Identifying Students At Risk Accurately and Early

Everaldo Aguiar[1], Nasir Bhanpuri[2], Himabindu Lakkaraju[3], David Miller[4], Ben Yuhas[5], Rayid Ghani[2,5]

[1]University of Notre Dame, [2]University of Chicago, [3]Stanford University, [4]Northwestern University, [5]Mentor

The Eric & Wendy Schmidt Data Science for Social Good Summer Fellowship

THE UNIVERSITY OF CHICAGO
CENTER FOR DATA SCIENCE AND PUBLIC POLICY

The problem

Increasing high school graduation rates improves outcomes for students, their families, and economies of local neighborhoods

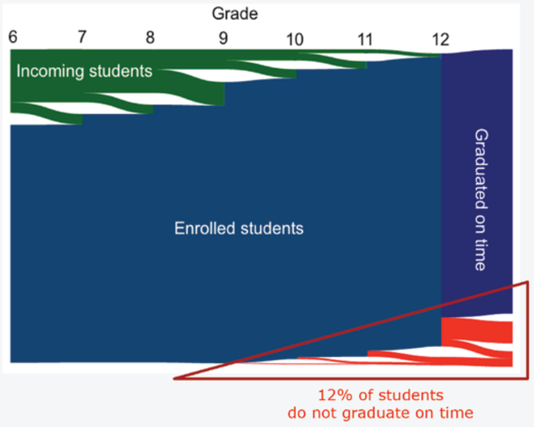
MONTGOMERY COUNTY PUBLIC SCHOOLS

We used machine learning methods to help Montgomery County Public Schools (MCPS) answer:

- **Who?** Which students are at risk of not graduating high school on time?
- **When?** When will student go off track?
- **Why?** How can schools better identify particular student needs?

Longitudinal data

The dataset comprises of information on students' grades, absences, suspensions, and other related information



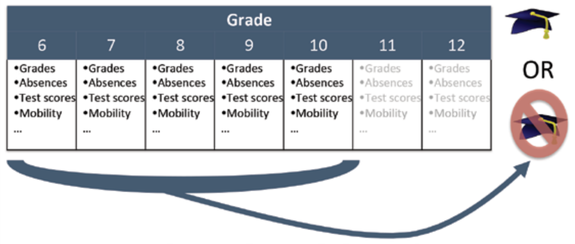
Three analytic goals

- **Who?** Develop predictive models of graduating high school on time
- **When?** Predict when students will first go off track using survival analysis methods
- **Why?** Characterize typologies of students using clustering methods and interactive visualizations

Methods

Who?

- Our goal was to identify at-risk students in different grades (e.g., develop a grade 10 model)



- We evaluated model performance based on the accuracy among students identified as high risk (e.g., top 10% of risk scores)

When?

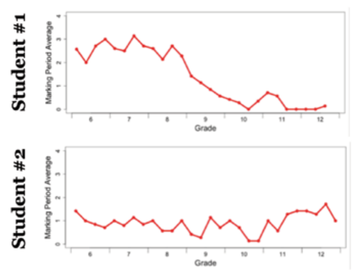
Problem: How do we prioritize students based on the urgency of their needs ?

Solution:

- Identify high-risk students using the above models
- Predict when those students are likely to go *off-track* (be held back a grade or drop-out)
- Ordinal regression trees outperformed standard survival analysis methods (e.g., Cox regression)

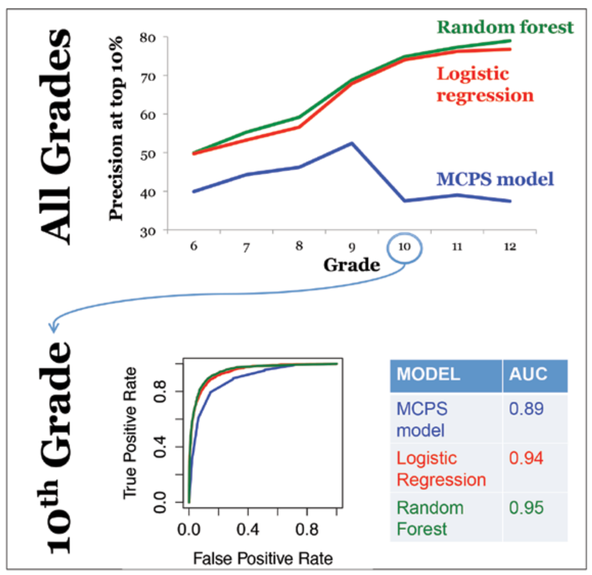
Why?

- Use clustering (e.g., k-means, latent class analysis) to group similar types of students

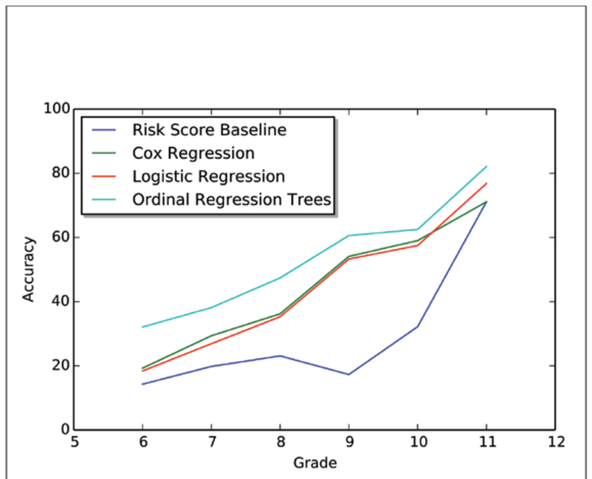


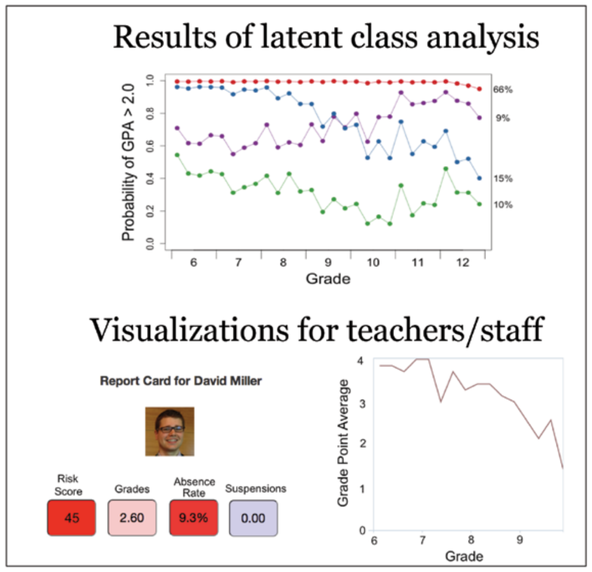
- Use interactive visualizations to show individual student profiles to teachers/staff

Results



MODEL	AUC
MCPS model	0.89
Logistic Regression	0.94
Random Forest	0.95





Summary

- Classification with logistic regression or random forest led to a sizeable improvement over rule-based model
- Ordinal regression trees can predict off-track year with about 50-80% accuracy
- Clustering revealed different typologies of at-risk students, and dashboard highlights student history succinctly

Conclusions

Machine learning methods can help school districts more efficiently use limited resources by:

- Identifying at risk students more accurately (**who**)
- Prioritizing students based on urgency of students' needs (**when**)
- Characterizing different student support needs (**why**)

Future directions

Match students to interventions

- Run experiments of evidence-based interventions
- Develop models of students' responsiveness to particular interventions

Test robustness and expand

- Test accuracy using future cohorts of students
- Expand to school districts with varying student needs and graduation rates

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