ABSTRACT

One of Purdue University’s foremost initiatives is to increase the four-year graduation rate. Improvement in this area enhances other major goals such as increased state funding of the University and decreased student debt. This paper uses well-established statistical methods, such as logistic regression and contingency tables, to systematically investigate and interpret relationships between variables and four-year graduation rates. With the relationships detailed in this paper, applicants with a high probability of success could more easily be identified and admitted, and current students could be directly assisted by advisors when recognized through data that they are at risk of missing the four-year graduation target.

Introduction

The results and conclusions from this analysis will be presented in three parts. The first part consists of the study of variables that represented the student before attending Purdue. In essence, this is the data that Purdue sees before deciding whether or not to admit a particular student. These variables range from demographics like gender and underrepresented minority classification to high school GPA and SAT scores. The second part of this paper provides information on student outcomes with respect to four-year graduation based on the variables describing student success and behavior while at Purdue. Examples of such variables are credit hours taken per semester, study abroad participation, and residence hall data. Finally, part three will construct a predictive model corresponding to the previous two parts. This model, combining variables and their interactions, could be used to identify students who are at risk of not graduating within four years.

Part I – Demographics and High School Achievement

The first variable under scrutiny, gender, has a clear relationship with four-year graduation and makes for many interesting comparisons of different variables when controlling for gender. This variable will be present in many of the following variable analyses. The first important feature of gender to realize is that there are many (882) more males in this cohort than females (almost exactly \( \frac{1}{3} \) more). The two-way contingency table of gender vs. the four-year graduate indicator reveals significant differences in outcomes, highlighting the dependence of four-year graduation rates on gender. In Figure 1 below, the mosaic plot provides a visualization for the contingency
table. Interpretation of these mosaic plots is as follows: there are four squares, each corresponds to one of the combinations (2 genders x 2 four-year graduation outcomes = 4 combinations). The area of the squares corresponds to the percent of students present in that combination of variables. If there was no dependence structure present, squares of the same color would be of very similar area, in contrast, when there is a dependence structure, there will be moderate to large differences between the areas of squares of the same color. It is important to note that differences in square area between different colors does not show dependence, that simply shows one entire variable category has fewer observations than the other, in this case fewer females than males does not imply dependence at all, but the difference in area between the red squares does. Returning now to Figure 1, we see that the relative size of females with a “Yes” in the four-year graduate column to the “No” column reveals dependence. Figure 2 simply shows how much greater the four-year graduation rate is for females (roughly 15 percentage points). Gender with respect to other variables will be consistently highlighted in future variable analysis.

SAT data analysis revealed a few very interesting features, including relationships with gender. First it should be noted where this data comes from. Many students provided SAT scores, and that data has been used. Another large group of students provided their ACT scores. In the case of these students, the ACT score used the University’s conversion formula to compute equivalent SAT scores for this analysis. In a relatively small number of students, both an SAT and ACT score was transferred; in that case, the ACT score was ignored and the SAT scores were used as provided. Finally, some students did not have any ACT or SAT data. This group of students was quite small and thus they were discarded from this analysis.

For each of the three SAT categories (Math, Writing, and Reading), the scores were placed into 5 groups according to their quintile. This allowed substantial numbers (roughly 20%) of the cohort to be in each group.

Figure 3 below shows the four-year graduation rates for the students in each of the five quintiles. It is clear that the upper quintile of Math SAT scores corresponds to the highest four-year graduation rates. Likewise, the bottom quintile had the worst four-year graduation outcomes. The middle three quintiles are relatively even, although there is an interesting result that the second
quintile performed better than both the third and fourth quintiles. The combination of Figures 4 and 5 may explain this. Figure 4 separates the five quintiles by gender, with males represented by red bars, and females by blue. Interestingly and importantly, for the bottom four quintiles, the relative difference between males and females is close to constant, but that gap shrinks in the upper quintile. This implies that some difference between males and females four-year graduation outcomes is explained by controlling for SAT Math Category, i.e. quintile. Additionally, looking at Figure 5, it should be clear that there is a significant dependence structure between the SAT Math Category and Gender. In other words, females are overrepresented in the bottom two quintiles and underrepresented in the upper two quintiles. Numerically, nearly 43% of this cohort is female, yet more than 57% of the students in the bottom two quintiles and 27% in the upper two quintiles are female. This may describe the outcome seen in Figure 3 because, the females are overrepresented in the bottom quintiles, yet they still have higher four-year graduation rates, thus the weighted average of males and females skews the bottom quintile outcome higher.

![Figure 3: 4yr Grad Rate vs. SAT Math Category](image1)

![Figure 4: 4yr Grad Rate vs. SAT Math Category by Gender](image2)

![Figure 5: Gender vs. SAT Math Quintile](image3)
Due to the similarity of the three SAT categories, the analysis of Writing and Reading categories will be brief compared to the Math analysis. While the difference in graduation rates between the top and bottom quintile of the SAT Writing scores is not as large as it is in SAT Math, the graduation rates by category (Figure 6) is more intuitive, for the same reasons Figure 3 was not. Figure 7 shows the typical difference between male and female four-year graduation rates. The unusual drop in male four-year graduation in the fourth quintile weighed down the overall graduation rate of the fourth quintile (seen in Figure 6). Lastly, the odd shape present in Figure 3 is not present (Figure 8) due to the exceptionally consistent distribution of males and females among the five quintiles. Only the bottom quintile has a small overrepresentation of males.

The results from the SAT Reading scores had similarities to both the SAT Math and Writing scores. The primary unique feature of this data (Figure 9) is that the third quintile had the lowest four-year graduation rate. There is no clear explanation for that. Nonetheless, it is still useful to know that the upper quintile had the best four-year graduation rate; consistent with both Math and Writing SAT score data. Overall, in contrast to Writing and Math SAT scores, Reading scores do not have as strong of a relationship with four-year graduation rates. An important
takeaway from the analysis of SAT scores is that the upper quintiles generally have much better four-year graduation rates than the bottom quintiles. Additionally for females, while having better outcomes than males, they are still significantly underrepresented in the upper quintiles of SAT scores. This feature is most prominent in Math (Figure 5), but is also present in Reading (Figure 11). Males and females appear nearly indistinguishable in Writing (Figure 8) scores.

Next, High School GPA will be analyzed for its relationship with four-year graduation rates. As new variables are analyzed, their interactions with those previously investigated will also be analyzed. Just as gender was investigated, then SAT scores, then gender and SAT scores.

There is a very clear and significant relationship between high school GPA and four-year graduation. Students in the upper quintile of high school GPAs graduated within four years at roughly twice the rate of those in the bottom quintile. Figure 12 highlights this relationship by plotting the four-year graduation rate for each of the five quintiles. There is a steady increase in four-year graduation rates as high school GPA increases from one quintile to the next. Figure 13
shows the gender outcomes underlying Figure 12. Like the SAT data, the gap between males and females shrinks the most in the upper quintile for the high school GPA.

The clear differences in outcomes between males and females in Figure 13 do not show the distribution of gender in the high school GPA quintiles. Figure 14, which is a mosaic plot of gender vs. high school GPA quintile shows a dependence structure. Specifically, females are significantly overrepresented in the upper quintile of GPAs, accordingly, males are highly overrepresented in the bottom quintile.

The relationship between SAT scores and high school GPA is natural to investigate. The question this attempts to answer is, do the metrics agree with each other in their assessment of the quality of a student? In other words, does knowledge of one imply the other? Based on Figure 15, Math SAT score generally agrees with the high school GPA. This is exemplified in the corners of the plots. Very few students were in the bottom quintile for Math SAT score and the upper quintile for high school GPA (top right of mosaic plot).
Likewise, many students were in matching categories (bottom right) where they are in both the upper quintiles of Math SAT score and high school GPA. This plot shows that the two metrics generally agree, however, there is still a significant amount of variation, so both variables should be used in conjunction, rather than as substitutes for each other. It is important to note that, while only the Math SAT score was shown plotted vs. high school GPA, the other two SAT categories had strikingly similar plots and thus were redundant.

For both completeness and a way to look at a specific subject, rather than overall high school GPA, now data for high school math GPA will be presented. In general, the first two plots look very similar to the overall high school GPA. Specifically, four-year graduation rates (Figure 16) increase with the quintile of the students’ math GPA. Additionally, there is a significant difference between male and female four-year graduation rates (Figure 17) in each of the quintiles except the upper quintile, where the difference is relatively minimal. Finally, Figure 18 provides very little evidence of a dependence structure between gender and high school math GPA. This is in contrast to the overall high school GPA (Figure 14), where females were overrepresented in the upper quintiles. This necessarily implies females are achieving better grades in high school courses other than math.
Moving on to more demographic variables, starting with underrepresented minority status, exposes some interesting relationships with four-year graduation rates. Naturally, underrepresented minorities are few in number, which results in significant variation in outcomes in both four-year graduation as well as other categorical variables previously studied. This reality results in some large differences in outcomes not being found statistically significant. Without larger numbers of underrepresented minorities for study, it will be difficult to find specific significant relationships that are present but are currently hidden in noise. Nonetheless, the results will be presented with the important caveat that some differences that look significant may not be due to the few observations under study.

A clear dependence structure (Figure 19) is found between URM status and successful four-year graduation rate. This single factor relationship is statistically significant. Non-URM students’ four-year graduation rate is 47.7% (17 percentage points) greater than that of URM students (Figure 20). Dissecting the variable by gender (Figure 21) shows a reasonably consistent decrease in four-year graduation rates from females to males in both non-URM and URM. There is not enough data to identify significant differences in four-year graduation rates for URM students when comparing quintiles of SAT scores or high school GPAs. However, ignoring four-year graduation rates and simply looking at the relationship between URM status and those metrics, reveals insightful relationships. Furthermore, based on previous analysis showing the dependence of four-year graduation rates on those quintile metrics, similar results for URM students could reasonably be assumed. Figure 22 is a mosaic plot of URM status versus SAT Math Quintile. It should be clear that there is a relationship. Looking at the numbers, URM students are significantly overrepresented in the bottom quintile of SAT Math. Since these are quintiles, 20% of students would be expected to be in that category, however, URM students were in the bottom quintile nearly 50% of the time and in the upper quintile less than 5% of the time. This likely explains the four-year graduation rate disparity between non-URM and URM students. While not presented below, the other SAT categories had similar distributions; however, Math was the most extreme case. In addition to SAT scores, GPA quintiles also show a dependence structure (Figure 23). Again, instead of roughly 20% of URM students present in the bottom quintile, more than 36% were had GPAs in that quintile.
Legacy is another demographic variable with interesting relationships with four-year graduation. This variable has four possible outcomes: not being a legacy student, being the child of a Purdue alumni, being the sibling of a Purdue alumni, or having multiple relationships. More than 40% of students have some legacy relationship, so it is an important variable to study. Figure 24 presents the four-year graduation rates of the four legacy types. A statistically significant difference is found between the four possible classifications. Both sibling and parental legacies resulted in the largest improvement in four-year graduation rates. Figure 25 shows essentially the same plot with gender included as an additional factor. As expected, females have higher four-year graduation rates but the pattern is a little different. Females see improved four-year graduation rates with either a parental or a sibling legacy. Males, however, generally only see a significant improvement with a sibling legacy. While these were found to be statistically significant, it was beneficial to group the three types of legacies into one so that any legacy could be compared to no legacy. This allowed for more total observations in each combination of factors. Otherwise, the few observations resulted in too much noise to find significance for URM students. While both underrepresented minority students and non-underrepresented minority students alike appear to benefit from legacy relationships (Figure 26), a significantly larger improvement in
four-year graduation rates is associated with legacy students who are also underrepresented minorities. The final legacy relationship presented (Figure 27) is the four-year graduation rates of students in each high school GPA quintile separated into legacy and non-legacy students. There is a clear upward trend in four-year graduation rates for both legacy and non-legacy students as GPA increases; however, at each quintile the legacy students had better four-year graduation rates.

The final variable presented for the Part I analysis of high school and demographic data is student residency. There are significant differences in four-year graduation rates between the three types of residency (Figure 28). Foreign students visibly out-perform in-state and out-of-state students. There is also a statistically significant difference between out-of-state students and in-state students. When combining the residency factor with gender (Figure 29) more information is gained about where the differences are occurring. Female foreign students stand far out from the other combinations of students. Comparing high school GPAs and residency could have been an interesting analysis, alas, foreign students generally did not transfer a GPA. They did, however, transfer SAT scores which were previously found to generally agree with GPA (Figure 15). As a substitute for GPA, and perhaps an interesting analysis in its own right,
residency vs. Math SAT score (quintile) shows (Figure 30) a significant dependence structure. More than 50% of foreign students placed in the upper quintile of Math SAT scores. In contrast, about 12% of in-state students were in the upper quintile and 30% for out-of-state students. Looking at the bottom quintile, under 2% of foreign students scored in that quintile, 13% of out-of-state students and more than 25.5% of in-state students. While not presented, an analysis of the other two SAT categories and its relationship with residency was performed. In both cases, in-state and out-of-state students’ distribution in the quintiles were similar. In other words, out-of-state students were distributed in the five quintiles roughly the same for each category. The same is true for the in-state students. Foreign students, however, had vastly different distributions from Math to Reading and Writing. In Math foreign students were significantly overrepresented in the upper quintile. In Reading and Writing foreign students were overrepresented in the bottom quintile, particularly in Reading with 46.5% in the bottom quintile. Finally, comparing residency status by legacy status reveals the interesting relationship that perhaps being a legacy student positively influences four-year graduation rates for in-state students the most; the absolute increase is the largest for in-state students (Figure 31). Nonetheless, as concluded previously, legacy status is associated with higher four-year graduation rates.
Part II – Purdue Student Success and Behavior
The information gleaned in Part I of this paper could be used to identify promising students based on their particular combination of the variables analyzed. Part II, however, could be used to monitor students once they are at Purdue, in an effort to reach out to students that are, based on data, at risk of not graduating within four years.

The first variable studied, transfer credits, could have been studied in either Part I or Part II, but because often these credits were earned during summers after starting at Purdue, it will be presented now. More than \( \frac{1}{3} \) of students transfer at least one credit to the University. Transfer credits for this analysis were defined as credits earned from any transfer institution and transferred during or before Summer term 2013 (the last term to be considered a four-year graduate). Dichotomizing the transfer credit earned variable into students who did not transfer any credit and those who transferred some credit produces substantial evidence (Figure 32) that four-year graduation rates increase with transfer credit earned. The amount of transfer credit was wide ranging (10% of students transferred more than 10 credits), so perhaps presenting the data in two groups does not satisfactorily describe the underlying structure. Separating the data into intervals of 10 credits results in five groups (Figure 33). Students that transferred between 1 and 30 credits had four-year graduation rates substantially higher than those students who transferred no credits and those who transferred more than 30. The very majority of the students who transferred credits were in the range from 1 to 10 credits. In fact, fewer than 3% of students were in the 21 to 30 credits and 31 or more credits categories combined. The benefit of presenting this data in two figures is that Figure 32 shows a positive relationship between transfer credit earned and four-year graduation rates, and the second shows more of the underlying structure. Furthermore, relatively few students transferred more than 21 credits, so the strength of the evidence is stronger in Figure 32, where many students are included in making the estimate.

There is additional knowledge that can be gained by studying transfer credits. For example, studying the different effects of transfer credit for the particular college a student is in, or understanding what kind of courses are being transferred. While insightful, Figure 34 requires some explanation of what it is depicting. First students were separated into two groups, one group of students had transferred at least one course (computed as at least one credit earned
outside of Purdue) and the other group of students transferred zero courses. Further separating those students into their respective college (as of their first term, fall 2009) there were then two groups of students for each college. With two groups in each of 10 colleges, the four-year graduation rates were computed for each of the 20 total groups. The figures plotted are the ratios of four-year graduation rates within colleges of the students who transferred at least one course, to those who did not transfer any. All but one college (College of Education) had a ratio greater than 1, implying the general improvement in four-year graduation rates for students who transfer credit. Nonetheless, there is substantial variation in improvement between colleges. As an example, First Year Engineering students that transferred credit graduated within four-years at a rate 23% higher than those who did not transfer credit, College of Agriculture students, however, had a 63% improvement. The average improvement, represented by the horizontal line, was a 40% increase in four-year graduation rates.

An additional question that the study of transfer credits can answer is what kind of courses are being transferred. This analysis is revealing the department and college equivalent of the transfer course, not the college the student was in. As a simple example, if a mathematics student in the College of Science transferred an English course it would be classified as English in the College of Liberal Arts. First, looking at the total number of courses transferred to the various colleges (Figure 35), it is clear that a minority of colleges receive a vastly disproportionate number of transfer credits. More than half of the courses were in the College of Liberal Arts and the College of Science; not even accounting for the fact that a significant portion did not have a specified college, surely many of which would have been distributed to those two colleges. There are many departments, so it would be impractical to plot all of them (Figure 36), but using the top 13 (this is not completely arbitrary, these were the departments with at least 100 transfer courses), it...
is still clear that two departments receive the majority of transfer courses, English and Mathematics. The final analysis with regard to transfer credits is identifying the most common transfer institutions. The transfer institution with, by far, the largest number of courses transferred (Figure 37) is Ivy Tech in Lafayette. There were many institutions that had credit earned transferred to Purdue, and figure 37 shows those institutions that had more than 100 of such courses.

The next variable under study is participation in Explorers (Undergraduate Studies Program). Directly comparing Explorers and non-Explorers (Figure 38) highlights the significant difference in four-year graduation rates between them. Controlling for more variables may reveal differences between the Explorers themselves. Because there are relatively few Explorers in this cohort (460 out of 6,166 students), the number of subgroups will be limited to prevent subgroups with too few students in them to draw statistical inference.

To be consistent with the paper thus far, gender will be the first variable controlled (Figure 39). Of the students that are not Explorers, females had a four-year graduation rate about 36.5%
greater. Female Explorers, however, have a four-year graduation rate about 72% greater. This implies that male Explorers perform significantly worse than non-Explorers, relative to females.

The first variable studied in Part II of this report was transfer credit earned. Controlling Explorers for transfer credit earned (Figure 40), uncovers a considerable increase in four-year graduation rates among the Explorers that transferred at least one earned credit. Explorers transferred credit earned at roughly the same rate as non-Explorers (30.8% and 33.8%, respectively).

Due to the relatively small number of Explorers, instead of controlling for five High School GPA quintiles, Explorers are compared to non-Explorers based on classification as above or below the overall median of high school GPAs. This resulted in more than 100 students in each of the four combinations; enough for reasonable statistical inference. After controlling for High School GPA, the gap between Explorers and non-Explorers four-year graduation rates shrinks markedly.
There is a statistically significant dependence structure of high school GPA on Explorer status. Specifically, Explorers are significantly underrepresented above the median high school GPA, and thus overrepresented below the median (roughly 27% and 73% respectively). This implies that Explorers were much more likely to be below the median GPA than above it, further explaining the four-year graduation rate disparity between Explorers and non-Explorers.

To summarize this analysis, although Explorers have relatively poor four-year graduation rates, the program is clearly not the cause, as the gap between the Explorers and non-Explorers approaches insignificance when controlling for the quality of student. It appears to be a self-selection issue, where students of lower quality enter the Explorers program at higher rates than students of higher quality. This quality determination here is based purely on high school GPA.

The variable with arguably the strongest relationship with four-year graduation rates is the GPA. Prior analyses showed the dependence of four-year graduation rates on high school GPA (overall and mathematics), now Purdue term and cumulative GPAs will be investigated.

Comparing four-year graduation rates to eight GPA intervals, each 0.5 points wide, highlights the obvious trend that four-year graduation rates increase with GPA. Figure 43 shows this with respect to the cohorts first term, fall 2009.
Controlling for gender, the usual distribution appears (Figure 44), where females outperform males in every category. There is a notable decrease in the gap between males and females at the highest GPA level. After seeing GPA’s large impact on four-year graduation rates, it is interesting to look at the GPA distribution between males and females (Figure 45). Without even seeing the numbers, it is apparent that females are overrepresented in the higher GPAs and underrepresented in the lower GPAs. In the highest GPA range, females account for nearly 49% of that subpopulation, despite being roughly 43% of the population. Keep in mind that these are not quintiles, and thus the bottom has a much smaller number of students. Nonetheless, females only account for about 30% of GPAs (after first term) 2.0 or below (thus males account for 70%). The distribution evens out for the upper GPAs (above 2.0). Here females account for about 45% of these GPAs, a moderate overrepresentation.

Continuing with studying GPA’s relationship with other variables, two more demographic variables, residency and underrepresented minority status, are presented for brief analysis. In studying residency status vs. first term GPA (Figure 46), the primary unique feature is that foreign students are heavily overrepresented in the highest GPA interval. Approximately 38% of foreign residency students maintained between a 3.5 and 4.0 after first term. Non-residents and residents both had almost 25%. Interestingly, foreign residency students simultaneously had a relatively larger share of students in the bottom three GPA categories. In other words, the percent of foreign students in the highest GPA interval and the bottom three intervals was larger than both non-residents and residents alike. Non-residents and residents were similar in their percent of students within each category.

At least partially explaining the gap in four-year graduation rates between underrepresented minority students and non-URM students (Figure 20), the percent of URM students in the highest GPA interval is about 40% (11 percentage points) smaller than the percent of non-URM students in that interval (Figure 47). Although the difference is not as significant as the highest interval, the second highest interval also has non-URM students having a larger proportion in that category than the proportion of URM students. This naturally implies the URM students are
represented in higher proportions in lower GPA intervals, which was shown to be associated with lower four-year graduation rates.

Although not presented in plots here, both legacy and explorers were studied with respect to first term grades. For legacy status, the differences between those with a legacy and those without was minimal. Overall students with legacies had better grades, but it was not a drastic relationship as seen in other variables. Then, the three types of legacies compared to themselves shows that they had very similar outcomes. Explorers, however, had a relationship that was surprisingly similar to that of URM students. The percentages were notably similar. A natural next relationship to study was between Explorers and URM students. If Explorers were also largely URM students, this similarity in relationship with grades would be reasonably explained. The result is, however, that there is an insignificant relationship between URM and Explorers. In other words, being an URM student does not make one more likely to be an Explorer.

Thus far, analyses of grades (at Purdue) only included the first term of this cohort. Due to the value of GPA as a predictor variable for four-year graduation, and because it changes over time, the variable could be studied periodically to identify students with low probability of four-year graduation. To highlight that the GPA matters just as much first term as it does in later terms, four-year graduation rates vs. GPAs are plotted for both the end of sophomore and junior year (spring cumulative). These plots do not include the students who dropped out prior to or during that term, therefore the horizontal line representing the average four-year graduation rate at that time, is increasing due to exclusion of the dropout students weighing down the four-year graduation rate.

Students with grades between a 3.5 and 4.0 at the end of either sophomore or junior year graduate within four-years more than 75% of the time. Students with between a 3.0 and 3.5 also graduate at high rates of nearly 70%. It is the students with a GPA below 3.0, that begin to graduate within four-years at a rate less than that particular term’s average, and it drops significantly from there.
Another key variable for successful four-year graduation is the amount of credit attempted and earned each term. The focus of this analysis will be on credit attempted, particularly for fall and spring semesters. In this cohort’s first term, and each of the cohort’s ensuing terms, the amount of credit had a significant relationship with four-year graduation rates. During the cohort’s first two semesters (Figures 50 and 51, respectively), students attempting more than 15 but less than or equal to 17 credits had the highest rate of four-year graduation. This analysis is important due to the large number of students taking an amount of credits outside of that range with the highest four-year graduation rate. Slightly more than 21% of students took 14 or fewer credits their first term. Those students had a four-year graduation rate of approximately 40%, or about 6.5 percentage points below the cohort average.

Only 10% of students attempted more than 17 credits their first term. Furthermore, their four-year graduation rate was nearly 45%. This implies that attempting too few credits, in other words 14 or fewer, is much riskier to achieving four-year graduation than attempting more than the 15 credits necessary to stay on-track for four-year graduation.
In the cohort’s second year, the conclusions are generally the same as in the first year, with the exception that students four-year graduation rate increases with number of credits attempted (Figures 52 and 53), instead of peaking at 16 or 17 credits. The horizontal line, “Average 4yr Grad Rate”, in each of the graphs is not constant across graphs. It increases each term to account for the new four-year graduation rate of the students remaining at Purdue after each term.

While this cohort’s fifth term (not including summers) is generally more of the same pattern, interesting changes being in spring term of 2012 (Figures 54 and 55). Both tails of the distribution increase substantially over the previous terms while the middle begins to flatten out. A reasonable explanation for this occurrence is that students who are taking many credits are (successfully at about 77% four-year graduation rate) attempting one final push to make it going into their final term. At the other end of the credit spectrum, the considerable jump in four-year graduation rate of those taking 11 or fewer credits, could be explained by two things, a couple students graduated at this point (a year early), but most of the difference comes from being ahead of the necessary credit amounts and being able to coast to the end.
In the cohort’s final year (while still able to graduate within four years), the shape of the four-year graduation rate vs. credits attempted changes drastically (Figures 56 and 57). The graphs have intuitive features. At this point in the students’ academic career (seventh term), they know how many credits it will take to graduate. This is a reasonable explanation for the flatness of the plot. The students graduate at roughly the same rate for any amount of credits (exception for students who are not full-time). Then in the eighth term, the students who can coast by taking 12 or fewer credits have excellent four-year graduation rates, those taking more have some differences that are not too difficult to interpret. Those taking more than 18 credits are trying hard to graduate, hence the spike in four-year graduation rates.

There is an important and interesting relationship between grades and the number of credits attempted. In the following analysis, cumulative GPA and credits attempted that term will be the variables studied over time. Though not as concise as a bar chart, the following heat maps are revealing of the underlying relationship. Each of the following four plots represents one of the four spring semesters’ credits attempted that term versus the cumulative GPA interval. The rectangle color indicates the row specific concentration in that combination of credits attempted.
and cumulative GPA. It is not necessary to look at individual rectangles, rather it these plots are to show general trend.

For the first two spring semesters, there is a strong and clear diagonal trend of increasing grades with increased credits attempted. Thus, larger concentrations (red) move toward higher GPAs (x-axis) as the credits attempted (y-axis) increase. One way to look at these plots is to identify where the largest concentration is for each row, indicated by the rectangle that is the darkest red in that row. For example, during spring semester 2010, the largest concentration of GPAs for students taking more than 18 credits was in the very top right of the plot, corresponding to the GPA interval of 3.5 to 4.0.

An important note is that only students taking courses during the respective semester are included in these plots, so students who have dropped out are not included. This results in many white squares, particularly in the following two plots, where no students were present, so the concentration is zero.

In Figures 60 and 61, it appears that the differences in concentrations diminish across credit totals as the average GPA tends to converge somewhere between 2.5 and 3.5. Naturally, very few students were located in the blue sections because very low GPAs have a strong negative correlation with four-year graduation.

Figure 60: Spring 2012 concentration of Credits and Grades
Figure 61: Spring 2013 concentration of Credits and Grades

Until this point in the report, the study of credits has focused on the amount attempted. This was useful in showing what the students were willing to try for, and the outcomes based on that willingness. Now the focus will turn to credits earned to show what was possible for the students to achieve. Before specifically studying credit earned directly, which would look very similar to the amount of credits attempted, the outcomes of summer school participants will be presented.

Simply plotting the four-year graduation rate of those who attended summer school at least once at Purdue versus those who did not (Figure 62), unveils a statistically significant increase in four-year graduation rates among those who participated in summer school. This, however, is extremely misleading. There exists a statistical phenomenon called survivorship bias, which unless accounted for, biases conclusions about some variable taking place over time. Summer
school is perfect example of a variable that is influenced by survivorship bias. To elaborate, a student needs several semesters of study to graduate. In this particular dataset, the fastest time to graduation was 5 semesters; three students with a lot of transfer credit accomplished this. Thus, because multiple terms are necessary in order to graduate, the more terms taken the more likely a student is to graduate. Therefore, in the vast majority of cases, a student will have been at Purdue for two semesters (Fall and Spring) before the first opportunity to partake in summer school. This has two primary results, (in general) if a student attempted summer school, they were probably already at Purdue for a full two semesters and thus had not dropped out. On the contrary, if a student did not attempt summer school due to dropping out before the opportunity, that student is indistinguishable from those who graduated without needing summer school. So, the group of students who did not attend summer school (via the “No” indicator) includes those students who dropped out before having the opportunity. This explains the misleading feature that was mentioned above. Survivorship bias is present because the students who attended summer school generally had a minimum of two semesters of school already and were more likely to graduate in four years than the students who did not go to summer school, which includes dropouts, and successful students alike.

There is no standard or widely accepted methodology for controlling survivorship bias. The method used to control for survivorship bias in this study was simply to control for students who had the opportunity to do variable X (e.g. summer school) and then compare the outcomes of those who chose to and those who chose not to participate in that variable.

In contrast to Figure 62, when controlling for students who were present during or at any point after Fall term of 2010 (one year after starting with one summer semester opportunity), the four-year graduation rate of students attending summer school is lower than those who did not. This difference, however, is not statistically significant and thus no plot is shown. This method of controlling for attendance after one summer semester opportunity ignores the fact that there are three more opportunities (including summer 2013) to attend summer school that are not controlled for. If, instead of controlling for fall 2010 attendance, all four summer semesters are captured by controlling for either graduation or attendance during spring 2013, a significant
difference results between students who did and who did not attend summer school (Figure 63). The conclusions are the opposite of those made without controlling for survivorship bias.

There is a natural relationship between summer school and credits earned. Students who did not earn enough credit during the regular school year to stay on track for four-year graduation have the opportunity to earn credit during the summer to essentially catch-up or even to get ahead. The following analyses will focus on the relationship between these two variables.

To do the following analysis, two variables were created, one that is an indicator of needing summer school, the other indicating if summer school was successful. In the former, if the sum of the credit earned prior to fall 2009 (first term), during fall 2009, or during spring 2010 is less than 30 credits, that student is indicated as being “off-track”. For the summer school success variable, the student is indicated as either being off-track before summer and then on-track after summer (that is a success), or off-track before summer and remaining off-track after summer.

Beginning with the indicator variable for needing summer school after the first regular school year, a very strong dependence structure is found between four-year graduation rates and on or off-track status at the end of the first year. Of course, survivorship bias is affecting these results. The students who dropped out during fall or spring term are generally not going to be on track, unless they had a lot of transfer credit, so they are artificially weighing down the four-year graduation rate. After correcting for the survivorship bias by only including those students who attempted credit during spring 2010 or any semester thereafter, the conclusions are largely unchanged. Students who did not earn at least 30 credits during fall 2009 and spring 2010 only graduated in four years approximately 20% of the time. These statistics are most troubling because of how many students were off-track after their freshman year; in fact, nearly 1/3 of students were off-track after spring of freshman year. Had those students that were off-track, graduated at the same rate as the on-track students, an additional 780 students from this cohort would have graduated on time.

Figure 64: 4yr Grad Rate vs. On-track Indicator

Figure 65: 4yr Grad Rate vs. Survivorship Bias Corrected On-track Indicator

Despite a very large portion of students becoming off-track during their freshman year, a small (12%) minority of those students attempted summer school the following term. Although the
four-year graduation rate of the off-track students who attempted summer school in 2010 was larger than the students who did not, the difference was not statistically significant.

There was a significant difference in four-year graduation rates, however, between the students that were off-track before summer school and then became on-track during and the students who were off-track and stayed that way until at least fall 2010. The four-year graduation rate of those students who essentially caught up was more than 10 percentage points higher than the students who did not catch-up that following summer. Interestingly, the students who caught up still had a four-year graduation rate significantly less than the students who were on-track to begin with.

Any number of reasons could account for this; one of them being that getting off-track early on is indicative of getting off-track again later on.

It is insightful to identify what type of students will get off track, so several variables will be combined to identify significant patterns. Looking at gender, for example, females accounted for approximately 37% of the off-track students, but they are 43% of the cohort as a whole, so they are underrepresented in the off-track group. Although the catch-up rate is low for both females and males, females caught up during summer school at twice the rate the males did. Amazingly, the four-year graduation rates of females who did not catch-up is about four percentage points greater than males who did catch-up.

There are also substantial differences between the colleges the cohort started in, in terms of the percent of students off-track after spring term of freshman year. As previously mentioned, nearly \( \frac{1}{3} \) of students were off-track after spring of freshman year, so colleges with a larger percent are thus doing worse than average and colleges with a smaller percent are doing better. Plotting the percent of students that were off-track after spring term of freshman year versus the student’s college (Figure 66) reveals the following differences. Students that were in Undergraduate Studies (Explorers) were off-track more than 50% of the time. In contrast, Pre-Pharmacy students were off-track just 16% of the time.

It is clear that there are both significant numbers
It has been demonstrated that a significant portion of the cohort was off-track of four-year graduation by the end of their first year and that being off-track is detrimental to four-year graduation rates. The prior analysis focused on credit earned, because that is the true determination of on or off-track status. However, it would be interesting to find out if the students set themselves up to be on or off-track from the beginning (not attempting enough credit), or if it is a result of failed courses (thus no credit earned) that they are off-track at the end of the first year. By adding transfer credit earned before summer 2010 and then the number of credits attempted during fall 2009 and spring 2010 the final figure describes whether a student would be on track even if they passed all of their classes. It turns out that more than 16% of students did not attempt enough credit to be on track at the end of the freshman year (controlling for those students who were still present during or after spring 2010). This explains roughly half of the number of students who were not on track to graduate after their first two semesters. The other half, then, can be explained by course failures.

The patterns described above continue in the years ahead. Students failing to earn enough credit have a difficult time catching back up and thus have poor four-year graduation rates. This year by year analysis controls for survivorship bias in the usual way by making comparisons only between those students who were still present at the time, in other words dropouts who of course do not have enough credit earned, are not weighing down the four-year graduation rate. This also explains why the average four-year graduation rate line on each of the plots increases yearly to account for the graduation rate of the students still present. Figures 68 and 69 depict the four-year graduations rates vs. the students’ on-track status for year 2 and year 3, respectively. Despite the benefit of four-year graduation associated with continued presence, the rate actually decreases substantially for students off-track from year 2 to year 3 to year 4. This could be
caused by a number of factors including continued or increased course failures, a credit deficit that is difficult to overcome, or simply a willingness to go at a slower pace and graduate in four plus years.

Studying abroad is another interesting variable to study that is susceptible to survivorship bias. Without controlling for the bias, the difference in four-year graduation rates between students who studied abroad and students who did not is drastic (Figure 70). Again, this variable includes students studying abroad as late as summer 2013. Controlling for students who either graduated already or were present during or after spring 2013 enables direct comparisons between students who had the opportunity to study abroad. Although the difference in four-year graduation rates is large and statistically significant (Figure 71), it is not as large as the biased outcome.

This difference in outcome is intuitively explained, at least partially, by the nature of the students who study abroad. Looking at the data, it is apparent that part of the explanation arises from females being overrepresented in the group of students studying abroad. 59% of study abroad participants are female (when controlling for survivorship bias) despite being just 43% of the
cohort. Another explanation is the fact that students who studied abroad have significantly better grades than those who do not (looking at first semester grades for reference). A final explanatory variable is that some colleges with above average four-year graduation rates are overrepresented in the pool of students studying abroad. An example of an overrepresented college with excellent four-year graduation rates is the College of Health & Human Sciences.

Aside from simply investigating whether students did or did not study abroad, the particular time (Year or Semester type) the study abroad took place could also be considered. Starting with the study of four-year graduation and its relationship with the year the study abroad was taken (Figure 72), it is clear that the first three years have far better four-year graduation rates than either not studying abroad or going during the fourth year. This is also true when controlling for survivorship bias. In fact, students who did not study abroad actually graduated within four years at a higher rate than the students who studied abroad their last year. While it appears that studying abroad during the first year is the best, it should be noted that only 35 students participated in that period, so the sample size is small and thus more variable. It is reasonable, though, that studying abroad in the first year would have the highest graduation rate if it is before students are in arguably more challenging upper-level courses.

![Figure 72: 4yr Grad Rate vs. Study Abroad Indicator](image1)

![Figure 73: 4yr Grad Rate vs. Study Abroad Survivorship Bias Corrected](image2)

This cohort had five study abroad timing options; they could go during winter break, spring break, summer, fall, and spring semester. The data for winter break, which looks exceptional (Figures 74 and 75), should simply be disregarded because in the entire cohort, only four students studied abroad during that time, thus inference about future students cannot reasonably be made based on that few observations.

Aside from the winter break period, spring break stands out as the best time to study abroad. The four-year graduation rates of those students who studied abroad during spring break were significantly higher than any other period, including not studying abroad. The other three periods are reasonably similar after controlling for survivorship bias (Figure 75).
Perhaps one of the more interesting and unintuitive variables is the count of terms in the residence halls. As should be clear from Figure 76, the four-year graduation rates of students who were in the dorms for an odd number of semesters was significantly lower than students who stayed in the dorms an even number of terms, including zero. Number of semesters in the residence halls is another variable that is intuitively impacted by survivorship bias. Figure 76 has been controlled for students who either graduated or were still present during or after spring 2013.

In an effort to explain the unusual shape the students who were in the dorms for an odd number of terms (roughly 5% of the cohort) will be studied more in-depth. Beginning with the 60 students only in the dorms for 1 semester (and who graduated or was still present spring 2013), 95% of those students had their first term as their only term in the dorm. Additionally, 80% of those students were taking classes the next semester (spring 2010). Therefore, 75% of students with one semester in the dorms were there just their first term, left the dorms but were still at Purdue during spring 2010, but they only graduated within four years at a rate of 42%.
Of the 103 students with 3 terms in the dorms, 88% had their first two semesters fall and spring of freshman year. Half of those students were in the dorms for the last time fall of sophomore year, and the other half came back to the dorm at a another time. The first half of those students graduated within four years 45% of the time. The second half only graduated in four years 32% of the time. For the 295 students that were in the dorms an odd number of terms (with a four-year graduation rate of 41%), some possible explanations for poor four-year graduation rates could be doing a Co-Op, joining Greek Life, or taking a break from school. Finally, making comparisons between just the odd students results in a statistically insignificant outcome. It looks like there is an upward trend, and there may be one, but with the relatively small sample sizes, it is not significant. Likewise, comparing students with even numbers of semesters in the dorms, there is not a significant difference in four-year graduation rates.

**Part III – Predictive Model Construction**

The usefulness of logistic regression versus linear regression is that probabilities (of four-year graduation) are modelled. In this way, responses are limited to values between zero and one. The theoretical aspects of logistic regression will not be presented in this report.

An immense number of combinations of variables and interactions between variables could be fitted to create an effective predictive model. Here two primary (with distinct purposes) models are presented. The first model more or less corresponds to the first part of this report. Using only the information that an applicant provides, prediction of four-year graduation success is reasonably (if not surprisingly) accurate. This model includes all of the variables presented in Part 1 including many of the investigated interactions. Some insignificant interactions were removed from the final model. Figure 77 below shows which variables and interactions were included in the final model. Some variables show insignificance, but are present because the Law of Hierarchy calls for including insignificant lower-order variables if significant higher-order variables are included in the model. Finally, Figure 78 shows a type of cross-validation. At a probability level of 50%, the model classifies each student into more than likely to be a four-year graduate or less than likely. The sensitivity value of 55.7 indicates that almost 56% of the time the model correctly classified a student as being a four-year graduate. Likewise, specificity of 73.1 indicates the model correctly identified students who were not four-year graduates.
In summary of the first model, before students ever step foot on Purdue’s campus, the model accurately determined four-year graduation classification at a rate of 65%.

Of course, a lot more information is gained as students spend time at Purdue. At the same time, it is not a fair model to input spring 2013 grades as a factor because it would be clear that many students dropped out and the model would correctly classify all of those students. The model could be used however, to continually monitor student progress by including the most recent variables. At that point, the dropout students will have already been realized, so that model could be valid for making actual predictions. Here, as an example, only information up to and including fall 2009 was included. In other words, new variables consist of fall 2009 GPA, fall 2009 credits attempted, etc. The model using data up to and including fall 2009 had a significant improvement in sensitivity (Figure 80), now nearly 68%, and a moderate decrease in specificity to 72.4%. The Type III sums of squares are also presented in Figure 79. Instead of cross-validating using the same dataset that created the model, future studies could look at another cohort (say, fall 2010), to see how well the model classified those students.

<table>
<thead>
<tr>
<th>Effect</th>
<th>DF</th>
<th>Wald Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range_Value</td>
<td>4</td>
<td>81.7648</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>Gender</td>
<td>1</td>
<td>0.0938</td>
<td>0.7594</td>
</tr>
<tr>
<td>Study_Abroad_IND</td>
<td>1</td>
<td>0.0041</td>
<td>0.9491</td>
</tr>
<tr>
<td>Gender*Study_Abroad</td>
<td>1</td>
<td>0.5091</td>
<td>0.4755</td>
</tr>
<tr>
<td>Frozen_College</td>
<td>10</td>
<td>32.674</td>
<td>0.0003</td>
</tr>
<tr>
<td>Gender*Frozen_College</td>
<td>9</td>
<td>26.3698</td>
<td>0.0018</td>
</tr>
<tr>
<td>Study_Abroad*Frozen_College</td>
<td>10</td>
<td>15.4675</td>
<td>0.1134</td>
</tr>
<tr>
<td>Gender<em>Study_Abroad</em>Frozen_College</td>
<td>9</td>
<td>18.0639</td>
<td>0.0344</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Prob Level</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Percentages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event</td>
<td>1536</td>
<td>2349</td>
<td>897</td>
</tr>
</tbody>
</table>

**Figure 79: Type III Sums of Squares for Fitted Model**

**Figure 80: Sensitivity and Specificity of Fitted Model**

**Conclusion**

The above discovered and interpreted interesting and important relationships between variables, in particular with respect to the response variable of four-year graduation. Additionally, predictive models were devised that could both assist the admissions process and the continual advising process.