Identifying and Influencing Students at Risk of Not Finishing High School

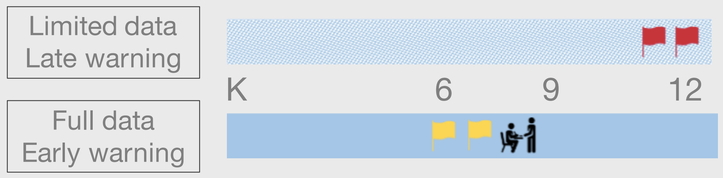
Eric & Wendy Schmidt Data Science For Social Good Summer Fellowship

Xiang Cheng — Emory University
Jacqueline Gutman — New York University
Johanna Torrence — University of Chicago
Zhe Zhang — Carnegie Mellon University
Chad Kenney — University of Chicago
R. Ali Vanderveld — University of Chicago
Kevin H. Wilson — University of Chicago

THE UNIVERSITY OF CHICAGO

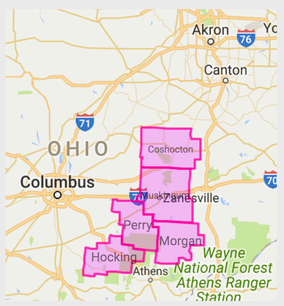
1 Introduction

1 in 5 high school students in the U.S. do not graduate high school within four years*. Schools are limited by teachers' intuition, incomplete information, and interventions that come too late. Districts rarely share data, and at-risk students are only identified late in high school after significant red flags. With access to more comprehensive data, predictive models can supplement educators' intuition to sooner identify which students are in need of targeted interventions.

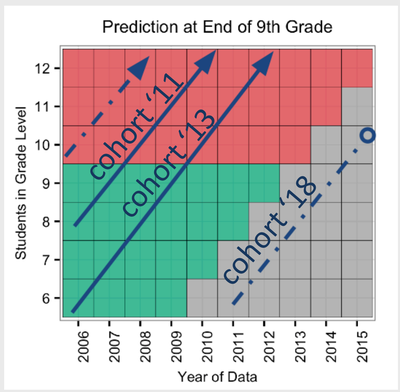


* U.S. Department of Education, National Center for Education Statistics. (2015). *The Condition of Education 2015* (NCES 2015-144),Public High School Graduation Rates.

2 Data



The data is provided by our partner Muskingum Valley Educational Service Center, which provides services to 16 school districts in 5 counties in rural southeastern Ohio, and which reaches 30,897 pre-K to 12 students each year.



Data Time Range: 2006 - 2016

Not used; Occurs after prediction
Use for machine learning
No outcome labels

Data to Identify Student Outcomes

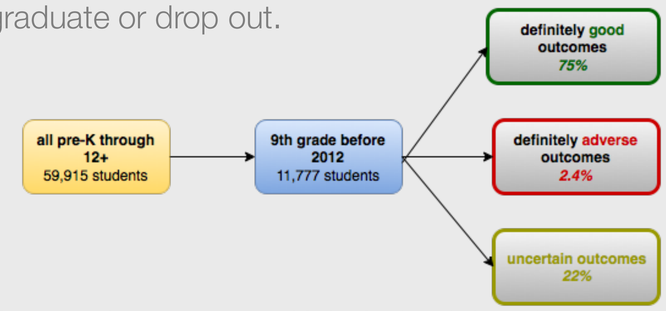
Graduation Dates	Withdrawal Reasons	Ohio Graduation Tests Scores

Data to Create Predictive Features

Demographics	School Grades	Elementary + Middle School Test Scores
Daily and Annual Attendance	Student Mobility (by District/Residence)	Services and Special Education Received
Club and Athletic Memberships	Teacher + Course Memberships	Home Addresses
Discipline and Suspensions	Disability Classification	Economic and Gifted Status

3 Labels

To train models to differentiate between students at risk of not graduating on time and those likely to graduate, we use labeled data. However, in high-mobility districts such as those serviced by MVESC, about 22% of the students disappear from the data without an indication of whether they eventually graduate or drop out.



To complete our set of labeled training data, we analyzed the performance of unlabeled students (in their last year of observed data) versus the performance of students with known outcomes. Using this comparison, we used cutoffs on grades, attendance, and test scores to classify these unlabeled students.

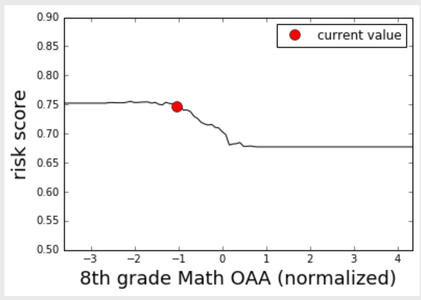
4 Methods

Predictive Models

We tested many model families and setups to estimate each student's risk of not graduating. Random forest and logistic regression models most accurately ranked students according to their risk. These models were evaluated on the percentage of actual at-risk students identified in the top 15% of predicted students. This metric allows educators to see how effectively interventions can be targeted to the students most in need of assistance.

Risk Factors

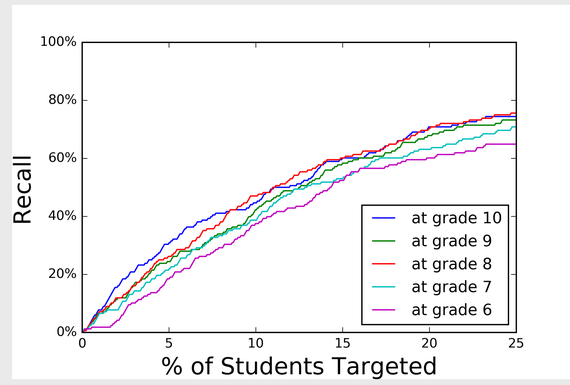
From each of these models, we can estimate the influence of individual student attributes on their predicted risk. These are not causal, but we highlight the strongest factors in the model to present to educators, to help them target appropriate support.



This graphic visualizes how variations in a particular student's score on the Ohio standardized math exam would influence our model's estimate of their risk.

5 Results

We built models to predict a student's risk of not graduating high school at the start of each school year. If schools can offer interventions to 15% of students, using our model they can already identify 53% of students at risk at the start of 7th grade. By the start of 10th grade, 60% of students at risk can be properly identified and offered help*.



A typical warning system based on cutoffs in the ABCs (attendance, behavior, and course grades) would only identify 45% of the at-risk students in 10th grade. Our predictive model successfully identifies an additional 15% of 10th grade students at risk who can then receive the support they need.

* These percentages are calculated based on members of the classes of 2015 and 2016 who are present in our data from 7th through 10th grade.

6 Impact

Three extra years. Our predictive model at 7th grade identifies more at-risk students than a typical ABCs system would identify in 10th grade. This allows teachers three additional years to provide effective support for these students.

Key	Risk		Top 3 Risk Factors		
student	score	level	factor1	factor2	factor3
23214	0.98	High	GPA	Disability	Absence
00621	0.86	High	Discipline	Absence	Mobility
18762	0.79	Med	Disability	GPA	Absence

Our model is designed to integrate with existing dashboards for teachers which provide access to a wide breadth of individual student data. Adding our data-driven predictions and interpretable, individual risk factors empowers educators' intuitions. With this, they can better target support to specific students, improving graduation and the number of students achieving desirable educational outcomes.

Muskingum Valley Educational Service Center @mvesc

This work was done during the Eric & Wendy Schmidt Data Science for Social Good Fellowship at the University of Chicago.