Data Analysis for Student Retention

Bilal Mohammed Khaja

Jasminder Singh

Queeny Neelu Mathayas

Sairaj Prakash Jadhav

Dr. Thuyuyen Newby

College Of Business and Economics   
Department of Information Systems and Decision Sciences

California State University, Fullerton

Table of Contents

[Introduction and Problem Background 2](#_Toc102915997)

[Data Description 2](#_Toc102915998)

[Maximum and Minimum 4](#_Toc102915999)

Understanding of variables………………………………………………………………. 5

[Analysis 5](#_Toc102916000)

[Data analysis for the students who will enroll in the seminar. 5](#_Toc102916001)

[Data analysis of the students who will drop out of college after the first year 6](#_Toc102916002)

[Split observations analysis 7](#_Toc102916003)

[Input Tables to solve seminar retention problem 9](#_Toc102916004)

[1. SEMINAR vs. GPA 9](#_Toc102916005)

[2. SEMINAR vs. UNITS ENROLLED 9](#_Toc102916006)

[3. COMMUTER vs. GPA 10](#_Toc102916007)

[4. COMMUTER vs. UNITS ENROLLED 10](#_Toc102916008)

[Recommendations 10](#_Toc102916009)

# **Introduction and Problem Background**

This paper is about students’ retention. Due to the high number of students who drop out at the end of their first year of college, the college provides hour-long seminars with their respective faculty to increase retention. According to the Student Support Administrator, this could also be due to a hectic schedule or the number of units they enrolled in in their first year.

In this paper, we selected 600 randomly selected students' data from their previous academic year to test whether the seminars increased retention rate and investigate all the factors contributing to student dropout.

# **Data Description**

We conducted descriptive analytics with a 90 percent confidence level to understand our data statistically before moving ahead with analyzing the data. This helped us understand the range, variance, mean, mode, sum, count, standard deviation, and skewness of the data. Below is the image outlining the descriptive stats of the data and the pivot tables used to understand the relationship between data.

**Figure 1**

**Descriptive Statistics**

Table

Description automatically generated

**Figure 2**

**Drop Out Vs. Gender**



**Figure 3**

**Drop Out Vs. Age**

****

**Figure 4**

**Drop out Vs. Units Enrolled**

****

**Figure 5**

**Drop out Vs. Commuter**

****

**Figure 6**

**Drop out Vs. Register Status**

****

**Figure 7**

**Drop out Vs. Seminar Attendance**

****

**Figure 8**

**Drop out Vs. GPA**

****

**Mean**

We derived the average values for variables provided from 600 randomly selected students’ data. The mean can also be evaluated by adding all the data points divided by the number of records in that field—the average GPA of 600 students from descriptive statics yields to 2.87465 or 2.88. The average unit enrolled by the students is 12.25833333 or 12. 26. Average age yield to 21.44833333 or 21.45. The average seminar attended yielded to 0.32 or 32 percent. The average dropped out yield to 0.208333333 or 20 percent. Average gender male yield to 0.471666667 or 47 percent. Average gender of female yields to 53 percent indicated that more female students enrolled in the college program than male. Average full-time students yield to 0.83333333 or 83 percent which indicates part time students are at 17 percent. Average students who commute yield to 0.323333333 or 32 percent which indicates that about 68 percent of students do not commute to attend classes.

# **Maximum and Minimum**

The maximum GPA among 600 randomly selected student data is 4 and minimum is 2.01. Maximum units enrolled is 18 units and minimum units enrolled is 4. The maximum age is 43 and minimum age 18.

**Understanding the Variables**

To understand the variables in a better way, we used pivot tables to draw out statistics to determine the relationship between them. As can be seen in image above, we created multiple tables and the below are some of the important observations we made:

* Dropped out Students Vs, Gender – The total number of students that dropped out is 125 but the number of females that dropped out is higher than that of the number of males (67:58).
* Age Vs. Dropped out Students – The average rate of dropout students is higher for students aged between 18-20 and lowest for students aged between 33-43. The highest number of students who dropped out are of age 19.
* Dropped out Students vs GPA – The higher the GPA, lower the dropped-out rate. Dropped out rate is higher for students with lower than 3 GPA, between 3-3.55 GPA the dropped out is lower, whereas the dropped-out rate is 0 for students above 3.55 GPA. The average GPA of students dropping out is 2.67 whereas average age of students not dropping out is 2.93
* Dropped out Students Vs. Units Enrollment – Students enrolling in more units have a higher rate of dropping out. The average number of units enrolled by dropouts is 13.67 whereas the average number of units enrolled by not dropouts is 11.89.
* Dropped out student's Vs. Commuter status – Non-commuter students are dropping out more than commuter students with the count of non-commuter students dropping out being 101 as compared to commuter students being 24
* Dropped out students Vs. Register Status – Full-time students are dropping out more than part-time students with the count of full-time students dropping out being 106 as compared to part-time students being 19.
* Dropped out students Vs. Seminar Attendance – Seminar not attending students are dropping out more than seminar attending students with the count of seminar not attending students dropping out being 82 as compared to seminar attending students being 43.

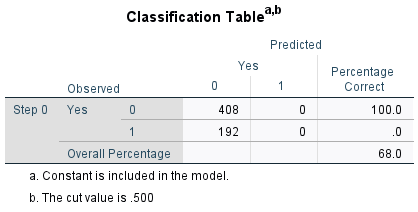
# **Analysis**

# **Data analysis for the students who will enroll in the seminar.**

The logistic regression equation relating to the predictive performance of the probability that students will enroll in the seminar is P-hat = 1/ (1+ exp (-(-0.940 -0.107\*HS\_GPA +0.013\*units enrolled +0.058\*Age -0.308\*Gender -0.435\*Full-Time -1.068\*Commuter))).

**Figure 5**

**SPSS: Classification Table**

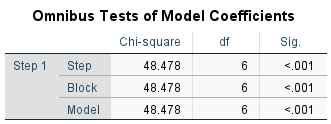
Table

Description automatically generated

From the classification tables, it can be seen how the accuracy increases from 68% to 69.7% (1.7% increase from Block 0 to Block 1), not a very high significant value.

**Figure 6**

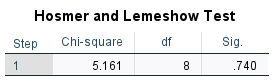
**SPSS: Omnibus Tests of Model Coefficients**



The goodness of fit test appears to be highly significant.

**Figure 7**

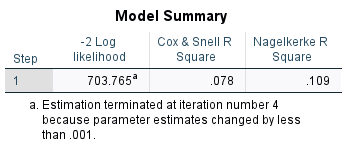
**SPSS: Hosmer and Lemeshow Test**



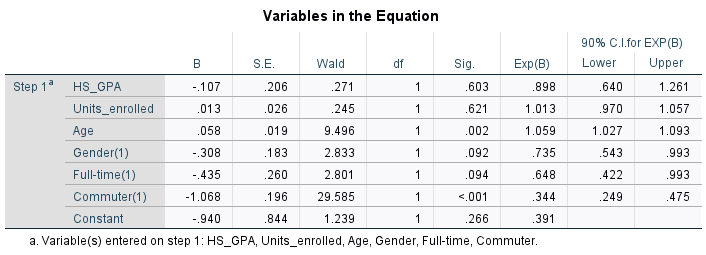
The poor fit fails, telling us that the values are insignificant **(0.740)**

**Figure 8**

**SPSS: Model Summary**



The pseudo- R, which is variance is suggested to be between **0.078 to 0.109**



For this model, students age, gender, register status and commuter status have significant relationship whereas students GPA and numbers of units enrolled have insignificant relationship.

Here, the number of units enrolled, and the student’s age will have an impact on enrollment in the seminar as the odd’s ratio is more than 1 . Students' GPA will not have an impact on enrollment in seminar because the odds ratio is less than 1. The odd ratio also tells us that females, non-commuters, and part-time students are more likely to enroll in the seminar.

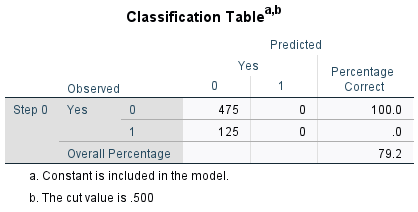
# **Data analysis of the students who will drop out of college after the first year**

The logistic regression equation relating to the predictive performance of the probability that students will drop out of college after the first year is

P-hat = 1/(1+ exp (-(-0.385 -1.502\*HS\_GPA +0.206\*Units enrolled +0.023\*Age -0.092\*Gender +0.355\*Full-time -1.017\*Commuter +0.316\*Seminar))).

**Figure 9**

**SPSS: Classification Table**

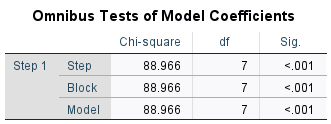
Table

Description automatically generated

From the classification tables, it can be seen how the accuracy increases from 79.2% to 80.2% (1.0% increase from Block 0 to Block 1), not a very high significant value.

**Figure 10**

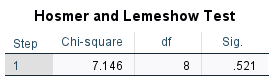
**SPSS: Omnibus Tests of Model Coefficients**



The goodness of fit test appears to be highly significant.

**Figure 11**

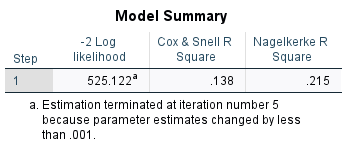
**SPSS: Hosmer and Lemeshow Test**



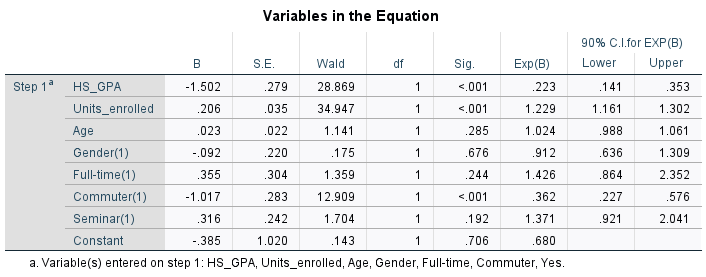
The poor fit fails, telling us that the values are insignificant **(0.521)**

**Figure 12**

**SPSS: Model Summary**



The pseudo-R, which is variance is suggested to be between **0.138 to 0.215**



For this model, students GPA, units' enrollment, and commuter status have significant relationship whereas students age, gender, register status and enrollment in seminar have insignificant relationship.

Here, the number of units enrolled, and the student’s age will have an impact on the drop-out status of the students as the odd’s ratio is more than 1. Students' GPA will not have an impact on drop-out status of the students because the odds ratio is less than 1. The odd ratio also tells us that females, non-commuters, full-time students and seminar attending students are more likely to drop-out.

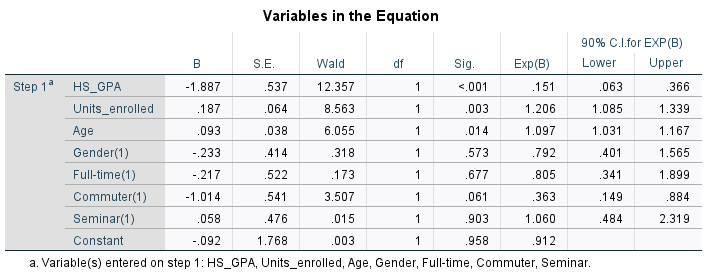
# **Split observations analysis**

When you split the data and compare the 600 samples to 200 observations and 400 observations, the following are the overall results:

The results for 200 observations are as follows:

**Figure 12**

**SPSS: Variables in the Equation (200 observations)**



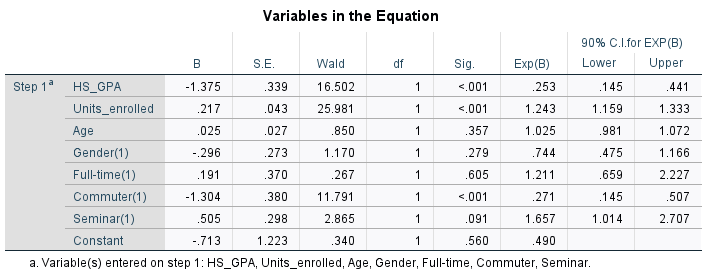
There is an overall percentage increase from 81% to 85% (Block 0 to Block 1), when we compare this percentage with 600 observations, it is an increase from 79.2% to 80.2% (1.0% increase from Block 0 to Block 1), a good significant increase from 1.0% to 4.0% for 200 observations.

For 600 observations, a student’s age is insignificant, whereas, for 200 observations, students' age has a significant relationship. Here, the odds ratio tells that the part-time students are more likely to drop out whereas, for 600 observations, full-time students are more likely to drop out

The results for 400 observations are as follows:

**Figure 13**

**SPSS: Variables in the Equation (400 observations)**



There is an overall percentage increase from 79% to 80.8% (Block 0 to Block 1), when we compare this percentage with 600 observations, it is an increase from 79.2% to 80.2% (1.0% increase from Block 0 to Block 1), a more significant change from 1.0% to 1.8%.

For 600 observations, student's enrollment in seminar is insignificant whereas for 400 observations, student's enrollment in seminar has significant relationship. The results for odds ratio for 600 observations and 400 observations are the same.

We would recommend using the Student Dropped out model rather than the seminar attended model as the overall percentage of the classification model is higher for the student dropped out model i.e., 80% whereas the percentage for Seminar attended model is ~ 69%.

Also, when we look at the model summary for both the models, we can see the R-squared values are much higher for the dropped-out model than the Seminar attended model.

Also, from the analysis of the 600, 400, and 200 observations of the drop-out model, we can conclude that the overall results are mostly similar, and the drop-out model can be used to evaluate the impact of different demographics on students dropping-out.

# **Input Tables to solve seminar retention problem**

Using the college dropout model, we have created four scenarios – SEMINAR vs. GPA,

SEMINAR vs. UNITS ENROLLED, COMMUTER vs. GPA, and COMMUTER vs. UNITS

ENROLLED to analyze the impact of variables on students dropping out.

We have completely discarded the Gender variable from all the four scenarios as it is an

insignificant variable, and the estimated p-hat value increases for all the scenarios on removing the Gender variable.

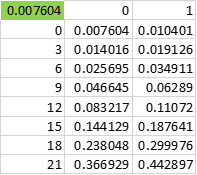
# **SEMINAR vs. GPA**

Table

Description automatically generated

The probability results show that GPA will have an overall impact on a student dropping out. Students with GPAs less than 1.7 will be the most liable to drop out, and those with GPAs of more than 3.0 will be the least liable to drop out.

# **SEMINAR vs. UNITS ENROLLED**



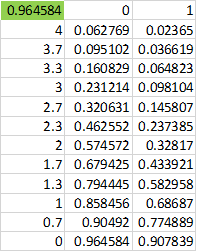
The probability results show that the number of units enrolled does not have much impact on students dropping out. Students enrolling in over 18 units will still be more susceptible to dropping out.

From the results above, we conclude that Seminar attendance will not have an overall impact on the retention of students, as students attending seminars are more liable to drop out than the students not attending seminars.

We believe that a commuter variable can also play a significant role in classifying the dropout model.

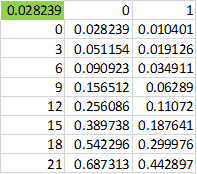
To analyze the results of commuter and non-commuter students' impact on the drop-out model, we have created two more 2-input tables.

# **COMMUTER vs. GPA**



The probability results show that the GPA will have an overall impact on a student dropping out. Students with GPAs less than 2 will be the most liable to drop out, and those with GPAs of more than 3.0 will be the least liable to drop out.

# **COMMUTER vs. UNITS ENROLLED**



The probability results show that the number of units enrolled does not have much impact on students dropping out. Students enrolling in over 18 units will still be more susceptible to dropping out.

From the results above, we conclude that the commuter variable Seminar will have an overall impact on the retention of students, as non-commuter students are more liable to drop out than the commuter students.

# **Recommendations**

We recommend the college stop enrolling students in seminars because students attending seminars are not being influenced by them, and the retention problem is not being resolved.

The college should instead focus on improving the number of commuter students or improving the involvement of non-commuter students in college activities. Results show that commuter students are less likely to drop out, and non-commuter students are more liable to drop out. The descriptive data indicates that there are 194 commuter students and 406 non-commuter students. The college can improve the number of commuter students by making more flexible schedules and providing an on-campus job to these students. This will help the college to resolve the student retention problem.

|  |
| --- |
|  |