Research in Higher Education (© 2006)
DOI: 10.1007/s11162-006-9021-8

A PREDICTIVE MODEL OF INQUIRY TO ENROLLMENT

Cullen F. Goenner*,† and Kenton Pauls**

The purpose of this paper is to build a predictive model of enrollment that provides data driven analysis to improve undergraduate recruitment efforts. We utilize an inquiry model, which examines the enrollment decisions of students that have made contact with our institution, a medium sized, public, Doctoral I university. A student, who makes an inquiry to our university such as by returning a request for information form, often provides far less information than is available from applicants. Despite this fact we find that characteristics of the student, as well as geographic and demographic data based on the student's zip code are significant predictors of enrollment. Accounting for uncertainty in our model's specification, we find that we are able to predict out of sample the enrollment decision of 89% of student inquiries. We also demonstrate how these findings can be used to improve marketing efforts.

KEY WORDS: predictive model; recruitment; geodemography; specification uncertainity.

INTRODUCTION

Colleges and universities operate in an environment in which they compete with each other to attract students. One way to increase demand for a good or service is through marketing. Marketing targeted at potential customers not only makes individuals aware of the availability of the product, but may also differentiate in a positive manner the product from those of competitors. For educational institutions such activities are increasingly relevant as recent tuition increases have left

*Department of Economics, University of North Dakota, Box 8369, Grand Forks, ND 58202, USA.

**Director of Enrollment Services, University of North Dakota, Grand Forks, ND 58202, USA.

†Address for correspondence to: Cullen F. Goenner, Department of Economics, University of North Dakota, Box 8369, Grand Forks, ND 58202, USA. E-mail: cullen.goenner@und.nodak.edu.

enrollment managers searching for ways to increase demand in order to maintain and increase enrollments. Furthermore, for a state such as North Dakota, which faces a declining number of high school graduates, the need for recruiting out of state and region is imperative. Adding to the recruitment challenge is the need not only to attract students, but those with certain characteristics. Enrollment managers are typically tasked with increasing the academic ability and ethnic diversity of the entering student body. Such lofty goals are increasingly difficult to achieve in an era in which institutions of higher education are expected to do more with fewer resources.

No matter the institution, resources are always limited, thus institutions must choose how to use their resources to best achieve their goals. With choice though often comes accountability. For instance in North Dakota a bill was passed in 2001 that allows the institutions that make up the North Dakota University System to determine the amount and use of tuition in exchange for greater accountability in reporting to the state board of higher education. Accountability requires one to report their actions and results, as well as the justification of those actions. Rather than base decisions on intuition and anecdotal evidence, data driven analysis is required to best target the use of resources in order to achieve one's goals.

In this paper we demonstrate the utility of predictive modeling in conducting data driven analysis of enrollment behavior that can be used to improve marketing efforts. Predictive modeling analyzes past data to make future predictions, or in econometric terms, data is analyzed to estimate a model which is used to make out of sample predictions. Predictive modeling as it pertains to applications in education has received little academic interest outside of the work by DesJardins (2002) and Thomas, Dawes, and Reznik (2001). One reason for this is the misperception that campuses lack the in house expertise to implement such methods without the help of outside consultants.

The framework that we adopt in this paper is in many ways similar to that of DesJardins (2002) and Thomas, Dawes, and Reznik (2001) in that we build a statistical model to predict enrollment and test our predictions out of sample. Our paper though differs in several important aspects. First, our analysis focuses on the enrollment decision of students who inquire of an institution rather than those who apply or are admitted. An inquiry may take the form of a student filling out an information card, attending a college fair, making a campus visit, sending an email, or phoning to request information. Focusing on student inquiries allows us to estimate student interest early in the recruitment process, which provides time to selectively target our marketing and

recruitment efforts based on interest and prior to students having made their enrollment decision. Inquiry models though offer several challenges as inquiry information may not be currently collected, and the data when collected contains far less information than data which is available from completed applications. Typically inquiries reveal only limited information about the prospective student, including such information as type and frequency of contact, program of interest, high school attended, home address, and whether the student was referred by an outside party.

This challenge though is not overwhelming. To counter the lack of information on factors such as ability, socio-economic status, and preferences we draw on geographic and demographic data by zip code to aid in our predictions. We borrow from geodemography by inferring demographic characteristics of a student based on the aggregate characteristics of the neighborhood in which they reside. The idea is that birds of a feather flock together, which is to say that likes tend to attract each other. In our context this implies that individuals with similar backgrounds and preferences tend to live in the same neighborhoods. This provides us with a large number of additional variables to consider for inclusion in our model specification.

A large number of theoretically interesting variables, as in our case, can create uncertainty as to the appropriate variables to include in the model's specification. *A priori* it is often theoretically difficult for researchers to select one model specification over another. This may result in different models being used, which produce coefficients, standard errors, and overall predictions that vary widely across models. In such cases one is uncertain as to the true effects of the parameters of interest. To account for this uncertainty we use techniques of Bayesian model averaging (Hoeting, Madigan, Raftery, and Volinsky, 1999), BMA, to form our predictions. BMA generates estimates that are a weighted average of the estimates from the individual models of interest, with weights determined by the support each model receives in the data. The technique not only incorporates uncertainty into one's estimates, but has also been shown to be better able to discern causal effects and predict out of sample relative to other variable selection techniques.

Our results indicate that geodemographic variables contribute to predicting the enrollment decision of students who inquire. To test the ability of our model to predict out of sample, we randomly split our data into two groups; one to build the model and the other to test the model's forecasts. Our model is able to predict enrollment out of sample correctly for 89% of inquires when a cutoff of .5 is used. The model though is better at predicting who will not enroll than who will enroll.

The ''sensitivity'' of the model, which is the model's ability to predict enrollment of students that do enroll is 36%, compared to its ''specificity'' of 97%, which is the model's ability to predict students that will not enroll and do not enroll. These findings as we discuss below can be used by decision makers to improve student recruitment.

COLLEGE CHOICE

A student's choice of college can be seen as a three stage process (Hossler and Gallagher, 1987) involving a student's predisposition for higher education, their search for potential schools, and finally their selection based on their choice set. Predisposition is influenced by encouragement from friends and family, preparation via coursework and cognitive ability. Typical ways in which students discover information about various institutions involve reviewing college guidebooks, making campus visits, talking to counselors and other students, among other activities. This process may limit the set of choices further via geographic limits and program availability. At this stage socio-economic factors tend not to play as important a role. Selection then comes down to choosing among those schools that one applied to and was admitted, which depends on ability, fit of the institution, and socio-economic status. Influencing each of these stages are economic, sociological, and psychological factors (Paulsen, 1990).

From the economic perspective enrolling in college is an investment in human capital (Becker, 1993) that offers benefits in the form of higher salaries, access to more prestigious occupations, and less likelihood of unemployment. The direct costs of college (tuition, fees, and books) are well known and have been increasing over time. These costs are influenced by the availability of financial aid as well as parental income. Indirect costs though also influence the decision to attend college. By enrolling in college, students spend time in class and studying, forgoing alternatives such as working a full time job. As unemployment of non-college graduates increases, the indirect cost of college falls. Manski and Wise (1983) find that college choice is equally sensitive to both types of costs.

College choice though is more than just an investment decision, as it is a decision to become part of a community. This community develops academic and non-academic relationships, which strengthen the individual as well as the community itself. The decision to join is influenced by the psychological desire to be in an environment that fulfills one's needs and aspirations. Characteristics that reflect the fit of the institution and the student, such as the size of the school, location, or program

offerings, are important in the decision. Furthermore, social interactions, which are dictated by societal norms, also play a critical role. Students whose parents attended college are expected to attend college as are those students whose friends attend college.

The existing empirical literature has focused primarily on the search and enrollment stages of college choice. The work of DesJardins, Dundar, and Hendel (1999) along with Weiler (1994) examines the factors that influence application behavior. Both sets of researchers take advantage of survey data provided by ACT takers in the former case and SAT takers in the latter, to focus on the fit between student and institution as it relates to the application decision. For instance, from the ACT Student Profile Questionnaire, students respond to questions that reveal information on the characteristics, such as size and type, of their preferred institution. These "fit" variables along with test scores, high school rank, age and parental income are all found by DesJardins et al. to be important to the application decision. Toutkoushian (2001) also contributes to the search literature by considering whether parental education and income influence which institutions students initially consider attending. Examining where students submit their SAT scores, Toutkoushian finds that parental education and income have little effect on the choice of schools that New Hampshire seniors consider attending. What matters most in this study is that students at this early stage tend to choose schools where their ability closely matches that of the institution.

Many of the factors that influence the decision to apply also influence the decision to enroll. Models of enrollment typically consider whether or not a student enrolls at a particular institution, while conditioning on a student's either being admitted (DesJardins, 2002; Leppel, 1993) or having applied (Bruggink and Gambhir, 1996; Thomas et al., 2001) to that institution. The choice is binary, thus a logistic model is used that controls for factors that are said to influence college choice. Variable selection differs across models, but generally reflects the academic and personal characteristics of the student in addition to variables that reflect the fit of the institution with a student's preferences. These variables may include individual factors such as gender, age, race, family income, high school size, residency status, test scores, and whether a major is indicated among other measures. Controlling for these factors, researchers then estimate the probability that an individual student, who has applied or been admitted, will enroll at the institution. DesJardins (2002) and Thomas, Dawes, and Reznik (2001) have demonstrated that these results can be used to segment students so as to more effectively target recruitment efforts.

BUILDING AN INQUIRY MODEL

The predictive model of enrollment that we develop in this paper builds on this existing literature, though largely differs in that it focuses on the decision making of students who have inquired about our institution. As noted above, college choice is a multi stage process in which inquiry is differentiated (Davis-Van Atta and Carrier, 1986; Ihlanfeldt, 1980) from post application and admission decision making. During the inquiry stage, students search for information about institutions in order to identify the handful of institutions to which they will apply for admission. A predictive model that can be utilized early in the recruitment cycle provides the admissions officer with a tool to craft recruitment and marketing efforts to solicit more applications and increase enrollment. A model that targets only applicants or admitted students, while useful, is limited to a much smaller population of students that have already made a significant gesture of interest toward our institution. Increasing the sample to include inquiries allows us to target those students who are most likely to be influenced by our recruitment efforts. As Paulsen (1990) along with Hossler and Gallagher (1987) note the exchange of information at the inquiry stage has the potential of making it the most important for institutions to increase applications and enrollment.

Data availability is the primary challenge of using an inquiry model. While institutions typically collect and have easy access to student application information, they tend not to collect information concerning student inquiries. Towards this end, enrollment services at UND created a software program to integrate prospective student information with application and enrollment data. This database allows us to keep track of the type of inquiry as well as the number of inquiries made. Included in the list of inquiries recorded are activities such as filling out an information card, attending a college fair, making a campus visit, sending an email, or phoning to request information. An additional difficulty using inquiry data is that prospective students generally provide less information than those who complete applications. For instance, in the case of a student that returns a request for information card, we may only obtain information on their high school attended, interested major, and address. The result is that for most student inquiries the researcher lacks much of the demographic data that is used in typical models that examine students that have applied or been admitted.

To overcome this difficulty we utilize tools from geodemography. Geodemography is the process of attaching demographic characteristics with geographic characteristics. The often stated notion is that birds

of a feather flock together, which is to say that people living in the same neighborhood tend to have similar characteristics (income, education, occupations, and family size) and thus behave in similar patterns. It is this predictability in behavior that marketers try to exploit by segmenting the market into clusters based on their characteristics. Aiding marketers in these activities are firms such as Claritas, which use US Census data as well as proprietary household survey data to segment the market geographically into clusters with catchy names like "Shotguns & Pickups" and "Kids & Cul-de-Sacs." Individuals in the Kids & Cul-de-Sacs cluster are described by Claritas as upscale, suburban, married couples with children. They tend to be college educated, white-collar professionals with administrative jobs and upper middle class incomes. For educational institutions this cluster is important because the combination of these characteristics result in large expenditures on education related products and services.

Despite the increasing popularity of geodemography as a marketing tool, little academic interest has been shown in its use. A search of the Web of Knowledge, Proquest, Wilson Web, Academic Search Elite, and ERIC databases revealed only one reference to the application of geodemography (Lang, Hughes, and Danielsen, 1997). One reason for the lack of academic research is that it is often perceived that such methods require the use of proprietary household data in order to get actionable results. We demonstrate though that data collected by the US Census, at the zip code level and publicly available, can provide useful generalizations of individual behavior that aid in the prediction of enrollment. Data reported by the Census is aggregated over particular geographic units (census tract, zip code, MSA) and thus we assume that the average demographic characteristics for a given zip code represent those of individuals residing in that area.[1] More specifically we assume that the college enrollment decision of individuals will be similar of that of individuals in similar neighborhoods.

The model that we develop here uses geodemography to build a predictive model of enrollment. Rather than segment individuals into geographic clusters, we use the demographic characteristics of geographic groups to aid our model's predictions. Individual demographic characteristics, which our data lack for most inquiries, are replaced with demographic characteristics from individuals' zip codes. Demographic variables from the census that we consider for inclusion are the average income, median age, college demographic, white demographic and total population for an inquiry's zip code. Even though neighborhoods divided by zip code are not composed of homogenous individuals, as income or age may differ, behavior tends to be the same throughout due

to sociological effects. Research (Ceballo, McLoyd, and Toyokawa, 2004; Datcher, 1982; Duncan, 1994; Garner and Raudenbush, 1991; Leventhal and Brooks-Gunn, 2000) has consistently shown the importance of "neighborhood effects" on educational choices. With respect to enrollment, the research of Bruggink and Gambhir (1996) showed that minority students from ethnically diverse areas were less likely than other minority students to enroll at their institution, which was located in a state with little ethnic diversity.

DATA ANALYSIS

The model that we develop predicts the probability that a student that inquires of our institution will enroll the following fall term. Our sample consists of 15,827 students that inquired into our institution and were interested in attending in the fall of 2003. Of these students 2067 actually enrolled. We model each student's decision as a binary choice, either they enroll at our institution or they do not. Given this assumption we use a logistic regression model to estimate the probability that individuals enroll, while controlling for a number of factors theoretically relevant to the decision. As discussed above, a large number of factors influence a student's decision to enroll at a particular college, and as such researchers have used various variables in their model specifications. Typically these variables include measures of a student's ability, their demographic background, interests, and fit with the institution.

Students inquire of our institution using a number of methods that we record. Controlling for a subset of these methods allows us to evaluate which types of interaction are more likely to indicate students that are interested in attending our institution. The types of interaction that we control for are whether students automatically submit their ACT scores, make a campus visit, are referred by an outside party, contact via the internet, contact via telephone, in addition to the overall number of contacts. While certain types of inquiries, such as test score submissions, yield more information about students than others, we utilize information that is available to each inquiry in our sample. This information includes the student's high school attended, their address, and their area of academic interest.

Using students' addresses we are able to consider for inclusion in our model a number of geographic and geodemographic variables. The geographic variables include measures of a student's distance to UND, distance to our closest regional competitor, attendance at a high yield high school, and whether they live in Minnesota or North Dakota.[2] Research by Leppel (1993) shows that distance is an important factor in

college choice as an increase in distance decreases interaction between the student and institution as well as increases the financial cost of enrolling. Increasing distance also increases the likelihood of more and closer institutions in a student's choice set. Along the lines of Leppel (1993) we allow the distance to UND to influence enrollment in a non-linear fashion. Four dummy variables are used that indicate breakpoints for various distances; 100–300 miles, 300–500, 500–1000, and more than 1000 miles. Those students closer than 100 miles serve as the reference group. We also consider for inclusion a variable that measures the distance in miles from our institution. Using each inquiry's zip code we merged US Census data of their neighborhood's characteristics with our variables to capture demographic characteristics. These variables included the average income, median age, college demographic, white demographic, and total population.

In addition to the above mentioned variables, we also considered for inclusion in our model a number of interaction terms. We interact campus visit with distance, distance with interest in our aviation program, income with distance, and finally income with interest in our aviation program. Special attention was given to our aviation program, given the national strength and recognition of the program as well as its much higher cost than our other programs. As such we have data on 28 variables that we will consider as candidates for inclusion in our model of enrollment. A description and summary statistics are provided in Table 1.

As is typical in empirical research, we have a wide array of control variables that we consider for inclusion in our model's specification. Theory should obviously guide our final choice, though theory, as in this case, often says very little about how one should measure these variables. This results in a large number of variables for our consideration. For instance, an individual's choice set is an important factor to choosing enrollment at an institution, but how exactly does one measure the ''choice set.'' Researchers have used a variety of measures such as demographic characteristics, family income, distance, and test scores to proxy for this effect. With theory silent as to the appropriate selection of control variables, researchers are left with the seemingly arbitrary task of choosing among variables. In such cases one is uncertain of the true model specification that generates the data as well as to the robustness of the results generated by a particular specification.

Rather than accept *a priori* a single model as the true model when several models are theoretically interesting, we utilize Bayesian model averaging to explicitly account for alternative specifications in our estimates. We assume, while we do not know the model specification that

TABLE 1. Variable Description and Summary Statistics

Predictor	Description	Mean	Std. Dev.
Contact			
contacts	Number of inquiries	1.647	1.164
autoact	1 if automatically submitted ACT score; 0 otherwise	0.3510	0.4773
visit	Number of campus visits	0.1511	0.3581
referral	1 if referred by faculty, coach, alumni; 0 otherwise	0.1053	0.3069
www	1 if inquiry made by internet; 0 otherwise	0.1037	0.3049
phone	1 if inquiry made by phone; 0 otherwise	0.0595	0.2365
Geographic			
distance	Distance in miles from our institution	396.4	437.3
hystate	Resident of MN or ND	0.6419	0.4795
hyschool	Historically high yield school	0.0965	0.2953
compete	Distance in miles to closest regional competitor	238.1	403.3
dist1	Distance between 100–300 miles	0.4731	0.4993
dist2	Distance between 300–500 miles	0.0933	0.2909
dist3	Distance between 500–1000 miles	0.1331	0.3397
dist4	Distance greater than 1000 miles	0.1149	0.3189
Geodemographic			
colldemo	% of population who completed some college	0.1606	0.0529
totalpop	Total population of zip code	16719	14897
medage	Median age of zip code	35.73	4.572
whitedem	% of population white (Non-Hispanic)	0.9104	0.1673
avginc	Average income in dollars of zip code	54803	22222
Academic			
acadint	1 if academic interest expressed; 0 otherwise	0.5570	0.4968
aviation	1 if academic interest is aviation; 0 otherwise	0.0816	0.2737
Interaction			
vismile	# of visits × distance	36.73	142.9
avitmile	Aviation × Distance	51.41	230.3
aviatinc	Aviation × Average income	30165	31372
incmile1	Average income × Distance 1	26058	31716
incmile2	Average income × Distance 2	4592	15017
incmile3	Average income × Distance 3	8021	22532
incmile4	Average income × Distance 4	7411	22512

generates the data, we know that some linear combination of the variables considered represent the true model. A Bayesian perspective provides a natural approach to uncertainty in model specification as it allows one to compare the relative evidence in favor of one model over another, rather than simply accept or reject the model as in the case of frequentist methods. For example, stepwise among other variable

selection methods choose a single model specification over others based on statistical criteria that lack a theoretical basis. Estimates from BMA are a weighted average of the estimates from the set of models of interest, with weights determined by the relative support that each receives from the data. Our inferences thus reflect a degree of uncertainty as to the true model specification.

While theoretically appealing, Bayesian model averaging also offers several practical advantages to other methods. Raftery et al. (1997) demonstrate using simulated data, where the true model is known, that BMA is better able to discern the true model specification than stepwise methods in the presence of uncertainty. Goenner and Snaith (2004) have also shown with respect to the analysis of graduation rates that this method provides a neutral way of dealing with variable selection and improves out of sample predictions relative to results based on a single model specification.

In order to implement Bayesian model averaging the researcher must choose the set of models that are theoretically relevant. As noted above, the theory of college choice provides limited guidance as to the appropriate choice of regressors, which results in uncertainty of model specification. For this reason we examine several models that consist of different linear combinations of the variables that we believe to be theoretically relevant. With k candidate variables for inclusion, the model space that we average over includes 2^k models. For each of the 2^k models in the model space, we must specify our prior beliefs with respect to the probability that each model is the true model. In the analysis below a uniform prior is used, which assumes that each of the K models is *a priori* equally likely and that $P(M_1) = \ldots P(M_K) = 1/K$. This implies that the prior probability for inclusion of each variable is 1/2. The uniform prior is typical in cases such as this that lack strong prior beliefs. Raftery (1995) notes that in large samples the choice of priors has very little influence on the posterior mean and variance of the estimates.

Combining our prior beliefs along with observing the data, we update our beliefs to obtain the posterior probability that each model is the true model that generates the data. These values are referred to as posterior model probabilities (PMP) and by Bayes' rule and the law of total probability are equal to

$$P(M_k/D) = \frac{P(D/M_k)P(M_k)}{\sum_{l=1}^{K} P(D/M_l)P(M_l)} \tag{1}$$

for each of the K models. $P(D/M_{\mathrm{k}})$ represents the likelihood of model k, and has been shown by Raftery (1995) to be approximately equal

to $\exp(-1/2\mathrm{BIC}_k)$, where BIC_k represents Schwarz's (1978) Bayesian information criterion for model k. $P(M_\mathrm{k})$ represents the prior probability that model M_k is the true model, which as discussed above is assumed to be $1/K$ for each model. Thus the posterior model probability becomes

$$P(M_k/D) \approx \frac{\exp(-\frac{1}{2}\mathrm{BIC}_k)}{\sum_{l=1}^{K}\exp(-\frac{1}{2}\mathrm{BIC}_l)} \tag{2}$$

The PMP is important because it provides information on the relative support that each of the theoretical models receives from the data. It should be noted that the model that receives the highest PMP is not necessarily the model that generates the data, for this model is assumed to be unknown.

BMA accounts for uncertainty in our estimated logistic regression coefficients, by taking a weighted average of the maximum likelihood estimates produced by each of the models. The weights are determined by the support that each model receives from the data, which is represented by the model's posterior model probability. The effects of our variables of interest on the dependent variables can then be summarized by the posterior mean and variance of the distribution. Raftery (1995) reports these values can be approximated for coefficientβ_1 as

$$\begin{aligned} E(\beta_1/D,\beta_1\neq 0) &\approx \textstyle\sum_{A_1}\hat{\beta}_1(k)P(M_k/D) \\ \mathrm{Var}(\beta_1/D,\beta_1\neq 0) &\approx \textstyle\sum_{A_1}[\mathrm{Var}(k)+\beta_1(k)^2]P(M_k/D) - E(\beta_1/D,\beta_1\neq 0)^2 \end{aligned} \tag{3}$$

where $\hat{\beta}_1(k)$ and $\mathrm{Var}(k)$ are the maximum likelihood estimates and variance of β_1 under model k, and the summation is over models that include β_1 (set A_1). Also of interest is the posterior effect probability (PEP) $\Pr[\beta_1 \neq 0/D]$, which is the posterior probability that β_1 is not equal to zero. Raftery (1995, 146) reports that the evidence in favor of a variable having an effect is weak, positive, strong, and very strong based on the breakpoints .5, .75, .95, and .99 on the probability scale.

Bayesian model averaging for logistic regression models is computed using the SPlus program BIC.logit written by Raftery and Volinsky (1996). With 28 regressors the number of models to average over in equation 3 would be nearly 27 million. In order to aid computation, the program eliminates a large number of models that receive very little support in the data, which are models that receive significantly lower posterior model probabilities than that of the highest. Inclusion of these models in our average would have little effect on our reported estimates

given the relative low weight that each would receive. The program reports the models supported by the data, their posterior model probabilities, the posterior mean and standard deviation of the coefficients, as well as the posterior effect probabilities.

RESULTS

The data indicates that there is uncertainty in the model's specification as 26 models are supported by the data, and the model that receives the highest posterior model probability receives only 21% of the total probability. The top ten model specifications appear in Table 2. Of the 26 models averaged over, eight variables did not appear in any of the model specifications, which indicates a posterior effect probability of zero. These variables included the median age, three of the distance breakpoints, three of the interaction terms between distance and income, and an interaction term between interest in aviation and income. This result is not particularly surprising given our uncertainty of interaction effects as well as the additional consideration of a non-linear relation with distance.

The signs of the coefficients are as one would expect from theory. Results appear in Table 3. Increasing the distance to our institution reduces the probability of enrolling, while increasing the distance from our regional competitors increases that probability. Students inquiring from high yield states as well as from high yield schools are also more likely to attend. Each of these geographic variables is highly significant, as indicated by posterior effect probabilities (PEP) of 100%. Somewhat surprisingly, students living between 300 and 500 miles (distance 2) are more likely to attend, though this variable receives little support (PEP 5.2%) from the data. Expressing interest in a particular major, as well as interest in the aviation program both had a positive effect on enrollment. In the case of major the result is highly significant (PEP 97.8%), whereas with aviation the result is not significant.

With respect to our inclusion of geodemographic controls, four of our five variables entered the models averaged over and two were highly statistically significant. The college demographic and average income of each inquiry's zip code both had a positive effect on the probability of enrollment. The fraction of the population in an inquiry's zip code that was white also had a positive effect on enrollment, though not statistically significant (PEP 7.9%). This indicates that the less ethnically diverse an inquiry's neighborhood the more likely they are to attend our institution. This result is not surprising given the high fraction of our state's population that is white (92%), and previous results that indicate

TABLE 2. Specifications of the 10 Models with the Highest Posterior Model Probability (PMP)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
Contact										
contacts	X	X	X	X	X	X	X	X	X	X
autoact					X			X		
visit	X	X	X	X	X	X	X	X	X	X
referral	X	X	X	X	X	X	X	X	X	X
phone		X		X	X			X	X	
Geographic										
distance	X	X	X	X	X	X	X	X	X	X
hystate	X	X	X	X	X	X	X	X	X	X
hyschool	X	X	X	X	X	X	X	X	X	X
compete	X	X	X	X	X	X	X	X	X	X
dist2										X
Geodemographic										
colldemo	X	X	X	X	X	X	X	X	X	X
whitedem							X		X	
avginc	X	X	X	X	X	X	X	X	X	X
Academic										
acadint	X	X	X	X	X	X	X	X	X	X
aviation			X	X				X		
Interaction										
vismile	X	X	X	X	X	X	X	X	X	X
avitmile	X	X			X	X	X		X	X
incmile2						X				
PMPs	0.2076	0.1714	0.1418	0.108	0.0276	0.0258	0.0248	0.0223	0.022	0.0213

likes tend to attract each other. Total population also had a positive effect, though very small posterior effect probability (4.9%).

Of the six student contact variables that we examined, each appeared in the models averaged over, with three of the six as highly significant. The number of inquiry contacts, campus visits, and whether the inquiry was referred all had positive effects on enrollment probability and posterior effect probabilities of 100%. The other contact variables, which examined phone and internet contacts as well as auto submission of act scores, had positive though statistically insignificant effects. Our interaction variables, whose inclusion we were largely uncertain of for the most part, received little support from the data. The one exception was the

TABLE 3. Results of BMA Applied to Prediction of Enrollment Full Sample

Predictor	*Mean* β/D	*Std Error* β/D	$Pr(\beta \neq 0/D)$
Contact			
contacts	0.1969	0.0299	100
autoact	0.0191	0.0690	8.3
visit	1.3386	0.0827	100
referral	1.7240	0.0745	100
www	0.0147	0.0665	5.6
phone	0.1650	0.1901	47.5
Geographic			
distance	−0.0040	0.0004	100
hystate	0.7726	0.1213	100
hyschool	0.9491	0.0819	100
compete	0.0033	0.0004	100
dist1	0	0	0
dist2	0.0155	0.0723	5.2
dist3	0	0	0
dist4	0	0	0
Geodemographic			
colldemo	2.8015	0.5395	100
totalpop	2.80E-07	1.34E-06	4.9
medage	0	0	0
whitedem	0.0578	0.2178	7.9
avginc	8.59E-06	1.49E-06	100
Academic			
acadint	0.2725	0.0803	97.8
aviation	0.1871	0.2478	38.1
Interaction			
vismile	0.0016	0.0002	100
avitmile	0.0005	0.0004	63.1
aviatinc	0	0	0
incmile1	0	0	0
incmile2	4.10E-07	1.59E-06	7.4
incmile3	0	0	0
incmile4	0	0	0

interaction between campus visit and the distance traveled, which was highly significant.

The primary interest in the model is not in the coefficients themselves but their ability together to predict enrollment outcomes. Towards this end we demonstrate our model's ability to make out of sample predictions. Keep in mind that the relevance of the model's predictions to future time periods will depend on the stability of the system and

TABLE 4. Results of BMA Applied to Prediction of Enrollment Build Sample

Predictor	*Mean* β/D	*Std Error* β/D	$Pr(\beta \neq 0/D)$
Contact			
contacts	0.1838	0.0347	100
autoact	0	0	0
visit	1.400	0.1163	100
referral	1.753	0.1042	100
www	0	0	0
phone	0	0	0
Geographic			
distance	−0.0042	0.0005	100
hystate	0.7216	0.1702	100
hyschool	1.014	0.1216	100
compete	0.0035	0.0006	100
dist1	−0.0164	0.0681	6.6
dist2	0.0078	0.0570	2.4
dist3	0	0	0
dist4	0	0	0
Geodemographic			
colldemo	1.955	1.250	77.1
totalpop	0	0	0
medage	0	0	0
whitedem	0.0328	0.1916	3.6
avginc	2.26E-06	0.00000381	27.8
Academic			
acadint	0.1156	0.1818	30.8
aviation	0.6785	0.1413	100
Interaction			
vismile	0.0017	0.0003	100
avitmile	0	0	0
aviatinc	4.64E-06	0.00000333	69.2
incmile1	0	0	0
incmile2	6.00E-07	0.00000225	8
incmile3	0	0	0
incmile4	0	0	0

require future updates as data become available. We randomly split our data in half, where one half of the data is used to build the model and the other half is used to test our model's predictions. The model built using the sub sample of data is largely the same as that from the complete data set, though averages over a smaller number of models (11). The build model's results appear in Table 4. From this one can see that

most of the highly significant variables from the full data set remained highly significant with little change in their coefficients. A number of variables that were insignificant in the full model, PEP $< .5$, became more insignificant in the build model, which is not surprising. Somewhat surprising though was the effect of inquiries from aviation students. In the build data the effect became very significant as did the interaction between aviation and income. The effect of income and interest in any major though became insignificant. These differences though, as seen below, have little impact on our model's overall predictions.

To determine the model's performance we examine its ability to correctly classify outcomes by its predictions. Logistic regression produces a prediction that is a probability, whereas the enrollment outcome is binary. In such a case one must make a determination as to what level of estimated probability predicts enrollment. As Greene (1997) discusses there is no correct cutoff to use, though the typical value is .5 due to the fact that the event, in this case enrollment, is more likely than not. The tradeoff is that in general a lower cutoff results in more students being predicted to enroll. This will increase the accuracy of inquiries predicted to enroll and who actually do enroll, but at the cost of inquires that are predicted to enroll and do not enroll (false positives). Other researchers (DesJardins, 2002; Thomas, Dawes, and Reznick, 2001) have utilized cutoffs lower than .5, which equal the fraction of students that enroll in their samples. Our choice of a .5 cutoff was guided by our intended use of the classification, which is to devote additional marketing at those inquiries that are more likely to enroll. Thus we view classifying a good candidate (true likelihood of enrolling high) as bad, to be less problematic than classifying a bad candidate (true likelihood of enrolling low) as good. The reason is that "good" candidates are likely to enroll without additional effort, whereas wasting limited resources on "bad" candidates is inefficient.

The results using the coefficients from the build model to classify observations in the testing model using a cutoff of .5 appear in Table 5. One can see that the model correctly predicts the enrollment behavior of 89.25% of inquiries. By comparison our model is relatively accurate. DesJardins' (2002) model correctly classifies 65.7% of his observations, while Thomas, Dawes and Reznik's (2001) model predicts 70% of their observations. The sensitivity of the model, which is the model's ability to predict enrollment of students that do enroll is 36%, compared to its "specificity" of 97%, which is the model's ability to predict students that will not enroll and do not enroll. This difference in predictive ability is not surprising given the model favors classification of the larger

TABLE 5. Out of Sample Predictive Performance—.5 Cutoff

	Actual Outcome		
Prediction	Enrolled	Did not Enroll	Total
Predicted to Enroll	370	194	564
	36%	2.8%	
Predicted not to Enroll	657	6693	7350
	64%	97%	
Total	1027	6887	7914

group, which in our sample are students that do not enroll (87%). The error associated with our prediction of enrollees is also due to idiosyncratic factors of individuals that influence the decision to enroll and are not captured by the model.

As noted earlier, one could lower the cutoff and increase the sensitivity, though this comes at the cost of lowering specificity and increasing the false positive rate. Lowering the cutoff to .13, which is the fraction of students that enroll, lowers the classification rate to 81% and increases the sensitivity to 78%. The false positive rate (1−specificity) increases from 2.8% to 18%.

Another way of evaluating the performance of our predictions, beyond the single specificity-sensitivity measures above, is to evaluate the relation between the two. A receiver operating characteristic (ROC) curve is a plot of the model sensitivity versus 1−specificity of the model, which is the false positive rate. The area under the curve and above a 45 degree line measure the model's ability to correctly classify those who enroll from those who do not based on their characteristics. A value of .5 indicates a model that predicts as well as chance, whereas a value of 1 indicates a model with perfect predicting power. Our model applied to the testing data generates an area under the ROC curve of .87, which is an indication of good out of sample predictive performance from our model. The corresponding measure of fit for DesJardins's (2002) model is .72.[3]

DISCUSSION AND CONCLUSION

The number of high school graduates is projected to decline over the next several years in North Dakota, Minnesota, South Dakota, Wyoming, and Montana. Thus it will become increasingly important for the University of North Dakota to attract students from outside the

region that we traditionally recruit from. As one moves further away from our institution this becomes more difficult as students are less likely to know about the academic quality and program offerings of our institution. To improve our presence outside our region requires us to leverage limited marketing and recruitment resources, where the challenge is to draw students from a much larger market in which we lack experience. Analysis of data via predictive modeling can provide insight into meeting these challenges. Our predictive model provides the ability to statistically determine who the best prospects are that inquire of our institution and locate geographic areas to target our marketing efforts.

Based on our model's estimates, we can determine the probability that inquiring students will actually enroll given their characteristics and actions. This allows us to identify individuals that may be influenced by additional recruitment efforts. One might think it would be prudent to exclude those with high or low scores, given these students are either likely to enroll or not enroll without additional effort. Our model though as discussed is more accurate at predicting who will not enroll than who will. Therefore our strategy is to focus additional recruitment efforts on those students who receive higher model scores. With the focus on a smaller segment of the population, tools such as the direct calling of prospective students becomes feasible and potentially influential on enrollment outcomes. Measuring the effects of such treatment though is inherently difficult and left for future research.

One way that UND expects to increase enrollment is by extending its traditional marketing campaign to areas outside our region. National marketing campaigns though are costly and our resources are limited. To leverage these resources, our institution has used our predictive model to segment the market of prospective students by where they live. The idea is to identify geographic areas in which to concentrate our recruitment efforts. The results of our model indicate geographic and demographic factors that contribute to student's enrolling at our institution. We then use these results to score each zip code in the United States. Of recent enrollees, we find that 70% fall in zip codes that score in the highest 15% of all zip codes. The results indicate that students in some states appear particularly worth targeting. For example, 83% of enrolled students from the state of Washington fell within the top scoring zip codes. Coupled with the fact that Washington is projected to have an increasing number of high school graduates has led UND to identify this state as a target for additional recruitment efforts.

From the perspective of an enrollment manager these results then influence where marketing and recruitment efforts are directed, such as the choice of which college fairs to attend, high schools to visit, and

where to send direct mailings. At UND our results were used in the summer of 2005 to guide the purchase of names of prospective students, who would be sent promotional material. Rather than buying names randomly from a national or state list as in the past, we were able to focus our purchase on names of juniors from high scoring zip codes. The expected benefit is to achieve a higher enrollment yield, and allow us to target our efforts in a larger number of states. Given the recent implementation of our model, the effect of this action will not be seen until the fall of 2006 enrollment numbers.

Implementing our predictive model presents a few challenges, none of which are overwhelming. Building an inquiry model requires data to be collected from inquiries. In addition we estimate our model using Bayesian model averaging, which requires statistical software that is not typically used by institutional researchers. The advantage of BMA relative to other variable selection techniques, such as stepwise methods, is that rather than base our predictions on a single model specification, we incorporate uncertainty in specification by averaging over the estimates of several models that are supported by the data.

While we recommend accounting for uncertainty, researchers motivated by our findings may build similar models using standard techniques that rely on a single model specification. In such cases we encourage researchers to check the robustness of their predicted coefficients and probabilities to changes in model specification. Analysis of our data indicated that the model chosen by stepwise selection resulted in classifications that were similar to those provided from BMA, but that the *p*-values of the coefficients tended to suggest significant effects for several variables that were not supported by BMA. This latter result is not surprising given Raftery's (1995) findings that stepwise methods and their reported *p*-values tend to overstate the evidence for a variable having an effect relative to BMA, when by construction that variable is known not to have an effect. No matter which method is used, an important aspect of using the model in the future will be to rebuild the model to account for new data and also account for any significant change in policy, such as tuition reciprocity agreements. In the case of BMA, estimates are more robust to change though require updating for the best performance.

A challenge of using our results is the timing between when students make inquiries and the decisions that enrollment managers make. Students inquire throughout the year, thus an inquiry's model score changes as an individual makes additional contacts. For this reason we constantly update model scores so that enrollment managers can act on the most current information.

Our paper has shown that by collecting data from student inquiries one is able to develop a predictive model of enrollment. We supplemented the limited information obtained from our inquiries with geodemographic data from the inquiry's zip code. Accounting for model uncertainty we estimated a model that was able to correctly classify out of sample 89% of our observations. The model allows us to identify geodemographic and other characteristics that contribute to student's enrolling at our institution. Armed with this information, enrollment managers may then target their recruitment efforts at students and geographic areas that are most likely to increase enrollment.

END NOTES

1. More refined data are available at the census tract level, though this requires one to transform street addresses to census tracts, which can be done for a cost.
2. Our regional competitors include U of MN Crookston, U of MN Twin Cities, U of MN Duluth, Bemidji State University, Minot State University, NDSU, Minnesota State Mankato, Bismarck State University, U of MN Morris, U of SD, SDSU, and St. Cloud State University.
3. The Hosmer-Lemeshow (1989, p. 141) test statistic as a measure of goodness of fit is inappropriate when using Bayesian model averaging as it assumes that the fitted logistic regression model is the correct model, while we assume the correct model is unknown.

REFERENCES

Becker, G. S. (1993). *Human Capital: A Theoretical and Empirical Analysis With Special Reference to Education*, University of Chicago Press, Chicago.

Bruggink, T. H., and Gambhir, V. (1996). Statistical models for college admission and enrollment: A case study for a selective liberal arts college. *Research in Higher Education* 37(2): 221–240.

Ceballo, R., McLoyd, V. C., and Toyokawa, T. (2004). The influences of neighborhood quality on adolescents' educational values and school effort. *Journal of Adolescent Research* 19(6): 716–739.

Davis-Van Atta, D. L., and Carrier, S. C. (1986). Using the institutional research office. In: Hossler, D. (ed.): *Managing College Enrollments*, New Directions for Higher Education No. 53, Jossey-Bass, San Francisco.

Datcher, L. (1982). Effects of community and family background on achievement. *Review of Economics and Statistics* 64(1): 32–41.

DesJardins, S. L. (2002). An analytic strategy to assist institutional recruitment and marketing efforts. *Research in Higher Education* 43(5): 531–553.

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

Duncan, G. J. (1994). Families and neighbors as sources of disadvantage in the schooling decisions of white and black-adolescents. *American Journal of Education* 103(1): 20–53.

Garner, C. L., and Raudenbush, S. W. (1991). Neighborhood effects on educational attainment – A multilevel analysis. *Sociology of Education* 64(4): 251–262.

Goenner, C. F., and Snaith, S. M. (2004). Accounting for model uncertainty in the prediction of university graduation rates. *Research in Higher Education* 45(1): 25–41.

Greene, W. H. (1997). *Econometric Analysis*, Prentice Hall, Upper Saddle River, NJ.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science* 14(4): 382–401.

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

Hossler, D., and Gallagher, K. S. (1987). Studying student college choice: A three phase model and the implications for policymakers. *College and University* 62(3): 207–221.

Ihlanfeldt, W. (1980). *Achieving Optimal Enrollments and Tuition Revenues: A Guide to Modern Methods of Market Research, Student Recruitment, and Institutional Pricing*, Jossey-Bass, San Francisco.

Lang, R. E., Hughes, J. W., and Danielson, K. A (1997). Targeting the suburban urbanites: marketing central-city housing. *Housing Policy Debate* 8(2): 437–470.

Leppel, K. (1993). Logit estimation of a gravity model of the college enrollment decision. *Research in Higher Education* 34(4): 387–398.

Leventhal, T., and Brooks-Gunn, J. (2000). The neighborhoods they live in: The effects of neighborhood residence on child and adolescent outcomes. *Psychological Bulletin* 126(2): 309–337.

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 No. 6. Washington, D.C.: The George Washington University, School of Education and Human Development.

Raftery, A. E. (1995). Bayesian Model Selection in social research. In: Marsden, P. V. (ed.): *Sociological Methodology 1995*, Blackwells Publishers, Cambridge, MA, pp. 111–163.

Raftery, A. E. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* 92(437): 179–191.

Raftery, A. E., and Volinsky C. T. (1996). Splus function Biclogit, version 2.0. (http://www.research.att.com/~volinsky/bma.html).

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* 6: 461–464.

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

Toutkoushian, R. K. (2001). Do parental income and education attainment affect the initial choices of New Hampshire's college-bound students?'. *Economics of Education Review* 20(3): 245–262.

Weiler, W. C. (1994). Transition from consideration of a college to the decision to apply'. *Research in Higher Education* 35(6): 631–646.

Received June 22, 2005.