Will They Come Back?

– An Interpretable Model for Student Retention at UNC

Consultants: Wei Angel Huang, Ruiping Ke, Qinghua Li, Zhaoqi Liu, Xinjie Qian

Client: Lynn Williford, Chris Eilers, Office of Institutional Research and Assessment (OIRA), UNC at Chapel Hill

Professors: Dr. Perry Haaland and Dr. Steve Marron

February 2, 2021

Abstract

This report created a statistical model and framework to describe the characteristics of students who are at risk of not returning in the next semester. Using Generalized Linear Models and Decision Trees, we identified several key factors that significantly impact the probability for a student to return to school in the next semester based on data from Spring 2019. Key conclusions are: the probability of returning is higher for a student with more credits earned, with fewer credits failed, having a double major status, being a new student (instead of a transfer student), being in certain programs (such as Business), being a junior, being an in-state student, and with fewer cumulative terms. This research will potentially help school officials to target and engage with students at risk early on in order to provide guidance and assistance to not only benefit the students but also enhance the University's reputation.

Executive Summary

An overview of the major findings with an indication of where details can be found is as follows:

- *The most important factor affecting student retention is their performance (e.g. credits earned and failed) (Result 1 and 2).*
- *Students who earned low credits are more sensitive to failed credits (Result 2).*
- *Having a double major improves retention rate based on student's performance (Result 3).*
- *New students have a higher retention rate than transfer students (Result 4).*
- *Retention rate depends on the program a student is in (Result 5).*
- *Some programs with high retention rates (e.g. Business) have higher retention rates even among their students at risk due to other factors. (Result 5).*

- *Transfering students are more sensitive to low credits earned, especially in certain programs (e.g. BA, BS) (Result 5).*
- *Juniors have the highest retention rate and seniors the lowest when other factors are not considered (Result 6).*
- *Freshmen have the lowest retention rate when controlling credits earned (Result 6).*
- *Earning more credits gives better retention for freshmen and seniors (Result 6).*
- *In-state students have a higher retention rate than out-of-state students (Result 7).*
- *Poorer performance has a greater negative impact on out-of-state students (Result 7).*
- *Among non-graduating seniors, those with more than 8 terms have the lowest retention rate (Result 8).*

Suggestions of future work are discussed in the Discussion/Future Directions section at the end of this report.

Background

University retention rate is the percentage of students who study full-time in the fall semester and keep on their study as full-time students in the school in the next fall semester. According to College Transitions, for first-year students who enrolled in Fall 2018 and returned for Fall 2019, the average nationwide retention rate was 78%, while UNC Chapel Hill had a retention rate around 96%, which ranked the 44th among all universities in the United States.[1] Even though the current ranking looks good, keeping and raising the retention rate is still an important goal for UNC, because a high retention rate always means high student satisfaction and recognition. For this, it is useful to investigate the reasons and signs of students dropping out of school.

The reason why students do not return can be multifarious. The Harvard Graduate School of Education's "Pathways to Prosperity" study suggests that the most frequent reasons contain poor academic performance, inability to strike a balance between study and life, and financial burden.[2] At UNC Chapel Hill, 12% of students on average didn't graduate in 6 years since the start of class 2001. Only 24% of them dropped out because of academic ineligibility to return; the rest left school without a degree because of their own difficulties.[3] We are interested in the reasons behind those statistics. On the one hand, the retention rate reflects students' maintained interest in school after enrolling. Knowing their reasons for leaving can help make school improvements. On the other hand, school revenues are generated by student credit hour enrollments. The ability to predict retention rate for the next semester can contribute to project

[1] "Retention and Graduation Rates." College Transitions, August 22, 2020, https://www.collegetransitions.com/dataverse/retention-and-graduation-rates.
[2] William C. Symonds et al, 2011. *Pathways to prosperity: Meeting the challenge of preparing young Americans for the 21st century.* 10-11. Cambridge, MA: Pathways to Prosperity Project, Harvard University Graduate School of Education.
[3] "Six-year Graduation Rates and Academic Eligibility Status of Non-completes." UNC Chapel Hill Office of Institutional Research and Assessment, last updated August 15, 2019, https://public.tableau.com/views/WebFactsSix-YearGraduationRatesandAcademicEligibilityStatus/UNC-ChapelHillSix-YearGraduationRates

enrollment and funding correspondingly. In our study, we took several academic measurements into consideration, aiming to reveal the underlying traits associated with the dropout rate and create a model providing insights into the retention rate for UNC Chapel Hill.

Data

Our dataset was from the Office of Institutional Research and Assessment (OIRA) by Chris Eilers at UNC Chapel Hill. The dataset was collected from Spring 2019 and contains 14,238 observations and 37 variables. Student retention is defined as whether they come back or not in Fall 2019.

Credits attempted were collected ten days after Spring 2019 started, credits were collected at the end of Spring 2019, and retention was collected ten days after the Fall 2019 semester started.

Variables we used in this analysis were:

- student's retention status (IS_RETAINED = Retained)
- total credit hours attempted (CREDIT_HOUR_ATT)
- credit hours earned (CREDIT_HOURS_EARNED = Credits)
- class years (CLASS_LEVEL_STUDENT = Class)
- double major or not (IS_DBL_MAJOR = DoubleMajor)
- program (MAJ_1_PROGRAM_CODE = Program)
- residency (RESIDENCY = Residency)
- student's enrollment status (STU_ORIG_ENROLL_STATUS = Enrollment)
- cumulative terms (CUM_RESIDENT_TERMS_BOT = CumulativeTerms)
- credits failed was calculated using total credit hours attempted minus credits earned.

There are 10 variables used in this analysis in total, including 6 categorical variables and 4 numerical ones. There are 3 categories for student's enrollment status: new student, new transfer student and unclassified. There are only 67 unclassified students. Therefore, we did not include unclassified in our analysis. Our final dataset has 14,171 rows and 10 columns.

In our final dataset, there are 459 observations in the non-retained group and 13,712 in the retained group.

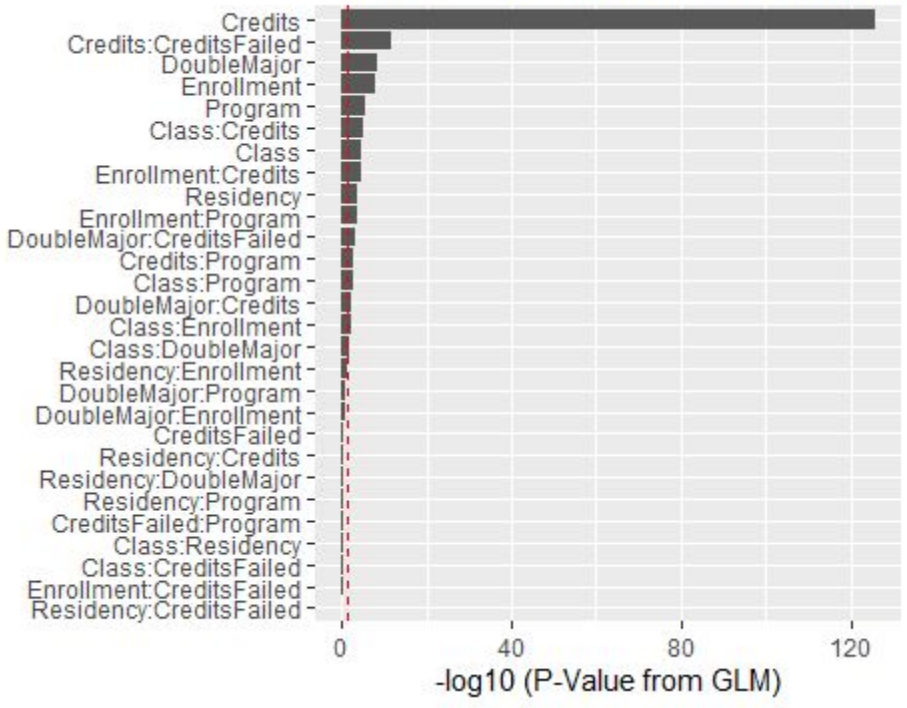
Results

1. Identification of most important factors for student retention

To determine the key factors affecting student retention, we built a Generalized Linear Model (GLM) model with Retention as the dependent variable and the variables: Credits, CreditsFailed, DoubleMajor, Enrollment, Program, Class, Residency, and their two-way

interactions as the independent variables. The model fits data very well (McFadden's Pseudo[4] R^2 = 0.25, where 0 indicates poor fit, 1 indicates perfect fit, and 0.2-0.4 indicates excellent fit[5]). Further, we ranked the effects by their significance (Fig 1.1). The most important factors are: Credits, Credits:CreditsFailed interaction, DoubleMajor, Enrollment, Program, Class:Credits interaction, Class, Enrollment:Credits interaction, Residency, and Enrollment:Program interaction (all with *p-value* < 0.001). We will investigate these factors in detail in the following subsections.

Figure 1.1 Ranking of the key effects by significance from GLM Model



Credits and Credits Failed are the most important predictors of retention. The Pareto plot above shows the significance level of each term from the GLM logistic regression fit with Retention as the response. The terms are sorted from top to bottom from most to least significant. The model contained the main effects and two-factor interactions: Credits, CreditsFailed, DoubleMajor, Enrollment, Program, Class, Residency. Red dotted line indicated where p-value = 0.05. The bars that exceed the red dotted line imply that the corresponding effect is statistically significant with α = 0.05.

[4] McFadden, D. 1974. "Conditional logit analysis of qualitative choice behavior." Pp. 105-142 in P. Zarembka (ed.), Frontiers in Econometrics. Academic Press.
http://eml.berkeley.edu/~mcfadden/travel.html

[5] Hensher, David A, and Peter R. Stopher. *Behavioural Travel Modelling*. *(McFadden) Chapter 15: Quantitative Methods for Analyzing Travel Behaviour on Individuals: Some Recent Developments.* London: Croom Helm, 1979. Print.

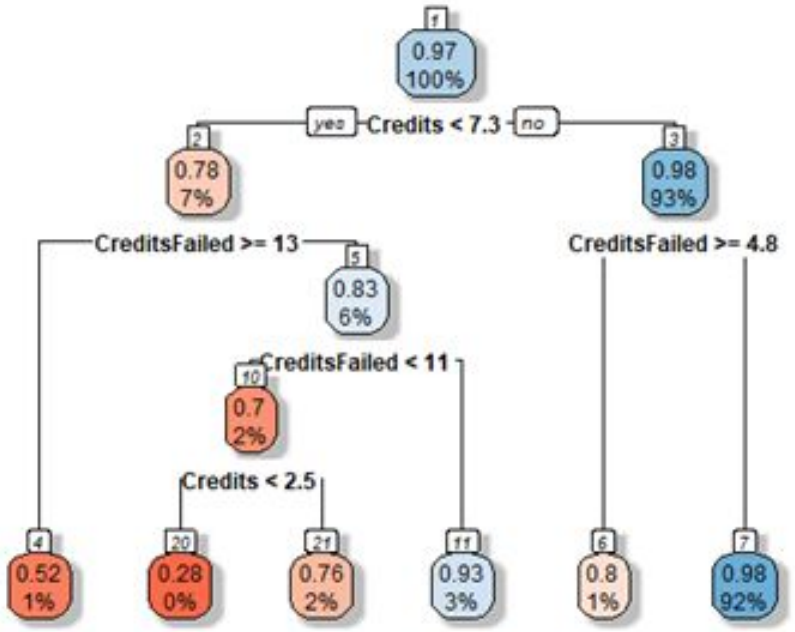
2. The most important factor affecting student retention is successful completion of credits

To find out what are the most important factors and how they affect retention, we built an interpretable model using a decision tree (Fig. 2.1). We fitted a decision tree model with retention rate being the dependent variable and class, residency, double major, enrollment, credits, credits failed, and program as the independent variables. Decision tree is a classification algorithm and uses the Gini impurity criterion to select splits, which divide the dataset into most homogenous sets. The fact that this tree splits mainly on Credits and CreditsFailed indicates that these are the two most important variables for classification, and that their interactions are also important for classification. In addition, we discovered the top splits used for classification, which is Credits less than 2.5, between 2.5 and 7.3, and greater than or equal to 7.3; splits for CreditsFailed are less than 4.8, between 4.8 and 13, and greater than 13. For easy interpretation, we will round the splits to integers for later analysis.

For better understanding of Figure 2.1, we first introduce how to read a decision tree. The tree starts from the entire dataset as the root (top node 1). For each node, the number inside the square on top of the node indicates the order of nodes. Earlier nodes (with smaller numbers) are more important in splitting the dataset than the later nodes. The first number inside a node represents the probability of retention (e.g. P(y=1) = 0.97 for the 1st node) for this node, the second number represents the proportion of total data that is in this node (e.g. There is 100% of the data in the 1st node, since no split has occurred yet). Under each node, there is a criterion for splitting the dataset, e.g. whether a student has fewer Credits than 7.3 or not; if yes, this student will be grouped into the next left node (e.g. node 2, with probability of retention being 78%, and 7% of the students are in this group); if not, this student will be grouped into the next right node (e.g. node 3, with probability of retention being 98%, and 93% of students are in this group). The color encodes the probability of retention, with blue representing high retention and red representing low retention. To estimate a student's probability of retention, one can follow each criterion and go down the decision tree until reaching the end node at the bottom, and to estimate the retention rate of students who are in that group.

The highest retention group (bluest, rightmost node in the bottom line) were students who earned at least 7.3 credits and failed less than 4.8 credits (retention rate = 0.98, 92% of students are in this group), which was consistent with our expectation since this represented the group with best academic performance. The lowest retention group (reddest, second node in the bottom line) were students who earned less than 2.5 credits and failed 11~13 credits (retention rate = 0.28, less than 1% of students are in this group), and students who failed at least 13 credits (first node in the bottom line, retention rate = 0.52, 1% of students are in this group). This suggests that the students who failed high numbers of credits are most likely to not return, especially the ones with very few earned credits.

Figure 2.1 Main splits of a decision tree classification model are based on credits earned and failed.



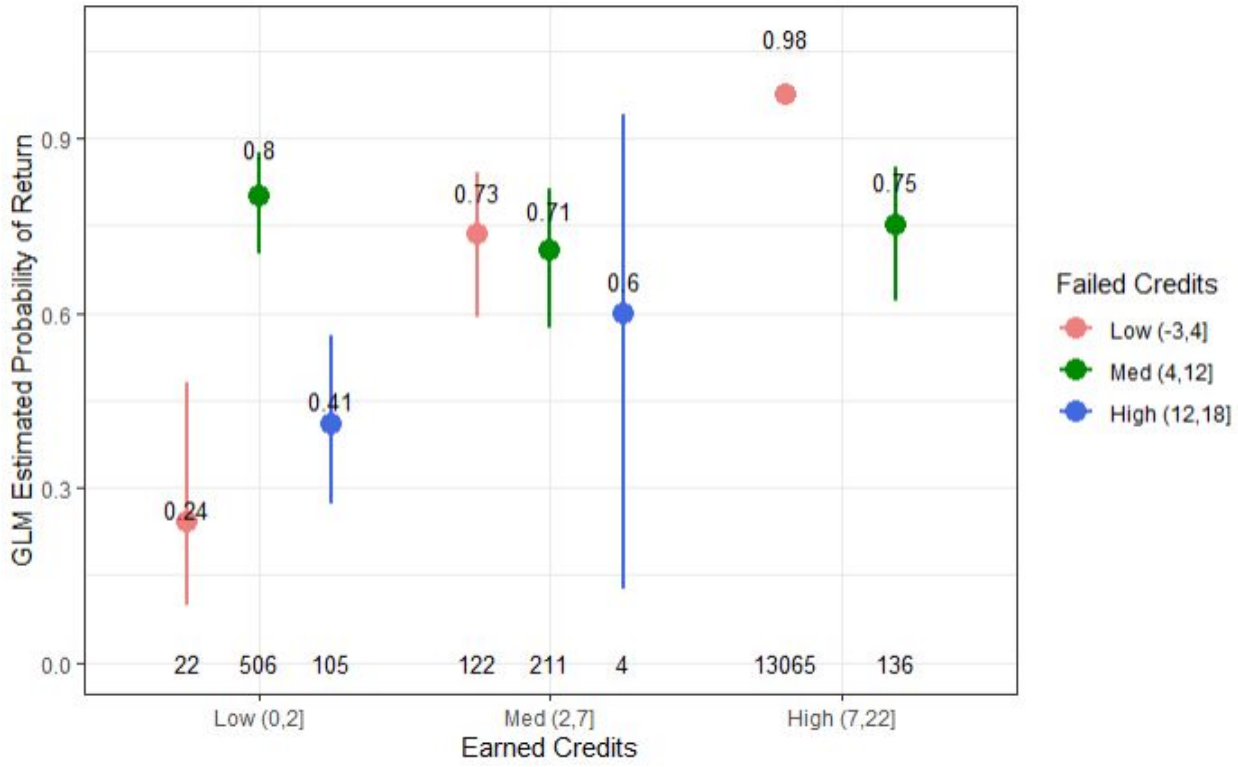
A decision tree model for explaining key factors and how they affect retention. The major splits for classification are based on credits earned and failed.

Since we identified a significant interaction effect between earned and failed credits on retention rate in both a GLM model (Fig. 1.1) and a decision tree model (Fig. 2.1), the next step was to investigate further into this interaction effect. For easier interpretation, we binned the values of Credits and FailedCredits (continuous variables) into 3 categories: Low, Medium, High. The categories for Credits and FailedCredits were defined by the decision tree splits shown above. This way, we can split the data into the most homogeneous groups while enabling easier interpretation.

We plotted the estimated retention rate for students in each category to visualize the interaction between earned and failed credits (Fig. 2.2). Students with low or high earned credits were more sensitive to failed credits, whereas students with median credits are less sensitive. Specifically, there were three different patterns for the three categories of earned credits. For the group earning no more than 2 credits (left column), the retention rate was low in general, but was the lowest (24%) when failed credit is no more than 4 (leftmost bar), probably because so few credits (no more than 6) were attempted in total. Interestingly, for the group that failed 4~12 credits, a student's return rate increases to 80%. Then when more credits are failed, the return rate drops again to around 41%. This implies a combination effect of how many credits are attempted in total versus how many are failed. When a student earns a medium (2~7) number of credits (middle column), the return rate is similar regardless of failed credits, with a slight trend

of higher retention rate with fewer failed credits. For the group earning more than 7 credits (right column), there was no student who failed more than 12 credits. This group responds to more failed credit, which is associated with the lower return rate ~75% (rightmost bar) compared to students with low failed credits (98% return rate). This trend is consistent with the finding from the decision tree model and the real retention rate by earned and failed credits, providing support for the GLM model estimates.

Figure 2.2 The interaction effect between earned and failed credits on predicted retention rate



Students who earned low credits are more sensitive to failed credits. The plot shows GLM estimated probability of return (Y axis) by earned credits (X axis) and failed credits (color). For each group, the mean probability of return is shown with a 95% confidence interval. Non-overlapping confidence intervals between two categories indicates a significant difference in probability of return between the two categories. In each category, the number on top of each bar denotes the retention rate, and the number at the bottom of each bar denotes the number of student samples.

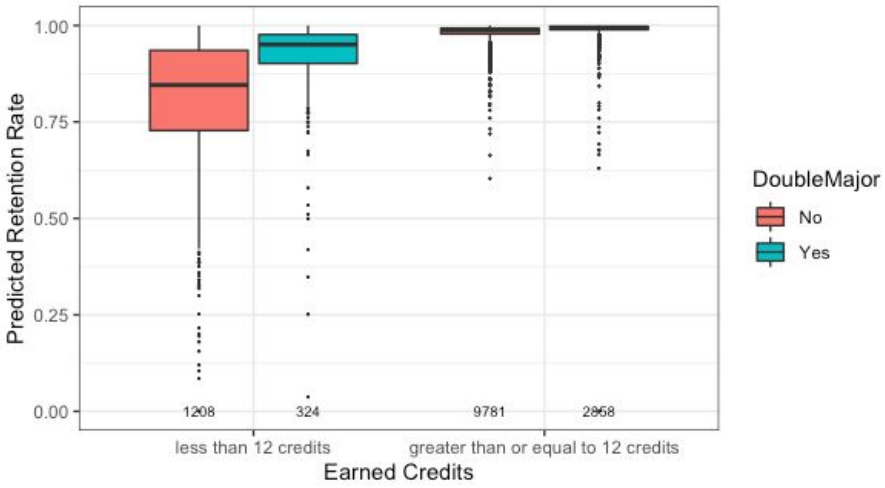
3. Double Major affects retention rate based on student's performance

In the GLM model, we found that the interaction term between Double Major and Credits has a significant positive effect on retention, which means for the same earned credits, students with a double major are more likely to return to school. This is reasonable because students with a double major tend to devote more time to studies and take more time to graduate. Then,

we are interested in how the interaction term affects the retention rate. At UNC, the minimum course load for undergraduates in a single semester is 12 academic credit hours. For students to register less than 12 credits in a semester, they need permission from their dean. If students fail some credits and finally earn less than 12 credits at the end of the semester, they may get themselves into academic trouble. So, in order to better visualize the interaction term, here we divide students into two categories based on whether the students earned at least 12 credits in a semester, and make this plot (Fig. 3.1).

To better understand the plot, we first introduce how to read the boxplot. Boxplot is a good way to display the distribution of data based on the five number summary (minimum, first quartile, median, third quartile and maximum). The first quartile is the median of the lower half of the data, and the third quartile is the median of the upper half of the data. The distance between the first and third quartiles is called the InterQuartile Range (IQR). IQR measures variability in a spirit similar to the median. In boxplots, the box is drawn from the first and third quartiles. A long box indicates a large IQR, so that the middle half of the data is more spread. The horizontal line in the box denotes the median of the data. The minimum and maximum are calculated based on the 1.5 IQR rule, which are the ends of two whiskers (vertical lines outside the box) in the boxplot. Data that are either greater than the calculated maximum or smaller than the calculated minimum are suspected outliers. In boxplot, those suspected outliers are shown in dots. In our plots, we also add some numbers to provide more information. The numbers at the bottom denote the number of students in each category. The numbers within the boxes denote the average estimated retention rate in each category.

Figure 3.1 Predicted retention rate for students with and without a second major based on their earned credits



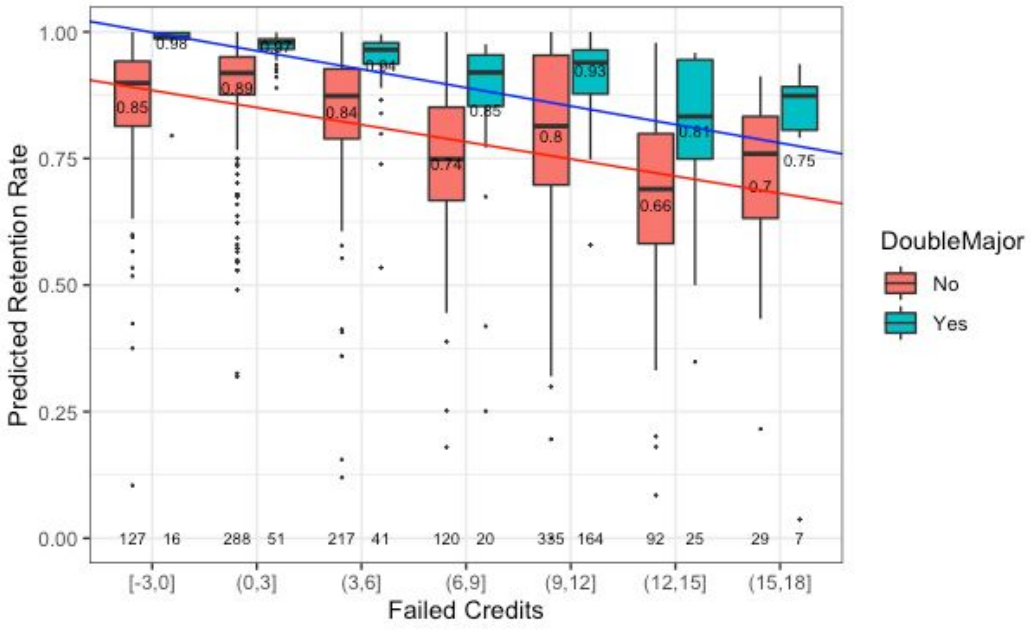
The plot shows estimated probability of return (Y axis) by earned credits (X axis) and double major (color). Students who earned less than 12 credits have substantially lower retention rate, and this is somewhat mitigated by the double major status.

We can see that double major affects the retention rate of students in two categories in different ways. For students who have earned enough credits, whether they have a second major doesn't have much impact on the predicted retention rate. They both have high retention probabilities, and their mean retention probabilities are 0.98 and 0.99. However, for students who have earned less than 12 credits, the two boxes are very different. The median value of the estimated retention rate of double-major students is even larger than the third quartile value of the estimated retention rate of single-major students. The shorter box and fewer outliers of double-major students also suggest that the predicted rate has less variability. We can see that the difference of retention rate between double-major and single-major students is more obvious for students who failed to earn enough credits.

We also want to know what role Failed Credits plays here. We classify Failed Credits into bins that are multiple of three since most classes are 3 credits. Then we divide students into two categories based on whether they earn enough credits at the end of semester, and make two plots respectively. Figure 3.2 plots the estimated retention rate for students who earned less than 12 credits, and Figure 3.3 is for students who earned enough credits.

For students who earned less than 12 credits (Fig. 3.2), there seems to be a negative linear trend between failed credits and retention rate for both single-major and double-major students. The more credits they fail, the less likely the students will return to school. In order to better visualize the trend, we fit regression lines for the mean predicted retention rate in the seven failed credits categories for both single-major and double-major students. The fitted regression lines look parallel to each other, which means the interaction term between the categorical variables Failed Credits and Double Major is not significant ($p = 0.2261$). In other words, the mean predicted retention rate decreases as failed credits increase no matter whether students have a double major or not. However, we can see that overall, according to their academic performance, double-major students are less likely to quit, because in all seven categories, double-major students all have shorter and higher boxes, and fewer outliers than that of single-major students.

Figure 3.2 Predicted retention rate with double major and failed credits for students who earned less than 12 credits

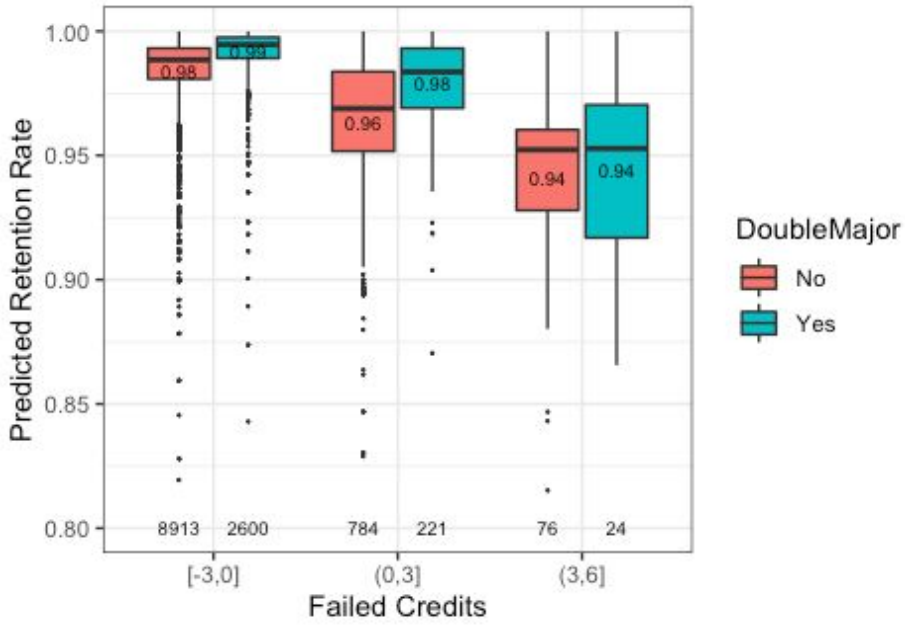


The plot shows estimated probability of return (Y axis) by failed credits (X axis) and double major (color) for students who earned less than 12 credits. The two lines are the fitted regression lines for the mean predicted retention rate in each category. Students with more failed credits have lower retention rate, a double major mitigates such drop to some extent. (The description of boxplots was given in Figure 3.1.)

For students who have earned enough credits this semester (Fig. 3.3), double major is not a very influential factor. Most of them did not fail any credits. We can see that the boxes lengthen as the failed credits increase, but we cannot tell a significant difference between single and double major. This coincides with our findings in the first plot; saying that, for students who have earned enough credits, whether they have a second major doesn't have much impact on the predicted retention rate.

In a nutshell, students with a double major are more likely to return, but the strength of this effect depends on students' academic performance. For students who have earned enough credits this semester, the effect of double major on retention probability would be small, whereas the effect would be strong for students who haven't earned enough credits this semester.

Figure 3.3 Predicted retention rate with double major and failed credits for students who earned enough credits

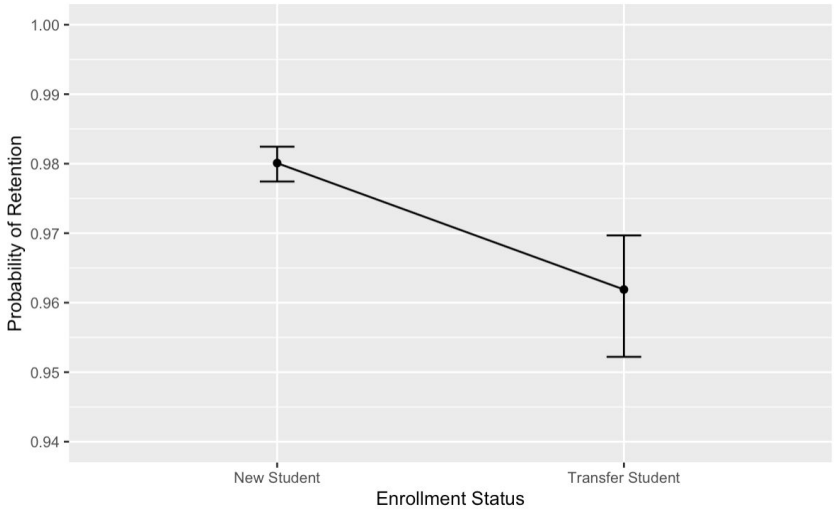


The plot shows estimated probability of return (Y axis) by failed credits (X axis) and double major (color) for students who earned no less than 12 credits. Since they can fail at most 6 credits to meet the academic minimum credits requirement of UNC, this plot only has three categories on the x axis. Students with more failed credits have lower retention rate, a double major mitigates such drop somehow. (The description of boxplots was given in Figure 3.1.)

4. Enrollment status affects students retention rate based on credits

In the GLM model, enrollment status is an important variable. We have two types of status, new student and transfer student. There are 12723 new students and 1448 transfer students in the dataset. For new students, the mean retention rate is 98% and for transfer students, the mean retention rate is 96.2%, which is significantly lower than new students (p = $3.3 * 10^{-7}$). As shown in Figure 4.1, the 95% confidence interval for the mean retention rate of the new student is much smaller than that for the transfer student, because there are many more new students than transfer students. By checking the standard deviation of two types of students (standard deviation of new students is 0.17 while that of transfer students is 0.24), we are more certain about the retention rate of the new students and the difference between these two types is significant, as reflected in the non-overlapping confidence intervals.

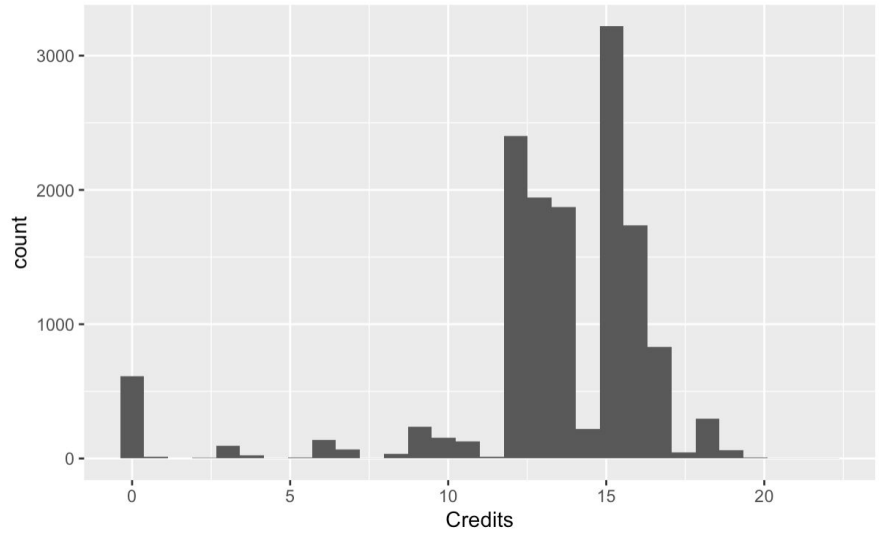
Figure 4.1 Predicted retention rate in different enrollment status



In the plot, Y axis represents probability of retention and X axis represents the enrollment status. For each status, the mean probability of retention (the black dot) is shown with 95% confidence interval.

From the GLM model, we identify a significant interaction effect between enrollment type and earned credits. We divided the earned credits into 5 parts based on a histogram for the earned credits (Figure 4.2). Most students earned 11 credits to 17 credits.

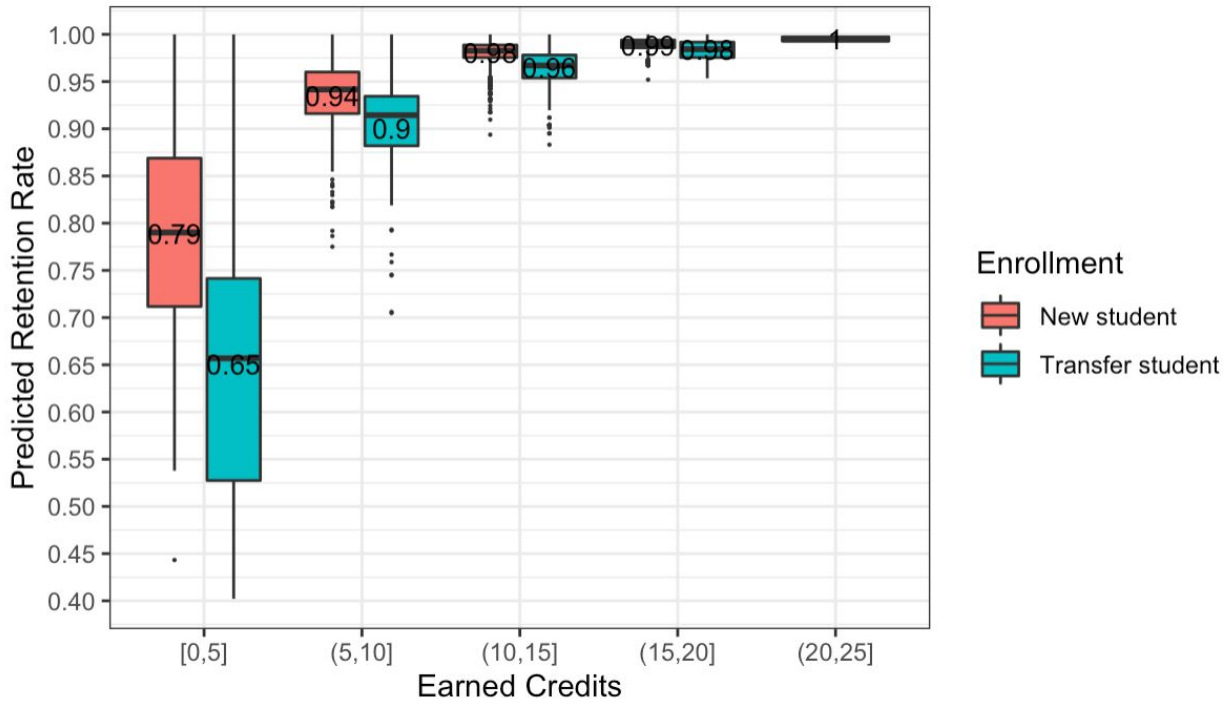
Figure 4.2 Histogram for the earned credits



In this histogram, X axis represents the earned credits and the Y axis represents the count of the students.

We plotted the predicted retention rate for each enrollment type with different earned credits groups to explore the interaction between retention rate, enrollment status and earned credits. As shown in Figure 4.3, earned credits has a positive relation with the retention rate in both enrollment types. For each earned credits group, new students have a higher predicted retention rate than transfer students. In particular, for the group with earned credits<=5, the mean retention rate of new students was 0.79 while that of transfer students was 0.65. For 5<earned credits<=10, the means were 0.94 and 0.9. Among transfer students, there is no one whose earned credits is more than 20.

Figure 4.3 Predicted Retention Rate with Enrollment Status and Earned Credits



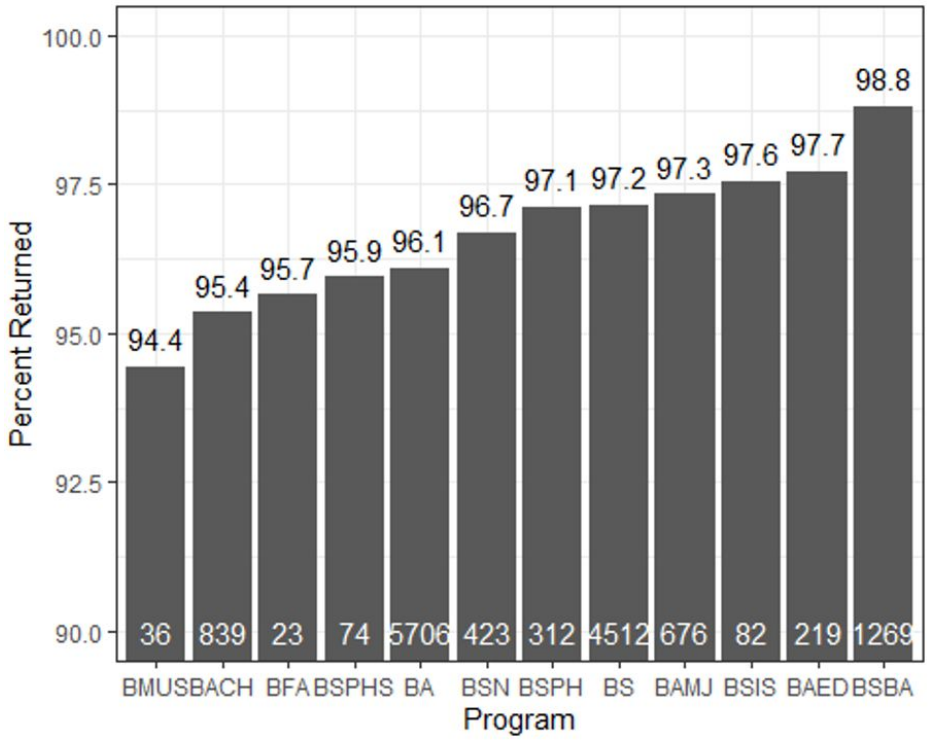
The boxplot represents predicted retention rate (Y axis) by earned credits (X axis) and enrollment status (Color). There are only 3 new students whose earned credits are more than 20.

5. Program affects students retention rate via enrollment status and sensitivity to credits

Program is one of the most important variables in the GLM model above in affecting retention rate. We first plotted the real percentage of students retained in each program in ascending order (Fig. 5.1). The programs with the lowest retention rate are BMUS (Music Performance, 36 students, 94.4% retained), BACH (College of Arts and Sciences, 848 students, 95.4% retained), BFA (Studio Arts, 23 students, 95.7% retained), BSPHS (Pharmaceutical Sciences, 74 students, 95.9% retained), but all with smaller sample sizes. The programs with the highest retention rate are BSBA (Business Admin, 1270 students, 98.8%), BAED (School of Education, 219 students, 97.7%), and BSIS (Information Science, 82 students, 97.6%). The in

between programs include BA (Bachelor of Arts), BSN (Nursing), BSPH (Public Health), BS (Bachelor of Sciences), and BAMJ (School of Media and Journalism).

Figure 5.1 Percent of students retained in each program



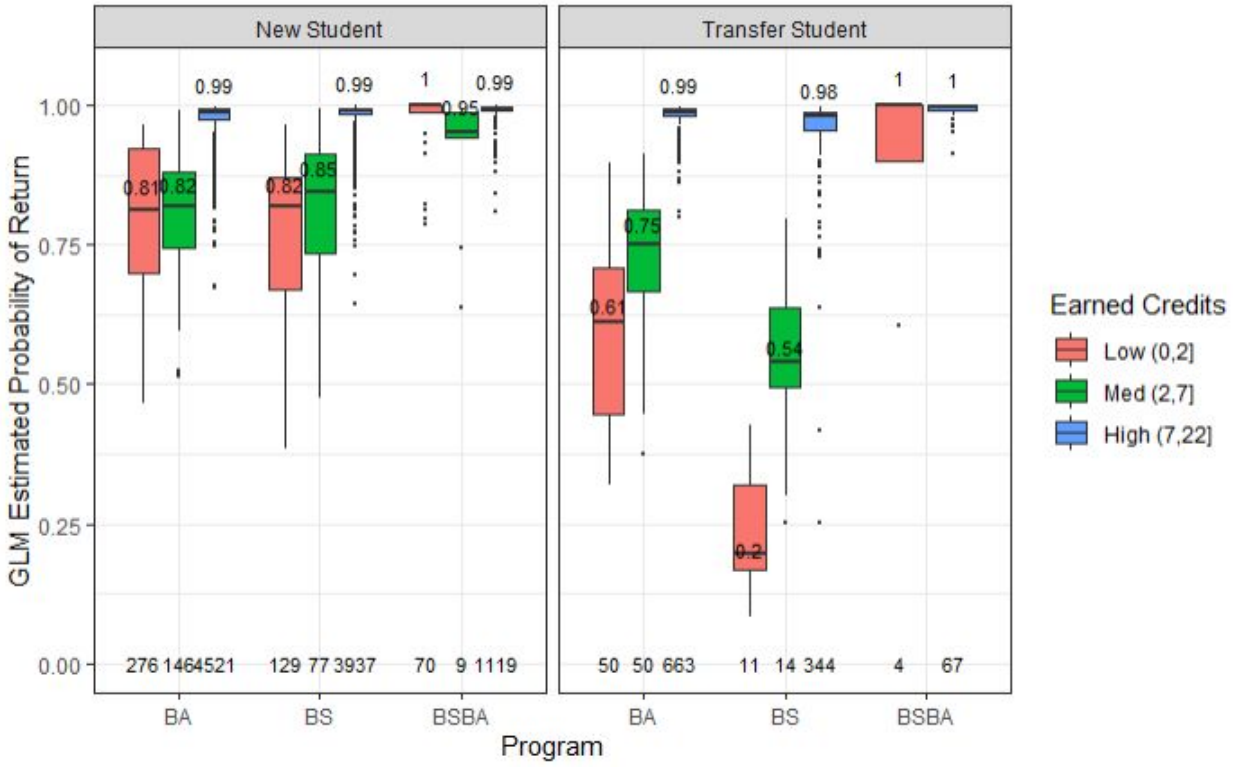
Bar plot of percentage of students retained (Y axis) by program (X axis) in ascending order of retention. Number on top of each bar denotes the retention rate, and the number at the bottom of each bar denotes the number of students in each program. This shows which programs have high (e.g. Business) or low retention rates (e.g. Music Performance).

The GLM model also revealed a significant interaction effect between Program and Enrollment. To be specific, students in different programs showed different retention profiles based on earned credits and enrollment status. We picked the 3 programs with the highest number of students as examples: science (BS), art (BA), and business (BSBA) (Fig. 5.2), because the GLM model is more likely to give accurate predictions when the sample size is large.

From Figure 5.2, we discovered several interesting patterns: First, there was a high return rate for the business program (BSBA columns), regardless of enrollment status and earned credits. This is consistent with the high return rate for that program (98.8%) observed in Figure 5.1. Second, regardless of program (column) and enrollment status (panel), the group with high earned credits had a very high probability of return (95%-99%), and the spread of data is relatively small, as indicated by the shorter boxes (blue bars for BA, BS, BSBA). However, for the group earning low to median credits, especially for BA and BS, new students were more likely to return (81%-85%) than transfer students (20%-75%) (red and green bars for BA and BS

columns). In addition, transfer students showed a gradient effect of increased credits on increasing the probability of return (right panel, red and green bars for BA and BS columns), whereas new students maintained a high probability of return for both low and median level of earned credits (left panel, red and green bars for BA and BS columns). In general, the transfer students are less likely to return when earned credits are not high and they are more sensitive to how many credits they earned, except for students from certain programs such as business. (Note that the categories with smaller sample sizes need to be interpreted with caution, as the prediction might not be robust.)

Figure 5.2 Program interacts with earned credits and enrollment status to affect the predicted retention rate

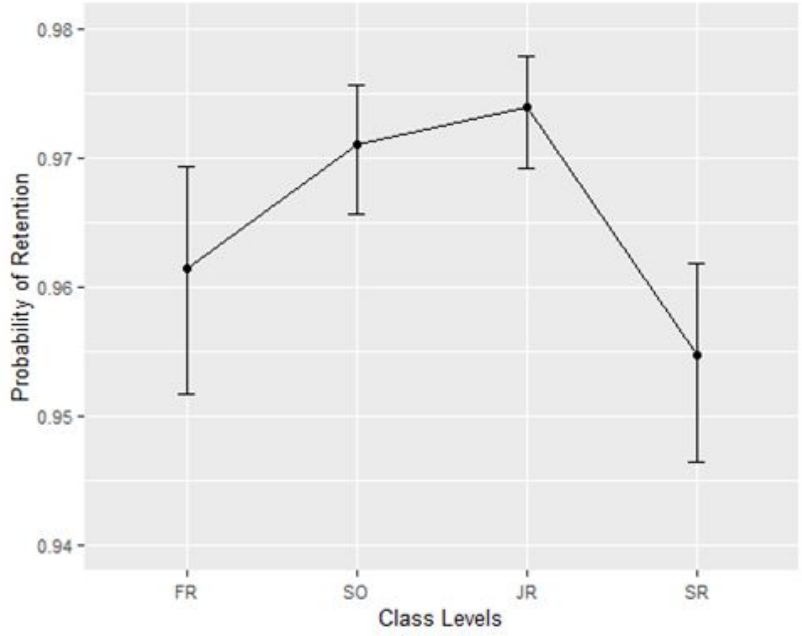


The plot shows GLM estimated probability of return (Y axis) by enrollment status (panel), program (X axis), and earned credits (color). Transfering students are more sensitive to low credits earned, especially in BA and BS programs. The business program has a higher retention rate even among transfer students and those who earned low credits. (The description of boxplots was given in Figure 3.1.)

6. Effect of class and interaction between class and credits on student retention rate

Class year is also an important variable for estimating student retention rate. Using the GLM model, the mean retention rates shown in Figure 6.1 were 96.15% for freshmen, 97.11% for sophomores, 97.39% for juniors and 95.47% for seniors. Statistical significant differences between these groups are indicated by the non-overlapping 95% confidence intervals. Compared with seniors, both sophomores and juniors had higher retention rates. Both sophomores and juniors had statistically significant higher retention rates than seniors. Freshmen had a higher mean retention rate than seniors, but the corresponding 95% CIs had some overlap with that of seniors.

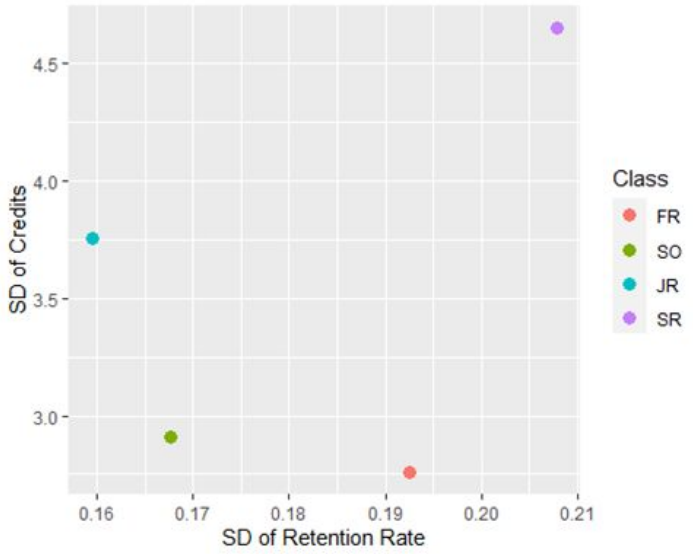
Figure 6.1 Student retention rates in different class years



X axis represents class years: freshmen (FR), sophomore (SO), junior (JR) and senior (SR). Y axis represents student retention rate. Black dots are mean retention rates, and the bars represent 95% CI. Juniors have the highest retention rate and seniors are significantly lower.

To further confirm the findings, a standard deviation (SD) scatter plot was made (Figure 6.2). Juniors had the lowest SD of retention rate and seniors the highest. SDs of retention rate for freshmen and sophomores were in between. Freshmen tend to have the least variation in terms of credits, but considerably more variation in terms of retention rate. On the other hand, juniors have much more variation in terms of credit, but less in terms of retention. Seniors have the highest overall retention in both respects.

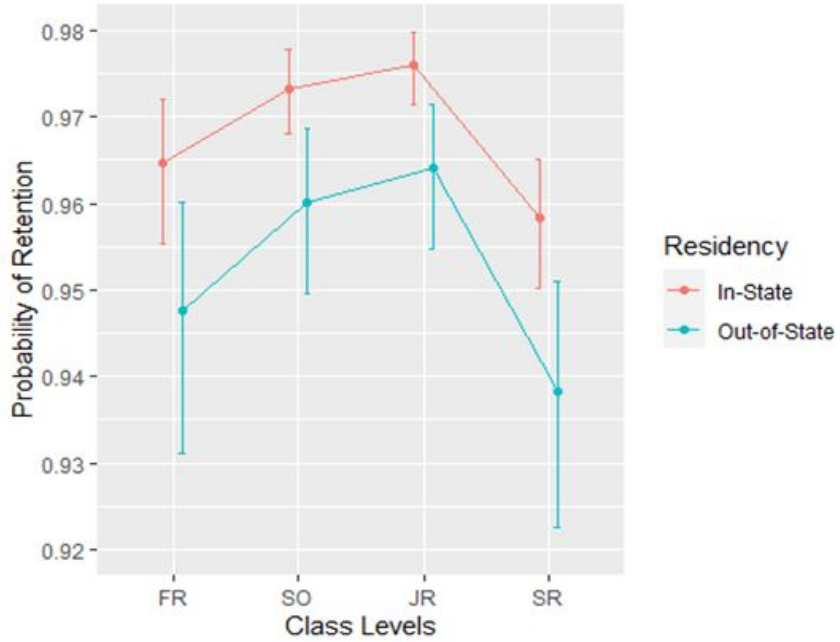
Figure 6.2 Standard deviation scatter plot



X axis represents the SD of retention rate, and Y axis represents the SD of credits. Shows large differences between classes with respect to these two types of variations.

Effect of class on student retention rate was also explored in In-State and Out-of-State groups. Regardless of the residency status, junior students had the highest retention rate and seniors the lowest (Figure 6.3). This is consistent with the finding in Figure 6.1. Interestingly, In-State students had a higher retention rate in every class year than the Out-of-State group (Figure 6.3).

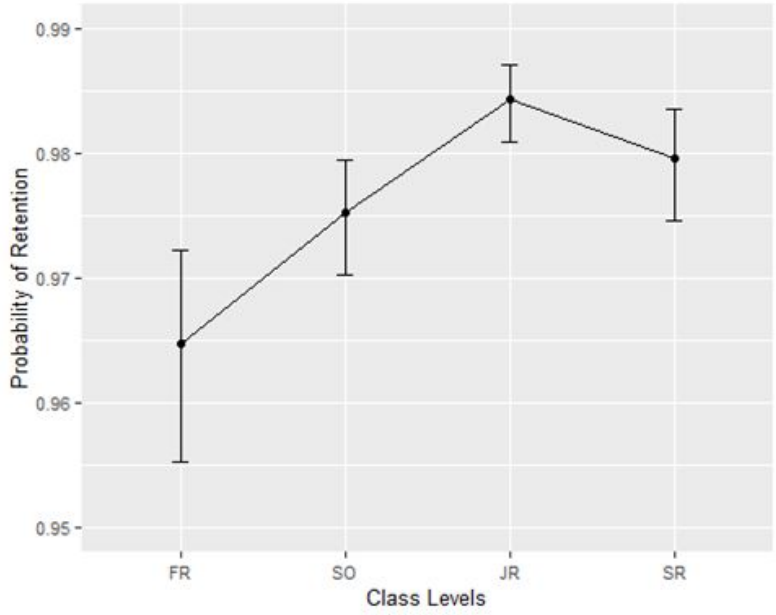
Figure 6.3 Student retention rate for class in different residency groups



X axis represents class years and color represents residency status. Y axis represents student retention rate. Dots are mean retention rates, and the bars represent 95% CI. Class year lesson is very similar to Figure 6.1, but there is a substantial difference between in-state and out-of-state students.

Another view of retention rate controls for credits earned. From this viewpoint, freshmen had the lowest retention rate. Juniors had a higher 95% CI than freshmen and sophomores, indicating juniors had a significantly higher retention rate. Similarly, in contrast to Figure 6.1 seniors have higher retention rates than freshmen (Figure 6.4).

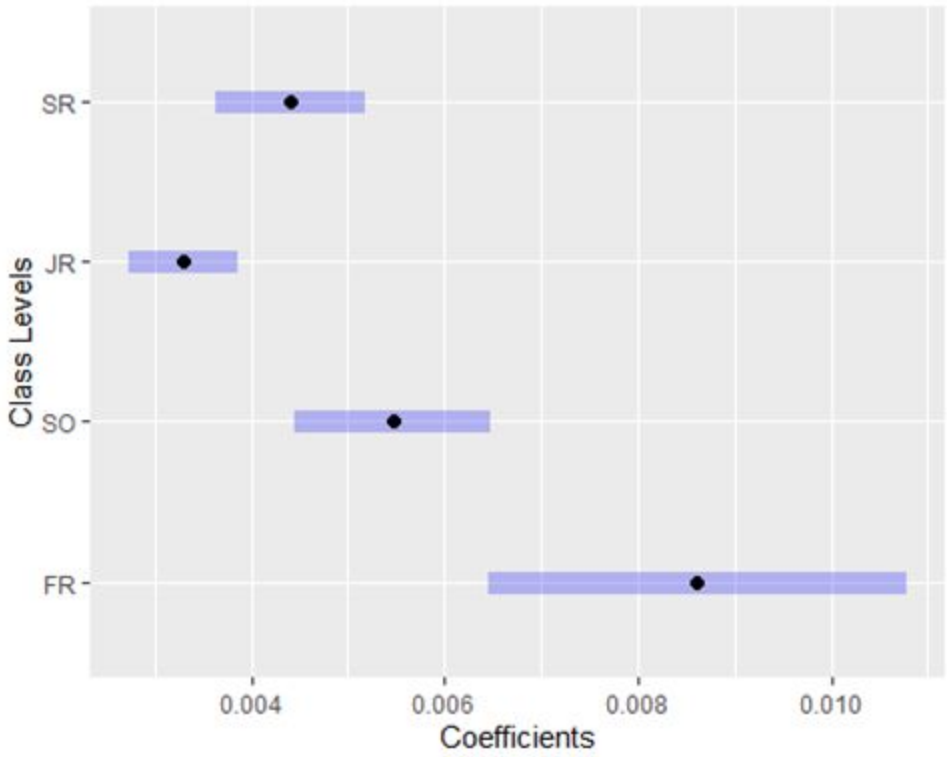
Figure 6.4 GLM main model effect plot for class year and credits earned



Retention rate controlling for credits earned. X axis represents class years and Y axis represents student retention rate. Black dots are mean retention rates, and the bars represent 95% CI. Shows different retention rates between classes than the uncontrolled rates in Figure 6.1.

Motivated by the above observation, we studied the interaction effect between class and credits on retention rate using the GLM model in Figure 6.5. Earning more credits has the greatest positive effect on freshmen and smallest positive effect on juniors reflected by the slopes and 95% CI. These findings further confirmed the results in Figure 6.4.

Figure 6.5 Class and Credits Interaction Effect



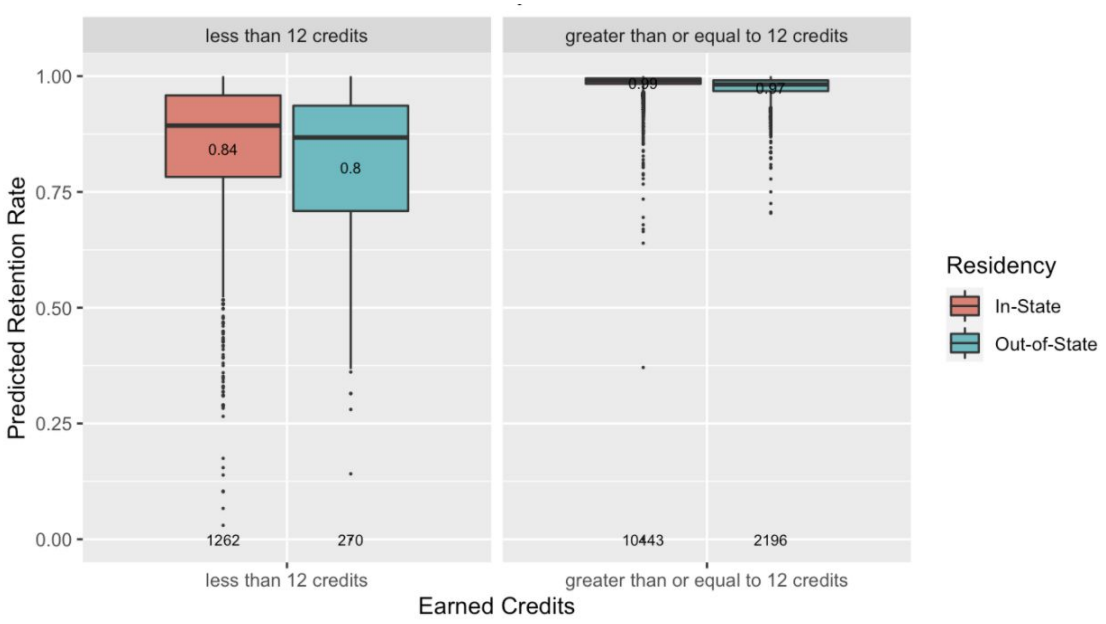
X axis represents coefficients (slope) for Credits, and Y axis represents class years. Black dots are the mean coefficients, and purple bars represent 95% CI. Freshmen are most impacted by the credits earned and juniors are the least impacted.

7. Residency status affects retention rate based on student's performance

In the GLM model, we found that Residency status is one of the most important variables in affecting retention rate. In Figure 7.1, we explored the impact of interaction between credits students earned and residency on retention rate. We divided credits earned into two categories: less than 12 and greater than or equal to 12. The mean retention rate for students who are in-state are higher than the students who are out-of-state, regardless of the earned credits category. This is reasonable not only because it's more convenient for students who are in-state to come back to university, but also because the students who are out-of-state are required to pay more tuition fees. In particular, in-state tuition is typically much cheaper than out-of-state tuition.

In addition, compared with the students who earned at least 12 credits, there is a bigger gap in the retention rate between out-of-state students and in-state students who earned less than 12 credits.That may be because the students who earned at least 12 credits have to concentrate more on their studies, leaving them with less energy to consider environmental and personal factors.

Figure 7.1 The interaction effect between earned credits and residency status



The plot shows GLM estimated retention rate (Y axis) by earned credits (X axis) and residency status (color). Description of the boxplot was in Figure 3.1. Shows credits earned have an impact on the differences between residency status.

The out-of-state student's variability of the predicted retention probability is larger, as shown by the bigger box, which means the data is more spread in this group. That also seems reasonable because out-of-state students tend to experience more uncontrollable factors, such as man-made (like family or emotional reasons) or external cause (like inconvenient transportation), and perhaps more loneliness and homesickness than students in-state.

Furthermore, from Figure 7.1, we can see that no matter the residency status, the student who earned at least 12 credits is more likely to return to UNC compared to those who earned fewer. This could be because students who have a positive experience are more likely to return to UNC.

Another way to investigate the impact of residency status is through a Chi-square test of the corresponding contingency table as shown in Table 7.1. This shows that students who are in-state have a significantly higher retention rate compared with students who are out-of-state (p=0.0002, from the Chi Square test).

Table 7.1 Comparison of the effectiveness of Residency Status

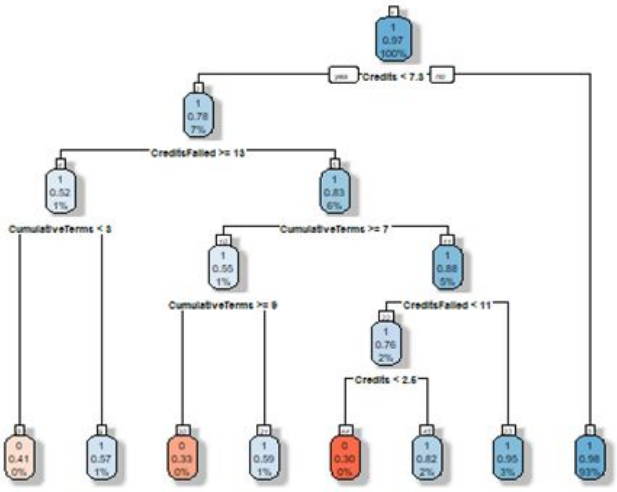
Residency Status \ Retention Status	Not return	Return
In-State	349	11356
Out-of-State	110	2356

Table shows the number of students based on residency status (rows) and retention status (columns).

8. Effect of Cumulative Term on Student Retention Rate

The number of semesters a given student has been at UNC is called CumulativeTerms. In Figure 8.1, we investigate the impact of CumulativeTerms on retention rate using a classification tree, as in Figure 2.1. In addition to credits earned and failed being important variables, this shows that the CumulativeTerms variable is also very important. This makes sense because CumulativeTerms provides additional information.

Figure 8.1 Investigation of the impact of Cumulative Terms using a Classification Tree Model



CumulativeTerms can be classified into 4 groups based on the classification tree: (0, 3], (3, 7], (7, 9] and (9, 13]. Description of a classification tree model was given in Result 2. CumulativeTerms is also important for estimating student retention rate.

To better understand the relationship between cumulative terms and class years, a contingency table was made (Table 8.1). As can be seen in the table, seniors have the broadest distribution of CumulativeTerms. We investigated how this affects retention rate in the next analysis.

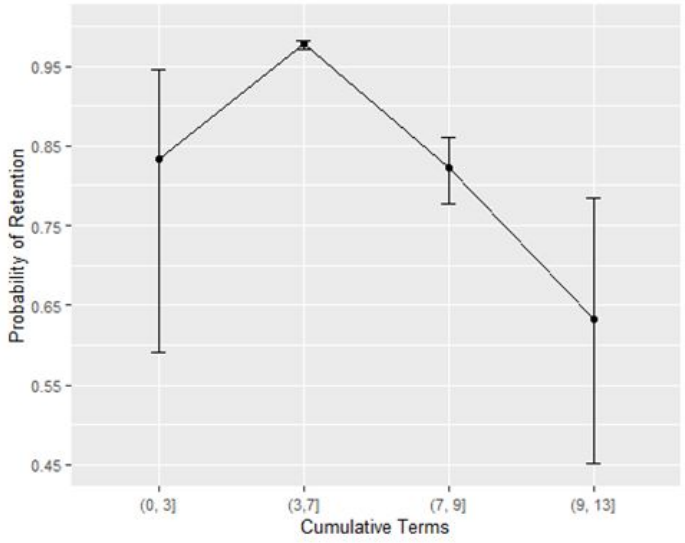
Table 8.1 Contingency Table of Class and Cumulative Terms

	(0, 3]	(3, 7]	(7, 9]	(9, 13]	Sum
FR	1841	1	0	0	1842
SO	2293	1960	0	0	4253
JR	398	4827	22	1	5248
SR	18	2447	333	30	2828
Sum	4550	9235	355	31	14171

Number of students in each class year (rows) and CumulativeTerms (columns). Shows seniors have the most variation.

As shown in Figure 8.2, for seniors only, there is a downward trend in retention rate as a function of CumulativeTerms with the exception that the (0,3] group (i.e. transfer students) have a lower rate.

Figure 8.2 GLM main effect of CumulativeTerms in seniors only



X axis represents cumulative term groups: (0, 3], (3, 7], (7, 9] and (9, 13]. Y axis represents student retention rate*. Black dots in the graph are mean retention rates, and the bars represent 95% CIs. Shows generally lower retention for more cumulative terms.*

Discussion/Future Directions

In this report, we used Generalized Linear Models and Decision Trees to build our model, from which we identified several important factors that significantly impact the probability of retention. We assume that the data we used for this model is representative of a typical year and it would be interesting to see if it applies to years following the pandemic.

The analysis in this report was based on a set of carefully selected variables. We believe there is the potential for many more interesting variables (such as GPA and financial aid information), which could be studied using a variable selection method such as Lasso in future work.

Another potential approach would be to analyze a dataset based on the specific courses a student took, instead of the summaries analyzed here. A challenge to this approach is that this would be a much larger dataset.

In the future, we could also construct a model aimed at predicting a specific student's retention probability. We could also evaluate the prediction accuracy of such a model by dividing the dataset into training and test sets.

Appendix: Methods

GLM: In statistics, the generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables having error distribution models other than a normal distribution. In the case of a 0-1 response (0=students who didn't return, 1=students who returned) as in this report, we did logistic regression, which is a special case of GLM.

Decision Tree: Decision tree is a useful and interpretable tool for classification and prediction. A decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label(student would return or not return).

Chi-square test: The Chi-Square test of independence is used to determine if there is a significant relationship between two categorical variables. The data can be displayed in a contingency table where each row represents a category for one variable and each column represents a category for the other variable, like Table 7.1. In this report, we examined the relationship between residency status (in-state vs. out-of-state) and retention rate (high vs. low). The Chi-square test of independence can be used to examine this relationship. The null hypothesis for this test is that there is no relationship between residency status and retention rate; while the alternative hypothesis is that there is a relationship between residency status and retention rate. If the p-value is less than 0.05, then we reject the null hypothesis, which means that there is a relationship between residency status and retention rate.