IES Institute of Education Sciences

STUDY SNAPSHOT
October 2021

Regional Educational Laboratory Mid-Atlantic
At Mathematica

Predicting Early Fall Student Enrollment in the School District of Philadelphia

Predicting incoming enrollment is an ongoing concern for the School District of Philadelphia (SDP) and similar districts with school choice systems, substantial student mobility, or both. Inaccurate predictions can disrupt learning as districts adjust to enrollment fluctuations by reshuffling teachers and students well into the fall semester. This study explored how machine learning algorithms might improve predictions of incoming cohort size in SDP. More accurate predictions could help SDP better anticipate incoming cohort sizes and, consequently, reduce instability among students and staff.

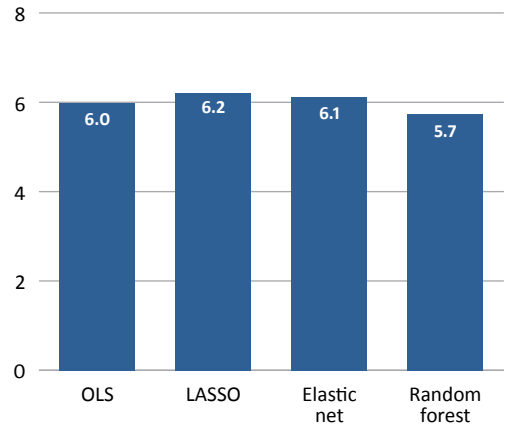
This study compared the accuracy of four prediction methods—ordinary least squares (OLS) regression (a commonly used statistical technique) and three theoretically more powerful methods (least absolute shrinkage and selection operator, elastic net, and random forest)—in forecasting fall enrollment at the school-by-grade level, using school-by-grade-level data from prior years, to assess which approach might be the most useful for planning school staffing in SDP. The study also assessed whether predictive accuracy varied across cohorts with different demographic compositions.

Key findings

- **All four algorithms have similar predictive accuracy, with the random forest slightly outperforming the others.** The median prediction error for all four algorithms is about 6 students, or roughly 10 percent of a typical cohort of 60–70. About 22 percent of cohorts would be subject to reshuffling under each algorithm. If the goal is to increase predictive accuracy at the typical school, the choice of algorithm is not consequential.
- **Four predictors provide the most meaningful contribution to accurately predicting school-by-grade enrollment.** Of the 259 predictors assessed in the models, 4 stand out as the most important: prior cohort size, in-school suspensions, out-of-school suspensions, and absences. Regardless of the methods used, improved accuracy is likely to require additional predictors that include stronger signals of incoming cohort sizes.

Typical error is six students for each algorithm

Median absolute deviation (number of students)



OLS is ordinary least squares. LASSO is least absolute shrinkage and selection operator.

Note: Values are from the extrapolation set (2018/19 school year).

Source: Authors' calculations based on administrative data for 2015–19 provided by the School District of Philadelphia.

For the full report with technical details, see https://go.usa.gov/xMQhG.