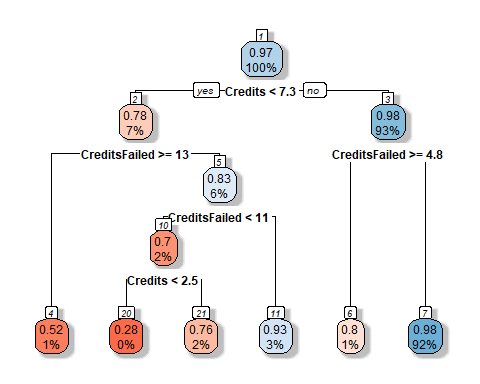
## Story 1: The most important factor affecting student retention is performance

To find out what are the most important factors and how they affect retention, we built an interpretable model using decision tree (Fig. 2). 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 Gini impurity criteria to select splits, which divide dataset into most homogenous sets. The fact that this tree splits mainly on Credits and CreditsFailed indicates that these are the two of the 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 integer for later analysis.

The highest retention group (bluest, right-most 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 number of credits are most likely to not return, especially the ones with very few earned credits.

**Figure 2. Main splits of a decision tree classification model are based on credits earned and failed.**



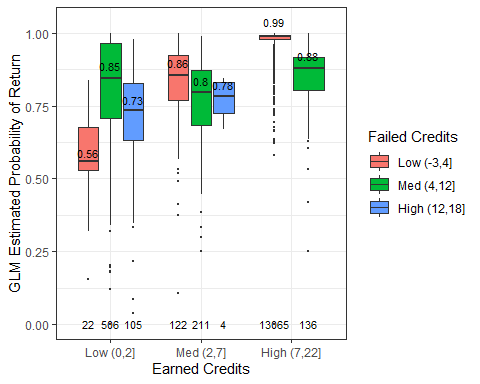
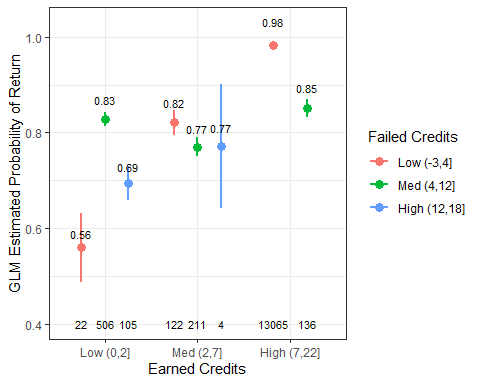
*Legend. A decision tree model for explaining key factors and how they affect retention. The tree starts from the entire dataset as the root (top node 1). For each node, the number inside the square on top of the node indicates the order of nodes, earlier nodes (with smaller numbers) are more important in splitting the dataset. The first number inside a node represents the probability of retention (eg. 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 (eg. There is 100% of data in the 1st node, since no split has occurred yet). Under each node, there is a criterion for splitting the dataset, eg. whether a student has fewer Credits than 7.3 or not; if yes, this student will be grouped into the next left node (eg. node 2, with probability of retention being 78%, and 7% students are in this group); if not, this sample will be grouped into the next right node (eg. 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 the same group.*

Since we identified a significant interaction effect between earned and failed credits on retention rate in both a GLM model (Fig. 1) and a decision tree model (Fig. 2), 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.

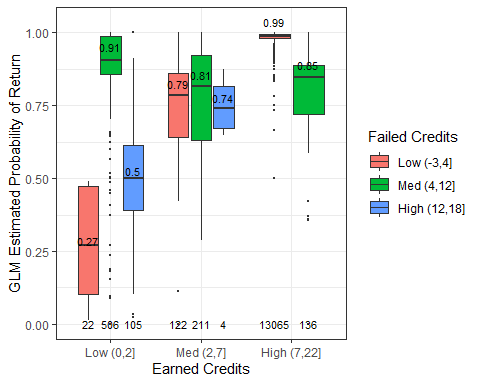
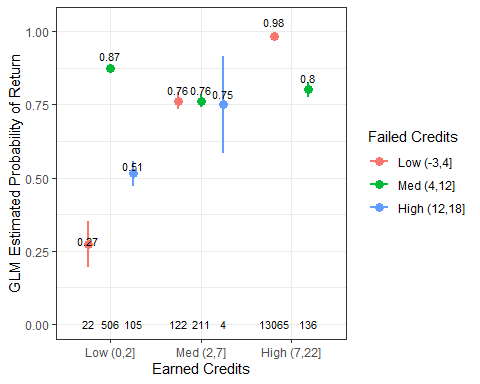
We plotted the estimated retention rate for students in each category to visualize the interaction between earned and failed credits (Fig. 3). Students with low or high earned credits were more sensitive to failed credits, whereas students with median credits are less sensitive. Specifically, there were three different patterns for the three categories of earned credits. For group earned no more than 2 credits (left column), the retention rate was low in general, but was the lowest (56%) if failed credit is no more than 4 (left red bar), probably because so few credits (no more than 6) were attempted in total. Interestingly, for group failed 4~12 credits, a student’s return rate increases to 83%, then when more credits are failed, the return rate drops again to around 69%. This implies a combination effect of how many credits are attempted in total versus how many are failed. When a student earned median (2~7) number of credits (middle column), the return rate is similar regardless of failed credits, with a slight decreasing trend with increased failed credits. For the group earned more than 7 credits (right column), there was no student fail more than 12 credits. This group was sensitive to lower failed credit, which associated with the highest return rate ~98% (right red bar), which is also the category with the most number of students. 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 3. The interaction effect between earned and failed credits on predicted retention rate.**

Left: median; Right: mean (Mean seems to look better)

Using categorical Earned and Failed credits for GLM

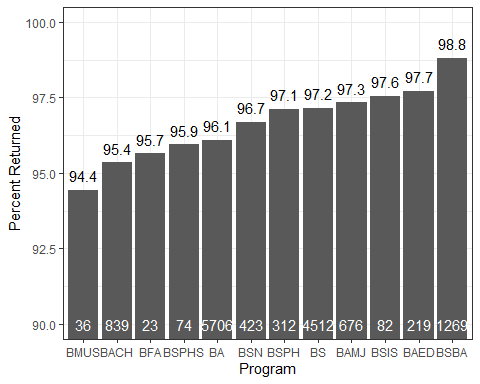
 

*Legend. The plot shows GLM estimated probability of return (Y axis) by earned credits (X axis) and failed credits (color). For each group, the mean probability of return is shown with 95% confidence interval. Non-overlapping confidence intervals (i.e. lines) between two categories indicates a significant 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.*

## Story 2: How does program affect retention rate?

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. 4). The programs with the lowest retention rate is 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 is BSBA (Business admin, 1270 students, 98.8%), BAED (School of education, 219 students, 97.7%), and BSIS (Information Science, 82 students, 97.6%).

**Figure 4. Percent of students retained in each program.**

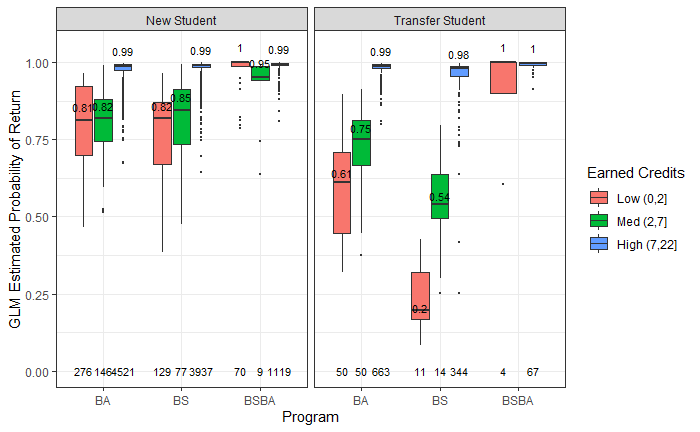


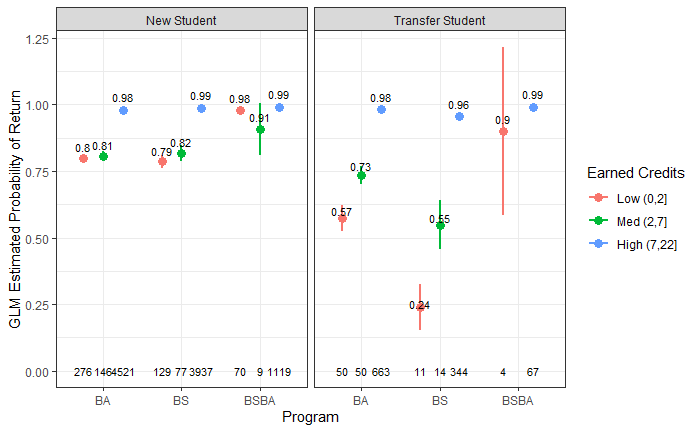
*Legend. Bar plot of percentage of students retained (Y axis) by program (X axis) in ascending order of retention. Number on top of each bar denotes the retention rate, and the number at the bottom of each bar denotes the number of student samples in each program. Programs ranked from lowest to highest retention rate are: BMUS (Music Performance), BACH (College of Arts and Sciences), BFA (Studio Arts), BSPHS (Pharmaceutical Sciences), BA (Bachelor of Arts), BSN (Nursing), BSPH (Public Health), BS (Bachelor of Sciences), BAMJ (School of Media and Journalism),* BSIS (Information Science), *BAED (School of Education), BSBA (Business Administration).*

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 3 programs with the highest number of students as examples: science (BS), art (BA), and business (BSIS) (Fig. 5), because the GLM model would more likely to give accurate predictions when the sample size is large.

From Figure 5, we discovered several interesting patterns: First, there was a high return rate for business program (both panels, BSBA column), regardless of enrollment status and earned credits. This explains why there is such a high return rate from business program (98.8%) in general. Second, regardless of program and enrollment status, the group with high earned credits had a very high probability of return (96~99%), and the variance in the group is very low (both panels, blue bars for BA, BS, BABS, except for 2 bars in BABS due to small sample size). However, for group earned low to median credits, especially for BA and BS, new students were more likely to return (79~82%) than transfer students (24~73%) (both panels, 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 double major students remained 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.

**Figure 5. Program interacts with earned credits and enrollment status to affect the predicted retention rate.**





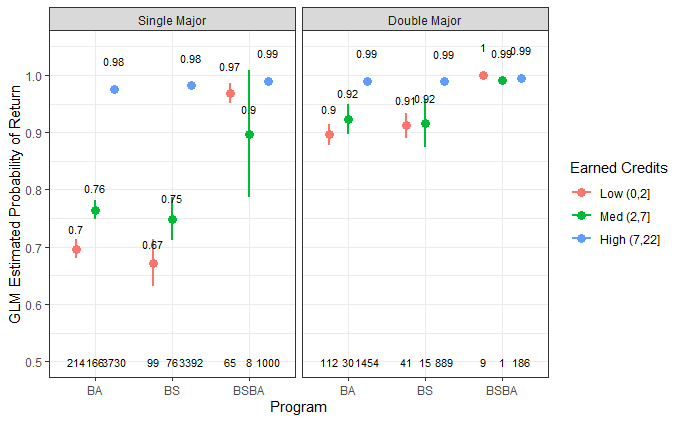
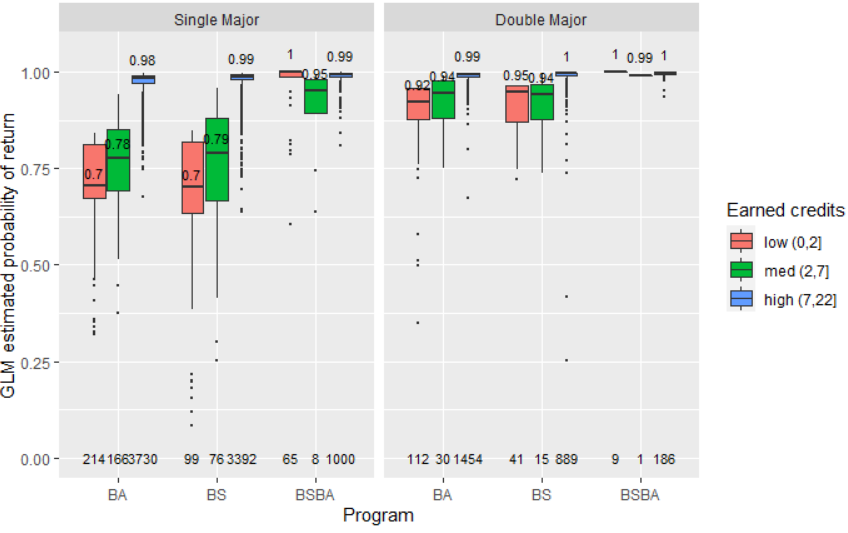
*Legend. The plot shows GLM estimated probability of return (Y axis) by enrollment status (panel), program (X axis), and earned credits (color). Dot represents the mean probability of return of the group, lines extending from the dot represent the 95% confidence interval. Non-overlapping confidence intervals (i.e. lines) 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.*

(Program \* DoubleMajor \* Earned Credit)

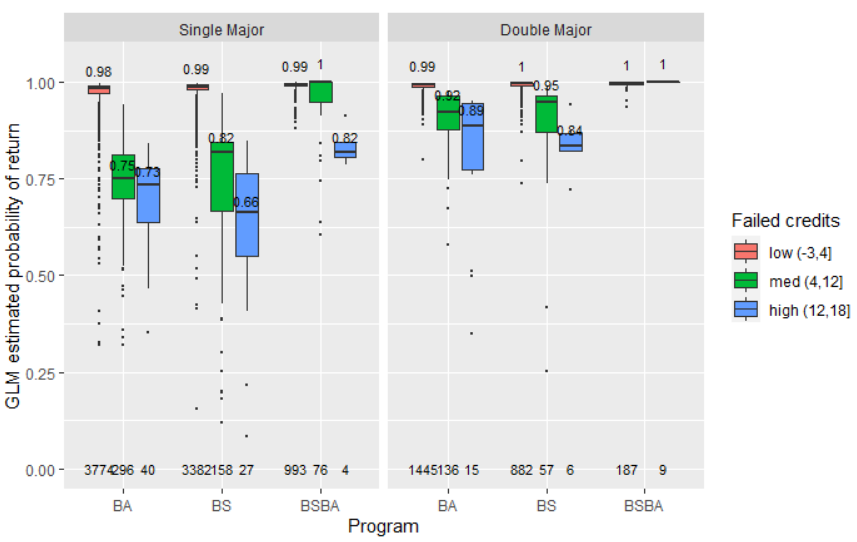
Interestingly, students in different program also showed different retention profile based on earned credits and double major status. We picked 3 programs with the highest number of students as examples: science (BS), art (BA), and business (BSIS) (Fig. 5), because the GLM model would more likely to give accurate predictions when the sample size is large.

From Figure 5, we discovered several interesting patterns: First, there was a high return rate for business program (both panels, BSBA column), regardless of double major status and earned credits. This explains why there is such a high return rate from business program (98.8%) in general. Second, regardless of program and double major status, if a student earned high credits, there was a very high probability (97~99%) that they will return, and the variance in the group is very low (both panels, blue bars for BA, BS, BABS). However, for group earned low to median credits, especially for BA and BS, double major students were more likely to return (90~92%) than single major students (60~80%) (both panels, red and green bars for BA and BS columns). In addition, single major students showed a gradient effect of increased credits on increasing the probability of return (left panel, red and green bars for BA and BS columns), whereas double major students remained a high probability of return for both low and median level of earned credits (right panel, red and green bars for BA and BS columns). In general, the single major students are more sensitive to how many credits they earned, except for students from certain programs such as business.

**Figure 5. Earned credits interact with program and double major to affect predicted retention rate.**

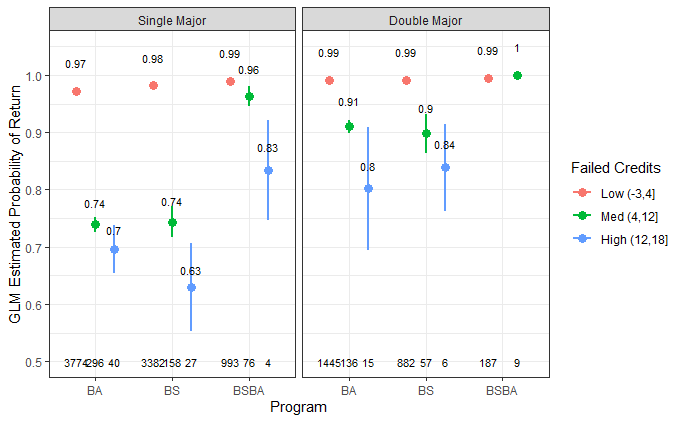


*Note. The plot shows GLM estimated probability of return (Y axis) by double major status (panel), program (X axis), and earned credits (color). Dot represents the mean probability of return of the group, lines extending from the dot represent the 95% confidence interval. Non-overlapping confidence intervals (i.e. lines) 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.*



Unlike earned credits, when we look at failed credits, it seems like all students are demotivated (have lower probability of return) by increased failed credits, as shown by the gradient decrease in probability of return as failed credits increase (Fig. 6). Consistent with previous findings, double major students are more resilient to failed credits in two ways. One is that they have a generally higher probability of return at each failed credit level, which is 10~20% higher than single major students. Second is that they are less sensitive to increased failed credits (less steep slope) compared to single major students. In addition, business students are still the most resilient and have higher probability of return. However, even single major business students were demotivated by a high level of failed credits (>12 credits failed).

**Figure 6. Failed credits interact with program and double major to affect predicted retention rate, but differently from earned credits.**



*Note. The plot shows GLM estimated probability of return (Y axis) by double major status (panel), program (X axis), and failed credits (color). Dot represents the mean probability of return of the group, lines extending from the dot represent the 95% confidence interval. Non-overlapping confidence intervals (i.e. lines) 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.*