First Year Retention Analysis

Summer Fall Full-Time, First Time in College Cohorts 2016-17 through 2018-19

Esther Wilkinson, Institutional Research Analyst II

Edited by Andre Watts, Interim Director of Institutional Analytics

A cohort analysis of First Year Retention (students who returned for their second Fall) and the identification of key indicators that may predict retention.

Contents

BACKGROUND 3
KEY FINDINGS 3
METHODOLOGY 4
ANALYSIS RESULTS 7
Fall Model 9
Fall/Spring Model 9
Factor: Summer Enrollment Prior to Second Fall 10
Factor: Any Probation Status End of 1st Fall 10
Factor: Fall UCF GPA 11
Factor: Number of W Grades in the Fall 11
Factor: Number of F Grades in the Fall 12
Factor: Challenge Course Combinations in First Year 12
Factor: Participation in the LINK Program 13
Factor: Fall Online Credits Taken 13
Factor: Spring Term GPA 14
Factor: Number of DFW Grades in Spring Term 14
Factor: Number of Hours Earned in the Fall 15
Factor: Days between Matriculation and Fall Start 16
DISCUSSION 17
APPENDIX 18
Table 1. First Year Retention of First-Time-In-College, Summer-Fall-Full-Time 18
Table 2. Fall Model 18
Table 3. Fall/Spring Model 18
Table 4. Challenge Courses 19
Table 5. Correlations of Select Variables 20
Table 6. Variables in Dataset 21
References 24

BACKGROUND

First Year Retention is a nationally known success metric in higher education that institutions use in reporting to the public and to stakeholders. The retention metric is the calculated rate of first year, full-time students that are retained into their second fall semester. This metric applies to UCF through the Preeminent Research University Metrics as the "Freshman Retention Rate" defined as, *metric (c) A freshman retention rate of 90 percent or higher for full-time, first-time-in-college students, as reported annually to the Integrated Postsecondary Education Data System (IPEDS),*(Florida Legislature, 2019*)*. As the Legislature and Board of Governors emphasize retention as a success marker for higher education institutions, we can understand why here at UCF a number of student success initiatives have made retention a focus. At UCF, these initiatives have helped to steadily increase the retention rate over the last ten years to its current high of over 90%. In order to maintain and keep improving retention, it is essential that we understand the factors that contribute to retention when there exists constraints on university time and resources. Information backed by current analysis and prior literature can guide decision-making and institutional actions which produce high yield.

This study takes the historical data of three FTIC Summer Fall Full-Time cohorts 2016-17 to 2018-19, and through modeling, finds the most important features that contribute to retention. Sub-setting the data created two models that are not distinct, but taken from the perspective of different points in time in the first academic year. The fall dataset with 19,866 students have variables that primarily belong to fall academic performance and fall activities, but also includes incoming data such as demographics and test scores. The Fall/Spring dataset has fewer students due to end of fall attrition, with 19,263 students and data pertaining to fall and spring academic performance and activities, as well as incoming data.

KEY FINDINGS

Positive Impacts

1. The variable with the largest positive effect on retention was enrolling in the summer term prior to the second fall (both Fall and Fall/Spring models); 97.5% of those enrolled in summer prior to second fall were retained.
2. There was a positive effect earning a UCF GPA in the fall above 2.60 (Fall model).
3. Participating in the LINK[1] program had a positive effect on retention; students who participated were 27% more likely to be retained than students who did not participate.
4. Time between the matriculation date[2] and the start of the first fall semester had a positive impact. Students who committed before May prior to fall start were more likely to be retained and had average retention rates over 90% (Fall and Fall/Spring models).

[1] Learning and Interacting with New Knights (LINK) is an education and involvement-based program to help students new to UCF get involved on campus, https://link.ucf.edu/.

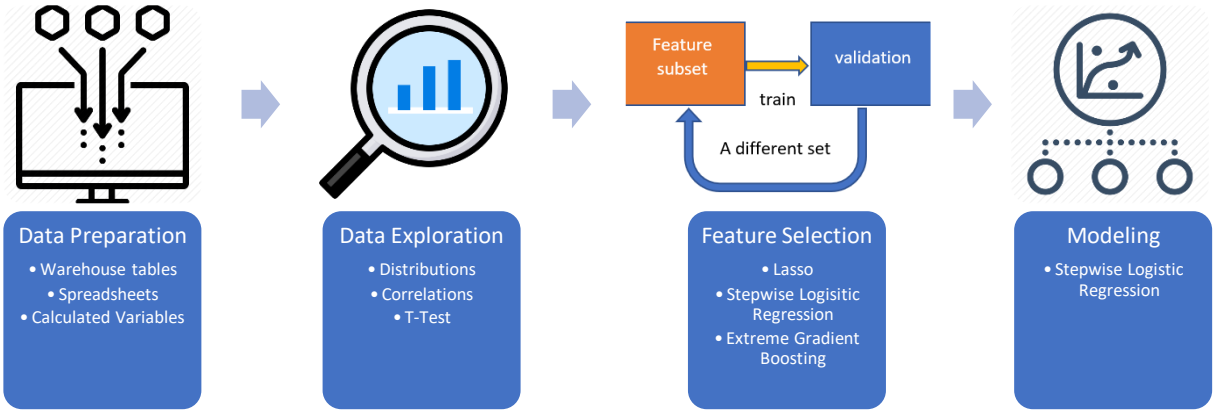
[2] The matriculation date is defined as the day the student makes their first financial deposit to confirm their plan to attend UCF.

5. Taking challenge courses[3] had a positive effect on retention (both Fall and Fall/Spring models).
6. There was a positive effect earning a term GPA in the spring above 2.40 (Fall/Spring model).
7. There was a positive effect in the number of credit hours earned in the fall, at or above 12 credit hours (Fall/Spring model).
8. Attendance at the Recreation and Wellness Center, RWC, had a small but positive effect on the retention models (both Fall and Fall/Spring models).

Negative Impacts

1. The largest negative effect on retention was having any probation status at the end of fall (Fall model). A student was 42% less likely to be retained if they had any probation status for fall.
2. There was a negative effect earning one or more W or F grade in the fall (Fall model).
3. Out-of-state status had a negative effect on retention (Fall model); 14% of out-of-state students were not retained. The estimated odds of out-of-state students being retained was 43% less likely than a student from the state of Florida.
4. There was a negative effect earning more than one D, F, or W grade in the spring term (Fall/Spring model).
5. Major changes had a negative effect on retention (Fall/Spring model).
6. Increasing the number of online courses to 6 credit hours in either the fall or spring had a negative effect on retention (both Fall and Fall/Spring models).

METHODOLOGY



Data Preparation

Students from the most recent three FTIC Summer Fall Full-Time cohorts, 2016-17 through 2018-19 were chosen to be used in this analysis. Each of the three cohorts were part of a new retention initiative starting in the academic year 2016-17 where more than half of each cohort received extra contact

[3] Credit bearing undergraduate courses based on the 2017-18 General Education Program catalog year, with at least 20% DFW rate historically, and have at least 50 students in them. See appendix, Table 4.

through retention intervention efforts by members and departments of the Retention Intervention Team. A total of 151 variables and 19,886 records were collected for the analysis and were compiled using data from the Institutional Knowledge Management warehouse tables, PeopleSoft, and data from various external departments across campus who collect student data for those students they serve (e.g. mentoring programs, volunteer UCF, intramurals, and LINK participants). The dataset was subset for different models based on the semester data i.e., Fall dataset contained only fall academic and engagement activity, demographics and financial characteristics. A Fall/Spring dataset contained the variables from the fall dataset with the addition of spring activity. Due to attrition at the end of the fall semester, the Fall/Spring dataset had 623 fewer records for a total of 19,263 records. In all, only a few variables had missing data such as *family income* and *unmet need* since this data was collected only from students who filled out a FAFSA, for a total of 2,137 missing records. The variable *distance to UCF* (distance calculated from high school zip code to UCF zip code), had 209 missing records due to a missing zip code or international students without zip codes; and 9 missing records for the variable *high school gpa,* were imputed with the average gpa. Analyses were run two ways with consideration of these variables, once with removing records for each variable with missing financial or zip code data, and again with the omission of these variables. All statistical tests for this study were performed using R version 3.6.2.

Data Exploration

The overall retention rate in the data for all cohort years (2016-17 – 2018-19) was 90.5% with 18,002 students retained verses 1,884 not retained from the total 19,886 records.

Imbalanced data: There is considerable debate among scientists over the treatment of imbalanced data, which in this study is true as we have 90.5% of students in one class (retained), and the other 9.5% in the other class (not retained). Most statistical algorithms work best when the number of samples in each class are about equal to maximize accuracy. Treatments such as oversampling (adding more copies of the minority class), undersampling (removing some observations of the majority class) or using the Synthetic Minority Oversampling Technique (SMOTE) would have considerable drawbacks compromising the data by adding noise, or losing valuable information. In recent studies, Ensemble methods such as bagging or boosting are becoming popular solutions for imbalance problems (Haixiang et al., 2017, p. 226). Decisively, no sampling treatment to balance the data was implemented in this study.

Dealing with highly correlated variables: The features of the data contained calculated fields to provide many ways to look at what might be significant, it also contained several measures of gpa (UCF cumulative, UCF term, and cumulative all). For example, a field for each semester for each count of D grades, F grades or W grades was also calculated to incorporate a sum total of DFW grades for the semester. Also, various separate student group participation variables i.e., LEAD scholars or Athlete were also calculated for a field that served as a flag for participation in “any” group. Other calculated fields served as flags for “any” probation per semester, or for the year. With the assumption that calculated fields or different combinations of these fields would be highly correlated and perhaps multicollinear, correlation tests were performed for each data subset. Pearson correlation tests were used as well as point-biserial correlation (also a Pearson correlation) which is a special case of the product-moment correlation in which one variable is continuous and the other variable is dichotomous (Kennedy, 2020). Analysis of Variable Inflation Factors (VIF) were used to test for multicollinearity.

Feature Selection

To narrow down the number of variables and to keep only those that had the largest effect on being retained, three methods were employed as a feature selection tool, Lasso, Stepwise Logistic Regression, and Extreme Gradient Boosting (XGBoost). The Lasso method similarly like forward or backward stepwise selection will select models that include just a subset of the available variables. The Lasso with its constraint function, penalizes the coefficient estimates and shrinks them towards zero so that only variables that produce the greatest impact on the outcome variable will remain in the model referred to as sparse models (James, Witten, Hastie, & Tibshirani, 2013, p. 220). Extreme Gradient Boosting is a tree boosting algorithm that is a highly effective and widely used machine learning method. It is a preferred and ten times faster algorithm over other existing gradient boosting algorithms among data scientists (Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y., 2015). Gradient boosting has the advantages of high predictive accuracy, and it works with both categorical variables (treated with one-hot encoding creating dummy variables) and numerical values without scaling. It also handles missing data. Three data subsets were subjected to these algorithms, Fall dataset with 19,886 records and 41 variables, Fall/Spring subset with 19,263 records and 81 variables, and a Fall/Spring/Summer2(prior to second fall) subset with 9,736 records and 88 variables. The datasets were split 50/50 for train and test sets. The Fall/Spring/Summer2 data did not provide additional insight at this stage which was primarily contributed to the fact that 97.5% of the students enrolled in summer2 were also retained to the fall.

Modeling

Considering attrition that occurs at the end of the first fall semester, along with the abundance of communication students received in the spring and summer through retention initiatives, a two model approach was used for this study. The Fall and the Fall/Spring dataset would each provide their own insights at different points in time of the first academic year. Two methods were used for the training and testing data; the first as a 50/50 split, and the second method used the 2016-17 and 2017-18 cohorts as training sets to test against the 2018-19 cohort. A repeat of Stepwise Logistic Regression was performed for the final model on variables with the most impact on retention resulting from the feature selection methods. Accuracy of models were measured by the Akaike information criterion (AIC), area under the Receiver Operating Characteristic (ROC) curve (AUC), and confusion matrices. Confusion matrices provide an easy to interpret count of how many were predicted correctly but the overall accuracy can be misleading on imbalanced data. For this reason, the specificity and sensitivity of the confusion matrix are the best performance metrics (Luque, Carrasco, Martín, & de las Heras, 2019, p. 226). In this study tests for accuracy showed high accuracy rates (specificity) over 98% predicting students that were retained, and roughly 38% accuracy (sensitivity) in predicting students that were not retained.

ANALYSIS RESULTS

Feature Selection Results

During the feature selection process, each method produced a list of 5-12 variables that were picked by the algorithms as having the most impact on retention. Variables known to UCF as potentially impacting retention from prior analyses were included in the data but were not selected by any of the three algorithms as having a large or significant effect. Among those variables were ethnicity, first generation, Pell eligible, and living in UCF housing to name a few. Lasso, XGBoost and Stepwise Logistic methods had slight variances between them in the variables that surfaced as important, with some variables surfacing in all three or two methods. Figure 1 compares the variable selection by each method for the Fall dataset, and comparisons for Fall/Spring dataset are shown in Figure 2.

Fall Data

Rank	Lasso	XGBoost	Logistic Stepwise
1	FRST FALL ANY PROB	Enrolled Summ2	Enrolled Summ2
2	FALL W GRADES	FRST FALL CUR GPA	FRST FALL ANY PROB
3	FALL SUM DFW	ANY PROB	FRST FALL CUR GPA
4	FRST FALL UCF GPA	FRST FALL UCF GPA	FALL W GRADES
5	FRST FALL CUR GPA	FRST FALL TOT HRS ERN	Any Support Group
6	FALL F GRADES	Y1 Challenge Crs COMBO	LINK Participation
7	Out of State	FALL SUM DFW	FALL F GRADES
8	FRST FALL TOT HRS ERN	MatricDays Prior Fall	Y1 Challenge Crs COMBO
9	Y1 Challenge Crs COMBO	Distance to UCF	Fall Online CRDS
10	LINK Participation	LINK Participation	FRST FALL RWC
11	FRST FALL RWC	AY UnmetNeed	Distance to UCF

Figure 1

Fall/Spring Data

Rank	Lasso	XGBoost	Logistic Stepwise
1	SP SUM DFW	FRST SP UCF GPA	Enrolled Summ2
2	FRST SP CUR GPA	TOT HRS ERND YEAR1	FRST FALL ANY PROB1
3	TOT HRS ERND YEAR1	FRST SP CUR GPA	Out of State
4	FRST SP TOT HRS ERN	Enrolled Summ2	SP SUM DFW
5	SP ONLINE CRDS	Y1 Challenge Crs COMBO	Y1 Challenge Crs COMBO
6		FRST FALL UCF GPA	FRST SP TOT HRS ERN
7		FRST SP TOT HRS ERN	FRST SP CUR GPA
8		Distance to UCF	MajorChanges

Figure 2

Model Results

Reworking the top variables selected by the three methods, and checking correlations and multicollinearity, the final logistic regression results can be bucketed into two main categories, academic performance and engagement (Figure3 and Figure 4). Based on the results, we could infer that students who were enrolled in the summer prior to the second fall, which was the variable with the most impact on both models, may be more actively engaged including making good progress through their academics. Noted earlier, these students had a retention rate of 97.5%. It was also observed that 63% lived in UCF housing during the spring, 35% lived in housing off campus/not affiliated, and the remaining 2% lived in affiliated housing during the spring (housing status in the summer was unknown). However, there were little differences among housing locations in the retention rates of students enrolled in the summer (97.5%, 97.6%, and 97.4% respectively).

Engagement on campus was observed as a significant positive impact on retention with participation in the LINK program or the Recreation and Wellness Center (RWC). Students participating in the LINK program (59% of the population, n=11,786) were 27% more likely to be retained than non-participants. The retention rate for participants averaged 92% compared to 88.4% of non-participants (n=8,100). Students who were engaged at the RWC also saw a positive impact to retention. Attendance in activities at the RWC had a small but significant impact on retention model, and differences in retention rates of participant's verses non-participants were large. More than half of the total population (54%, n=10,737) who attended the RWC in both fall and spring at least two times had a retention rate of 94.4%. Conversely, students who did not attend the RWC more than once in both fall and spring had a retention rate of 86.5%. It was also observed that students participating in either the RWC or the LINK program had higher average Fall UCF GPA's than non-participants (participants of RWC or LINK were 3.27 and 3.25 respectively; non-participants of RWC or LINK were 3.14 and 3.13 respectively).

Students who ended the fall semester in a probation status, either placed on probation, continuing on probation or academically disqualified had the largest negative impact on retention. This was most significant in the Fall model. The results showed that academic performance is the most important factor overall in retention for this population which also related to the number of DFW grades negatively affecting retention. We saw that GPA was important and positively impacted retention, and too many online courses negatively affected retention. Retention rates increased 5% when students take two challenge course combinations verses zero combinations (n=6,930 and 91% retention compared to n=6,496 and 86% retention respectively), and rates increased another 4% to 95% retained when taking 4 challenge course combinations. These results are shown in Figure 10.

Fall Model *(N=19,886)*

Significant Factors for Retention of Summer Fall Full-Time FTIC

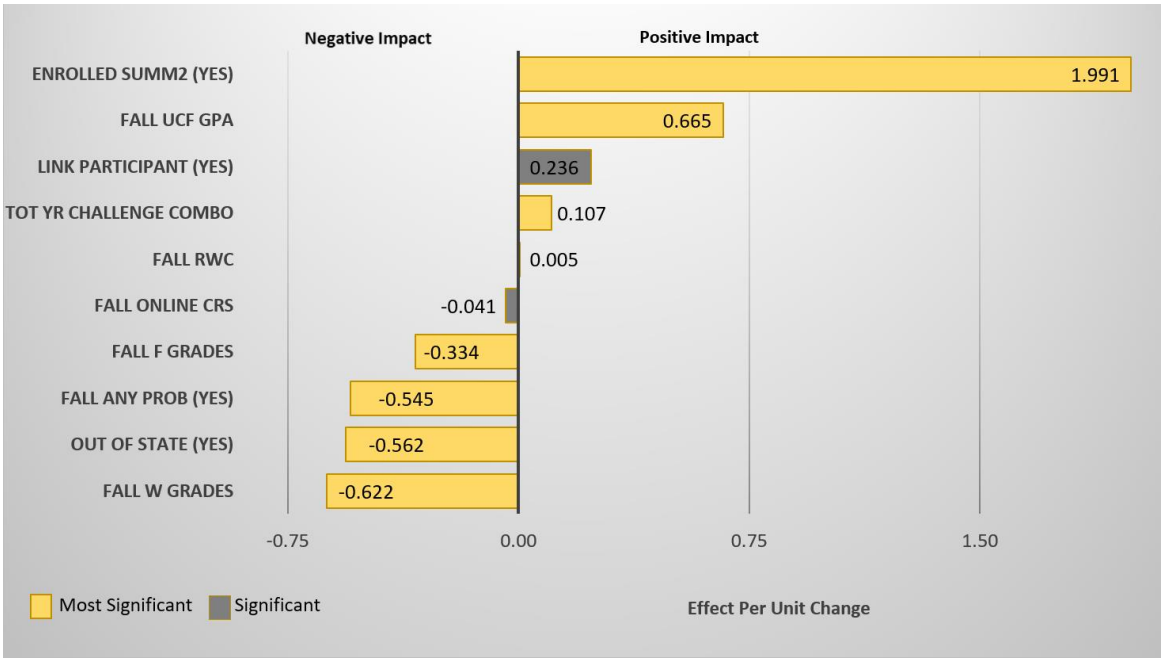


Figure 3

Fall/Spring Model *(N=18,740)*

Significant Factors for Retention of Summer Fall Full-Time FTIC

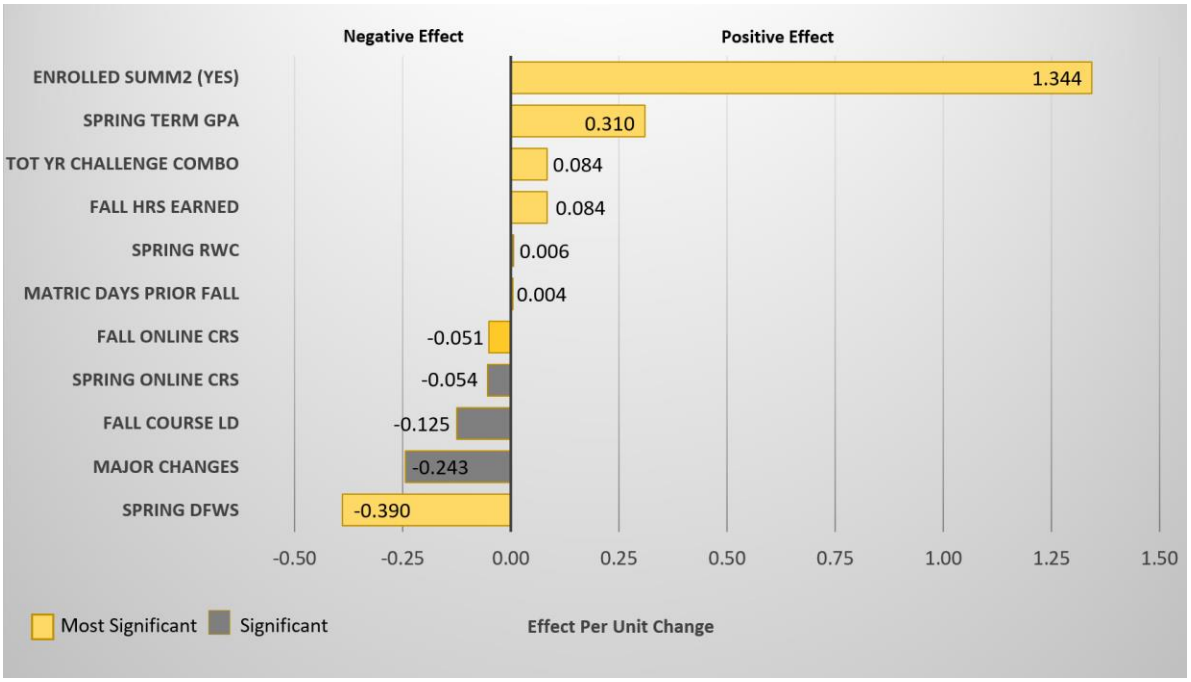


Figure 4

Factor: Summer Enrollment Prior to Second Fall

The estimated odds that a student was retained and who also enrolled in the summer prior to second fall (summer2), was 3.8 times greater than a student not enrolling in summer2.

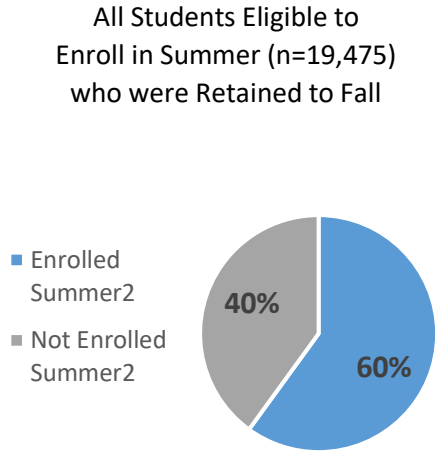


Figure 5

Factor: Any Probation Status End of 1st Fall

The results showed that students who had any probation status at the end of fall were 42% less likely to be retained than students with no probation status. A total of 6.7% (n=1,330) of the entire population in the data had a probation status at the end of fall.

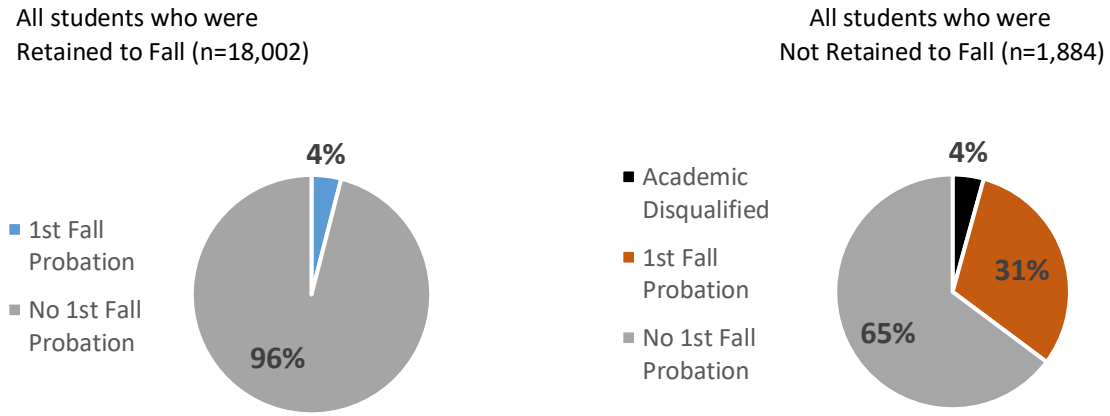


Figure 6

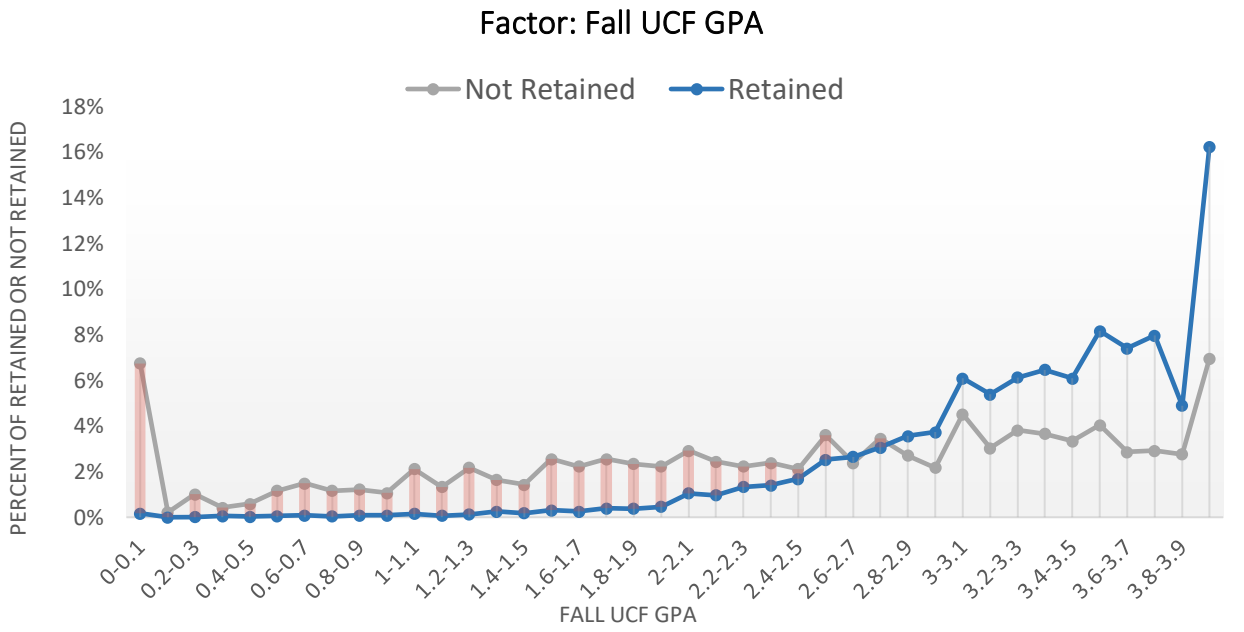


Figure 7

There was a positive impact to retention when Fall UCF GPA is above 2.60. The region below 2.60 represents a larger proportion of the population not retained (51%) compared to the proportion of the population retained (12%). The estimated odds that a student was retained is 94% greater for each unit increase in Fall UCF GPA.

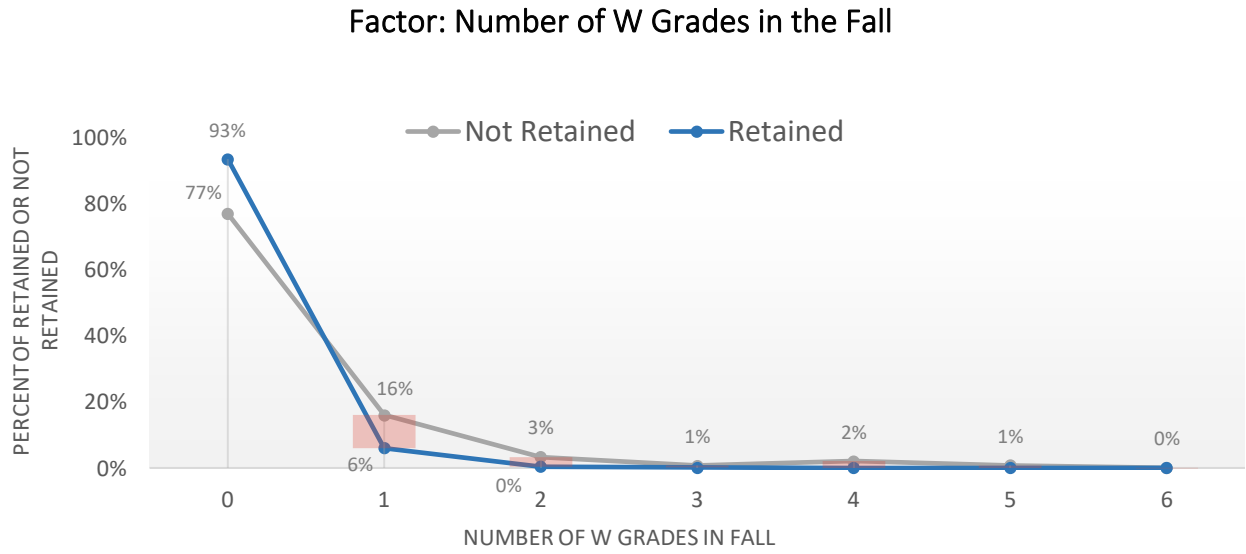


Figure 8

The estimated odds for a student who had one W grade in the fall is 46% less likely to be retained than a student who had zero W grades in the fall. Of the proportion of students retained, 93% (n=16,820) compared to 77% (n=1,449) of the students not retained earned zero W grades in the fall.

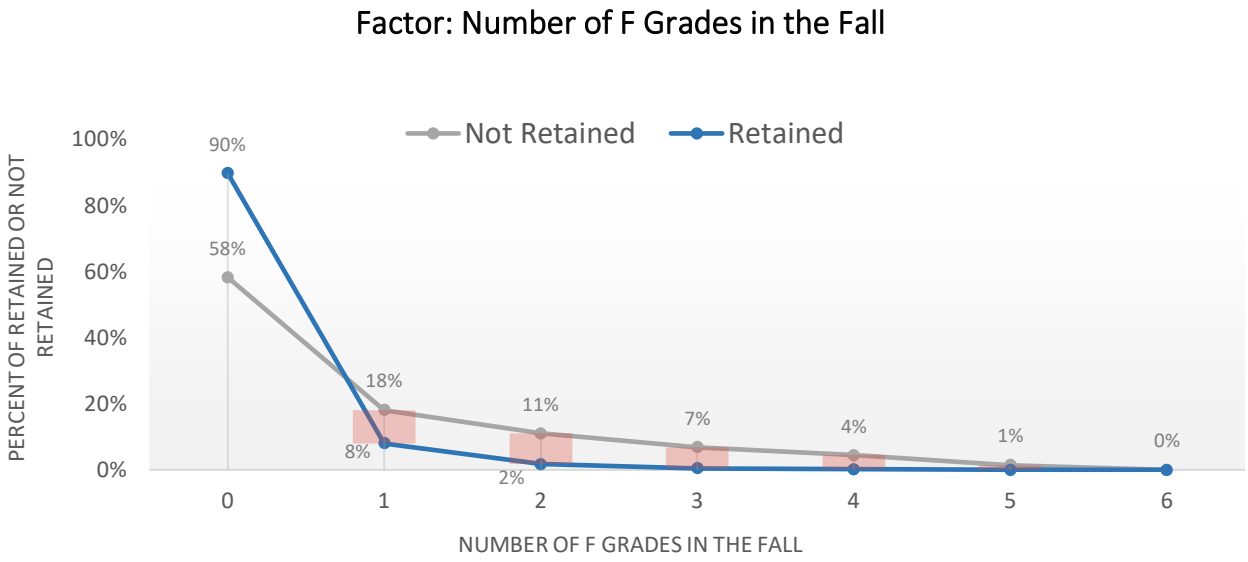


Figure 9

The estimated odds for a student who had one F grade in the fall is 28% less likely to be retained than a student who had zero F grades in the fall. Of the proportion of students retained, 90% (n=16,143) compared to 58% (n=1,097) of the students not retained earned zero F grades in the fall.

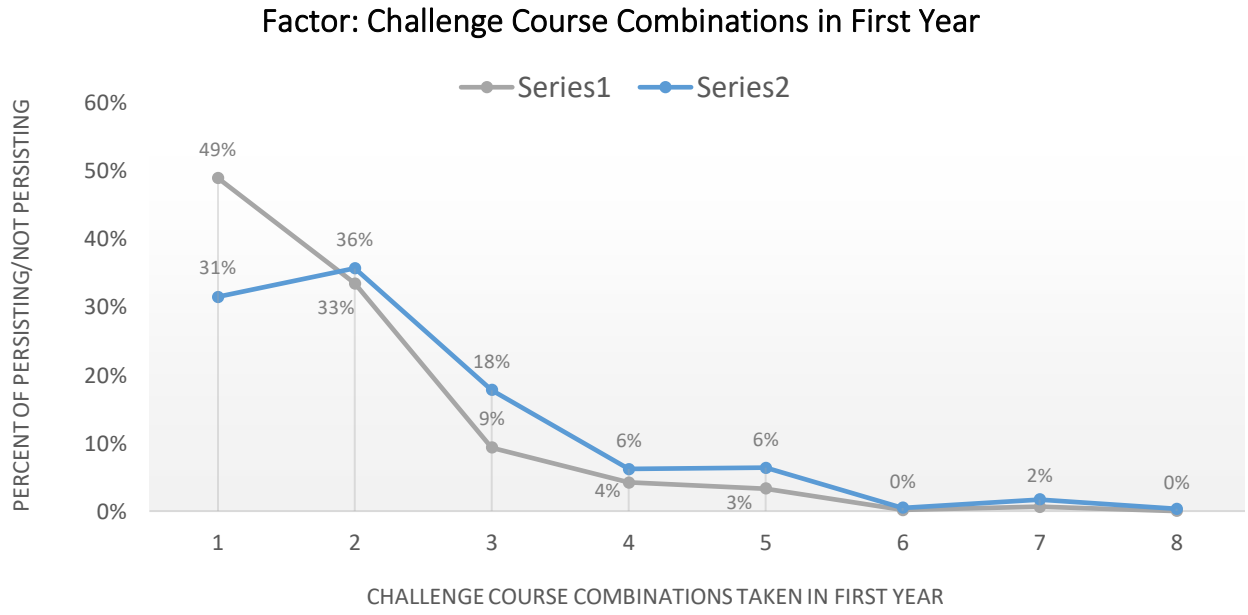


Figure 10

Challenge course combinations had a positive impact on retention. This factor was also correlated to being in a STEM program (*r*= 0.36).

Factor: Participation in the LINK Program

There were 11,786 students (59% of the population) participating in the LINK program. The retention rate for the students who participated was 92% versus 88.4% for the non-participants. Program participants were 27% more likely to be retained.

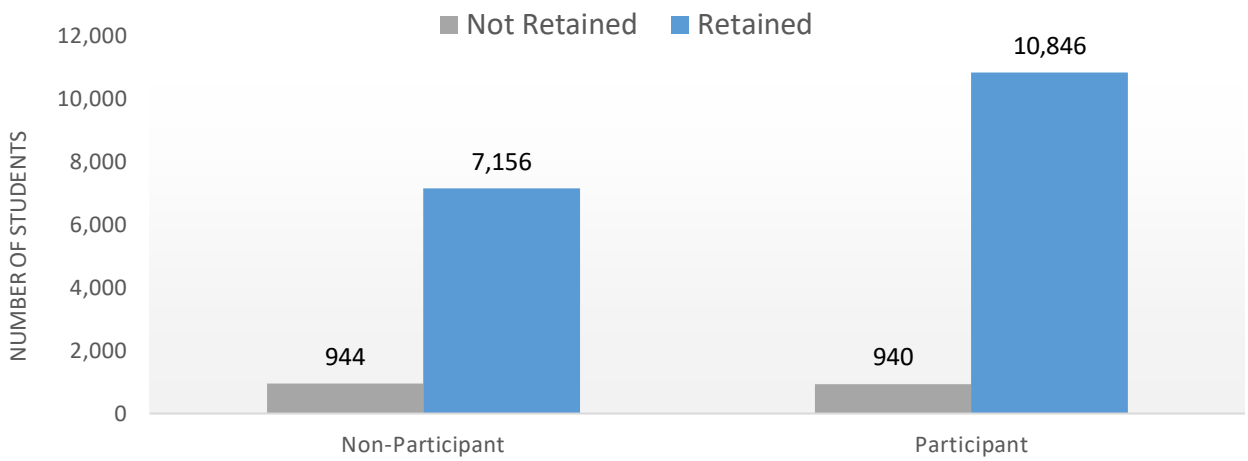


Figure 11

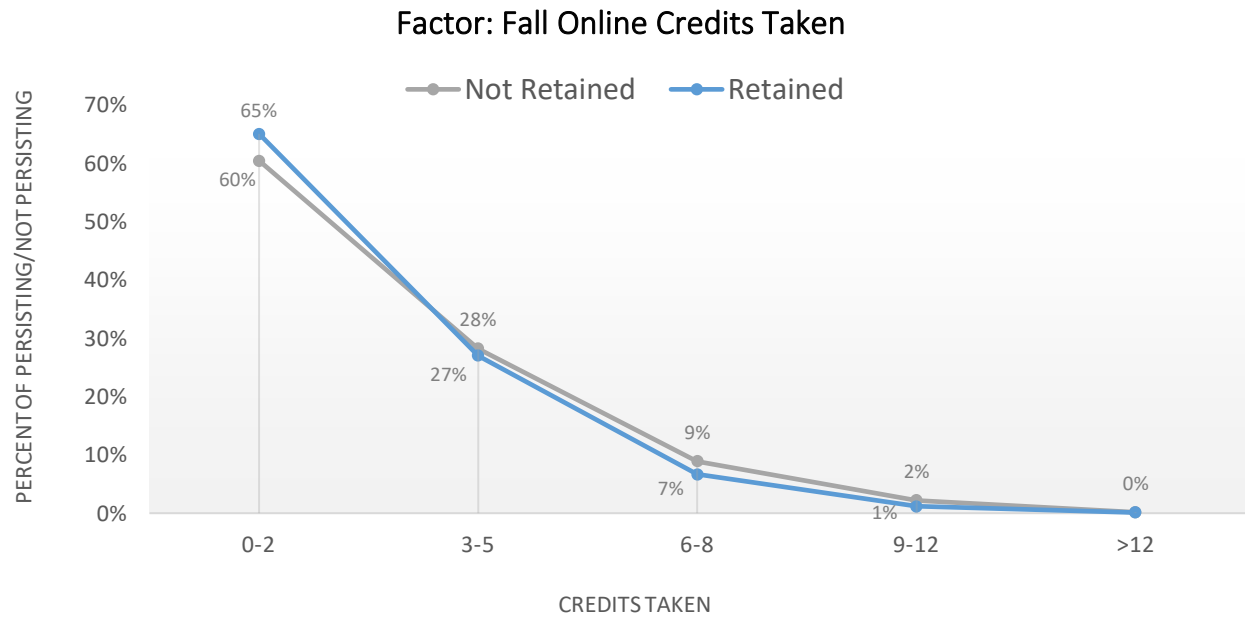


Figure 12

Taking online credits in the fall had a small but statistically significant negative impact on retention. A larger percentage of students who were not retained took online courses in the fall compared to students who were retained.

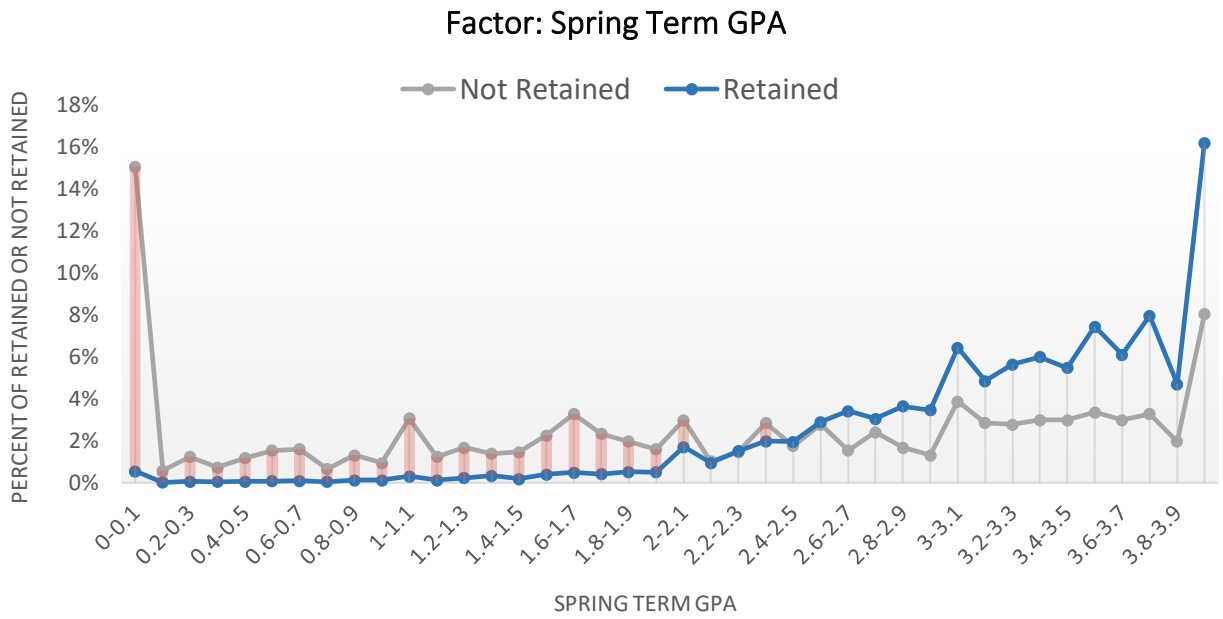


Figure 13

There was a positive impact to retention when Spring Term GPA is above 2.40. The region below 2.40 represents a larger proportion of the population not retained (53%) compared to the proportion of the population retained (11%).

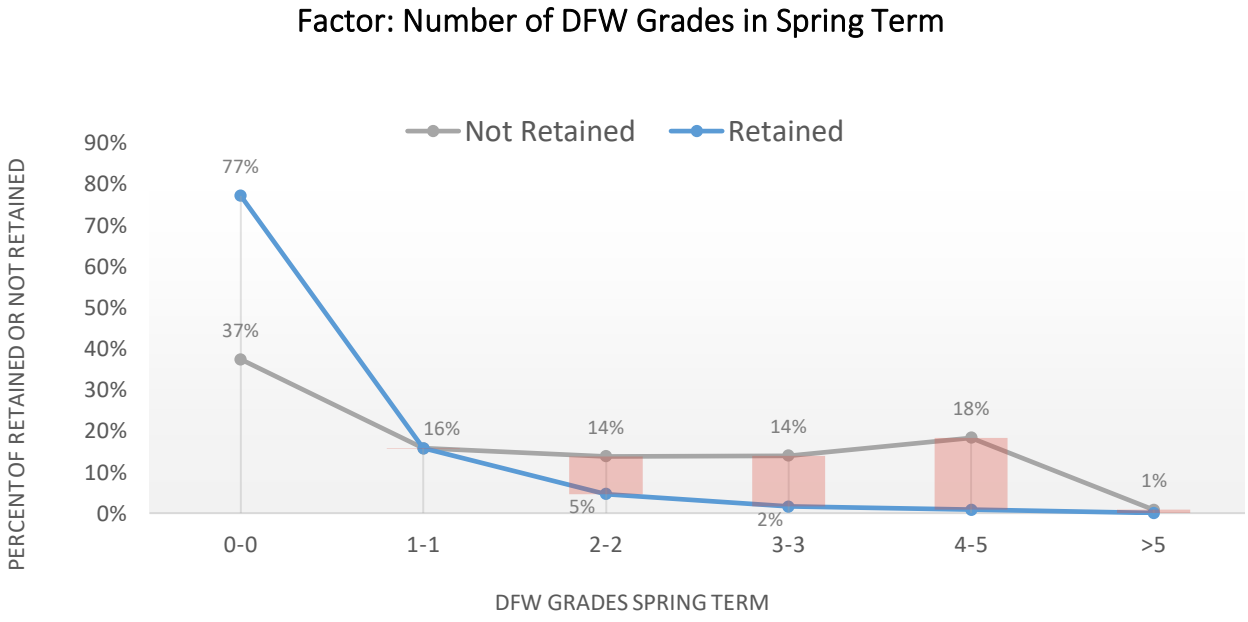


Figure 14

Of the proportion of students enrolled in spring and were retained 77% (n=13,796) compared to 37% (n=512) of the students not retained earned zero DFW grades in the spring. The estimated odds of being retained was 32% less likely for a student who earned one D, F or W in the spring than for a student who earned zero DFW grades in the spring.

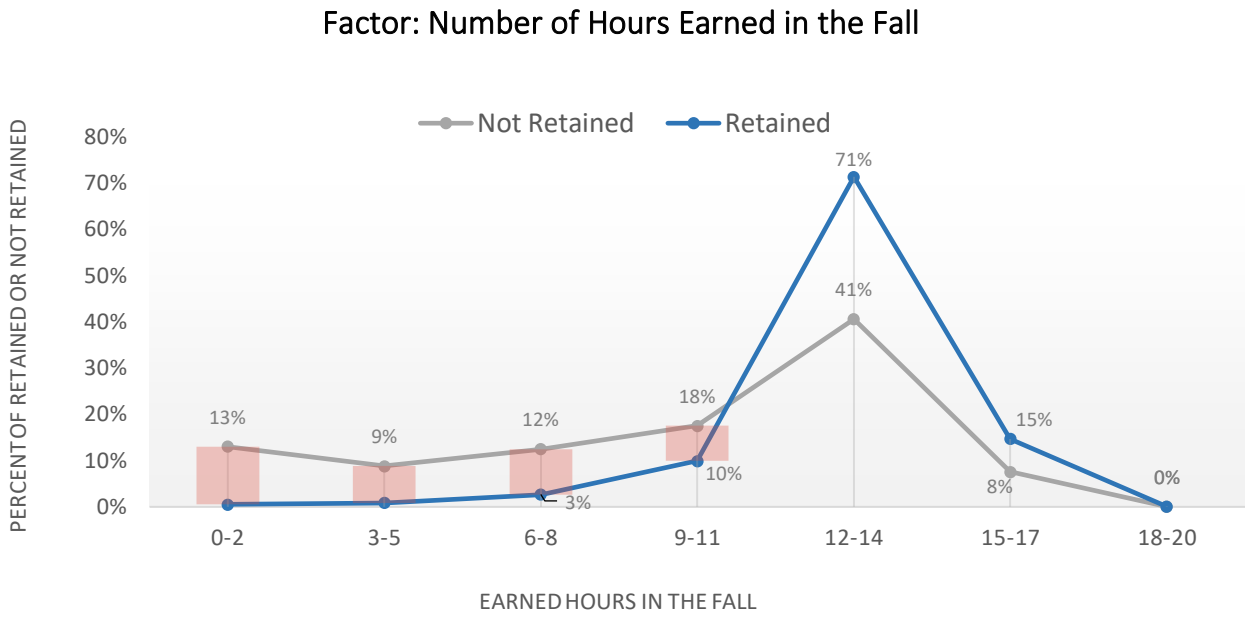


Figure 15

Students who were not retained had a higher percentage of their population earning less than 12 credit hours in the fall term. The region below 12 credit hours represents a proportion of the population not retained (52%) compared to the proportion of the population retained (14%).

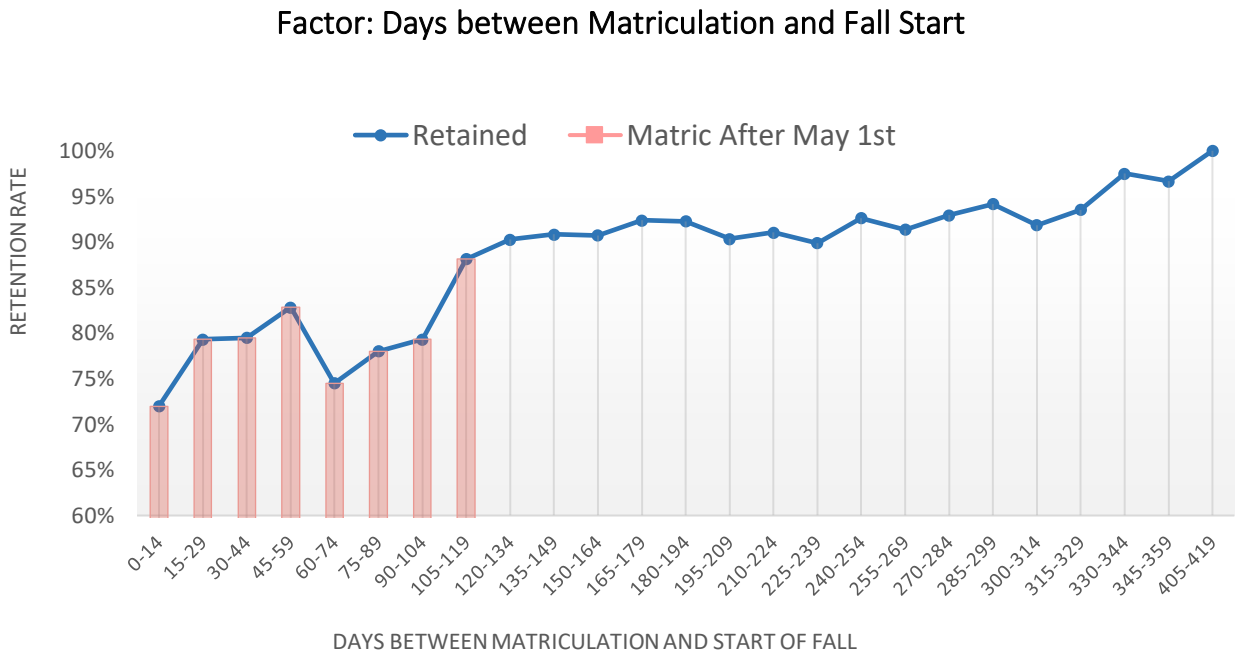


Figure 16

Retention rates increase as the number of days between a student's matriculation date and the start of fall increases. Students who commit to UCF after May 1, or approximately within 115 days to start of fall (n=2,185 in the data) had an average retention rate of 84.9%.

Days after May 1st and before Fall start	Not Retained	Retained	Total
0-14	7	18	25
15-29	6	23	29
30-44	8	31	39
45-59	11	53	64
60-74	26	76	102
75-89	31	110	141
90-104	61	234	295
105-119	295	2,191	2,486
>120	1,439	15,266	16,705
Total	**1,884**	**18,002**	**19,886**

DISCUSSION

The results of this study have shown that academic performance along with progression and engagement, are the most important factors in retention. The lack of qualitative data in this study, hearing the student's perspective, or the perspective of their advisors, plus access to data on student's working hours in jobs off campus or additional financial and family concerns is a limitation to this study. We observed traditionally, factors such as HS GPA, gender, ethnicity or first generation status have differences in retention rates but were not significant in these models, and efforts focusing on academic performance could potentially increase retention outcomes on a larger scale. The underlying reason for differences in retention rates among select populations, some which are less than 30% of the total FTIC population, has more to do with their performance and progression and not because of "who" they are. For example, we saw in the data that first generation students (17% of the total FTIC population in the dataset) had a 89.6% retention rate compared to 90.7% of those not first generation. Those retention rates have increased for first generation students with each cohort in the data, but so have the rates for non-first generation students. Relatively close retention rates of first generation to non-first generation students are also a factor of the various services and attention given to first generation students to promote their success. Would those retention rates be different if no additional focus was spent on this population? One would think so, but keeping things in perspective of the big picture, all students as a whole contributes to the health of UCF and so it's important to know where staff and faculty can make the biggest difference in our student's success.

Other limitations were noted such as missing financial data, which is only captured by UCF if the student had filled out a FAFSA, or whether a student had lost Bright Futures before the start of the second year. The variable *Major Changes* had its own set of limitations. Major change was a negative factor in the spring model, however not having a data capture available of every major change within the semester was a significant limitation. Information about majors for a student could only be noted if it was different from semester to semester without a time stamp of when it was changed or how often it was changed. Future research should consider the addition of mandatory surveys of students or advisors which may provide more insight into some of these gray areas. Additionally, further analysis with removing the academic performance variables which dominate the models might indicate on a more granular level other factors that rise to the top. For example, in a quick analysis without the academic performance variables, it was noticed that students in the data who did not have a Bright Futures scholarship had a negative effect on retention. This makes senses as we saw retention rates for students without Bright Futures was 87.4% (n=7,204) compared to 93% (Academic Scholar, n=6,768) or 91.6% (Medallion n= 5,914).

APPENDIX

Table 1. First Year Retention of First-Time-In-College, Summer-Fall-Full-Time

Summer-Fall-Full-Time (SFFT) First-Time-In-College (FTIC)	Retained	Percent Retained	Not Retained	Percent Not Retained	Total Cohort
2016-2017	5,503	89.6%	641	10.4%	6,144
2017-2018	6,042	90.4%	643	9.6%	6,685
2018-2019	6,457	91.5%	600	8.5%	7,057
Total	**18,002**	**90.5%**	**1,884**	**9.5%**	**19,886**

Data retrieved 12/9/2019

Table 2. Fall Model

Effect	Estimate	Std Error	P-Value	Odds Ratio	95% CI
FRST_FALL_UCF_GPA	0.665	0.076	<.001	1.944	(1.677 , 2.254)
Enrolled_Summ2	1.991	0.104	<.001	7.320	(5.970 , 8.974)
FALL_W_GRADES	-0.622	0.071	<.001	0.537	(0.467 , 0.617)
Y1_Challenge_COMBO	0.107	0.018	<.001	1.113	(1.074 , 1.153)
FALL_F_GRADES	-0.334	0.068	<.001	0.716	(0.626 , 0.818)
MatricDays_prior_Fall	0.003	0.001	<.001	1.003	(1.001 , 1.004)
Out_of_State	-0.562	0.130	<.001	0.570	(0.442 , 0.736)
FRST_FALL_ANY_PROB	-0.545	0.163	<.001	0.580	(0.421 , 0.797)
LINK_Participation	0.236	0.082	<.01	1.266	(1.078 , 1.487)
FRST_FALL_RWC	0.005	0.002	<.05	1.005	(1.001 , 1.011)
FALL_ONLINE_CRDS	-0.041	0.018	<.05	0.960	(0.926 , 0.994)

Table 3. Fall/Spring Model

Effect	Estimate	Std Error	P-Value	Odds Ratio	95% CI
SP_SUM_DFW	-0.391	0.067	<.001	0.677	(0.593 , 0.771)
Enrolled_Summ2	1.344	0.110	<.001	3.833	(3.091 , 4.757)
Y1_Challenge_COMBO	0.084	0.020	<.001	1.088	(1.046 , 1.131)
FRST_SP_CUR_GPA	0.310	0.084	<.001	1.364	(1.156 , 1.607)
MatricDays_prior_Fall	0.004	0.001	<.001	1.004	(1.002 , 1.006)
Distance_to_UCF	-0.001	0.0001	<.001	0.999	(0.999 , 0.999)
FRST_FALL_TOT_HRS_ERN	0.084	0.020	<.001	1.088	(1.046 , 1.131)
FRST_FALL_TOT_CRS_LD	-0.125	0.043	<.01	0.883	(0.811 , 0.960)
SP_ONLINE_CRDS	-0.054	0.018	<.01	0.947	(0.915 , 0.981)
FALL_ONLINE_CRDS	-0.050	0.023	<.05	0.951	(0.909 , 0.995)
FRST_SP_RWC	0.006	0.003	<.05	1.006	(1.000 , 1.102)
MajorChanges	-0.243	0.109	<.05	0.784	(0.633 , 0.971)

Challenge Courses

1. UGRD courses only
2. At least 20% DFW rate historically
3. At least 50 students taken the class historically
4. Course is credit-bearing (SCH > 0)
5. GEP courses based on 2017-18 catalog year

Table 4. Challenge Courses

ARH2050
ARH2051
CHM1032
CHM2040
CHM2041
CHM2045
CHS1440
COP3502
COT3100
ECO2013
ECO2023
GLY1030
MAC1105
MAC2311
MGF1106
MGF1107
STA2023

Table 5. Correlations of Select Variables

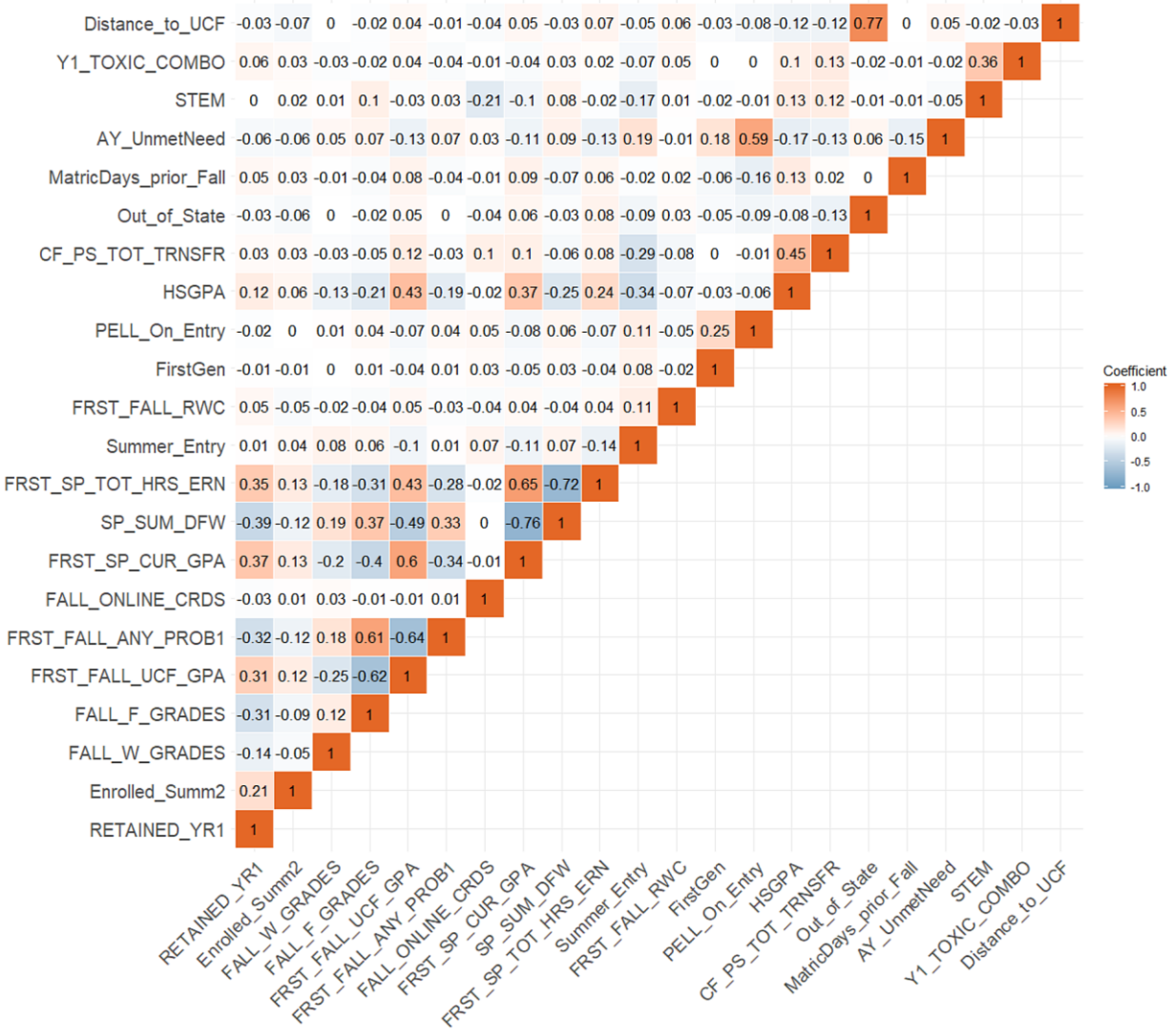


Table 6. Variables in Dataset

No.	Variable in FINAL dataset	Description
1	COHORT_YEAR	FTIC Summer-Fall Full-Time, 2016-17, 2017-18, 2018-19
2	EMPLID	
3	**RETAINED_YR1**	**Dependent Var. Based on cohort returning to UCF for the second fall.**
4	CF_PS_ADMIT_TYPE	8 levels, People Soft admissions type (i.e. UFF, UEA, UWL…)
5	GENDER_REC	0=male, 1=female
6	CF_BOE_RACE_DESCR	9 levels, Race/Ethnicity
7	HSGPA	Any missing (9 in total) were imputed with average 3.96
8	FX_SAT_IR_TOT	SAT calculated by IKM
9	CF_ACT_CTOT_AD	ACT Total score
10	FX_SAT_ACT_FLAG	2 levels, A or S
11	ENTRY_STRM	Entry Terms 1570,1580,1600,1610,1630,1640
12	Entry_CIP	91 levels
13	Entry_CIP_4	67 levels
14	DATE_ENTRY	Entry Term formatted i.e. 201608
15	SUMMER_FALL_FULL_TIME	All are Summer Fall Full Time, 1= SFFT
16	CF_BOE_CNTY_RES	68 levels, County of Residence
17	CF_PS_ACAD_GROUP	13 levels, College at the time of entry
18	CF_PS_ACAD_PLAN_DS	114 levels, Major at the time of entry
19	LAST_INST	58 levels. The last institution the student was enrolled in before UCF
20	LAST_INST_DESCR	Description of last institution the student was enrolled in before UCF
21	AA	Binary. Entered UCF with an AA
22	AS	Binary. Entered UCF with an AS
23	AA_AS	Binary. Entered UCF with an AA/AS flag
24	LAST_ACAD_GROUP	Most recent college student was enrolled in
25	LAST_ACAD_PLN_DESCR	Most recent major student was enrolled in
26	LAST_CIP	94 levels, Most recent 6 digit CIP
27	LAST_CIP_DESCR	95 levels, Description of Most recent 6 digit CIP
28	LAST_ENRL_STRM	Most recent enrolled term
29	LAST_ENRL_TERMID	Most recent enrolled term formatted, i.e. 201608
30	MatricDays_prior_Fall	Number of days between matriculation date and first day of fall semester
31	CF_PS_TOT_TRNSFR	Total number of undergraduate credits the student brings coming into UCF
32	FIRST_TRM_SUMM_STRM	Entry term if a summer term, otherwise blank
33	FIRST_TRM_SUMM_TOT_TRNSFR	Total number of undergraduate credits the student brings entering in the summer term.
34	FIRST_TRM_SUMM_ONLINE_CRDS	Total number of online credits student took in their entering summer term
35	FIRST_TRM_SUMM_TOT_CRS_LD	Total course load for entering summer term
36	FIRST_TRM_SUMM_TOT_HRS_ERN	Total hours earned in entering summer term
37	SUMM1_D_GRADES	Number of D grades in entering summer term
38	SUMM1_F_GRADES	Number of F grades in entering summer term
39	SUMM1_W_GRADES	Number of W grades in entering summer term
40	SUMM1_SUM_DFW	Sum of DFW grades in entering summer term
41	FIRST_TRM_SUMM_UCF_GPA	UCF cumulative gpa-entering summer term
42	FIRST_TRM_SUMM_CUR_GPA	UCF term gpa-entering summer term
43	FIRST_TRM_SUMM_CUM_GPA	Cumulative gpa-entering summer term
44	FIRST_FALL_STRM	Fall term
45	FRST_FALL_TOT_TRNSFR	Total number of undergraduate credits the student brings entering into the first fall term.
46	FALL_ONLINE_CRDS	Total number of online credits student took in their first fall term
47	FRST_FALL_TOT_CRS_LD	Total course load for first fall term
48	FRST_FALL_TOT_HRS_ERN	Total hours earned in first fall term
49	FALL_D_GRADES	Number of D grades in first fall term
50	FALL_F_GRADES	Number of F grades in first fall term
51	FALL_W_GRADES	Number of W grades in first fall term
52	FALL_SUM_DFW	Sum of DFW grades in first fall term
53	FRST_FALL_UCF_GPA	UCF cumulative gpa-first fall term

54	FRST_FALL_CUR_GPA	UCF term gpa-first fall term
55	FRST_FALL_CUM_GPA	Cumulative gpa-first fall term
56	FIRST_SP_STRM	Spring term if enrolled, otherwise blank
57	FRST_SP_TOT_TRNSFR	Total number of undergraduate credits the student brings entering into the spring term.
58	SP_ONLINE_CRDS	Total number of online credits student took in their spring term
59	FRST_SP_TOT_CRS_LD	Total course load for spring term
60	FRST_SP_TOT_HRS_ERN	Total hours earned in spring term
61	SP_D_GRADES	Number of D grades in spring term
62	SP_F_GRADES	Number of F grades in spring term
63	SP_W_GRADES	Number of W grades in spring term
64	SP_SUM_DFW	Sum of DFW grades in spring term
65	FRST_SP_UCF_GPA	UCF cumulative gpa-spring term
66	FRST_SP_CUR_GPA	UCF term gpa-spring term
67	FRST_SP_CUM_GPA	Cumulative gpa-spring term
68	SUMM2_STRM	Second summer term (before 2 fall) if enrolled, otherwise blank
69	SUMM2_TOT_TRNSFR	Total number of undergraduate credits the student brings entering into second summer term.
70	SUMM2_ONLINE_CRDS	Total number of online credits student took in their second summer term
71	SUMM2_TOT_CRS_LD	Total course load for second summer term
72	SUM2_TOT_HRS_ERN	Total hours earned in second summer term
73	SUMM2_D_GRADES	Number of D grades in second summer term
74	SUMM2_F_GRADES	Number of F grades in second summer term
75	SUMM2_W_GRADES	Number of W grades in second summer term
76	SUMM2_DFW	Sum of DFW grades in second summer term
77	SUMM2_UCF_GPA	UCF cumulative gpa-second summer term
78	SUM2_CUR_GPA	UCF term gpa-second summer term
79	SUMM2_CUM_GPA	Cumulative gpa-second summer term
80	TOT_ONLINE_YEAR1	Total online credits for year 1
81	TOT_CRD_LD_YEAR1	Total course load for year 1
82	TOT_HRS_ERND_YEAR1	Total hours earned for year 1
83	TOT_W_GRDS_YEAR1	Total W grades for year 1
84	TOT_F_GRDS_YEAR1	Total F grades for year 1
85	TOT_D_GRDS_YEAR1	Total D grades for year 1
86	TOT_DFW_YEAR1	Total sum of DFW grades for year 1
87	FRST_FALL_15_SCH	Binary, Think30. 15 or more credit hours taken in fall
88	FRST_SP_15_SCH	Binary, Think30. 15 or more credit hours taken in spring
89	TOT_SCH_GE30	Binary, Think30. 30 or more credit hours taken in year 1
90	FALL_TO_SP_GPA_DIFF	Calculated difference in gpa from fall to spring
91	FRST_SUM_ACAD_STNDNG_ACTN1	3 levels, Probation status end of entering summer term
92	FRST_FALL_ACAD_STNDNG_ACTN1	8 levels, Probation status end of fall term
93	FRST_SP_ACAD_STNDNG_ACTN1	8 levels, Probation status end of spring term
94	SECND_SUM_ACAD_STNDNG_ACTN1	8 levels, Probation status end of second summer term
95	FRST_SUM_Any_PROB1	Binary, any probation entering summer term
96	FRST_FALL_ANY_PROB1	Binary, any probation fall term
97	FRST_SP_ANY_PROB1	Binary, any probation spring term
98	SECND_SUM_ANY_PROB1	Binary, any probation second summer term
99	ANY_PROB	Binary, Any probation flag year 1
100	MajorChanges	Count of major changes in year 1
101	DQ_flag	Binary, Academic disqualified flag (not used in analysis)
102	Y1_TOXIC_COMBO	Count of challenge courses taken year 1
103	FALL_HSG	3 levels, First fall housing type
104	SP_HSG	3 levels, First spring housing type
105	FRST_FALL_BUILDING	112 levels, First fall housing building
106	FRST_FALL_COMMUNITY	13 levels, first fall housing community
107	FRST_SP_COMMUNITY	13 levels, Spring housing community
108	FRST_SP_BUILDING	111 levels, Spring housing building

109	Recoded_Fall_HSG	Housing type Affiliated, Non Affiliated/UCF, UCF (A, N, U)
110	RECODED_SP_HSG	Housing type Affiliated, Non Affiliated/UCF, UCF (A, N, U)
111	Intramurals	Binary, Participant in intramurals Fall or Spring
112	Intramurals_Fall	Binary, Participant in intramurals Fall
113	FALL_LLC1	18 levels, Fall living learning community description
114	SP_LLC1	18 levels, Spring living learning community description
115	SP_LLC_Flag	Binary, Spring living learning community flag
116	FALL_LLC_Flag	Binary, Fall living learning community flag
117	ANY_LLC_Flag	Binary, Any living learning community flag (fall or spring)
118	KW_Term	Knight Watch participant term
119	ANY_KW_Flag	Binary, Knight Watch participant flag
120	ATHLETE	Binary, Athlete flag
121	OSSM_Part	Binary, Out of State Mentoring Program participant
122	Out_of_State	Binary, Out of State Student
123	LEARN_Part	Binary, LEARN program participant
124	Excel_part	Binary, EXCEL program participant
125	COMPASS_Part	Binary, COMPASS program participant
126	MAPP_Part	Binary, MAPP program participant
127	Other_Spec_Pop	Binary, LEAD Scholars, Honors participants
128	Any_Support_Group	Binary, Participant in any group (LLC,KW, Athlete, OSSM, LEARN, EXCEL, COMPASS, MAPP, LEAD, Honors)
129	FRST_FALL_RWC	Number of times attending RWC in fall
130	FRST_SP_RWC	Number of times attending RWC in spring
131	LINK_Participation	Binary, Participating in the LINK program
132	VUCF_Fall_	Binary, Volunteer UCF participant in fall
133	VUCF_Sp	Binary, Volunteer UCF participant in spring
134	Any_VUCF	Binary, Any Volunteer UCF fall or spring participation
135	PELL_On_Entry	Binary, Pell eligible upon entry to UCF
136	FirstGen	Binary, First Generation student
137	BF_SCHOLAR_LEVEL	3 levels, Bright Futures Scholar Level
138	BrightFutures	Binary, Bright Futures flag
139	AID_YEAR	Aid Year if applied to FAFSA, missing if no FAFSA
140	DEPNDNCY_STAT	3 levels, Dependency status for FAFSA (missing if no FAFSA)
141	Parent_AGI	Parent Income YR1, missing if no FAFSA
142	Student_AGI	Student Income YR1, missing if no FAFSA
143	FAMILY_INCOME1	Parent plus student income YR1, missing if no FAFSA
144	Family_Income_CAT	6 levels, Family income category(missing="unknown")
145	AY_UnmetNeed	Any unmet need for YR1, missing if no FAFSA
146	STEM	Binary, STEM flag
147	Summer_Entry	Binary, Summer Entry flag
148	Enrolled_Spring	Binary, Enrolled in Spring flag
149	Enrolled_Summ2	Binary, Enrolled Second Summer flag
150	_2_Digit_CIP	24 levels, 2 digit CIP upon entry
151	Distance_UCF	Distance to UCF calculated from zip code of high school attended

References

Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1-4. Retrieved 4/23/2020

Florida Board of Govenors, https://www.flbog.edu/finance/performance-based-funding/

Florida State Legislature, http://www.leg.state.fl.us/Statutes/index.cfm?App_mode=Display_Statute&URL=1000-1099/1001/Sections/1001.7065.html

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications, 73, 220–239. https://doi.org/10.1016/j.eswa.2016.12.035

Integrated Postsecondary Education Data System (IPEDS), https://nces.ed.gov/ipeds/

Integrated Postsecondary Education Data System (IPEDS), *Graduation and Retention Rates: What is the full-time retention rate in postsecondary institutions?* https://nces.ed.gov/ipeds/TrendGenerator/app/answer/7/32?f=1%3D1

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). Linear Model Selection and Regularization. In An Introduction to Statistical Learning: with Applications in R (1st ed. 2013, Corr. 7th printing 2017 ed., pp 219–222). New York, New York: Springer

Kennedy, W. (2020, February 10). NCSS Statistical Software Documentation | NCSS Software Help. Retrieved from https://www.ncss.com/software/ncss/ncss-documentation/#Correlation

Luque, A., Carrasco, A., Martín, A., & de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition, 91, 216–231. https://doi.org/10.1016/j.patcog.2019.02.023