Logistic Regression Analysis of Mont St. Michel College

Date: May 5th, 2024
To: Professor Callaghan
From: Pranavi Chinthireddy, Nandini Koppunuru, Temi Oyefeso, Sheetal Padmanabhan, Ishani Parkar
Subject: Analysis and prediction of dropout rates at Mont St. Michel

Descriptive Analysis

The Student Support Office at Mont St. Michel, a local community college, has identified a concerning trend in student retention rates. Over the past few years, there has been an increase in students leaving Mont St. Michel College at the end of their first year. In response to this issue, the college implemented voluntary one-credit seminars aimed at helping first-year students establish campus connections. These seminars' effectiveness in improving student retention is under scrutiny, with the college seeking evidence to support this initiative's continuation.

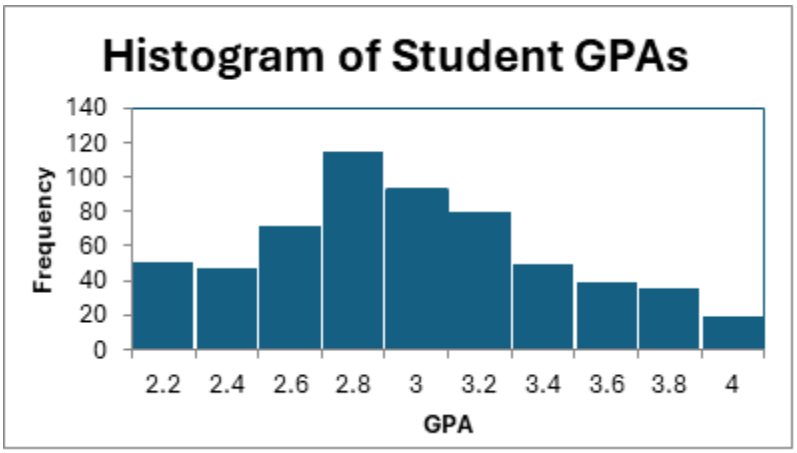
To remedy the scrutiny they have been under, a comprehensive analysis of data from 600 sample students from the previous academic year was conducted. The variables included in this analysis are high school GPA, seminar enrollment, units enrolled, student status (full-time or part-time), demographics (gender, age, commuter status), and dropout status. The Student Support Office will use descriptive statistics and data visualization, evaluate predictive models, and recommend future initiatives to tackle retention based on the analysis results. They will use logistic regression models to predict seminar enrollment and student dropout probabilities.

In the sample collected, there were 8 variables. 3 of those variables were continuous and the other 5 variables were dichotomous. The 3 continuous variables included High School GPA, Units Enrolled, and Age. The 5 dichotomous variables were Seminar Attendance, Enrollment Status, Dropout Rate, Gender, and Age.

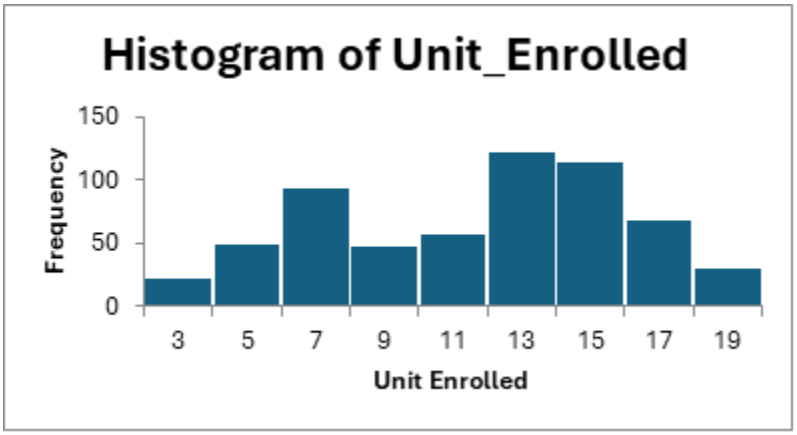
Statistic	HS_GPA	Unit_Enrolled	Age
Mean	2.875	11.190	21.458
Standard Error	0.019	0.174	0.194
Median	2.830	12.000	19.000
Mode	2.620	13.000	19.000
Standard Deviation	0.466	4.252	4.760
Sample Variance	0.217	18.077	22.656
Kurtosis	-0.516	-1.042	4.039
Skewness	0.299	-0.279	2.114
Range	1.990	16.000	24.000
Minimum	2.010	3.000	18.000
Maximum	4.000	19.000	42.000
Sum	1724.790	6714.000	12875.000
Count	600.000	600.000	600.000
Confidence Level (90.0%)	0.031	0.286	0.320

To understand the data, the team utilized descriptive statistics based on continuous variables to describe the essential features of the dataset. Some important insights can be drawn based on the sample of 600 students. The average GPA for these students is 2.88 with a variance of 0.22, meaning that on average the students in this obtain a B- GPA. There is also a wide range between the max and min GPA. The highest GPA earned at Mont St. Michel is a 4.0 and the lowest is a 2.01, indicating there is a bit of dispersion. Secondly, we see that the average unit enrolled is 11.19, which can be attributed to the combination of part-time and full-time students. The last data summary provides demographic data on the age of the sample participants. We see that the average age of participants is 21.46, indicating that most participants are young. The youngest

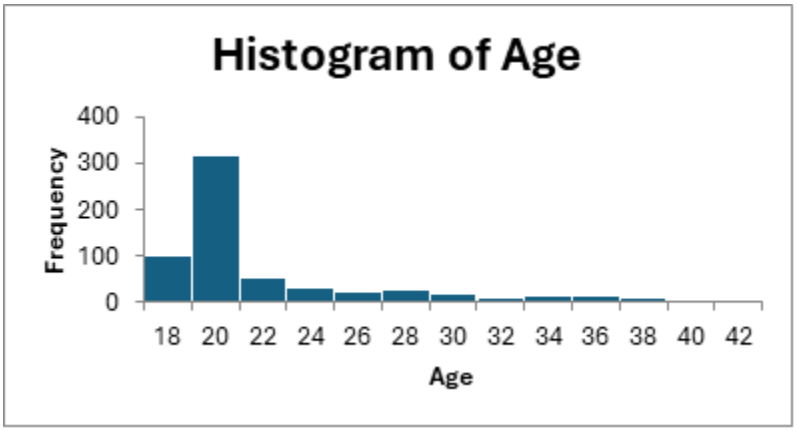
student in the same is 18 and the oldest is 42, indicating that some people followed the more traditional path of going to college after high school while some are return students.



To visualize this data, we utilized the histograms to understand central tendency and distribution. Our analysis of the Student GPA histogram indicated a mostly normal distribution. This conclusion was affirmed by the proximity of the mode, median, and mean in the data summary table. We saw that most students fall between the range of (2.8,3.2). Additionally, 28% of students fall in the lower GPA range of (2.2,2.6) and 15% fall in the upper range of (3.6,4.0).

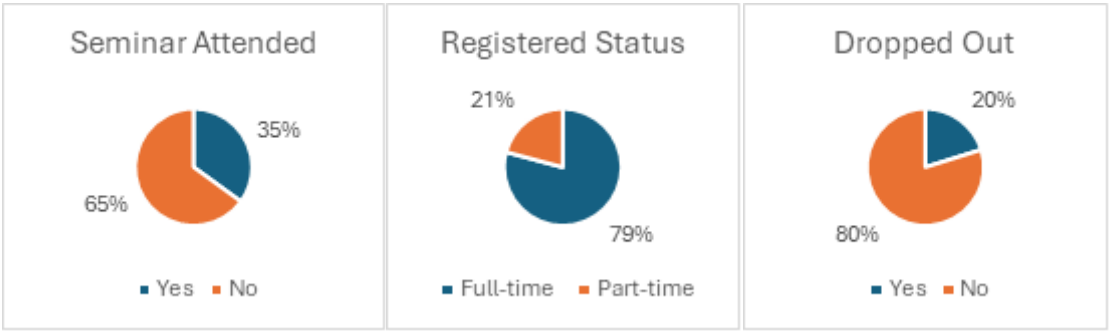


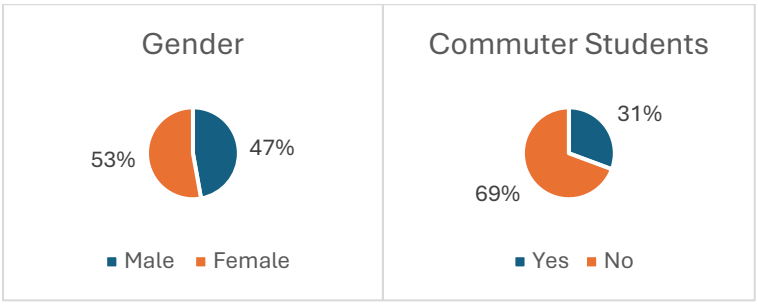
This histogram depicting units enrolled, is left-skewed. Indicating that the mean is less than median. This means that most of the students are enrolling for higher number of units, and more students are full-time than part-time. 11.8% of students are taking between (3,6) units and 39% of students are taking between (13,16) units.



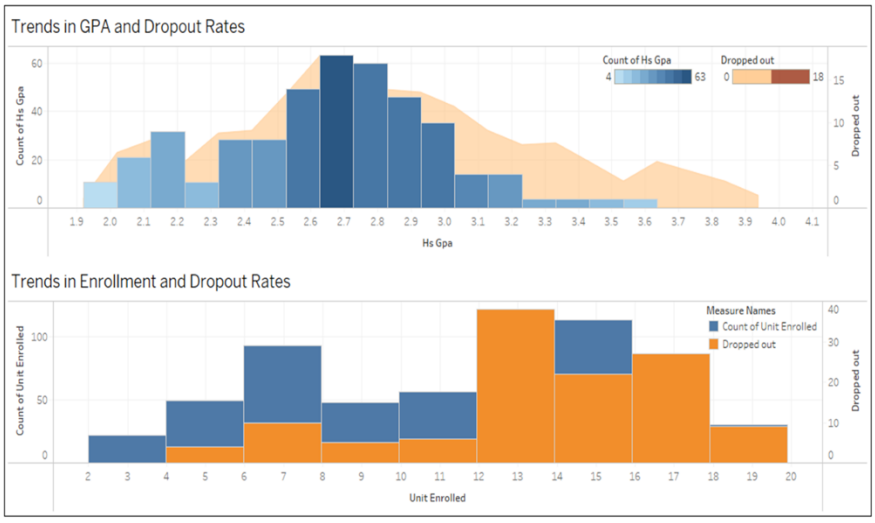
According to the histogram of age, the distribution is heavily skewed to the right. Therefore, the median is less than the mean and there are much more younger students than older. Approximately 1% of students are within the ages of (40,43) and 69% of students are between the ages of (18,21).

The following pie charts provide an overview for the dichotomous variables: Seminar Attended, Registered Status, Dropped Out, Gender, and Commuter Students.

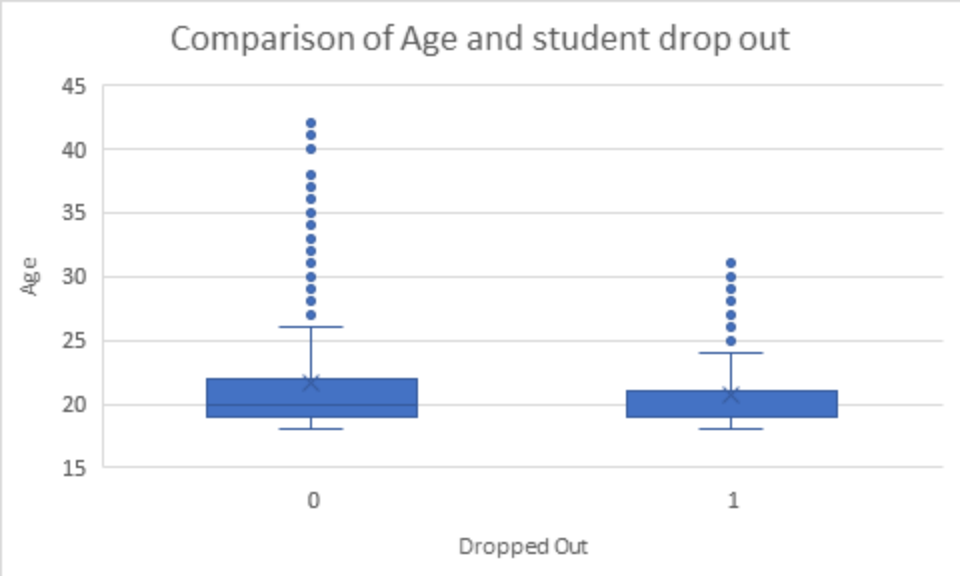




During a student's first year, they have the option to attend an hour-long seminar to establish a connection on campus. The attendance of this seminar is quite low and only 35% of the first-year students attend. 79% of the students are full-time students and take about 12 units a semester. This is significantly greater than the students who attend part-time. Of all students 80% go on to receive a degree, however 20% of the students opt to drop out. This is quite high for a university, hence Mont St. Michel's concern. The demographic data includes gender, and we can see that at Mont St. Michel there are more female than male students. Finally, we can see that 31% of students are commuter students and 69% live close to campus indicating they will be more likely to be more involved with on-campus.



Descriptive statistics analysis reveals a notable trend indicating that students who enroll in multiple units concurrently might be more likely to drop out compared to those who enroll in a lower number of units. There might be a potential correlation between higher course load and dropout rates, attributed to student burnout.



Students who did not drop out (represented by 0 on the x-axis) tend to be much older on average, with ages ranging from around 25 to over 40 years old. The distribution of ages for non-dropouts is much wider and more spread out compared to dropouts. The distribution of ages for dropouts is tight and concentrated. The graph shows that most of the dropped-out participants fall under the age group less than 25, observation is also backed up by the Histogram of Age mentioned above.

Model to predict Seminar Enrollment

Classification Table[a]

			Predicted		
			Seminar_Attended		Percentage Correct
	Observed		0	1	
Step 1	Seminar_Attended	0	371	19	95.1
		1	176	34	16.2
	Overall Percentage				67.5

a. The cut value is .500

Figure 1

Classification Table[a,b]

			Predicted		
			Seminar_Attended		Percentage Correct
	Observed		0	1	
Step 0	Seminar_Attended	0	390	0	100.0
		1	210	0	.0
	Overall Percentage				65.0

a. Constant is included in the model.
b. The cut value is .500

Figure 2

There is an improvement in the overall percentage of the model from Block 0 (Figure 1) to Block 1 (Figure 2) as seen in the classification table, hence we can say that we can consider the model to predict our output after examining other factors.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	25.082	6	<.001
	Block	25.082	6	<.001
	Model	25.082	6	<.001

Figure 3

From the Omnibus test (goodness of fit test) we can say that the model is significant since the p-value is less than alpha which we have set as 0.1.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	13.297	8	.102

Figure 4

From the Hosmer and Lemeshow Test (poor fit test) we can say that it is insignificant that is the p-value is greater than

0.1, hence confirming that the model is a good fit. This output supports the goodness of fit test.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	751.854[a]	.041	.056

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

From the model summary we can say that the existing independent variables used in the model can predict the correct probability for the independent variable between 4.1% to 5.6%.

Figure 5

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	90% C.I.for EXP(B) Lower	Upper
Step 1[a]	HS_GPA	-.295	.196	2.272	1	.132	.744	.539	1.027
	Unit_Enrolled	.032	.030	1.203	1	.273	1.033	.984	1.084
	Age	.048	.018	7.309	1	.007	1.050	1.019	1.081
	Gender_F(1)	-.356	.176	4.091	1	.043	.701	.525	.936
	Registered_Full_Time(1)	-.036	.314	.013	1	.909	.965	.575	1.618
	Commuter_Student(1)	.675	.193	12.230	1	<.001	1.963	1.430	2.697
	Constant	-1.189	.755	2.477	1	.116	.305		

a. Variable(s) entered on step 1: HS_GPA, Unit_Enrolled, Age, Gender_F, Registered_Full_Time, Commuter_Student.

Figure 6

From the Variables table we can see that the High school GPA, Units enrolled, and the Registered status variables are insignificant in predicting whether the student will enroll for the seminar or not since they have a p-value greater than 0.1. Only the Age, Gender and Commuter status variables are significant in predicting the seminar enrollment since they have a p-value less than alpha (0.1). Out of the significant variables Age and Commuter status variables have a positive relationship with the dependent variable that is the elder participants are more likely to attend the seminar. Gender variable has a negative relationship with the dependent variable which means that females are less likely to attend the seminar.

The equation for the model to predict Seminar Enrollment is as follows:

$$\hat{p} = -1.189 - 0.295 * x_1 + 0.032 * x_2 + 0.048 * x_3 - 0.036 * x_4 + 0.0675 * x_5 - 0.356 * x_7$$

<u>Model to predict Drop-out status</u>

The team worked on 4 different models namely:

1. Model with all variables (High School GPA, Units Enrolled, Age, Registered Full Time, Commuter Student, Seminar Attended, and Gender).

Classification Table[a,b]

			Predicted		
			Dropped_out		Percentage Correct
	Observed		0	1	
Step 0	Dropped_out	0	479	0	100.0
		1	121	0	.0
	Overall Percentage				79.8

a. Constant is included in the model.

b. The cut value is .500

Figure 7

Classification Table[a]

			Predicted		
			Dropped_out		Percentage Correct
	Observed		0	1	
Step 1	Dropped_out	0	461	18	96.2
		1	101	20	16.5
	Overall Percentage				80.2

a. The cut value is .500

Figure 8

There is an improvement in the overall percentage of the model from Block 0 (Figure 7) to Block 1 (Figure 8) as seen in the classification table, hence we can say that we can consider the model to predict our output after examining other factors.

Variables in the Equation

								90% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	HS_GPA	-1.651	.290	32.371	1	<.001	.192	.119	.309
	Age	-.078	.028	8.078	1	.004	.925	.884	.967
	Gender_F(1)	.098	.226	.187	1	.666	1.102	.761	1.598
	Commuter_Student(1)	-.671	.277	5.847	1	.016	.511	.324	.807
	Unit_Enrolled	.105	.038	7.781	1	.005	1.110	1.044	1.181
	Registered_Full_Time(1)	1.491	.559	7.119	1	.008	4.442	1.771	11.136
	Seminar_Attended(1)	.585	.233	6.285	1	.012	1.794	1.223	2.633
	Constant	2.194	1.113	3.886	1	.049	8.970		

a. Variable(s) entered on step 1: HS_GPA, Age, Gender_F, Commuter_Student, Unit_Enrolled, Registered_Full_Time, Seminar_Attended.

Figure 9

When we see at all the variables, we can see that the Gender variable seems to be insignificant i.e. it has a p-value greater than alpha (0.1), hence a model by eliminating the Gender variable was created and then compared all the factors.

2. Model eliminating Gender variable.

Classification Table[a,b]

			Predicted		
			Dropped_out		Percentage Correct
	Observed		0	1	
Step 0	Dropped_out	0	479	0	100.0
		1	121	0	.0
	Overall Percentage				79.8

a. Constant is included in the model.

b. The cut value is .500

Figure 10

Classification Table[a]

			Predicted		
			Dropped_out		Percentage Correct
	Observed		0	1	
Step 1	Dropped_out	0	461	18	96.2
		1	101	20	16.5
	Overall Percentage				80.2

a. The cut value is .500

Figure 11

There is an improvement in the overall percentage of the model from Block 0 (Figure 10) to Block 1 (Figure 11) as seen in the classification table, hence we can say that the variables we used to build the model can be considered to predict output after examining other factors.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	103.777	7	<.001
	Block	103.777	7	<.001
	Model	103.777	7	<.001

Figure 12

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	103.589	6	<.001
	Block	103.589	6	<.001
	Model	103.589	6	<.001

Figure 13

Figure 12 and Figure 13 give us information about the goodness of fit test for model with all variables and model without the gender variable respectively. We can see there is not much difference in the Chi-square value and that the p-value for the model is less than alpha (0.1) hence proving the model is significant.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	499.469[a]	.159	.250

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Figure 14

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	499.656[a]	.159	.250

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Figure 15

From the model summary for the model with all variables (Figure 14) and model with no Gender variable (Figure 15), we can say that the existing independent variables used in the second model can predict the correct probability for the independent variable between 15.9% to 25% which is similar to the model with all variables.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	6.774	8	.561

Figure 16

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	9.857	8	.275

Figure 17

We see that the Chi-square has increased for the model with no gender variable (Figure 17) and the p-value has decreased. We still go ahead to consider the model with no gender variable as even though the p-value decreased it is greater than 0.1 making the poor fit test insignificant and supporting the goodness of fit test.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	90% C.I.for EXP(B) Lower	Upper
Step 1[a]	HS_GPA	-1.645	.289	32.368	1	<.001	.193	.120	.310
	Age	-.078	.027	7.959	1	.005	.925	.884	.968
	Commuter_Student(1)	-.665	.277	5.753	1	.016	.514	.326	.811
	Unit_Enrolled	.105	.038	7.803	1	.005	1.111	1.044	1.181
	Registered_Full_Time(1)	1.486	.558	7.082	1	.008	4.419	1.764	11.069
	Seminar_Attended(1)	.575	.232	6.132	1	.013	1.777	1.213	2.602
	Constant	2.219	1.111	3.988	1	.046	9.194		

a. Variable(s) entered on step 1: HS_GPA, Age, Commuter_Student, Unit_Enrolled, Registered_Full_Time, Seminar_Attended.

Figure 18

This is the variable table for the model with no gender variable and it shows that all the variables are now significant few having a negative relationship and few having a positive relationship with the dependent variable – dropped out. We consider this model to predict the Drop out status as all the tests have similar results to the model with all variables which proves that the Gender variable had no importance in the model. The variables Units Enrolled, Registered Full Time and Seminar Attended have a positive relationship that means students who are registered full time and/ or have enrolled in multiple units are more likely to drop out. Similarly, students who attend the seminar are more likely to drop out as well. High school GPA, Age and Commuter Student status have a negative relationship which means higher the GPA lower are the chances of dropping out, the older the student is lower are the chances of dropping out, and if the student is a commuter the chances of dropping out is lower.

The equation for the model to predict drop out status is as follows:

$$\hat{p} = 2.219 - 1.645 * x_1 + 0.105 * x_2 - 0.078 * x_3 + 1.486 * x_4 - 0.665 * x_5 + 0.575 * x_6$$

3. Model eliminating Gender and Registered Full Time or Gender and Units Enrolled variables.

	HS_GPA	*Unit_Enrolled*	*Age*	*Gender_F*	*Registered_Full_Time*	*Commuter_Student*	*Seminar_Attended*
HS_GPA	1.00	-	-	-	-	-	-
Unit_Enrolled	0.01	1.00	-	-	-	-	-
Age	0.12	0.04	1.00	-	-	-	-
Gender_F	0.02	0.00	0.03	1.00	-	-	-
Registered_Full_Time	0.05	0.72	0.01	0.02	1.00	-	-
Commuter_Student	0.25	0.03	0.03	0.04	0.03	1.00	-
Seminar_Attended	0.04	0.06	0.11	0.07	0.04	0.13	1.00

Figure 19

From the correlation matrix we found that the Units Enrolled and Registered Full Time variables are highly correlated. Hence, we tried out models by eliminating these variables one by one along with having the Gender variable eliminated.

The model built by eliminating the Gender and Units Enrolled variables showed that the overall percent from Block 0 (Figure 20) to Block 1 (Figure 21) deteriorated, which means the variables used to build the model are not good predictors.

Classification Table[a,b]

			Predicted		
			Dropped_out		Percentage Correct
	Observed		0	1	
Step 0	Dropped_out	0	479	0	100.0
		1	121	0	.0
	Overall Percentage				79.8

a. Constant is included in the model.

b. The cut value is .500

Figure 20

Classification Table[a]

			Predicted		
			Dropped_out		Percentage Correct
	Observed		0	1	
Step 1	Dropped_out	0	463	16	96.7
		1	110	11	9.1
	Overall Percentage				79.0

a. The cut value is .500

Figure 21

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	507.940[a]	.147	.232

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Figure 22

The model built by eliminating the Gender and Registered Full Time variables, the R squared value reduced from the range of 15.6% to 25% to 14.7% to 23.2%. Hence, we decided to choose the model by just eliminating the Gender variable.

Data Split on Final model

We chose Model 2 (model built by eliminating the gender variable).

Data Split (80%)

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	389.203[a]	.168	.266

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Figure 23

When we split the data and build the final model, we find out that the model stands true with all the factors improvising from the original model. The R squared goes up to 16.8% to 26.6%.

Data Split (20%)

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	95.431[a]	.232	.365

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

Figure 24

When we split the data to test on 20% of the data, we found all the factors improving but except the GPA and the Seminar enrollment variable all variables were insignificant. The R squared increased to 23.2% to 36.5%.

These statistics prove that the final model chosen is stable.

The final statistics for the model in terms of the ratio of data used are as follows:

	100% Data	80% Data	20% Data
Classification Table (Block 1)	80.20%	81%	84.40%
Omnibus Test	Significant	Significant	Significant
Hosmer Lemeshow Test	Insignificant	Insignificant	Insignificant
Model Summary	15.9% - 25%	16.8% - 26.6	23.2% - 36.5%

Data Input Table

		Seminar Attended	
		1	0
Units Enrolled	0	0.94	0.90
1	3	0.96	0.93
4	6	0.97	0.95
7	9	0.98	0.96
10	12	0.98	0.97
13	15	0.99	0.98
16	18	0.99	0.98

Figure 25

		Seminar Attended	
	High School GPA	1	0
A+, A	4	0.02	0.01
A-	3.7	0.04	0.02
B+	3.3	0.07	0.04
B	3	0.11	0.06
B-	2.7	0.16	0.10
C+	2.3	0.27	0.17
C	2	0.38	0.26
C-	1.7	0.50	0.36
D+	1.3	0.66	0.52
D	1	0.76	0.64
D-	0.7	0.84	0.74
F	0	0.94	0.90

Figure 26

From the above data input tables, we can infer that the seminar hasn't had major impact in retaining the students from dropping out, but the number of units enrolled, and the GPA has some major impacts in the probability of dropping out.

	Units Enrolled						
High School GPA	0	3	6	9	12	15	18
4	0.01	0.02	0.02	0.03	0.04	0.06	0.08
3.7	0.02	0.03	0.04	0.05	0.07	0.09	0.12
3.3	0.04	0.05	0.07	0.09	0.12	0.16	0.21
3	0.06	0.08	0.11	0.15	0.19	0.24	0.30
2.7	0.10	0.13	0.17	0.22	0.28	0.34	0.42
2.3	0.17	0.22	0.28	0.35	0.42	0.50	0.58
2	0.26	0.32	0.39	0.47	0.55	0.62	0.69
1.7	0.36	0.43	0.51	0.59	0.66	0.73	0.79
1.3	0.52	0.60	0.67	0.74	0.79	0.84	0.88
1	0.64	0.71	0.77	0.82	0.86	0.90	0.92
0.7	0.74	0.80	0.85	0.88	0.91	0.93	0.95
0	0.90	0.93	0.95	0.96	0.97	0.98	0.98

Figure 27

From the table in Figure 27 we can see that higher number of units enrolled and lower GPA play a major role in determining the drop out probability of a student.

Recommendations

Based on the results from the logistic regression and classification data analysis, it appears that the traditional focus on seminar attendance as a universal method to improve student retention may need to be reconsidered. The data indicates that attending seminars may be linked to a higher likelihood of dropping out, especially among students with lower GPAs, as shown by a notably high odds ratio. This finding suggests that the college should shift away from broadly promoting seminar participation and instead adopt a more targeted approach. Focusing on personalized academic support, rather than indiscriminately advocating seminar attendance, could be more effective, particularly for older or advanced students who are capable of self-directed learning.

Here's are some suggestions that the administration can implement to reduce the drop out ratio:

- Instead of pushing seminars for everyone, the college can offer different types of help based on each student's situation as the reason can be varied including financial issues. Providing scholarship opportunities might help students who are willing to study but cannot afford it to continue their education.
- Students with lower grades, especially those taking a lot of classes, need more support. The college should offer them regular one-on-one meetings with advisors to help them succeed based on their individual requirements.
- The administration can suggest the implementation of constraints on the maximum number of course units that students can enroll in per semester. Imposing such constraints can mitigate the risk of student burnout and subsequently reduce dropout rates.
- This restriction can be varied based on students High School GPA, i.e. students below a certain threshold can only enroll to lower number of units which can be then increased based on their performance in college.
- Some students, especially older or more advanced ones, might not need seminars and would benefit more from other types of support. Trying to analyze some additional variables like disability of any form can be a factor to understand the drop out rate and the administration can work on improving their disability services to accommodate students and help them continue.