Scott Hirabayashi

Data 602 (Krishnamoorthy)

**Applying Machine Learning Strategies to Educational Outcomes:**

**Modeling Usage on Chicago Public Schools**

**Abstract**

This summary proposes possible models and applicable considerations for a model-based approach to understanding graduation rates within Chicago’s Public Schools High Schools. Using a classification model based on a graduation rate threshold, central numerical and categorical features are cleaned, imputed, and standardized, then run through several model types, and checked for accuracy, precision, and F1 scores from their results. Models with the highest overall performance are then analyzed for feature importance, and re-modeled by categorizing types of features within the original. While exploratory, this paper and study conclude by specifying the importance of different factors within a school, and their effects on schools meeting graduation targets.

## **Introduction**

**1.1 Background**

The American K-12 educational system in 2021 has found itself thrust into the spotlight, with concerns regarding teaching styles (does online education work?), content, such as the debate over critical race theory, and societal valuation, where many teachers feel underappreciated and overworked. And, more than ever, school systems, schools, and teachers are constantly posed with the question, “How well are our schools performing?”.

By student expenditure, America spends considerably on its educational system. In 2017, an OECD study on student expenditures showed that, within the OECD, the US was 4th in expenditure per student out of 38 member countries. These expenditures compared to the OECD average were about 37% higher (The US’ $14,100 against the OECD average of $10,300). Yet, in terms of student performance, 2018 PISA results yielded reading scores and science scores slightly above average compared to the OECD average, and under the average for mathematics. In terms of secondary graduation rates, OECD studies found the US as 7th among the OECD nations, still trailing a higher expectation for expenditure, but performing better than PISA scores.

It is easy and popular to assert that the US educational system is performing poorly for the amount we pay. However, the education question of value to efficacy and outcomes is rife with nuance. Straus (2018) notes that confidence in the US’s schools dropped, but the drop is not indicative of quality change. She writes “It seems, then, that abstract perceptions of schools — the nation’s schools — have suffered, while satisfaction with actual schools remains fairly constant.”

The reality of the education question is that, despite high student expenditures, schools and districts are limited in what actions they can take to improve their educational outcomes for students.

**1.2 Focus**

This paper analyzes the Chicago Public School system as a subset of the US’s public education, and analyzes current educational data to look for important considerations in delegating and distributing its funds and emphasis areas for improvement, with the hopes that the US’s funding-to-outcomes could compare more favorably to those in fellow OECD countries.

Thus, the focus and analysis of this project and paper is to see whether or not high schools within the Chicago Public Schools system are meeting the state and national standards for performance (meeting the graduation threshold of 85%, the metric for same-year graduation in Illinois and the larger US), and what factors have the greatest importance in determining if a school can meet this criteria. In this case, a special emphasis is placed on graduation rate, a universal metric for school performance.

**1.3 Research Questions**

The research questions to be addressed are:

1. Can a skillful model be created which can predict school graduation rates in Chicago Public Schools for the 2017-2018 school year?
2. What school features have high importance within a skillful model? What school features may seem intuitive, but have little effect on meeting graduation rates?
3. What conclusions can be drawn from the data in terms of what areas schools should place the most focus on?

**Methodology**

**2.1 Data Cleaning**

The public high school data was obtained from Chicago’s public data website, which collects data on the 654 schools within the Chicago Public School systems, and tracks 182 different features for each school. Data was first cleaned; features in the dataset were dropped if they had high percentages of nulls or did not meet a high enough correlation threshold to the target classifier variable.

As several of the features within the dataset were survey data, these were converted into a likert-scale score. All kept columns were imputed with mean values for nulls. After cleaning all relevant features, features were min-max scaled to standardize each feature’s importance.

Lastly, the target graduation rate variable was set as a binary classifier, where a high school met or did not meet the National and Illinois standard (graduation rate of 85%). Due to the obvious importance of this classifier, any school that lacked graduation rate statistics was dropped from the dataset.

**2.2 Model Building**

After data cleaning, the remaining features in the dataset could be categorized within roughly three categories:

* Standardized testing data (SAT and PSAT scores from numerous grades),
* Survey data on school performance (school safety, teacher collaboration, leadership, and family factors),
* General non-test performance statistics (attendance and truancy).

Using these features, the dataset of these features for the 140 Chicago public high schools was run against several different models using Scikit Learn’s inbuilt model-building functionalities (Logistic Regression, Ridge Regression Classifier, Bayesian, Tree, Random Forest, Ada Boost, and Gradient Boost), and analyzed for precision, recall, and F1 score as a means of evaluating model performance.

Last, the features were split modeled based on the category of the feature data into three models to compare to performance of the main model.

**Results**

Model performance across the different models varied, with model weighted accuracy scores ranging from around 60-80%. Overall, logistic regression and ridge classifier models appeared to perform the best on predicting whether or not schools were meeting the target 85% graduation rate.

* Logistic Regression: Weighted Accuracy of 82%
* Ridge Classifier: Weighted Accuracy of 82%

However, it is also important to note that *none* of the models performed well on schools that were meeting the 85% graduation rate threshold classifier (ridge classifier was best at an F1 score of .63).

**Discussion:**

Within context, logistic regression suggests that continuous explanatory variables are of significant importance to this classification model (Science Direct). This is similarly indicated by the strength of testing metrics on the logistic regression model.

An analysis of coefficients was subsequently run after model creation on the logistic regression model. Feature importance suggests that test scores and test-related measures have high importance within this model. Three of the five top features (Student Attainment Rating, 9th Grade PSAT attainment, and all-grades Attainment, all test performance based factors). The other two important features were student mobility score and Freshmen on Track.

On the opposite end of the spectrum, student growth, teacher attendance, the number of suspensions, and teacher collaboration were the weakest predictors of school performance. Additionally, an effective leadership survey feature also had very weak importance, and barely had any feature importance with high schools meeting the graduation threshold.

When modeled in the three categories, standardized testing data again showed its superiority in classification, with a weighted accuracy average of 75%, compared to 63% for the survey and non-test performance models. More significantly, while poor, the standardized test data was able to occasionally predict the threshold-meeting schools, whereas neither the survey nor non-test performance models were successful with a single classification on threshold-meeting schools (0.0 for precision/recall/f1 score).

**Conclusion:**

Returning to our original question, are we able to create a skillful model? Due in large part to the strength of standardized testing scores within the features in this dataset, a semi-skillful model can be created using logistic regression. Ultimately, model performance was satisfactory for warranting future analysis, but performance scores make it challenging to draw larger conclusions on the data.

Still, these findings suggest that any school should take into account the surprising importance that standardized test scores can have on predicting graduation rates. This is not to say that they are causal- more research.

Additionally, the logistic regression model suggests how important it is to get students off to a good start. 9th grade PSAT scores, and freshmen (9th grade) students being on track are both significant predictors (both in the top-5 features in importance, Attainment PSAT = 1.0, Freshmen on Track=.965) for predicting a school meeting the graduation target. Equivalent scoring metrics for 10th and 11th grade PSAT/SAT scores were nowhere near as important (10th grade PSAT = .685, 11th grade SAT = .614).

Last, the mobility of students within a school appears to have a high negative correlation to a school meeting a graduation ranking. There are several ways to interpret this- perhaps that students with poor academic and behavioral records may often transfer in and out of lower performing schools. However, this could also be indicative of structural issues within Chicago Public Schools that could benefit from further analysis- student homelessness and housing insecurity can contribute to high mobility, and could have significant impact on the system’s chances of their graduation.

**Bibliography**

Camera, L. (2019, June 19). *International Survey: U.S. Teachers Are Overworked, Feel Underappreciated*. US News and World Report.

*Coe—Education expenditures by country*. (2021, May). https://nces.ed.gov/programs/coe/indicator/cmd

*Illinois Report Card*. (n.d.). Https://Www.Illinoisreportcard.Com/District.Aspx?Source=studentcharacteristics&source2=mobility&Districtid=15016299025; Illinois Report Card.

*Logistic regression—An overview | sciencedirect topics*. (n.d.). Retrieved December 20, 2021, from https://www.sciencedirect.com/topics/computer-science/logistic-regression

*Questions about CPS*. (n.d.). Retrieved December 20, 2021, from https://avault.github.io/example-expedition.html

*Secondary graduation rate*. (n.d.). Secondary Graduation Rate; OECD. https://data.oecd.org/students/secondary-graduation-rate.htm

Strauss, V. (2018, October 15). Perspective | How are America’s public schools really doing? *Washington Post*. https://www.washingtonpost.com/education/2018/10/15/how-are-americas-public-schools-really-doing/