New techniques could help identify students at risk for dropping out of online courses
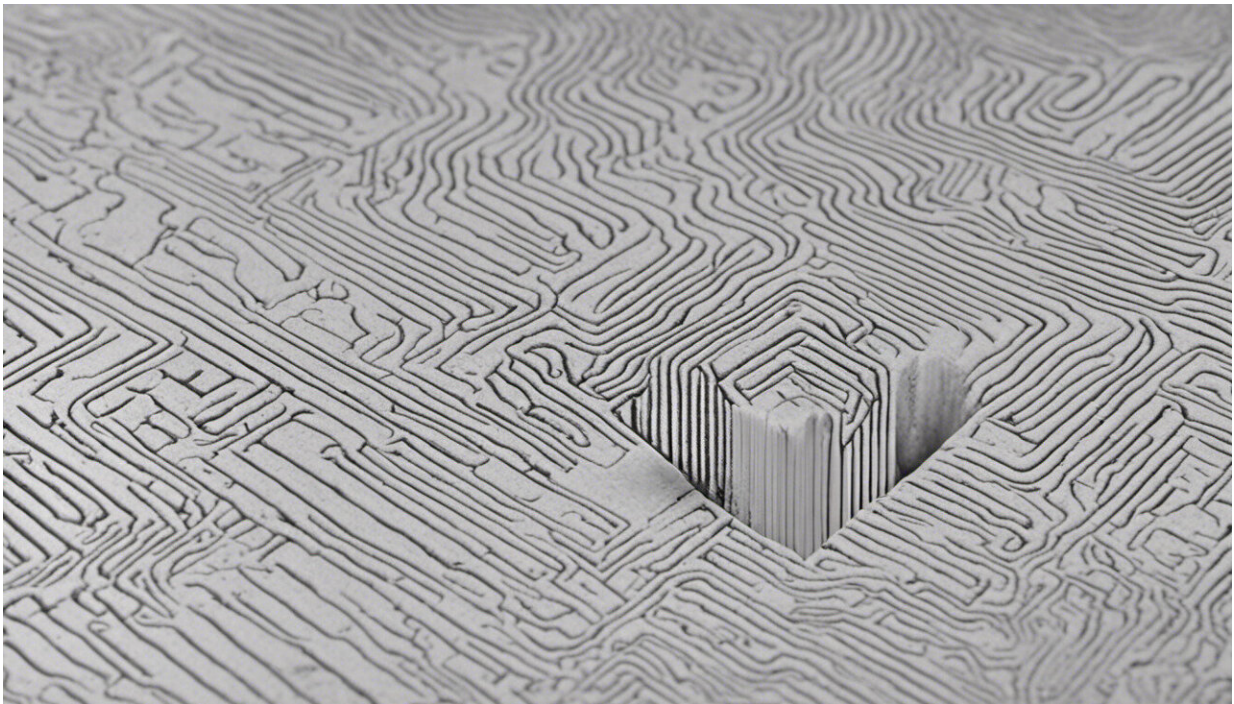
July 1 2015, by Larry Hardesty

Credit: AI-generated image (disclaimer)

MOOCs—massive open online courses—grant huge numbers of people access to world-class educational resources, but they also suffer high rates of attrition.

To some degree, that's inevitable: Many people who enroll in MOOCs

may have no interest in doing homework, but simply plan to listen to video lectures in their spare time.

Others, however, may begin courses with the firm intention of completing them but get derailed by life's other demands. Identifying those people before they drop out and providing them with extra help could make their MOOC participation much more productive.

The problem is that you don't know who's actually dropped out—or, in MOOC parlance, "stopped out"—until the MOOC has been completed. One missed deadline does not a stopout make; but after the second or third missed deadline, it may be too late for an intervention to do any good.

Last week, at the International Conference on Artificial Intelligence in Education, MIT researchers showed that a dropout-prediction model trained on data from one offering of a course can help predict which students will stop out of the next offering. The prediction remains fairly accurate even if the organization of the course changes, so that the data collected during one offering doesn't exactly match the data collected during the next.

"There's a known area in machine learning called transfer learning, where you train a machine-learning model in one environment and see what you have to do to adapt it to a new environment," says Kalyan Veeramachaneni, a research scientist at MIT's Computer Science and Artificial Intelligence Laboratory who conducted the study together with Sebastien Boyer, a graduate student in MIT's Technology and Policy Program. "Because if you're not able to do that, then the model isn't worth anything, other than the insight it may give you. It cannot be used for real-time prediction."

Generic descriptors

Veeramachaneni and Boyer's first step was to develop a set of variables that would allow them to compare data collected during different offerings of the same course—or, indeed, offerings of different courses. These include things such as average time spent per correct homework problem and amount of time spent with video lectures or other resources.

Next, for each of three different offerings of the same course, they normalized the raw values of those variables against the class averages. So, for instance, a student who spent two hours a week watching videos where the class average was three would have a video-watching score of 0.67, while a student who spent four hours a week watching videos would have a score of 1.33.

They ran the normalized data for the first course offering through a machine-learning algorithm that tried to find correlations between particular values of the variables and stopout. Then they used those correlations to try to predict stopout in the next two offerings of the course. They repeated the process with the second course offering, using the resulting model to predict stopout in the third.

Tipping the balance

Already, the model's predictions were fairly accurate. But Veeramachaneni and Boyer hoped to do better. They tried several different techniques to improve the model's accuracy, but the one that fared best is called importance sampling. For each student enrolled in, say, the second offering of the course, they found the student in the first offering who provided the closest match, as determined by a "distance function" that factored in all the variables. Then, according to the closeness of the match, they gave the statistics on the student from the first offering a greater weight during the machine-learning process.

In general, the version of the model that used importance sampling was more accurate than the unmodified version. But the difference was not overwhelming. In ongoing work, Veeramachaneni and Boyer are tinkering with both the distance function and the calculation of the corresponding weights, in the hope of improving the accuracy of the model.

They also continue to expand the set of variables that the model can consider. "One of the variables that I think is very important is the proportion of time that students spend on the course that falls on the weekend," Veeramachaneni says. "That variable has to be a proxy for how busy they are. And that put together with the other variables should tell you that the student has a strong motivation to do the work but is getting busy. That's the one that I would prioritize next."

  More information: "Transfer Learning for Predictive Models in Massive Open Online Courses." groups.csail.mit.edu/EVO-Desig … eeramachaneni228.pdf

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology