Predictive Analytics Goes to College – to Predict Student Success

- Posted by William Vorhies on November 19, 2015 at 10:30am
- View Blog

Summary: *Higher education has been a little slow on the uptake to use advanced analytics to improve student success but now with the technology that allows us to marry and analyze structured and unstructured data, including streaming data, a number of successful projects are underway.*

Earlier this week I had the pleasure of moderating a webinar focusing on the work of two

Pivotal data scientists working with a prestigious mid-west university to use data to predict student success. It's a topic that has long interested me as I devoted a good deal of time trying to promote this type of project in the early 2000's.

At the time, I didn't have much luck. That could reflect on my skills as a salesman but on consideration it also illustrates how fast and how far our big data technologies have brought us. So after hosting the webinar (which I recommend for your viewing, *you can see it here*) I did a quick literature search and was gratified to see that in fact many colleges and universities are undertaking these studies.

One thing stood out just by examining the dates of the published literature. Prior to about 2007 what predictive analytics was performed tended to focus on the sort of static data you can find in a student's record: high school GPA, SAT scores, demographics, types of preparatory classes taken, last term's grades, and the like. That's pretty much all there was to draw on and there was some success in that period.

What changes in the most current studies is the extensive use of unstructured data integrated with structured data. It wasn't until about 2007 that our ability to store and analyze unstructured data took off and now we have data from a variety of new sources.

Learning Management Systems is one of the most important new sources. These are the on line systems used to interact with students outside of the classroom. From these we can learm for example when they submitted assignments relative to the deadline, how they interact with instructors and classmates in the chat rooms, and a variety of click stream data from library sites and the like.

Sensor and Wi-Fi data showing frequency and duration on campus or at specific locations like the library.

Student Information Systems. These aren't necessarily new but greatly improved in level of detail regarding classes enrolled and completed with regular grade markers.

Social Media. What is standard now in commerce is becoming a tool for assessment of progress or markers for concern. Positive and negative social media comments are evaluated for sentiment and processed as streaming data that can be associated with specific periods in a student's term or passage through to graduation.

The goals of each study are slightly different. Some are seeking better first year integration programs which are so important in student long term success. Some are focused on the transition from Community College to four year institution. But universally they tend to look at some similar markers that would allow counsellors and instructors to intervene. Some of those common markers are:

- Predicting first term GPA.
- Predicting specific course grades.
- Predicting reenrollment.
- Predicting graduation likelihood, some focused on getting students through in four years, others getting them through at all.

As in any data science project, each institution seems to have identified its own unique set of features drawn from both the traditional structured and new unstructured data sources. Paul Gore who headed one of these studies at the University of Utah had a nice summary of the categories that's worth considering. He says the broad categories of predictive variables fall into these six groups:

Measures of academic performance:

Academic engagement or academic conscientiousness: in other words, how seriously does the student take the business of being a student? Does the student turn in assignments on time? Attend class diligently? Ask for help when needed?

Academic efficacy: the student's belief and confidence in their ability to achieve key academic milestones (such as the confidence to complete a research paper with a high degree of quality, or to complete the core classes with a B average or better, or their confidence in their ability to choose a major that will be right for them).

Measures of academic persistence:

Educational commitment: This refers to a student's level of understanding of why they are in college. Students with a high level of educational commitment are not just attending college because it is "what I do next" after high school (i.e., in order to attain a job or increase their quality of life); these students have a more complex understanding of the benefits of their higher education and are more likely to resist threats to their academic persistence.

Campus engagement: This is the intent or desire to become involved in extracurricular activities. Does the student show interest in taking a leadership role in a student organization, or participating in service learning opportunities, intramural sports, or other programs outside of the classroom?

Measures of emotional intelligence:

Resiliency: How well does the student respond to stress? Do small setbacks throw the student "off track" emotionally, or are they able to draw on their support network and their own coping skills to manage that stress and proceed toward their goals?

Social comfort*:* Gore notes that "social comfort is related to student outcomes in a quadratic way -- a little bit of social comfort is a good thing, while a lot may be less likely to serve a student well, as this may distract their attention from academic and co-curricular pursuits." (aka too much partying).

Where the studies were willing to share, the fitness measures of the predictive models look pretty good, achieving classification success rates in the 70% to 80% range.

From our data scientist friends at Pivotal who are featured in the webinar we also learn that administrators and counsellors are generally positive about the new risk indicators. There was always the possibility that implementation might be hampered by disbelief but there are some notable examples where there is good acceptance.

Some of the technical details are also interesting. For example, there are instances where the models are being run monthly to update the risk scores. This allows the college to act within the current term and not wait for the term to be over, which might be too late.

And there are examples in which the data is being consumed not only by administrators and counsellors but also being pushed directly to the students through mobile apps.

I originally thought to include a listing of the colleges that were undertaking similar projects but a Google search shows that there are a sufficiently large number that this is no longer a completely rare phenomenon. In its early stages to be sure but not rare.

Finally I was struck by one phenomenon that is not meant as a criticism, just an observation. Where the research and operationalization of the models was funded by say a three year grant, it took three years to complete the project. But where our friends at Pivotal were embraced by their client, four data scientists, two from Pivotal and two from the university had it up and running in three months. Just sayin.

November 19, 2015

Bill Vorhies, President & Chief Data Scientist – Data-Magnum - © 2015, all rights reserved.

About the author: Bill Vorhies is President & Chief Data Scientist at Data-Magnum and has practiced as a data scientist and commercial predictive modeler since 2001. Bill is also Editorial Director for Data Science Central. He can be reached at:

Bill@Data-Magnum.com or Bill@DataScienceCentral.com

The original blog can be seen here.