5-1-2014

Enrollment Trends at the College at Brockport

Jourdan Giambona
The College at Brockport, jourdanaly@gmail.com

Enrollment Trends at the College at Brockport

A Senior Honors Thesis

Submitted in Partial Fulfillment of the Requirements
for Graduation in the Honors College

By
Jourdan Giambona
Mathematics Major & Finance Minor

The College at Brockport
May 1, 2014

Thesis Director: Dr. Pierangela Veneziani, Associate Professor, Mathematics

Abstract

We have analyzed enrollment patterns at the College at Brockport, State University of New York, between 2008 and 2013. The percentages of students attending mapped by SAT score and high-school GPA over time shows a shift in the composition of our freshman cohort. The college has concentrated its efforts to improve enrollment rates through financial leveraging. Because the purpose of our analysis is to guide the College in its enrollment and marketing efforts to accepted students, we evaluate pre-enrollment variables as predictors of one-year retention of first-time students. Information about family background (parental education and socioeconomics), individual attributes (academic ability, race and gender), characteristics of the student's high school, and high school academic records are incorporated in our model.

Table of Contents

Acknowledgements

I would like to thank my Thesis Director, Dr. Pierangela Veneziani, my parents, Fran and Paul, and my brother, Jason, for continuously encouraging and supporting me throughout my four years here at Brockport.

Introduction and Problem Description

We have analyzed enrollment patterns at the College at Brockport, State University of New York, between 2008 and 2013. Data from over 7,500 freshman students including personal, academic, financial and family background information has been studied. Our analysis shows that while overall enrollment has remained constant, first-year retention rates are declining. The percentages of students attending mapped by SAT score and high-school GPA over time shows a shift in the composition of our freshman cohort. This shift is worse for clusters of students with the strongest academic credentials. The college has concentrated its efforts to improve enrollment rates through financial leveraging. We seek to support the College's efforts to identify the clusters of students on which to concentrate marketing and financial outreach by analyzing retention rates. The two prevalent models available in the literature modeling student retention are Tinto's Student Integration Model (1975, 1993) and Bean's Student Attrition Model (1980, 1983, 1990). Tinto theorizes that college environment, the quality of faculty-student interactions, and students' social integration into the school – in essence the degree of a school's commitment to students – are key factors in retaining students. Bean (1983) theorizes that the students' perspective on their academic experience is central to their persistence and recognizes an element missing from Tinto's analysis: the impact on retention of factors external to an institution. Because the purpose of our analysis is to guide the College in its marketing efforts to accepted students, we evaluate pre-enrollment variables as predictors of one-year retention of first-time students. Information about family background (parental education and socioeconomics),

individual attributes (academic ability, race and gender), characteristics of the student's high school, and high school academic records are incorporated in our model.

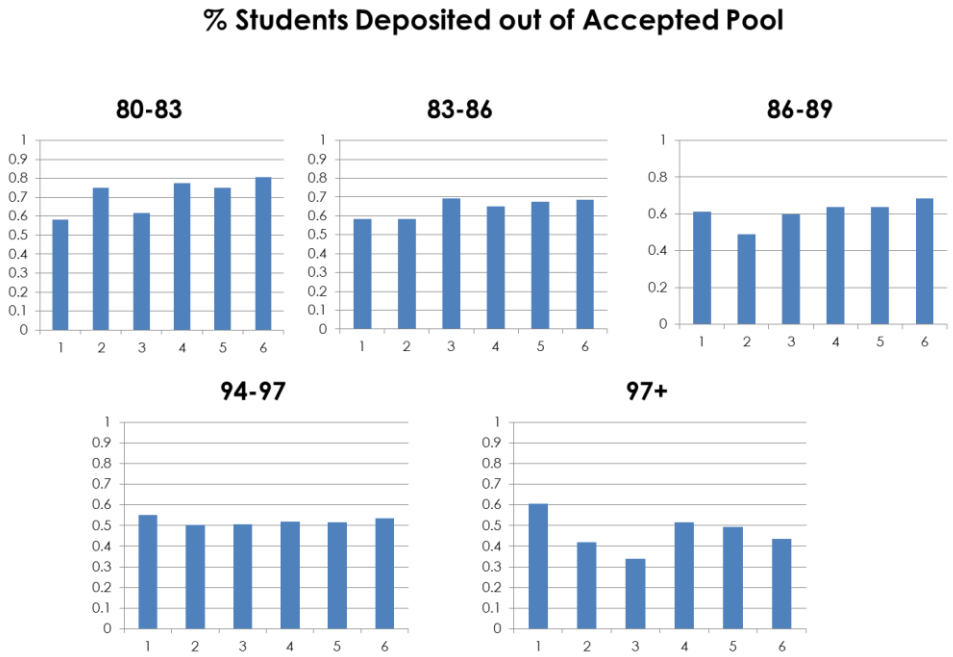
Upon completion of the analyses, we can made a few conclusions regarding predictions of the first year cumulative GPA of a randomly chosen freshman, the retention of a freshman through to sophomore year and the decision of a high school senior to enroll with the College or not.

Problem Solution

The goal of my thesis is to study and predict enrollment patterns based on academic, demographic and personal characteristics of entering freshman cohort.

The preliminary descriptive statistics of this work displayed students were clustered according to High School GPA. In other words, Admissions personnel at the College at Brockport were initially dependent on these two parameters alone in order to cluster student applicants. This is important because these were the sole two characteristics used to award merit scholarships to incoming freshman as incentive to enroll. Enrollment statistics (% of student coming to Brockport out of the number of students accepted) were computed for the last 5 years.

The following column graphs display samples illustrating the shift of freshman enrollment trends from students with higher high school GPA's (high nineties) to students with lower high school GPA's (low eighties). The horizontal axis represents in the years 2008 through 2013 while the vertical axis represents the percentage of accepted freshman who decided to enroll with the College.



I then used MINITAB statistical software to perform a regression analysis using the provided data. Personal, academic, and geographic data was gathered about entering freshmen for the past three academic years. The first goal is to identify which of these variables have a significant effect on the GPA of a student at the end of his/her freshman year. Regression analysis lets us see how multiple factors affect an outcome. There was a plentiful amount of information provided about each student, which included:

Academic predictors:

- ACTC: Composite ACT score used for admissions decisions
- ADMIT TYPE: Admission type of student
- AP Credit: Number of earned Advanced Placement credits
- CEEB: The College Board six-digit high school code for the high school the student graduated from
- College Credit: Number of earned College credits
- GPA: Raw cumulative high school GPA from high school transcript
- Housing: Answers whether student lives on or off campus
- MAJ1/CONC: Student's first choice major/concentration
- MAJ2/CONC: Student's second choice major/concentration
- PCT Rank: High school rank percentage
- SAT Total: Verbal SAT score plus Math SAT score
- TIER: The academic tier group classification for the student

Financial predictors:

- EFC: Expected Family Contribution based on Federal Methodology

- EOP APP: Student applied with the Educational Opportunity Application (EOP)
- FAM INC: Family income as reported on the FAFSA
- Scholarships: Amount of scholarship
- Total Aid: Amount of total financial aid package including scholarships, grants (e.g., Pell, TAP, etc.), and loans
- Unmet Need: The amount of remaining financial need the student has to pay for college after grants and scholarships have been accounted for

Personal predictors:

- Age: Age of student
- First Gen: Whether the student is a first generation student or not
- Gender
- Hispanic: Answer to the question on the SUNY application "Are you Hispanic/Latino? “
- Race Code: Race of student
- ZIP: Student’s home zip code

First, we are interested in predicting the first year GPA of students prior to their enrollment because we would rather enroll students who have a higher probability of succeeding in school thus leading to a higher retention rate. We need to determine which predictors, among the available ones, best contribute to the first year GPA of incoming freshmen. When the number of predictors is large, we need to identify variables that are highly collinear, which can make one of the variables almost redundant in some cases. To understand whether any one variable is correlated to another we could use standard

correlation analysis, which in this case only shows that the verbal, math, and essay SAT scores are highly correlated (as expected). Due to the large number of predictors, we can run Principal Component Analysis to order the components from most to least significant. Whether we aggregate the SAT scores or not, a student's high school rank and the number of AP credits taken were found to be the principal components in our data. We further study which predictors to include in the best model by carrying out variable selection by best subset regression. Among all models output by best subsets, we found that this technique produced the following model to, in fact, be the best one which includes the first two principal components of our data.

Regression Analysis: (Y) first_year_c versus AP CREDIT, SAT TOTAL, ...

```
The regression equation is
(Y) first_year_cum_gpa = 1.63 + 0.0196 AP CREDIT + 0.000185 SAT TOTAL
                         - 0.000025 UNMET NEED* + 0.0154 PCT RANK


1920 cases used, 727 cases contain missing values


Predictor           Coef     SE Coef      T      P
Constant         1.63452     0.08695  18.80  0.000
AP CREDIT       0.019643    0.003716   5.29  0.000
SAT TOTAL     0.00018477  0.00006421   2.88  0.004
UNMET NEED*  -0.00002549  0.00000398  -6.40  0.000
PCT RANK       0.0154458   0.0008885  17.38  0.000


S = 0.640596   R-Sq = 22.9%   R-Sq(adj) = 22.8%


Analysis of Variance

Source            DF        SS      MS       F      P
Regression         4   234.001  58.500  142.56  0.000
Residual Error  1915   785.845   0.410
Total           1919  1019.846
```

The analysis shows that unmet need, high school ranking, number of AP credits taken in high school, and SAT total score are all significant contributors to first year GPA

and account for approximately 1/4 of the variation in a student's freshman GPA (R-Square = .228). The R-Square statistic tells us what percentage of the variation in the data (predictor variables) is explained by this model. Once again, the following model for First Year Cumulative GPA was obtained.

First Year Cum GPA =1.63 + 0.0196 AP CREDIT + 0.000185 SAT TOTAL - 0.000025 UNMET NEED + 0.0154 PCT RANK

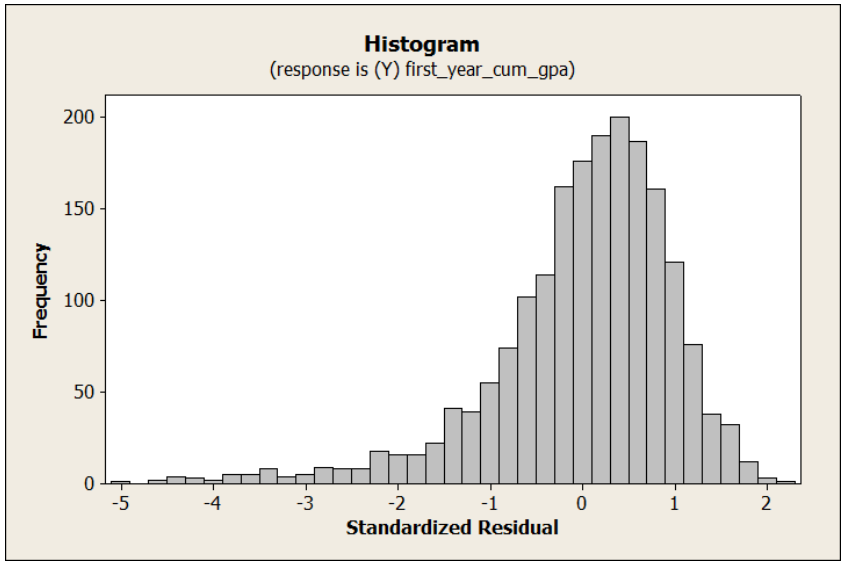
Furthermore, the model helps identify the fact that AP credit, SAT total and high school percentile rank contribute positively to a student's first year cumulative GPA. More specifically, for a one unit increase in AP credits, the first year GPA increases by 0.0196; for a one unit increase in SAT total, the first year GPA increases by 0.000185; for a one percent increase in high school percentile rank, the first year GPA increases 0.0154. Conversely, the amount of unmet need of the student contributes negatively to a student's first year cumulative GPA. More specifically, for a one unit increase in amount of unmet need of the student, the first year GPA decreases 0.000025.

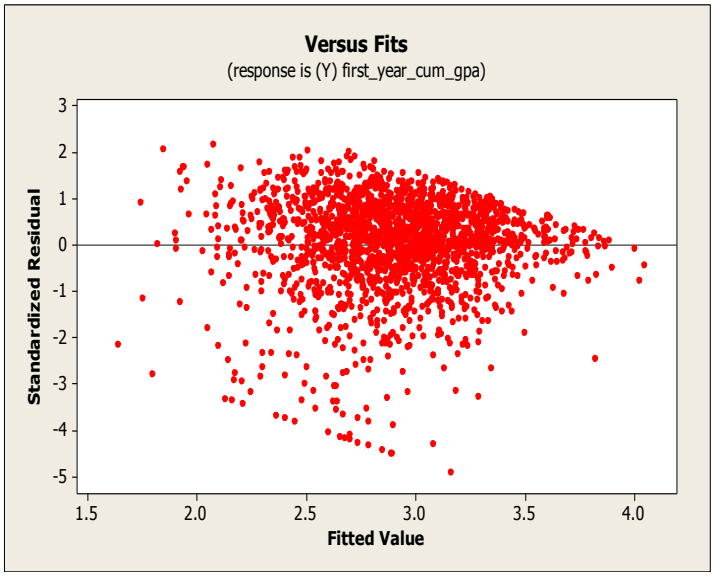
As an illustration of the model, consider the following example: the student has earned 3 AP credits, has an SAT total score of 1200, has $1,000 of unmet need and was placed in the 90^{th} percentile of his/her high school class. The model, therefore, predicts that this particular student will have a GPA of 3.27 during his/her first year at the College at Brockport, as shown below.

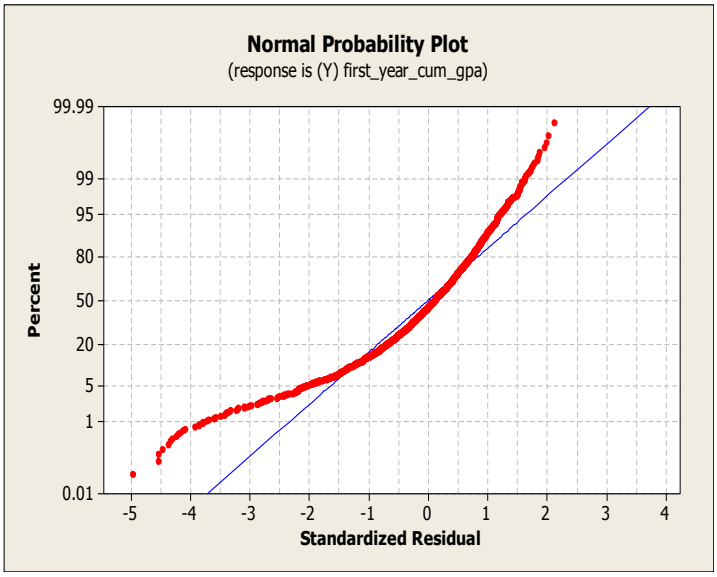
First Year Cum GPA =1.63 + 0.0196*3+0.000185*1200 - 0.000025*1000 + 0.0154*90

=3.27

Before we can make use of a model, we must first verify that the regression assumptions have been met. The following plots represent the distribution and variance of the residuals i.e. error terms associated with this model. The first image is a histogram of standardized residuals which tells us that the error terms of this model follow a normal distribution with a central tendency to zero. In short, the normality assumption is met.







The image on the left is the Versus Fits plot of the standardized residuals which determines whether the residuals have a constant variance or not. The best scenario would be that all of the residual data points form a strictly unwavering horizontal band around mean zero. Unfortunately, the residuals in this image, as you can see, create a "funneling in" effect. Thus, we can't confidently say that the constant variance assumption is met.

Our analyses have produced results that are similar to the ones obtained at other institutions when using the same type of data. Surprisingly, neither age nor gender nor family income nor number of parents with a college degree had significance to first year GPA. Attempts to improve the model with quadratic terms and interaction terms did not produce success. To improve the fit of the data, other institutions have expanded their data sets to include:

- The number of hours worked by a student during the semester
- Whether a student is enrolled in a freshman seminar
- The grade obtained by a student in a set of core freshman classes (whether Math or English or both)

Next, we will model the likelihood of sophomore retention at the College at Brockport. This is of particular interest to university admission personnel because retention rates reflect on the image of the school as well as the fact that retaining students minimizes marketing costs among other expenses.

Multivariate Models to predict Retention have become popular as a cost-effective way to identify students who could benefit from targeted interventions.

Broadly applied efforts, such as targeting all freshmen, can be wasteful in that they include people who would have come back anyway. Narrowly applied efforts, such as targeting students on probation or on one or two other risk factors, fail to take advantage of the fact that multiple characteristics combine to create retention. Multivariate methods help identify students who are *most likely* to leave based on a variety of factors. We are interested in understanding the effect of our predictors upon the retention rate, formulated as the odds of a student being retained versus the odds of a student dropping out. First, we must determine which predictors, among the available ones, best forecast whether or not a freshman student will return for his/her sophomore year at the College. Because the dependent variable is binary (student drops out or not), a binary logistic regression analysis is appropriate.

Binary Logistic Regression: Y retained_s versus HISPANIC, PCT RANK, ...

```
Response Information

Variable         Value         Count
Y retained_soph  Retained        654  (Event)
                 Not Retained    124
                 Total           778
Logistic Regression Table

                                                        Odds     95% CI
Predictor          Coef    SE Coef      Z      P  Ratio  Lower  Upper
Constant        2.93275   0.901703   3.25  0.001
HISPANIC
 Yes           -1.58514   0.528223  -3.00  0.003   0.20   0.07   0.58
PCT RANK       0.0131498  0.0064471   2.04  0.041   1.01   1.00   1.03
UNMET NEED*  -0.0001155  0.0000283  -4.08  0.000   1.00   1.00   1.00
ACTC         -0.0910456  0.0378075  -2.41  0.016   0.91   0.85   0.98
AP CREDIT     0.0913093  0.0371730   2.46  0.014   1.10   1.02   1.18
Log-Likelihood = -321.172
Test that all slopes are zero: G = 40.188, DF = 5, P-Value = 0.000
Goodness-of-Fit Tests

Method           Chi-Square   DF      P
Pearson             759.559  772  0.618
Deviance            642.344  772  1.000
Hosmer-Lemeshow       4.754    8  0.784

Table of Observed and Expected Frequencies:
(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)
                                     Group
Value            1     2     3     4     5     6     7     8     9    10  Total
Retained
  Obs           51    59    62    63    70    65    66    72    73    73    654
  Exp         49.7  59.9  63.1  64.9  66.3  66.6  68.5  69.7  71.1  74.2
Not Retained
  Obs           26    19    16    15     8    12    12     6     5     5    124
  Exp         27.3  18.1  14.9  13.1  11.7  10.4   9.5   8.3   6.9   3.8
Total           77    78    78    78    78    77    78    78    78    78    778
Measures of Association:
(Between the Response Variable and Predicted Probabilities)

Pairs       Number  Percent  Summary Measures
Concordant   53872     66.4  Somers' D              0.34
Discordant   26464     32.6  Goodman-Kruskal Gamma  0.34
Ties           760      0.9  Kendall's Tau-a        0.09
Total        81096    100.0
```

The MINITAB output above represents a binary model for predicting sophomore retention. Moreover, the model is proven significant since all of the Goodness-of-Fit Tests have a p-value which is greater than 0.05. The predictors in this model – namely Hispanic, percentile rank, unmet need, ACT score and AP credit – are all significant

since each of their individual p-values is less than 0.05. The fit of this model can be reflected through the Concordant Percent towards the bottom of the MINITAB output. This percent is comparable to the R-Squared of a linear regression model in that it represents the percentage of variation in the probability of retention that can be explained by this model. Since this model has a Concordant Percent of 66.4%, we are confident that this model can prove useful in predicting sophomore retention of our students. The model is shown explicitly below.

$$ln\left(\frac{Prob(retained)}{Prob(not\ retained)}\right) = +2.93275 - 1.58514*\text{HISPANIC} + 0.0131498*\text{PCT RANK} - 0.0001155*\text{UNMET NEED*} - 0.0910456*\text{ACTC} + 0.0913093*\text{AP CREDIT}$$

The model shows that the probability of retention is negatively correlated to a student declaring themselves to be Hispanic. The model also shows, however to a lesser extent, that the probability of retention is mildly negatively correlated to unmet need and ACT score. Conversely, the model shows that the probability of retention is positively correlated to the student's percentile rank in high school and the number of AP credits that the student has earned – which are measures of academic strength. The extent to which each predictor affects the probability of retention is reflected in the variable's coefficient which is presented in the MINITAB output. Larger coefficients signify a larger correlation whereas smaller coefficients signify a more mild correlation to retention.

The following example illustrates that the probability that a specific student will be retained with the College can be predicted and solved for by using this model.

Consider a student who is Hispanic (represented as a binary number 1), has been placed in the 75th percentile of his/her high school, has $2,000 of unmet need, has an ACT score of 16 and has earned no AP credits. The model produces a value equal to the natural log of the probability retained divided by the complement of probability retained. Thus, by taking the exponential of this equation and solving for the numerator, we have predicted that the probability that this student will be retained into his/her sophomore year is 65.63%, as shown below.

$$ln\left(\frac{Prob(retained)}{Prob(not\ retained)}\right) = +2.93275 - 1.58514*1 + 0.0131498*75 - 0.0001155*2000 - 0.0910456*16 + 0.0913093*0$$

$$= 0.6470$$

$$Prob(retained) = 0.6563$$

The following output represents a second binary model that is equally good in predicting sophomore retention.

Binary Logistic Regression: Y retained_soph versus logUNMET, AP CREDIT

Link Function: Logit

Response Information

Variable	Value	Count	
Y retained_soph	Retained	1310	(Event)
	Not Retained	327	
	Total	1637	

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	2.80772	0.297627	9.43	0.000			
logUNMET	-0.467528	0.0880635	-5.31	0.000	0.63	0.53	0.74
AP CREDIT	0.0545879	0.0204160	2.67	0.008	1.06	1.01	1.10

Log-Likelihood = -797.063

Test that all slopes are zero: G = 43.081, DF = 2, P-Value = 0.000

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	1440.00	1424	0.378
Deviance	1430.03	1424	0.450
Hosmer-Lemeshow	64.17	8	0.000

Table of Observed and Expected Frequencies:
(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

	Group									
Value	1	2	3	4	5	6	7	8	9	10	Total
Retained											
Obs	82	119	129	130	144	144	140	137	144	141	1310
Exp	117.1	121.0	123.6	125.2	127.8	133.0	133.2	136.8	141.3	151.0	
Not Retained											
Obs	81	45	35	33	20	23	23	26	19	22	327
Exp	45.9	43.0	40.4	37.8	36.2	34.0	29.8	26.2	21.7	12.0	
Total	163	164	164	163	164	167	163	163	163	163	1637

Measures of Association:
(Between the Response Variable and Predicted Probabilities)

Pairs	Number	Percent	Summary Measures	
Concordant	276519	64.6	Somers' D	0.30
Discordant	147511	34.4	Goodman-Kruskal Gamma	0.30
Ties	4340	1.0	Kendall's Tau-a	0.10
Total	428370	100.0		

This second model for predicting student retention is also considered adequately significant since the two out of three of the Goodness-of-Fit Tests have p-values much higher than 0.05. Also, the individual predictors, log unmet need and AP credit, have p-values less than 0.05 which indicates that these are significant contributors to retention. More specifically, the model shows that log unmet need (a transformation of the original unmet need) is negatively correlated to student retention whereas number of AP credits earned is positively related to student retention. The Concordant Percent for this model is 64.6% meaning that 64.6% of the variation in the probability of student retention can be predicted using this model. The model is shown explicitly below.

$$ln\left(\frac{Prob(retained)}{Prob(not\ retained)}\right) = +2.80772 - 0.467528*\text{logUNMET NEED} + 0.0545879*\text{AP CREDIT}$$

Similarly to the linear regression model above, we can solve for Prob (Retained) when given an arbitrary student's data. Consider for example a student with $3,000 of unmet need and 3 AP credits has a 31.61% probability of being retained with the College, as shown by the calculation below.

$$ln\left(\frac{Prob(retained)}{Prob(not\ retained)}\right) = +2.80772 - 0.467528*\log 3000 + 0.0545879*3$$

$$= -0.7717$$

$Prob(retained) = 0.3161$

Lastly, we will use binary regression analysis to predict whether or not an accepted student will enroll with the College or not. This model can be useful in the fact that financial incentives can be used to target students who are "on the fence" about their decision to enroll with our university. The ability to predict whether or not a student will enroll or not before it actually happens would be beneficial because the better (and reasonably attainable) students can be predominantly targeted.

First, we determine which predictors, among the available ones, best forecast whether or not an accepted freshman student will deposit and enroll with the College.

Binary Logistic Regression: DEC DESC versus sqrtSAT, AP CREDIT

```
Link Function: Logit

Response Information

Variable  Value  Count
DEC DESC  1       2466  (Event)
          0         96
          Total   2562

Logistic Regression Table
                                                    Odds     95% CI
Predictor      Coef    SE Coef      Z      P  Ratio  Lower  Upper
Constant   -7.81199    1.88603  -4.14  0.000
sqrtSAT    0.342318  0.0606722   5.64  0.000   1.41   1.25   1.59
AP CREDIT  0.457322   0.175492   2.61  0.009   1.58   1.12   2.23

Log-Likelihood = -372.298
Test that all slopes are zero: G = 74.326, DF = 2, P-Value = 0.000

Goodness-of-Fit Tests

Method           Chi-Square   DF      P
Pearson             251.330  352  1.000
Deviance            107.058  352  1.000
Hosmer-Lemeshow      28.624    8  0.000
```

Table of Observed and Expected Frequencies:
(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

				Group						
Value	1	2	3	4	5	6	7	8	9	10
1										
Obs	248	300	307	274	277	249	246	259	256	50
Exp	256.8	292.7	300.8	273.8	273.2	249.8	254.2	258.7	255.9	50.0
0										
Obs	39	12	9	11	5	7	12	1	0	0
Exp	30.2	19.3	15.2	11.2	8.8	6.2	3.8	1.3	0.1	0.0
Total	287	312	316	285	282	256	258	260	256	50

Value	Total
1	
Obs	2466
Exp	
0	
Obs	96
Exp	
Total	2562

Measures of Association:
(Between the Response Variable and Predicted Probabilities)

Pairs	Number	Percent	Summary Measures	
Concordant	168273	71.1	Somers' D	0.44
Discordant	63224	26.7	Goodman-Kruskal Gamma	0.45
Ties	5239	2.2	Kendall's Tau-a	0.03
Total	236736	100.0		

The above output displays a binary model that predicts the probability that an accepted student will enroll with the College at Brockport. The model is a good fit since the two out of the three Goodness-of-Fit Tests have p-values greater than 0.05. The individual predictors, the square root of SAT total score and AP credit, are significant contributors to the model since their p-values are lower than 0.05. The overall model is useful in that the Concordant Percent is 71.1% meaning 71.1% of the variation in probability of an arbitrary student's decision to enroll is defined by this model. The model is explicitly shown below.

$$ln\left(\frac{Prob(deposited)}{Prob(not\ deposited)}\right) = -7.81199 + 0.342318*\text{sqrt SAT} + 0.457322*\text{AP CREDIT}$$

The above model shows that a positive correlation exists between SAT score and student enrollment as well as between AP credits earned and student enrollment. The Prob(Deposited) can be solved for similarly to the way described previously for the Retention model. The below example illustrates that a student with an SAT total score of 1200 and 6 AP credits has a probability of 99.89% of depositing with the College.

$$ln\left(\frac{Prob(deposited)}{Prob(not\ deposited)}\right) = -\ 7.81199\ + 0.342318*\text{sqrt } 1200 + 0.457322*6$$

$$= 6.7902$$

$Prob(deposited)=0.9989$

Results and Final Remarks

The models we have produced to predict first year cumulative GPA, probability of retention and probability that a student will deposit with the College are less concise than we would have hoped. However, these results are in-line with other studies done at various universities around the country. In essence, it is difficult to predict human behavior since it can be quite random. Also, many factors that may help contribute to our models are unmeasurable such as how well the student copes with stress and other personal considerations.

In conclusion, the lack of concise results was not due to a lack of information available but a lack of the most relevant information available. Transforming the data (both predictors and the response), including interaction terms and/or quadratic terms did not further improve the fit of the models. Ultimately, student behavior and student performance is difficult to predict quantitatively when there are infinite variables, both measurable and unmeasurable, that should be considered. Even then, the accuracy of the model may be at question.