DOCUMENT RESUME

ED 433 753		HE 032 303
AUTHOR	Thomas, Emily; Reznik, Gayle; Dawes, William	
TITLE	Using Predictive Modeling To Target Student Recruitment: Theory and Practice. AIR 1999 Annual Forum Paper.	
PUB DATE	1999-06-00	
NOTE	18p.; Paper presented at the Annual Forum of the Association for Institutional Research (39th, Seattle, WA, May 30-June 3, 1999.)	
PUB TYPE	Opinion Papers (120) -- Speeches/Meeting Papers (150)	
EDRS PRICE	MF01/PC01 Plus Postage.	
DESCRIPTORS	College Choice; Decision Making; Higher Education; Institutional Research; Models; *Predictor Variables; *Regression (Statistics); State Universities; *Statistical Analysis; *Student Recruitment	
IDENTIFIERS	*AIR Forum; *State University of New York Stony Brook	

ABSTRACT

This paper argues that a typical use of regression models to target student recruitment efforts is theoretically unsound and may therefore be operationally inefficient. It presents results from a study using a predictive model to identify the prospective students on whom recruitment efforts have the greatest impact. The model uses four kinds of variables: demographic, academic, geographic, and behavioral. Application of the model at the State University of New York (Stony Brook) found that predictive variables significantly and positively related to student enrollment at the 1 percent level were: high-yield high school average, high-yield Scholastic Assessment Test scores, high-yield zip code, and open house attendance. Significant predictive variables related negatively to student enrollment included White or Hispanic ethnicity, U.S. citizenship, regular admission status, early application, and on-campus housing request. The model was used to demonstrate the efficacy of an experimental program of increased contact with admitted students and their parents. Findings indicated that a modest increase in recruitment activity increased the enrollment of students with relatively low enrollment probabilities but did not improve the recruitment of students identified by the regression model as more likely to enroll. Results suggest that the typical use of predictive modeling to identify "hot prospects" may be inefficient and ineffective. (DB)

USING PREDICTIVE MODELING TO TARGET STUDENT RECRUITMENT: THEORY AND PRACTICE

Emily Thomas*
Director of Planning and Institutional Research
310 Administration
State University of New York
Stony Brook, New York 11974-0701
(516)632-7272

Gayle Reznik
Research Assistant
Office of Institutional Research
488 Administration
State University of New York
Stony Brook, New York 11794-1307
(516)632-6980

William Dawes
Undergraduate Director
Department of Economics
S-641 Social and Behavioral Sciences
State University of New York
Stony Brook, New York 11794-4384
(516)632-7540

1

* This project was a collaborative effort between Institutional Research, Admissions and the Department of Economics. Essential contributions were made by Gigi Lamens and Manuel London who made the experimental use of predictive modeling an integral part of their recruitment plan, Paula Pelletier and Dave Taiclet who assisted with data management, and Mark Montgomery who provided statistical advice.

AIR

for Management Research, Policy Analysis, and Planning

This paper was presented at the Thirty-Ninth Annual Forum of the Association for Institutional Research held in Seattle, Washington, May 30-June 3, 1999.
This paper was reviewed by the AIR Forum Publications Committee and was judged to be of high quality and of interest to others concerned with the research of higher education. It has therefore been selected to be included in the ERIC Collection of AIR Forum Papers.

Dolores Vura
Editor
Air Forum Publications

USING PREDICTIVE MODELING TO TARGET STUDENT RECRUITMENT: THEORY AND PRACTICE

ABSTRACT

This paper argues that a typical use of regression models to target student recruitment efforts is theoretically unsound and may therefore be operationally inefficient, and it presents results from a controlled experiment that support this conclusion. Whereas regression models are frequently used to identify the prospective students most likely to enroll, we sought instead to identify those on whom recruitment efforts have the greatest impact. A modest increase in recruitment activity increased the enrollment of students with relatively low enrollment probabilities but did not improve the recruitment of students identified by the regression model as more likely to enroll. These results suggest that using predictive modeling in admissions to identify "hot prospects" may be inefficient. It may also be ineffective because the students most likely to enroll in a college or university are not likely to be the most desirable applicants.

USING PREDICTIVE MODELING TO TARGET STUDENT RECRUITMENT: THEORY AND PRACTICE

In the competitive market of student recruitment, college admissions offices are experimenting with the use of predictive models to increase the effectiveness of their recruitment efforts. Regression analysis is used to estimate students' probability of enrollment. Then different recruitment activities are directed at students with different enrollment probabilities. This paper argues that a typical use of predictive modeling is theoretically unsound and may therefore be operationally inefficient. To test this hypothesis and explore an alternative use of predictive modeling we designed and assessed an experimental recruitment program. The first-year results confirm our perspective and identify a valuable role for statistical modeling in recruitment management.

Theoretical Considerations

Predictive modeling is frequently used to identify the students most likely to apply or to enroll in a college or university so that admissions staff can concentrate their attention on these "hot prospects" in order to enroll more students (Gose, 1999). While this is an attractive approach it may not be an efficient one.

Consider, for example, the hypothetical responses shown in Table 1 to a recruitment initiative such as a special mailing or invitation to a campus open house. After the intervention almost all the students in Group A enroll. It increases their average enrollment probability from 80% to 85% and adds 5 students to the entering class. In Group B, the students have only a 30% chance of enrolling after the recruitment intervention, but it increases their probability by 10% and enrollment by 10 students. The students in Group A are "hot prospects" in that they are likely to enroll, but devoting admissions office efforts to their recruitment diverts resources away from the target population on which they would have the greatest impact.

Table 1. Hypothetical Recruitment Intervention

	Group A	Group B
Number of students	100	100
Probability of enrollment w/o intervention	80%	20%
Probability of enrollment with intervention	85%	30%
Yield without intervention	80	20
Yield with intervention	85	30
Effect of intervention	5	10

To be efficient, recruitment programs should be directed at prospective students wavering on the brink of an enrollment decision and most susceptible to the additional encouragement provided by recruitment efforts. Admissions resources should be targeted where they will cause the greatest increase in the probability of students' enrolling, and those may or may not be the students with the highest probability of enrollment. In the language of economics, admissions office resources will be used efficiently if they are used where they have the greatest marginal impact. It is these "fence sitters" rather than the "hot prospects" that predictive modeling should help identify.

Focusing on high-probability students may also be ineffective by directing attention away from the most institutionally desirable prospective students. It is likely that high-achieving students have more attractive alternative admissions offers to consider and therefore lower enrollment probabilities. Conversely, the "hot prospect" high-probability students are likely to have weaker academic credentials.

Though theoretically sound, targeting recruitment programs to students susceptible to persuasion is problematic in practice. Neither admissions personnel nor researchers know the efficacy of different interventions or how prospective students' response to recruitment efforts varies with their absolute enrollment likelihood. Predictive modeling can, however, be used to develop efficiently targeted recruitment programs. By providing estimates of students' pre-intervention enrollment probabilities, it can support experiments to test the marginal impact of various recruitment initiatives. The rest of this paper describes a test of the value of this approach. This project began with the development of a predictive model, used the model to select samples of students on which to test an experimental recruitment initiative, and conducted the experiment to determine its differential effect on students with different enrollment probabilities.

Prediction In Practice

Developing a predictive model of students' enrollment decisions requires selecting a study population, identifying variables likely to affect students' enrollment probabilities, developing a prediction procedure, and using the model to test the effects of experimental recruitment activities. This section describes how we completed each of these steps to develop a better understanding of our university's applicant pool and support targeted recruitment efforts.

Population

The model we developed predicts the probability that a student offered admission to our university as a full-time freshman in the fall will enroll. The focus on admitted students distinguishes this project from another common application of predictive modeling. Models can also be used to assess the likelihood that students who inquire about admission will actually complete an application. We chose instead to focus on admitted students because increasing the enrollment yield from this pool is a priority for our Admissions Office. Moreover, far more information is available about admitted students, increasing the likelihood of accurate predictions.

Predictive Variables

The selection of variables for a predictive model of students' enrollment decision depends on a combination of theoretical considerations—regarding the student characteristics likely to affect enrollment in a specific college—and practical considerations—regarding the availability of data. Our model includes four kinds of variables: demographic, academic, geographic, and behavioral (Table 2). Some are likely to be important with any student population while others may be relatively specific to our campus. Statistically insignificant variables are included because our goal is accurate prediction, and retaining all the variables increases the model's overall accuracy.

Most of the predictive variables are significant at at least the 10% level. The highly significant variables (1% level) positively related to enrollment are high-yield high school average, high-yield SAT score, high-yield math SAT, high-yield verbal SAT, high-yield zip code, and open house attendance. Dummy variables indicating whether a student's high school average and SAT scores are within ranges that historically yield a high number of enrollees are used instead of the raw values of these variables because there is no reason to believe they are linearly related to the probability of enrollment. Categorical variables with large numbers of values—like the student's high school and zip code—can only be included in the regression through a classification scheme reflecting the historical relationship between students' geographic origins and enrollment.

Variables significant at the 1% level and negatively related to enrollment include White or Hispanic ethnicity, US citizenship, regular admission status, early application and on-campus housing request. The model's insignificant variables are age, gender, Asian ethnicity, having an intended major, having an intended science major, speaking English as a native language, low family income, and residing in a high-yield zip code.

Additional variables could improve the model. For example, the behavioral variables—relatively early application and open house attendance—offer very limited information about how eager students are to attend our university, and we hope to add other indicators in the future.

The model's accuracy would almost certainly be improved by financial aid data such as the percentage of a student's financial aid need met, the types of aid offered, and the date on which the financial aid offer was mailed. However, including these variables would limit the model's usefulness because financial aid information does not become available until relatively late in the recruitment process. A model including financial aid measures could not be used to identify students' enrollment probabilities early enough to allow time for targeted recruitment efforts, and we chose not to include them.

Prediction procedures

We used logistic regression to predict students' enrollment probabilities. The theory is simple. Consider, for example, a simplified case in which the only thing known about students is whether or not they are female. If in one year 100 female students are admitted and 60 enroll, the probability of a female student's enrolling is 60%. Hence we can predict that the enrollment probability of a female student admitted in subsequent years will also be 60%. Logistic regression merely permits a large number of student characteristics to be incorporated in this type of calculation.

Three years of data are required to develop the predictive model, test its stability, and use it to analyze recruitment initiatives. We used 1996 data to develop the model by estimating the values corresponding to 60% in the simplified example. Based on those values we predicted the

Table 2. Variables in the Logistic Regression Model

Variable	Coefficient	Description
	DEMOGRAPHIC VARIABLES	
Age	-0.2285	college age: 17-19 years old
Gender	0.0448	male = 1, female = 0
White	**-0.3770*** **	**ethnicity is White**
Asian	-0.0777	ethnicity is Asian-American
Hispanic	**-0.4139*** **	**ethnicity is Hispanic-American**
Black	**-0.2705* **	**ethnicity is African-American**
Citizen	**-0.7161*** **	**United States citizen**
Permanent resident	**-0.4206* **	**US permanent resident**
English	-0.1056	English is native language
High income	**-0.1794** **	**high self-reported family income (≥$75,000)**
Low income	0.0919	low self-reported family (<$39,000)
Status	**-0.4339*** **	**1=regular admission, 0=special**
	ACADEMIC VARIABLES	
HY HS average	**0.2389*** **	**high-yield high school average**
HS missing+	**1.0382*** **	**missing high school average**
HY SAT	**0.3857*** **	**high-yield SAT combined score**
SAT missing+	**-0.9218*** **	**missing SAT score**
HY math	**0.3048*** **	**high-yield SAT Math score**
HY verbal	**0.2433*** **	**high-yield SAT Verbal score**
Major	-0.0718	application indicated an area of interest
Science major	0.1224	application indicated science interest
	GEOGRAPHIC VARIABLES	
HY HS	**0.3929*** **	**high-yield high school**
HY Zip	0.0902	high-yield zip code
NYC/LI	**0.4418*** **	**lives in NYC or Long Island**
	BEHAVIORAL VARIABLES	
Applied early	**-0.3872*** **	**applied before December 1**
Open house	**0.9706*** **	**attended an open house**
Campus housing	**-0.7676*** **	**requested on-campus housing**
CONSTANT	0.0594	

* significantly different from zero at the 10% significance level.
** significantly different from zero at the 5% significance level.
*** significantly different from zero at the 1% significance level.

\+ These variables permit cases with missing data to be included in the model. Actual scores are used where available, and an alternative coefficient is assigned to missing data.

enrollment decisions of students matriculating in fall 1997 and compared students' predicted and actual behavior to evaluate the model's accuracy. We then used the model to predict the enrollment behavior of students entering in fall 1998, with and without an experimental recruitment initiative.

Prediction results

For logistic regression there is no simple statistic measuring the accuracy of a model comparable to the R-square statistic computed for linear regressions. Instead the model's predictive power can be assessed by measuring goodness-of-fit through classification tables that compare results predicted by the model with students' actual enrollment decisions.

In order to draw this comparison it is necessary to decide what counts as a "prediction to enroll" since the regression predicts enrollment probabilities as a continuous variable. A cut-off probability level must be selected, above which a student is counted as being predicted to enroll and below which a student is counted as predicted not to enroll. The selection of this cut-off value is somewhat arbitrary. We used 0.30 because about 30% of our admitted students enroll and because we prefer the way predictions using this cut-off compare with students' actual enrollment decisions. Lower cut-off points result in lower overall accuracy while higher cut-off points substantially underestimate the number of students who enroll.

Table 3 shows the results of using the model estimated using fall 1996 data to predict fall 1997 enrollment decisions.

Table 3. Predicted and Actual Enrollment

ACTUAL BEHAVIOR	PREDICTION: Predicted to enroll	PREDICTION: Predicted not to enroll	TOTAL
Enrolled	**1,219** **16%**	917 12%	2,136 27%
Did not enroll	1,450 19%	**4,223** **54%**	5,673 73%
TOTAL	2,669 34%	5,140 66%	7,809 100%

The upper left and lower right quadrants represent correct predictions by the model: 1,219 students (16% of the total) were predicted to enroll and actually enrolled, while 4,223 students (54% of the total) were predicted not to enroll and did not enroll. The model is accurate 70% of the time, which is a significant improvement over the prediction possible without the regression model. An uninformed projection would predict enrollment correctly 27% of the time, since 27% of all admitted students actually enrolled.

Table 4 offers further assurance that the model accurately assigns probabilities to admitted students by confirming that the actual enrollment decisions of students in different predicted probability ranges corresponds to the prediction. For example, the first row shows that 9% of the students with predicted enrollment probabilities between 0% and 10% actually enrolled.

Table 4. Predicted and Actual Enrollment by Probability Range

predicted enrollment probability	percent enrolled	number of students
0%-10%	9%	1,495
10%-20%	17%	2,219
20%-30%	28%	1,426
30%-40%	39%	1,051
40%-50%	46%	768
50%-60%	51%	450
60%-70%	58%	270
70%-80%	57%	103
80%-90%	57%	23
90%-100%	0%	1
TOTAL		7,806

In ranges up to 60%, the percentage of students who actually enrolled falls within the predicted probability range. Above that there is a discrepancy between the predicted probability and the actual percentage who enrolled, but there are not many students in those ranges and more than 50% of the students in each range actually enrolled, double the overall enrollment percentage. The estimated model fits the data quite well.

These results confirm the importance to our university of recruitment initiatives targeting students with low predicted enrollment probabilities. Most of both the admitted students and the students who actually enroll have low enrollment probabilities.

Other institutions' admissions pools may display different patterns. Our large number of admitted students with low enrollment probabilities may, for instance, be attributable to a systemwide application system that makes it easy for students to apply to several campuses. The distribution is institutionally important, however, because knowing its shape has helped the Admissions Office better understand its target population.

The analysis of admitted students by enrollment probability also confirms the hypothesis that targeting recruitment efforts to the students most likely to enroll does not focus attention on those with the strongest academic credentials. Enrollment probability is inversely related to average SAT score (Table 5).

Table 5. Average SAT Score by Probability Range

predicted enrollment probability	Average SAT score	number of students
0%-10%	1250	1,495
10%-20%	1217	2,219
20%-30%	1169	1,426
30%-40%	1120	1,051
40%-50%	1058	768
50%-60%	1034	450
60%-70%	1012	270
70%-80%	1020	103
80%-90%	976	23
90%-100%	-	1
TOTAL		7,806

In this population of admitted students, targeting hot prospects does not direct recruitment efforts to high-achieving students. Effective recruitment requires attention to low-probability students. To investigate whether recruitment initiatives aimed at students with low enrollment probabilities would also be efficient, we designed a recruitment experiment.

A Recruitment Experiment

Experimental design

With enrollment growth an institutional goal, our Admissions Office was interested in identifying recruitment efforts that would increase freshman enrollment by increasing the relatively low percentage of admitted students who enroll. Specifically, Admissions wished to test the efficacy of increased contact with admitted students and their parents through activities requiring modest staff effort so that, if successful, those activities could be expanded to a large group of students.

To carry out this test, we designed an experiment to compare the enrollment decisions of students to whom the new program was directed—the experimental group—and students not affected by it—the control group. The experiment focused on students with enrollment probabilities between 30% and 60%—the middle of the distribution—because the experimental recruitment initiative was designed for use with a large number of prospective students.

The experimental and control groups are shown in Table 6. The experimental group initially included 200 students, the largest number to which the Admissions staff felt they could devote additional attention, divided into equal-size experimental groups to facilitate comparisons among the three probability ranges. This experimental group was subsequently expanded to include half the students in the highest predicted probability ranges. The control groups were all students with comparable enrollment probabilities who had been admitted at the time the samples were drawn.

Table 6. Experimental Groups

Probability range	Experimental group	Control group
30-40%	67	447
40-50%	67	178
50-60%	67	64
> 60%	125	130

Students in the experimental group received an additional invitation to visit the campus, and their parents received two special mailings. Most of these students also received expedited financial aid packaging, and as many as could be reached were contacted in a financial aid telethon. The experiment was not perfectly controlled in that it included several different interventions, some of which did not include every student in the experimental group. It was, however, sufficient to provide initial evidence of the program's efficacy.

Experimental results

The experimental program increased enrollment in the groups with relatively low enrollment probabilities. Of the students with enrollment probabilities between 30% and 50%, 42% of the experimental group enrolled compared to only 33% in the control group (Table 7), and this difference is statistically significant.

Table 7. Effects of the Recruitment Experiment

Probability of enrollment		Enrolled	Not enrolled	Number
30-50%	Experiment	42%	58%	134
p = .04	Control	33%	67%	625
50-60%	Experiment	33%	67%	67
p=.05	Control	50%	50%	64
60-90%	Experiment	49%	51%	125
p=.24	Control	56%	44%	130

In the groups with higher enrollment probabilities, the experimental program did not increase enrollment. Additional attention appears to have had no effect on students with the highest enrollment probabilities—the 60% to 90% group. In this probability range fewer students in the experimental group enrolled, but the difference between the experimental and control group is not statistically significant. In the 50-60% range fewer students in the experimental group enrolled, and the difference is statistically significant, an unexpected result that may be a product of small sample size.

While the accuracy and reliability of these results is limited by the small sample and imperfectly controlled experiment, they strongly suggest that students with relatively low enrollment probabilities are more susceptible to increased recruitment efforts. We plan to replicate and refine the experiment in future years with the hope of confirming this conclusion and determining the contribution of the different elements included in the experimental recruitment program.

Conclusion

The recruitment experiment indicates that to support enrollment growth our Admissions Office should include students with relatively low enrollment probabilities among its recruitment targets. A modest increase in the attention these students receive appears to have a significant effect on their behavior. Extending the experimental recruitment program to all students with enrollment probabilities between 30% and 50% would have increased the freshman class by about 70 students or 3%.

These results confirm the importance of the insight on which this project was based. It appears that recruitment efforts should not focus on "hot prospects," though further research with larger samples, more strictly controlled experiments, and different admissions pools is needed to verify this conclusion. Concentrating recruitment efforts on the "hot prospect" students appears to be inefficient—by diverting resources away from the population on which they have most effect—and ineffective—by focusing on students who are not the strongest candidates in the admissions pool.

The experiment also demonstrates a more general point: recruitment initiatives are relatively easy to assess. Compared to other assessment targets the outcome to be measured is simple: students either do or do not enroll. By providing a baseline prediction of students' behavior, predictive modeling can be a valuable research tool in an admissions office willing to experiment and assess the results of different recruitment activities.

Technical feasibility is not, however, the only issue in implementing assessment-based recruitment management. This approach makes significant demands on admissions staff who must have good data in a usable form and be willing to take some unusual risks. An experimental approach requires devoting resources to the recruitment of students who are unlikely to enroll. While a controlled experiment is in progress it also requires excluding some students from recruitment efforts that could increase enrollment. These are difficult actions for admissions staff under pressure to meet enrollment targets. In this context the research orientation of institutional research staff can provide encouragement while their technical expertise supports innovation. A project such as this requires active collaboration between institutional research and admissions to insure that predictive modeling is more than an academic exercise, but it can be a fruitful collaboration.

References

Gose, B. (1999). Colleges turn to consultants to shape the freshman class. Chronicle of Higher Education XLV, 35. May 7, 1999.

NOTICE

REPRODUCTION BASIS

- [x] This document is covered by a signed "Reproduction Release (Blanket) form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

- [ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").