Spring 2018

Modeling Term 3 Retention After Term 1 Completion

Torrence Savage

MODELING TERM 3 RETENTION AFTER TERM 1 COMPLETION

Torrence Savage

SPRING 2018

ST. CLOUD STATE UNIVERSITY BELONGING TEAM
720 4th Ave S, Saint Cloud MN, 56301

Table of Contents

Executive Summary

The St. Cloud State University Belonging Team is investigating how students' sense of belonging affects their retention at the college. It was found that Belonging Index, Term 1 Cumulative GPA, Term 1 Credit Completion Rate and other Demographic variables were useful in prediction.

The goal for this project is to create two models for predicting the likelihood of Term 3 retention after the student's first semester. These will be based on the Fall 2014 and 2015 cohorts of new first-year domestic students. The data we have includes measures of academic success, referred to as Academic Outcomes, including GPA, credits attempted, and credits completed among others. The demographic information has variables of interest including gender, whether or not SCSU is the closest university to the student's home, financial aid gap, etc. Students' sense of belonging was measured by a series of questions taken by online survey designed by the St. Cloud State University Belonging Team. There were 10 questions ranging from a student's commitment to complete their degree to how much they regret leaving home. The answers to these questions were measured on 7-point scale and converted to a new 5-point scale variable called Belonging Index.

One of the models created has predictors that include just Belonging Index and Academic Outcomes. The second has Belonging Index, Academic Outcomes, and Demographics as predictors. These two models are compared with the previous model, created in Fall 2017 by the author, that only had Belonging Index and Demographics as predictors. The best model will be chosen based on model accuracy, statistical and practical significance of the predictors, and the complexity of the model. The best model was applied to the Fall 2017 cohort of new first-year domestic students to find their average predicted probability of Term 3 enrollment and compare it with the Fall 2014-2015 cohorts.

It was found that a student's original Belonging Index is still a very strong predictor of Term 3 enrollment after a student's first semester. Also, Term 1 Cumulative GPA and Term 1 Credit Completion Rate were better predictors of Term 3 enrollment than all of the other previous variables involved.

The average predicted probability of enrollment for the Fall 2017 cohort, after Term 1 and excluding Term 2 dropouts was 79.9% compared to an enrollment of 79.1% in the 2014-2015 cohorts.

The predicted probabilities of enrollment for all students of the 2017 cohort were saved and sent to Dr. Robinson. Now they can be grouped according to their predicted probability of enrollment, and targeted for help based on who is most at risk.

Background

In total, four data sets were received from Dr. Robinson. The first two had more Demographic and Belonging information, hereafter referred to as the “Belonging sets” One was for the Fall 2014 and 2015 cohorts of new entering freshmen at St. Cloud State University and one for the Fall 2017 cohort. They were received in Fall 2017 and had formatting done that was listed in the previous report. The next two had more in-depth Academic Outcome information like Credits Completed and Cumulative Quality Points, hereafter referred to as the “Academic Outcome sets”. Again, one for the Fall 2014 and 2015 cohorts and another for the Fall 2017 cohort.

Past Results

The goal for the previous Belonging Team Term 3 Retention was to be able to predict Term 3 retention at the beginning of a student’s 1st semester. Three logistic regression models were created, one for all students, one for Non-STEM students, and one for just STEM students. The predictor variables included Belonging Index, Quality Points Predicted (QPP), demographic information, as well as several interactions between the terms. It was found that for demographic information, Student of Color (SOC), Gender, Closest to SCSU (Closest), ACE, Advising and Registration Month of August (AR Month August), Pell Grant Eligibility (Pell Eligible), and STEM were all useful in the overall model. However, Pell Eligible was slightly weak and could possibly be replaced by a different financial variable. The interactions between Belonging and Closest, Belonging and Gender, QPP and SOC, and STEM and ACE were also useful in the overall model. The results for the Non-STEM group were similar to the Overall group. However, for the STEM group, Gender and its interaction weren’t useful in the model.

Goals

Create a predictive model based on the Fall 2014 and 2015 cohorts Belonging Index, Academic Success measures, and potentially demographic information to predict Term 3 enrollment. This model would be created after students’ first semester is complete when Academic Outcomes are available. Examine predictors based on statistical significance and effect size. Use this model to find the predicted probabilities of Term 3 enrollment for the Fall 2017 cohort.

Data Formatting

First, the two data sets for each year group were merged. This was done using a join in JMP and matching the Student ID from the Belonging set, with the Tech ID of the Academic Outcome set. The 2014-2015 Belonging data set originally had 2639 students and the 2017 Belonging data set had 616 students. After merging, the 2014-2015 data set had 2639 students and the 2017 data set had 615 students. One student must have been missing from the 2017 Academic Outcomes set for some reason, but it couldn't be determined why.

Then, categorical and numerical versions of each possible predictor variable were created for data exploration and analyzation respectively for the 2014-2015 data.

Also, to apply the final model to the 2017 data, the predictor variables chosen were standardized to be the same type of column and have the exact same column name in both 2014-2015 and 2017 data sets.

It was determined that only students who had been enrolled in Term 2 were going to be used in this model. Obviously not being enrolled in Term 2 was highly related to not being enrolled in Term 3. There were 275 students who weren't enrolled in Term 2, and of those students, only 17 were enrolled in Term 3. All of these were excluded from our data set. Therefore, the data set was left with 2364 students to base these models on.

There are a lot more students than you might expect in the 2014 and 2015 data set compared to the 2017 data set. This is partially due to a lower enrollment size in 2017, but mostly because the response rate was much lower. For the 2014-2015 survey, the response rate was about 80%. For 2017 survey, the response rate was closer to 50% giving us a smaller sample size. Better incentives to respond to the survey could improve our ability to predict enrollment.

The variable "Drop in Credits Attempted" was created using a column formula that took the number of credits attempted in Fall (Term 1) minus the number of credits attempted in Spring (Term 2).

Since T1 Credit Completion Rate had a range of 0-1, and effect shows the difference in one unit of change in the predictor, it was multiplied by four to have approximately the same relative size effect as Belonging Index and GPA whose ranges were also four. This ensures that its effect didn't appear overly large in the model.

Process

Different Predictor variables will be examined to create the best possible model to predict Term 3 enrollment. One of the predictor variables will be Belonging Index. They will also include Academic Outcomes for Term 1 and the beginning of Term 2. These Academic Outcome predictors may include Term 1 Cumulative GPA and some form of Credits Completed. Also, the drop in credits attempted from Fall to Spring could be useful as well. Demographic information will be examined as well. Demographic variables could include the following: Student of Color, Gender, Closest to SCSU, ACE, AR Month of August, a financial variable, and STEM. Each of these variables will be tested for statistical significance and effect size.

Then, two different models will be created based on these variables. One will include all of the terms and one will just have Belonging and Academic Outcomes. These models may have possible interactions between the terms included. These will be compared to the previous model to see which is best based on model accuracy, effect sizes, p-values, and the complexity of the model.

Finally, the prediction formula for the best model will be saved to a column formula and applied to the Fall 2017 data to predict the probability of enrollment.

Preliminary Information

For the 2014-2015 data, retention rate for all 2639 students at the beginning of the first semester was 71.5 %. Of the 2364 students who made it through their first semester and were enrolled in Term 2, the Term 3 retention was 79.1%.

P-values, Effect Sizes, and Interactions

The typical measurement of statistical significance is the p-value. It is a measure of the likelihood of observing results as extreme as the given data, given that the null hypothesis is true. In this case, it is testing that the individual predictor has an effect in the given model. Low p-values indicate strong statistical significance and high p-values indicate no statistical significance. In this case, it's judging whether or not the predictor has an effect in the model, when all of the other terms are included. It can change in a model based on whether or not other terms are included. For example, models with more terms will tend to have higher p-values because it is harder to determine each predictor's individual statistical significance with all the other terms included.

Practical significance is the size of this effect. A predictor could be determined to have an effect in the model, but this effect could be very small. Effect size shows for an average student, with all other predictors remaining the same, the difference of the predicted probability of enrollment between two students who have a one point difference in that predictor. For example, comparing males and females, males tend to have higher retention rates. Everything else remaining the same, a female will have a lower predicted probability of enrollment than a male. Since Gender is 0 for Male and 1 for Female, Gender will have a negative effect size. It's important to not just throw away terms based on high p-values because some of those terms could have large effect sizes that are important in the model.

Interactions can increase p-values of both the main effect and the interaction term because we are adding more terms that are correlated and increasing multicollinearity in the model. They also change how we look at effect size. It makes more sense to look at how Belonging changes affect retention differently for Male and Female, rather than considering effect sizes of each predictor individually.

Because a goal of this project was to compare the effect size of the main effects and interactions before and after Term 1, some of the p-values will be quite higher than what we would generally look at. Some of these predictors had interactions that inflated their p-values and some had large effect sizes and so were kept in a model for comparison purposes.

Previous Results

The following are the main predictors used in the previous model: Belonging Index, QPP/10, AR Month of August, Closest to SCSU, Gender, Student of Color, Pell Eligibility, STEM, and ACE.

Belonging Index is a measure of a student's sense of belonging on a scale of 1-5. It is scored based on 10 questions from the Belonging survey sent to incoming freshmen.

QPP is Quality points predicted. It's a measure of students past academic performance and has a range of about 100 points, so it was divided into units of 10 to give it a more comparable effect size.

The rest of these are binary variables of either 0 or 1.

AR Month of August is a 1 if their advising and registration is in August, otherwise a 0.

Closest is a 1 if St. Cloud State is the closest university to the student's home and 0 if not.

Gender is 1 for Female, 0 for Male.

Student of Color is 1 for SOC and 0 for Non-SOC.

Pell Eligibility is 1 for students eligible for the Pell Grant and 0 for those who are not.

STEM is 1 for a STEM Majors and 0 for non-STEM majors.

ACE is 1 for at-risk students receiving help through the ACE program, 0 for everyone else.

There were also several interactions that were found to be important in the data. For example, Belonging was more important for females than it was for males. Increasing Belonging Index for those who are Female increases retention rate more than the same increase in Belonging for those who are Male. This holds true going the other way as well. Lower Belonging Index for Female has a larger negative effect than it does for Male. For students who are not close to home, Belonging is more important than it is for those who are close to home. For Students of Color, QPP is less important than it is for those who are not. For students who are STEM, being in the ACE program helps more than for non-STEM students. There weren't a lot of students who were both STEM and ACE, so the statistical significance is going to be weak. But, the effect size is quite large so it was kept in the model for now.

Here are the predictors and their p-values from the previous model.

Previous Model (Table 1)

Predictors	P-values
QPP	0.0001
Belonging Index	0.0001
AR Month	0.0033
Closest to SCSU	0.0095
SOC	0.0283
QPP/SOC	0.0388
STEM/ACE	0.0391
Belonging/Closest	0.051
Gender	0.0629
ACE	0.0682
Belonging/Gender	0.0874
Pell Eligible	0.1397
STEM	0.1683

The p-values for STEM and Pell Eligible are slightly higher than is desirable. Remember this model was created based on information from early in first semester. Attempting to apply that exact model after Term 1 is completed could be a poor decision. It doesn't take in to account when students are dropping out before Term 3. It could be in Term 1 or it could be in Term 2. We could apply the same model and get different results because of the timing of dropouts. Also, while past information is good, current academic performance information like GPA and Credit Completion Rate is very important.

These are the p-values from the previous model not including students who were not enrolled in second semester.

Previous Model after Term 2 Dropouts (Table 2)

Predictors	P-values
QPP	0.0001
Belonging Index	0.0001
AR Month	0.0045
Closest to SCSU	0.0533
SOC	0.0601
QPP/SOC	0.0865
STEM/ACE	0.09
Belonging/Closest	0.145
Gender	0.1826
ACE	0.1828
Belonging/Gender	0.259
Pell Eligible	0.3594
STEM	0.5646

These values could be elevated because of the timing of dropouts, model overfitting, or because 10% of the sample size is gone and it's harder to determine statistical significance. Regardless, a new model specifically for application after Term 1 completion will be useful.

New Results

There were three predictors that were examined for the new model: GPA, Credit Completion Rate, and Drop in Credits attempted. The version of GPA looked at was cumulative GPA after Term 1 because it included some students' college credits they had taken in high school and includes past success at the college level. Credit completion rate was the students credits completed divided by their credits attempted during Term 1. That resulted in a score of 0 to 1. This was multiplied by 4 to give it a scale from 0-4 to more easily compare it to belonging index and GPA which also had a range of 4. For example, if a student is taking 16 credits, one 4 credit class can be compared to 1 point in GPA for effect size. Drop in credits attempted was the student's credits attempted in the Fall minus the credits attempted in Spring.

I looked at new interactions briefly but I didn't think they were necessary in the model.

Academic Outcomes Model

The first version of the Academic Outcomes Model is shown here. I used Belonging Index, Drop in Credits Attempted, and T1 Cumulative GPA as the predictors.

Academic Outcomes Model 1 (Table 3)

Predictors	Effect	P-values
T1 Cumulative GPA	7	0.0001
T1 Credit Completion Rate	9	0.0001
Belonging Index	9	0.0001
Drop in Credits Attempted	0	0.8966

Drop in Credits Attempted was not statistically or practically significant with a large p-value and small effect size. We know that Drop in Credits Attempted definitely has a correlation with Term 3 enrollment. But, it's not useful in the model when considering Credit Completion Rate and GPA.

Here is the final version of the Academic Outcomes Model with Drop in Credits Attempted removed. T1 Cumulative GPA, Credit Completion Rate, and Belonging Index are all excellent predictors of Term 3 enrollment and provide a good base for a prediction model.

Academic Outcomes Model 2 (Table 4)

Predictors	Effect	P-values
T1 Cumulative GPA	7	0.0001
T1 Credit Completion Rate	9	0.0001
Belonging Index	9	0.0001

Academic Outcomes and Demographics Model

Here are the p-values of a model fit with the useful Term 1 academic outcomes plus all of the previous demographic predictors.

Academic and Demographics Model 1 P-values (Table 5)

Predictors	P-values	Power (Increase)	Power (Decrease)
T1 Credit Completion Rate	0.0001	93%	83%
Belonging Index	0.0001	67%	52%
T1 Cumulative GPA	0.0003	94%	84%
QPP	0.0012	100%	99%
STEM	0.0371	53%	40%
Closest to SCSU	0.0494	6%	6%
QPP/SOC	0.1054	92%	80%
SOC	0.1087	7%	6%
ACE	0.1245	44%	33%
Belonging/Closest	0.1385	40%	30%
Gender	0.2624	7%	6%
AR Month	$\boxed{0.2949}$	44%	32%
STEM/ACE	0.3607	13%	10%
Belonging/Gender	0.3989	49%	37%
Pell Eligible	$\boxed{0.7493}$	77%	62%

There are several of these p-values that are too high, in part because of how many terms are in the model. Remember some of these terms had interactions and large effects and so were included for comparison. However, when we look at AR Month and Pell Eligible, they had no interactions, small effect sizes, and very weak statistical significance. They were removed from the model because they were not useful in this case.

The power shows how likely an effect of five percentage points is to be detected, given that there is an effect to detect. If there was an effect for Pell Eligible, it is likely to be determined statistically significant, indicating that Pell Eligible does not have an effect of five percentage points. AR Month however, has a weak power. The large standard deviation of the term gives it a weak power, and it is difficult to detect an effect. A larger sample size would decrease this standard deviation, increase the power of the tests, and be able to determine more predictors statistically significant in the model. This would improve the model's prediction capabilities, thereby improving efforts at improving retention. For more questions regarding power, see Appendix.

Academic and Demographics Model 2 Estimates and P-values (Table 6)

Terms	Estimate	P-values
Intercept	-6.8276	0.0001
T1 Credit Completion Rate	0.7254	0.0001
Belonging Index	0.7052	0.0001
T1 Cumulative GPA	0.3512	0.0002
QPP/10	0.2253	0.0008
STEM	0.3428	0.0362
Closest to SCSU	1.4709	0.0509
ACE P	0.2929	0.1095
QPP/SOC	-0.1575	0.1107
SOC	1.1296	0.1235
Belonging/Closest	-0.2896	0.1382
Gender	-0.7429	0.2654
STEM/ACE	0.3313	0.3996
Belonging/Gender	0.1438	0.4035

Here is the model used for prediction with p-values and estimates included. It can be seen that only the terms that involve interactions have elevated p-values, indicating that it's more difficult to determine statistical significance because of the correlation between the interaction term and the original predictor term. That being said, all of the same interactions were still present in the data after Term 1 and the terms are still useful in the model.

Comparison of Models and Predictors

This is a comparison between the main effects of the new model (Academic and Demographic), and the previous model.

Main Effects Comparison (Table 7)

Predictor	New	Previous
T1 Cumulative GPA	5	-
T1 Credit Completion Rate	10	-
SOC	13	19
Gender	-15	-24
Closest to SCSU	15	21
Pell Eligible	-	-3
AR Month	-	-9

Again T1 cumulative GPA and Credit completion rate were great predictors, but they can't be compared to the previous model because they aren't in it.

All of these terms in the new model have smaller effect sizes due to the new terms that are in the model.

SOC actually had a positive effect. However, it is known that in this data set students of color have lower retention rates. This is because of confounding variables that have a negative effect that lowers SOC retention rate.

Being Female again had a large negative effect on enrollment. But women tend to have higher GPA's than men which improves retention.

Closest to SCSU was similar in both models. Being close to home had a large positive effect on enrollment.

Pell Eligibility was weak when predicting enrollment before Term 1 completion and was not useful when predicting enrollment after Term 1 completion. A different financial variable should be considered.

Again, AR Month was not useful in prediction after Term 1. It was important and had a large effect in the previous model before the end of Term 1. But, it was not important after Term 1 when also considering T1 Cumulative GPA and Credit Completion Rate.

Interaction Effects Comparison (Table 8)

Predictor	New	Previous
Belonging/Female/Closest	6	11
Belonging/Female/Not Closest	12	18
Belonging/ Male/Closest	4	6
Belonging/Male/Not Closest	9	13
QPP/SOC	1	4
QPP/Non-SOC	3	8
ACE/STEM	14	20
ACE	4	5
STEM	5	4

Slashes were used to denote different categories. So, for someone who is Female and Closest to SCSU, increasing Belonging by one point has a 6 percentage point increase for predicting retention.

Again, these effect sizes are smaller and the interactions seem to be less pronounced in the new model due to adding GPA and Credit Completion Rate. However, the same interactions are still present.

Belonging Index was still a very good predictor of retention on its own as well as with its interactions. We can see that its interactions with both Closest and Gender are still present. The effect size of Belonging is larger for Female than Male. It is also smaller for Closest than Not Closest.

QPP is still a good predictor, but it involves a student's academic performance similar to GPA and Credit Completion Rate. However, it looks at past performance which is still useful in predicting future performance. The same interaction with SOC is still present, but it is less important.

STEM is a useful predictor with a moderate effect size. The interaction term with ACE was useful again since there was a definite difference between students who were STEM and ACE and students who were one or the other. It's much larger than you would expect if there was no interaction. However, this interaction with ACE is questionable because of the small sample size. After Term 1, there were only 76 students who were STEM and ACE.

ACE increases retention. These students are receiving extra help compared to students who aren't at-risk. The confounding variable in this case is that ACE students have lower QPP's or GPA and are overall going to have lower retention rates.

Prediction for 2017 Data

Prediction Percentages (Table 9)

Year	Before Term 1	After Term 1
2014-2015 (Actual)	71.5%	79.1%
2017 (Prediction)	74.1%	79.9%

After eliminating the students who dropped out after Term 1, there were 567 students left. Applying the new model to this data gives an average predicted probability of 79.9%.

The predicted probabilities of enrollment for all students of the 2017 cohort were saved and sent to Dr. Robinson. Now they can be grouped according to their predicted probability of enrollment, and targeted for help based on who is most at risk.

Further Studies

An interesting area of study for this project could be in the timing of the dropouts mentioned earlier. Students could be targeted for help when they are most at risk based on their demographic information whether it be before or after completion of their first semester.

Also, how the performance of these models on the 2017 cohort will be evident this fall when their third term enrollment data will be available. Changes can be made to the model, and the school's efforts at increasing retention can be improved.

Appendix

Type 1 Error and Type 2 Error

There are two types of errors that are going to be discussed here. These are type 1 errors and type 2 errors. These errors can be affected by several different factors including alpha levels, sample size, and number of tests run.

A type 1 error occurs when we reject the null hypothesis when we should not reject it. That is, the p-value accepted for our t-test is due to random error and the null hypothesis is true. Typically an alpha level .05 is the accepted value for statistical significance. Any value less than that is seen as statistically significant, and any value larger is typically seen as not statistically significant. This is really an arbitrary value that can be changed to be higher or lower as you desire. Raising the p-values that are accepted as significant will increase type 1 errors. Lowering the p-values that are accepted will decrease type 1 errors.

A type 2 error occurs when the null hypothesis is accepted when it should have been rejected. This can be related to how many type 1 errors are occurring. As the alpha level increases, a type 1 error becomes more likely and a type 2 error becomes less likely. As it decreases, a type 1 error becomes less likely and a type 2 error becomes more likely.

The size of the sample can also affect whether or not there is an error. In a t-test, a z-score for the predictor's usefulness in the model is calculated by its estimate divided by its standard deviation. The standard deviation of the predictor is found using the size of the sample as part of the formula. The larger the sample size, the smaller the standard deviation will be. So, if there is a very small sample size, the standard deviation will likely be large. With a large standard deviation, the z-score will be small. With a small z-score, the predictor will have a large p-value and a type 1 error may occur. If there is a very large sample, the alpha level can be set to be low, and type 2 errors can be decreased. Increasing sample size is one way to decrease the number of errors that occur.

The number of tests run can also affect whether or not errors occur. For example, if there is an alpha level of .05, running one test for significance will have a .05 probability of a type 1 error occurring if the null hypothesis is true. However, running 20 tests with .05 probability of a type 1 error occurring increases the chances that at least one of these tests will have a type 1 error. A test with alpha .05 has a .95 probability that a type 1 error will not occur. The probability that none of these tests will have a type 1 error is $(.95)^{20} = .3585$. That is, there is an approximately 64% chance that at least one of these tests will have a type one error! Running multiple tests trying to achieve the results that you want isn't a statistically sound practice that can lead to false positives in the results.

Power

Statistical power is the likelihood that an effect of a given size will be detected, given that there is an effect to detect. Power can be increased by requiring a larger effect size to be important, decreasing alpha level, or increasing sample size.

Power is equal to one minus the probability of a Type 2 error. To calculate power in a logistic regression setting, it is necessary to determine what effect size is deemed to be important. Suppose that an effect size of 5% is determined to be practically significant. Then the difference of the logit score of the predictor and the logit score of the null hypothesis needs to be calculated. This is divided by the standard error to obtain a z-score of the difference between the two. Finally, the power is determined by taking that z-score minus the z-score of the alpha level, and finding that value in a normal distribution with mean zero and standard deviation one gives you the power level.

Power can be used before analysis as well as after. Before running a statistical test, power can tell you what sample size needs to be collected to have a certain level of power. For example, to detect an effect size of 5% a sample size of n needs to be collected if a power of 75% is desired. Calculating power after analysis is done can indicate which predictors have high p-values because they are not useful in the model or if they are high because the power of the test is weak.