XRDS SPRING 2023
FEATURE

Learning Analytics: An opportunity for education

How analytics is transforming the future for millions of learners.

By Ryan S. Baker

Analytics, data mining, machine learning, and, increasingly, broadly applicable artificial intelligence are revolutionizing various fields. Education is no exception. Though rhetoric around education often emphasizes how much classrooms look like the classrooms of a century ago, what children are doing to learn is changing rapidly. Increasingly, education is shaped by artificial intelligence and data. Today, growing amounts of K-12 and undergraduate learning involve adaptive learning platforms that let students work at their own pace, provide detailed feedback on student errors, and inform teachers what students are struggling with. More and more schools and universities are using predictive analytics to determine which students are at risk of dropping out (or other negative outcomes), and why. These advances—where they are used—are producing better learning outcomes and putting students on the path to more promising futures. These advances build upon a growing body of research, involving using analytics to better understand and model learning (aka learning analytics or educational data mining) and using the findings to redesign and improve learning systems (aka learning engineering).

In this article, I'll briefly discuss the methods used in learning analytics, highlight key applications, and discuss important challenges facing the field. I'll also discuss ways you can get involved in the field of learning analytics.

METHODS

Many of the methods used in learning analytics are drawn from analytics and data science more generally. However, there are specific areas (such as a learning system figuring out what a student knows) that are more specific to education. In this section, I'll briefly outline some of the key areas; for a more in-depth treatment, there is our chapter within the *The Cambridge Handbook of the Learning Sciences* or my (still totally free) edX course, Big Data and Education.[1]

The most frequently used type of method in learning analytics is prediction modeling. The goal of prediction modeling is to develop a model that can infer a single aspect of the data (the predicted variable) from some combination of other aspects of the data (predictor variables). Developing a prediction model typically depends on first finding out what the predicted variable is for a small set of data; a model is then created for this small set of data and statistically validated so that it can be applied at greater scale. However, there are now a few examples where emerging techniques, such as zero-shot learning, have been applied to entirely new areas. Researchers in the field use various types of prediction modeling, including classification,

[1] https://www.edx.org/course/big-data-and-education

regression, and sequence prediction. One that receives considerable attention, specifically in learning analytics, is knowledge tracing. Knowledge tracing involves inferring what a student knows, based on their correct and incorrect answers. An effective model of what a student knows will also predict their future performance. Prediction modeling is often used to develop models that can be embedded in learning platforms to make real-time decisions. Additionally, it can also be used to develop variables for research analyses.

A second major area is structure discovery. Structure discovery algorithms attempt to find structure in the data without focusing on a specific variable in advance. This is a very different goal than in prediction. Common approaches to structure discovery in learning analytics include clustering/latent class analysis (to find groups of data points that are similar), factor analysis (to collapse a large number of variables into a smaller number of variables), domain structure discovery (to discover how learning content groups into skills or topics), and the analysis of networks (including social networks and also epistemic networks, which study relationships in qualitative data). Structure discovery methods are used for a broad range of purposes, from gaining initial insight into a dataset to building structural models (such as knowledge graphs, for instance) that can be used within a learning platform.

A third major area is relationship mining. In relationship mining, the goal is to discover relationships between variables when you have a dataset with a large number of variables. Broadly, there are four types of relationship mining, which each look for a distinct type of relationship: association rule mining (finding if-then rules in data), sequential pattern mining (finding if-then rules over time), correlation mining (finding relationships between variables), and causal data mining (trying to understand causes and effects). Relationship mining is often used to derive insights relatively early in a design process in order to find ways to improve a system.

A variety of other learning analytics methods exist that do not fit into these three broad categories. Reinforcement learning algorithms attempt to learn what choices a system should make (for instance, whether a system should give a hint or an explanation). Discovery with models approaches use the predictions made by prediction models as components of models of complex phenomena. Visualization and explainable artificial intelligence methods attempt to make data meaningful to users such as teachers and students. Finally, generative artificial intelligence approaches (also called foundation models), such as large language models, are now being used in a multitude of ways, from enhancing prediction models (for instance, through zero-shot learning and the creation of synthetic data) to the auto-generation of learning content.

KEY APPLICATIONS

Learning analytics methods are being used for a wide variety of purposes in education. One of the most prominent uses is within adaptive learning systems (sometimes also referred to as intelligent tutoring systems). While the term adaptive learning system is sometimes used in a commercial context to describe systems based on crude or out-of-date technology, the most advanced such systems use learning analytics to model a range of aspects of the student, and then use these models to dynamically adapt to the student's needs. Contemporary adaptive learning systems have been developed that can recognize and adapt to what a student knows, what learning strategies the student is using, and even what emotions the student is experiencing. Even a decade ago, these systems were already having a significant positive impact on student learning [3], and the technology is getting better all the time.

The same models that provide key information on students can also be provided to teachers and other stakeholders through dashboards. Today, a growing number of learning platforms provide real-time data to teachers on which skills their students are struggling with, and which students need support right now. Teachers use this information to redesign their upcoming class sessions and choose which students to call on (both to illustrate correct solutions and to review misconceptions across the class). They also use this information in real time to identify if a specific student is struggling, in order to talk with that student and proactively offer assistance.

Dashboards are also used for another major application of learning analytics—predicting which students are at risk of negative outcomes and why. At-risk prediction modeling has become an important part of education in several countries, from models predicting which students are at risk of dropping out of high school or not on track to graduate college on time to models predicting poor performance or failure in individual courses. The best of these models provide instructors, guidance counselors, academic advisors, and others with concrete and clear information on why a student is at risk (for instance, perhaps one student is struggling with mathematics while another student is getting into fights) and potential steps to take to assist the student.

Finally, learning analytics (particularly structure discovery and relationship mining) is used in a wide variety of ways to engineer better learning. For example, learning analytics models can help to identify content that needs to be improved, such as content that is too difficult, videos that do not help students learn, and items that students are too likely to be able to guess on. Learning analytics can also be used to better identify which content a student needs to know before moving on to other content, helping to create the knowledge graphs that are the backbone of many contemporary adaptive learning systems.

KEY CHALLENGES

Despite the overall successes of learning analytics in the 15 years since the first conference in the area (the International Conference on Educational Data Mining), there remain a range of interesting challenges for the field to solve.

A growing concern in the last few years is the risk of algorithmic bias. Evidence has emerged that a range of learning algorithms are less effective for specific groups of learners [2]. The field does not yet know the full scope of which groups of students are impacted or the full scope of how they are impacted (which learning situations they are impacted in, in which ways, etc.). A particularly prominent gap is research on algorithmic bias in education worldwide—most of the research on algorithmic bias in education has taken place in the United States.

A second challenge is in how to better integrate the full scope of knowledge that a learning system has about a student. A given student may use five different online learning technologies at their school in the course of a year. Each of these systems learns important things about the student; but the systems do not talk to each other and cannot leverage what the other systems learn. They also cannot leverage what the teacher knows and cannot provide the teacher with unified information on what has been learned about the student. Developing more interoperable systems, which can help different learning systems and the teacher work together in concert, has considerable potential for creating more sensitive and efficient learning experiences for students.

A third challenge is to detect the right things. It is now relatively easy to detect when a student is learning in a straight-forward subject, such as mathematics problem-solving. However, detecting a student's ability to apply their creativity, or their ability to regulate their learning, or their ability to work in a team is more challenging, although there is existing research in these areas. Ultimately it is essential to support students in developing these key skills for the future of work. In general, it has been easier to optimize learning systems for immediate learning and engagement than to develop systems that benefit aspects of learning and engagement that will carry forward throughout their lives.

A fourth challenge is generalization. Right now, although software for running algorithms and fitting models is reused, the specific models developed for specific learning platforms often do not generalize to other learning platforms. Sometimes they do not even generalize to new content within the original learning platform. Since models remain expensive to develop, poor generalizability slows the uptake of new ways to benefit students. Some technologies, such as engagement detection, have been demonstrated to benefit students, but their cost and poor generalizability have prevented them from being used at scale. Addressing this challenge could therefore have considerable benefits.

A final important challenge—or perhaps it is an opportunity—comes from the recent advances in generative artificial intelligence, such as large language models. Adaptive learning systems will need to recognize when student inputs (such as essays or computer programs) are generated by generative AI to avoid overestimating student competence. At the same time, great opportunities are present for advancing the quality of models and breadth of content that can be supported in next-generation learning systems. When learning analytics first came onto the scene, it was accompanied by a new generation of learning technologies and educational support technologies. The same can be anticipated over the next few years, as we figure out the full scope of ways that generative AI can be applied to education.

GETTING IN INVOLVED

If this article motivates you to get more involved in learning analytics, there are several ways to do so. Perhaps the easiest thing is to join relevant mailing lists, like the Society for Learning Analytics Research mailing list, edm-announce, edm-discuss, and the Learning Analytics google group. There are now several master's programs in learning analytics, several of them fully online, including programs at the University of Pennsylvania, Teachers College Columbia University, the University of Wisconsin, and the University of Texas at Arlington. Other universities offer certificate programs. The top scientific journals and conferences in the area offer open-access articles and proceedings.

Learning analytics is growing, in terms of the size of the field and opportunities for impact. We hope you will join us.

References

[1] Baker, R. and Siemens, G. Educational data mining and learning analytics. In Sawyer, K. (Ed.) *The Cambridge Handbook of the Learning Sciences: 3rd Edition.* Cambridge University Press, Cambridge, 2022. Preprint available online at https://learninganalytics.upenn.edu/ryanbaker/Chapter12BakerSiemensv3.pdf

[2] Baker, R. S. and Hawn, A. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education* 2022, 1–41.

[3] VanLehn, K. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist* 46, 4, 2011, 197–221.

Biography

Ryan S. Baker is the director of the University of Pennsylvania Center for Learning Analytics, and professor of Education (courtesy appointment in Computer and Information Science). Baker was the founding president of the International Educational Data Mining Society and has used machine learning to develop models of student engagement, knowledge, and academic risk embedded in several educational platforms used at scale.