ENVISION

THE 5 WAYS ENROLLMENT PREDICTIONS ARE DRIVING COLLEGES OUT OF BUSINESS

Thom Golden, Ph.D.
Vice President of Data Science

Brad Weiner, Ph.D.
Director of Data Science

Pete Barwis, Ph.D.
Senior Data Scientist

capture

Table of Contents

The 5 Ways Enrollment Predictions Are Driving Colleges Out of Business

Strategic enrollment management (SEM) is, in its simplest form, a prediction business. The modern enrollment manager's employability depends on estimating and then producing a class with certain characteristics and financial dependencies. Incorrect estimates can cause dramatic consequences. Higher education institutions that struggle to forecast net-tuition revenues are experiencing new pressures for survival, especially in a competitive landscape that is increasingly divided between the "haves" and the "have-nots." Remarkably, the availability of advanced tools and methods to address this prediction challenge have been laggard, even for resource-flush institutions. This is especially puzzling given the techno-data revolution that is occurring in hundreds of other fields. In many of the same cities in which local law enforcement agencies are using terabyte-sized data sets and machine learning to anticipate crime before it happens, many universities still struggle to accurately forecast needed housing capacity the following fall. Even across campus, physicians at the university medical center are utilizing prescriptive analytics to diagnose and treat the most intractable forms of disease, while the enrollment manager two blocks away is presented with shockingly out-of-date prediction solutions dressed up as new innovations.

At Capture Higher Ed, we have argued that innovation in enrollment management currently finds itself at a "horse versus locomotive" moment. In our published **Capture Manifesto**, we contend that the conventional methods of university enrollment management can no longer keep pace with the demands of the market and that new ideas are imperative. The more modern equivalent however is that **the field of enrollment management prediction sits at a "cab versus Uber" moment.** While both transportation options will get you from one place to another, there is a greater awareness now that you have a choice as compared to years ago when cabs ruled. A more direct question: Is the lack of innovation in the higher education prediction industry a coalescing around best practices, or simply an artifact of an industry that seems to believe it is the only game in town? Either way, we believe that higher education deserves a choice.

Strategic Enrollment Managers don't have the luxury of waiting for new thinking from old players in the higher ed prediction space. Now, more than ever, predicting enrollment is an integral precursor to university financial forecasting and has never been more important.

Dr. Thom Golden

> Higher education deserves a choice."

Given the critical importance of accurate predictions, it is alarming that the lion's share of prediction methodology used by enrollment managers is (charitably) decades old, and dominated by linear models and frequentist statistics, which are (charitably) outmoded compared to newer methods of prediction. There are signs that modern data science is forcing an evolution of these traditional methods, as hinted by the American Statistical Association's recent suggestion of retiring hypothesis testing with a p value, a mainstay of social science statistics responsible for the idea that some findings are "significant" and therefore more likely to be published.

Unfortunately, SEM practitioners don't have the luxury of waiting for new thinking from old players in the higher ed prediction space. An *Inside Higher Ed* survey found that 1 in 4 American private colleges are threatened with closure in the foreseeable future, according to those institutions' CFOs. In the same study, 88% of university chief business officers will be looking at increasing or otherwise optimizing enrollment to right the ship. Unfortunately, the enrollment prediction industry that would be enlisted to assist with this endeavor has been dominated by a small number of established firms with very little competition. In fact, according to *IBISWorld*, the educational consulting industry is remarkably small, with less than half as many firms as the healthcare consulting industry (7,268 compared to 16,086) and is barely a blip within the broader U.S. management consulting market (799,884). Despite projected growth in the higher education advisory services sector, an argument can be made that lower levels of competition has flattened the innovation curve in a space desperate for new solutions.

So, now more than ever, predicting enrollment is an integral precursor to university financial forecasting and budgeting and has never been more important. Consider this 2014 *Inside Higher Ed* survey:

- **61%** of institutions missed their enrollment goals last year
- **32%** were still recruiting applicants after May 1
- But . . . **73%** of 4-year private institutions (67% of 4-year publics) are using a statistical model to predict enrollment outcomes (according to a 2016 Ruffalo Noel-Levitz study)

Something doesn't add up.

So what is happening here? We believe there are at least 5 reasons the SEM prediction industry has dramatically missed the mark and is putting higher education, a field we love immensely, at great risk:

1) Treating adolescent decision-making as linear

Anyone who has ever met an adolescent knows that the teenage years are many things, but one thing they are not, is linear. There is no point A to point B, mostly because point A is not speaking to point B because that [expletive] didn't like point A's recent Instagram post. Still, nearly all of the predictive models in higher education utilize linear and log-linear regression, a statistical methodology rooted in Pearson's famous 1875 garden of pea pods that has served as a social science workhorse for well over a century.

This is not to say that linear regression models are only used by old farts who studied stats in the '70s. They are used extensively today, even though they are older, because they are reliable, interpretable, efficient to run, and because logistic regression is very often the final section in an intro stats class.

Despite their widespread use, they operate on several questionable assumptions when used to predict the behaviors of teenagers. First, they both assume a linear (or log-linear for logit models) relationship between each predictor and the outcome. Predictors, known as "independent variables" are assumed to have their own isolated, linear effects on an outcome. For example, each additional time a teenager visits campus, their likelihood of enrolling increases by the size of the coefficient for campus visit. If the true relationship between the number of campus visits and the probability of enrolling is not linear, then it is impossible to estimate the effect size of a campus visit in any meaningful way.

As it turns out, **the assumption of a linear relationship between predictors and an outcome rarely holds up in real life.** Real life is usually more complicated than a teeter-totter way of modeling each predictor against an outcome. These complications can be thought of as 'non-linearities,' or non-linear relationships between a predictor and an outcome. For example,

a campus visit may matter in a different way for athletes as compared to non-athletes. Recruited athletes might get a meal in the cafeteria, front row tickets to a football game, and meal at the training table, while non-athletes might get a 30-minute information session, validated parking, and a student-led tour. Surely the effects of those visits are different. **Linear models can be informative when little information is available.** However, given the increasingly vast array of online and offline data available to enrollment managers, there is no reason to hinder predictive accuracy due to data limitations.

2) Choosing interpretability over accuracy

The Weather Channel's Jim Cantore loves the weather. He. Loves. It. What else would compel a sane person to hold a microphone in an electrical storm while bracing 40 MPH winds? Unless you are like Jim though, you typically don't care how the weather prediction gets made; you just want to know if you need an umbrella when heading outside. We'll talk more about weather forecasting as an analogy for enrollment prediction later, but for now, consider this statement: As a predictive model's complexity increases, its interpretability decreases. This tension exists at the heart of a significant nerd battle (the best kind if you ask us) between traditional statisticians and computer scientists. In the SEM space, the interpretable model often is framed as a "7 factor model" with weighting of certain variables that form the recipe of exactly what is going into the prediction and in what proportion. The machine learning approach is less concerned with providing such interpretable results—which may be exceedingly complex, in deference to prediction accuracy. Thus goes the argument:

Traditional Statistician: After reading 20 academic papers, and understanding their results, it seems pretty clear that distance from campus is strongly correlated with propensity to enroll.

Machine Learning Practitioner: After testing 10,000 variables using 75 algorithms, we found that the price of coffee exports during the rainy

> "As a predictive model's complexity increases, its interpretability decreases."

season in Suriname is actually a better predictor than the campus visit. The punch line though is that the truth lies somewhere between these technocratic Jets and Sharks. It all depends on the outcome you are seeking and the strategic value thereof. In situations where understanding *what* makes a student more likely to enroll is more critical to recruitment strategy, a frequentist statistical model with its relatively clean explanations can be quite valuable. Likewise, when accuracy is of paramount interest, then bring on the "algos."

The central issue however is that the SEM prediction industry has never provided much of a choice between interpretable models and accurate ones. Maybe there are some enrollment managers out there who care more about *why* that email from the Nigerian prince was deemed to be spam than they do about the overall accuracy of their spam filter. I am betting though, that most enrollment managers value predictive accuracy above all else, especially in our overstuffed email inboxes. Those of us who make predictions for a living should plan accordingly.

3) Not adequately testing models in real world scenarios

Simply put, prediction is a method for using known information about the past to estimate a similar outcome for an unknown future. In practice, this involves fitting a mathematical equation to a set of your historical data, and then applying that same equation to data for which the outcome is still unknown.

The challenge, of course, is that the outcome of the future is likely to be very similar, but not identical, to the outcomes of the past. Therefore, it is necessary to build a model that is closely aligned with the historical data, but not too closely aligned. This process is called "specifying" or "fitting" a model, and a clothing metaphor is apt. A model that is underfitted to your data is like a baggy sweatshirt; its fit is too non-specific to show off a great body. A model that is "overfitted" is like a neoprene wetsuit leaving very little room for adjustments (and less to the imagination). The former will not take full advantage of the predictive power of the data, and the latter will take so much advantage that it falls apart when it is applied to data it

"The outcome of the future is likely to be very similar, but not identical, to the outcomes of the past."

So how do we solve this problem?

Glad you asked.

hasn't seen before. So how do we solve this problem? Glad you asked. In the world of enrollment prediction, the body of historical data is called the training set—in that the modeler uses it to "train" the model to know what to expect when it comes time to consider new data at prediction go time. But, how do you choose how much data to use for model specification and how do you know how well the model performs? The method to which we are partial starts with historical data of no less than 5 years, which is then sub-divided into a training set and a testing set that is considered a "holdout" because it is completely firewalled from the data used to build the model. Once the model is trained, the model then gets a dry run on the holdout set and thus provides us a deep understanding of that model's accuracy, which can help us make the model even better. Here again, the nerd battle rages in terms of what is the ideal way to build a more accurate model. Traditional prediction methods often utilize only 2-3 years of data, but use the entire historical training set to build the model. Any testing that is conducted is done through a method called resampling, in which data are randomly selected from the training set itself and tested against the model. **There are two issues with this:**

1. Resampling may miss methodological errors baked into the analysis whereas a holdout set provides a methodological fail safe.
2. Resampling is at a greater risk of overfitting the model, which is likely to dramatically reduce accuracy when applied to new data.

Traditional statistical models that currently dominate the enrollment prediction space are simply inadequate at generating the most accurate models primarily because they are not tested in game-day scenarios. Their predictions are most likely fitted to the data that was used to construct them, which will lead to less accurate predictions later. Instead, these traditional models rely on "goodness of fit" statistics that laud the amount of variance in the outcome that is explained by the model, rather than on fit statistics that assess the accuracy of each prediction.

4) Trying to forecast October's weather on January 1

The *Old Farmer's Almanac* is a veritable treasure trove of goodness. Want to know the order in which you should plant your beans and cabbage, your horoscope a year in advance or the best recipe of Irish Soda Bread? Covered. You know what it's not good at? Weather forecasting.

Indeed, the *Old Farmer's Almanac*, published out of Lewiston, Maine, has long been issuing extended-term weather predictions based on the positioning of planets, sun spots, and tidal patterns. How colloquial, you might say. A quasi-scientific industry clinging to folk theories that at one point dominated meteorological science to predict Halloween's weather before Martin Luther King Day. Endearing even. Grandmothers and Farmer Greyhair relying on the reputation of the almanac to accurately predict complex weather patterns months in advance. In many ways though, the SEM prediction industry has taken this very approach to forecasting the highly volatile and evolving behavior of adolescents. Many of the predictions being made are calculated at the beginning of an admissions cycle and rarely, if ever revisited.

Why is this a horrible idea?

Many conditions are simply not yet known, which is why Nate Silver—in his ground-breaking book *The Signal and the Noise: Why Most Predictions Fail but Some Don't*—cited the increase in accuracy between a 10-day weather forecast and the 24-hour forecast. Higher education must plan its prediction methodologies similarly. Student behavior ebbs and flows and students move toward and away (and then potentially back again) from the schools they are considering. Famous on-air salesman and purveyor of the "Chop-o-Matic," Ron Popiel may have been on to something with his "set it and forget it" rotisserie cookers, but then again, Ron was never a dean of admissions. Set it and forget it might work for a pot roast but is not an effective strategy for building a diverse and prepared class.

Our colleagues at Capture Higher Ed have been working on an alternative to the current practice of training a model once and filtering analysis through it over the course of one (or even two) year(s). Instead, we believe that the dynamic nature of SEM demands a more flexible approach in which a process called "feature selection" can adjust how the model performs mid-stream. Enrollment predictions are a special type of "self-cancelling predictions" in that an admissions shop that receives these predictions will naturally take action on them to either make sure that prediction does or does not (if it's less than optimal) occur. In this way (and hopefully only in this way) enrollment predictions are similar to flu outbreak predictions. When public health organizations like the Centers for

Set it and forget it might work for a pot roast but is not an effective strategy for building a diverse and prepared class.

Dr. Thom Golden

Disease Control (CDC) predict greater than normal flu levels in the coming year, the broader health industry reacts by producing more vaccines and publicizing the need for people to get their flu shots. As a result, incidences of flu decrease from their original prediction. The original predictions were not inaccurate per se, only that there was an intervening collection of actions that changed the trajectory of the predicted outcome. So goes enrollment predictions.

How then, can any serious enrollment prediction, be made once and not adjusted as conditions unfold throughout an enrollment cycle?
It represents a fundamental misunderstanding of not only the dynamic nature of enrollment management but of the nature of prediction itself.

5) Bringing too few tools to the job site

As we've discussed previously, the challenges presented by the non-linear relationships in enrollment management render a "linear-or-bust" toolset obsolete. The good news is that recent advances in computer science and statistics have produced a large set of predictive algorithms that can apprehend and utilize these non-linear relationships in data to make more accurate predictions than can be made when we assume relationships are linear. But these algorithms all pick up on relationships in different ways, making some algorithms better suited for certain types of data. **There is no one-size-fits-all master algorithm that always works best regardless of what we are predicting.** So we must be smart with how we build our models to reflect this reality.

Rather than testing our data against one or two statistical approaches, Capture Higher Ed uses an approach that selects from dozens of algorithms in order to make the most accurate predictions using your data. This type of approach is made possible through advances in the field of machine learning, a subfield of computer science concerned with helping computers make increasingly optimal decisions using past information.
To put machine learning in the context of human learning, imagine when a child touches a hot stove. In an instant, she receives data ("ouch, that hurts"), considers what that data might mean ("stoves are hot and can hurt me") and will likely make a better decision concerning stoves in the future. Computers

At the core of machine learning is an emphasis on an empirical approach to prediction that focuses on deeply "listening" to the data to discern the signal from the noise, rather than a top-down, theoretical approach that creates a hypothesis to be tested from a set of "known" (or at least previously assumed) predictors.

Dr. Thom Golden

are capable of analyzing data in much the same way by looking at past data to predict future behaviors. You have surely seen examples of machine learning everywhere from your Netflix queue, to Google's self-driving car, to Watson's complete annihilation of Ken Jennings on *Jeopardy*.

At the core of machine learning is an emphasis on an empirical approach to prediction that focuses on deeply "listening" to the data to discern the signal from the noise, rather than a top-down, theoretical approach that creates a hypothesis to be tested from a set of "known" (or at least previously assumed) predictors.

The enrollment patterns of adolescents worldwide are complex and the most sophisticated tools should be utilized to accurately make predictions in this environment. With advances in computer science, statistics and computational processing, there is no reason why enrollment managers should limit their use of the most advanced technologies, which in the hands of a capable modeler, can produce highly accurate predictions.

Uber or Cab

A lack of competition in the market tends to favor longstanding, entrenched products and services. Just a few years ago, taxis were the only way to get rides if you didn't own a car. And while few people would champion the experience of riding in a cab, there was simply no other comparison until other, better services entered the market.

In the enrollment management space, there have been few options to challenge orthodoxy in the predictive modeling space. This has led to prediction products that use outdated technology and institutional leaders without the option to compare products or demand results based on what really matters: out-of-sample model accuracy.

At Capture Higher Ed, we believe that our technology and know-how are truly superior, and as partners, we are willing to help you ask the right questions to more completely understand how our predictive engine will help your institution meet its goals.

To start, here are 7 questions to ask your current predictive modeling provider:

Is your current predictive modeling provider:

1. Giving you reports on model performance and accuracy and explaining what they mean?
2. Exclusively utilizing logistic regression for predicting yes/no outcomes?
3. Using multiple types of data: institutional, contextual (publicly available demographic), and behavioral data (such as those provided by Capture Behavioral Engagement)?
4. Using automatic feature selection to optimize model training?
5. Dynamically scoring each prospect, every time new data becomes available?
6. Re-training your model within the enrollment cycle?

If every answer is yes, then there is one last question:

7. **"Is your current provider willing to compete?"**

In the world of data mining and predictive analytics, accuracy is the only currency that matters and competition is the best way to find the most accurate, highest performing model. Capture Higher Ed is willing to analyze your historical enrollment data and evaluate the accuracy of an Envision predictive model at no cost to you. Put us up against your current model and see how we fare. We are confident that you will see enrollment predictions in a whole new way.

> "Accuracy is the only currency that matters and competition is the best way to find the most accurate, highest performing model."

Contact

Thom Golden, Ph.D., Vice President of Data Science
Email: tgolden@capturehighered.com
Twitter: @Doctor_Thom

- 14-year career in higher education, expertise in enrollment management, strategic marketing, psychometrics and developmental psychology
- Former Senior Associate Director, Vanderbilt University

Brad Weiner, Ph.D., Director of Data Science
Email: bweiner@capturehighered.com
Twitter: @brad_weiner

- 13 years in higher education including admissions, research, and analytic advisory roles
- Former Analyst, Office of the Vice Provost for Undergraduate Education, University of Minnesota

Pete Barwis, Ph.D., Senior Data Scientist
Email: pbarwis@capturehighered.com

- Over 11 years experience as a statistician and research analyst for a variety of sectors, including higher ed
- Former Statistics and Technology Analyst for the University of Notre Dame

References

[I.] Widening Wealth Gap. Inside Higher Ed. https://www.insidehighered.com/news/2015/05/21/rich-universities-get-richer-are-poor-students-being-left-behind

[II.] Capture Manifesto. Capture Higher Ed. https://capturehighered.com/storage/app/media/Resources/Manifesto.pdf

[III.] Embracing 'Messy' Science. Inside Higher Ed. https://www.insidehighered.com/news/2016/03/15/american-statistical-association-seeks-usher-new-era-statistical-significance

[IV.] Closure Concerns and Financial Strategies: a Survey of College Business Officers. Inside Higher Ed. https://www. insidehighered.com/news/survey/closure-concerns-and-financialstrategies-survey-college-business-officers

[V.] Universities, hurt by falling funding, hire consultants for help. http://www.chicagobusiness.com/article/20150117/ISSUE01/301179978/universities-hurt-by-falling-funding-hire-consultants-for-help

[VI.] 2016 Marketing and Student Recruitment Practices Benchmark Report. Ruffalo Noel Levitz. https://www.ruffalonl.com/papers-research-higher-education-fundraising/2016/marketing-student-recruitment-practices-benchmark-report

[VI.] Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors. Journal of Statistics Education. http://www.amstat.org/publications/jse/v9n3/stanton.html

[VII.] Probing Question: Is the Farmers' Almanac accurate. Penn State. http://news.psu.edu/story/141165/2007/09/24/research/probing-question-farmers-almanac-accurate

[VIII.] Capture Behavioral Engagement. Capture Higher Ed. https://capturehighered.com/what-we-do/behavioral-engagement/

Capture Higher Ed
2303 River Road, Suite 201
Louisville, KY 40206

502.585.9033 | capturehighered.com