SESUG Paper 184-2023

Performing Higher Education Enrollment Management Predictions Using SAS

Phil Moore, Marla Mamrick, Sara Reinhardt, University of South Carolina

ABSTRACT

With the increased demand for data-driven decision-making in higher education, analytics researchers are expected to provide state-of-the-art data analyses to augment dashboards and descriptive statistics. This session will introduce building higher education AI predictions models in SAS and will discuss the role of predictive models within the enrollment management prediction and decision-making process. Issues encountered when predicting in higher education using both AI and traditional methods will also be addressed. Finally, an application will be demonstrated that was designed to assist in enrollment management decision-making. Enrollment Management Utilities (EMU) allows select users to access institutional data via a point-and-click environment to perform data manipulation and analyses used in higher education enrollment management. This homegrown SAS system includes applications such as At-Risk Student Locator (ARSL), matrix visualization, and student flow chart, to mention just a few. In addition, EMU can perform typical statistical analyses and create graphics often used in higher education.

Products used include Base SAS, SAS/STAT, SAS/AF, and SAS Viya. Since we will focus on application and not statistical theory and calculations, this session is appropriate for all levels and institutional types.

INTRODUCTION

Within the past 10 years, the role of higher education research has changed. Our reason for existence is no longer to merely count things, like students, faculty, and number of graduates, nor to submit institutional data to meet state and federal reporting requirements. We are now expected to also use state-of-the-art data analytics to explain the “why” behind the numbers to assist in institutional improvement. While many times well-designed dashboards will meet the needs of high-level educational administrators such as presidents, provosts, and vice presidents, when supporting the data needs of the registrar and the directors of admissions and financial aid, researchers are expected to perform complex data analyses to create predictive models. However, as will be described in this paper, the results from predictive models are simply one of many factors to be considered in any decision-making process, and the importance of predictive models can be easily overstated. Many other analytical approaches can and should be used to augment predictive modeling.

Like every other aspect of life in the 21st century, AI has permeated research in higher education with varying levels of success. For predicting retention, enrollment, graduation, and first-year college GPA, AI, and in particular, gradient boosting algorithms generally outperform logistic regression or regression by one to two percentage points in terms of accuracy (Albreiki, Zaki, & Alashwal, 2021; Bilquise, Abdallah,& Kobbaey, 2020; Delan, Davazdahemami, & Dezfouli, 2023) This is a far cry from most AI success stories. AI search engines can locate relevant web pages from an almost limitless source and return the results in a second or two, correcting our spelling in the process. AI can drive cars, recognize enemy targets, detect fraudulent banking transactions, find relevant DNA sequences to predict disease, and can even beat Ken Jennings in Jeopardy. But for one or two notable exceptions, there have been few AI success stories in higher education, and arguably no new insights, which leads to the first topic of this paper.

WHY AREN’T WE GETTING BETTER PREDICTIONS AND INSIGHTS INTO HIGHER EDUCATION WITH THESE NEW AI ANALYTICAL TECHNIQUES?

The lack of breakthrough findings in higher education through AI can perhaps be attributed to the nature of higher education data and the manner in which we use AI techniques, to our relative inexperience with the nuances of AI techniques, and to the very nature of higher education.

DATA AND THE USE OF DATA

Most agree that AI works best with large, complex data. While it is difficult to quantify either term, we will start with big data. Jonathan Zack (n.d) states that the tools and techniques used to analyze 100,000 records are far different from those used to analyze one billion records. Oracle (2023) defines big data as being so large that traditional analysis methods can't handle these massive data sets. In addition, Hadoop is a system that clusters multiple computers to analyze data sets that range in size from gigabytes to petabytes. Hadoop's parallel processing algorithm does not come into play for datasets less than 128 megabytes. Finally, Zack mentions big data in terms of wide data sets, defined as having an inordinate number of columns. Pharmaceutical data may contain more columns than rows, making traditional analyses daunting. Using these opinions and observations as a guide, the typical institutional research data set should probably not be considered big data.

As for complex data, Barry (n.d) considers data with many-to-many relationships as complex. Others describe complex data as having intricate variable relationships that are difficult to determine with traditional analytics (SAS, n.d). By these definitions, typical institutional data sets currently used for building models to predict enrollment, retention, graduation, and first-year GPA are probably not considered particularly complex.

THE MANNER IN WHICH INSTITUTIONAL RESEARCHERS USE AI

Consider the following observations and definitions:

> Deep learning is only going to be used when it really makes sense—where it can quickly find intricate, variable relationships hidden in large volumes of data that we have not been able to pull out in any other way yet (Ainsworth, as cited in SAS, n.d.).
>
> …artificial intelligence refers to the general ability of computers to emulate human thought and perform tasks in real-world environments, while machine learning refers to the technologies and algorithms that enable systems to identify patterns, make decisions, and improve themselves through experience and data (Columbia Engineering, 2023).

Does this describe how we use AI in enrollment management modeling? Or are we just building static models with the same data we always used but with a different analysis toy? In short, are we using AI to its fullest potential?

INEXPERIENCE WITH AI MODELING

Whenever a new technology is introduced to a discipline, there will generally be two types of individuals applying the technology: those who are experts in using the new technology but new to the discipline, and those who are experts in the discipline but new to the technology. Both types will have their struggles. Terenzini (2013) explains this phenomenon by positing that there are three types of intelligence required for institutional research: 1) technical/analytical intelligence, which includes research, statistical, and computer skills and is usually acquired in the classroom; 2) issues intelligence, which includes knowing the very essence of the discipline and an understanding of how the past and the current environment will impact the future and is acquired by working within the discipline; and, 3) contextual intelligence, which consists of the nuances of the issues within an institution and is acquired by working at that institution. We will refer to these three types of intelligence throughout this paper.

Technical Experts New to the Discipline

These professionals are probably data scientists who will attempt to apply their technical knowledge to the field of education. Without familiarity with enrollment management terms, data cycles, and nuances of variable values, model worth can be jeopardized due to this lack of issues intelligence. For example, one study attempted to predict whether a student would graduate on time. The model was quite accurate. Unfortunately, it included semester hours earned as a predictor variable. In essence, an AI model was created to inform us that students who do not have enough credit hours to graduate will not graduate on time (Pang, Judd, O'Brien, and Ben-Avie, 2017).

In addition, in a meeting recently, a very competent data analyst was “techsplaining” to a group of seasoned enrollment management experts on how to predict which applicants will be most successful. His model had a high accuracy rate. Unfortunately, his largest “pre-enrollment” predictor variable was first-year college GPA. While the enrollment management experts did not understand the nuances of the model building process, they certainly understood the severe shortcomings of the model. Due to a lack of understanding of enrollment cycles, the model built contained a predictor variable that would be unavailable when the model would need to be run.

Discipline Experts New to Technology

These experts have the issues and contextual intelligence to understand all variables available and what constructs need to be included in the model for accurate prediction. However, their lack of technical/analytical intelligence may lead to poorly constructed models. For example, practitioners new to AI may be mesmerized by the incredible training accuracy with little regard to or understanding of the need for test datasets. These individuals may also not necessarily comprehend the fact that Area Under the Curve and accuracy are not the only measures to determine the usefulness of a model. When we first started building AI models, we were somewhat pleased to discover that our tried and tested logistic regression and regression models outperformed the AI models. However, after much more practice at building AI models, we determined that AI was not the issue, but rather the problem stemmed from our inability to build quality AI models.

THE NATURE OF HIGHER EDUCATION

Building accurate prediction models in higher education is perhaps more difficult than in other disciplines due to the cyclical nature of the data and also due to the difficulty of predicting human behavior.

Cyclical Data

In higher education, data are usually grouped by cohorts. For instance, a freshman class is generally followed for six years. While we would like to think that prediction models that predict for one cohort will predict for other cohorts, this is not always the case. Admissions officers are constantly tweaking the criteria for accepting new students. High ranking university officials may also change the goals for the new cohort profile. Notable examples include cohorts affected by a pandemic, switching to test optional admissions, and a heavier reliance on the holistic review of applicants which will all result in changes to cohorts. When building prediction models, researchers must ensure that a model does not overfit the data for a cohort and must also make sure that it does not overfit to a particular cohort.

Predicting Human Behavior

Kuhn and Johnson (2016) observed that one of the major reasons that models fail is the inability to account for the complexity of human behavior. This is most particularly true in higher education where we are dealing with 18–24-year-olds who are entering the most transitional period of their lives. They will leave the familiar confines of their home, family, and friends to begin a new life on their own with hundreds to thousands of other people their age, probably in a new city or state. According to Astin’s IEO model (1993), inputs such as demographic characteristics and academic preparedness as well as interactions within the college environment will affect college outcomes. This appears to describe the complex data best addressed with AI. Unfortunately, typical institutional research data sets will not contain the complex variables required for accurate prediction. To address this complex interaction of input and environmental variables, data such as social media, financial, cellular data, and transactional information available via student identification cards will need to be blended with institutional data to better predict student behavior.

AI IN SAS

While it may seem as though we are not proponents of AI, this is not the case. Becoming proficient in the use and interpretation of AI techniques and output can only improve model building. In addition, taking the average of different model predictions can lead to better accuracy. Finally, studying the variables that are considered important in some models but not in other models can lead to a better understanding of variable interactions and can assist in better featurizing the data.

Fortunately for SAS users, SAS has developed a platform called VIYA that allows users to directly access R, Python, and SAS AI procedures to conduct AI modeling. Dan Vesset of the International Data Corporation (as cited in SAS, 2022) described SAS VIYA as, "… one of the most comprehensive analytics platforms on the market today." Concerningly, this new focus on VIYA puts into question the continued support for SAS 9.4. However, there is no set retirement date for 9.4, and SAS will continue to support this release at least through January 2028. Whether there will be a SAS 9.5 or 10.0 remains to be determined at this time. Most new development appears to be directed at VIYA.

For those wishing to perform AI in SAS 9.4 but cannot afford SAS VIYA, there are several options. First, academic departments within a university can purchase the academic version of VIYA at a greatly discounted price. Unfortunately, institutional research and other administrative offices will have to purchase the more costly administrative VIYA license. To gain experience with VIYA, researchers can use SAS VIYA for Learners at no cost. However, VIYA for Learners should probably not be used to perform the duties of the institutional research office due to ethical issues as well as the fact that only 5 gigabytes of information can be uploaded to the SAS cloud.

There is one option for the institutional researcher to perform AI in SAS 9.4 without VIYA. All R code and packages can be run using SAS PROC IML in SAS 9.4. If you are proficient in R, you may want to perform all data cleaning and featurizing as well as execute AI techniques using the R language. Since we are a SAS shop, we perform data cleaning and featurizing within SAS code, export the SAS data set into an R data space, perform the analysis via an R package, import the R output and associated data spaces into SAS data sets, and create our own graphics and output using SAS code.

Unfortunately, while SAS has several AI procedures such as GRADBOOST, HPSPLIT, and NNET as well as PROC PYTHON which allows you to run Python code and packages within SAS, VIYA is required to access these features. While disappointing, there is probably little difference in model accuracy among R, SAS, and Python.

Example

To illustrate some of the issues involved in building prediction models in higher education, we built models to predict admissions yield rates for Fall 2023. The question posed was not who will enroll but rather how many will enroll. This information is required for both financial planning as well as for capacity planning. For instance, high enrollment will make the president happy by increasing net revenue, but the provost may be unhappy if more freshmen classes are required or if the requests for on-campus housing exceeds residence hall capacity. The predictions were due by May 2023 so that plans could be implemented based on the size of the freshman class.

Three models were built using different techniques on the 2022 cohort data. A logistic regression model was built using the Hosmer-Lemeshow model building process (2013). A second logistic regression model was built using oversampling because the data set is unbalanced. Since the majority of applicants do not enroll, the logistic regression model overemphasized not enrolling. Oversampling is one way to address unbalanced data sets. The third model was built using R XGB.CV with nfolds = 10 to avoid overfitting. The SHAP.VALUES subroutine was used to generate the SHAP version of gain scores and to produce probabilities of enrolling. Once the models were created, they were used to score the 2021 and 2023 cohorts for cohort cross-validation. Table 1 lists the outcomes for in-state wave 2 (second group of admits) applicants.

	Logistic Regression			Logistic Regression Oversample			R – XGB.CV		
	Train	Test 2021	Test 2023	Train	Test 2021	Test 2023	Train	Test 2021	Test 2023
AUC	0.6423	0.5994	0.5793	0.6426	0.6014	0.5803	0.6778	0.6163	0.5857
Accuracy	0.6902	0.7007	0.6688	0.6004	0.6134	0.5765	0.6952	0.7037	0.6375
Precision	0.5622	0.4756	0.4910	0.5931	0.3719	0.3883	0.6321	0.5132	0.3884
Recall	0.1171	0.0656	0.0868	0.5478	0.4357	0.4912	0.0993	0.0656	0.1704
Specificity	0.9575	0.9640	0.9557	0.6495	0.6886	0.6185	0.9730	0.9737	0.8677
F1 Score	0.1939	0.1153	0.1475	0.5696	0.4012	0.4337	0.1717	0.1163	0.2369
Predicted	1349	1199	1206	2698	1807	1801	1349	1211	1411
Actual	1349	1189	1256	2698	1189	1256	1349	1189	1256

Table 1. In-state, Wave 2 Applicant Outcomes

As expected, the XGB.CV model outperformed both logistic regression models in terms of prediction accuracy for both the train and test datasets. Unfortunately, all model predictions were at least four percentage points better for the train data set, which technical/analytical intelligence would indicate that the models overfit the 2022 data and were not generalizable to 2021 and 2023. Issues intelligence can offer more insight by pointing out that 2021 was still a COVID year while 2022 was not. In addition, contextual intelligence would suggest that enrollment trends differed between 2021 and 2022 because, at this institution, few classes were taught in person in 2021 and in 2022 all were taught in person. These conditions would probably have an impact on enrollment. While AI has been shown to predict many things, no technique can predict in the future what has not happened in the past. If changes to the external environment alter covariate pattern behaviors, prediction accuracy will drop.

As can be seen in Figure 1, no model significantly outperformed the other models as measured by AUC.

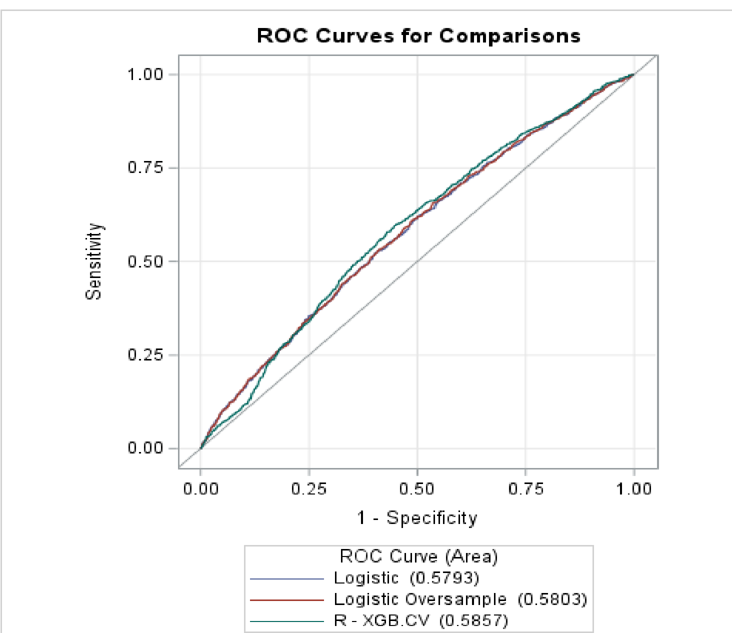

Figure 1. ROC Curves for Comparisons

However, as alluded to earlier, AUC is not always the best measure of model worth. This premise is substantiated by comparing predicted enrollment to actual enrollment. In further comparing the models, the logistic regression predictions for 2021 and 2023 in-state wave 2 enrollment were within 10 and 50 students respectively of the actual enrollments. Regarding the logistic regression oversample model, the predicted enrollments were within 618 and 545 students of actual enrollments for 2021 and 2023. While it appears that oversampling would improve the prediction of who will enroll by improving recall, oversampling was inadequate for predicting overall student enrollment. Finally, the R XGB.CV model predictions were within 22 and 155 students of the actual enrollments for 2021 and 2023. The inaccuracy of this model can be attributed to changes in behaviors of covariate patterns. Consider the gain scores in Figure 2 and Figure 3.

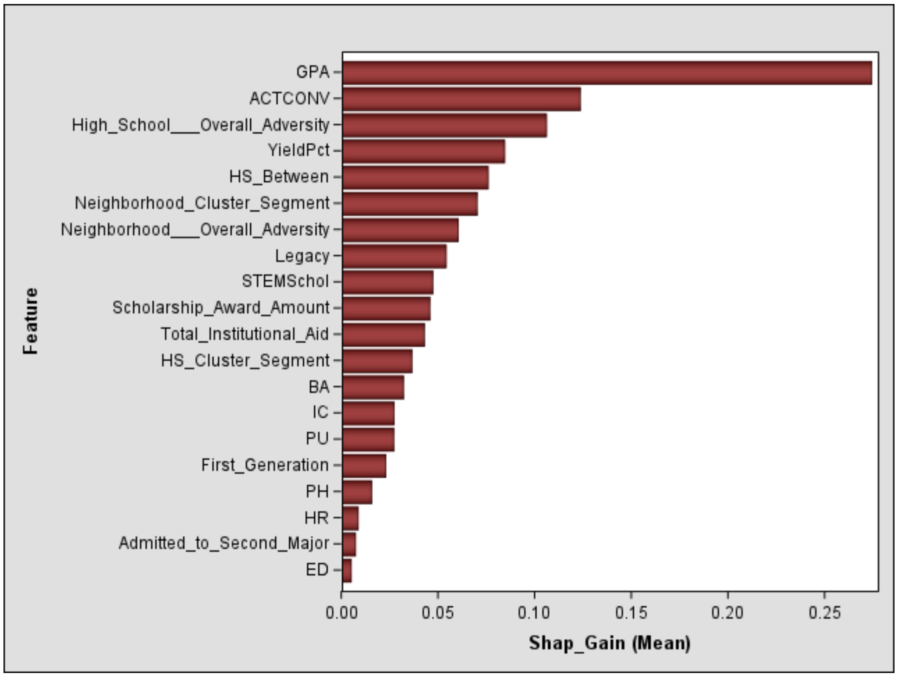

Figure 2. R XGB.CV Gain Scores

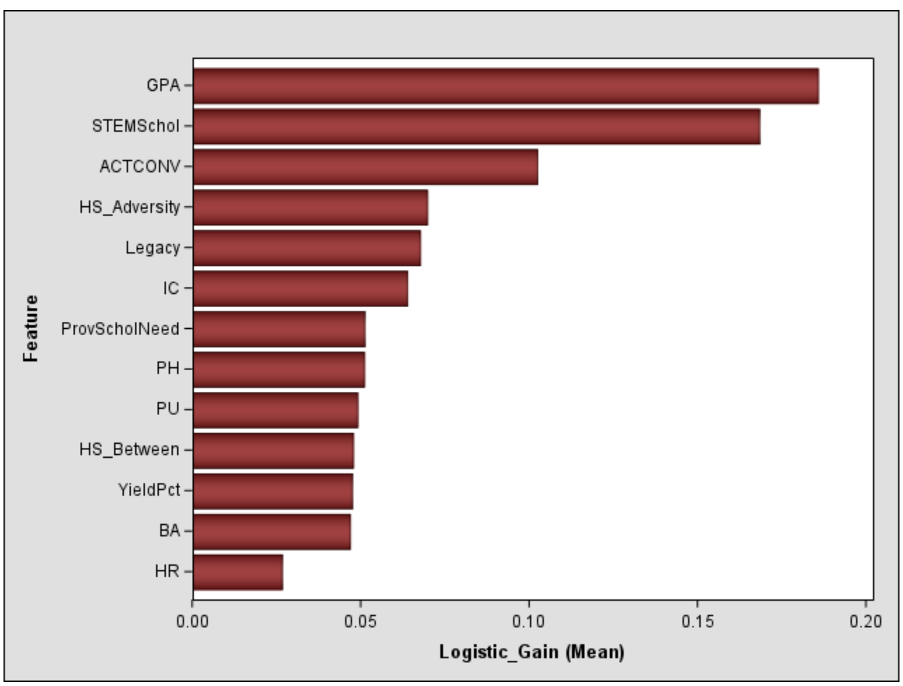

Figure 3. Logistic Regression Gain Scores

The R XGB.CV model emphasizes high school GPA more than the logistic regression model does. In 2022, the average high school GPA was 4.37 and the mode was 4.35 while in 2023 the average high school GPA was 4.35 and the mode was 4.30. Unfortunately, the delta-p values indicate that for every one-point increase in high school GPA, the applicant's estimated probability of enrolling will drop 13.5% in 2022 while in 2023 the same grade point increase decreased an applicant's estimated probability of enrolling by only 10.9%. Since the model was formed on 2022 data, 2023 estimates will be erroneously high. This is no indication of a shortcoming of XGB.CV since the changes in covariate patterns could have adversely affected a logistic regression model in other instances.

To determine the relative worth of the models, refer to Figure 4.

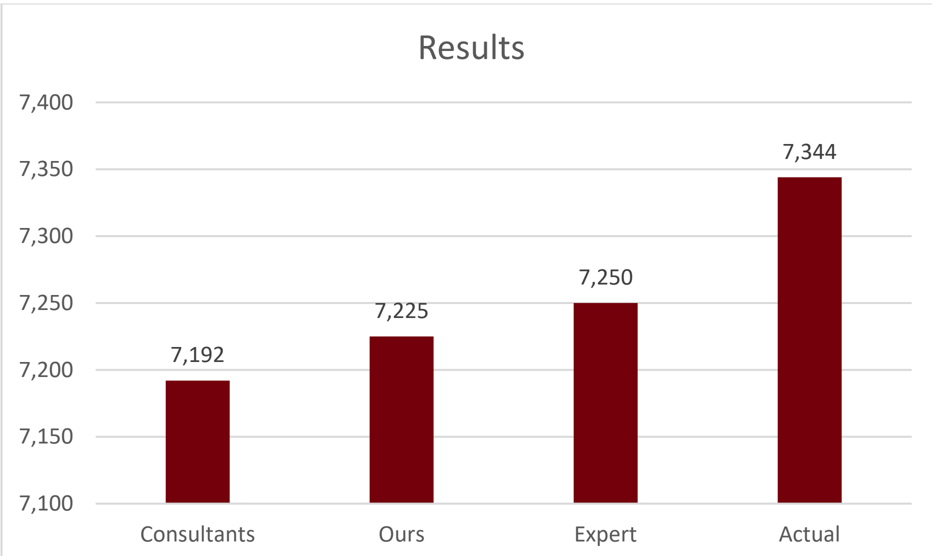

Figure 4. Model Worth Comparison

The university paid a consulting firm a lot of money to predict the number of new freshmen who will enroll for Fall 2023. Our prediction was based on the sum of the logistic regression model predictions for first wave and second wave in-state and out-of-state probabilities of enrollment as well as an estimate of the number of post-May applicants who will enroll. The expert estimate was submitted by the Executive Director of Undergraduate Admissions and Assistant Vice President for Enrollment Management. Her estimate was based on years of experience, careful monitoring of enrollment data, the institution's typical market share, and knowledge of changes in admissions procedures at a nearby institution which could positively affect enrollment at our institution. As can be seen, the expert had the more accurate prediction. For all intents and purposes, however, the differences in predictions were minimal and would not affect any university planning. The successful prediction by the expert can be explained by Svenja Szillat (2022) who believes that when rule-based systems and a spreadsheet lead to accurate prediction then machine learning is unnecessary.

WHERE DO WE GO FROM HERE?

We have a saying in our office, "If you find a surprising result in higher education, check your answer. You probably did something wrong." Admissions practitioners have been successfully admitting students to colleges for well over a century by using issues and contextual intelligence. By adding technical/analytical intelligence, we can verify the expert's predictions, provide the expert with more information on which to base decisions, automate some tasks of the expert, and we can investigate institutional policies and practices that experts perceive as being detrimental to student outcomes. We have developed a SAS point-and-click application called Enrollment Management Utilities (EMU) to further assist in meeting the technical/analytical aspect of enrollment management.

EMU

When clicking on the desktop icon for the application, SAS will run in the background and Display 1 will be shown.

Display 1. EMU Main Menu

Verifying Experts Predictions

As stated earlier, prediction in higher education is difficult due to differences in cohorts and trying to predict the behavior of 18–24-year-olds. Even the most grizzled veteran of enrollment management will at times arrive at inaccurate predictions.

Provide the Expert with More Information on which to Base Decisions

Great models are seldom built by providing an AI technique with existing higher education data sets to find relevant patterns in the data. Additional data needs to be added and variables redefined to enrich the data set.

According to Tom Keldenich (2022), “Featurization is the set of techniques used to obtain new information from pre-existing data in a dataset.” For instance, a typical longitudinal cohort data set will have housing information for each year of enrollment. Adding additional information such as on-campus housing for the first year and the number of years enrolled in which the student was living in on-campus housing may lead to better prediction of graduation. One tool to assist in adding features to data sets is an application called At-Risk Student Locator (ARSL).

ARSL can be used to determine the retention and graduation rates of subpopulations in a cohort. For instance, it can quantify the relative success of in-state, first generation engineering students who place into the math curriculum at a level lower than calculus. To demonstrate ARSL, we will investigate outcomes for students who are academically eligible for admission to the university but who underperformed in high school. These students are referred to as Non-AWE, indicating that they do not have a particularly strong academic work ethic. The results from ARSL are presented in Figure 5, Figure 6, and Table 2. The plots represent the student flow for this cohort over six years.

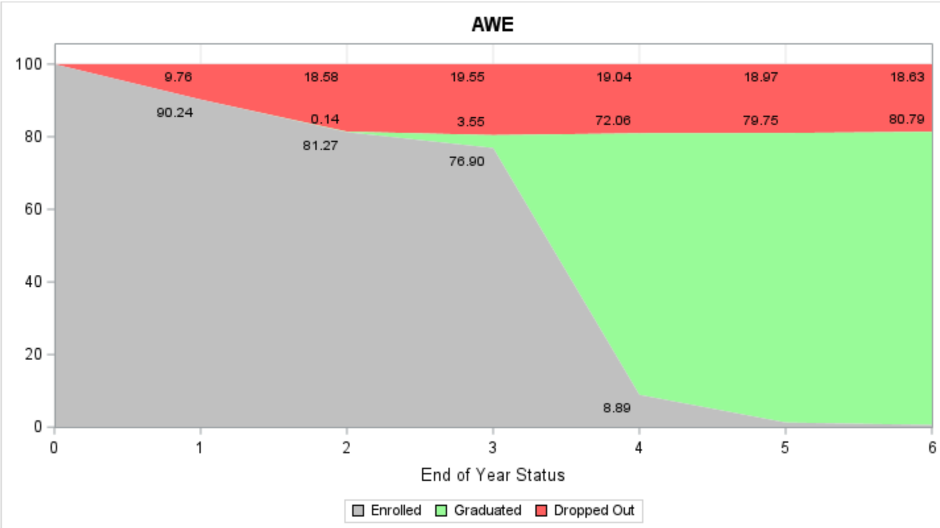

Figure 5. AWE Student Outcomes

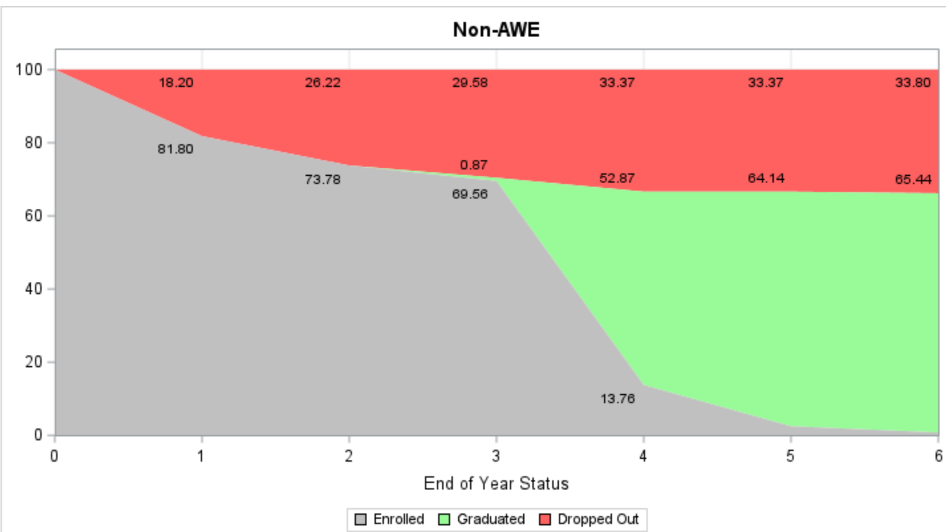

Figure 6. Non-Awe Student Outcomes

Metric	Selected Population	AWE	Non-AWE	Revised Selected Population	Impact on Cohort
Number of Students	5,061	4,138	923	5,061	5,061
Retention	0.89	0.90	0.82	0.90	0.02
Six-Year Graduation Rate	0.78	0.81	0.65	0.81	0.03
Net Tuition	$288.0M	$237.5M	$50.5M	$290.5M	$2.5M

Table 2. Comparison of Outcomes Between AWE and Non-AWE Students

As can be seen, AWE students are retained at a higher rate than the Non-AWE students and are much more likely to graduate. Referring to Table 2, AWE students have a retention rate 8% higher than Non-AWE students and graduate at a 16% higher rate. If the Non-AWE students performed at the same level as the AWE students, the university retention rate would increase by two percentage points, the graduation rate would increase by three percentage points, and the university net tuition revenue would increase by $2.5M. This information may be used to improve model accuracy but can also be used by experts to assist in the areas of admissions, student success initiatives, and advising.

Automate Tasks of the Expert

EMU also includes applications to prepare data for Clearinghouse batch file submission, to create high school report cards, to generate custom random samples, and to create custom enrollment management matrices. In addition, a suite of programs is available to assist in ranking students for scholarships and admission to the honors program and includes measures of inter-rater reliability and the need for third reviews.

Investigate Institutional Policies and Practices that Experts Perceive as Being Detrimental to Student Outcomes

Institutional policies are put into place with the intention of helping students; however, this is not always the case. Likewise, well-intentioned practices may actually be detrimental to the student. Technical/Analytical intelligence can be used to investigate conflicting suppositions by experts concerning policies and practices. For example, at most institutions, students are placed into the math curriculum based on high school math courses taken or by scores on a math placement test. Students may petition to be placed into a higher-level math course to reduce the number of math courses that they are required to take. While some may feel that this practice is beneficial to the student by allowing the student to have some influence over their educational experience, others question the student's judgement as it pertains to curriculum matters. In short, is the practice of overriding the math placement policy detrimental or beneficial to the student? Insights into this question can be provided by revealing what happens to students who fail their first math course. Refer to Figure 7, Figure 8, and Table 3.

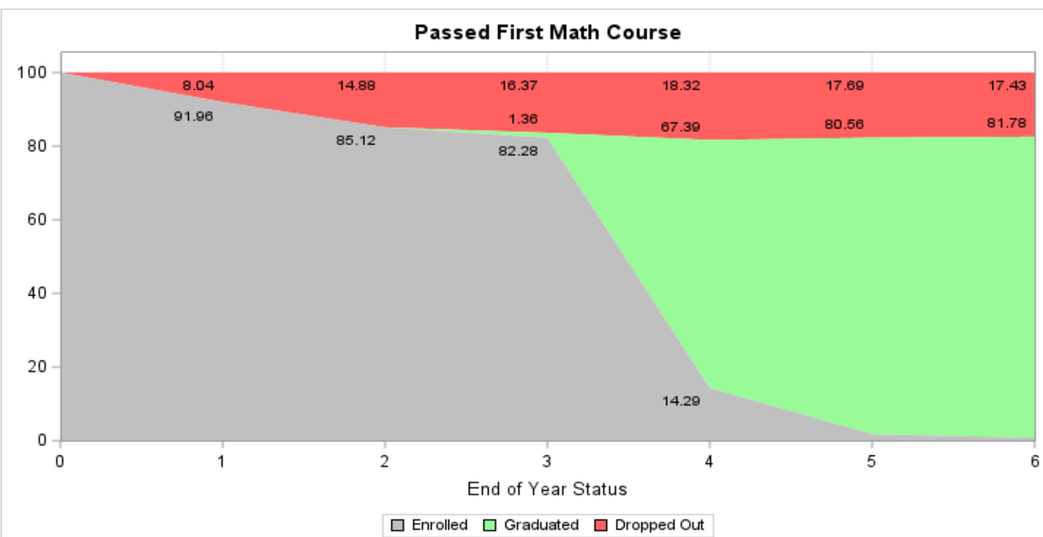

Figure 7. Outcomes for Students Passing First Math Course

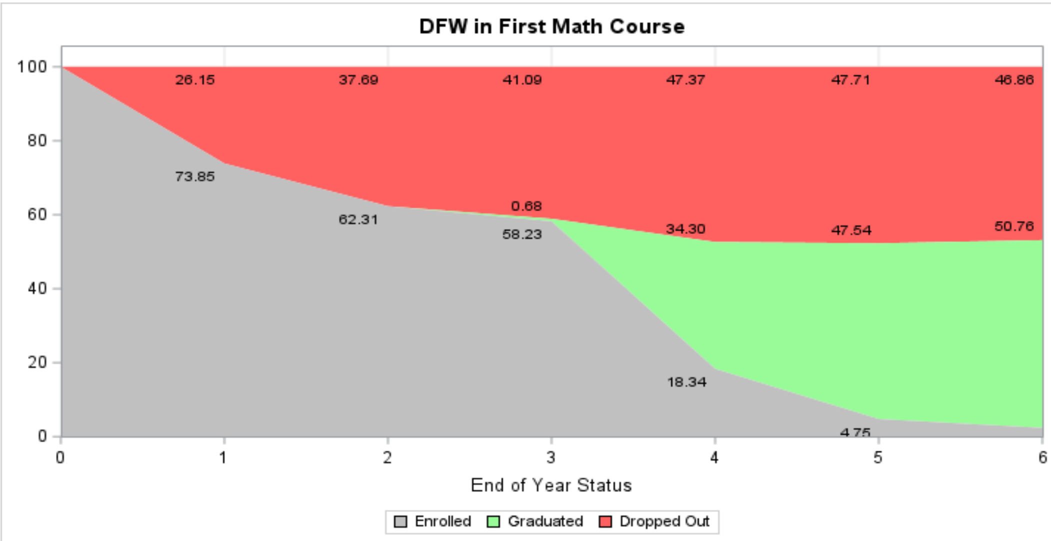

Figure 8. Outcomes for Students Earning a DFW in First Math Course

Metric	Selected Population	AWE	Non-AWE	Revised Selected Population	Impact on Cohort
Number of Students	3,613	3,024	589	3,613	3,613
Retention	0.89	0.92	0.74	0.92	0.03
Six-Year Graduation Rate	0.77	0.82	0.51	0.82	0.05
Net Tuition	$211.7M	$181.7M	$30.0M	$217.1M	$5.4M

Table 3. Outcome Comparisons Based on DFWs in First Math Course

As can be seen, students who fail their first math course have an 18% lower retention rate and a 31% lower graduation rate than students who pass their first math course. While the lack of success in a student's academic career cannot be solely attributed to failing their first math course, the results indicate that overriding the math placement policy should not be taken lightly.

CONCLUSION

In the early 1950's, tastes in America changed. The president of the Master Brewers Association of America implored brewers to quit making beers they are proud of and start making beers people will buy (Ogle, 2006). Perhaps we should challenge ourselves to quit building models that we are proud of, in terms of accuracy and esoteric methods used, and start building models practitioners can and will use to improve the educational experience of students at our institutions. If models are too complex or too difficult to explain to constituents, the experts may choose a simpler alternative which may be less accurate but more beneficial for the intended purpose. Accuracy for the sake of accuracy should not be the goal. According to Kuhn and Johnson (2016), the definition of predictive modeling should be changed from "…the process by which a model is created or chosen to try to best predict the probability of an outcome" to "…the process of developing a mathematical tool or model that generates an accurate prediction" (p. 2).

The results of predictive modeling are just one of many factors to be considered by the expert and should not be considered an absolute truth. Ayers (2007) views predictive modeling as a complement and not a substitute for intuition. Institutional researchers can assist the expert in decision-making by building accurate usable models and by adding additional pertinent information to augment the expert’s prior experiential knowledge. This may include results from complex AI models, traditional models, and relevant research studies. In the future, AI will probably play a bigger role in higher education as we gain more experience building these complex models and learn to better collect and use relevant student data, but it is doubtful that AI will replace the higher education expert anytime soon.

REFERENCES

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of students' performance prediction using machine learning techniques. *Education Sciences, 11*(9), 552. https://doi.org/10.3390/educsci11090552

Ayers, I. (2007). *Super crunchers: Why thinking-by-numbers is the way to be smart*. Bantam Books.

Barry, D. (n.d). *Many-to-many relationships*. Service Architecture. https://www.service-architecture.com/articles/object-oriented-databases/many-to-many-relationships.html

Bilquise, G., Abdallah, S. Kobbaey, T. (2021). Predicting student retention among a homogeneous population using data mining. In Hassanien, A.E., Darwish, A. (Eds.), *Machine learning and big data analytics paradigms: Analysis, applications and challenges. Studies in big data*, vol 77. Springer, Cham. https://doi.org/10.1007/978-3-030-59338-4_13

Columbia Engineering. (2023).*Artificial intelligence (AI) vs. machine learning.* https://ai.engineering.columbia.edu/ai-vs-machine-learning/

Delan, D., Davazdahemami, B., & Dezfouli, E. (2023). Predicting and mitigating freshmen student attrition: A local-explainable machine learning framework. *Information Systems Frontier*. https://doi.org/10.1007/s10796-023-10397-3.

Hosmer Jr., D.W., Lemeshow, S. and Sturdivant, R.X. (2013). *Applied Logistic Regression*. 3rd Edition, John Wiley & Sons.

Keldenich, T. (2022). *Featurization – How to use it to boost your model*. https://inside-machinelearning.com/en/featurization-how-to-use-it/

Kuhn, M., & Johnson, K. (2016). Applied predictive modeling. Springer.

Magolda, M. & Astin, A. (1993). What matters in college: Four critical years revisited. *Educational Researcher*. *22*. 10.2307/1176821.

Ogle, M. (2006). *Ambitious Brew: The Story of American Beer.* Harcourt Inc.

Oracle (2023). *What is big data?* https://www.oracle.com/big-data/what-is-big-data/

Pang, Y., Judd, N., O'Brien, J. and M. Ben-Avie, M. *Predicting students' graduation outcomes through support vector machines*, in 2017 IEEE Frontiers in Education Conference (FIE), Indianapolis, IN, USA, 2017 pp. 1-8. doi: 10.1109/FIE.2017.8190666

SAS. (n.d). *Machine learning and artificial intelligence in a brave new world*. https://www.sas.com/content/dam/SAS/documents/marketing-whitepapers-ebooks/third-party-whitepapers/en/machine-learning-ai-brave-new-world-109489.pdf

SAS. (2022). *SAS Viya now available in Microsoft Azure marketplace*. https://www.sas.com/en_us/news/press-releases/2022/september/sas-viya-now-available-in-the-microsoft-azure-marketplace--.html.

Szillat, S. (2022). *Four situations where you should not use machine learning*. Qymatix. https://qymatix.de/en/situations-without-machine-learning/#:~:text=No%20Data%20%E2%80%9CCold%20Start%20Problem%E2%80%9D&text=You%20want%20to%20predict%20your,cannot%20generate%20recommendations%20from%20it.

Terenzini, P. T. (2013). On the nature of institutional research revisited: Plus ca change…? *Research in Higher Education*, *54*(2), 137-148.

Zack, J. (n.d) *7 signs you are dealing with complex data*. Sisense. https://www.sisense.com/blog/7-signs-youre-dealing-with-complex-data.