USING BIG DATA TO PREDICT STUDENT DROPOUTS: TECHNOLOGY AFFORDANCES FOR RESEARCH

David Niemi[1] and Elena Gitin[2]
[1]Kaplan Inc., Fort Lauderdale, FL
[2]Kaplan University, Fort Lauderdale, FL

ABSTRACT

An underlying theme of this paper is that it can be easier and more efficient to conduct valid and effective research studies in online environments than in traditional classrooms. Taking advantage of the “big data” available in an online university, we conducted a study in which a massive online database was used to predict student successes and failures. We found that a pattern of declining performance over time is a good predictor of the likelihood of dropping out, and that having dependents or being married or in the military reduces the risk of dropping out. The risk of dropping out was higher for older students, females, and students with previous college education or transfer credits. These results provide a foundation for testing interventions to help students who are at risk and will also help to inform the development of a “research pipeline” that will enable rapid experimental studies of new tools and strategies.

KEYWORDS

Virtual education, Online learning, Technology

1. INTRODUCTION

The early years of the 21st century have witnessed an explosion of interest in new technologies for learning and teaching at all levels of education. Notable examples of applications of technology in education include the Khan Academy and the massive open online courses (MOOCS) offered by edX, Coursera, Udacity and other entrepreneurial partnerships. Non-profit and for-profit online universities now provide a wide variety of postsecondary training, certificates, course credits, and degrees, and traditional colleges and universities have greatly expanded their online learning opportunities. These and a startling proliferation of other online opportunities have made it possible for students and educators to pursue a range of alternative learning and teaching paths not available to them before.

Whether new forms of online instruction can be effective in educating students on a large scale across the world remains a vexing question, however. At the moment, the evidence is lacking, but the opportunity to build evidence does exist. The growth of online instruction across the globe in fact opens new opportunities for data collection, analysis and reporting and for studies of the effectiveness of research-based instructional strategies, among other things (e. g., Baker, in press; Campbell et al., 2007; Cen et al., 2006; Desmarais et al., 1996; Means et al., 2010; Romero et al., 2008; Romero et al., 2011).

These opportunities include the possibility of using routinely collected data to evaluate the relative effectiveness of current and innovative instructional approaches, as well as the opportunity to provide better information to guide learners and their teachers. In online environments it is possible to collect detailed real-time data on every action taken by every learner. The existence of these data for thousands, tens of thousands, and even millions of students studying the same topics or using the same curriculum under different conditions, gives us new leverages for studying the influence of contextual factors on learning and learners. This “big data” affordance can help learners by identifying which learning paths might be best for them, teachers by recommending approaches for helping students who are struggling, and researchers by enabling them to test principles of learning and instruction in authentic learning environments at scale.

As a first step toward capitalizing on these opportunities, we conducted an initial investigation intended to use large existing datasets to predict student success and failure in an online university program. Having developed algorithms to flag students needing additional support, we then plan to use both data mining and experimental methods to test which types of support will be most effective and efficient for which students.

2. BODY OF PAPER

2.1 Analyzing "Big Data" to predict Student Dropouts

The study we conducted involves the mining and analysis of "big data", which in our case refers to large existing datasets that can be analyzed to discern patterns in student and teacher performance, identify at-risk students, and study relationships among important variables, such as attendance, learning, and student satisfaction.

We conducted this study in an online university that offers approximately 1000 different online courses to about 60,000 students and that has assembled an extensive database of information on students and their performance. This university tends to enroll students who are older and for various reasons need an alternative to a traditional classroom-based university program.

The database we analyzed contains student and faculty background data as well as measures of learning, student satisfaction, retention, engagement, teacher performance, and postgraduate success. These data make it possible to test a wide range of research and evaluation questions on a very large scale, as well as to monitor and respond to student performance, engagement, and motivation.

The study was specifically designed to examine how students' academic and demographic characteristics relate to their dropout rates. We analyzed academic and demographic characteristics of degree-seeking students (N = 14791) enrolled during a two-year period in the online university. Demographic variables analyzed were age, gender, marital status, military status, previous college education, estimated family financial contribution, and number of transfer credits from other universities. Academic variables included measures of student performance available in the online eCollege platform, such as final exam, discussion, project, and other assignment scores.

Survival analyses (a type of logistic regression analysis) revealed that measures of student performance that declined over time were significant predictors of the likelihood of dropping out. With respect to the demographic predictors, we found that age was a significant predictor of retention, with older students more likely to drop out. Military and married students retained at a higher rate than non-military and unmarried students, respectively, and students with prior college experience and higher financial contributions from their families retained at lower rates than students without these characteristics.

Results of the analyses are presented in Table 1, with the last column showing the effect of each predictor variable on the risk of students dropping out. The outcome we modeled is student retention, coded as a dichotomous variable with students who dropped out receiving a "1" score on the variable.

For comparison purposes we also calculated odds ratios based on a logistic regression analysis; this analysis produced the same outcomes as the survival analysis.

Table 1. Predicting student dropouts

Predictor	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Effect on dropout risk
Age between 29 and 38	1	-0.50719	0.04485	127.9116	<.0001	0.602	Reduces risk by 40%
Age between 38 and 45	1	-0.07808	0.04767	2.6833	0.1014	0.925	Reduces risk by 7.5%
Older than 45	1	0.26797	0.04479	35.7943	<.0001	1.307	Increases risk by 30.7 %
Has transfer credits	1	0.86241	0.04427	379.5799	<.0001	2.369	Increases risk by 236 %
Enrolled in 200 level courses	1	0.07112	0.07777	0.8365	0.3604	1.074	Increases risk by 7.4%
Enrolled in 300 level courses	1	-0.31259	0.07711	16.4345	<.0001	0.732	Reduces risk by 26.8%
Enrolled in 400 level courses	1	-1.07894	0.08296	169.1268	<.0001	0.34	Reduces risk by 66%
In the military	1	-1.3035	0.08964	211.4631	<.0001	0.272	Reduces risk by 72.8%
Previous college education	1	0.12529	0.04324	8.3962	0.0038	1.133	Increases risk by 13.3 %
Female	1	1.20342	0.06681	324.4385	<.0001	3.331	Increases risk by 330 %
Estimated financial contribution from family	1	0.15667	0.03569	19.275	<.0001	1.17	Increases risk by 17 %
Married	1	-0.44225	0.03807	134.98	<.0001	0.643	Reduces risk by 35.7%
Has dependents	1	-0.86645	0.05448	252.9138	<.0001	0.42	Reduces risk by 58%
Above median score on discussions	1	-0.72133	0.0497	210.6254	<.0001	0.486	Reduces risk by 51.4%
Above median score on final exam	1	-1.4674	0.16955	74.9057	<.0001	0.231	Reduces risk by 76.9%
Above median score on projects	1	-0.96057	0.06153	243.6801	<.0001	0.383	Reduces risk by 61.7%
Above median score on course review	1	-1.04862	0.05681	340.7227	<.0001	0.35	Reduces risk by 65%
Above median score on other teacher-graded assignments	1	-0.98675	0.05474	324.9456	<.0001	0.373	Reduces risk by 62.7%

3. CONCLUSION

3.1 Moving Toward Rapid Experimentation

The finding that declining performance over time is related to dropout tendencies, while not surprising, is a particularly useful one for us. It represents a first step on the path toward generating data for faculty and administrators that will enable them to provide additional support to students who have a high likelihood of failing or dropping out. Results on student background variables will serve a similar function, enabling us to build profiles of students who may need more support to succeed in the program. Some of our findings for these variables were surprising; for example that females, unmarried students and students with transfer credits or dependents are more likely to drop out. We can speculate about these relationships but it will be more useful to conduct follow-up investigations, including qualitative studies with representative samples of students, to explore why some groups drop out at higher rates.

As we continue to collect data over time we can also confirm or disconfirm the strength of the relationships we have found and possibly discover new ones. We are currently developing prior knowledge measures in several courses and expect, based on extensive previous research on the effects of prior knowledge on learning (e. g., National Research Council, 1999; Sweller et al, 2011), that these will be strong predictors of student success, failure and retention in online courses. Ultimately we will be able to determine which combinations of indicators and performance patterns constitute “red flags” requiring immediate, strong intervention and which may call for less intensive strategies.

As valuable as these analyses of routinely-collected big datasets may be, however, they give us only part of the information needed to build more effective evidence-based educational programs. Another critical piece is the testing of new instructional, motivational and support strategies to help students who are having difficulty and may be likely to drop out. To determine what kinds interventions will be most effective for which students will require experimental studies in which students are randomly assigned to different instructional conditions, so that alternative hypotheses for observed changes or differences in student performance (such as differences in initial student ability levels) can be ruled out. In this case we can take advantage of another technology affordance for research: online learning systems make it possible to run large numbers of randomized control trials far more rapidly than would otherwise be possible (Shadish and Cook, 2009).

We are currently developing “research pipelines”, or platforms and processes that will to enable us to randomly assign individual students, class sections, or classes to different instructional experiences. The pipelines will make it possible to build up knowledge on scalable improvements incrementally, testing the impact of one variable at a time, but quickly over time. Big data analytics in combination with experimental testing will thus enable us take advantage of the tremendous scale of the online university to find out what works more quickly than would otherwise be possible–and have the data to prove it. (As a final note, it is worth mentioning, however, that the online university in this study primarily serves a high percentage of students who for one reason are another are not able to attend or are not interested in attending traditional college and university programs, so results may not generalize to students in those programs.)

REFERENCES

Baker, R.S.J.d, in press. Data Mining for Education. To appear in McGaw, B., Peterson, P.,Baker, E. (Eds.) *International Encyclopedia of Education (3rd edition)*. Elsevier, Oxford, UK.

Campbell, J. P., DeBois, P. B., and Oblinger, D. G., 2007. Academic Analytics: A New Tool for a New Era. *EDUCAUSE Review*. EDUCAUSE, http://net.educause.edu/ir/library/pdf/erm0742.pdf

Cen, H., Koedinger, K., Junker, B., 2006. Learning Factors Analysis - A General Method for Cognitive Model Evaluation and Improvement. *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, Jhongli, Taiwan, pp. 12-19.

Desmarais, M.C., Maluf, A., and Liu, J., 1996, User-expertise Modeling with Empirically Derived Probabilistic Implication Networks. *In User Modeling and User-Adapted Interaction*, Vol. 5, No. 3-4, pp. 283-315.

Mayer, R. E., amd Alexander, P. A. (Eds.), 2011. *Handbook of Research on Learning and Instruction*. Routledge, New York, USA.

Means, B., Toyama, Y., Murphy, R., Bakia, M. & Jones, K., 2010. *Evaluation of Evidence-based Practice in Online Learning: A Meta-analysis and Review of Online Learning Studies*. US Department of Education Office of Planning, Evaluation, and Policy Development. http://www2.ed.gov/rschstat/eval/tech/evidence-based-practices/finalreport.pdf

National Research Council, 1999. *How People Learn.* National Academy Press, Washington DC, USA.

Romero, C., Ventura, S., Pechenizkiy and Baker, R.S.J.d, (Eds.), 2011. *Handbook of Educational Data Mining.* CRC Press, Boca Raton, FL, USA.

Romero, C., Ventura, S., Espejo, P.G., Hervas, C., 2008. Data Mining Algorithms to Classify Students. *Proceedings of the First International Conference on Educational Data Mining*. Montreal, Canada, pp. 8-17.

Shadish, W. R. and Cook, T. D., 2009. The Renaissance of Field Experimentation in Evaluating Interventions. *In Annual Review of Psychology*, Vol. 60, No. 1, pp. 607–629.

Sweller, J., Ayers, P., & Kalyuga, S., 2011. *Cognitive Load Theory*. Springer, New York, USA.