Eric & Wendy Schmidt
Data Science For Social Good
Summer Fellowship

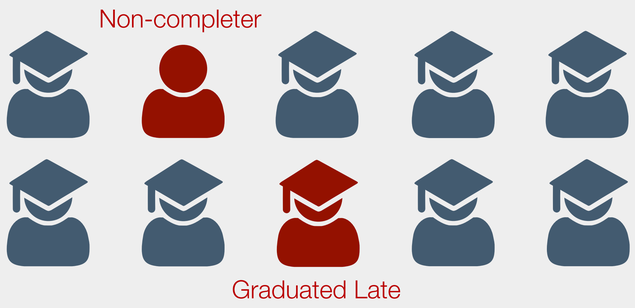
Data-Driven Strategies for Predicting On-Time High School Graduation

THE UNIVERSITY OF CHICAGO

Kerstin Frailey Cornell University
Ruobin Gong Harvard University
Siobhan Greatorex-Voith Harvard University
Reid A. Johnson University of Notre Dame
Anushka Anand Tableau Research
Alan Fritzler Northwestern University

1 Introduction

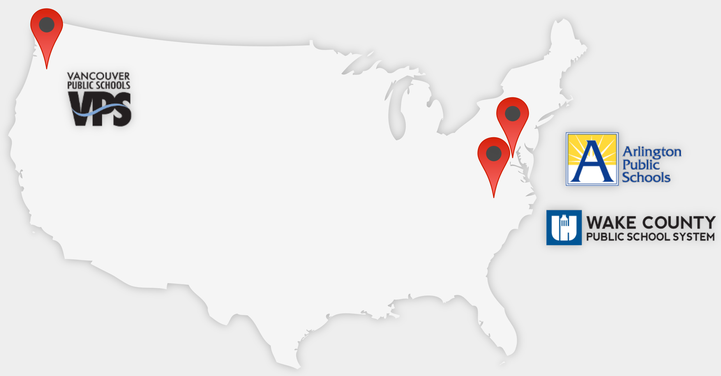
About one in five high school students in the U.S. fail to graduate on time*, either not completing high school or graduating late. To best apply intervention programs for these off-track students, schools need to identify students at risk as early as possible and understand the factors that contribute to this risk.



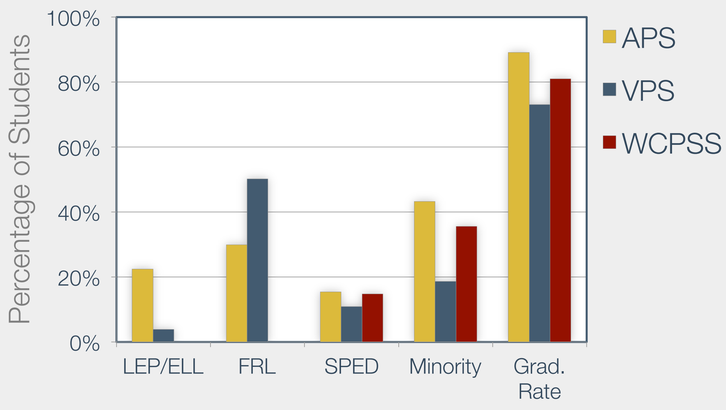
*G. Kena, L. Musu-Gillette, J. Robinson, et al. *The Condition of Education 2015*. (NCES 2015-144). U.S. Department of Education, National Center for Education Statistics, Washington, D.C., 2015.

2 Data

Data for this work comes out of a partnership with three public school districts—Arlington Public Schools (APS), Vancouver Public Schools (VPS), and Wake County Public School System (WCPSS)—each already recognizing the importance of identifying students at risk and applying interventions.



In addition to being geographically disparate, these districts represent markedly diverse student populations with different graduation outcomes.



3 Methods

With the provided data, we applied machine learning models to deliver accurate, interpretable, and data-driven predictions of on-time high school graduation.

Raw Data

The data provided by our partnering school districts primarily consists of longitudinal student records.

Grade Level						
6	7	8	9	10	11	12
Absences Grades Tardies Tests …	Absences Grades Tardies Tests …	Absences Grades Tardies Tests …	Absences Grades Tardies Tests …	Absences Grades Tardies Tests …	Absences Grades Tardies Tests …	Absences Grades Tardies Tests …
Date of Birth, Gender, Race/Ethnicity, …						

Feature Generation

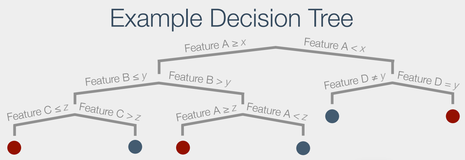
From these records, we generated student-level features potentially predictive of on-time graduation.

Grade Level						
6	7	8	9	10	11	12
Features	Features	Features	Features	Features	Features	Features

Data-Driven Modeling

We then used these student-level features to predict on-time graduation via a suite of predictive models, generating a model for each applicable grade level.

Example Decision Tree



Model Validation

These predictive models were evaluated upon the precision with which they predict off-track students.

Risk Scoring

Name	Score	Actual
Jayne Cobb	.97	Yes
River Tam	.92	Yes
Olivia Dunham	.90	Yes
Maeby Funke	.87	No
Inara Serra	.82	Yes
Malcolm Reynolds	.79	No
Kaylee Frye	.79	No
Maggie Lizer	.62	Yes
Nina Sharp	.40	No
Peter Bishop	.21	No

Good: Off-track students have the highest predicted risk.

Bad: On-track student predicted higher risk than off-track student.

Precision:

$$\frac{\text{true off-track predictions}}{\text{true off-track predictions + false off-track predictions}}$$

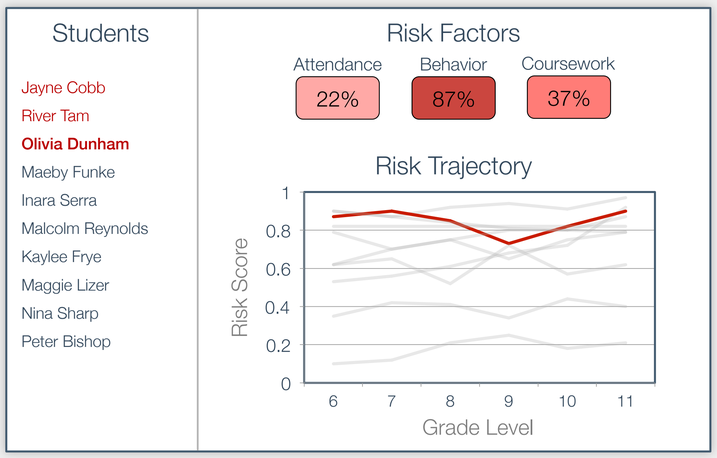
Precision at *k*: Precision on the top *k* off-track predictions.

4 Results

Our data-driven modeling approach significantly improves upon existing baselines. Here we summarize the results produced for one of our partner districts, where we predict on-time high school graduation for students in 6th through 11th grade. The results indicate that we can identify off-track students with relatively high precision, which increases as the prediction time-frame decreases. The error bars illustrate statistical significance.

5 Impact

Our work provides school district personnel with empirical, data-driven predictions of whether students will graduate on time. These predictions can be leveraged to identify key factors contributing to a student's struggles. In turn, these key factors can be used to facilitate focused interventions, which may have a profoundly positive effect on a student's academic trajectory—and, ultimately, his or her life.



Arlington Public Schools
VANCOUVER PUBLIC SCHOOLS VPS
WAKE COUNTY PUBLIC SCHOOL SYSTEM

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