A Logistic Regression Model For The Enhancement Of Student Retention: The Identification Of At-Risk Freshmen

Joseph G. Glynn, (glynn@canisius.edu), Canisius College
Paul L. Sauer, (sauer@canisius.edu), Canisius College
Thomas E. Miller, (tmiller@sa.usf.edu), University of South Florida

Abstract

A logistic regression model will be developed to provide early identification of freshmen at risk of attrition. The early identification is accomplished literally within a couple of weeks after freshman orientation. The dependent variable of interest is persistence, and it is a binary, nominal variable. Students who proceed from freshman matriculation to graduation without ever having dropped out are labeled persistors. Freshman matriculates who leave college either temporarily or permanently are classified as dropouts. The independent variables employed to predict attrition include demographics, high school experiences, and attitudes, opinions, and values as reported on a survey administered during freshman orientation. The model and its results will be presented along with a brief description of the institutional intervention program designed to enhance student persistence.

Introduction

A concern of many administrators is the ability to predict as early as possible the likelihood of a student dropping out of school. This concern is particularly critical in smaller, moderately to poorly funded private schools that depend upon tuition payments as the largest and primary source of operating funds. For such schools retention of students is financially critical. The purpose of this research is to develop a model of persistence with maximum predictability based on data obtained as early as possible in the application/matriculation process. The earlier the likelihood of dropping out of school is detected, the sooner intervention can occur to reduce that likelihood.

Because most dropouts occur during the freshman year, our goal is to maximize the predictive validity of the model while employing data collected prior to or concurrent with matriculation. The potential predictor variables are limited to demographics concerning the student and the parents of the student, high school academic and social experiences, and attitudes and opinions over a wide range of topics collected via a survey administered during freshman orientation. Thus, all data are measures of characteristics, opinions, attitudes, and values formed prior to the college experience.

Variable Development And Related Institutional Research

Beginning in the late 1970s, the dean of students and senior student affairs officer at a medium-sized private college in the northeastern part of the U.S. directed student oriented research employing a combination of established external instruments and a series of ad hoc studies which required the development of in-house survey instruments.

External Survey Research

CIRP Survey – This national study of freshmen is part of the Cooperative Institutional Research Program (CIRP), and was administered annually by the institution to incoming students prior to the beginning of classes. The focus was

upon freshman social, political, religious, and moral values, high school experiences, and expectations concerning academic and career goals.

College Student Experiences Survey – This national research was sponsored by the Higher Education Research Institute, UCLA Graduate School of Education, and was administered every four years to a cross-section of undergraduate students. Information collected included demographics, academic and social activities, and opinions concerning the value of various aspects of the college experience.

In-House Survey Research

In addition to these national student research studies in which we participated, several internal data collection procedures were also devised and implemented.

CIRP Senior Follow-up Survey – A survey including portions of the CIRP instrument, and a few additional items added by the college was administered to seniors as a part of their graduation experience. The goal was to track and assess changes in student perceptions between the freshman and the senior year.

Dropout Survey – Students who dropped out of the college were surveyed with respect to the reason(s) for their withdrawal. Results were compiled in a newly created dropouts database.

Focus Groups – The College contracted with an independent research organization to conduct focus groups wherein students were asked to evaluate their educational experience with respect to issues that address student services.

Ad Hoc Studies – Several ad hoc studies were undertaken to survey students over a wide array of topics potentially related to retention/attrition. Some of the projects were:

1. Change of major study,
2. Alcohol/substance abuse study,
3. Residence hall preferences study,
4. Student life study (focused upon the complete array of college services, i.e., freshman orientation, bursar, parking, academics, financial aid, career planning, etc.),
5. Faith and justice study (attempted to assess moral values),
6. Residence hall retention study, and
7. Student "work" research (analyzed characteristics of students working varying amounts on and/or off campus).

Student Retention Instrument and Attrition Database

In 1984 the senior student affairs officer and a faculty colleague developed a new survey instrument to be administered to all freshman matriculates. Survey items were derived in part from the CIRP Study and also from information gleaned from the ad hoc studies previously described. The survey data were combined with selected data from various campus administrative databases – financial aid, registrar, admissions, bursar, and the aforementioned new dropouts file. The intent of the researchers was to develop a set of variables that would be of value in predicting risk of attrition.

Freshmen are required to fill out the survey as a part of freshman orientation, and this is the only path to become a member of the attrition database. Through 1995, the response rate had been about 95 percent, meaning that approximately 95 percent of freshman matriculates are represented in the attrition database. The resulting attrition database houses approximately 250 variables on all freshmen that completed the survey.

Logistic Regression Model

Our operational definition of attrition is the act of dropping out at any time in a student's college experience. Thus a student who leaves the college during any semester, or completes a semester but fails to register for the following semester (without having graduated) is a dropout. Students who never drop out are referred to as persistors. Our fundamental goal was to, at the time of matriculation, assess the probability of attrition.

Because the dependent variable, persistence, is binary and categorical, a logistic regression was used. The logistic regression model was built on data collected from the freshman classes of 1988 to 1995 and was employed to assign a chance or probability of attrition to each freshman matriculating in the fall of 2000. The maintenance of relative balance between the number of dropouts and the number of persistors (i.e., 50% dropouts, 50% persistors) in the analysis sample was attained by random selection of an appropriate number of persistors from the population of all persistors. The result was an overall analysis sample of 5,221 students.

Principal Components Analysis of Survey Data

The student retention survey contains several sections that seek to elicit respondent attitudes/opinions with respect to a host of variables such as physical fitness, study habits in high school, expectations with respect to college and future career, family values, politics, religion, social responsibility, financial considerations, social relationships, etc. These data, in combination with the responses to the importance of selected attributes in the decision to attend this college, form the subset of variables to be initially analyzed via principal components analysis (PCA) prior to incorporation as independent variables in the logistic regression model. The exploratory PCA model was utilized for two specific purposes: (a) data reduction, and (b) the identification of factors to serve as uncorrelated potential predictor variables.

Other Potential Predictor Variables

In addition to the factors derived from the PCA model, a set of potential predictor variables for the logistic regression model consisting of background variables, external factors, financial factors and influences of significant others were taken from other internal college databases (e.g., admissions, registrar, etc.) as well as from ad hoc studies previously cited.

Results

Principal Components Analysis of Survey Items

Seventy-nine items from the survey were included in the PCA model. All items were scaled one-to-five, with higher response numbers always signifying more agreement or more importance. Variables that did not appear to comport with the concept of simple structure were dropped from the analysis. A scree plot was used to determine the number of factors to retain for subsequent rotation. The final model utilized 37 of the 79 survey items, and identified 12 principal components (factors) which accounted for 62.8 percent of the total variance among the 37 variables. A varimax rotation was used, and the resulting factor loading matrix met all of our expectations with respect to parsimony and a clear, interpretable factor pattern. The 12 rotated factors (SF1 - SF12) and their eigenvalues derived from the principal components analysis of the survey items are listed below. Factor scores (to be used as inputs to the logistic regression model) were computed by the regression method.

SF1 – Academic reasons for choosing this school (4.92)
SF2 – Unsure of career. (2.79)
SF3 – Enjoy politics (2.26)
SF4 – Cost was an important factor in choice of this college (2.10)
SF5 – Moral/religious values (1.75)
SF6 – Bad academic attitude in high school (1.68)

SF7 – Good study habits in high school (1.60)
SF8 – Good relationships with high school teachers (1.46)
SF9 – Concern for the disadvantaged (1.30)
SF10 – Physical fitness (1.19)
SF11 – Expect academic problems in college (1.13)
SF12 – Computer skills (1.07)

Other Predictor Variables

Selection of other possible independent variables for inclusion in the logistic regression model was accomplished via a series of tests comparing persistors vs. dropouts on: (a) mean differences (t-tests) for interval/ratio and assumed-interval/ratio variables, or (b) the likelihood of a relationship (chi-square tests-of-independence) for nominal variables. Variables with observed significance levels of .05 or higher ($p \leq 0.05$) were included as independent variables in the logistic regression. The final tally of independent variables to be included in the logistic regression model was 62, and was comprised of:

- 38 interval/ratio or assumed-interval/ratio variables (identified via t-tests),
- 12 categorical/nominal variables (identified via chi-square tests), and
- 12 factors: SF1 – SF12 (identified by the PCA analysis of the survey data).

Many of these 62 variables had high multicollinearity. For example, consider the combination of the three SAT variables – verbal SAT, math SAT, and total SAT scores. All 62 variables were originally used in exploratory backward stepwise logistic regression runs. We relied upon the backward stepwise procedure to eliminate variables that displayed high correlations with variables or combinations of variables in the model.

Logistic Regression

The 5,221 freshmen who matriculated between 1988 and 1995 formed the original analysis sample. A two-part procedure was employed in an iterative manner to arrive at the final model of significant predictor variables: (a) the backward stepwise logistic regression procedure to remove variables on the basis of the Wald statistic being less than 0.10 for removal; and, (b) our own criterion of a studentized residual greater than 1.90 for removal of a case as an outlier. The initial logistic regression run included all 62 potential predictor variables and all 5,221 cases. On successive runs variables were eliminated on the basis of poor predictive power (i.e., criterion Wald < 0.10 to leave the model) as determined by the backward stepwise procedure, and cases were disqualified as outliers on the basis of a studentized residual greater than 1.90.

Cases that had missing data for any predictor variable(s) were excluded from the analysis. The composition of the analysis sample appears below:

• Original cases	5,221
• Removed as outliers	- 920
• Selected Cases	4,301
• Excluded for missing data	-1,057
• Cases classified	3,244

Ultimately, the following nine variables and seven factors from the attrition database surfaced as potentially the most effective predictors of student attrition.

Background Variables

HSAVE – Eighth-term high school average
OFFWORK1 – Off campus hours worked per week at time of matriculation (range = 1-5)

MATRCAGE – Age at time of matriculation
EDPARNTS – Total of mother's and father's education (range = 1-10)
ZIPCODE – Live in local MSA (1) or not (2)
PARMAR – Parents alive, married and living together? YES (1) or NO or Missing (2)
GENDER – Male (1) or Female (2)

Financial Considerations

CONCERN1 – Concern for financing education at time of matriculation (range = 1-5)

Goal Commitment

SF5 - Moral/religious values
SF6 - Bad academic attitude in high school
SF7 - Good study habits in high school
SF8 - Good relationships with high school teachers
SF9 - Concern for the disadvantaged
SF11 - Expect academic problems in college
MARFUT – Student's expectations concerning marriage within one year after graduation. YES (1) or NO or Missing (2)

Institutional Commitment

SF1 - Academic reasons for choosing this school

The results of the logistic regression to predict student attrition are represented in Table 1. Each predictor variable is listed, along with:

1. The R statistic [a measure of the partial correlation between the variable and the likelihood of dropping out],
2. The B value [logistic regression coefficient],
3. The SE B [the standard error of B],
4. The Wald Sig [the observed significance level for the B coefficient via the Wald statistic], and
5. The Exp (B) [the factor by which the odds of dropping out change when the value of the predictor variable is increased by one unit].

Variables in Table 1 have been listed in order of the magnitude of the R statistic, the strength of the partial correlation between the predictor variable listed and the dependent variable - likelihood of dropping out. Please note the signs of the R values, and the signs of the regression coefficients B. The signs of R and B are negative for HSAVE (eighth-term high school average). The interpretation is that as high school average increases, the likelihood of dropping out decreases. The magnitude of the R statistic for HSAVE is larger than that of any other predictor. This means that HSAVE is more closely associated with chances of dropping out than are any of the other predictor variables. The second variable listed in Table 1 is SF6, the factor variable that represented a bad academic attitude in high school. The positive sign of its coefficient B (and R value) indicates that higher scores on this variable are associated with higher chances of dropping out. Alternatively, bad academic attitudes in high school are likely to lead to attrition in college.

Table 1

Logistic Regression Results

Variable	R	B	S.E. B	Wald Sig	Exp(B)
Constant	-	22.367	2.049	.000	-
Hsave	-.326	-0.381	0.017	.000	0.683
Sf6	.252	1.405	0.083	.000	4.075
**Parmar	.212	2.132	0.149	.000	8.435
Sf9	.181	0.818	0.067	.000	2.266
Sf7	-.179	-0.800	0.066	.000	0.449
**Gender	-.179	-1.578	0.131	.000	0.206
Sf5	-.178	-0.799	0.067	.000	0.450
Edparnts	-.165	-0.375	0.034	.000	0.687
Offwork1	.161	0.712	0.065	.000	2.038
Concern1	.145	0.516	0.053	.000	1.676
**Marfut	-.115	-1.287	0.165	.000	0.276
Sf8	-.108	-0.454	0.061	.000	0.635
Matrcage	.090	0.529	0.085	.000	1.697
Sf11	.086	0.368	0.062	.000	1.445
**Zipcode	.072	0.743	0.147	.000	2.102
Sf1	-.018	-0.113	0.060	.061	0.893

** Categorical Variables

Examination of the R statistic (partial correlation coefficient) and the B value (logistic regression coefficient) reveals that variables positively related to the probability of dropping out included SF6, PARMAR, SF9, OFFWORK1, CONCERN1, MATRCAGE, SF11, and ZIPCODE. Variables that were positively related to greater likelihood of persistence included HSAVE, SF7, GENDER, SF5, EDPARNTS, MARFUT, and SF8.

The criterion validity of the results of the logistic regression can be assessed by the percent of students correctly classified as dropouts or persistors (see Table 2). In our model 83% were correctly classified. To assess criterion validity, a cut point of 0.5 was used to form a classification table. Each student whom the model indicated had a probability of 0.5 or greater of dropping out was classified as a predicted dropout. Each student who was forecast to have a probability of any value less than 0.5 was classified as a predicted persistor.

Table 2
Classification of Predicted Versus Actual Dropouts and Persistors

	Predicted		
Observed	Persistors	Dropouts	Percent Correct
Persistors	1319	273	82.9%
Dropouts	279	1373	83.1%
		Overall Percent	83.0%

Intervention Program

There are always a host of potential problems that may contribute to a student's decision to drop out of college. In response, there must be a wide array of possible solutions that will increase the chances of retention. Recall that the primary goal of this research is the early identification of students who may be or may become candidates for dropping out. The logistic regression model provides the Director of Student Retention with invaluable information concerning which students are at highest risk, and which variables may be influencing that high likelihood of attrition.

As soon as a satisfactory logistic regression model has been attained, each student in the new freshman class is assigned a predicted probability of attrition based upon responses and scores on the predictor variables. This information is then promptly and directly communicated to the Director of Student Retention in the form of a list of all freshmen ordered by chances of dropping out. The director evaluates the projected at-risk students beginning at the top of the ordered list, or those students with assigned probabilities of or near 1.0. The director examines the scores on the predictor variables and creates a physical file on each student. In addition to the scores on the logistic

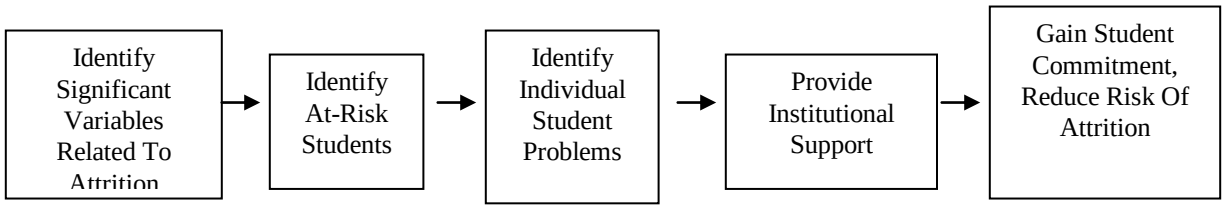
regression predictors, many student attributes are noted. Some examples are athletic status, major, extracurricular activities, commuter/resident status, health considerations, job commitments, and actual financial need.

Students with the higher chances of attrition are contacted by telephone to make an appointment with the Office of Student Retention. If a meeting is not scheduled or missed, up to two follow-up phone calls will be made to secure an appointment. The purposes of the face-to-face meetings with at-risk students are to:

1. Identify problems,
2. Provide support, and
3. Reduce the potential of attrition.

When successfully executed, this process enhances the ability of the director to prescribe an effective solution to the problem. For example, if a student acknowledges the difficulty of living at home with younger siblings, the director might suggest a move to on-campus housing. If the student discusses his/her off-campus job, the director may present the alternative of work on campus. A student with ambivalence about career issues can be referred to career services.

The objective of the director is to assist the student to identify and resolve any real or perceived problems, and to provide the student with a sense of involvement and ownership in the educational process. Schematically, the Office of Student Retention performs the following activities:



Limitations

The model is not able to provide any information with respect to the changes in likelihood of dropping out over time. That is, we are unable to track differences in the likelihood of attrition from the freshman to sophomore to junior to senior years. Though past research indicates that the predictors of attrition vary over the tenure of the student, we are unable to adjust the model for this variation. Hence the results reflect an average of effects over a four-year period (or longer) during which the student may have elected to dropout.

The research provided pertains to a private university in an urban location with a significant, though declining number of commuter students. For freshman matriculates, the percent living in college residences has risen from about 37 percent in the late 1980s to 57 percent in 2000.

References

1. Astin, Alexander, Korn, William, and Green, Kenneth (1987). Retaining and Satisfying Students. *The Educational Record*, 68, (1), 36.
2. Bean, John P. (1982). Student Attrition, Intentions, and Confidence: Interactions Effects in a Path Model. *Research in Higher Education*, 17, 291-319.
3. Miller, Thomas E., Glynn, Joseph G., and Neuner, Jerome L. (1988). Reducing Attrition: A College At Work In Research and Practice. *NASPA Journal*: 25, (4), 236-243.
4. Pascarella, Ernest T. and Terenzini, Patrick T. (1980). Predicting Freshman Persistence and Voluntary Dropout Decisions from a Theoretical Model. *Journal of Higher Education*, 51, (1), 60-75.

5. Stage, Frances K. (1988). University Attrition: LISREL with Logistic Regression for the Persistence Criterion. *Research in Higher Education*, 29, (4), 343-357.
6. Tinto, V. (1993). *Leaving College: Rethinking the Causes and Cures of Student Attrition.* The University of Chicago Press.
7. Wetzel, James N., O'Toole, Dennis, and Peterson, Steven (1999). Factors Affecting Student Retention Probabilities: A Case Study. *Journal of Economics and Finance*, 23, (1), 44-55.

Notes