weaker and stronger points of the educational system in Chicago. It is not surprising that education is impacted by a plethora of factors. Schools usually report certain objective quantitative metrics as well as surveys done by their students to who assess the perceived quality of their education. Those sources provide us with insight for the quality of education one receives, and hence it would be fruitful to try and predict how one would rate their education given the circumstances they live in. Furthermore, there is a lot of socio-economical and healthcare factors that indeed might affect the quality of education one receives, and this can be observed in day-to-day life by noticing the divide between public - private education, or the differences of the quality of education in certain regions.

Analysis goals. Knowing that socio-economic factors impact every aspect of our lives, I am analyzing which metrics specifically have the highest relevance when talking about elementary school public education. Alongside this, I analyze how some more directly related statistics such as misconduct and attendance relate to overall quality of education in order to assess an approximate proportion of the influence of socio-economic factors on the public elementary schools system from Chicago.

Significance. I hope that this analysis will provide fruitful insight for education legislators and school administrations as to what are the principal reasons that might stagnate the education process in a school. Furthermore, aside from describing the hard-to-change socio-economic factors that influence education, I present whether more direct qualitative factors about education play a bigger role in final quality outcomes.

3 Data

3.1 Data sources

My dataset is merged from three sources: the 2008 - 2012 selected census data for specific community areas in Chicago, the 2011 - 2012 public schools progress in Chicago and historical health data. The datasets come from the Chicago Data Portal. The data regarding the socioeconomic factors was pulled from https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2/data. In order to ensure a wide variety of explanatory variables in our dataset, we include all the explanatory variables in it.

The other dataset I used is pulled from https://data.cityofchicago.org/Education/Chicago-Public-Schools-Progress-Report-Cards-2011-/9xs2-f89t and contains multiple metrics that assess the progress of high schools between 2011 and 2012, but we do not really use those metrics in our analysis since we focus on the more stable statistics concerning quality scores and attendance/misconduct.

The final data I used is pulled from https://data.cityofchicago.org/Health-Human-Services/Public-Health-Statistics-Selected-public-health-in/iqnk-2tcu and contains multiple social metrics and health metrics which present crucial insight in the particularities of Chicago community areas.

3.2 Data cleaning

The two most important tasks of cleaning data was selecting the relevant metrics from the educational dataset and merging the three pulled files. To merge them, I first joined the health and census data by community area name, and then used community are number, to connect the school data. Selecting the useful metrics from the schools and health dataset has been particularly difficult and a lot of the data has been disposed given the nature of certain values, and the fact that most of them have been undereported. ## Data description

3.2.1 Observations

The initial dataset has 566 observations and the cleaned one has slightly above 400

3.2.2 Response Variable

The response variable is the aggregate score, which is the mean of other three surveyed metrics, instruction, safety and environment scores. Those surveys have been reported by the students.

3.2.3 Features

Drawing on data regarding the socio-economic status, health history and public schools report, I included 15 explanatory variables in my models.

3.3 Data allocation

Before building my predictive models, I first removed observations from the dataset for which any variables had NA/NDA values. This has been done for consistency purposes as a lot of schools udnerreported their metrics. The I split the dataset into two subsets: a training dataset used for building the predictive models and a test dataset for evaluating the models. I used an 80-20 split, in favor for the training dataset. I used the same seed for everytime I need randomize certain splits or other types of actions.

3.4 Data exploration

3.4.1 Response

We first sought to understand the response variable's distribution. As seen in the histogram aggregate scores (Figure 1), the data appears to be right-skewed.The median aggregate score is around 0.477.

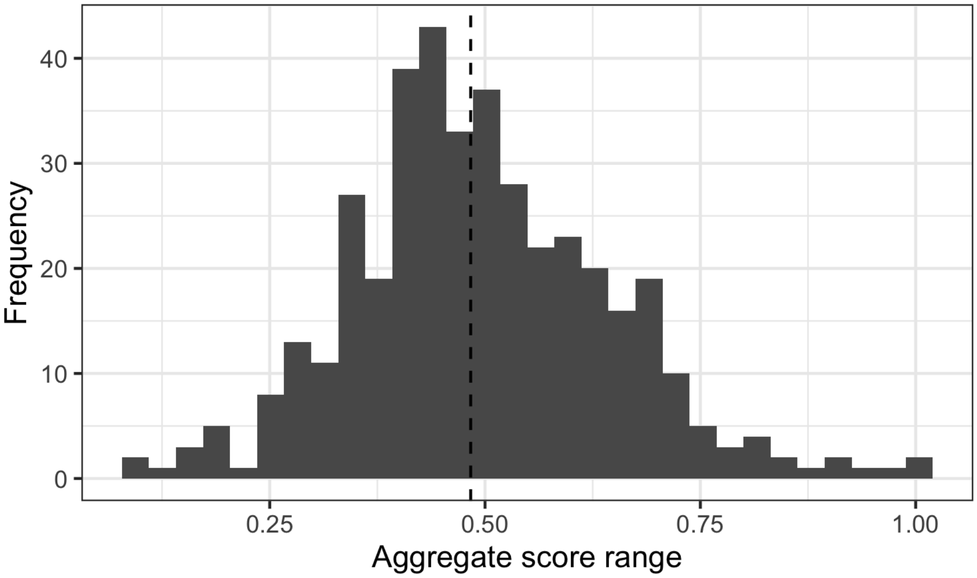

Figure 1: Distribution of aggregate scores.

Next, we group by community areas analyze the means for each of them in particular to see if there are trends between better-off communities and education quality scores. First, we print the top 10 communities by aggregate scores together with their socio economical data. As we can see, there isn't a clear very strong relationship between the scores and the two chosen socio-economical factors, except that the schools in middle for both indices seem to be more prevalent in the top 10.

Table 1: Top ten communities by aggregate score

Community	Score	Per-capita income	Hardship index
Forest Glen	0.76	44164	11
North Center	0.68	57123	6
Mount Greenwood	0.68	34381	16
Edison Park	0.65	40959	8
Clearing	0.64	25113	29
Norwood Park	0.64	32875	21
Lincoln Park	0.64	71551	2
North Park	0.63	26576	33
West Ridge	0.62	23040	46
Hermosa	0.62	15089	71

To explore this relationship further, we create two scatterplots of score against hardship index and per-capita income to asses a broader distribution. We notice a clear, yet not very trend in both of the scatterplots of figure (Figure 2) - as income tends to rise, mean aggregate scores of communities increases abruptly in the lower ends of income distribution and then slightly decreases. For hardship index, it seems that as it approaches 100 the aggregate score slowly, but very monotonically decreases.

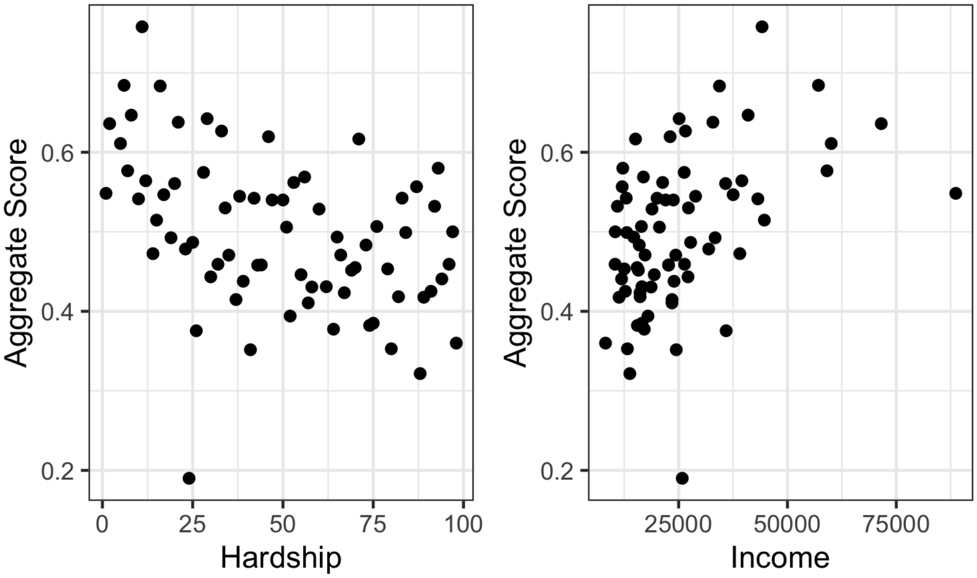

Figure 2: Scatterplots of aggregate score.

3.4.2 Features

To illustrate the relationship between the features we use a heatmap. Unsurprisingly, a good chunk of the socio-economic factors are highly-correlated which leads us to believe that those factors will be crucial in the construction of the models. Furthermore, the abscence of correlation between healthcare and socio-economic factors is in itself also surprising.

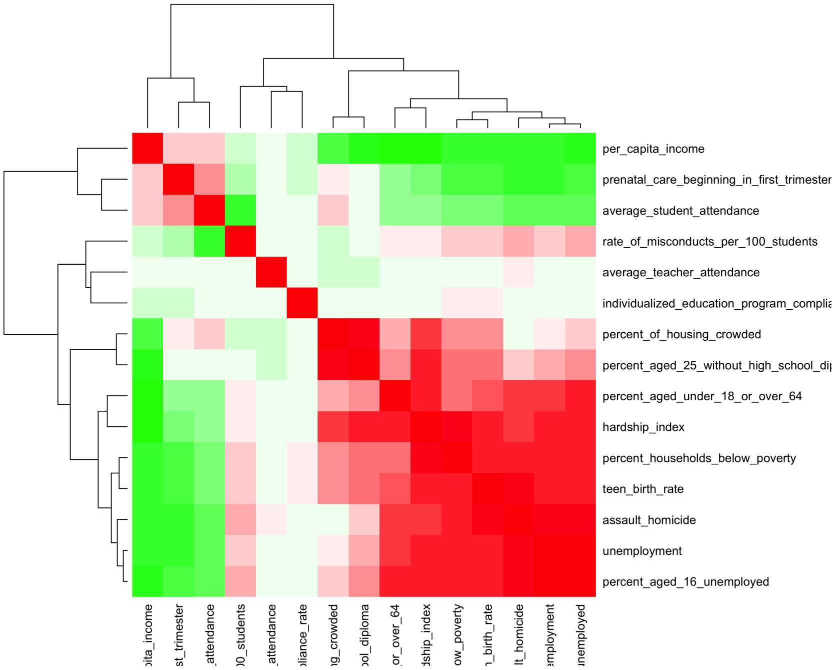

Figure 3: Heatmap.

4 Modeling

4.1 Regression-based methods

4.1.1 Ordinary least squares

The OLS revealed the following factors as having significat explanatory with level 0.05: per_capita_income, average_student_attendance, percent_aged_16_unemployed and unemployment. ### Penalized regression

Since the OLS seems to show that a good chunk of predictors are not very significant, I decided to use LASSO and ridge in order to see whether those factors would still remain insignificant under some penalties, or if adding a penalty would actually shift the explanatory relevance towards other variables.

For the lasso, Figure 4 shows the CV plot, Figure 5 shows the trace plot, and Table 2 shows the selected features and their coefficients. LASSO selected 3 variables according to the 1se rule and 14 as the optimal number without adjusting to errors.

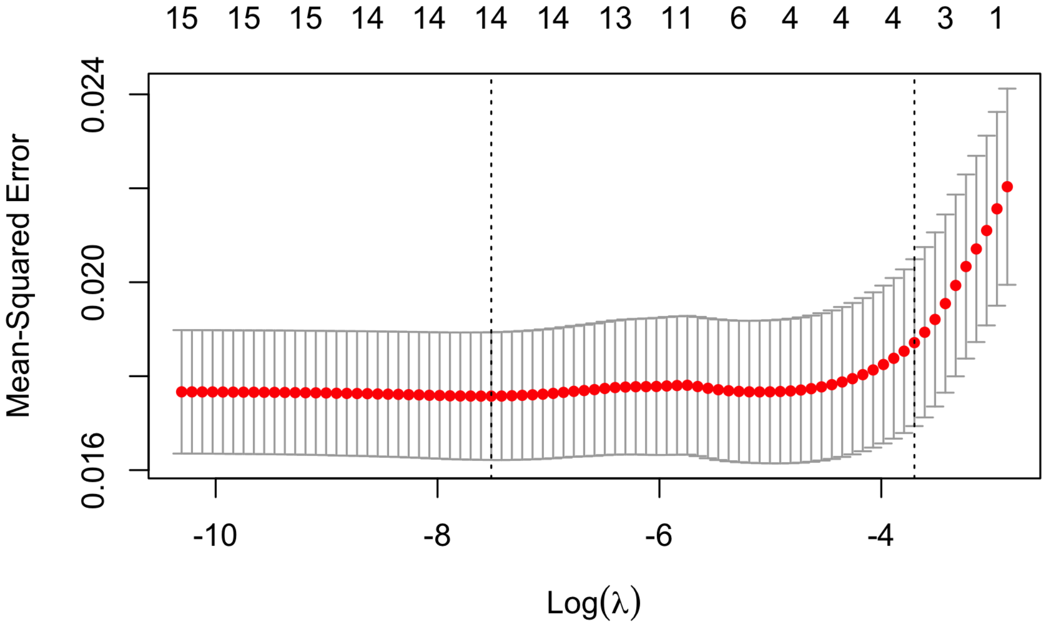

Figure 4: Lasso CV plot.

It looks like both LASSO and ridge agree on most of the top 6 features, which seem to be slightly different from the OLS selected features. In particular they agree on student attendance, rate of misconduct, per capita income, and the percent of individuals aged 25 without a high school degree. Below are attached the trace plots of both models

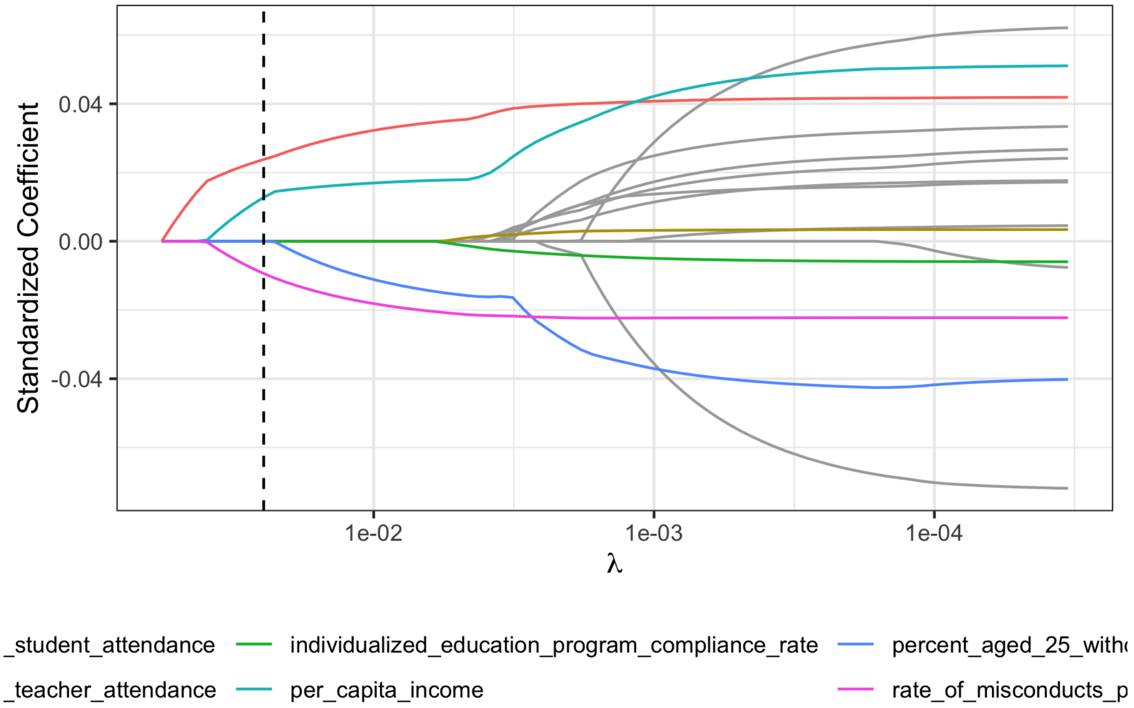

Figure 5: Lasso trace plot.

Next, we can look at some coefficients for the LASSO models that are chosen based on the 1se rule.

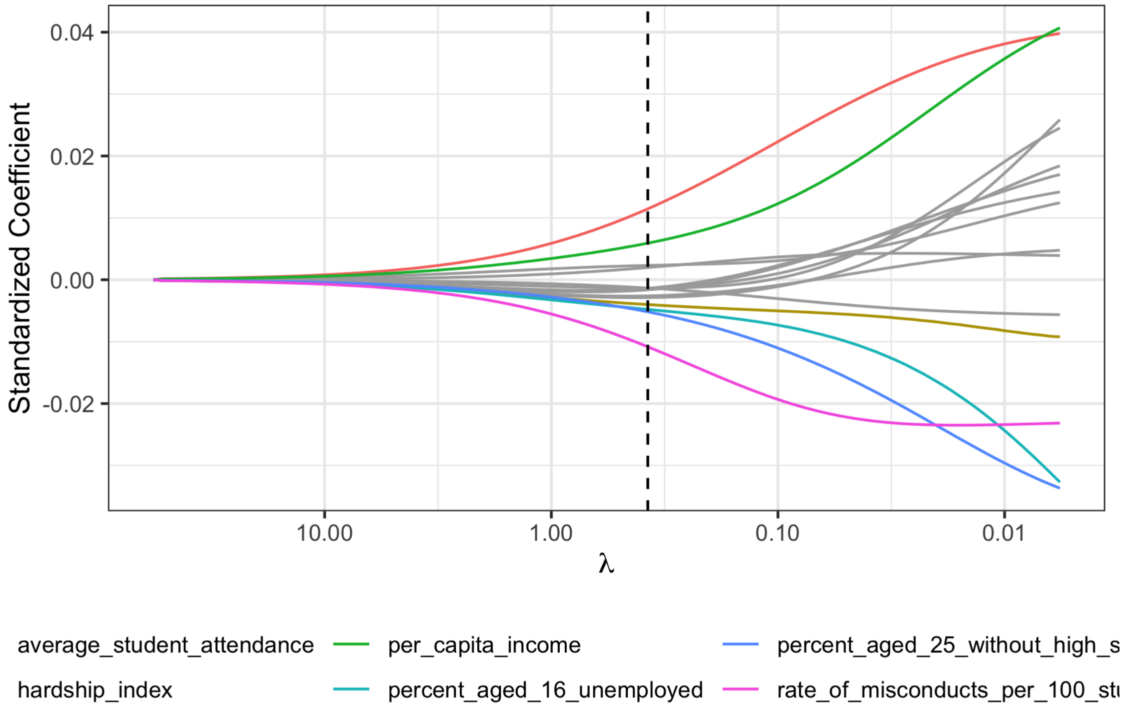

Figure 6: Ridge trace plot.

Table 2: Standardized coefficients for features in the lasso model based on the one-standard-error rule.

Feature	Coefficient
average_student_attendance	0.02
per_capita_income	0.01
rate_of_misconducts_per_100_students	-0.01

4.2 Tree-based methods

4.2.1 Decision trees

Now, we use decisions tree. First, I employ a simple tree model without using a minimum split. The resulting categories are similar to what the regressions have pinpointed

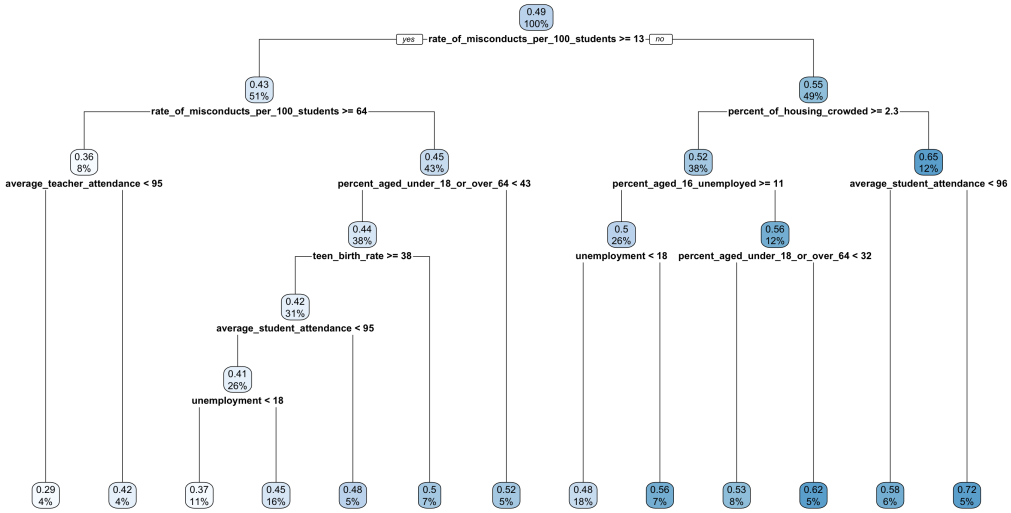
0.49
100%
yes
rate_of_misconducts_per_100_students >= 13
no
0.43
51%
rate_of_misconducts_per_100_students >= 64
0.55
49%
percent_of_housing_crowded >= 2.3
0.36
8%
average_teacher_attendance < 95
0.45
43%
percent_aged_under_18_or_over_64 < 43
0.52
38%
percent_aged_16_unemployed >= 11
0.65
12%
average_student_attendance < 96
0.44
38%
teen_birth_rate >= 38
0.5
26%
unemployment < 18
0.56
12%
percent_aged_under_18_or_over_64 < 32
0.42
31%
average_student_attendance < 95
0.41
26%
unemployment < 18
0.29
4%
0.42
4%
0.37
11%
0.45
16%
0.48
5%
0.5
7%
0.52
5%
0.48
18%
0.56
7%
0.53
8%
0.62
5%
0.58
6%
0.72
5%

Figure 7: Decision tree.

Next, I perform a minimum split at 100. Again, it seems that we have the same variables being relevant as in the case of regressions except in different orders.

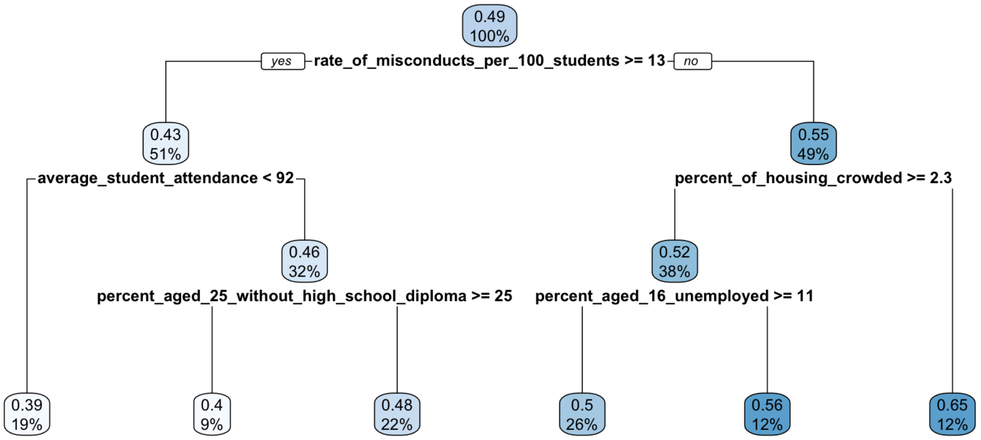
0.49
100%
yes
rate_of_misconducts_per_100_students >= 13
no
0.43
51%
average_student_attendance < 92
0.55
49%
percent_of_housing_crowded >= 2.3
0.46
32%
percent_aged_25_without_high_school_diploma >= 25
0.52
38%
percent_aged_16_unemployed >= 11
0.39
19%
0.4
9%
0.48
22%
0.5
26%
0.56
12%
0.65
12%

Figure 8: Decision tree with split at 100.

To wrap up the decision tree analysis, I include the CV error depending on the number of terminal nodes, which picks an optimal CV error for 3 nodes.

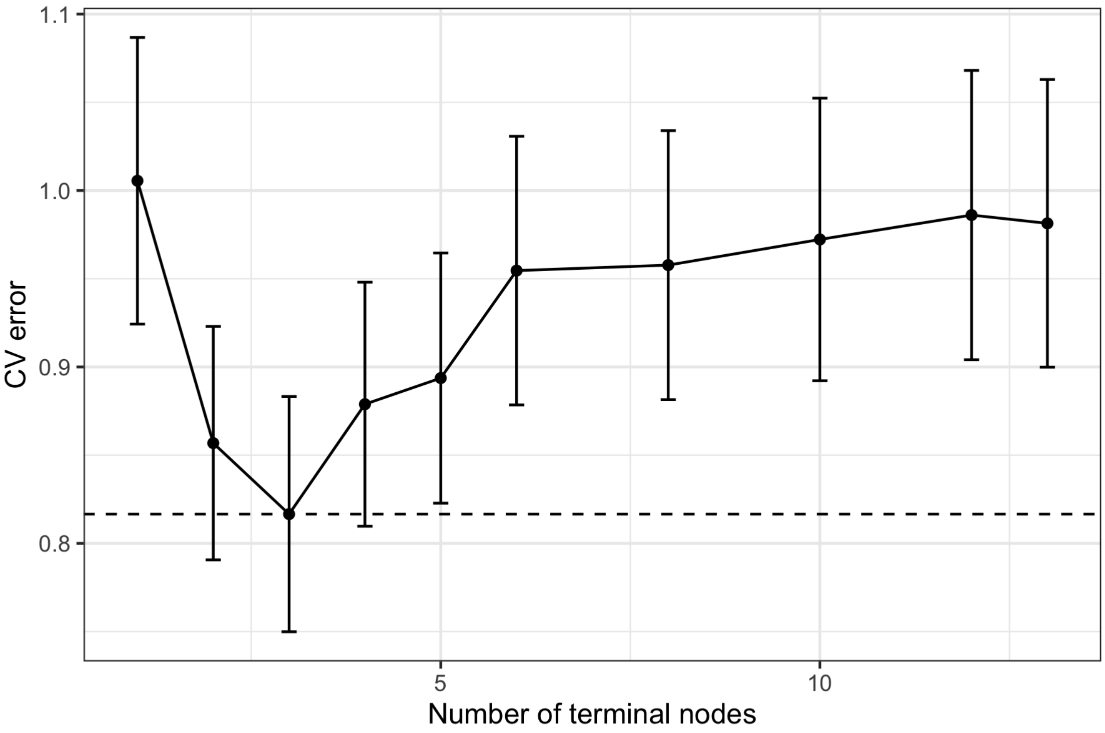

Figure 9: CV error graph.

4.2.2 Random forest

Next, to wrap up the tree-based methods, I fit a random forest to the data. Bagging is achieved when all 15 variables are considered at each tree split, which leads to a decent amount of vriation due to the fact that we don't have too many explanatory variables. Next, I tune the random forest model and analyze the error for each value of m from 1 to 15 and this leads to $m = 11$ being the OOB error is minimized. Next, we tune the B parameter which plateaus starting at around $B = 140$.

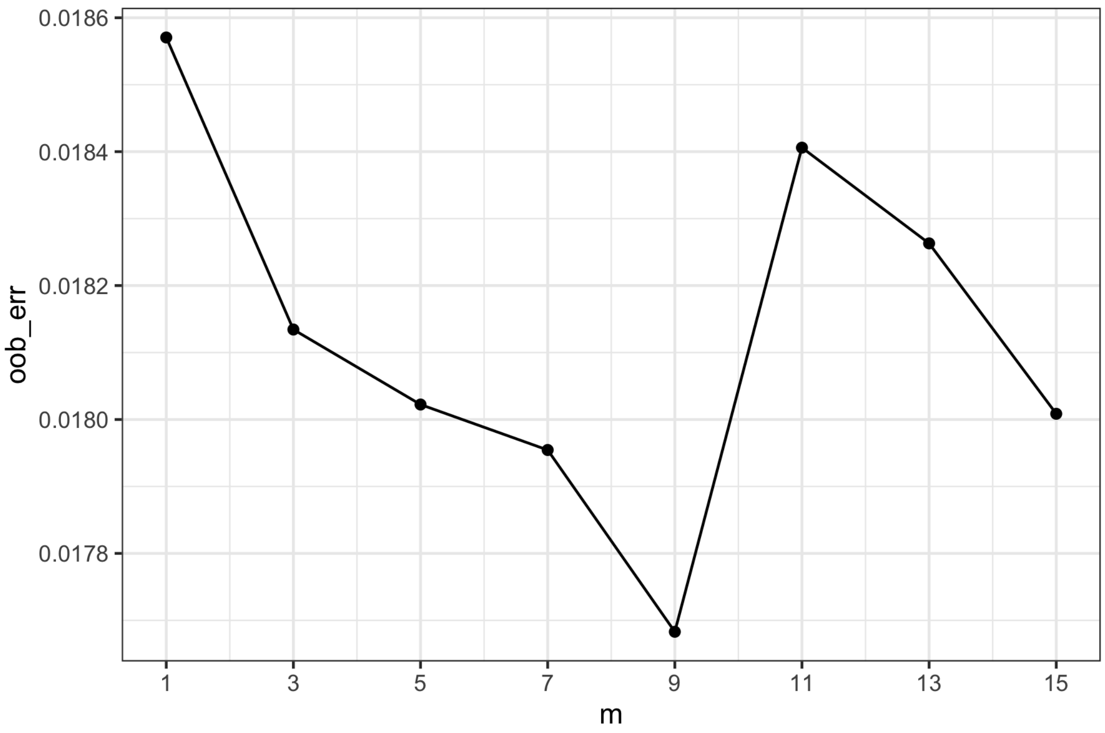

Figure 10: OOB against m.

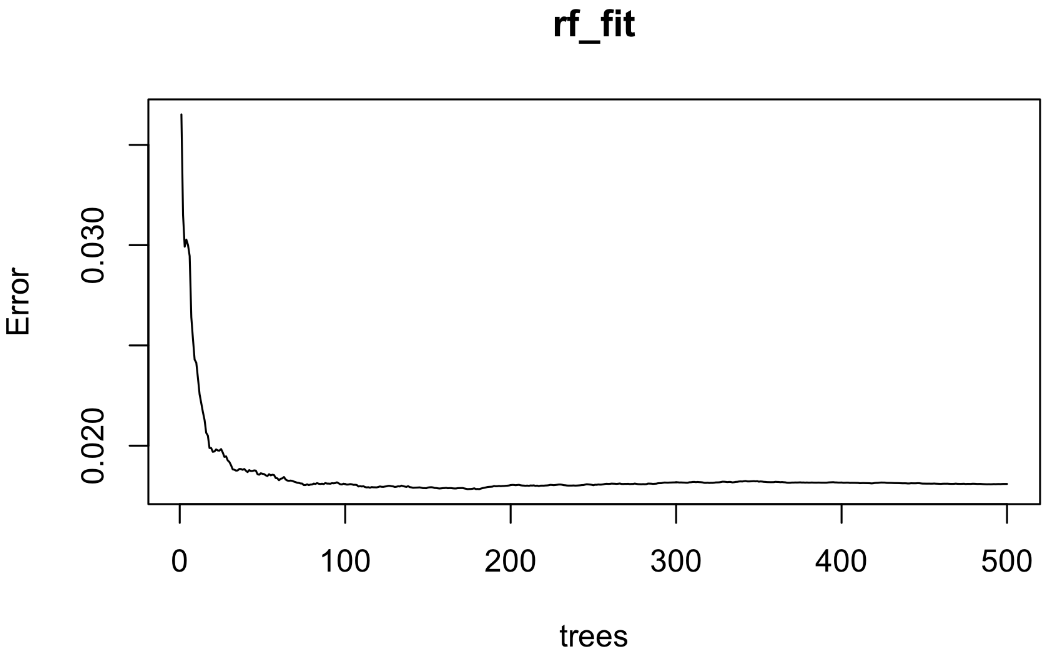

Figure 11: Error and number of bags.

I use those values to finally conclude the importance of each node. Looking at the node purity graph, we notice that social factors and school-specific factors are more important in predicting the scores in the random forest model. This is suggested by the fact that the first three variables with highest importance are all school-related metrics, and then follows housing crowdedness.

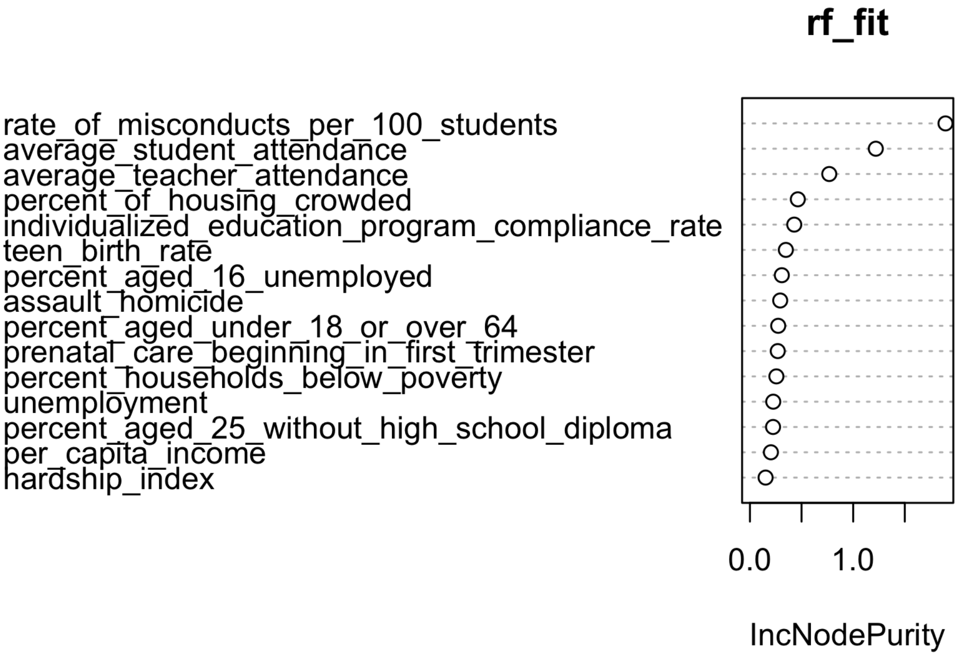

Figure 12: Node purity.

5 Conclusions

5.1 Method comparison

Table 3: Root-mean-squared prediction errors for models.

Model	Test error
OLS	0.16
LASSO	0.16
Ridge	0.16
Decision Tree	0.16
Split decision tree	0.15
Random Forest	0.02

Table 3 shows the test RMSE for all the methods considered. Random forest has the lowest error, while all the other models have a relatively equal RMSE. The regressions and the basic tree models have around 0.16 RMSE each, which falls behind to the superior error of the random forest model, which is 0.02.

The methods overlap significantly in their rsults and chosen variables, mostly focusing on the socio-economic factors and giving little to no importance for the healthcare related factors. It would be possible to argue that those results follow from the high correlatedness of socio-economic factors taken from the census data, except upon running methods individually that included only the hardship index in them, I identified that the results do not differ by much.

5.2 Takeaways

In general we notice that factors directly tied to the school environment itself are the strongest indicators of a better aggregate score in general. Misconduct rates and student attendance seem to be one of the highest, if not the two highest relevance explanatory variables in all our models. Other socioeconomic variables that are more general and defyining of social realities, such as hardship index and per capita income seem to be also important in assessing the predictive value of the aggregate score. Unfortunately, social factors that are closer tied to healthcare do not seem to have an important role, with some notable exceptions.

Given that the strongest factors were most directly school-related variables such as attendance and misconduct, it appears that the performance of a school is highly related to how it can enforce its discipline between students. Furthermore, this result is mostly surprising given the fact that a lot of socio-economic factors are relatively highly-correlated and hence they could have skewed the models in a certain direction, yet the internal school factors played a bigger role. If my findings are accurate in a certain measure, this implies that the crucial factor in improving public elementary school education quality is the reinforcing of adequate school policies that promote consistency and rules-obeying. However, one can't not notice that the rate of misconduct has an above average correlation with almost all socio-economic factors which implies that fixing the underlying societal factors and social policy would in an end also facilitate lowering down the misconduct rates. However, targeting misconduct directly could be a shortcut to a faster improvement of instruction, safety and environment of teaching in Chicago public elementary schools. ## Limitations

5.2.1 Dataset limitations

Finding data that corresponds to the same period of time can be very challenging in particular. The census data is from 2008-2012, while the schools data is from 2011-2012, so the overlap in those two periods could prove to be very benificial to the quality of the study. However, the historical health data is taken over a wider period of time, which in turn could have been slightly detrimental towards the results. Assuming the fact that the historical health data from past 10 years did not change significantly in relative terms in each community area, we could safely say that the detriment for this particular caveat is not big. A lot of schools have been removed from the analyzed data set due to under reportation of results, and also I could have not added high schools or elementary schools to the study even though I had some of their data, since it does not make much sense to compare metrics of high schools with metrics of elementary schools, even though they are the same. There is some correlation between the explanatory variables, especially between the socio economic data pulled from the census - this definitely creates some confounding which is not ideal to the models. One final limitation is that only very few schools actually reported certain metrics in the initial schools data set that seemed interesting, but I had to drop them due to severe lack of data.

5.2.2 Analysis limitations

Even though I fit a lot of models to the data and all of them produce similar outcomes, it might be the case that the type of explanatory variables I have chosen could not be the only ones that significantly affect the quality of teaching in a school. For example, the health data I have pulled had little to none variables that would talk explicitly about illnesses in children. Furthermore, some more specific socio-economic variables that might concern welfare programs could have shifted the results given that this would balance out the importance of per capita income. Finally, a big limitation could be the fact that we used an assessment made by students about the education they received, so there could be some human bias in the distribution of the aggregate scores that we used as the predicted variable.

5.3 Follow-ups

Definitely, more explanatory variables can be added to the dataset in order to connect the outcomes of an elementary schools education with other social aspects. For instance, it would be helpful to study the relation between crime rates and the aggregate scores presented in the dataset. Also analyzing subsets of the observations (i.e public elementary schools) that had a more systematic reporting of all their reports.

6 Apendix

6.1 Data from the public schools dataset:

SAFETY ICON: Student Perception/Safety category from 5 Essentials survey

SAFETY SCORE: Student Perception/Safety score from 5 Essentials survey

FAMILY INVOLVEMENT ICON: Involved Families category from 5 Essentials survey

FAMILY INVOLVEMENT SCORE: Involved Families score from 5 Essentials survey

ENVIRONMENT ICON: Supportive Environment category from 5 Essentials survey

ENVIRONMENT SCORE: Supportive Environment score from 5 Essentials survey

INSTRUCTION ICON: Ambitious Instruction category from 5 Essentials survey

INSTRUCTION SCORE: Ambitious Instruction score from 5 Essentials survey

LEADERS ICON: Effective Leaders category from 5 Essentials survey

LEADERS SCORE: Effective Leaders score from 5 Essentials survey

TEACHERS ICON: Collaborative Teachers category from 5 Essentials survey

TEACHERS SCORE: Collaborative Teachers score from 5 Essentials survey

PARENT ENGAGEMENT ICON: Parent Perception/Engagement category from parent survey

PARENT ENGAGEMENT SCORE: Parent Perception/Engagement score from parent survey

AVERAGE STUDENT ATTENDANCE: Average daily student attendance

RATE OF MISCONDUCTS (PER 100 STUDENTS): # of misconducts per 100 students

AVERAGE TEACHER ATTENDANCE: Average daily teacher attendance

INDIVIDUALIZED EDUCATION PROGRAM COMPLIANCE RATE: % of IEPs and 504 plans completed by due date

PK-2 LITERACY: % of students at benchmark on DIBELS or IDEL

PK-2 MATH: % of students at benchmark on mClass

GR3-5 GRADE LEVEL MATH: % of students at grade level, math, grades 3-5

GR3-5 GRADE LEVEL READ: % of students at grade level, reading, grades 3-5

GR3-5 KEEP PACE READ: % of students meeting growth targets, reading, grades 3-5

GR3-5 KEEP PACE MATH: % of students meeting growth targets, math, grades 3-5

GR6-8 GRADE LEVEL MATH: % of students at grade level, math, grades 6-8

GR6-8 GRADE LEVEL READ: % of students at grade level, reading, grades 6-8

GR6-8 KEEP PACE MATH: % of students meeting growth targets, math, grades 6-8

GR6-8 KEEP PACE READ: % of students meeting growth targets, reading, grades 6-8

GR-8 EXPLORE MATH: % of students at college readiness benchmark, math

GR-8 EXPLORE READ: % of students at college readiness benchmark, reading

ISAT EXCEEDING MATH: % of students exceeding on ISAT, math

ISAT EXCEEDING READ: % of students exceeding on ISAT, reading

ISAT VALUE ADD MATH: ISAT value-add value, math

ISAT VALUE ADD READ: ISAT value-add value, reading

ISAT VALUE ADD COLOR MATH: ISAT value-add color, math

ISAT VALUE ADD COLOR READ: ISAT value-add color, reading

STUDENTS TAKING ALGEBRA: % of students taking algebra

STUDENTS PASSING ALGEBRA: % of students taking algebra

9TH GRADE EXPLORE (2009): Average EXPLORE score, 9th graders who tested in fall 2009

9TH GRADE EXPLORE (2010): Average EXPLORE score, 9th graders who tested in fall 2010

10TH GRADE PLAN (2009): Average PLAN score, 10th graders who tested in fall 2009

10TH GRADE PLAN (2010): Average PLAN score, 10th graders who tested in fall 2010

NET CHANGE EXPLORE AND PLAN: Difference between Grade 9 Explore (2009) and Grade 10 Plan (2010)

11TH GRADE AVERAGE ACT (2011): Average ACT score, 11th graders who tested in fall 2011

NET CHANGE PLAN AND ACT: Difference between Grade 10 Plan (2009) and Grade 11 ACT (2011) COLLEGE

ELIGIBILITY: % of graduates eligible for a selective four-year college

GRADUATION RATE: % of students who have graduated within five years COLLEGE/ ENROLLMENT RATE: % of students enrolled in college

COLLEGE ENROLLMENT (NUMBER OF STUDENTS): Total school enrollment FRESHMAN ON TRACK RATE: Freshmen On-Track rate

6.2 Census data:

community_area_number: The number of the community area

community_area_name: The name of the community area

percent_of_housing_crowded: The percent of occupied housing units with more than one person per room

percent_households_below_poverty: The percent of households living below the federal poverty level

percent_aged_16_unemployed: The percent of persons in the labor force over the age of 16 years that are unemployed

percent_aged_25_without_high_school_diploma: The percent of persons over the age of 25 years without a high school diploma

percent_aged_under_18_or_over_64: The percent of the population under 18 or over 64 years of age (i.e., dependency);

per_capita_income: The mean yearly per capita income

hardship_index: The hardship index