MINING DATA TO CREATE A SYSTEM TO IDENTIFY AT-RISK FRESHMEN

Nora Galambos, PhD
Senior Data Scientist
Office of Institutional Research, Planning & Effectiveness
Stony Brook University

Abstract

Data mining is used to develop a series of three models, deployed during orientation through week six, to identify low GPA freshmen in order to improve their outcomes. Customized dashboards are developed to enable users to segment, filter, and list students to assign them to the appropriate advising plans and interventions. Previous modeling has been successful in the early identification of low GPA students and has demonstrated a strong association between learning management system (LMS) logins and GPA outcomes. Factors entered into the predictive models include advising visits, freshmen course-taking activity, LMS logins, college activity participation, SAT scores, high school GPA, demographics, and financial aid.

Introduction

The goal of this data mining effort is to predict as soon as possible, which first-time full-time freshmen students will receive a low GPA in their first term as soon as possible so they can be assigned to interventions. The fall 2012 through fall 2015 freshmen cohort students at our institution who are in the lowest first semester GPA decile had one-year retention rates that ranged from 26 to 34 percentage points lower than those in the second decile. The differences between decile 2, decile 3, and the other deciles combined were much more modest (see figure 1) The results for two-year retention were similar, with differences between decile 1 and decile 2

ranging from 24.6 to 26.3 percentage points. Again, the differences between the higher deciles were much smaller (see figure 2).

The study utilizes information gained and expands upon a fall 2015 study (Galambos 2015) that predicts fall 2015 first-time full-time freshmen GPA's by week 6 of their first semester. That study was our first to use learning management system (LMS) logins in a predictive model. It was determined that learning management system logins did, in fact, have predictive utility and were the were top GPA predictor among students having a high school GPA less than 92.0 (Galambos 2015). Further, the decision tree model provided useful early freshmen GPA estimates, as well as demonstrating differences in the set of predictors for students with different pre-college profiles, most notably high and low high school GPA. A limitation of that study was the lack of archived LMS logins, so only fall 2014 login data was available, leaving only one semester's worth of data available for modeling.

This current study combines fall 2014 and fall 2015 first-time full-time freshmen data to develop three models to predict first semester GPA and builds on methodological information gathered in the development of the previous model. (See the variable list in the appendix for a list of the measures entered into the models.) The first model uses data available on or before orientation, which includes course and major selections, to allow advisors to have an early view of students' possible GPA outcomes to aid in early advising. Course selection and early campus interactions, such as tutoring service utilization and LMS logins, were used to update the model at week three after the end of the drop and add period, and a final model was developed utilizing data through week six. K-fold cross validation was again used to avoid over-fitting, and average squared errors were used to compare the models. Based on the results of the prior study, CART

and CHAID decision tree methods were used for the models with the relative importance measure used to evaluate the relative strength of the variables that are entered into the model.

Figure 1. One-year retention rates of first-time full-time freshmen by first semester GPA deciles and cohort

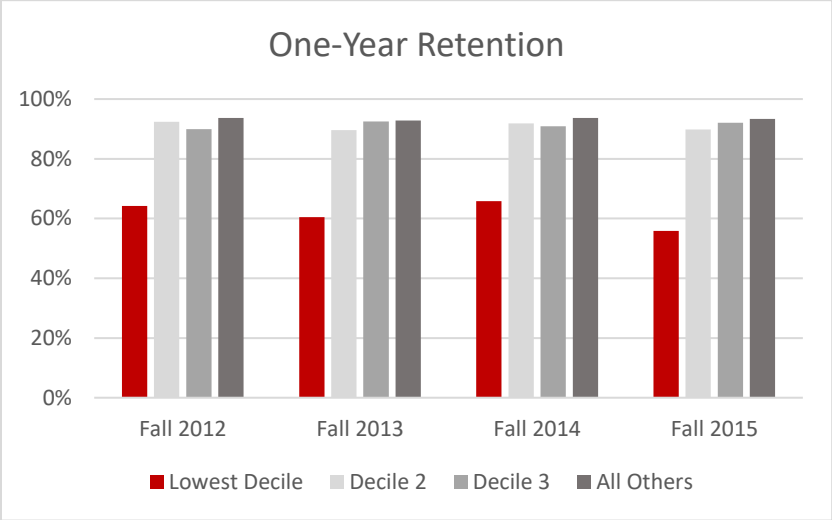
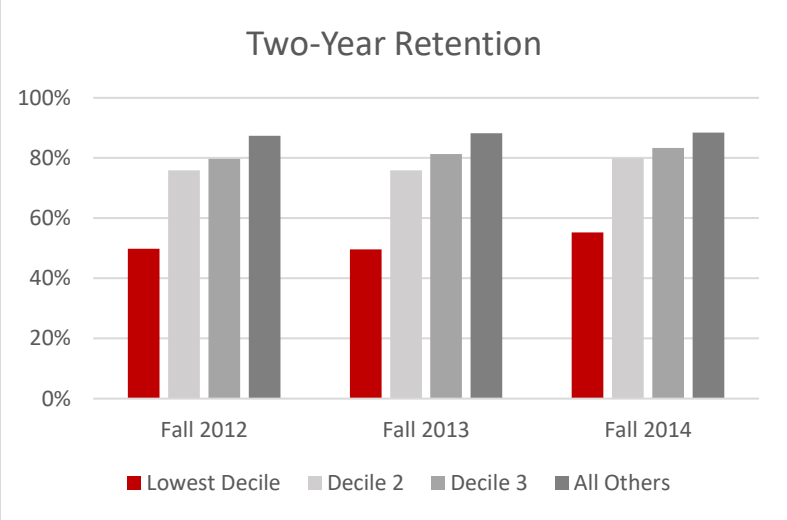


Figure 2. Two-year retention rates of first-time full-time freshmen by first semester GPA deciles and cohort



Dashboards allow users to visualize the predictions and select students for assignment to interventions. Most of the student record data are collected from the university warehouse system, however at present custodians of transaction data are contacted separately to provide LMS logins, advising and tutoring center visits, and other data. These data will eventually be placed in a designated, more easily accessible repository for both data access and archiving purposes.

Literature Review

The study has cast a wide net in terms of assembling a variety of data for use in studying academic, social, and economic factors to determine elevated risk of a low GPA, which can translate to increased risk of early attrition or longer time to degree. Consistent with the retention study of Tinto (1987), we evaluate many types of data representing students' interactions with their campus environment to determine if higher levels of campus engagement are predictive of improved freshmen outcomes. These measures of engagement include interactions with the learning management system, intramural sports and fitness class participation, and academic advising and tutoring center visits. More recently researchers at North Carolina State University presented a study demonstrating that academic achievement is improved by increasing physical activity by just one hour each week (EAB 2016).

It appears that students who are identified to be at risk in their first term and remain at the institution, continue to be at risk, with greater numbers leaving in the subsequent term (Singell and Waddell 2010). This is consistent with the results at our institution which are presented in Figures 1, 2, and 3. Methods capable of more accurate predictions will result in more effective utilization of campus resources, and higher retention and graduation rates. Course-taking

behavior is also important, particularly math readiness. Herzog (2005) found math readiness to be "more important than aid in explaining freshmen dropout and transfer-out during both first and second semesters." To account for the effects of both math readiness and course taking behavior on GPAs, we included our institution's math placement exam results, since the placement exam is administered to all newly enrolled students at our institution. Additionally, we tallied the number of credits of high failure rate courses in which the students were enrolled. Herzog also focused on both merit and need-based aid, and the role that the interaction of aid and academic preparedness plays in student retention. Living within a 60 mile radius of the institution, the percent of students at a high school who take the SAT, along with the percentage at the high school receiving free lunches was explored by Johnson (2008) underlining the need to examine the role of the secondary school and socio-economic factors in developing a model. Persistence increases among students closer to the institution and not surprisingly, decreases among those who were from schools having a high percentage of students receiving free school lunches. The role of differing stop-out patterns exhibited by grant, work-study, and loan recipients (Johnson 2010) demonstrated that grants have the highest positive effect on persistence, but its effect decreases more than that of loans after controlling for other factors.

Resource utilization was studied (Robbins et al. 2009) using a tracking system. Services and resources were grouped into academic services, recreational resources, social measures and advising sessions, with all but social measures demonstrating positive associations with GPA even after controlling for other demographic and risk factors. We have included tutoring center and academic advising visits, and, as previously mentioned, the recreation center usage. The relationship of learning management system usage with student outcomes is of particular interest. A study of five online biology courses (Macfadyen and Dawson, 2010) examined a

variety of LMS tracking measures including the number of discussion messages, new discussion posts, assignments read, and time spent on assignments. Of the 22 LMS metrics evaluated, 13 were significantly correlated with the students' final grades. Further, a regression analysis found that total discussion posts, total mail messages sent, and total assessments completed accounted for 33% of the variation in student achievement scores in the course, and logistic regression correctly categorized as "at risk" 17 of 21 (80.0%) of students who ended up failing the course. In 2007 Romero, et. al. examined a number of data mining methods to demonstrate how they can be used to study outcomes in an open-source LMS online course environment.

These papers have demonstrated that researchers are examining a range of factors in studying and modeling risk. The research highlights the fact that student success is a complex interaction of student engagement, academic service utilization, financial metrics, demographics, combined with student academic characteristics that include high school GPA and SAT scores. Data mining is ideal for developing a model with a large diverse number of predictors.

Methodology

A broad list of data was selected for model development. The more traditional data include demographics, pre-college characteristics, and financial aid measures. In addition to those items the list of college measures includes major groupings, number of AP courses accepted for credit, and number of courses with large proportions of D, F, and W grades, i.e., high DFW courses. A course was coded as a high DFW course if it has an enrollment of at least 70 students with 10% of its grades consisting of D's, F's, or W's. Service utilization data includes Learning Management System (LMS) logins, tutoring center visits, academic advising interactions, and recreation center usage. Studying the use of LMS logins is consistent with research that has shown that engagement with the campus environment improves student outcomes. LMS logins were tabulated as follows. One login per course per hour per student was

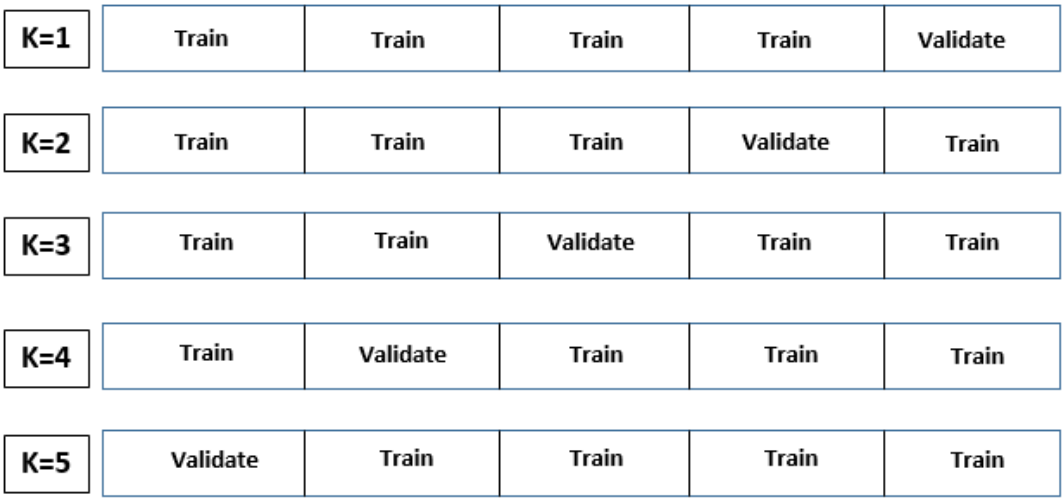
counted, so each student can have a maximum of 24 logins in each course per day. This eliminated multiple logins in the data that occurred just seconds apart. Total logins (using the previous definition) were tabulated for each time period, weeks 1 to 3, and weeks 1 to 6. In addition, total logins were divided by the number of courses utilizing the Learning Management System in which the student was enrolled to create an additional "logins per course" metric. The optimal method for utilizing the LMS data remains an area of active research. Other measures include the average SAT scores of the high schools to control for high school GPA, a variety of financial aid measures, number of enrolled credits grouped by STEM and non-STEM, and AP courses accepted for credit. (See the Appendix for a more complete listing of the data.)

Considerable effort was expended in developing the model to predict the fall 2015 freshmen GPA at week 6 (Galambos 2016). Five different methods were compared with gradient boosting, classification and regression trees (CART), and chi-squared automatic interaction detection (CHAID) having the lowest average squared errors in that order. Because the gradient boosting method yields scoring code, with no explicit, easily understood algorithm or decision tree, and additionally did not demonstrate a substantive error rate reduction, it was not used. Being able to understand how the predictors contribute to student GPA outcomes is useful for selecting and assigning students to interventions and monitoring measures to help keep students on track. The graphic decision tree display is compelling in that regard.

With LMS data available for both fall 2014 and fall 2015, two years of data were used to develop the three models to predict the fall 2016 freshmen GPA's. The total number of first-time full-time fall 2014 and fall 2015 freshmen was 5,664 after 34 students who withdrew prior to the end of the term were removed from the sample. In order to avoid overfitting the model the data are typical divided into training and validation sets. The model is developed using the

training set after which the model is run on the hold out validation sample. We expect similar error results in both the training and validation sets if the model is performing well. Our sample has close to 5,700 students, which may seem sufficient for a 60/40 training to validation data split, however if one considers that over 50 variables are being entered into the model and we are mainly focused on obtaining accurate predictions for the bottom GPA decile, only about 350 students in the group of interest would be left in our sample. As with the previous year's model, K-fold cross validation was used, which allows us to subdivide the sample into 5 groups or folds and run the model five times using 80% of the data and then validating it on the remaining 20%, with a different hold out sample used each time the model is run. Figure 3 shows the 5-fold cross validation scheme. The error results are obtained by taking the average of the five average squared errors (ASE)[1] generated for the training and validation samples for each fold.

Figure 3. K-fold cross-validation sampling design.[2]



[1] ASE = SSE/N or ASE = (Sum of Squared Errors)/N

[2] From Galambos N., (2015). Using data mining to predict freshmen outcomes. *42nd NEAIR Annual Conference Proceedings*, February 2016, p. 89.

The current data were modeled using both CART and CHAID[3]. CART does an exhaustive search for the best binary split at each node. For interval targets the variance is used to assess the splits; for nominal targets the Gini impurity measure is used. The result is a set of nested binary decision rules to predict the outcome. CHAID on the other hand uses the chi-square test to determine categorical splits and F tests for intervals. It allows multiple splits in continuous variables and allows categorical data to be split into more than two categories.

Results

In the fall 2015 study the focus was on identifying measures that can be used to predict freshmen GPA by mid-semester and how those predictors differed by student academic profiles or characteristics. For this set of models, the focus is on the small sections of the decision trees providing the low GPA predictions. The resulting algorithms will be used to assign GPA predictions to the fall 2016 freshmen cohort data for use by the appropriate stakeholders. The average squared errors for the cross validation results, presented in Table 1, are similar to those obtained for the fall 2015 model, but with slightly more concordance between the training and validation errors. A GPA prediction was made on day 1 and forwarded to advisors and others in contact with students so they could take early action or monitor students' progress. Though the average ASE of the CHAID model for day 1 was slightly higher than that of the CART model, the decision was made to use the CHAID method for the day 1 low GPA model, since it had a high level of agreement between training and validation results for the nodes of interest. One of

[3] The CHAID and CART methods have been closely approximated by using Enterprise Miner settings. SAS Institute Inc. 2014. *SAS® Enterprise Miner™ 13.2: Reference Help*. Cary, NC: SAS Institute Inc. p. 755-758.

the predictors, the number of high DFW rate courses in which a student is enrolled, is highly actionable at the beginning of the term.

Table 1. Average Squared Error (ASE) Results for the Three Data Mining Methods

	Day 1 Model (CHAID)		Week 3 Model (CART)		Week 6 Model (CART)	
K Folds	**Validation ASE**	**Training ASE**	**Validation ASE**	**Training ASE**	**Validation ASE**	**Training ASE**
1	0.46	0.41	0.44	0.46	0.43	0.46
2	0.48	0.41	0.45	0.44	0.43	0.44
3	0.50	0.40	0.45	0.46	0.44	0.43
4	0.51	0.40	0.45	0.43	0.46	0.43
5	0.56	0.39	0.51	0.43	0.51	0.42
Average ASE	**0.50**	**0.40**	**0.46**	**0.44**	**0.45**	**0.44**

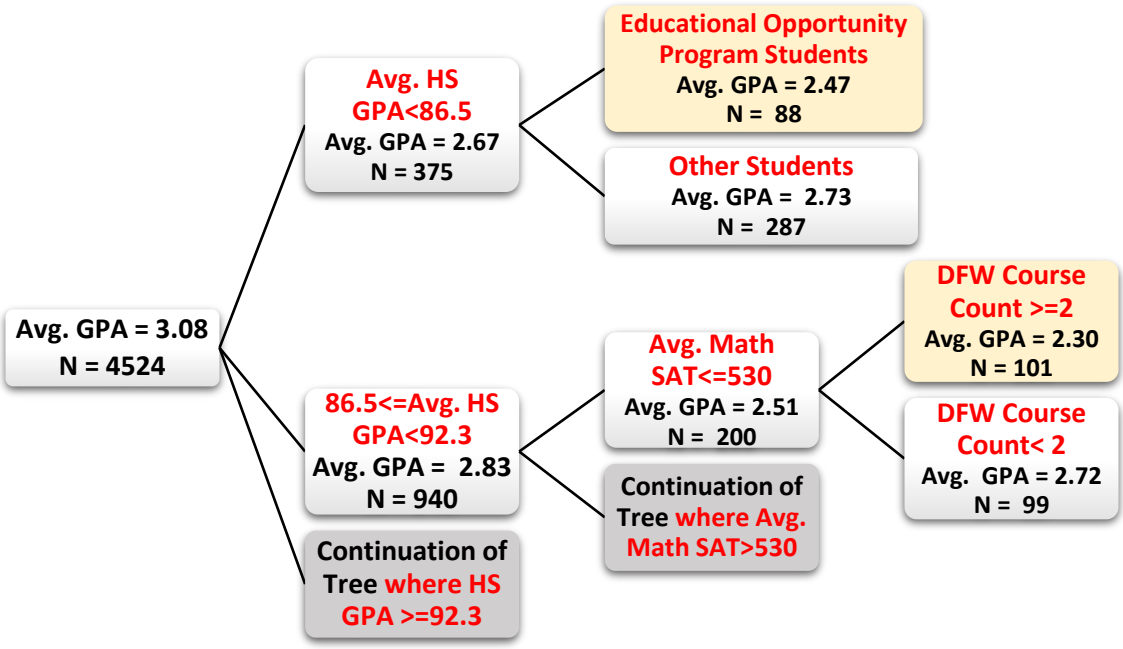


Figure 4. Low GPA Portion of CHAID Model for the Predicting Freshmen GPA on Day 1.

For each of the three models the low GPA section of the corresponding decision tree is shown. Nodes having an average GPA of less than 1.00 are highlighted in orange and those with an average GPA of 1.00 to below 2.50 are highlighted in yellow. Lists of fall 2016 students selected by the decision rules of the highlighted nodes were provided to the appropriate entities on campus. The gray nodes indicate where sections of the decision trees have been truncated to facilitate the graphic presentation, and because they do not contain any nodes in the low GPA range of interest. The node frequencies reflect the fall 2014 and fall 2015 training samples which were used for the model. The model predictors are displayed in red in each node, below which is the predicted GPA for students falling within the corresponding decision rule for the node.

The Educational Opportunity Program (EOP) which figures prominently in the day 1 model (Figure 4) is a program for students whose circumstances, both economic and educational, may have limited their options for obtaining a post-secondary education. Students accepted into the program are typically from historically disadvantage backgrounds and have demonstrated potential for finishing college although they may not have been accepted through the traditional admission process. The program provides financial assistance, tutoring, and mentoring. Because students are admitted to that group by virtue of their lower academic profile, it is not surprising that in the day 1 model some of the EOP students, those with a high school GPA below 86.0, are predicted to have low GPA outcomes. Students having a high school GPA in the 86.5 to 92.3 range, math SAT scores of 530 or less, and are enrolled in 2 or more high DFW rate courses are also predicted to have a low GPA. The average GPA prediction for the EOP students with the lowest high school GPA is 2.47, and is 2.30 for the students with a slightly higher high school GPA, low math SAT scores and two or more high DFW courses. Those are the lowest average

GPA nodes in the entire model. There are none as low or lower within the nodes not displayed in figure 4. A list of those students was provided to the campus stakeholders who provided them with tutoring and peer mentors, as appropriate.

The score distribution table, table 2, part of the decision tree output, has 20 equally spaced bins created by dividing the interval between the highest and lowest predictions by 20, and presents the average GPA and number of students in each interval. Bins with no observations are removed from the table. The model scores are calculated by taking the mid-point of each interval. Since the table shows the number of students at each average GPA level, it can assist in choosing GPA cut points for intervention groups. The number of values in each row are based on a fall 2014 and fall 2015 training sample used for the model.

Table 2. Day 1 Model Score Distribution Table

Prediction Range	Average GPA	N	Model Score
3.80 - 4.00	3.94	3	3.90
3.60 - 3.80	3.71	233	3.70
3.40 - 3.60	3.56	426	3.50
3.20 - 3.40	3.32	1312	3.30
3.00 - 3.20	3.04	932	3.10
2.80 - 3.00	2.91	362	2.90
2.60 - 2.80	2.74	981	2.70
2.40 - 2.60	2.46	147	2.50
2.20 - 2.40	2.31	104	2.30
2.00 - 2.20	2.18	8	2.10
1.80 - 2.00	1.96	5	1.90
1.60 - 1.80	1.61	1	1.70
1.40 - 1.60	1.52	2	1.50
0.60 - 0.80	0.73	4	0.70
0.40 - 0.60	0.50	2	0.50
0.20 - 0.40	0.34	5	0.30
0.00 - 0.20	0.00	1	0.10

As part of the modeling process relative importance measures are calculated and provided as part of the output. “The relative importance measure is evaluated by using the reduction in the

sum of squares that results when a node is split, summing over all of the nodes.[4] In the variable importance calculation when variables are highly correlated they will both receive credit for the sum of squares reduction, hence the relative importance of highly correlated variables will be about the same. For that reason, some measures may rank high on the variable importance list, but do not appear as a predictor in the decision tree."[5] The top importance measures have been included in tables presented below and include measures that may only appear in the portions of the decision trees that have nodes with higher average GPA's. For the day 1 model, high school GPA heads the list, followed by average high school SAT scores, which controls for high school quality, SAT math plus critical reading, size of scholarship received, math placement scores, total DFW STEM credits, and overall total STEM credits.

Table 3. Variable Importance Table for Day 1 Model

Variable	Relative Importance
High School GPA	1.0000
Avg. High School SAT Critical Reading, Math Score	0.5208
Avg. High School SAT Score	0.4921
SAT Math and Critical Reading Score	0.2939
Total Disbursed Scholarship Funds	0.2834
Math Placement Score	0.2806
Total DFW STEM Credits	0.2557
Total STEM Enrolled Credits	0.2145

At week 3 the second model was developed. High school GPA again was the measure that was most associated with average GPA outcomes, with total LMS logins at week 3 associated

[4] SAS Institute Inc. 2014. *SAS® Enterprise Miner™ 13.2: Reference Help*. Cary, NC: SAS Institute Inc. p. 794.
[5] Galambos N., (2015). Using data mining to predict freshmen outcomes. *42nd NEAIR Annual Conference Proceedings*, February 2016, p. 89.

with the GPA outcomes for students with a high school GPA less than 94 (figure 5). Those with less than 61 logins through week3, who attended a school where the average combined SAT math, critical reading, and writing score was less than 1570, and finally had less than 2.1 logins per course as of week three, had an average GPA of slightly below 1.00. If instead they had more than 2.1 logins per course at the week 3 time point, and 5 or more credits in high DFW rate courses, they were predicted to have a GPA of 2.27. If they went to a high school that had an average math, critical reading, and writing exam over 1570 and less than 5.2 LMS logins per course in the first 3 weeks, their average GPA was 2.30.

Table 4. Score Distribution Table for Week 3 Model

Prediction Range	Average GPA	N	Model Score
3.48 - 3.61	3.59	525	3.55
3.35 - 3.48	3.46	383	3.41
3.22 - 3.35	3.29	705	3.28
3.08 - 3.22	3.16	514	3.15
2.95 - 3.08	2.98	1006	3.02
2.82 - 2.95	2.90	694	2.88
2.68 - 2.82	2.72	290	2.75
2.55 - 2.68	2.66	120	2.62
2.28 - 2.42	2.35	93	2.35
2.15 - 2.28	2.27	184	2.22
0.95 - 1.09	0.95	10	1.02

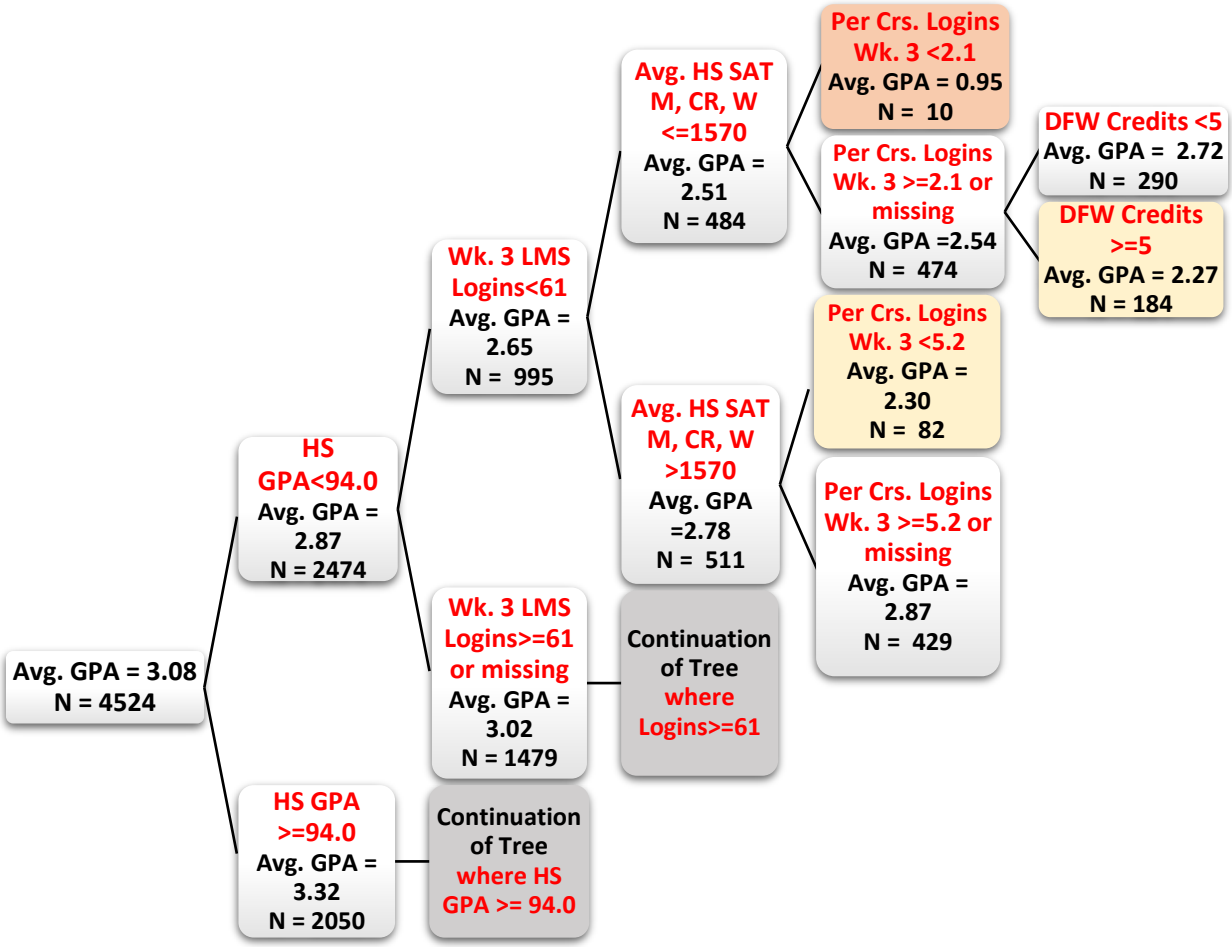


Figure 5. Low GPA Portion of the CART Model Predicting Freshmen GPA: Week 3.

The high school GPA was again at the top of the relative importance list (table 5) and LMS logins also appear on the list with high variable importance scores. The inclusion at week 3 of actual data on the students' interactions with the campus and their courses has strengthened the magnitude of the importance measures and has altered the variables included on the list considerably. Note that although ethnicity, geographic location, and tuition did not appear on the

day 1 model, those items (and other demographic and pre-college measures) were, in fact, entered into the day 1 model.

Table 5. Variable Importance for Week 3 Model

Variable	Relative Importance
High School GPA	1.0000
IPEDS Ethnicity	0.8600
Academic Level	0.8281
Area of Residence at Admissions--6 Categories	0.8152
Total LMS Logins at Week 3	0.8016
Major Type—Major, Undeclared, Area of Interest	0.7516
Residency Tuition	0.7483
SAT Math and Verbal Combined	0.6246
Per Course STEM LMS Logins, Week 3	0.5867
Per Course STEM Total LMS Logins, Week 3	0.5836
Total DFW STEM Units	0.4527
Avg. SAT CR+M+W Avg. for the High School	0.4417

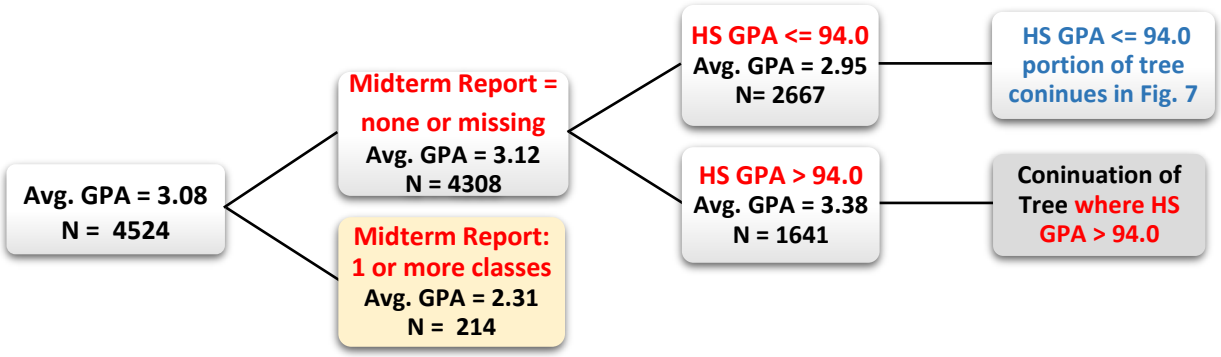


Figure 6. Low GPA Portion of the CART Model Predicting Freshmen GPA at the End of Week 6: Part 1.

The final model, using data as of the end of week 6, is presented in two parts, shown in figures 6 and 7. Part 2, figure 7, continues from top, right node with blue text.

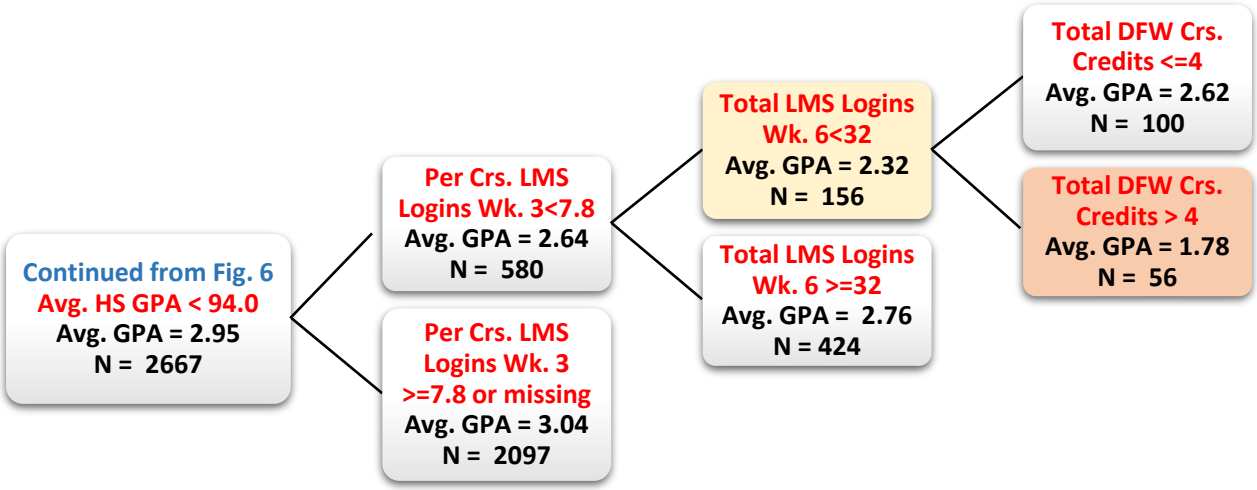


Figure 7. Low GPA Portion of the CART Model Predicting Freshmen GPA at the End of Week 6: Part 2.

In the week 6 model the top measure associated with the GPA outcome was the midterm grade report. The midterm report is a requested from professors by Academic Advising around week 6. Not all professors respond, but those that do provide midterm grade information for students in their classes. Academic Advising reaches out by email to all students on the list. The midterm report data element in the model represents the number of courses for which academic advising received a report pertaining to the student. The fall 2015 midterm report list was the first one available for the modeling process, hence the measure is missing for fall 2014. The average GPA for the midterm report node is 2.31. Those for whom there was no midterm report and additionally had a high school GPA of 94.0 or less, the model continues in figure 7. For those low high school GPA students who additionally had low LMS logins at week 3 and week 6, and more than 4 credits of high DFW rate courses, the predicted GPA was 1.78 (see figure 7).

In the week 6 model, an LMS login measure has risen to the top of the variable importance table and we see the table populated with a number of LMS login measures along with high school GPA and the midterm report.

Table 6. Week 6 Model Score Distribution Table

Prediction Range	Average GPA	N	Model Score
3.406 - 3.556	3.50	1091	3.48
3.257 - 3.406	3.38	295	3.33
3.107 - 3.257	3.21	302	3.18
2.958 - 3.107	2.99	2040	3.03
2.808 - 2.958	2.96	157	2.88
2.509 - 2.659	2.64	367	2.58
2.360 - 2.509	2.45	198	2.43
1.762 - 1.912	1.78	56	1.84
0.567 - 0.716	0.57	16	0.64

Although there is some variety in the measures predicting the GPA outcomes, we find high school GPA, number of high DFW rate courses, and LMS logins playing a prominent predictive role in all three models. In terms of the lowest high school GPA EOP students, we notice that the EOP student group did not appear again in the week 3 and week 6 models. As previously discussed, students in that program receive tutoring, peer mentoring, and other academic assistance, so clearly once that the semester progressed those EOP students as a group were no longer predicted to have a low GPA. In fact, the academic support program for the EOP students can serve as a model in designing interventions for other students. With profiles that resulted in some of them being predicted to have a low GPA in the day 1 model, it is important to note that the fall 2009 freshmen cohort EOP students had a six-year graduation rate of 79.7%, well surpassing 68.3 %, the rate for the entire fall 2009 cohort.

Table 7. Variable Importance for Week 6 Model

Variable	Relative Importance
Per Course Total LMS Logins, Week 3	1.0000
High School GPA	0.9731
Total LMS Logins, Week 3	0.8932
Total LMS Logins, Week 6	0.7606
Academic Level	0.6842
Midterm Report	0.6300
Area of Residence at Admissions--6 Categories	0.6102
Dorm Housing Indicator	0.5923
Women in Science and Eng. Program	0.5848
Total LMS Non-STEM Logins, Week 6	0.5047
Per Course Non-STEM LMS Logins, Week 6	0.4952
Non-STEM Total Logins, Week 3	0.4345
Per Course Total LMS STEM Logins, Week 3	0.4339

Data Delivery

Samples of data delivery methods with filters and the ability to drill down to the student level data can be found in the Appendix. Dashboards can easily be customized depending upon the user. Advisors may want to be able to easily find the students in various predicted low GPA groups, then drill down and view their schedules and other information. As evidenced by the graphs in the introduction, intervening early in imperative because roughly half may be gone by the end of their second year. Departments may also want to determine how many majors they have who are predicted to have low GPA's to motivate their own interventions and department advising. Since the number of students predicted to be on the lowest end of the GPA spectrum is only 10 to 15 percent of the freshmen, providing the data in spreadsheet form can also suffice.

Conclusion

The modeling process has demonstrated that measures most strongly associated with low GPA outcomes are related to how individuals perform as students, as evidenced by the variable

importance scores of the high school GPA and LMS logins. Providing predictive information to those providing support services can head off a potentially damaging low GPA outcome. Additionally, being alerted to the lower high school GPA students, who may be taking multiple difficult courses may help advisors to be more pro-active in terms of helping students with challenging course schedules. Peer mentoring and tutoring have already been suggested to some such students at our institution.

Since the data used is being pulled in from many sources, the next logical step is to create a repository allowing easier access, which will in turn streamline the modeling process. Additionally, the data being collected contains information on services being provided to students such as tutoring and advising. These data not only have predictive utility and can be used to track interventions, but are also a gold mine of information that can be used to understand and study what students are utilizing the services we are providing to enable them to succeed.

References

Bahr, P.R. (2008). Does mathematics remediation work?: a comparative analysis of academic attainment among community college students. *Research in Higher Education*. 49:420-450.

Bean J. (1983). The application of model of turnover in work organizations to the student attrition process. *Review of Higher Education*, 6(2), 129-148.

Breiman, L., Friedman, J., Olshen, R., Stone, D. (1984): *Classification and Regression Trees*. Wadsworth Books.

Chen R. (2012). Institutional characteristics and college student dropout risks: a multilevel event history analysis. *Research in Higher Education*. 53:487–505.

Friedman, J., Hastie, T., Rosset, S., Tibshirani, R., Zhu, J. (2003) Discussion of Boosting Papers. Retrieved from http://web.stanford.edu/~hastie/Papers/boost_discussion.pdf

"Gym time correlated with graduating, making better grades." Retrieved from: https://www.eab.com/daily-briefing/2016/05/05/gym-time-correlated-with-graduating-making-better-grades?WT.mc_id=Email|Daily+Briefing+Headline|DBA|DB|May-05-2016|||||&elq_cid=1621083&x_id=003C000001xsbjvIAA . n.p. May 5, 2016.

Galambos N. (2015). Using data mining to predict freshmen outcomes. *42nd NEAIR Annual Conference Proceedings*, February 2016, p. 89.

Herzog S. (2005). Measuring determinants of student return vs. dropout/stopout vs. transfer: a

first-to-second year analysis of new freshmen. *Research in Higher Education*. 46:883-928.

Johnson I. (2006). Analysis of stop-out behavior at a public research university: the multi-spell discrete-time approach. *Research in Higher Education*. 47:905-93.

Johnson I. (2008). Enrollment, persistence and graduation of in-state students at a public research university: does high school matter? *Research in Higher Education*. 49:776-793.

Macfadyen LP, Dawson S. Mining LMS Data to Develop an "Early Wrining System" for Educators: A

Proof of Concept. *Computers & Education*. 2010, 54: 588-599.

Parker M. (2005). Placement, retention, and success: a longitudinal study of mathematics and retention. *The Journal of General Education*. 54:22-40.

Robbins S, Allen J, Casillas A, Akamigbo A, Saltonstall M, Campbell R, Mahoney E, Gore P. (2009). Associations of resource and service utilization, risk level, and college outcomes. *Research in Higher Education*. 50: 101-118.

Romero C, Ventura S, Garcia E. Data Mining in Course Management Systems: Moodle Case Study and

Tutorial. *Computers & Education*. 2008, 51: 368-384.

SAS Institute Inc. (2014). *SAS® Enterprise Miner™ 13.2: Reference Help*. Cary, NC: SAS Institute Inc.

Singell L, Waddell, GR. (2010). Modeling retention at a large public university: can at-risk students be identified early enough to treat? *Research in Higher Education*. 51:546-572.

Stater M. (2009). The impact of financial aid on college GPA at three flagship public institutions.

American Educational Research Journal. 46:782-815.

Stinebrickner R, Stinebrickner T. (2014). Academic performance and college dropout: using longitudinal expectations data to estimate a learning model. *Journal of Labor Economics*. 32:601-644.

Thomas EH, Galambos N. (2004). What satisfies students? mining student-opinion data with regression and decision-tree analysis. *Research in Higher Education*. 45:251-269.

Tinto, V. (1987). Leaving college: rethinking the causes and cures of student attrition. Chicago, IL: The University of Chicago Press.

Zwick R, Sklar JG. (2005). Predicting college grades and degree completion using high school grades and SAT scores: the role of student ethnicity and first language. *American Educational Research Journal*. 42:439-464.

Appendix

Variable List

Demographics
Gender
Ethnicity
Area of residence at time of admission: Suffolk County, Nassau County, New York City, other NYS, other US, International

Pre-college Characteristics
High School GPA
College Board SAT Averages by High School
- Average High School Critical Reading
- Average High School SAT Math
- Average High School SAT Critical Reading + Math

SAT: Math, Critical Reading, Writing, Math+Critical Reading

College Characteristics
Number of AP STEM courses accepted for credit
Number of AP non-STEM courses accepted for credit
Total credits accepted at time of admission
Total STEM courses
Total STEM units
Total Non-STEM courses
Total No-STEM units
Class level
Dorm Resident
Intermural Sports Participation
Fitness Class Participation
Honors College
Women in Science and Engineering
Educational Opportunity Program
Stony Brook University Math and Writing Placement Exams
College of student's major or area of interest: Arts and Sciences, Engineering, Health Sciences, Marine Science, Journalism, Business
Major Group: business, biological sciences health sciences, humanities and fine arts, physical sciences and math, social behavioral science, engineering and applied sciences, journalism, marine science, undeclared, other
Major type: declared major, undeclared major, area of interest
High DFW Rate Courses: enrollment >= 70, percent DFW >=10%
- Total high DFW STEM units
- Total high DFW non-STEM units
- Highest DFW rate among the DFW Courses in which the student is enrolled
- Highest DFW rate among the DFW Courses in which the student is enrolled

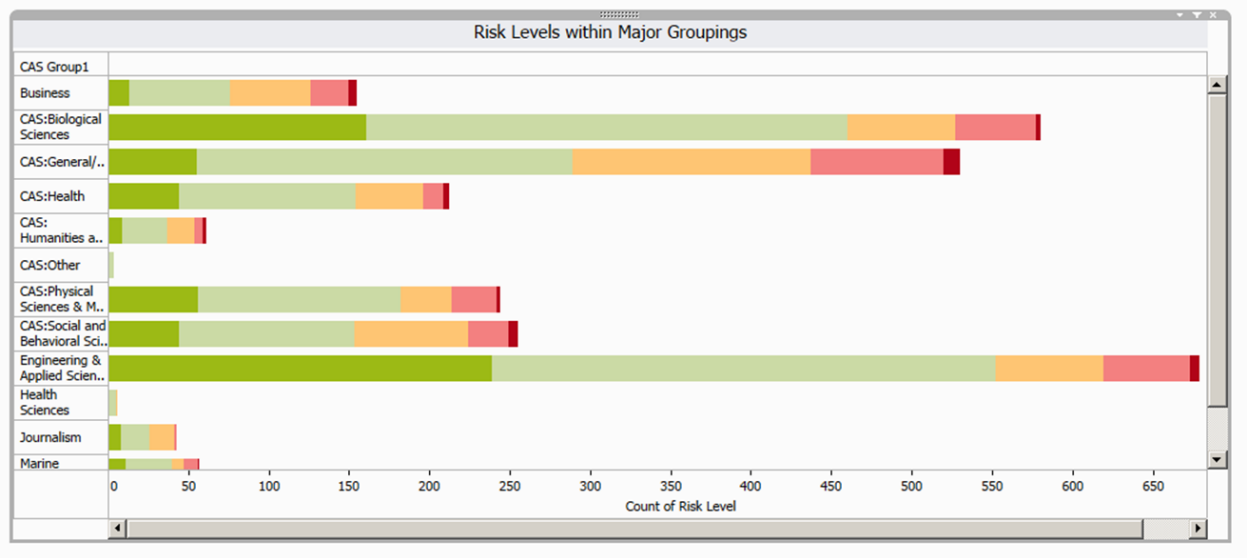
Proportion of freshmen in a student's highest DFW rate STEM course
Proportion of freshmen in a student's highest DFW rate non-STEM course
Type of math course: high school level, beginning calculus, sophomore or higher math

Financial Aid Measures
Aid disbursed in the Fall 2014 and Fall 2015 academic years
Total grant funds received
Total Loans recorded by the Financial Aid Office
Total scholarship funds received
Total work study funds received
Total athletics aid received
Athletic aid, grant, loan, PLIS loan, subsidized/unsubsidized loan, scholarship, work study, TAP, Perkins, Pell indicators
Adjusted Gross Income
Federal Need
Federal Expected Family Contribution
Dependent status

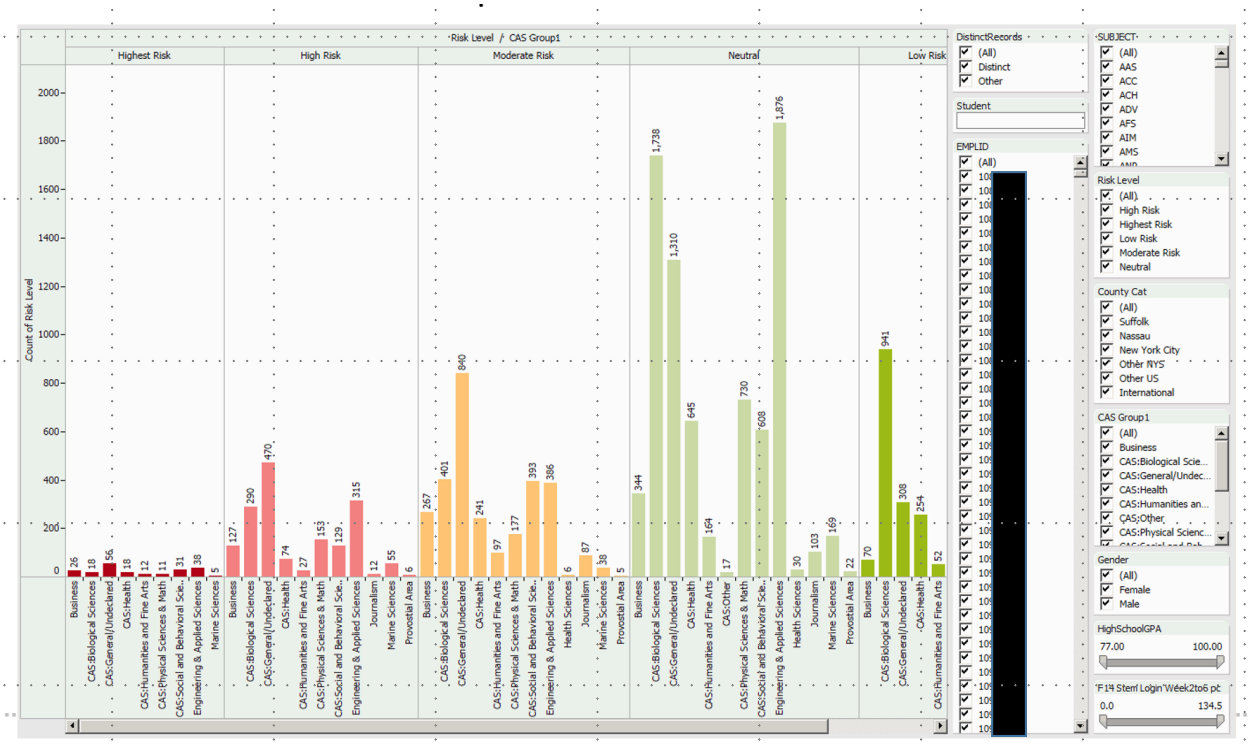
Services/Learning Management System (LMS)
Advising Visits/Tutoring Center Usage
Tutoring center appointment no shows
Number of STEM Course Center Visits, weeks 1 to 6
Number of non-STEM Course tutoring Center visits, weeks 1 to 6
Advising Visits during week 1- 3
Advising visits during weeks 3 – 6
Course Management System Logins
F14 and F15 Stem Logins
F14 and F15 NonStem Logins Weeks 1 -3
Non-STEM course related logins during weeks 3 - 6
Non-STEM Course related logins during week 1 -3
STEM Course related logins during week 1 -3
STEM Course related logins during weeks 3 to 6
Number of STEM course logins per STEM course using the CMS, weeks 1 – 6.
Number of non-STEM course logins per non-STEM courses using the CMS, weeks 1 – 6.

Dashboard Samples

Red and pink colors represent the lowest GPA levels. Checkbox lists allow filtering.



Users can choose majors, id number, can find a student by entering their name. Sliders at the bottom right allow selection of a GPA and/or LMS login range.



Double clicking on any bar above, allows drilling. down to student data.