Machine Learning Guided Evaluation of a College Program for Under-Prepared Students

Abstract for 2018 SREE Conference

William J. Gehring
University of Michigan Department of Psychology (wgehring@umich.edu)

Julian Hsu
The College Board (**Contact Author**: jhsu@collegeboard.org)

Wei Ai
University of Michigan School of Information (aiwei@umich.edu)

Informing Policies for Under-Prepared College Students

Context

Under-prepared college students are at risk of dropping out (Bettinger & Long, 2005) and having lower academic achievement to their better-prepared peers. Yet they would likely benefit most from higher education (Jaegar & Page, 1996; Kane & Rouse, 1995), highlighting the need for institutional support. While developmental courses in two-year community colleges are well studied (Martorell & McFarlin, 2011; Scott-Clayton & Rodriguez, 2015), four-year colleges have other tools at their disposal.

Program and Population

We study one selective four-year state institution's (anonymized as the University) program, anonymized as the No-MisMatch Program (NMP). NMP identifies under-prepared students and requires them to take a variety of summer prep courses before enrollment, gives them extensive tutoring resources throughout the year, offers specialized course sections, and requires multiple advisor meetings each semester. Students are notified of their NMP assignment with their admission decision.

Research Objective and Preview

Using machine learning with a regression discontinuity design, we estimate the effect of NMP assignment on academic outcomes. Our key preliminary finding is that assignment to NMP increases the likelihood students subsequently enroll at the University around 20 percentage points. In future work, we will look at impacts on college GPA, cumulative credits, and graduation.

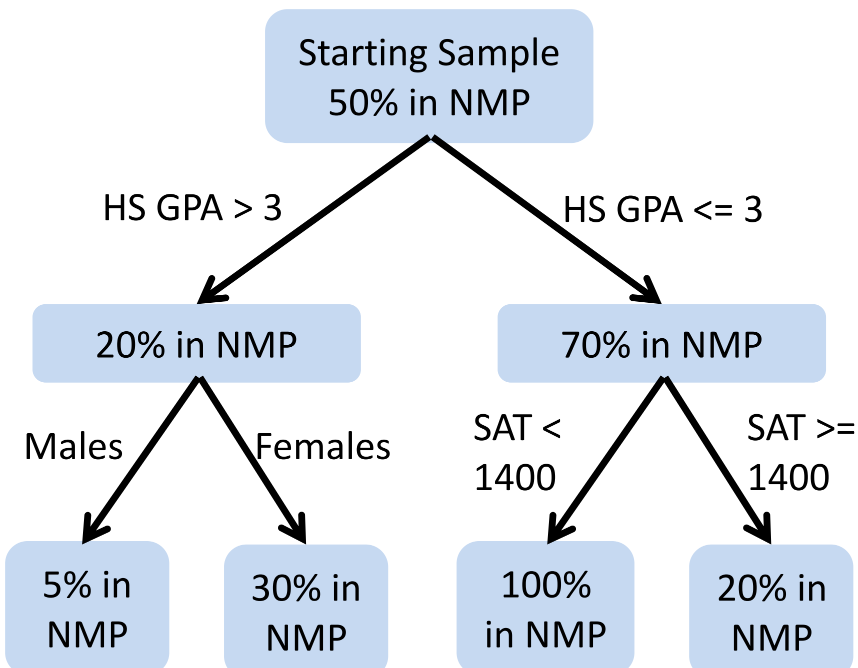
Research Design: Combining Machine Learning with Regression Discontinuity

The first methodological problem is that the University only assigns under-prepared students to NMP, so a straightforward comparison is likely not causal. Using discontinuities in NMP assignment based on students' characteristics, we employ a regression discontinuity framework to estimate the causal impact of NMP assignment on academic outcomes. The regression discontinuity framework argues that among students in the small interval around the discontinuity, assignment to NMP is as good as random, facilitating causal inference (Thistlewaite and Campbell, 1960).

The second problem is that the decision rule for assignment to NMP is not transparent, as the process is potentially sensitive to the University. We use standard machine learning techniques to approximate the assignment mechanism and identify discontinuities. We use decision tree algorithms, which iteratively bifurcate the data to match the actual probability students are assigned to a treatment group. Where previous implementations use machine learning to strengthen propensity score matching methods (Athey and Imbens, 2015) or heterogenous treatment effects (Davis and Heller, 2017), we use decision tree algorithms to identify discontinuities in how students are assigned to NMP.

In the sample figure below, the decision tree algorithm identifies having a high school GPA less than or equal to 3, and having an SAT score less than 1400 as potential discontinuities. We would then do a graphical test of a discontinuity.

Figure 1: Graphical Example of Decision Tree Algorithm Output



We use administrative admissions information from the University for students applying for the enrolling classes of Fall 2002 to Fall 2012. Except for college essays and recommendation letters, we use the same administrative data used by the University to assign students to NMP. This data includes students' ACT and SAT scores, high school GPA, recruiter recommendations, and socio-economic characteristics.

Using the decision tree algorithm, we find four discontinuities in how students are assigned based on their ACT scores. Figures 2 to 5 below show these discontinuities across four admission cohorts:

Figure 2. Probability of NMP Assignment over ACT Composite Score Fall 2003 Cohort

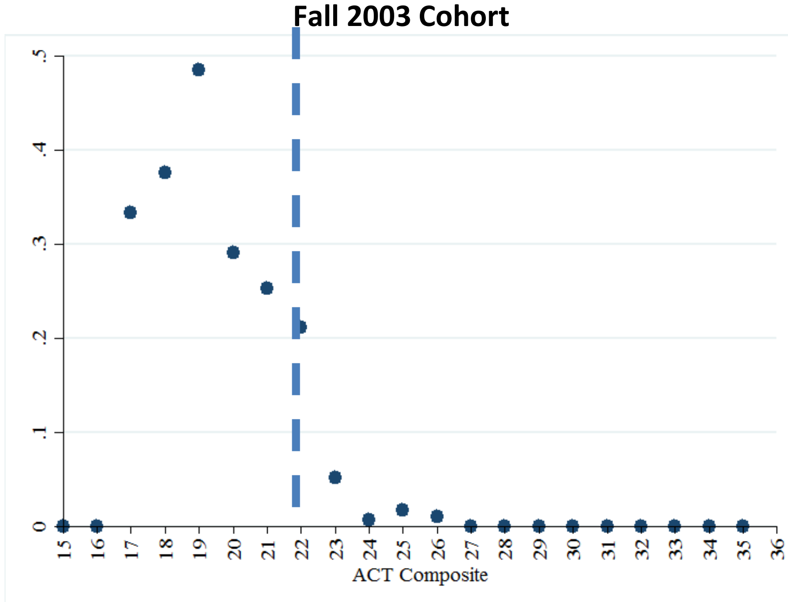


Figure 3. Probability of NMP Assignment over ACT Composite Score Fall 2005 Cohort

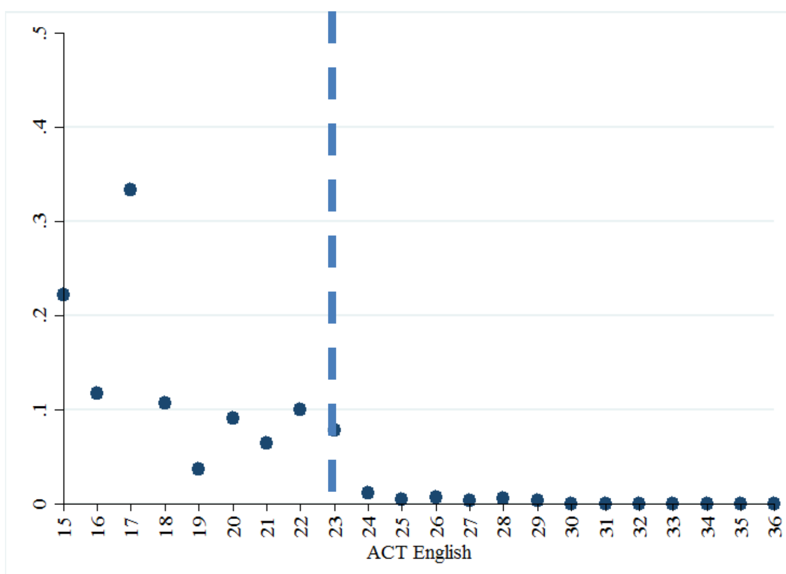


Figure 4. Probability of NMP Assignment over ACT Math Score
Fall 2007 Cohort

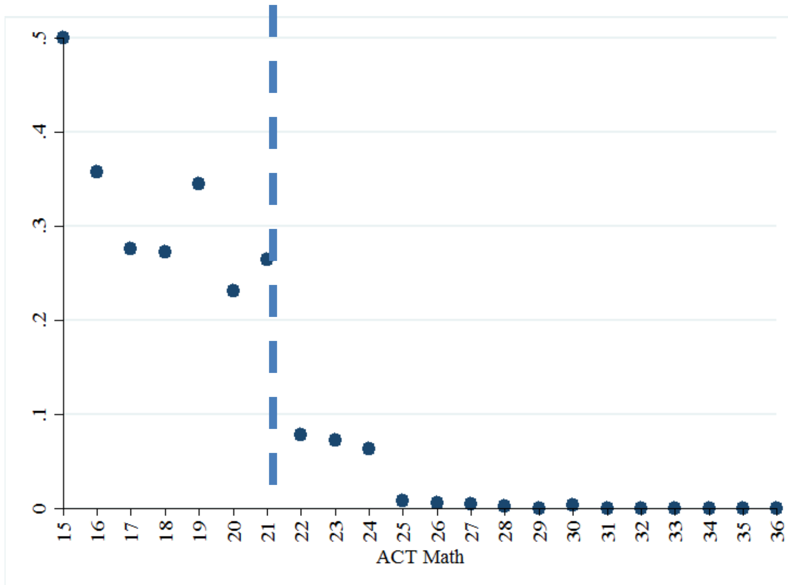
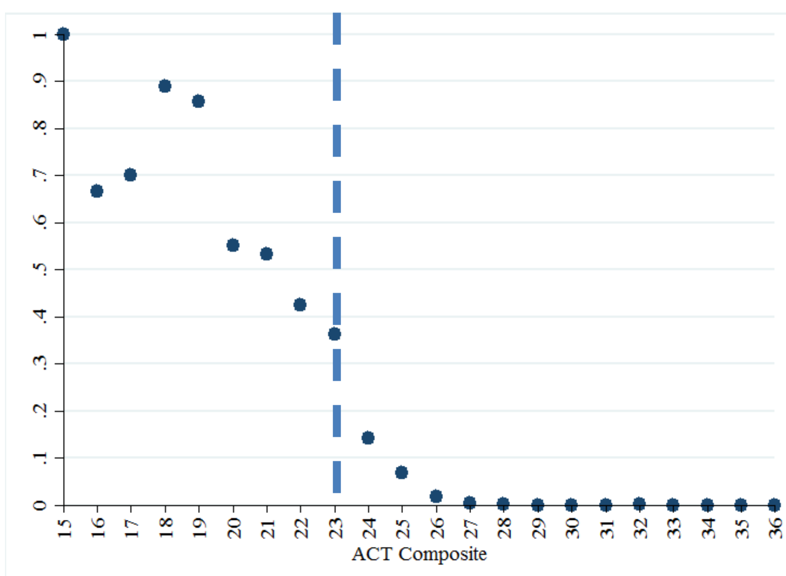


Figure 5. Probability of NMP Assignment over ACT Composite Score
Fall 2012 Cohort



Estimation Strategy

Across these four cohorts, we collapse applicants ACT Composite or ACT Math scores into a single running variable and center it at zero. We compare students 5 points around the cutoff.[1] We use whether students have a running variable lower than the cutoff as an instrument for whether students are assigned to NMP.

$$(1) \qquad NMP_i = \alpha_0 + \alpha_1 1\{RV_i \leq 0\} + \alpha_2 RV_i + \nu_i$$
$$(2) \qquad Y_i = \beta_0 + \beta_1 \widehat{NMP}_i + \beta_2 RV_i + \epsilon_i$$

Our first stage equation (1) regresses NMP assignment on an indicator of whether the student has a running variable less than or equal to zero and the continuous running variable. The second stage equation (2) regresses the predicted NMP assignment on outcome Y_i. Under the monotonicity assumption (Angrist et al., 1996), β_1 estimates the causal impact of being assigned to NMP on the outcome.

Our preliminary results on Table 1 look at the outcome students enroll at the University. In future work, we will look at academic outcomes such as college GPA, attempted credits, and time to graduation.

Table 1.
Whether NMP Assignment Affected Enrollment at The University

	Instrumental Variable	First Stage	Intent-to-Treat
NMP	0.218		
	[0.221]		
1{Running Variable $\leq$ 0}		0.125***	0.021
		[0.013]	[0.026]
F- Statistic		54.732	5.248
Sample Size	6219	6219	6219
Control Mean	0.600	0.600	0.600

Implications for Policy Makers

While not statistically significant, this substantially significant estimate suggests that NMP can act as a potential enrollment management tool for policy makers. Students' admission letters positively frame being assigned to NMP, offering students resources for success, rather than indicating under-preparation. Reaching out to these under-prepared students importantly promotes the institution's and higher education's mission to close potential achievement gaps.

We believe that this approach of reaching out to students with the offer of academic resources is reasonably scalable. In addition to students who are under-prepared, institutions offer similar resources to students admitted to the Honors Program.

---

[1] Current bandwidth selection mechanisms do not allow for discretized running variables.

References

Angrist, J., Imbens, G., Rubin, D. (1996). Identification of Causal Effects Using Instrumental Variables. Journal of the American Statistical Association, 91(434), 444-455.

Athey, S. & Imbens, G. (2015). Machine Learning for Estimating Heterogeneous Causal Effects. Working Paper.

Bettinger, E. P., & Long, B. T. (2009). Addressing the Needs of Underprepared Students in Higher Education. Journal of Human Resources, 44(3), 736–771.

Davis, J. M.D. & Heller, S. B (2017). Using Causal Forests to Predict Treatment Heterogeneity: an Application to Summer Jobs. American Economics Review: Papers and Proceedings, 107(5), 546-550.

Jaegar, D., & Page, M. (1996). Degrees Matter: New Evidence on Sheepskin Effects in the Returns to Education. Review of Economics and Statistics, 78(4), 733–740.

Kane, T. J., & Rouse, C. E. (1995). Labor-market returns to two-and four-year college. The American Economic Review.

Martorell, P., & McFarlin, I. (2011). Help or Hindrance? The Effects of College Remediation on Academic and Labor Market Outcomes. Review of Economics and Statistics, 93(2), 436–454.

Thistlethwaite, D. L., & Campbell, D. (1960). Regression-discontinuity analysis: An alternative to the ex-post facto experiment. Journal of Educational Psychology, 51(6), 309–317.