Research in Higher Education, Vol. 43, No. 5, October 2002 (© 2002)

AN ANALYTIC STRATEGY TO ASSIST INSTITUTIONAL RECRUITMENT AND MARKETING EFFORTS

Stephen L. DesJardins*,†

This article demonstrates how to fit a statistical model to historical data, test whether the model can accurately predict enrollment out-of-sample, and use the results to segment admitted students into groups so that different recruitment and marketing interventions can be applied. Conceptual and practical issues are discussed, as well as policy considerations.

KEY WORDS: out-of-sample prediction; development and validation samples; Hosmer-Lemeshow goodness-of-fit test.

INTRODUCTION

Recruitment of college students is an increasingly important policy issue for many institutions of higher education. Enrollment-generated revenue has become an increasingly important part of university budgets as financing burdens have shifted from federal and state governments to institutions and students:

> [O]perating costs have escalated and public-sector financial support has flattened. As a result, many colleges and universities have had to sharply increase tuition and fees and look for ways to control costs to avoid financial disaster. (Council for Aid to Education, 1997, p. 10)

The overall effect of these financing trends "has been to increase reliance on tuition generally" (Pozdena, 1997, unnumbered).

Given the increased reliance on tuition revenue, the pressure to enroll more high-ability students, and the desire to have a diverse student body effective recruitment and enrollment of students, is an even more important function than it was a decade ago. Given the increased importance of being able to effectively

*Stephen L. DesJardins, Educational Policy and Leadership Studies, The University of Iowa.

†Address correspondence to: Stephen L. DesJardins, The University of Michigan, Center for the Study of Higher Postsecondary Education, 2108D School of Education Building, 610 E. University Avenue, Ann Arbor, MI 48109-1259. E-mail: sdesj@umich.edu

0361-0365/02/1000-0531/0

recruit and enroll prospective students, especially high-ability students, a number of firms (e.g., The College Board and Noel-Levitz) and consultants (e.g., Stephen Brooks and Associates) provide services to assist institutions in optimizing their recruitment and enrollment efforts. These firms typically assist institutions of higher education with all aspects of the enrollment process—from early contacts with prospective students to predicting which students will eventually enroll.

One stage of the enrollment process that is particularly important is the "admitted-to-enrollment" stage. At this stage, institutions spend a great deal of time and resources on admitted students encouraging them to enroll in the institution. Rather than focusing on students who are very likely to enroll, enrollment consultants often direct institutions to focus on students who are at the margin with respect to enrollment. These "fence sitters" (as some analysts call them) are students who may be convinced to matriculate to the client's institution. In collaboration with enrollment management professionals, consulting firms use statistical techniques to estimate each admitted student's probability of enrollment. These students are then segmented into groups to whom specific messages are targeted. Segmenting based on the probability of enrollment may even allow decision makers to eliminate some groups of students from some (or all) future recruitment efforts. Properly administered, this segmentation strategy makes institutional recruitment efforts more efficient by targeting recruitment resources toward segments that have the potential to produce additional enrollments.

Even though the general strategy outlined above may be known and practiced in some circles, there is scant literature available with regard to how this analysis actually takes place. Thomas, Dawes, and Reznik (2001) note that "an ERIC search reveals little research" (p. 1) about how this modeling strategy can be used to improve enrollment management. Consulting firms and analysts within institutions of higher education who conduct this type of research often do so under a veil of secrecy. Many consultants view their techniques as proprietary (for an exception, see Brooks, 1994, 1996), and institutional analysts are often reluctant to divulge methods that give their institution a comparative advantage in recruiting students (another exception is Thomas et al., 2001). The relative lack of information about this analytic strategy does little, however, to assist institutions that are unable to afford the high price of enrollment management consultants or who have institutional research offices lacking the analytic capabilities to conduct such research. Given the above, the objectives of this article are (a) to present a theoretically based approach that pierces this analytic veil of secrecy, (b) to present a detailed exposition of some of the statistical tests available when using this analytic strategy, and (c) to discuss how the results of this effort can be used to make student recruitment efforts more efficient and effective. Specifically, I demonstrate how this analytic strategy can be used to help enrollment managers target marketing and recruiting efforts once students have been admitted but before financial aid offers have been made.

It is important to stress, however, that the analytic strategy described herein is simply a tool designed to help enrollment management professionals make better decisions. The results of statistical modeling such as what follows should complement the experience of enrollment management professionals, not substitute for sound professional judgment about recruitment and telemarketing efforts. As an anonymous reviewer of an earlier draft of this article noted, "it is important to remember that enrollment management, even with the very best of tools, is far more 'art' than science."

LITERATURE REVIEW

The Theory of Student Choice

The conceptual underpinnings of models used when conducing studies of enrollment behavior are typically based on the student choice literature (Chapman, 1979; Hossler, Braxton, and Coppersmith, 1989; Kohn, Manski, and Mundel, 1976; Leppel, 1993; Paulsen, 1990; Welki and Navratil, 1987). Studies of college choice indicate that student enrollment behavior is related to students' individual characteristics and their preferences about the institution(s) they are considering (Fuller, Manski, and Wise, 1982; Manski and Wise, 1983). Other researchers have conducted macro-level studies of college-going behavior. These student demand models explain enrollments as a function of measures characterizing the population of potential enrollees and the characteristics of a relevant set of existing schools (Hoenack and Weiler, 1979).

The student-choice literature details that the college-choice process involves three broad stages. The first stage is the formation of college aspirations, which typically takes place from early childhood through high school, but may last well beyond high school for some individuals. The second stage involves identification, selection of, and application to a set of colleges (known as the "choice set"). This stage typically takes place in a student's junior or senior year in high school, but for nontraditional students this stage may occur after high school. The final phase is admission to an institution of higher education (or a number of institutions) and eventual matriculation to a single institution.

Given the increased importance and complexity of managing the enrollment enterprise within institutions, many colleges and universities have established high-level positions (typically at the vice president or vice provost level) responsible for enrollment management. Enrollment managers typically oversee staff in the admissions and financial aid offices, but may also be responsible for other departments involved in student recruitment and retention (i.e., residential services and academic advising). One of the main objectives of these administrators is to more effectively manage and coordinate the enrollment enterprise, since effective strategic enrollment management (SEM) depends on coordination among these departments (Hossler, 1984; Hossler and Bean, 1990). Effective coordination among the organization's various functions is not, however, the only role of enrollment managers. They must also improve their understanding of the way students discover, evaluate, and choose (or fail to choose) their institutions. Paulsen (1990) notes that college-choice studies can help institutional policymakers in this endeavor.

METHODOLOGY

The Samples

The study institution is a large, public Research I institution located in the Midwest. This institution enrolls approximately 3,800 new freshmen each fall; about 40% of these students are from out of state. In recent years, the institution has implemented strategic enrollment management techniques in an effort to make their recruitment and retention policies and practices more efficient and effective.

The Office of Admission at the study institution provided the data used in this study. The samples are restricted to (a) students who were admitted by early January of the year in which they plan to enroll in the institution, (b) students who were not recruited athletes, and (c) students for whom the institution had ACT Student Profile Questionnaire (SPQ) information. The data are restricted to admitted students because the objective is to build a model that will provide information about the admitted-to-enrollment stage of the student college-choice process. The samples are further restricted to students who applied to and were admitted by the end of the first week of January. This restriction is imposed since the results of the estimated model will be used in January of subsequent years to help enrollment managers at that point in the recruitment cycle (before financial aid offers are made). Recruited athletes are eliminated since the recruitment process for student-athletes is markedly different than for students in general.

Two entering cohorts are used in the analysis conducted here. The first data file contains students who were admitted for enrollment in the fall of 1999 ($N = 7{,}603$); the second file is comprised of students who were admitted for the fall 2001 class ($N = 6{,}810$).[1] (Data for the 2000 entering cohort was not yet available when this analysis was conducted.) To reiterate, both of these files are restricted as mentioned above. Each of these files contains information from three sources: the undergraduate application form, the student's high school transcript, and the SPQ. Data from the application includes demographic information, such as the student's home address, whether the student's parents graduated from college, and whether either or both parents graduated from the study institution. Transcript information used includes the name and size of the high school from which the student graduated. The SPQ section of the ACT Assessment is administered when the ACT entrance test is taken and is a rich source of information that is often underutilized by educational researchers. (SAT also collects this type of information when administering their test.)

The variables used in the multivariate models estimated here are described in Table 1, and descriptive statistics are presented in Table 2. The dependent variable in this study is a discrete variable indicating whether or not a student enrolled in the study institution. Independent variables hypothesized to affect a student's probability of enrollment include the student's personal, background, educational characteristics, and college preferences and intentions. Also included is information about the size of the student's high school and the enrollment to application "yield" that the study institution realized from high schools in recent years. These variables were chosen after surveying the student-choice literature.

The Conceptual Model

The conceptual model used in this study is based on human capital theory (Becker, 1964, 1993). The human capital model states that students make college-choice decisions by weighing the benefits and costs of various schooling and nonschooling alternatives. Students first decide whether to attend college or pursue noncollege options such as labor force participation. Conditional on deciding to enter college, the next decision involves making a choice between applying or not applying to a particular college from a set of preferred institutions. Human capital theory posits that at this stage, students choose to apply to a particular institution when the expected benefits of doing so outweigh the anticipated costs. Once admitted to a particular institution (or group of institutions) a similar utility calculation is involved in choosing whether to enroll in a particular institution. It is assumed that students enroll in a particular institution when the utility derived by this action is greater than the utility gained by matriculating to another institution.

The process of college choice also entails an institutional-decision component. Once students have applied to a college or university, institutional decision makers (typically admissions office staff and representatives of the faculty) decide which students to admit. Admissions decisions are usually based on a student's academic potential (as measured by ACT or SAT test scores, high school grades or rank percentile, or some combination of these), special talents (e.g., athletic or musical), and other institution-specific measures (whether the student's parents are alumni; known as "legacies"). Institutions have a variety of enrollment objectives. For instance, institutions typically attempt to enroll a freshman class large enough to meet revenue needs. They also try to craft a class diverse in a number of areas, such as race/ethnicity or academic discipline, and of sufficient academic quality to meet the institution's academic objectives.

TABLE 1. Definitions of the Variables

Variable	Description
Enrolled	1 if enrolled, 0 otherwise
African American	1 if African American, 0 otherwise
Asian	1 if Asian American, 0 otherwise
American Indian	1 if American Indian, 0 otherwise
Latino	1 if Latino/Hispanic/Chicano, 0 otherwise
Other ethnicity	1 if Other ethnicity or missing, 0 otherwise
White	1 if Caucasian, 0 otherwise (reference group)
Male	1 if male, 0 if female
Low income	Family income less than $42,000 (reference group)
Middle income	Family income $42,000 to $80,000
High income	Family income above $80,000
Admissions index score	[(2 * ACT Composite) + high school rank %]
Square of index score	Square of admissions index score
High school size	Number of students in the high school the student attended
Historical yield from H.S.	Yield (enrolled/applied) from each high school; 5 year average
Study institution first choice	1 if study institution was the students first choice, 0 otherwise
Institution supplemental choice	1 if supplemental score sent to the study institution, 0 otherwise
Anticipated major	Major code, taken from the Student Profile Questionnaire
Iowa resident	1 if Iowa resident, 0 otherwise
Illinois resident	1 if Illinois resident, 0 otherwise
Parent(s) are alumni	1 if either/both parents are study institution alumni, 0 otherwise
Want to attend public college	1 if the student wants to attend a public institution, 0 otherwise
Attended public high school	1 if the student attended a public high school, 0 otherwise
Interest in varsity athletics	1 if student interested in varsity athletics in college, 0 otherwise
Applied before August	1 if the student applied before August, 0 otherwise
Applied in August	1 if the student applied in August, 0 otherwise
Applied in September	1 if the student applied in September, 0 otherwise
Applied in October	1 if the student applied in October
Applied in November	1 if the student applied in November
Applied after November	1 if applied after November, 0 otherwise (reference group)

TABLE 2. Descriptive Statistics of the Developmental Sample

Variable	N	%	Mean	S.E.	Min.	Max.
Enrolled	1237	44.5				
African American	43	1.5				
Asian	93	3.4				
American Indian	7	0.3				
Latino	51	1.8				
Other ethnicity	98	3.5				
Male	1114	40.1				
Middle income	971	35.0				
High income	944	34.0				
Admissions index score	2776	—	127.07	19.84	63	169
Square of index score	2776	—	16541.68	5007.92	3969	28561
High school size	2776	—	334.71	203.76	17	1176
Historical yield from H.S.	2776	—	37.34	19.14	0	100
Study institution first choice	987	35.6				
Institution supplemental choice	699	25.2				
Anticipated major	2776		List of Majors from Student Profile Questionnaire			
Iowa resident	1290	46.5				
Illinois resident	1175	42.3				
Parent(s) are alumni	584	21.0				
Want to attend public college	2176	78.4				
Attended public high school	2290	82.5				
Interest in varsity athletics	964	34.7				
Applied before August	24	0.9				
Applied in August	226	8.1				
Applied in September	498	17.9				
Applied in October	552	19.9				
Applied in November	1006	36.2				

During a typical admission cycle, enrollment managers attempt to “build” a freshman class using a variety of methods. Early in the admission cycle, which may begin a few years before the freshman class plans to enter, institutions purchase names of potential applicants and use direct mail and high school visits to contact these students. The strategy is to produce applicants from this pool of prospective students. By early in the calendar year that the class will enter (January), most students who are going to apply have already done so. Also, a majority of eventual admits have already been admitted (82% of students admitted to the study institution in a given year are admitted by early January). Therefore, at this point in the admission cycle, enrollment managers begin to shift their focus of attention from increasing the pool of applications to marketing to admitted students. The objective is to try to optimize enrollments along a number of dimensions (quality and diversity, for example) given the capacity constraints of the institution.

It is important to note that at this point in the enrollment cycle, financial aid offers have not yet been made to students. Thus, models built to inform enrollment management staff about the propensities of admitted students to enroll must be developed without financial aid information. The analytic strategy presented here is designed to assist enrollment managers in developing such a model.

The Statistical Model

In studies like the one conducted herein, logistic regression is typically used. Specifying a theory-based model like the one discussed here permits us to explain and to predict enrollment behavior. The results can be used to explain enrollment behavior by examining how the independent effects of the included regressors affect the probability of enrollment. This analytic approach can also be used to predict each student's probability of enrollment (conditional on having been admitted), thereby allowing us to understand better the enrollment propensities of different groups of students.

Logistic regression is an appropriate technique because of the dichotomous nature (enrolled/otherwise) of the dependent variable (Dey and Astin, 1993; Hanushek and Jackson, 1977).[2] The logistic regression model used in this study is specified as

$$log \frac{P}{1 - P_i} = \alpha + \beta_i X_i + \delta_i Y_i + \gamma_i Z_i + \varepsilon_i \quad (1)$$

where P_i is the probability that student i will choose to enroll in the study institution; X_i is a vector of personal and demographic characteristics, such as socioeconomic background and academic ability; Y_i is a vector of prior educational characteristics, college intentions and preferences; Z_i is an institutional-level variable indicating the historical enroll/applicant ratio for the individual's high school; α, β_i, δ_i, and γ_i are estimated coefficients; and ε_i represents a random error term that is logistically distributed. The dependent variable is the logarithm of the odds that a particular student will enroll in the study institution. The model is estimated using maximum likelihood estimation (MLE). Both SAS and Stata are used to verify the consistency of the results, and copies of the programming code are available from the author.

The Analytic Strategy

Many analysts use the following strategy to assist enrollment managers to optimize their recruitment efforts. First, they estimate a logistic regression model of enrollment using a file of historical data containing information about admitted students (such as the 1999 cohort described previously). Then, they use the estimates produced by this model to calculate a probability of enrollment for each student admitted for a subsequent incoming class (such as the 2001 sample described). Producing estimates of enrollment for the incoming class is often known as "scoring" the data set. Once the incoming class is scored, these students are categorized into segments that contain a (roughly) equal number of students (deciles are typically used). Enrollment managers then differentially target these ten groups or segments, with the objective being to influence students who have probabilities that are "at the margin" with respect to enrolling (more on the strategies employed later).

The problem with this analytic approach is that the researcher does not know how accurate the model estimated using the historical admitted-student data will be in predicting enrollments for subsequent incoming classes. "When you use the same data to test the predictive accuracy of your model that you use to fit the model, it biases your results" (SAS Institute, Inc., 1995, p. 36). Even if the (pseudo) R^2 and the correct classification rate statistics produced by the model built on the historical data are high, one has little statistical evidence that the model will be effective in correctly predicting on a different sample (or "out-of-sample"). Hosmer and Lemeshow (1989) note that the estimates obtained using historical data will always appear "to perform in an optimistic manner" (p. 171) on other samples, leading the researcher to believe that the model will accurately predict out-of-sample. There is, however, no statistical justification for believing that this will be the case.

One way to improve on the above strategy is to randomly split the historical data into a "developmental" sample and a "validation" sample. Another option is to use one cohort of historical data as the developmental sample (e.g., 1998 admits) and then validate the model on a more recent cohort (e.g., all admits from 1999).[3] In either case, the developmental sample is used to estimate the statistical model, and the validation sample (sometimes known as the "holdout" sample) is used to test the predictive accuracy of the model. This general strategy is the approach used here. The historical data (students admitted for the class of 1999) is randomly split into two files, a developmental ($N = 3{,}801$) and validation file (3,802). Statistical models are estimated using the former, and the model is cross-validated using the latter.

Tests of Model Fit

Before using the estimates produced by the developmental sample model to score incoming admits, one needs to test how well the model fits the validation or holdout sample. There are alternative ways to test for model fit.

One "useful summary of the predictive ability of the model is a 2×2 table of the hits and misses of a prediction rule" (Greene, 1993, p. 651). The prediction rule may take the form:

$$y^* = 1 \text{ if } p > p^* \text{ and 0 otherwise} \tag{2}$$

where y^* classifies individuals as enrolled if their predicted probability (p) is greater than some threshold value (p^*). Typically, the default threshold value in logistic regression programs is 0.5, under the logic that we want to predict enrollment if the probability of enrolling is greater than one-half. However, using 0.5 as the threshold (or cutoff) value may not be appropriate in some cases. For instance, if the developmental sample is unbalanced, that is, a high percentage of admitted students enroll (or do not enroll), then this prediction rule might not accurately classify students (see Greene, 1993, p. 652 for an example). Thus, the results obtained using a 2×2 classification table are very sensitive to the choice of the cutoff score. Generally, there are two types of errors to consider when assigning a cutoff score: the incorrect classification of enrollees and incorrectly classifying non-enrollees. Changing the threshold value reduces the probability of one of these errors, but necessarily increases the probability of the other type of error. When studying enrollment behavior, the choice of an appropriate cutoff value depends on the costs of incorrectly classifying enrollees or non-enrollees. Enrollment managers should probably be most concerned with incorrectly classifying actual enrollees. The implications of this are that to error on the "safe" side one should choose a (relatively) low threshold or cutoff score when developing 2×2 classification tables. Using a low cutoff score will result in relatively more students who actually did not enroll being classified as enrollees. (Classification tables can be produced using the logistic regression procedures in Stata, SAS, and SPSS by requesting the LSTAT, CTABLE, and classification function, respectively.)

A more sensitive way to examine how accurately the model classifies students is to regroup "the data by ordering on the predicted probabilities and then forming, say, ten nearly equal-size groups" (StataCorp, 2001, p. 229). Once these new groups are formed, one can table the actual vs. predicted enrollments by decile. This seems to be a common method in other applications of logistic regression to classification (see Hosmer and Lemeshow, 1989, or Lemeshow and LeGall, 1994, for details.) This type of table provides very detailed information about the accuracy of the model within each decile. These tables are easily produced when using the logistic regression procedures in many mainstream statistical packages such as SAS, SPSS, and Stata [see the documentation for the LACKFIT option in SAS, the Hosmer-Lemeshow (HL) goodness-of-fit option in SPSS, and the LFIT option in Stata].

Hosmer and Lemeshow (1989) developed a way to test whether the model fits the data using decile groupings like that described previously (pp. 140–145). Basically, the statistic tests the accuracy of the model in each grouping. The HL goodness-of-fit statistic tests the null hypothesis that the model fits the data (thus, a significant HL test provides evidence that the model does not fit the data).

A measure that is also used to indicate how accurately a model predicts the occurrence of events (such as enrollment) is the Brier score (Brier, 1950). The Brier score is a unitless index of predictive accuracy defined as:

$$\text{avg } [(p_i - y_i)^2] \tag{3}$$

where p_i is the predicted value of enrollment for individual i obtained from the estimated model and y_i is the actual or observed value of enrollment for each individual in the developmental sample. "The Brier score is a strictly proper scoring rule, which means that it is minimized for predicted probabilities that are equal to the true probabilities" (SAS Institute, Inc., 1995, p. 35). The Brier score ranges from 0 to 1, with smaller scores indicating better predictive accuracy. "The Brier score can also be used to compare the predictive accuracy of different models" (SAS Institute, Inc., 1995, p. 44).

A graphical way to test the predictive accuracy of a logistic regression model is to plot the receiver operating characteristics (ROC) curve (Hanley and McNeil, 1982). The ROC curve plots the sensitivity vs. one minus the specificity of the estimated model using a number of prespecified cutoff or threshold values. Sensitivity (specificity) is the percent of enrollees (non-enrollees) who are correctly predicting as enrolling (not enrolling; see Hosmer and Lemeshow, 1989; or SAS Institute, Inc., 1995, for more about these terms). The statistic used to assess the model fit measures the area between a 45-degree diagonal and the ROC curve (Fig. 1). This value, sometimes known as the "c" statistic, is automatically produced by the SAS logistic regression procedure. A model that would be no better than flipping a coin (no predictive accuracy) would have a ROC value of 0.5, and the ROC curve would fall exactly on the 45-degree line plotted in Fig. 1. A model that perfectly predicted the outcome would have a "c" statistic equal to 1. The greater the predictive power of the model, the more the curve tends to bow toward the upper left-hand corner of the graph and the higher the associated statistic.

Using the cross-validation approach discussed previously, in conjunction with the aforementioned statistical tests, provides a way to more thoroughly assess the efficacy of the statistical model built using the developmental sample (Hosmer and Lemeshow, 1989), and this is the procedure used here.

THE RESULTS

Table 3 contains relevant statistical information about the logistic regression results produced using the developmental sample. Included are the standard re-

TABLE 3. Developmental Sample Results

Model fitting information and testing global null hypothesis BETA = 0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	3824.826	3383.35	
-2 LOG L	3822.826	3329.35	493.476 with 26 DF ($p = 0.0001$)
Score			462.189 with 26 DF ($p = 0.0001$)
R^2	0.16		

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Odds Ratio
Intercept	3.2396	1.5982	4.1089	0.0427	—
African American	0.0584	0.3531	0.0273	0.8687	1.06
Asian	–0.3051	0.2482	1.5109	0.2190	0.74
American Indian	1.3759	1.2008	1.3130	0.2518	3.96
Latino	0.2156	0.3041	0.5028	0.4783	1.24
Other ethnicity	–0.1241	0.2448	0.2570	0.6122	0.88
Male	0.0519	0.0887	0.3429	0.5582	1.05
Middle income	0.1308	0.1079	1.4684	0.2256	1.14
High income	0.2284	0.1130	4.0893	0.0432	1.26
Admissions index score	–0.0636	0.0254	6.2892	0.0121	0.94
Square of index score	0.0002	0.0001	3.2426	0.0717	1.00
High school size	–0.0005	0.0003	3.9380	0.0472	1.00
Historical yield from H.S.	0.0073	0.0029	6.4337	0.0112	1.01
Study institution first choice	0.6889	0.1025	45.2085	0.0001	1.99
Institution supplemental choice	0.5337	0.1123	22.6034	0.0001	1.71
Anticipated major	0.0001	0.0003	0.2231	0.6367	1.00
Iowa resident	1.0243	0.1787	32.8375	0.0001	2.79
Illinois resident	0.1520	0.1552	0.9586	0.3275	1.16
Parent(s) are alumni	0.2892	0.1082	7.1402	0.0075	1.34
Want to attend public college	0.3201	0.1112	8.2829	0.0040	1.38
Attended public high school	0.0106	0.1272	0.0069	0.9337	1.01
Interest in varsity athletics	–0.4364	0.0921	22.4377	0.0001	0.65
Applied before August	0.7736	0.4556	2.8830	0.0895	2.17
Applied in August	0.6634	0.1859	12.7343	0.0004	1.94
Applied in September	0.4128	0.1460	7.9939	0.0047	1.51
Applied in October	0.2539	0.1442	3.1004	0.0783	1.29
Applied in November	0.1208	0.1276	0.8951	0.3441	1.13

TABLE 3. (*Continued*)

Group	Total	Enrolled		Did Not Enroll	
		Actual	Predicted	Actual	Predicted
1	278	40	42.6	238	235.4
2	278	56	61.5	222	216.6
3	279	76	76.9	203	202.2
4	278	103	93.7	175	184.3
5	278	110	110.5	168	167.6
6	279	136	131.9	143	147.1
7	278	156	152.5	122	125.5
8	278	165	173.8	113	104.2
9	278	196	194.2	82	83.8
10	272	217	217.5	55	54.5
Hosmer-Lemeshow goodness-of-fit statistic = 3.9007 with 8 DF (p = .866)					
Cutoff (prior probability)	44.5				
Correct classification rate	66.8				
Sensitivity	66.9				
Specificity	66.7				
Brier score	0.19				
"c" statistic	0.74				

sults produced by logistic regression models (coefficient estimates and their respective standard errors, Wald statistics, *p*-values, and odds ratios) and statistics assessing the model fit. Also included in Table 3 is the actual vs. predicted enrollments by decile, the HL goodness-of-fit test results, and information about how accurate the model is in classifying students.

The main objective of this study is to determine the predictive accuracy of the results of the model, and not to focus on the independent effects of the included variables. I briefly mention a few of the more interesting results.

As noted in Table 3, students who have high admissions index scores are less likely to enroll than their lower scoring counterparts. Specifically, *ceteris paribus*, for a one-point change in the admissions index the odds of enrolling are expected to change by a factor of 0.94. High school size and enrollment are also negatively related, but students from high schools with historically high yield rates have odds of enrollment that are higher than students from high schools with relatively low yield rates. Students for whom the study institution was their first choice or students who are residents of the home state of the study institution have odds of enrollment that are approximately 2.0 and 2.8 times higher (respectively) than other students. Students who are legacies have odds of enrolling about 1.3 times higher than students whose parents did not graduate from the study institution. The logistic regression results also provide evidence that students who apply before September of the year preceding enrollment have odds of enrolling that are about two times that of students who apply in December or early January.

Of note is the result that students who are interested in participating in varsity athletics at the study institution have odds of enrollment of approximately 65% that of students in general (keep in mind that the sample does not include recruited athletes). Additional analyses were conducted to better understand this relationship. The results indicate that these students are almost exclusively males, from Illinois or Wisconsin, who are from middle- and upper-income families, attended public high schools with less than 600 students, and participated in varsity athletics. An athletic administrator at the study institution suggested that many high school students have unrealistic expectations that they can participate in Division I athletics. Thus, it may be that as these students discover that they are unlikely to participate in Big Ten athletics, they gravitate toward institutions where they have a better chance of "making the team."

Cross-Validation

It is important to test whether the model estimates displayed in Table 3 will be useful in predicting enrollments for other samples of students. Known as cross-validation or testing for constancy, one simply uses the estimates from Table 3 to produce probabilities of enrollment for each person in the validation sample (this is quite easy to do in SAS and Stata). Then, the statistical tests of model fit discussed previously are used to test how well the developmental sample results fit the validation sample (see Table 4 and Fig. 1).

The HL goodness-of-fit statistic displayed at the bottom of Table 4 is not significant ($p = 0.6572$) providing hypothesis-based statistical evidence that the model fits the validation sample. This is in contrast to Thomas et al. (2001), who also test a predictive model on a holdout sample, produce a HL-like table, state that the "estimated model fits the data quite well" (p. 5), but fail to provide the appropriate HL statistic to justify this claim.

The correct classification rate (CCR) indicates that 65.7% of all cases are correctly predicted using a cutoff of 44.5, which is the percentage of enrollees in the developmental sample. The model correctly predicts 64.9% of enrollees (known as sensitivity) and 66.4% of non-enrollees (known as specificity). That the model is more accurate in predicting non-enrollees is not surprising given that there are more non-enrollees (55.5%) than enrollees (44.5%) in the sample used to develop the model.

The Brier score is 0.21, providing further evidence that the model is effective in predicting enrollments in the validation sample. The ROC curve is plotted and displayed in Fig. 1. The "c" statistic, or area between the ROC curve and the 45-degree line, is 0.72, also providing evidence that the model fits the validation sample.[4]

TABLE 4. Validation Sample Results

		Enrolled		Did Not Enroll	
Group	Total	Actual	Predicted	Actual	Predicted
1	278	54	45.0	224	233.0
2	278	62	62.7	216	215.4
3	280	80	78.5	200	201.5
4	278	88	93.5	190	184.5
5	278	112	110.3	166	167.7
6	278	121	129.2	157	148.8
7	278	146	150.4	132	127.6
8	278	174	173.9	104	104.1
9	278	192	195.7	86	82.3
10	273	211	219.6	62	53.4
Hosmer-Lemeshow goodness-of-fit statistic = 5.911 with 8 DF (p = 0.6572)					
Cutoff (prior probability)		44.5			
Correct classification rate		65.7			
Sensitivity		64.9			
Specificity		66.4			
Brier score		0.21			
"c" statistic		0.72			

A final way to assess the efficacy of the logistic regression model estimated above is to simply examine how accurately the model predicts within each decile of the validation sample (Table 4). In Group 1 there are 54 (224) students who enrolled (did not enroll), and the model predicted that 45 (233) would enroll (not enroll). The model predicts much more accurately, however, in other deciles. For instance, in Group 8 the actual and predicted observations are virtually identical. Providing the results in a table like this allows decision makers to observe whether the model over- or underpredicts for different groups, and this can be valuable information for enrollment managers when they are deciding to which groups to target their marketing efforts.

The combination of all of these different statistical tests should provide individuals with a clear indication of whether the model they have developed will be useful in predicting out-of-sample. Using the combination of tests described above is preferable to using any single test, such as the overall CCR or a HL-like table without the appropriate hypothesis-based test of model fit.

The final step in this analysis is to use the developmental model results to predict the probability of enrollment (and associated group or decile membership) for each student in the admit pool of the incoming freshman class (the 2001 sample). Newly scored incoming students are then grouped into deciles.

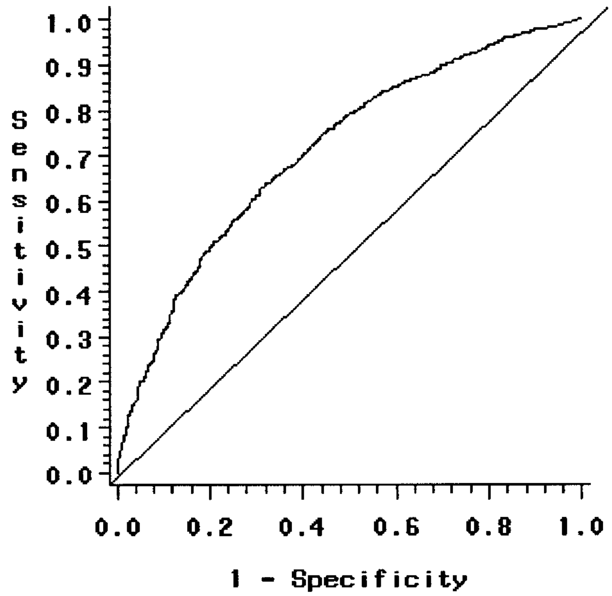

FIG. 1. Receiver operating characteristics curve for the validation sample.

These groups, and their associated probability-of-enrollment ranges, are displayed in Table 5. Since this analysis was conducted in early 2001, we have no way of knowing which students actually enrolled. However, given the tests of model fit conducted previously, we are confident that the actual results will be similar to the estimates displayed in Table 5.

CAVEATS

Some of the variables used in the models discussed here were constructed from the questionnaire filled out when students take the ACT test. For institutions that require the ACT test for admissions purposes, this data meets Thomas et al.'s (2001) suggestion that to implement assessment-based recruitment management the institution "must have good data in a usable form" (p. 7). At the study institution, the ACT information was easy to incorporate with institutional data because more than 90% of admitted students take the ACT test and fill out the SPQ. If, however, an institution gets a more even split of admits who took the SAT or ACT only, then using the survey information from two different sources may be problematic. In this case, one would have to obtain the survey information from both SAT and ACT and use items common to both surveys.

TABLE 5. Out-of-Sample Prediction for 2001 Admitted Students

		Predicted Probabilty	
Group	*N*	Min.	Max.
1	681	0.00	0.07
2	681	0.07	0.13
3	681	0.13	0.20
4	681	0.20	0.27
5	681	0.27	0.36
6	681	0.36	0.47
7	681	0.47	0.59
8	681	0.59	0.72
9	681	0.72	0.86
10	681	0.86	1.00

Although the techniques presented above are easy to perform using some statistical programs (Stata), they require more sophisticated programming expertise in other programs (SAS and SPSS). Fortunately, most institutions of higher education have faculty and/or staff who have the capabilities of performing this type of analysis. If in-house expertise is not available, the institution may be forced to hire an outside consultant. The savings produced by more efficiently targeting scarce recruiting resources should, however, more than cover the cost of the model development. Also, the model can often be used for multiple years, thereby reducing the cost with each use.

Allison (1999) notes that the HL goodness-of-fit test is popular "because it fills a major need and has no serious competition. But it's a rather *ad hoc* statistic and its behavior has not been extensively investigated" (p. 55). Allison conducted simulations that suggest that the HL test may not be very powerful. He notes, however, that his simulations hardly constitute "a definitive study" (p. 56). Nonetheless he suggests caution in "concluding that a model is OK just because the HL test is not significant" (p. 56). This suggests a strategy like that undertaken in this study where multiple tests of model fit are used.[5]

POLICY IMPLICATIONS

Enrollment managers should be interested in increasing the probability that marginal, or fence-sitting, students will enroll. Institutions have groups of admitted students who have very high probabilities of enrolling (Group 10 in Table 4; sometimes known as "hot prospects") and many of these students (77% in this case) will enroll regardless of what recruiters do. Thus, it is probably inefficient to spend a great deal of scarce recruiting and marketing resources on this group. A more effective strategy is to try to "move" students who are "wavering on the brink of an enrollment decision" (Thomas et al., 2001, p. 2). There is no purely empirical way, however, to determine who these fence sitters are. Identifying these prospects involves a combination of empirical methods, such as those previously discussed, as well as subjective analysis by enrollment management professionals. Once results like those presented in Table 4 are available, the analyst and enrollment professionals should meet to discuss which groups to target and in what order. Although the analytic strategy described herein provides information about students' enrollment propensities, this information must be combined with the experience of enrollment professionals to determine on which students to focus the majority of recruitment and telemarketing efforts.

Even though a majority of students from low-probability groups will not enroll, there are still a number of students in these groups who will eventually enroll. The information provided by this analytic technique can also be used to effectively target specific groups of students from low-probability decile groups. For instance, if the institution is interested in increasing the ethnic/racial diversity of its entering class, they could target relatively low-probability (Groups 3 and 4) minority students with an interest in a field of study for which the institution is highly regarded. These students could be contacted by the program's faculty and/or sent information about different sources of aid for students interested in this program. The general analytic strategy of combining information about a student's probability of enrolling with other characteristics allows enrollment managers a great deal of flexibility in deciding which students to target, at what time, and with what message. By segmenting students in this way, enrollment managers are limited only by the data available and their imaginations.

A great deal of money is spent, even in the admitted-to-enrollment part of the admission cycle, on mailings, telemarketing, and other forms of student contact. Being able to target direct mail, or to eliminate some students from mailings, has the potential of saving institutional resources. For instance, from January to March enrollment managers send information to admitted students about career fairs, computer resources, orientation schedules, the local community, and housing options. Eliminating low-probability students from these mailings could save a substantial amount of money. For example, if decision makers eliminated Groups 1 and 2 (Table 5) from these mailings, a rough estimate of the possible savings is about $7,000 (1,362 students in Groups 1 and 2, who would have been sent five mailings at $1.00/mailing). The savings could be even more substantial if more than five mailings were done, or if the mailings involved more expensive content (i.e., glossy viewbooks that often cost $2–$3 per item). Institutions that contract these services to direct mail marketers may reap even larger savings by more effectively targeting these mailings.

Similar savings can also be realized by prioritizing and targeting telemarketing efforts. For instance, the results of this study have changed the way telemarketing is done at the study institution. In the past, admitted-student telemarketing lists were largely prioritized by where the institution was in the student's choice set. First- and second-choice students were called first, and all other admitted students were called later (if time and resources permitted). Now telemarketing efforts are prioritized depending on a student's probability of enrolling, not simply on one or two observed characteristics (like choice or a demographic characteristic). The telemarketing operation now prioritizes calls based on probability group, in conjunction with other relevant variables. For example, high-ability students in Groups 6 and 7 (see Table 5) who had the study institution as their second-choice college and who were from contiguous states were among the students that enrollment managers targeted. Providing these students with information about the institution's scholarship for high-ability nonresidents may convince them to enroll in the study institution.

Enrollment managers need to be conscious of when to implement models like the one discussed above. When working the admitted-to-enrollment stage it is important to allow enough time to have an impact, yet not begin one's efforts too early in the cycle. The statistical model can be estimated anytime because historical data are being used. *When* to score one's admitted students is a more critical issue. We waited until early January, because many students apply and are admitted to the institution during the holiday season. By the second week of January, the institution has admitted over 80 percent of the total number of admits for the following fall. (Since this percentage may vary by institution, analysts need to determine the appropriate strategies for their institution.) Shortly after the holiday break, the Office of Admissions produced a file of all students who had already been admitted for the fall 2001 semester. This data file was passed to the author for scoring. The scored data file, containing estimates of the probability of enrollment for each student and their decile-group membership, was returned to enrollment managers in a matter of hours. Eventually the model estimates will be incorporated into the admission's database, and admits will be automatically scored as they enter the database.

Strategic enrollment management should incorporate all aspects of student recruitment, selection, and retention. The optimum analytic strategy is to develop a comprehensive model that links student choice and selection with student retention and success at the institution (for more about student application behavior, see DesJardins, Dundar, and Hendel, 1999; Weiler, 1994; Welki and Navratil, 1987). Such a model would more appropriately incorporate the concept of "student–environment fit," which has been found to increase student success in college (Tinto, 1975). For instance, Tinto and Wallace (1986) note, "the most effective retention programs begin with admissions" (p. 291). They argue, "an effective retention policy that includes the work of admissions officers can be as important, if not more important, to the long term survival of institutions than a policy that relies exclusively on recruitment" (p. 291). Tinto and Wallace believe that admissions officers must be more than just good recruiters and marketers. These staff members must also be adept at helping students choose an institution that fits their needs and be able to assist students in developing reasonable expectations about their academic careers. An admissions staff that emphasizes this "*educative* function" (emphasis theirs) can help to "engender among those students who enroll a growing commitment to the institution," which is not only the "key to institutional retention" but also a key to attracting future recruits (Tinto and Wallace, 1986, p. 292).

Finally, the analytic strategy detailed above is based on well-established statistical and conceptual theory. The independent variables included in the model were chosen based on previous studies of student choice, and the statistical technique is widely used to study education-related events like student enrollment. Even though it seems obvious that one should construct such a model based on sound conceptual theory, this is not always the case. Some enrollment management consultants often build these models using atheoretical techniques such as stepwise regression. They collect a large amount of institutional and geodemographic data, create a large number of explanatory variables from these data, and then include many or all of these variables in a stepwise regression model. One of the justifications commonly cited for using such an approach is that it helps the researcher to determine the variables that are most important in explaining enrollment behavior, thereby leading to the discovery of the "best" model (Hanushek and Jackson, 1977). However, in logistic regression, the determination of model fit is based on likelihood statistics. These measures are, however, sample specific and determined not only by the strength of the relationship between all the independent variables but also by the variance of each regressor included in the model. The stepwise procedure typically operates on increments to the log-likelihood, but the variation attributable to any single covariate is dependent on when that variable is entered into the model, which regressors are already included in the model, and the order in which other variables are entered into the model. For instance, a variable entered at an early stage may not be related to the dependent variable but may be correlated with other variables that are related to the dependent variable. Depending on the level of correlation among these variables, the "true" variables may never be included in the final model. Also, it is quite possible that some very important variables will *never* be included in the final model. This can happen if variables have "offsetting" effects, that is, the variables have similar effects but are negatively related in the sample, or if each has an opposing effect on the dependent variable. Hanushek and Jackson (1977) note that there is "little assurance that the final model—the model selected at the end of the entire stepwise procedure—bears any relationship to the underlying population model" (see p. 96 for an example of the problems that arise when using stepwise procedures). For a host of reasons, if one is interested in explanation as well as accurate prediction, theory-based models are preferred.

ACKNOWLEDGMENTS

I would like to thank the staff of the Office of Admissions at the study institution for helpful comments and for being supportive with regard to putting this article in the public domain. I would also like to thank Rob Toutkoushian and Don Hossler for comments on an earlier draft of the article. Any errors or omissions are, however, my responsibility.

ENDNOTES

1. The data files are not restricted to "students offered admission as a full-time freshman" as in the Thomas, Dawes, and Reznik article (2001, p. 2). If Thomas et al. restricted their samples to students who indicated that they would enroll full time, then their results could not be used to assess the enrollment propensities of all admitted students in later cohorts. Whether this is a matter for concern depends in part on the percentage of students who enroll part time. At the study institution in this article, about 8% of new freshmen enroll part time in their first semester. Thus, the samples used in the analysis include both full- and part-time students.
2. Probit regression would also be appropriate to use to estimate dichotomous dependent variable models.
3. For institutions with smaller pools of admitted students, it may be necessary to use a full entering cohort of admitted students as the developmental sample, and a different cohort as the validation sample to avoid estimation problems that arise when using small samples. In this article, I present the results of randomly splitting a single cohort into developmental and validation samples, but I also used the other approach, whereby one year was used to develop the model (1998 cohort) and another sample (1999) was used to validate the results. The outcomes were invariant to the method used, so I elected to present only one set of results.
4. When the sample is relatively balanced, that is, half of the sample has the event of interest, ordinary least squares (OLS) regression is an alternative to logistic regression (Dey and Astin, 1993; Hanushek and Jackson, 1977). The sampling theory underlying OLS procedures is well established, therefore, there are tests of parameter constancy (out-of-sample model fit) that are available when using OLS that are not available when using logistic regression. Thus, OLS models were also estimated using the developmental sample and Chow tests (Chow, 1960) were calculated. Two different forms of the Chow Forecast test were produced to test the predictive accuracy of the developmental model (see Johnston and DiNardo, 1997, pp. 113 and 116 for specifics about the tests). The results of these two tests support the findings noted above that the model estimated using the developmental sample accurately predicts out-of-sample (the Chow test results are available from the author on request).
5. In particular, if the proportion of enrollees/non-enrollees in the developmental sample is relatively balanced, one should test the stability or constancy of the model using Chow tests. The sampling theory underlying OLS regression is much more developed than that of nonlinear models like logistic regression, and the Chow tests (and a related test called the Hansen test, 1992) are quite powerful.

REFERENCES

Allison, P. D. (1999). *Logistic Regression Using the SAS System: Theory and Application,* SAS Institute, Inc., Cary, NC.

Becker, G. S. (1964). *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education,* National Bureau of Economic Research, New York.

Becker, G. S. (1993). *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education,* National Bureau of Economic Research, New York.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Mon. Weather Rev. **75:** 1–3.

Brooks, S. H. (1994). *Unlocking the enrollment puzzle: Using advanced techniques to forecast yields.* Unpublished manuscript.

Brooks, S. H. (1996). Econometric modeling of enrollment behavior. *J. Stud. Financ. Aid* **26**(3)**:** 7–17.

Chapman, R. C. (1979). Pricing policy and the college choice process. *Res. Higher Educ.* **10:** 37–57.

Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica* **52:** 211–222.

Council for Aid to Education (1997). *Breaking the Social Contract: The Fiscal Crisis in Higher Education,* Rand Corporation, Santa Monica, CA.

DesJardins, S. L., Dundar, H., and Hendel, D. (1999). Modeling the college application decision process in a land-grant university. *Econ. Educ. Rev.* **18**(1)**:** 117–132.

Dey, E. L., and Astin, A. W. (1993). Statistical alternatives for studying college student retention: A comparative analysis of logit, probit, and linear regression. *Res. Higher Educ.* **34**(5)**:** 569–581.

Fuller, W. C., Manski, C. F., and Wise, D. A. (1982). New evidence on the economic determinants of postsecondary school choice. *J. Hum. Resour.* **17**(Fall)**:** 477–498.

Greene, W. (1993). *Econometric Analysis,* Macmillan, New York.

Hanley, J. A., and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristics (ROC) curve. *Radiology* **143:** 29–36.

Hansen, B. E. (1992). Testing for parameter instability in linear models. *J. Policy Model.* **14:** 517–533.

Hanushek, E. A., and Jackson, J. E. (1977). *Statistical Methods for Social Scientists,* Academic Press, San Diego, CA.

Hoenack, S. A., and Weiler, W. C. (1979). The demand for higher education and institutional enrollment forecasting. *Econ. Inq.* **17:** 89–113.

Hosmer, D. W., and Lemeshow, S. (1989). *Applied Regression Analysis,* John Wiley and Sons, New York.

Hossler, D. (1984*). Enrollment Management: An Integrated Approach,* College Entrance Examination Board, New York.

Hossler, D., and Bean, J. (1990). *The Strategic Management of College Enrollments,* Jossey-Bass, San Francisco.

Hossler, D., Braxton, J., and Coppersmith, G. (1989). Understanding student college choice. In: Smart, J. C. (ed*.*)*, Higher Education: Handbook of Theory and Research* (Vol. V), Agathon Press, New York, pp. 231–288.

Johnston, J., and DiNardo, J. (1997). *Econometric Methods,* McGraw-Hill, New York.

Kohn, M. G., Manski, C. F., and Mundel, D. S. (1976). An empirical investigation of factors which influence college-going behavior. *Ann. Econ. Soc. Meas.* **5**(4)**:** 391–419.

Lemeshow, S., and LeGall, J. R. (1994). Modeling the severity of illness of ICU patients: A systems update. *J. Am. Med. Ass.* **272:** 1049–1055.

Leppel, K. (1993). Logit estimation of a gravity model of the college enrollment decision. *Res. Higher Educ.* **34**(3)**:** 387–398.

Manski, C. F., and Wise, A. D. (1983). *College Choice in America,* Harvard University Press, Cambridge, MA.

Paulsen, M. B. (1990*). College Choice: Understanding Student Enrollment Behavior* (ASHE-ERIC Higher Education Report 90-6), The George Washington University, Washington, DC.

Pozdena, R. (1997). *Power to the Student: An Alternative to Higher Education Funding Increases,* Cascade Policy Institute, Portland, OR.

SAS Institute, Inc. (1995). *Logistic Regression Examples Using the SAS System, Version 6, First Edition,* SAS Institute, Inc., Cary, NC.

StataCorp (2001). *Stata Statistical Software: Release 7.0,* Stata Corporation, College Station, TX.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Rev. Educ. Res.* **45:** 89–125.

Tinto, V., and Wallace, D. (1986). Retention: An admission concern. *Coll. Univ.* **61:** 290–293.

Thomas, E., Dawes, W., and Reznik, G. (2001). Using predictive modeling to target student recruitment: Theory and practice. *AIR Prof. File* **78**(Winter)**:** 1–8.

Weiler, W. C. (1994). Transition from consideration of college to the decision to apply. *Res. Higher Educ.* **35**(6)**:** 631–646.

Welki, A. M., and Navratil, F. J. (1987). The role of applicants' perceptions in their choice of college. *Coll. Univ.* **62:** 147–160.

Received April 19, 2001.