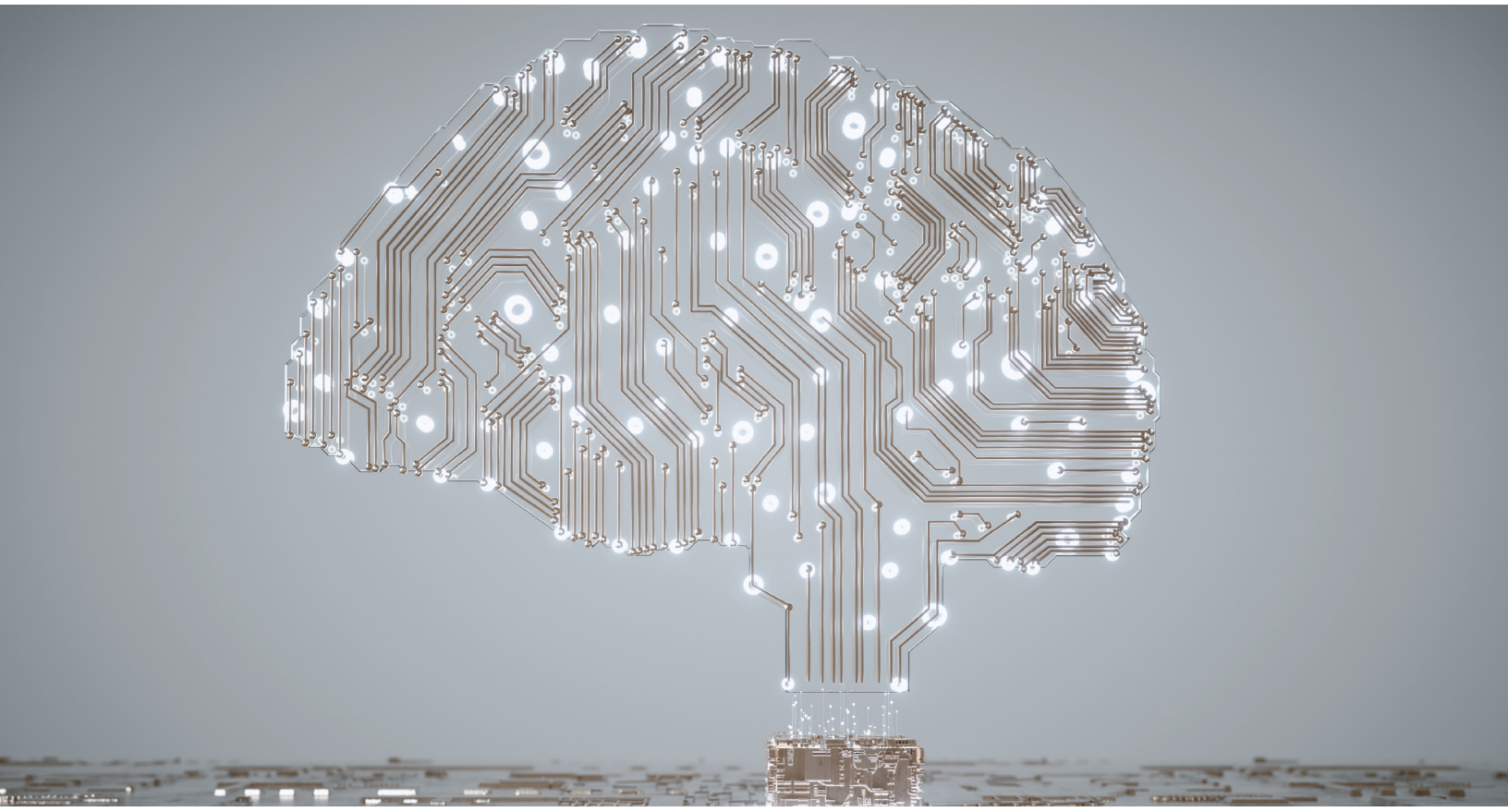
McKinsey & Company

Public & Social Sector Practice

Using machine learning to improve student success in higher education

Deploying machine learning and advanced analytics thoughtfully and to their full potential may support improvements in student access, success, and the overall student experience.

This article is a collaborative effort by Claudio Brasca, Nikhil Kaithwal, Charag Krishnan, Monatrice Lam, Jonathan Law, and Varun Marya, representing views from McKinsey's Public & Social Sector Practice.



April 2022

Many higher-education institutions are now using data and analytics as an integral part of their processes. Whether the goal is to identify and better support pain points in the student journey, allocate resources more efficiently, or improve student and faculty experience, institutions are seeing the benefits of data-backed solutions.

Those at the forefront of this trend are focusing on harnessing analytics to increase program personalization and flexibility, as well as to improve retention by identifying students at risk of dropping out and reaching out proactively with tailored interventions. Indeed, data science and machine learning may unlock significant value for universities by ensuring resources are targeted toward the highest-impact opportunities to improve access for more students, as well as student engagement and satisfaction.

For example, Western Governors University in Utah is using predictive modeling to improve retention by identifying at-risk students and developing early-intervention programs. Initial efforts raised the graduation rate for the university's four-year undergraduate program by five percentage points between 2018 and 2020.[1]

Yet higher education is still in the early stages of data capability building. With universities facing many challenges (such as financial pressures, the demographic cliff, and an uptick in student mental-health issues) and a variety of opportunities (including reaching adult learners and scaling online learning), expanding use of advanced analytics and machine learning may prove beneficial.

Below, we share some of the most promising use cases for advanced analytics in higher education to show how universities are capitalizing on those opportunities to overcome current challenges, both enabling access for many more students and improving the student experience.

The potential of advanced analytics in higher education

Advanced-analytics techniques may help institutions unlock significantly deeper insights into their student populations and identify more nuanced risks than they could achieve through descriptive and diagnostic analytics, which rely on linear, rule-based approaches (Exhibit 1). Advanced analytics—which uses the power of algorithms such as gradient boosting and random forest—may also help institutions address inadvertent biases in their existing methods of identifying at-risk students and proactively design tailored interventions to mitigate the majority of identified risks.

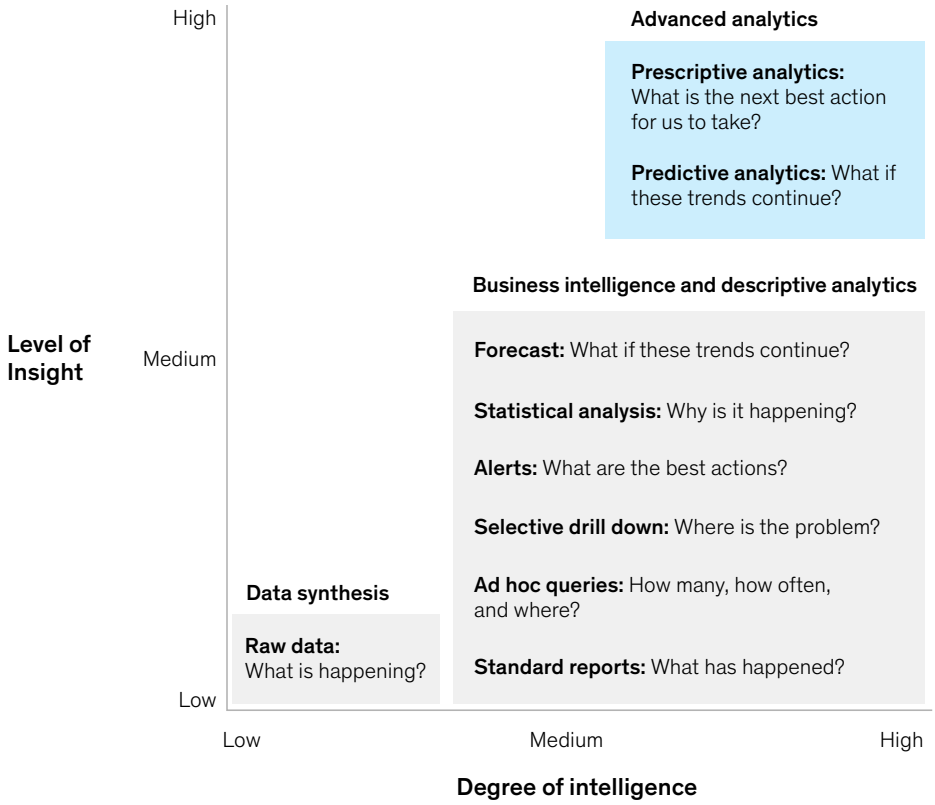
For instance, institutions using linear, rule-based approaches look at indicators such as low grades and poor attendance to identify students at risk of dropping out; institutions then reach out to these students and launch initiatives to better support

[1] "Available on-demand: Improving student success with a unified approach to data analytics and AI," Databricks, accessed December 2021; "About graduation rates," Western Governors University, December 2, 2021.

Data science and machine learning may unlock significant value for universities by ensuring resources are targeted toward the highest-impact opportunities to improve access for more students, as well as student engagement and satisfaction.

Exhibit 1

Advanced analytics is more sophisticated than other common approaches and could provide a competitive advantage.



them. While such initiatives may be of use, they often are implemented too late and target only a subset of the at-risk population. Linear approaches also fail to address two of the primary problems facing student success leaders at universities. First, risk of attrition is affected by too many variables to assess using a linear approach (such as academic, financial, and mental-health factors, and sense of belonging). Second, while it's easy to identify notable variance on any one or two variables, it is challenging to identify nominal variance on multiple variables. Linear, rule-based approaches therefore may fail to identify students who, for instance, may have decent grades and above-average attendance but have been struggling to submit their assignments on time or have consistently had difficulty paying their bills (Exhibit 2).

A machine learning model could address both of the challenges described above. Such a model looks at ten years of data to identify factors that could help a university make an early determination of a student's risk of attrition. For example, did the student change payment methods on the university portal? How close to the due date does the student submit assignments? Once the institution has identified students at risk, it can proactively deploy interventions to retain them.

Though many institutions recognize the promise of analytics for personalizing communications with students, reducing attrition, and improving student experience and engagement, institutions could be using these approaches for the full range of use cases across the student journey—

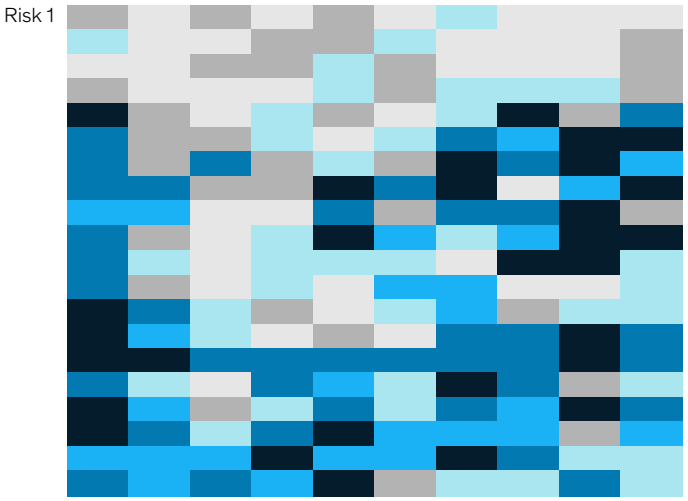
Exhibit 2

Machine learning techniques can surface insights using complex and unstructured data sets.

Real historical data on prospective candidates, likelihood of attrition

High likelihood Low likelihood

Real-life phenomena exhibit complex nonlinear patterns



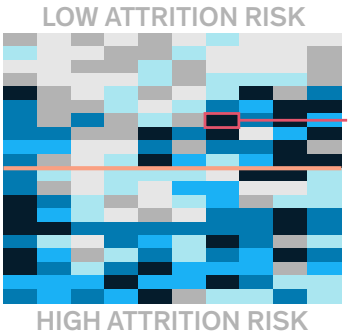
Example
Machine learning can be used in making the decision to reach out to a particular student at risk of attrition:

- Student has a 3.2 GPA
- Student has excellent community engagement and extracurriculars, and has demonstrated leadership
- Excellent attendance and on-time assignment submissions
- First-generation college student and has some delayed payments

---

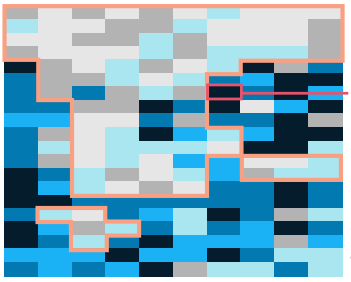
Traditional analytics rules

- Fit the phenomenon data into a predetermined "shape"
- Incorrectly identify the case in point as "low likelihood"



Machine learning algorithms

- Spot and record subtle patterns without clinging to a predetermined "shape"
- Correctly identify the case in point as "high likelihood"



Traditional analytics

- Algorithm is designed by humans to achieve expected output
- Any update of the algorithm based on changes in real-world conditions requires manual detections and updates
- Number of variables is restricted, limiting the amount of information taken into consideration in each decision

Machine learning

- The algorithm is learned by a computer through iterations of slightly different algorithms matching input and output
- Algorithm is continuously updated and adjusted based on difference between model output and real-world feedback
- High number of variables taken into account when making decisions, leading to a better-informed decision

for prospective, current, and former students alike. For instance, advanced analytics can help institutions identify which high schools, zip codes, and counties they should focus on to reach prospective students who are most likely to be a great fit for the institution. Machine learning could also help identify interventions and support that should be made available to different archetypes of enrolled students to help measure and increase student satisfaction. These use cases could then be extended to providing students support with developing their skills beyond graduation, enabling institutions to provide continual learning opportunities and to better engage alumni. As an institution expands its application and coverage of advanced-analytics tools across the student life cycle, the model gets better at identifying patterns, and the institution can take increasingly granular interventions and actions.

Deploying machine learning to harness this potential

Institutions will likely want to adopt a multistep model to harness machine learning to better serve students. For example, for efforts aimed at improving student completion and graduation rates, the following five-step technique could generate immense value:

1. Analyze 150 or more attributes from multiple years of historical data to understand the characteristics of a "successful student"—that is, someone who graduated within four years.
2. Define the goal for improving student success for key student segments as compared with a baseline; for example, an institution might aim to improve the graduation rate by five percentage points within a particular time frame.
3. Build an initial machine learning model using historical data to identify 30 to 50 attributes that indicate a high risk of attrition, then measure the model's effectiveness against a baseline, such as the university's existing methods for measuring student attrition.
4. Based on these attributes, build archetypes of students at risk of attrition and backtest for population skews or biases.
5. Develop and implement tailored interventions best suited for students in each archetype.

Institutions could deploy this model at a regular cadence to identify students who would most benefit from additional support.

Institutions could also create similar models to address other strategic goals or challenges, including lead generation and enrollment. For example, institutions could, as a first step, analyze 100 or more attributes from years of historical data to understand the characteristics of applicants who are most likely to enroll.

Institutions will likely want to adopt a multistep model to harness machine learning to better serve students.

Advanced analytics in action: How institutions have improved enrollment, retention, and, ultimately, the student experience

The experiences of two higher-education institutions that leaned on advanced analytics to improve enrollment and retention reveal the impact such efforts can have.

A private nonprofit university's effort to reach more students

One private nonprofit university had recently enrolled its largest freshman class in history and was looking for the next frontier of improvements. The institution wanted to both reach more prospective first-year undergraduate students who would be a great fit for the institution but don't currently apply and improve conversion in the enrollment journey in a way that was manageable for the enrollment team without significantly increasing investment and resources. The university took three important actions:

— *Allocating 'top of funnel' marketing spending to those most likely to apply.* The university developed a machine learning model using advanced analytics to predict which leads (prospective students) were most likely to apply. As a result, the university could identify the top 10 percent of leads, which accounted for about 90 percent of applicants. This enabled the team to immediately pivot its outreach efforts for the subsequent fall to prioritize the top 10 percent of leads yet to apply and ensure a higher return on investment for that outreach. In the future, this gives the institution the flexibility to either decrease its marketing spending to achieve the same number of applicants or maintain levels of spending to create a larger and potentially more competitive applicant pool.

— *Focusing yield efforts on archetypes that predict a high likelihood of matriculation.* To complement the advanced-analytics model for predicting which prospective students would apply, the institution developed a similar model for predicting which applicants would enroll. The model incorporated the wealth of additional data generated in the application process and broader demographic data, enabling the university to identify the top 40 percent of applicants, who accounted for about 85 percent of enrollment. Advanced analytics could then segment the high-potential applicants into five archetypes, with varying levels of expected conversion. For example, one archetype was characterized by students who sought out the university (that is, they came from organic sources) based on strong interest in particular arts programs, with roughly one in three of these applicants enrolling. This archetype segmentation enables the university to better prioritize and tailor its approach to applicants during the yield period. It also gives the institution future flexibility in targeting enrollment growth versus other strategic enrollment management priorities.

— *Identifying undertapped 'look-alike' markets*: The integration of demographic and other regional data enabled the institution to not only prioritize high-potential future enrollees within the markets where it currently recruits but also identify "look-alike" markets. Look-alike markets share predictive characteristics with markets that tend to have a high share of enrolled students, but they are not actively prioritized for recruitment by the college for various reasons, such as one-off past experiences or because they're less obvious fits. Through list buys that target specific counties, the university could increase its reach in look-alike markets and expand its applicant pool by 15 to 20 percent overall by prioritizing spending in look-alike markets over those with a lower likelihood of conversion.

For this institution, advanced-analytics modeling had immediate implications and impact. The initiative also suggested future opportunities for the university to serve more freshmen with great enrollment management spending efficiency. When initially tested against leads for the subsequent fall (prior to the application deadline), the model accurately predicted 85 percent of candidates who submitted an application, and it predicted the 35 percent of applicants at that point in the cycle who were most likely to enroll,

assuming no changes to admissions criteria (Exhibit 3). The enrollment management team is now able to better prioritize its resources and time on high-potential leads and applicants to yield a sizable class. These new capabilities will give the institution the flexibility to make strategic choices; rather than focus primarily on the size of the incoming class, it may ensure the desired class size while prioritizing other objectives, such as class mix by program or financial-aid allocation.

An online university's aspiration to enable more student success

Similar to many higher-education institutions during the pandemic,[2] one online university was facing a significant downward trend in student retention. The university explored multiple options and deployed initiatives spearheaded by both academic and administrative departments, including focus groups and nudge campaigns, but the results fell short of expectations.

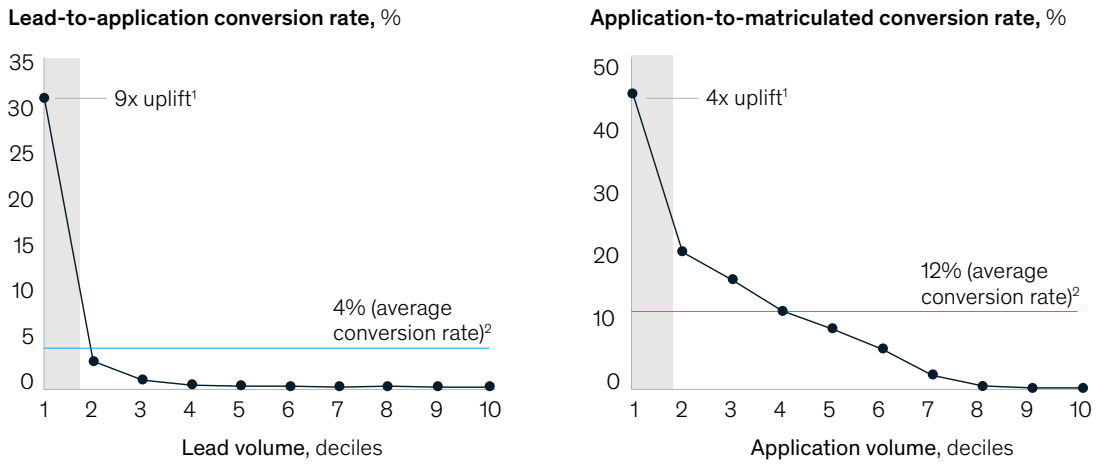
The institution wanted to set a high bar for student success and achieve marked and sustainable improvements to retention. It turned to an advanced-analytics approach to pursue its bold aspirations.

To build a machine learning model that would allow the university to identify students at risk of attrition early, it first analyzed ten years of historical data to understand key characteristics that differentiate students who were most likely to continue—and thus graduate—compared with those who unenrolled. After validating that the initial model was multiple times more effective at predicting retention than the baseline, the institution refined the model and applied it to the current student population. This attrition model yielded five at-risk student archetypes, three of which were counterintuitive to conventional wisdom about what typical at-risk student profiles look like (Exhibit 4).

[2] "Persistence and retention," National Student Clearinghouse Research Center, July 8, 2021.

Exhibit 3

Advanced-analytics modeling can predict lead-to-application and application-to-matriculation rates.



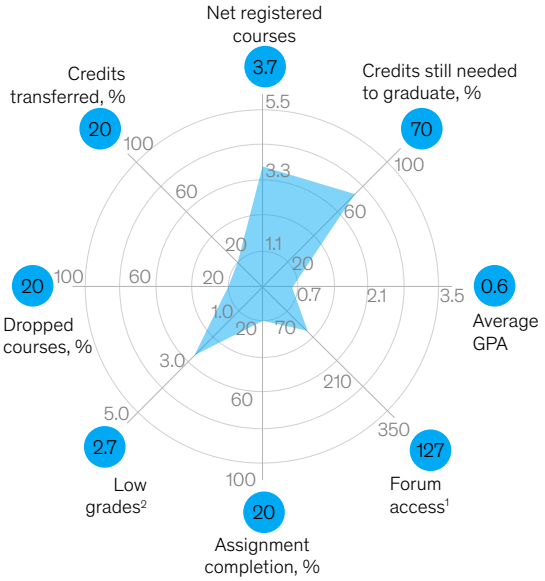
[1]"Uplift" indicates the conversion achieved in the top decile, as ordered by predictive model score, divided by current average conversion rate.
[2]Average conversion rate based on 2019 client data.
Source: McKinsey analysis

Exhibit 4

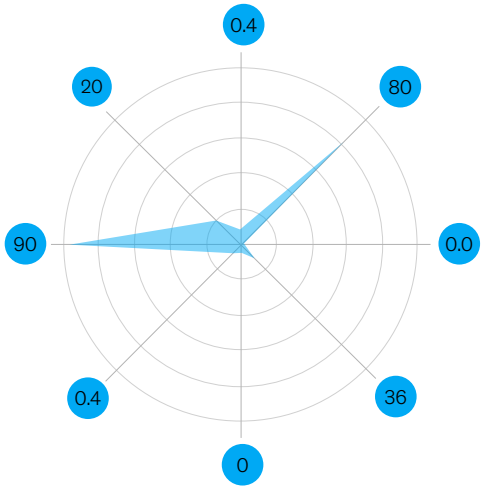
The advanced-analytics model identified five at-risk student archetypes, three of which would not have emerged based on linear rules.

Identifiable through linear rules (~30% at-risk students)

1. Students with academic needs

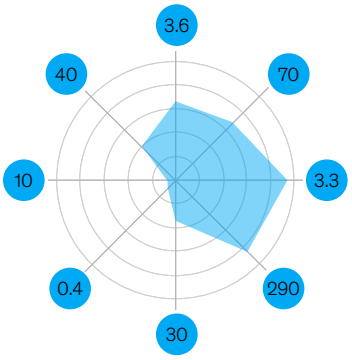


2. Noncommittal students

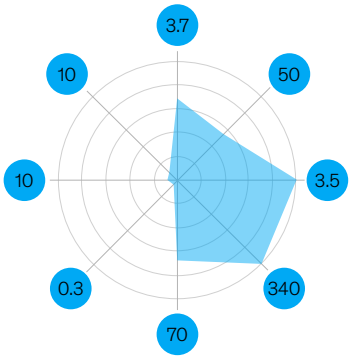


Additional archetypes identified only through advanced analytics (~70% at-risk students)

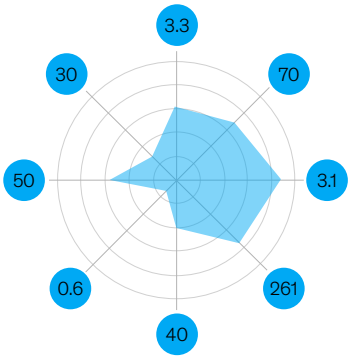
3. Committed to current path



4. Distinctive academic achievers



5. Average academic performers



[1] Number of times a student has accessed online student forums.
[2] Low grades are defined as follows: for undergraduate students, any grade below D is considered a low grade. For graduate students (for master's degrees), any grade below C is considered a low grade. Note that this definition is specific to the client for whom the case study was performed. Universities may choose to define this differently.
Source: McKinsey analysis

Together, these three counterintuitive archetypes of at-risk students—which would have been omitted using a linear analytics approach—account for about 70 percent of the students most likely to discontinue enrollment. The largest group of at-risk individuals (accounting for about 40 percent of the at-risk students identified) were distinctive academic achievers with an excellent overall track record. This means the model identified at least twice as many students at risk of attrition than models based on linear rules. The model outputs have allowed the university to identify students at risk of attrition more effectively and strategically invest in short- and medium-term initiatives most likely to drive retention improvement.

With the model and data on at-risk student profiles in hand, the online university launched a set of targeted interventions focused on providing tailored support to students in each archetype to increase retention. Actions included scheduling more touchpoints with academic and career advisers, expanding faculty mentorship, and creating alternative pathways for students to satisfy their knowledge gaps.

Advanced-analytics risks to keep in mind

Advanced analytics is a powerful tool that may help higher-education institutions overcome challenges, spur growth, and better support students. However, machine learning is complex, with considerable associated risks. While the risks vary based on the institution and the data included in the model, higher-education institutions may wish to take the following steps when using these tools:

1. Build and train models to ensure they don't accidentally introduce biases informed by race, age, or gender. Also ensure that new models are not inadvertently building on inherent accidental biases in current methods.

2. Focus models on use cases that involve supporting and including students as opposed to any decisions that suggest excluding students from certain interventions; the models also should explicitly test factors to remove unconscious bias from any decision making connected to the point above.

3. Use results and insights from machine learning models together with, and as input for, existing student support processes. Machine learning models provide additional insights to inform interventions; they should not be used as a replacement for existing structures and methods.

4. Consistently check the performance of the model for different student segments to ensure it performs relatively similarly for all segments and is not skewed toward any particular group.

While many higher-education institutions have started down the path to harnessing data and analytics, there is still a long way to go in realizing the full potential of these capabilities in terms of the student experience. The influx of students and institutions that have been engaged in online learning and using technology tools over the past two years means there is significantly more data to work with than ever before; higher-education institutions may want to start using it to serve students better in the years to come.

Claudio Brasca is a partner in McKinsey's San Francisco office, where **Varun Marya** is a senior partner; **Nikhil Kaithwal** is an associate partner in the London office; **Charag Krishnan** is a partner in the New Jersey–Summit office; **Monatrice Lam** is a consultant in the Bay Area–Silicon Valley office; and **Jonathan Law** is a senior partner in the Southern California office.

The authors wish to thank Inès Garceau-Aranda, Emily Cohen, Katie Owen, Xiaowo Sun, Xuecong Sun, and Shyla Ziade for their contributions to this article.