Impacting Student Persistence and Retention: A Journey of Self-Discovery

Kimberlyn Brooks, M.Ed.
Associate Director
Undergraduate Education
kbrooks@bgsu.edu

Cynthia Roberts, M.A.
Assistant Director
Undergraduate Advising and Academic Services
crobert@bgsu.edu

Andrew Alt. M.A.
Assistant Vice Provost
Academic Affairs
awalt@bgsu.edu

Abstract: Recently, BGSU has experienced a culture shift toward greater use of data. This shift enhances data collection, resulting in further interest in analyzing reliable data among various constituents. A primary institutional focus has been on improving persistence and retention rates of first-time, full-time (FTFT) freshman cohorts. To further this goal, a graduate analytics student was hired to build a rudimentary logistic model predicting retention with various demographic groups. A completed model was shared with the campus community and interest quickly grew, resulting in an increase in data contributions and emphasizing the benefits of collecting reliable data from departments. Additional research led to pursuing data points that are actionable by campus partners. The next model iteration expanded the number of graduate students, including statistics and computer science majors. These students suggested other statistical models, resulting in the use of multiple models, each with its own benefits. A new partnership between Academic Affairs and Student affairs allowed data and strategies to be shared between the units in a multiphase approach. In Phase 1, the analyzed data was used to provide intentional outreach to students and to identify a sampling of students who could potentially qualify for additional scholarship dollars. The next phase includes identifying areas of potential interventions that connect students to resources that already exist while also identifying gaps in services or resources and finding ways to fill those gaps. Both phases rely upon research, data analysis, and careful study of current University resources to deliver actionable insights. This multiphase approach has already identified promising initiatives in the immediate future and continues to promote the development of new strategies.

Introduction

In 2011-12 a cross-divisional office dedicated to student academic success and retention was created and charged with monitoring student academic success and persistence. Thus began a journey of self-discovery at BGSU; one in which the gains we made in terms of data and predictive analytics were matched by learning, and in some cases, stumbling upon, heretofore undiscovered and untapped technology, people, initiatives, and offices, as well as new and exciting collaborations to help us build bridges between technology and people in support of student success.

As part of this office, in collaboration with Registration & Records, in 2011 an Excel spreadsheet was created, lovingly entitled the "OGRE" (due to its size). At its inception, the OGRE contained 45+ columns of data points determined at that time to be critical to our knowledge of our incoming cohort of first-year students. Our original charge was, simply, to gather as much salient data as we could about our incoming first-year cohort, and to store it in a central repository where we could utilize it to guide our efforts in determining which students might be at risk for academic success and persistence. Further, we were to design targeted outreach and support early enough in the academic year to make a difference in outcomes.

By 2013, the OGRE contained close to 80 columns of data points, gathered from Admissions, Registration & Records, Institutional Research, academic departments, academic support offices, and other units across campus. In 2013, academic advising underwent a reorganization, and the efforts of the OGRE's original home office became absorbed under the newly-formed Office of Undergraduate Advising and Academic Services. As our original office grew to include the new model of academic advising, it became clear that the OGRE would need to evolve as well, as it was of limited use to academic advisors in its nascent form. While managing the sheer volume of information concerning our first-year students has always been an ongoing obstacle to providing truly efficient and timely outreach, follow-up, and support of our students, upkeep and maintenance of our homegrown OGRE was cumbersome, required frequent updates, and access was limited to academic advisors in the college offices. Information sharing between college advisors and program advisors was often random, ad hoc, and incomplete.

Nevertheless, and as a result of a strong collective effort across campus, retention of first-year students increased from just under 70% in 2011 to nearly 76% in 2015. It is here, in what began as an effort to learn about students and how we can best serve them that we also learned about ourselves and our institution.

We discovered from our early attempts at outreach and intervention that to have a direct and positive impact on student success, the entire advising community (college advisors, program advisors, and others working with at-risk populations) needed access to student data, to allow them the ability to share advising notes, retrieve reports, monitor progress of populations based on identified criteria, and assess student success. In our quest to become an institution where data-driven decisions guided our efforts, we had to begin with sometimes uncomfortable self-discovery, certainly in terms of where our gaps in understanding, collaborations, data, and initiatives lie. Certainly, our advisors, now key agents of student persistence, and student support personnel therefore had to become knowledgeable about who the key collaborators and initiatives were on campus. Our operating philosophy was, and continues to be, that the earlier at-risk students are identified and engaged, the greater the likelihood of issue remediation, academic success, and persistence.

Early Outreach

From 2011-early 2015, the OGRE provided a range of descriptive input and early collegiate performance data on our first year population, leading to the question of how we use the data for early identification of at-risk students and advisor outreach and support of BGSU's first-time-full-time students. This question, and the desire to be more strategic and improve response time in these efforts, resulted in the launching of the Microsoft SharePoint informal pilot initiative in spring of 2015 as a vehicle to provide advisors with a real-time way for way to prioritize, contact and record outreach to students who were not yet registered for the subsequent term. The process worked in this way:

- At the start of the registration cycle, the OGRE EXCEL spreadsheet was uploaded to the advisor SharePoint site
- Advisors filtered to the fields they wanted to see
- As they reached out, advisors inputted responses into the SharePoint OGRE in real time, allowing for all advisors to see results quickly and easily, and add additional outreach and responses as necessary
- Categories of responses included: RETURNING, NOT RETURNING (and why), UNSURE ABOUT RETURNING, AND UNABLE TO REACH
- Academic and other barriers to registration were quickly identified
- Senior advisors and others could quickly sort and prioritize responses for further outreach in an iterative loop, and other offices including the Bursar and Student Financial Aid and Scholarships could be brought on board quickly and easily to troubleshoot and remediate non-academic barriers to registration
- EXCEL pivot tables were used to quickly gauge students in each response category to create categories of students still viable for return

In 2016, as the OGRE grew with more data categories, advisors became wary of filtering through the voluminous fields and categories, and asked if a less cumbersome mechanism and process was possible. In response, plans were made for the move to a singular retention note in the online advising platform, SSC Campus.

As we moved forward with more organized, efficient, and timely student success efforts, a greater sense of urgency in the early identification and support of at-risk students was created, and our thoughts began to coalesce around just how early could we go? What essentially began as an off-hand conversation quickly grew and mushroomed into predictive modeling.

Evolution of Predictive Modeling at BGSU

As the amount of data being collected on the cohorts grew, so did curiosity about how BGSU could use the data to improve retention rates. There had been much descriptive and diagnostic analyses done about prior outcomes but there was a yearning for more. Could BGSU use the data to try to predict outcomes of incoming cohorts and could something be done to impact any predicted results? To get at the answer to these questions required something more than Excel spreadsheets and pivot tables. It was time to explore machine learning methods for analyzing data.

In the spring of 2016, discussions around predictive modeling began. Logistic regression was selected as the first method used to produce a predictive model. It was chosen simply because the graduate student (a Master of Science in Analytics student) working on the project was familiar with logistic regression and how to use Python for programming.

The initial variables selected to create the model were thought by many in the Vice-Provost's office to have an impact on retention. The variables also were easily accessible for both the current cohort and the previous cohort. The model was built using 80% of the previous cohort data, leaving 20% to test the model. Once the final model was delivered, the current cohort data was entered into the model equation and a probability score was assigned to each student indicating the probability of the student retaining. The first data set consisted of 13 different variables, categorical and numerical, obtained from

two different sources. The final list of variables used were: gender, high school GPA, ACT score, age, ethnicity, in or out of state status, academic standing assigned after the completion of the first fall semester, fall semester GPA, the original college the student selected a major from, whether a housing deposit was made for the next academic year, commuter status, Math emporium enrollment, and whether the student persisted from fall to spring.

Once the model was built, it was presented to a few key stakeholders who pushed for action to be taken on a subset of the cohort to try to impact the overall probability of retention. At this late juncture in the spring semester, it was very difficult to develop strategies to impact students, as all of the predictors influencing the outcomes were already in the past. The advising staff did proactively reach out to students, but the outreach was similar to efforts that had taken place in previous years. It was time to start planning for the next model, but to do so meant hiring new and additional graduate students.

In the fall of 2016, the next round of graduate students were hired to help with the modeling work. They were from the Ph.D. program in statistics (2) and the Master of Computer Science program (1). As with many new graduate students, the learning curve can be very steep, but this set of graduate students was well appointed and eager to handle the task. The expertise in their respective fields positioned them well to jump in and begin the process of building the next logistic model.

Near the beginning of October of 2016, a persistence model was built for the 2016 FTFT cohort using logistic regression. Data points similar to those used for the last retention model were used, except for items that had not yet occurred, and the addition of first-generation status, unmet need, and the residence hall in which a student was housed. The only significant predictors for the model turned out to be first-generation status, commuter status, ACT score, and high school GPA. The outcomes from this model and continued research motivated us to look for other available sources of data. Also, presenting the model to other campus constituents increased their curiosity in if and how their data might be incorporated into the modeling. It also became clear that new data sources, especially data points related to student activity, and other methods for modeling needed to be explored.

The logistic model does provide reliable predictions, but the results are difficult to deliver in an easily digestible form to campus partners. While it is easy to target students in a particular probability range, it is difficult to know which strategies would be most beneficial to which particular group of students. An additional challenge of the logistic model is missing data. For every variable that has a missing piece of data, decisions have to be made on either finding a way to assign a value (by imputing or some other data manipulation) or exclude the student from the model.

Spring of 2017 brought an expansion to the model types and the number of data points gathered for use in the modeling. While logistic regression was still used, the decision tree model was incorporated. Decision tree allowed us a way to get results for all students and to present the model to campus partners in a format that was easier to understand. Decision tree also deals with co-linearity problems in the data. New factors included in the retention model were: BGSU fall term GPA, visits to the Learning Commons (tutoring), admit date, orientation survey questions, and linked course enrollment. Data collection also began on ability level (tied to BGSU scholarship funds received), Canvas (learning management system) page views, distance between home and BGSU, Rec Center visits, DTM (Diver, Thriver, or Maintainer-comparison of high school GPA to BGSU fall term GPA), and student rank, i.e. freshman, soph, etc.(used as a proxy for number of credits a FTFT students comes in with) to be used in future modeling work.

While interest was growing in the modeling work, so was the need to produce the insights earlier to allow time for interventions to impact students. This led to the production of the persistence model for the 2017 cohort to be produced in the summer of 2017. While the new cohort of students would not be run through the model until the data was available after the 15th day of classes, the model could be

produced and insights could be delivered so that departments could start thinking about ways to impact students. Models were built using logistic regression, decision tree, and newly introduced random forest.

The decision tree model was a hit with the campus community, but it delivered only a small number of actionable insights. The addition of the random forest analysis in the 2017 persistence model gave us a ranked list of top predictors with an importance level that could compare the scale of the factor against other factors. Having a list of the top 20 predictors gave many more campus constituents the ability to contribute to the conversations regarding how their work could impact student persistence and retention. The list also reinforced the idea that student retention is a campus-wide responsibility.

In the spring of 2018, the retention model and the predictions relating to the 2017 cohort were delivered in the third week of classes. Delivering the predictions in the third week has become the expectation, which positions us well for the next set of advances – trying to move from predictive modeling to prescriptive modeling.

Hierarchical modeling was also introduced in the spring of 2018, but the computers assigned to the graduate students are not equipped to process the data as quickly and efficiently as necessary for us to do enough trials and testing to implement the model. The BGSU ITS department became aware of the modeling work and began to work on a relational database for us to achieve better data structure, security, and perhaps better ways to collect data from our campus partners.

The first University Student Retention Summit was held in the summer of 2018. One of the goals of the summit was to present the persistence and retention models for the incoming 2018 cohort. With over 80 campus partners in attendance, including constituents from Academic and Student Affairs and key institutional leaders, many conversations and strategies were discussed. An organized follow-up plan was coordinated to assist all departments in devising retention plans which include strategies with clear goals and measurable outcomes.

The Data and Partnerships

Much of the data used in the predictive modeling, is gathered from multiple sources around campus. There is no single system used to gather and store the data. This certainly makes it difficult to manage the data requests that have to go out to capture the data, as well as manage all of the Excel spreadsheets that come in. Currently, data used to produce the models are received from:

- Office of Admissions – admission data, ability level (BGSU non-need based scholarship data)
- Visual Zen software – SOAR (orientation) data
- Office of Institutional Research – ethnicity, age, number of credits student starts with, housing assignment, commuter status, original college assignment, member of a residential learning community, high school GPA, ACT score, BGSU term GPA, math course enrollment
- Office of Financial Aid – unmet need, first gen status (currently considering getting this information from SSC Guide App)

- SSC Campus software – number of early alerts, visits to the Learning Commons (tutoring), graduation plan on file, note reasons associated with persistence/ retention
- Student Recreation Center – number of visits to the Student Rec Center
- Dean of Student's Office – results from orientation survey
- Office of Campus Activities - membership in a student organization, membership in fraternity/sorority
- Information Technology Services (ITS) – Canvas page views in the first two weeks of fall semester
- Office of Residence Life – PED data for Labor Day weekend (use of door fob for res hall entry), housing deposit data, number of interactions between RA and individual cohort member
- Office of Undergraduate Education – linked course enrollment
- Office of Registration and Records – college/advisor changes, withdrawal or cancellation information
- Query in PeopleSoft – daily enrollment information
- Various other offices – rosters of students involved in groups, programs, or activities which individual offices wish to monitor the enrollment status of

The graduates students from the math and computer science department have allowed us access to a level of expertise that were unavailable to us when the project first began. Faculty have been instrumental to the process by providing timely advice to overcome issues and provide guidance about next steps.

When the data collection and modeling began, not all data in the above list was readily available. Some data were not even collected, and some was collected but not used. Campus partners worked with us to develop systems and/or specific time-tables by which the data would be organized and collected. Not only did this help with the modeling, it also helped provide data for measurable outcomes.

Data is now being received from various sources around campus, and also delivered to many campus partners. Both academic and program advising groups or other departments assigned to work with various student groups wish to monitor the status of groups with whom they work. As a byproduct of the modeling work, structures have been put in place using an Access database and Excel spreadsheet to regularly deliver to advisors updated lists of students who are not registered with

selected pertinent information known about the student in order to assist with helping the student overcome obstacles to persistence or retention.

In the fall of 2018, each Resident Advisor (RA) was provided with a list of 5 students to provide additional attention. Each college was also given a list of 20 students to pay particular attention. At the time, no specific strategies were outlined beforehand or coordinated. As we moved through the enrollment period for both the spring and following fall semesters, it was clear that the lists of students were persisting and retaining at lower rates than the overall cohort. As campus partners became aware of the difference, confidence increased in the predictive models as well as the ability to select students that may be in need of intentional strategies. The future intent is to employ this strategy again in the fall of 2019, with the hope that each college and RA will have strategies in place to work with students at risk for persistence and retention.

Systems/Infrastructure

The modeling, monitoring, and data delivery is currently being done on a shared drive using R programming, an Access database, and multiple Excel spreadsheets. While the current system has served us well so far, it is becoming increasingly difficult to manage the volumes of data. There is also concern for the security of the data. Throughout the 2018-2019 academic year, we will be working with ITS to develop a better storage and delivery system for our current data.

Over the past two years, the amount of data and information shared between Academic Affairs (AA) and Student Affairs (SA) has substantially increased. An Academic Affairs Retention Committee, a Student Affairs Retention Committee, and a Student Success Council have been formed**.** All three groups are comprised of representation from both AA and SA, and collaboration is expected from all groups. Currently, the structure and scope of the Student Success Council is to provide oversight in retention efforts.

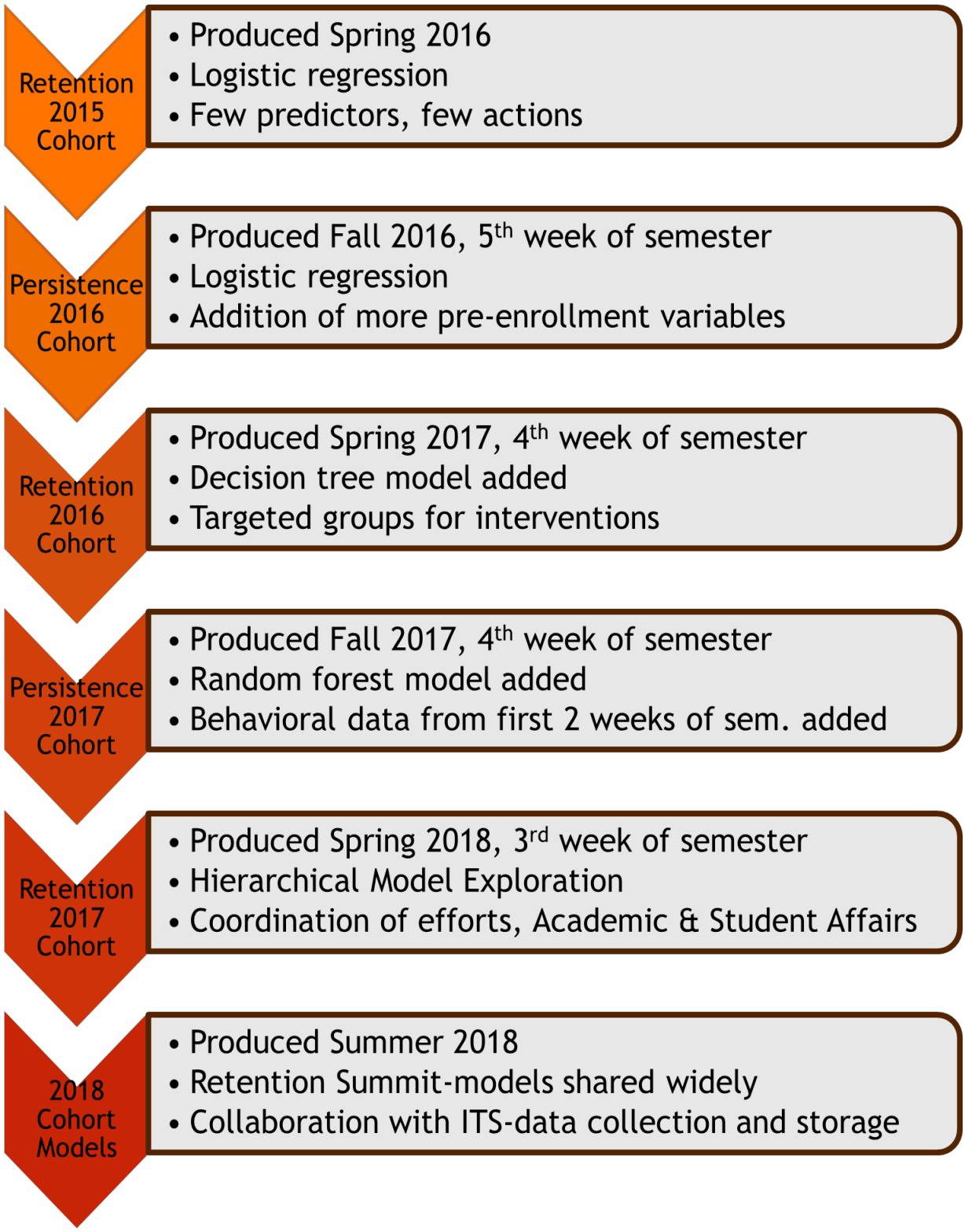
Conclusion

Like many other institutions, BGSU has been amassing data for quite some time. There has always been a need for descriptive data: data which tells a story. Over the years, there have also been more requests for diagnostic data: data which helps find out why something happened. There will always be a need for this type of data, but as an institution, we should find ways to harness the mass amounts of available data to improve student outcomes.

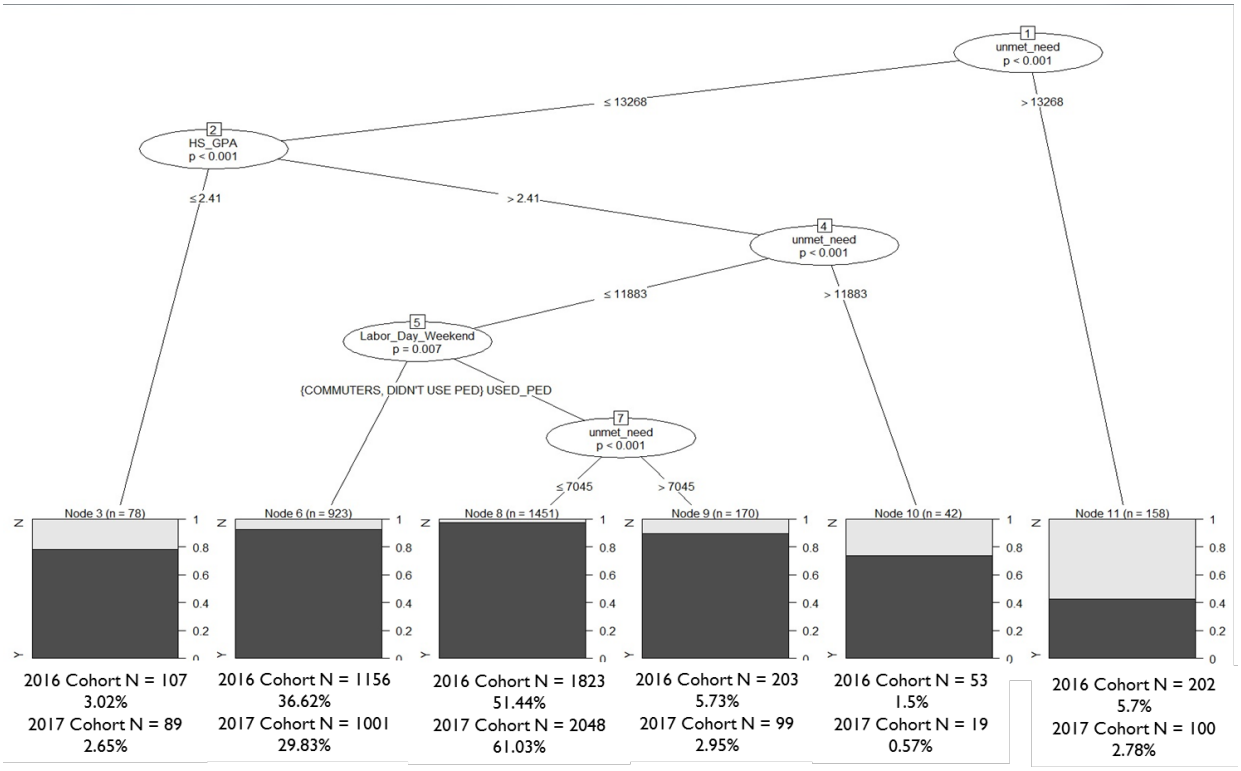
Throughout the process of predictive modeling, we have discovered that while we are doing many things very well, we have also uncovered several areas for improvement. We have detected places where we needed to do a better job of collecting and storing the data. We have found that we should be intentional about the data we collect. Most importantly, we have ascertained that the predictive modeling can help us uncover opportunities to develop strategies around student attributes that we hadn’t considered before, such as Canvas page views or number

of RA interactions. We have discovered that Academic Affairs and Students Affairs professionals must leverage the available data and work together to develop and implement strategies to positively impact student persistence and retention rates.

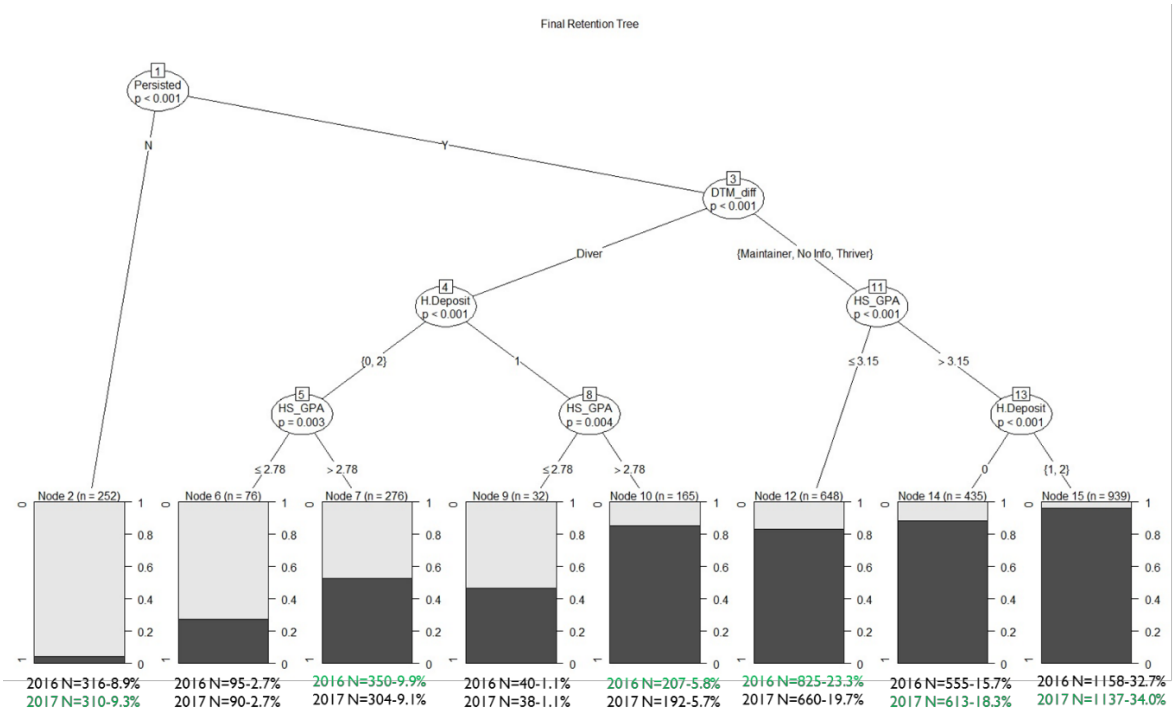
Modeling Timeline:



Decision Tree Persistence Model Output, 2017 Cohort



Decision Tree Retention Model Output, 2017 Cohort



Random Forest, Persistence Model Output, 2017 Cohort

Variable Names	Importance
unmet_need	71.8
HS_GPA	26.2
Canvas_Percentile	26.1
Age	24.6
Days_prior_rank	21.8
Soar	17.6
Acad_Prog	16.5
ACT_Merged_SAT	16.1
Rec_Center_Visits	14.3
FINANCE_CONCERN	11.8
NUMBER_OF_HOURS_PLAN_TO_WORK	11.4
NUMBER_OF_UNIVERSITIES_COLLEGES_ADMITTED	11.1
MathCrs	11.1
Ability_Level	10.9
College_Credit	9.4
OTHER_UNIV_APPLIED	8.3
HIGHEST_ACADEMIC_DEGREE_BGSU	6.9
DECISION_PARENTS_WANTED_ME_TO_GO	6.6
Labor_Day_Weekend	6.6
DECISION_WANTED_TO_GET_AWAY_FROM_HOME	6.3

Random Forest, Retention Model Output, 2017 Cohort

Variable Names	Importance
Persisted	159.32
DTM_diff	94.86
HS_GPA	74.68
Days_prior_rank	61.79
Canvas_Percentile	60.79
Age	59.25
unmet_need	49.39
H.Deposit	49.02
Rec.Total.Visits	47.16
Soar	46.69
ACT_Merged_SAT	41.22
Acad_Prog	35.90
College_Credit	25.52
MathCrs	24.26
LC_Visists_Counts	20.52
Ability_Level	20.10
Ethnicity	18.42
Grad_Plan	17.80
Learning_Communities	9.15
SEX	8.15
LC_Visisted	6.59
Linked_Courses	6.49
MathEnrol	5.63
GSW_1100	3.82
LIVING_CODE	3.69
Student_Athlete	1.94
OBOR_RANK	1.20